Exploration of the Topology of Chemical Spaces with Network Measures

Michael P. Krein and N. Sukumar[*]

Rensselaer Exploratory Center for Cheminformatics Research, and Department of Chemistry & Chemical Biology, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, New York 12180

AUTHOR EMAIL ADDRESS (nagams@rpi.edu)

RECEIVED DATE (to be automatically inserted after your manuscript is accepted if required according to the journal that you are submitting your paper to)

CORRESPONDING AUTHOR FOOTNOTE (nagams@rpi.edu).

ABSTRACT

Discontinuous changes in molecular structure (resulting from continuous transformations of molecular coordinates) lead to changes in chemical properties and biological activities that chemists attempt to describe through structure-activity or structure-property relationships (QSAR/QSPR). Such relationships are commonly envisioned in a continuous high-dimensional space of numerical descriptors, referred to as chemistry space. The choice of descriptors defining coordinates within chemistry space and the choice of similarity metric thus influence the partitioning of this space into regions corresponding to local structural similarity. These are the regions (known as domains of applicability) most likely to be successfully modeled by a structure-activity relationship. In this work the network topology and scaling relationships of chemistry spaces are first investigated independent of a specific biological activity. Chemistry spaces studied include the ZINC dataset, a qHTS PubChem bioassay, as well as the space of protein binding sites from the PDB. The characteristics of these networks are compared and contrasted with those of the bioassay SALI sub-network, which maps discontinuities or cliffs in the structure-activity landscape. Mapping the locations of activity cliffs and comparing the global characteristics of SALI sub-networks with those of the underlying chemistry space networks generated using different representations, can guide the choice of a better representation. A higher local density of SALI edges with a particular representation indicates a more challenging structure-activity relationship using that fingerprint in that region of chemistry space.

KEYWORDS. Graph theory, Network topology, QSAR, Chemistry Space, Molecular Similarity, Descriptors, protein-binding sites, atom types

1. Introduction

Graph theoretic concepts have a long history in chemistry, predating quantum mechanics and our modern conception of molecular structure. Early chemical formulas reflected only stoichiometry, with no underlying structural hypothesis. The gradual acceptance of the idea of molecules as real, physical objects in 3-D space led to the concepts of molecular structure and chemical bonding. From a very general standpoint, a bond is merely a connection between a pair of atoms. The detailed nature of the connection could be immaterial, provided there was general agreement on the convention used to define a bond. Thus one could define a bond on the basis of energetic criteria, based on a Cartesian distance cut-off between a pair of atoms, by counting valencies, or by some other algorithm. Richard Bader's theory of Atoms in Molecules[1] exploits the topology of the scalar electron density $\rho(r)$ field and its associated gradient vector field $\nabla\rho$ to define the presence or absence of a bond path between any pair of atoms. A bond path is defined as a trajectory of the gradient vector field $\nabla\rho$ connecting two atomic nuclei; surfaces satisfying the zero-flux criterion:

$$\nabla\rho \cdot n = 0 \tag{1}$$

define atomic boundaries, n being the surface normal. The network of bond paths then defines a molecular graph. The nodes or vertices of the molecular graph are atomic nuclei and the connections between them represent chemical bonds between the atoms. Each of the different definitions of connectivity could result in a slightly different bond network topology in specific cases, but the broad agreement between different methods demonstrates the robustness of the concepts of the chemical bond and the molecular graph derived from it.

Molecular descriptors derived from the molecular graph - the so-called topological descriptors[2-5] have had a rich history in chemistry, lending themselves to simple visualization and ease of computation. An abstract space of such descriptors (known as “chemistry space”) is often used to cluster molecules into similarity classes. Such representations have been successfully applied[6-10] to many drug discovery and materials design problems. Several kinds of networks can be defined within this chemistry space.

Depicting individual molecules as nodes within this space, one can calculate pairwise distances between nodes and employ distance cut-offs to define the presence or absence of a connection between any pair of molecules, thereby generating a network representation of a chemistry space[11]. A similarity network in chemistry space thus consists of a similarity relationship (connections) between individual molecules (forming the nodes of the network). Any characterization of similarity depends both upon the chemical-space (molecular descriptor) representation and upon the similarity assessment metric (distance measure) employed. In Section 2, we present some simple network measures and the similarity metrics used in this work, before exploring the network characteristics of chemistry spaces in Section 3.

Chemistry is a science of not just molecular properties, but of molecular transformations and reactions. Molecular transformations are described within Bader's theory of Atoms in Molecules[1] by partitioning nuclear configuration space into regions corresponding to distinct molecular graphs. Continuous transformations of molecular coordinates can lead to discontinuous changes in the molecular graph (through bond breaking and/or bond formation) that are described by catastrophe theory[12]. Such changes describe not only transformations between isomers corresponding to the same stoichiometric formula, but also to dissociation, association, substitution and elimination reactions. Depicting individual molecules as nodes, a network of molecular transformations can be constructed, with the connections between molecules represented by some measure of their transformability, i.e., some characteristic (thermodynamic, kinetic or heuristic, such as synthetic accessibility rules[13]) measure of the reaction connecting them. Of course, the two networks discussed above (the molecular similarity network and the molecular transformation network) are not unrelated, since the transformability between molecules is measured by the similarity between their scaffolds. Topological descriptors, being based upon properties of molecular graphs, provide a measure of the similarity between molecular scaffolds. A network graph constructed from such a similarity measure based on the Atomtyper algorithm[14] is presented in Section 2.2, and its scaling properties discussed in Section 4 .

The practical importance of molecular similarity measures[15] for drug design is summarized in the similarity principle, namely that similar molecules should exhibit similar activities in biological assays[16,17.] This principle constitutes a fundamental assumption implicit in most quantitative structure-activity relationship (QSAR) modeling. While such correlations have indeed been observed for simple physicochemical properties, very similar molecules may still exhibit very different activities in some assays, giving rise to so-called "activity cliffs"[16,17], and leading to deviations from the similarity principle. Such deviations arise on account of the complex nature of the activity landscape associated with biological assays. Measures such as the Structure-Activity Landscape Index (SALI)[18] and Structure–Activity Relationship Indices (SARI)[11,19] have been devised to characterize activity cliffs. Activity cliffs may be visualized as heat maps[20] or as network graphs[18,21] highlighting abrupt changes in biological activity associated with the steepest cliffs. These are the most interesting regions of a structure-activity relationship for purposes of drug design, but are also the most difficult regions to model quantitatively in a structure-activity relationship. The construction of such a SALI network graph for a high-throughput screening bioassay is described in Section 3.4 and contrasted with the parent graph (constructed based on fingerprint similarities, independent of a biological activity) in Section 4.

2. Network measures and similarity metrics

The most elementary characteristic of a node is its degree, which specifies the number of links between it and other nodes[22-27]. For a directed network, one can distinguish between the in-degree (the number of connections leading into a node from other nodes of the network) and the out-degree (the number of connections from a node to other nodes) of each node[28]. The degree distribution P(k) is the probability that a specified node has exactly k links. The (local) clustering coefficient is given by:

$$C_i = \frac{2n_i}{k(k-1)} \qquad (2),$$

where n_i is the number of links connecting the k neighbors of node i to each other. A global measure is the transitivity or average clustering coefficient C(k) of all nodes with k links. Assortativity is a preference for a network's nodes to attach to other similar nodes[29,30]. The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes; positive values signify correlation between nodes of similar degree, and negative values correlation between nodes of different degree. Assortativity is also measured by the neighbor connectivity, or the average degree of neighbors of a node. If the neighbor connectivity increases as a function of the degree of the node, nodes of high degree connect, on average, to nodes of high degree, and the network is assortative; if the neighbor connectivity decreases as a function of the degree, nodes of high degree tend to connect to nodes of lower degree, and the network is dissortative.

Three classes of networks have been characterized[22-27]: In a random or Erdös-Renyi network[31], the node degrees follow a Poisson distribution, indicating that most nodes have approximately the same number of links (close to the average degree). The tail of the degree distribution of a random network decreases exponentially $P(k) \sim e^{-k}$ with the degree k, indicating that nodes that significantly deviate from the average are extremely rare, and the mean path length is proportional to the logarithm of the network size, indicating a small-world property. Scale-free networks are characterized by a power-law degree distribution: the probability that a node has k links follows $P(k) \sim k^{-\gamma}$ (seen as a straight line on a log–log plot). The properties of a scale-free network are often determined by a relatively small number of highly connected nodes (hubs); the average path length of such a network follows a log log N distribution (where N is the number of nodes), which is substantially shorter than the log N behavior of a random small-world network. The third class, that of hierarchical networks, is found when clusters combine in an iterative manner where communication between highly clustered neighborhoods is maintained by a few hubs, leading to coexistence of modularity, local clustering and scale-free topology,. Hierarchical modularity is characterized by the clustering coefficient scaling as $C(k) \sim k^{-1}$ (a straight line of slope -1 on a log–log plot).

Disabling a substantial number of nodes in a random network leads to fragmentation of the network into small, disconnected islands of nodes. Scale-free networks, in the other hand, are characterized by topological robustness[22-27]; they do not have a critical threshold for disintegration, and are thus robust against accidental failures. Random failure affects mainly the numerous small degree nodes without disrupting the network's integrity, but this reliance on hubs induces vulnerability to targeted attack against a few key hubs in a scale-free network. Hubs in chemistry space are represented by molecules with high leverage in structure-activity relationships. Inclusion of such molecules would be important for maintaining the diversity[32-34] of a chemical library and ensuring good predictive performance of QSAR models across a wide domain of applicability. This ability to identify multiple diverse structures (spanning very different bond frameworks or structural scaffolds) with similar activities is referred to as scaffold hopping[10,35-37], and has tremendous implications to both drug and materials design.

2.1 Chemical fingerprints

Chemical fingerprints are descriptors that encode the presence or absence of specific structural patterns in the target molecule. Representation of these chemical features as bit strings allows for rapid comparison between molecules. Due to fast calculation and comparison, many fingerprints and fingerprint encoding schemas have been devised and their use in virtual high-thoughput screening of chemical databases is well established. [38,49]

2.2 Atomtyper Levels

A distance measure was also calculated using Atomtyper, an algorithm used for identification of atom types in the Transferable Atom Equivalent (TAE) RECON method[14,40-42] and for generation of the TAE library of atomic electron density fragments. The RECON method exploits the approximate transferability of topological fragment properties for high-throughput computation of molecular descriptors[42] when the fragments are defined using Bader's zero-flux criterion[1].

Atom types in Atomtyper are defined using several criteria, listed here in descending order of priority:

1. Element type or atomic number,
2. Coordination number (number of other atoms connected to the atom in question),
3. Atomic numbers and coordination numbers of (up to four) bonded neighbors,
4. Size of the ring system, if any, containing the atom,
5. Atomic numbers and coordination numbers of next-nearest neighbors for mono-coordinate atoms.

Atomtyper employs a sequential fallback procedure, using the best available representation for each atom (closest match found in the TAE library): a requested atom type is compared successively to each atom type in the TAE library, via string comparison to entries in a sorted TAE list file[14] until the library atom type string with the best match is found; this atom type from the TAE library is then used to represent the requested atom in the molecule. The Atomtyper algorithm thus categorizes the atom types in each molecule and determines a match level based on how well these atoms are represented in the TAE library. The match levels used in Atomtyper and their interpretation are shown in Table 1. The atoms in any molecule, library or database thus span a chemistry subspace that may be used to represent the atoms in any other molecule, library or database. The match level represents a similarity measure or distance (the "Atomtyper distance") between two molecules or two subspaces, which is then used as a cut-off to construct a network graph. The discrete "Atomtyper distance" from molecule A to molecule B is defined as the match level within which all atoms in molecule A can be represented by the set of atom types of molecule B. Note that this does not imply that all atom types in molecule B will be represented by the atom types of molecule A at the same level, and thus this similarity measure leads to a directed graph. Pairwise Atomtyper level matches were computed for washed structures. Level 3, level 2, and level 1 matches were used directly as edge lists. Molecule pairs with match levels equal to or exceeding the threshold were connected by a directed edge of the network graph. Lower levels of match lead to networks that are too densely connected for meaningful analysis.

2.3 USR shape signatures

The Ultrafast Shape Recognition (USR) algorithm[41] is a pure shape molecular similarity measure that uses the molecular centroid (ctd), the closest atom to the centroid (cst), the farthest atom from the centroid (fct) and the farthest atom from the fct (ftf), to compute the first, second and third moments with respect to these points. These 12 moments constitute a compact molecular shape fingerprint that is alignment-independent, extremely fast to compute and performs well at shape classification.

2.4 PESD signatures

Property-Encoded Shape Distribution[44] (PESD) signatures account for the distribution of polar and apolar regions, as well as electrostatic potential, on the surface of the protein binding site. Pairs of property-encoded points on this surface are chosen randomly from the surface, many times. For each pair of points, the properties and distance information is recorded and then binned according to a coarse binning scheme. The result is a histogram of property and distance information, a PESD signature. As they contain chemically-relevant information beyond pure shape, PESD signatures are potentially more useful for drug repositioning.[45,46]

2.5 SALI

Guha and Van Drie[18] defined the Structure-Activity Landscape Index (SALI):

$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)} \tag{3}$$

as a quantitative measure of activity cliffs in chemical models of biological activity, where A_i and A_j are the activities of the i^th and the j^th molecules, and sim(i,j) is the similarity coefficient between the two molecules. Steep activity cliffs in a data set are associated with high SALI values. Utilizing a cut-off value of the index enables one to represent sets of molecules through network graphs, with SALI edges between pairs of molecules highlighting abrupt changes in response associated with the steepest (most significant) cliffs. In order to assess QSAR models and modeling protocols, Guha and Van Drie[21] also defined the SALI curve, a plot of the SALI value at a given similarity threshold versus the value of the similarity threshold. While the SALI network graph orders each pair of molecules by activity, the SALI curve tallies how many of these orderings a model is able to predict.

3. Network topology of chemistry spaces

3.1 Network topology of a qHTS PubChem Bioassay

The Molecular Libraries Initiative from the National Institute of Health brought large volumes of quantitative high throughput assay results and methodologies into the public domain during the past decade. However, varying descriptor choices, similarity measures and modeling methods demonstrate that efficient exploitation of large amounts of often noisy quantitative High Throughput Screening (qHTS) data is far from a solved problem. Chemical structures for compounds in the qHTS PubChem Bioassay 361[47] were downloaded from PubChem[48]. These 51,415 compounds were preprocessed using the sdwash function in Molecular Operating Environment 2008.10[49]: salts were removed and neutral forms of protonation states were output with explicit hydrogens. From these structures, various chemical fingerprints, namely CDK and CDK extended, graph only, MACCS[38], E-State atom type, and PubChem fingerprints were computed via version 3.0.4 of rCDK[39] in R version 2.11.1[50]. Pairwise Tanimoto similarities of all molecules were computed for each set of fingerprints, and the corresponding edge lists were created for each fingerprint based on all pairs of molecules similar within a 70% threshold. Subsets of the edge lists were then computed based on tighter cutoffs: 75, 80, and 85% Tanimoto similarity, respectively. Atomtyper descriptors were calculated as described above, and USR descriptors were calculated using the software package *RECON*[14]. As USR is sensitive to molecular conformation, these descriptors were computed from minimum energy structures, as calculated in MOE, using the MMFF94x forcefield and default database energy minimization options. Euclidean distance cutoffs of 0.1 and 0.05 were chosen to form the USR edge lists.

All networks studied and their computed properties are listed in Table 2. In Figures 1, 2 and 3 are shown the degree distributions, neighbor connectivity distributions, and clustering coefficients respectively, for networks constructed from the AID361 dataset with a variety of chemical fingerprints at different thresholds. The assortative behavior of the networks can be clearly discerned. The degree and neighbor connectivity distributions with respect to USR signatures are shown in Figures 4 and 5, respectively; and the degree distribution, neighbor connectivity distribution, and clustering coefficient constructed with Atomtyper distances as shown in Figures 6, 7 and 8, respectively.

3.2 Network topology of the ZINC dataset

Euclidean similarity distances were computed for all pairs of molecules in the ZINC database[51] using the average surface electrostatic potential[52,53] and the local surface average ionization potential[54,55] from RECON[40-42], and using the logarithm of the octanol/water partition coefficient, the molecular refractivity and the polar surface area from Open Babel[56]. This is a database of several million commercially-available compounds for virtual screening (2,499,518 molecules were successfully processed through RECON and Open Babel). The degree distributions of the networks so obtained followed similar patterns as for the qHTS Bioassay in the previous section, and are shown in Figure S1 in the Supplementary Information. The degree distributions of the networks obtained using these distance measures and using USR signatures from the natural products subset of the ZINC database also followed similar trends, and are shown in Figure S2 in the Supplementary Information. Additionally, Atomtyper level matches were also computed, which were then used to generate a network graph at match levels 1, 2 and 3. Figure 9 shows the out-degree distribution of the ZINC database network constructed with Atomtyper distances, at match levels 1, 2 and 3. The linear behavior of the tail of the degree distribution on the log-log plot is apparent. This small world behavior of chemistry spaces has been noted in earlier work[57,58] but never before studied on such a scale as in the present investigation.

3.3 Network topology of protein binding sites

Protein networks have been extensively studied for several years from a wide variety of perspectives, including those of drug design and repositioning.[59,60] To examine the properties of a protein binding site network, the procedure of Das *et al.*[46] was followed to determine a subset of binding sites of all X-ray crystal structures obtained from the PDB as of October 30, 2009, including hemes. Structures were separated into protein and ligand segments via Molecular Operating Environment. Protein side chains were protonated at a pH of 7.0, and Gauss-Connolly surfaces of the protein were generated within 4.5 Å of the ligand. Ligands smaller than five heavy atoms were not considered. In total, 108,089 binding sites were available for comparison. USR-type pure-shape moment descriptors were calculated from the verticies of the triangulated surface of the binding site, as defined above. These descriptors were than standardized (mean centered, and divided by their standard deviation). PESD signatures were computed for each binding site with default settings.

Pairwise chi-squared[44] and Euclidean distances were calculated for PESD and USR descriptors, respectively. For PESD signatures, edge lists were calculated based on chi-squared distance cutoffs of 11,000 to 3000, in decrements of 1000. Edge lists at distance cutoffs of 2500, 2000 and 1500 were also computed. For USR descriptors, edge lists were created based on Euclidean distance cutoffs of 0.5, 0.4, 0.3, 0.2, and 0.1. Network properties and their plots were computed in R version 2.11.1 and the package igraph 0.5.4-1[61].

The degree and neighbor connectivity distributions of the protein binding site network using PESD signatures are shown in Figures 10 and 11, respectively. The degree and neighbor connectivity distributions with USR descriptors are shown in Figures 12 and 13, respectively.

3.4 SALI sub-network topology of a qHTS PubChem bioassay

The utility of visualizing and understanding the nature of structure-activity relationships through network measures has already been documented[18-21]. Display and examination of a network graph (and overlaid activity information) enables rapid assessment of the properties of chemical space near regions of interest, such as a known-active query molecule. For PubChem bioassay 361, the composite activity measure *RankScore* was used as the activity.

In comparing SALI sub-networks of the CDK, MACCS, and PubChem fingerprints, a cutoff of 95% of nonzero values was chosen as the SALI cutoff. Edges with SALI values greater than that 95% cutoff are hereafter called SALI edges. The physical layout was determined by using the force-based Fruchterman-Reingold method[62] for each parent (non-SALI) graph (with default settings in igraph). The Fruchterman-Reingold method gives equal weight to each vertex (molecule). Molecules with more connections and larger total neighborhoods thus tend to cluster near the center of the graph, while low degree nodes migrate towards the periphery. SALI edges were then overlaid on the original graph, so as to directly observe activity cliffs and their distribution in the network, providing complementary information to traditional plots of chemistry space, such as PCA plots. Figure 14(a) illustrates this for the 85% Tanimoto cutoff PubChem fingerprint network, where the SALI edges are shown thicker (red online). In figure 14(b), these edges are plotted by themselves. The topology of the SALI sub-network is observed to be grossly different from that of the parent network, which is borne out in the network statistics of all SALI networks, shown in Table 3. These differences reflect the ability of each structural representation to describe a unique mapping of a dataset in chemical space. A high degree of clustering of SALI edges in the network would thus reflect poor representation of that region of chemical space by that fingerprint with respect to activity, while sparsity would reflect inadequate coverage of the space. By examining the immediate neighborhood (closest connections) around targets of interest, multiple fingerprints can be compared in their ability to resolve those spaces. In Figures 15 and 16 this information is presented in the form of CDK, MACCS, and PubChem nearest-neighbor networks with SALI edges for molecules CID 893460 and 749132, respectively, that are reported active in the qHTS bioassay. The larger percentage of SALI edges present in the MACCS graph is an indication that greater care and attention should be taken in the construction of QSARs in these neighborhoods. The corresponding next-nearest-neighbor networks with the different fingerprints and SALI edges for molecules CID 893460 and 749132, are shown in shown in Figure S3 and S4, respectively, in the Supplementary Information.

Table 4 summarizes this information for CDK, MACCS, and PubChem fingerprints and USR descriptors for the AID361 assay. The data were median averaged over the full dataset and also over the top 200 actives that exist in all fingerprint comparison networks. This allows for a direct comparison of the "smoothness" of a representation for similarity measures, especially in representing the active molecules, the "hits" in the AID361 bioassay. The smoothness of representation in USR descriptor space indicates that the activity has a strong shape dependence.

4. Discussion and Conclusions

In this work we have investigated the network topology and scaling relationships of several chemistry spaces. The degree distributions for all the spaces investigated (Figures 1, 4, 6, 9, 10, 12, S1 and S2) show a qualitatively similar behavior. The large ZINC dataset with Atomtyper similarity measures seems to follow a clear power law out-degree distribution (Figure 9), as evidenced by the linear tail on the log-log plot (with a power law exponent approximately 1.5), indicating the small-world nature of the corresponding network. The small world behavior of chemistry spaces has been noted in earlier work[57,58] but never before verified on as large a scale as in the present investigation. The average Atomtyper out-degrees of nodes are two or more orders of magnitude smaller than the maximum out-degrees found in the corresponding networks, as seen from Table 2(a), further supporting the small-world nature of the Atomtyper networks.

Table 2(a) also shows that for the PubChem AID361 bioassay, the undirected networks constructed from any of the fingerprints, including the USR signatures, have positive global assortativity and transitivity. Figures 3 and 8 for the PubChem AID361 bioassay, using different similarity measures, show a bilinear behavior, with a clear decrease in the clustering coefficient C(k) with the degree k at high degree, although the transitivity data for this small dataset are too noisy to fit a power law with a great deal of confidence. The assortative nature of these networks is further evidenced by the nearest neighbor degree distributions in Figures 2 and 5, showing that nodes of high degree are more similar to (share an edge with) other high-degree nodes.

The extremely assortative behavior (and the high values of global assortativity and transitivity) of the PESD protein binding networks can be clearly discerned in figure 11, and rationalized: since the binding site network is redundant and not pruned, *all valid binding sites* are included, including those from oligomeric proteins and from proteins in wild-type vs. mutant studies. Thus, binding sites with self-similar structures occur quite frequently, leading to tight clusters of very similar protein binding sites. As expected, very tight cutoffs eliminate all but these pairs to form the network, revealing even greater global assortativity and transitivity. This behavior of the network mimics social and semantic networks, to some degree - which can be rationalized by noting that the PDB grows by deposition of experimentally-determined structures, rather than with the explicit goal of maximizing diversity. Elimination of these redundancies would be expected to lead to more disassortative behavior, as has been observed in protein-protein interaction networks[63,64]. The global properties of the USR protein binding site networks stand out in contrast to that of the PESD networks, due to higher density of edges, although the trends of increasing assortativity with tighter cutoffs remain. As the USR networks are much more densely connected, shape moments alone are a much less specific discriminant of similarity between binding sites than either shape or chemical environment. These apparent subtleties in dataset curation and choice of fingerprint will thus have a drastic impact on the performance of QSAR models. Differences in the network properties of binding site networks can thus help in the choice of an appropriate representation for resolving similar binding sites.

Differences in the characteristics of biological and chemical networks are of interest because the structure-activity landscapes associated with many biological assays are not smooth and often not even continuous. Activity cliffs, or discontinuous changes in biological activity resulting from small changes in the structural scaffold or descriptor space, lead to break down of simple QSAR models in their vicinity. There have been many attempts to assess and extend the applicability domains[6,65] of QSAR models, to design molecular libraries for chemical diversity and to develop models capable of scaffold hopping[10,66-68] across multiple structural motifs. However, diversity assessed via various fingerprints or descriptors lead to differing molecular libraries and molecular networks with different topological characteristics, as compared to each other and as compared to biological networks. This difference in the representation of biological networks and the networks of commonly used chemical libraries is a reason for encountering activity cliffs. Descriptors based on local molecular surface properties have been recommended[67,68] instead as a means to develop more general QSAR models capable of scaffold hopping. The use of unconventional similarity measures such as Atomtyper match levels, which lead to directed network graphs with different characteristics, as well as USR shape and PESD surface signatures to develop models favoring scaffold hopping, is currently under investigation.

Mapping out the locations of activity cliffs for different fingerprint representations, and comparing the global characteristics of SALI sub-networks with those of the underlying chemistry space networks generated using each representation, can guide the QSAR modeler in the choice of descriptor or fingerprint representation. A higher density of SALI edges in any given region of a chemistry space network graph with a particular fingerprint representation is an indication of a more challenging structure-activity relationship using that fingerprint in that region of chemistry space. A consensus view developed with the aid of the network representations discussed here can thus also aid in the development of better global models.

ACKNOWLEDGMENT: We thank Prof. Curt Breneman for many helpful discussions during the course of this work.

Supporting Information Available.

Figure S1. Degree distributions of the ZINC database network constructed with Euclidean similarity distances using (a) the logarithm of the octanol/water partition coefficient, the molecular refractivity and the polar surface area, and (b) the average surface electrostatic potential and the local surface average ionization potential, at different thresholds. Logarithms in the scale are natural logarithms.

Figure S2. Degree distributions of the natural products subset of the ZINC database network constructed with Euclidean similarity distances using (a) the logarithm of the octanol/water partition coefficient, the molecular refractivity and the polar surface area, and (b) USR signatures. Logarithms in the scale are natural logarithms.

Figure S3. (a) CDK, (b) MACCS, and (c) PubChem next-nearest-neighbor networks with SALI edges for molecule CID 893460, reported active in the assay. SALI edges are colored red.

Figure S4. (a) CDK, (b) MACCS, and (c) PubChem next-nearest-neighbor networks with SALI edges for molecule CID 749132, reported active in the assay. SALI edges are colored red.

FIGURE CAPTIONS

Figure 1. Degree distributions (cumulative frequency) for the PubChem bioassay 361 at 70 % Tanimoto similarity using EState, PubChem, MACCS, Graph, CDK and CDK Extended fingerprints. Degree distributions at other values of the Tanimoto similarity cutoff are qualitatively similar and are not shown.

Figure 2. Preferential attachment (nearest neighbor degree distributions) for the PubChem bioassay 361 using (a) PubChem Fingerprints, (b) MACCS Fingerprints, and (c) EState Fingerprints, at 70%, 75%, 80% and 85% Tanimoto cutoffs.

Figure 3. Local transitivity (averaged local clustering coefficient) for the PubChem bioassay 361, using PubChem and MACCS Fingerprints, at 70% Tanimoto cutoffs.

Figure 4. Degree distributions (cumulative frequency) for the PubChem bioassay 361, using USR descriptor distances, at 0.1 and 0.05 distance cutoff.

Figure 5. Preferential Attachment (Nearest neighbor degree distributions) for the PubChem bioassay 361, using USR descriptor distances, at 0.1 and 0.05 distance cutoff.

Figure 6. Out-degree distributions (cumulative frequency distributions) for the PubChem bioassay 361, at Atomtyper match levels 1, 2 and 3.

Figure 7. Preferential attachment (Nearest neighbor out-degree distributions) for the PubChem bioassay 361, at Atomtyper match Levels 1 and 2. Data for level 3 is not shown due to excessive noise at this sparse connectivity.

Figure 8. Local Transitivity (averaged local clustering coefficient) for the PubChem Bioassay 361, using Atomtyper directed Edges, at a level 1 match.

Figure 9. Out-degree distributions of the ZINC database network constructed with Atomtyper distances at Levels 1, 2 and 3. Logarithms in the scale are natural logarithms.

Figure 10. Degree distributions (cumulative frequency distributions) for PDB binding sites, using PESD descriptor distances, at 11000, 10000, 9000, 8000 and 7000 distance cutoffs.

Figure 11. Preferential attachment (nearest neighbor degree distributions) for PDB binding sites, using PESD descriptor distances, at 11000, 10000, 9000 and 8000 distance cutoffs.

Figure 12. Degree distributions (cumulative frequency distributions) for PDB binding sites, using USR descriptor distances, at 0.5, 0.4, 0.3, 0.2 and 0.1 distance cutoffs.

Figure 13. Preferential attachment (neighbor degree distributions) for PDB binding sites, using USR descriptor distances, at 0.5, 0.4 and 0.3 distance cutoffs.

Figure 14. (a) Bioassay 361 network graph as determined by pairwise comparisons of PubChem fingerprints at an 85% Tanimoto similarity threshold, in a Fruchterman-Reingold layout. Thick red lines represent SALI edges, chosen at a 95% cutoff of non-zero values. (b) is the network comprised solely of those SALI edges.

Figure 15. (a) CDK, (b) MACCS, and (c) PubChem nearest-neighbor networks with SALI edges for molecule CID 893460, reported active in the assay. SALI edges are colored red.

Figure 16. (a) CDK, (b) MACCS, and (c) PubChem nearest-neighbor networks with SALI edges for molecule CID 749132, reported active in the assay. SALI edges are colored red.

Table 1: Atomtyper matching

A requested atom type string is compared to each atom type string in the TAE library list in succession until, a) level designation is equal to 3 is found or, b) the level designation of current library atom string is less than that of the previous library atom string compared. The library atom type string with maximum level designation is used to model the requested atom in molecule.

Level	**Match**
3	Perfect match
2	Ring size differs
1	Hybridization of nearest neighbor does not match
0	Atomic number of nearest neighbor does not match
-1	Hybridization of atom does not match
-2	For monovalent atom, hybridization of nearest neighbor differs
-3	Atomic number of atom does not match

Table 2: Global Network characteristics (a) PubChem AID361 assay

			Max	Avg	Global	Global
Fingerprints	Vertices	Edges	Degree	Degree	Assortativity	Transitivity
CDK Fingerprint / 70 % Tanimoto Cutoff	44996	376693	277	16.74	0.82	0.57
CDK Fingerprint / 75 % Tanimoto Cutoff	40973	189909	134	9.27	0.83	0.58
CDK Fingerprint / 80 % Tanimoto Cutoff	36104	95364	84	5.28	0.82	0.61
CDK Fingerprint / 85 % Tanimoto Cutoff	30951	49107	39	3.17	0.81	0.67
CDK Extended Fingerprint / 70 % Tanimoto Cutoff	44955	374655	251	16.67	0.82	0.58
CDK Extended Fingerprint / 75 % Tanimoto Cutoff	40823	187985	134	9.21	0.82	0.59
CDK Extended Fingerprint / 80 % Tanimoto Cutoff	35849	92879	73	5.18	0.81	0.60
CDK Extended Fingerprint / 85 % Tanimoto Cutoff	30605	47029	39	3.07	0.81	0.66
Graph Fingerprint / 70 % Tanimoto Cutoff	50736	11759347	3806	463.55	0.54	0.45
Graph Fingerprint / 75 % Tanimoto Cutoff	49856	4681091	2011	187.78	0.58	0.43
Graph Fingerprint / 80 % Tanimoto Cutoff	48047	1722644	895	71.71	0.63	0.46
Graph Fingerprint / 85 % Tanimoto Cutoff	44601	592048	332	26.55	0.70	0.52
MACCS Keys / 70 % Tanimoto Cutoff	51266	32755396	8526	1277.86	0.29	0.40
MACCS Keys / 75 % Tanimoto Cutoff	50955	9790359	3469	384.27	0.38	0.36
MACCS Keys / 80 % Tanimoto Cutoff	49912	2513280	1017	100.71	0.50	0.36
MACCS Keys / 85 % Tanimoto Cutoff	46744	584712	320	25.02	0.63	0.41
EState Fingerprint / 70 % Tanimoto Cutoff	51365	72035823	12869	2804.86	0.28	0.47
EState Fingerprint / 75 % Tanimoto Cutoff	51183	32537089	7267	1271.40	0.37	0.46
EState Fingerprint / 80 % Tanimoto Cutoff	50788	17994483	5823	708.61	0.41	0.48
EState Fingerprint / 85 % Tanimoto Cutoff	50039	6514349	2544	260.37	0.58	0.42
PubChem Fingerprint / 70 % Tanimoto Cutoff	51125	21253120	5251	831.42	0.33	0.40
PubChem Fingerprint / 75 % Tanimoto Cutoff	50745	6433027	2043	253.54	0.40	0.39
PubChem Fingerprint / 80 % Tanimoto Cutoff	49736	1792249	735	72.07	0.55	0.45
PubChem Fingerprint / 85 % Tanimoto Cutoff	46590	489130	388	21.00	0.77	0.55
AtomTyper						
Atomtyper Level 1	50361	4271844	18638	169.65	-0.17	0.08
Atomtyper Level 2	43097	287544	4373	13.34	-0.15	0.02
Atomtyper Level 3	35424	115022	2413	6.49	-0.12	0.03
Standardized USR signatures						
Euclidean Distance Cutoff 0.1	51243	19780529	3178	772.03	0.68	0.45
Euclidean Distance Cutoff 0.05	48884	1054997	258	43.16	0.82	0.40

Table 2: Global Network characteristics (b) Protein Binding sites

PESD	**Vertices**	**Edges**	**Max Degree**	**Avg Degree**	**Global Assortativity**	**Global Transitivity**
Chi Squared Distance Cutoff 11000	93999	4986042	2185	106.09	0.68	0.44
Chi Squared Distance Cutoff 10000	90387	2952311	1586	65.33	0.69	0.46
Chi Squared Distance Cutoff 9000	86137	1715130	1026	39.82	0.71	0.50
Chi Squared Distance Cutoff 8000	81328	1018415	550	25.04	0.75	0.58
Chi Squared Distance Cutoff 7000	76298	647516	323	16.97	0.84	0.72
Chi Squared Distance Cutoff 6000	71410	452579	185	12.68	0.91	0.82
Chi Squared Distance Cutoff 5000	65813	329101	165	10.00	0.92	0.84
Chi Squared Distance Cutoff 4000	57693	220286	125	7.64	0.93	0.85
Chi Squared Distance Cutoff 3000	43212	111429	88	5.16	0.94	0.84
Chi Squared Distance Cutoff 2500	29353	60062	77	4.09	0.94	0.85
Chi Squared Distance Cutoff 2000	14500	24088	50	3.32	0.88	0.79
Chi Squared Distance Cutoff 1500	3577	5136	29	2.87	0.89	0.82
Standardized USR signatures						
Euclidean Distance Cutoff 0.5	94386	40772454	44396	863.95	-0.30	0.08
Euclidean Distance Cutoff 0.4	86004	12577411	39695	292.48	-0.31	0.03
Euclidean Distance Cutoff 0.3	72777	2663127	35543	73.19	-0.35	0.01
Euclidean Distance Cutoff 0.2	51466	229240	30626	8.91	-0.35	0.00
Euclidean Distance Cutoff 0.1	2692	2035	202	1.51	-0.11	0.00

Table 3: Network Characteristics of SALI networks: PubChem AID361 assay, at 95% non-zero SALI cutoff

Fingerprints	Vertices	Edges	Max Degree	Avg Degree	Global Assortativity	Global Transitivity
CDK Fingerprint / 70 % Tanimoto Cutoff	5223	4225	25	1.62	0.04	0.12
CDK Fingerprint / 75 % Tanimoto Cutoff	3172	2159	15	1.36	0.06	0.14
CDK Fingerprint / 80 % Tanimoto Cutoff	1766	1090	6	1.23	0.08	0.15
CDK Fingerprint / 85 % Tanimoto Cutoff	962	544	4	1.13	0.12	0.17
CDK Extended Fingerprint / 70 % Tanimoto Cutoff	5132	4170	24	1.63	0.03	0.12
CDK Extended Fingerprint / 75 % Tanimoto Cutoff	3148	2132	14	1.35	0.07	0.13
CDK Extended Fingerprint / 80 % Tanimoto Cutoff	1743	1057	4	1.21	0.1	0.18
CDK Extended Fingerprint / 85 % Tanimoto Cutoff	920	525	4	1.14	0.15	0.17
Graph Fingerprint / 70 % Tanimoto Cutoff	30874	157250	1775	10.19	-0.24	0.01
Graph Fingerprint / 75 % Tanimoto Cutoff	22231	59573	631	5.36	-0.22	0.02
Graph Fingerprint / 80 % Tanimoto Cutoff	13418	20747	135	3.09	-0.19	0.06
Graph Fingerprint / 85 % Tanimoto Cutoff	6723	6771	41	2.01	-0.09	0.16
MACCS Keys / 70 % Tanimoto Cutoff	41317	401313	4251	19.43	-0.28	0
MACCS Keys / 75 % Tanimoto Cutoff	30307	113501	1358	7.49	-0.24	0
MACCS Keys / 80 % Tanimoto Cutoff	16647	27999	285	3.36	-0.19	0.02
MACCS Keys / 85 % Tanimoto Cutoff	6710	6432	69	1.92	-0.02	0.09
EState Fingerprint / 70 % Tanimoto Cutoff	43318	793559	4491	36.64	-0.25	0.03
EState Fingerprint / 75 % Tanimoto Cutoff	37416	344452	2432	18.41	-0.23	0.05
EState Fingerprint / 80 % Tanimoto Cutoff	30304	176161	1552	11.63	-0.15	0.15
EState Fingerprint / 85 % Tanimoto Cutoff	19652	64588	531	6.57	-0.29	0
PubChem Fingerprint / 70 % Tanimoto Cutoff	38263	303256	2166	15.85	-0.3	0
PubChem Fingerprint / 75 % Tanimoto Cutoff	26493	87066	750	6.57	-0.23	0.01
PubChem Fingerprint / 80 % Tanimoto Cutoff	14064	22786	221	3.24	-0.12	0.03
PubChem Fingerprint / 85 % Tanimoto Cutoff	6096	5838	34	1.92	0.07	0.11
Standardized USR signatures						
Euclidean Distance Cutoff 0.1	10771	16466	80	3.06	-0.23	0.01
Euclidean Distance Cutoff 0.05	35742	309495	971	17.32	-0.29	0.01

Table 4: Comparison Between Neighborhood Network Characteristics

Medians of total edges for networks at 85% Tanimoto similarity cutoffs, SALI edges chosen at a 95% cutoff of non-zero values, and ratios between them were computed for the CDK, MACCS, and PubChem fingerprints and USR descriptors for the AID361 assay. Data was median-averaged over the full dataset and also over the top 200 actives that exist in all fingerprint comparison networks. This allows a direct comparison of the "smoothness" of representation for similarity measures, especially in representing the active molecules, the "hits" in AID361.

		Full Dataset Medians/Neighbors			Top 200 common actives Medians/Neighbors		
Fingerprints	**N**	Median Total Edges	Median SALI edges	Median Ratio	Median Total Edges	Median SALI edges	Median Ratio
CDK	1	3	0	0	1	0	0
CDK	2	4	0	0	3	0	0
CDK	3	5	0	0	3	0	0
CDK	4	5	0	0	3	0	0
CDK	5	6	0	0	3	0	0
CDK	6	6	0	0	3	0	0
MACCS	1	36	0	0	43.5	4	0.06
MACCS	2	522	6	0.006	380.5	16	0.033
MACCS	3	5410.5	61	0.006	2000	69.5	0.02
MACCS	4	43227.5	324	0.007	7543.5	265.5	0.013
MACCS	5	129938.5	923	0.008	43086.5	607.5	0.009
MACCS	6	171789.5	1491	0.009	145454.5	1358	0.01
PubChem	1	35	0	0	52.5	3	0.053
PubChem	2	328	3	0.005	561.5	15	0.03
PubChem	3	2147	18	0.007	4053.5	76	0.021
PubChem	4	9268	86	0.007	11234	197	0.017
PubChem	5	27408	246	0.008	29998	452	0.013
PubChem	6	61789	511	0.009	73038	939.5	0.012
Standardized USR signatures							
Euclidean Distance Cutoff 0.05	1	186	1	0.002	9	0	0
Euclidean Distance Cutoff 0.05	2	3160	17	0.005	114.5	2	0.0015
Euclidean Distance Cutoff 0.05	3	19366	111	0.007	3621	35.5	0.003
Euclidean Distance Cutoff 0.05	4	66354	399	0.007	24619.5	246.5	0.005
Euclidean Distance Cutoff 0.05	5	157364	1159	0.007	72773	819	0.006
Euclidean Distance Cutoff 0.05	6	301910.5	2472	0.007	176038.5	1927	0.007

REFERENCES

1. Bader, R. F. W., Atoms in Molecules: A Quantum Theory. Oxford Press: Oxford, 1990.

2. Kier, L. B.; Hall, L. H., Molecular Connectivity in Chemistry and Drug Research. Acadmic Press: New York, 1976.

3. Kier, L. B.; Hall, L. H., Molecular Connectivity in Structure-Activity Analysis. Research Studies Press: Letchworth, 1986.

4. Randic, M., The connectivity index 25 years after. *J. Mol. Graph. Model.* **2001,** 20 (1), 19-35.

5. Trinajstic, N., Chemical Graph Theory. CRC: Boca Raton, 1992.

6. Nikolova, N.; Jaworska, J., Approaches to measure chemical similarity - a review. *QSAR Comb. Sci.* **2003,** 22, 1006-1026.

7. Rupp, M.; Proschak, E.; Schneider, G., Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.* **2007,** 47 (6), 2280-2286.

8. Bergeron, C.; Hepburn, T.; Sundling, M.; Sukumar, N.; Bennett, K. P.; Breneman, C., Prediction of peptide bonding affinity: kernel methods for nonlinear modeling. *Protein Pept. Lett.* **2008**.

9. Guha, R.; Schuerer, S. C., Utilizing high throughput screening data for predictive toxicology models: protocols and applications to MLSCN assays. *J. Comput.-Aided Mol. Des.* **2008,** 22, 367-384.

10. Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J., Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010,** 53 (15), 5707-5715.

11. Wawer, M.; Peltason, L.; Weskamp, L.; Teckentrup, A.; Bajorath, J., Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure Activity Realtionship Indices. *J. Med. Chem.* **2008**, 51, 6075-6084.

12. Thom, R., Structural Stability and Morphogenesis. Benjamin: Reading, 1975.

13. Gasteiger, J., De novo design and synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, 21, 307-9.

14. Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D., Transferable Atom Equivalent Multi-Centered Multipole Expansion Method. *J. Comp. Chem.* **2003,** 24, 512-529.

15. Carbo, R. (Ed.), Molecular similarity and reactivity: from quantum chemical to phenomenological approaches. Kluver: Amsterdam, 1995.

16. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M., Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002,** 45, 4350-4358.

17. Maggiora, G. M., On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006,** 46 (4), 1535.

18. Guha, R.; Van Drie, J. H., Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008,** 48, 646−658.

19. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H., Navigating structure-activity landscapes. *Drug Discov. Today* **2009,** 14 (13-14), 698-705.

20. Agrafiotis, D. K.; Shemanarev, M; Connolly, P. J.; Farnum, M.; Lobanov, V. S., SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, 50, 5926-5937.

21. Guha, R.; Van Drie, J. H., Assessing How Well a Modeling Protocol Captures a Structure−Activity Landscape. *J. Chem. Inf. Model.* **2008,** 48 (8), 1716 - 1728.

22. Anbert, R.; Barabási, A.-L., Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, 74, 47-97.

23. Oltvai, Z. N.; Barabási, A.-L., Life's Complexity Pyramid. *Science* **2002**, 298, 763-764.

24. Barabási, A.-L.; Oltvai, Z. N., Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* **2004**, 5, 101-114.

25. Barabási, A.-L., Taming Complexity. *Nature Phys.* **2005**, 1, 68-70.

26. Barabási, A.-L., The Architecture of Complexity. I*EEE Control Sys. Mag*. Aug. **2007**, 33.

27. Park, J.; Barabási, A.-L., Distribution of node characteristics in complex networks. *Proc. Nat. Acad. Sci.* **2007**, 104(46), 17916-17920.

28. Bianconi, G.; Gulbahce, N.; Motter, A. E., Local Structure of Directed Networks. *Phys. Rev. Lett.* **2008**, 100, 118701.

29. Soffer, S. N.; Vásquez, A., Network clustering coefficient without degree-correlation biases. *Phys. Rev. E* **2005**, 71, 057101.

30. Holme, P.; Zhao, J., Exploring the assortativity-clustering space of a network's degree sequence. *Phys. Rev. E* **2005**, 75, 046111.

31. Erdös, P.; Rényi, A., On the evolution of random graphs. *Bull. Inst. Int. Stat.* **1961**, 38, 343.

32. Meinl, T.; Ostermann, C.; Berthold, M. R., Maximum-Score Diversity Selection for Early Drug Discovery. *J. Chem. Inf. Model.* **2011**, 51, 237–247.

33. Pearlman, R. S.; Smith, K. M., Novel Software Tools for Chemical Diversity, in *3D QSAR in Drug Design*, ed. ,Kubinyi, Hugo and Folkers, Gerd and Martin, Yvonne C., Springer Netherlands, 2002, 2, pp. 339-353.

34. Stanton, D. T., Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.*, **1999**, 39, 11–20.

35. Jenkins, J. L.; Glick, M.; Davies, J. W., A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, 47, 6144-6159.

36. Rush III, T. S.; Grant, J. A.; Mosyak, L; Nicholls, A., A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, 48, 1489-1495.

37. Barker, E. J. Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts,P.; Willett, P.; Gillet, V. J., Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, 46, 503-511

38. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002,** 42 (6), 1273-1280.

39. Guha, R., Chemical Informatics Functionality in R. *J. Stat. Software* **2007,** 18 (6).

40. Breneman, C. M.; Thompson, T., Modeling the Hydrogen Bond with Transferable Atom Equivalents. In Modeling the Hydrogen Bond, Smith, D., Ed. ACS Symposium Series: Washington, D.C., 1993; pp 152-174.

41. Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M., Electron Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. *Computers & Chemistry* **1995,** 19 (3), 161.

42. Sukumar, N.; Breneman, C. M., QTAIM in Drug Discovery and Protein Modeling. In The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design, Matta, C. F.; Boyd, R. J., Eds. Wiley-VCH: Weinheim, 2007; pp 471-498.

43. Ballester, P. J.; Richards, W. G., Ultrafast shape recognition for similarity search in molecular databases. *Proc. R. Soc. A* **2007,** 463, 1307-1321.

44. Das, S.; Kokardekar, A.; Breneman, C. M., Rapid Comparison of Protein Binding Site Surfaces with Property Encoded Shape Distributions. *J. Chem. Inf. Model.* **2009,** 49 (12), 2863-2872.

45. Das, S.; Krein, M. P.; Breneman, C. M., Binding Affinity Prediction with Property-Encoded Shape Distribution Signatures. *J. Chem. Inf. Model.* **2010,** 50 (2), 298-308.

46. Das, S.; Krein, M. P.; Breneman, C. M., PESDserv: a server for high-throughput comparison of protein binding site surfaces. *Bioinformatics* **2010,** 26 (15), 1913-1914.

47. Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P., Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Nat. Acad. Sci.* **2006,** 103 (31), 11473-11478.

48. Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H., An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **2010,** 38 (suppl 1), D255-D266.

49. Molecular Operating Environment, Version 2008.10; Chemical Computing Group, Inc.: Montreal, QC, 2008.

50. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2010, ISBN 3-900051-07-0, URL http://www.R-project.org.

51. Irwin, J. J.; Shoichet, B. K., ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, 45, 177-182. http://zinc.docking.org/

52. Politzer, P.; Murray, J. S.; Peralta-Inga, Z. Molecular surface electrostatic potentials in relation to noncovalent interactions in biological systems. *Int. J. Quantum Chem.* **2001,** 85 (6), 676-684.

53. Politzer, P.; Truhlar, D. G., Chemical Applications of Atomic and Molecular Electrostatic Potential; Plenum Press: New York, **1981**.

54. Politzer, P.; Murray, J. S. The average local ionization energy: concepts and applications, in *Theoretical Aspects of Chemical Reactivity*, Toro-Labbé, A., Ed.; *Theoret. Computat. Chem.,* Elsevier, **2007**, *19*, 119-137.

55. Murray, J. S.; Politzer, P.; Famini, G. R. Theoretical alternatives to linear solvation energy relationships. *J. Molec. Struct. (THEOCHEM)* **1998,** 454 (2-3), 299-306.

56. Open Babel: The Open Source Chemistry Toolbox. http://openbabel.org/

57. Benz, R. W.; Swamidass, J.; Baldi, P., Discovery of Power-Laws in Chemical Space. *J. Chem. Inf. Model.* **2008**, 48, 1138-1151.

58. Tanaka, N.; Ohno, K.; Niimi, T.; Moritomo, A.; Mori, K.; Orita, M., Small-World Phenomena in Chemical Library Networks: Application to Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2009**, 49(12), 2677-2686.

59. Yu, et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **2008**, 322, 104-110.

60. Yildrim, M. A.; Goh, K.I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M., Drug-target network. *Nature Biotech.* **2007**, 25, 1119-1126.

61. Csardi, G.; Nepusz, T., The igraph software package for complex network research. *InterJournal* **2006**, 1695.

62. Fruchterman, T. M. J.; Reingold, E. M., Graph drawing by force-directed placement. *Software Pract. Expr.* **1991**, 21 (11), 1129-1164.

63. Eguíluz, V. M.; Chialvo, D. R.; Cechi, G. A.; Baliki, M.; Apkarian, V., Scale-Free Brain Functional Networks. *Phys. Rev. Lett.* **2005**, 94, 018102.

64. Yook, S.-H.; Radicchi, F.; Meyer-Ortmanns, H., Self-similar scale-free networks and disassortivity. *Phys. Rev. E* **2005**, 72, 045105.

65. Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; O¨berg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, *J. Chem. Inf. Model.* **2008**, 48 (9), 1733–1746.

66. Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs, J. Chem. Inf. Model. ***2006,*** 46 (2), 503-511.

67. Ehresmann, B.; Groot, M. J. d.; Alex, A.; Clark, T. New molecular descriptors based on local properties at the molecular surface and a boiling-point model derived from them. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (2), 658 -668.

68. Clark, T. QSAR and QSPR based solely on surface properties? *J. Molec. Graph. Model.* **2004,** *22*, 519–525.

Figure 1

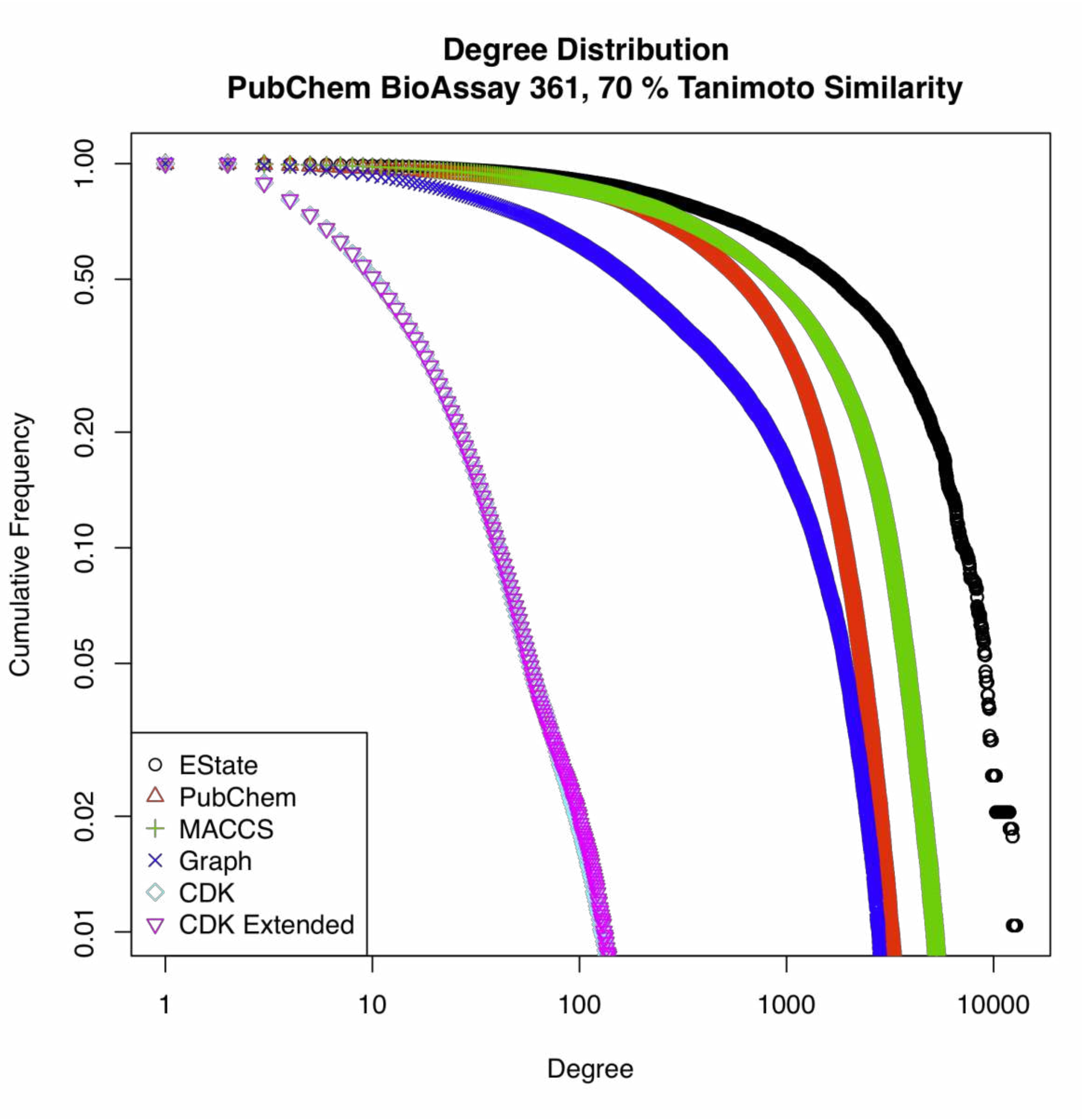

Degree Distribution
PubChem BioAssay 361, 70 % Tanimoto Similarity
Cumulative Frequency
1.00
0.50
0.20
0.10
0.05
0.02
0.01
Degree
1
10
100
1000
10000
○ EState
△ PubChem
+ MACCS
× Graph
◇ CDK
▽ CDK Extended

Figure 2a

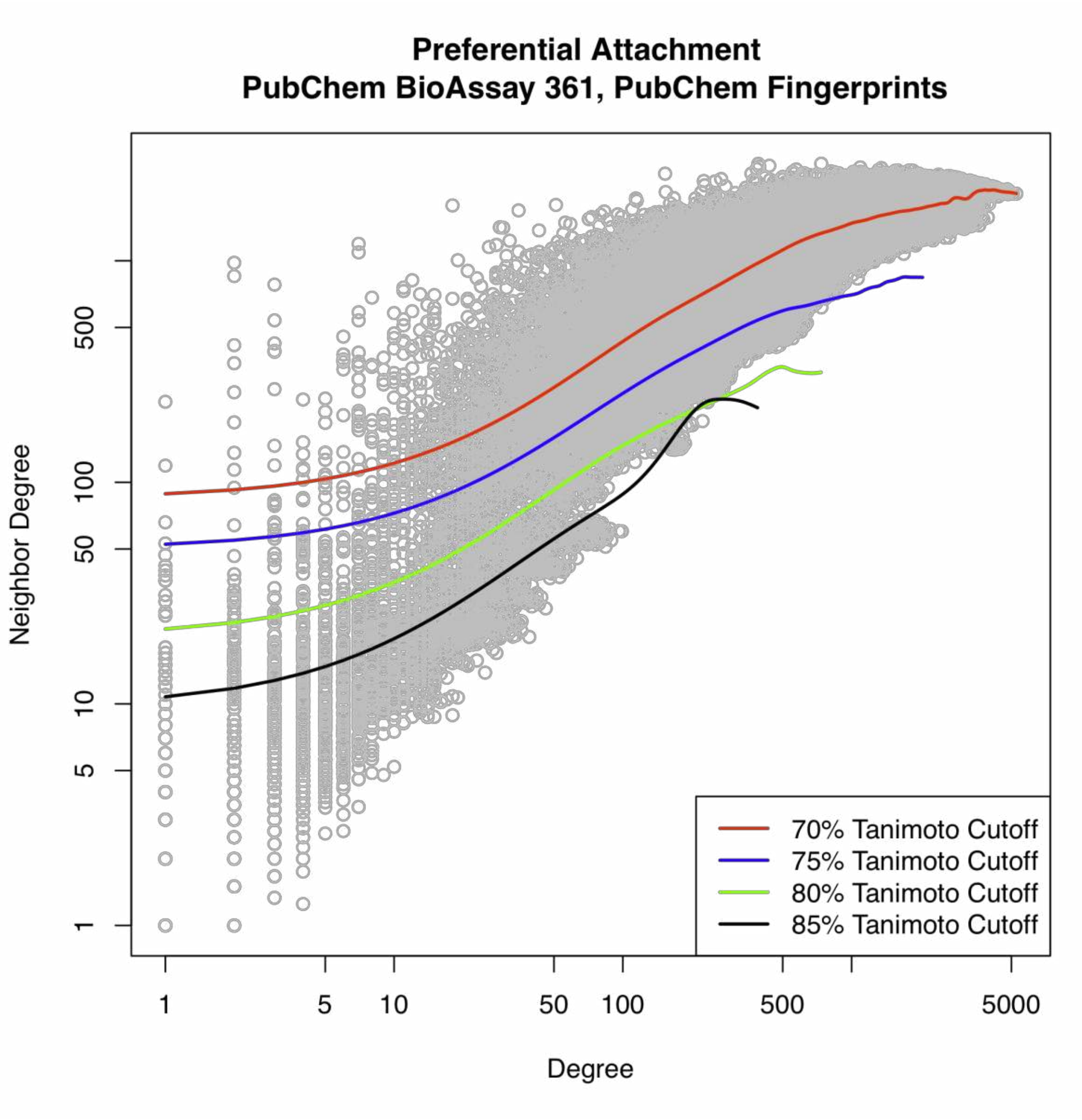

Preferential Attachment
PubChem BioAssay 361, PubChem Fingerprints
Neighbor Degree
1
5
10
50
100
500
Degree
1
5
10
50
100
500
5000
70% Tanimoto Cutoff
75% Tanimoto Cutoff
80% Tanimoto Cutoff
85% Tanimoto Cutoff

Figure 2b

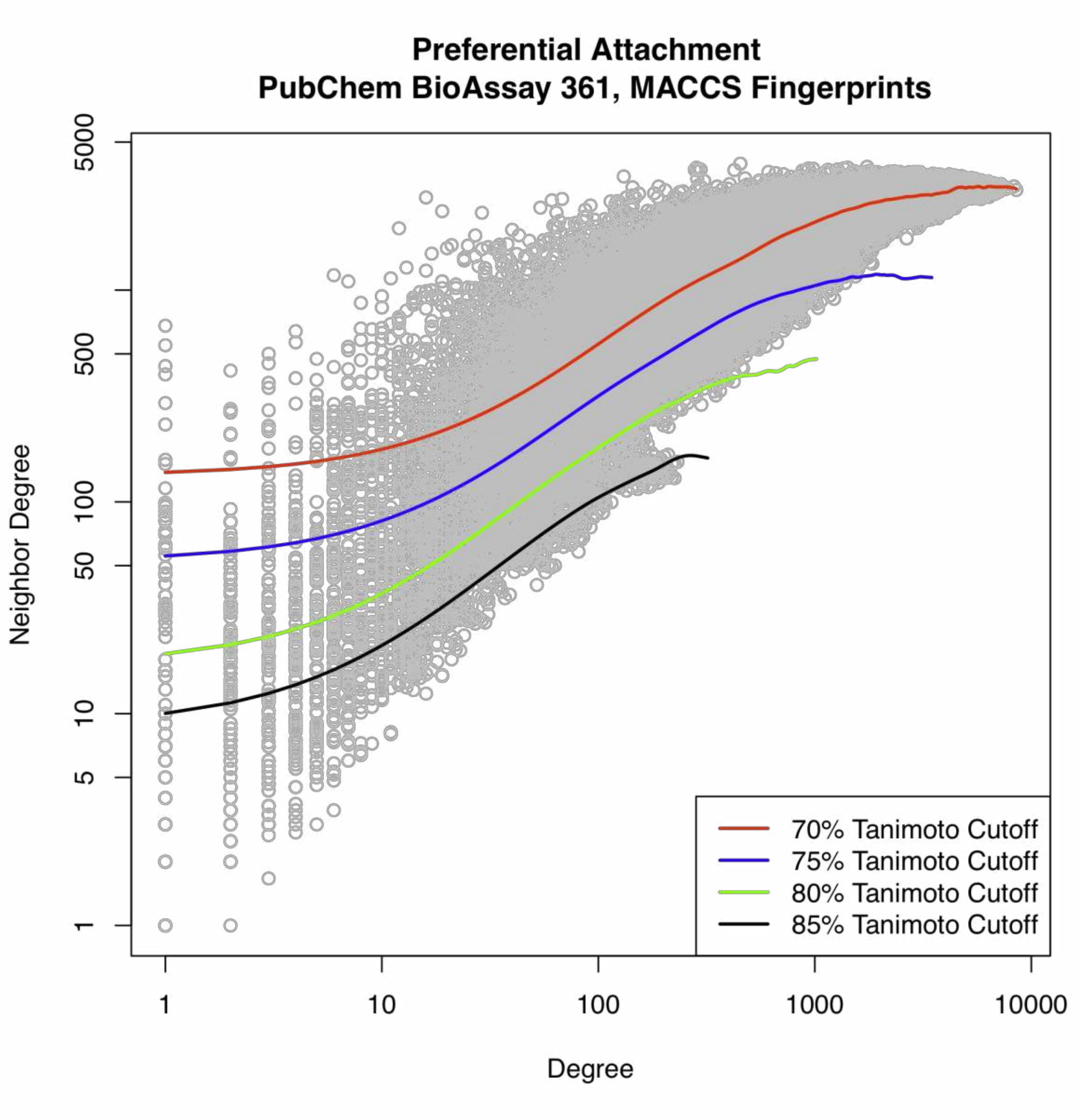

Preferential Attachment
PubChem BioAssay 361, MACCS Fingerprints
Neighbor Degree
5000
500
100
50
10
5
1
Degree
1
10
100
1000
10000
70% Tanimoto Cutoff
75% Tanimoto Cutoff
80% Tanimoto Cutoff
85% Tanimoto Cutoff

Figure 2c

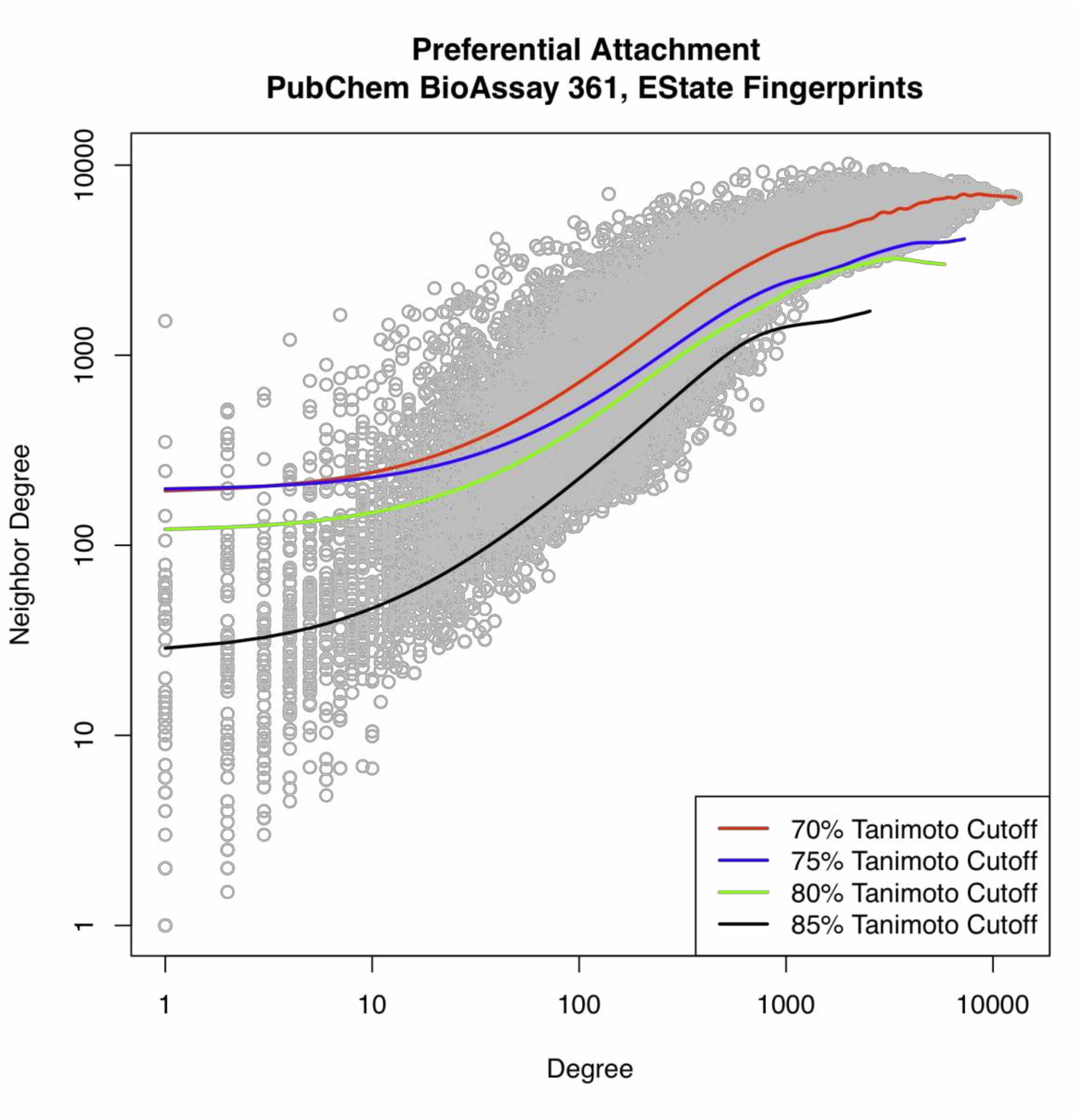

Preferential Attachment
PubChem BioAssay 361, EState Fingerprints
Neighbor Degree
10000
1000
100
10
1
Degree
1
10
100
1000
10000
70% Tanimoto Cutoff
75% Tanimoto Cutoff
80% Tanimoto Cutoff
85% Tanimoto Cutoff

Figure 3

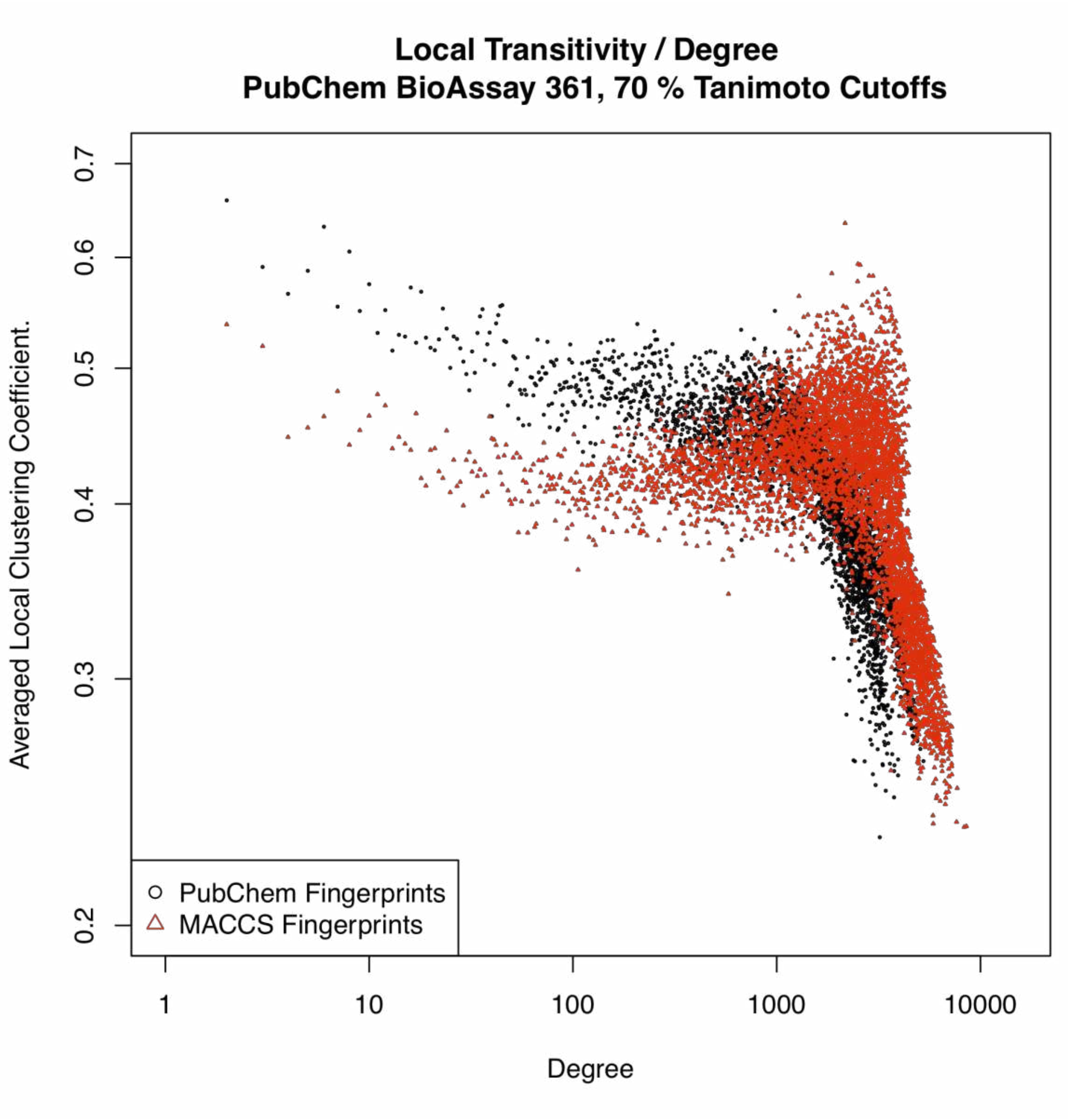

Local Transitivity / Degree
PubChem BioAssay 361, 70 % Tanimoto Cutoffs
Averaged Local Clustering Coefficient.
0.7
0.6
0.5
0.4
0.3
0.2
1
10
100
1000
10000
Degree
○ PubChem Fingerprints
△ MACCS Fingerprints

Figure 4

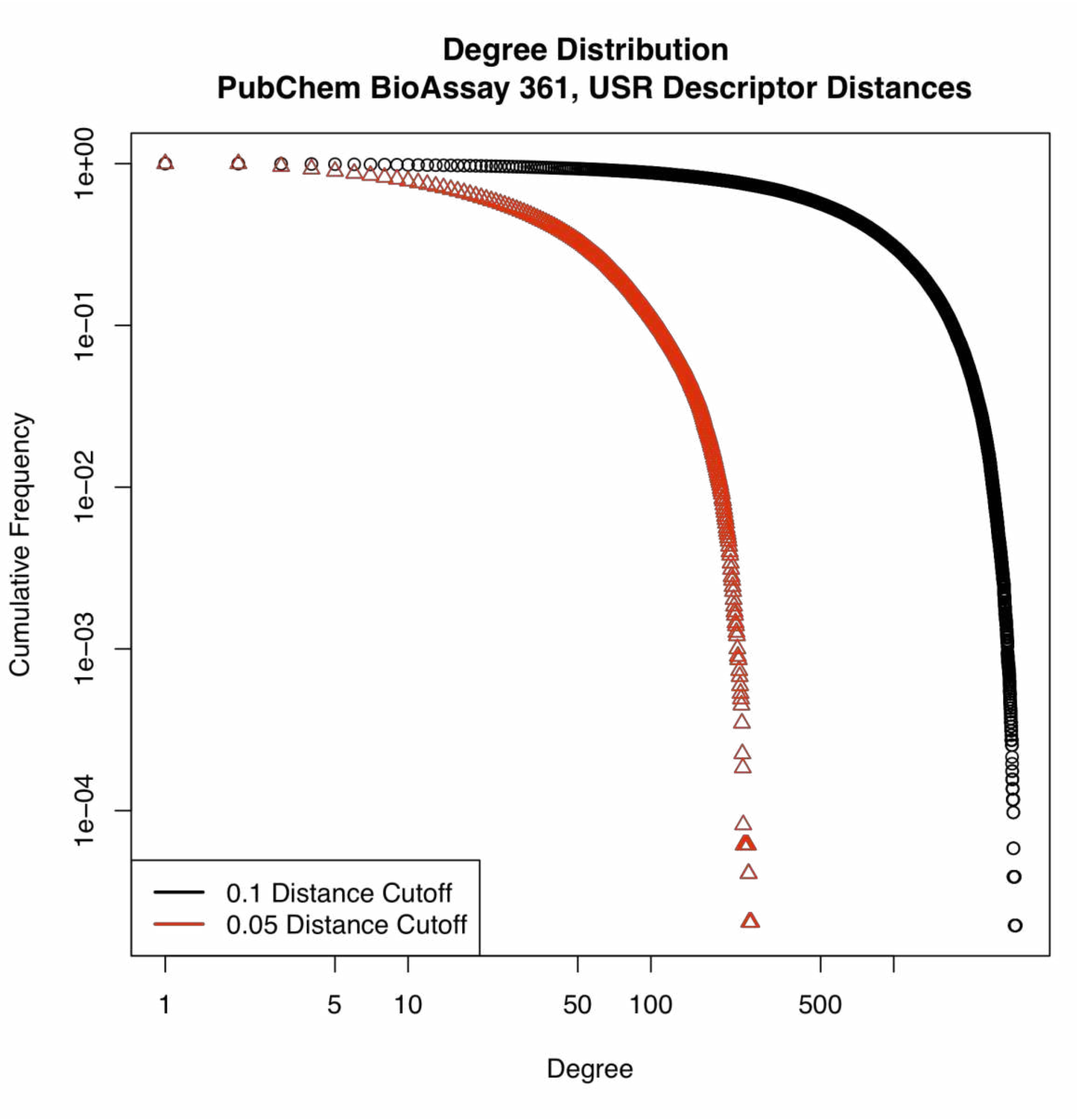

Degree Distribution
PubChem BioAssay 361, USR Descriptor Distances
Cumulative Frequency
1e+00
1e−01
1e−02
1e−03
1e−04
Degree
1
5
10
50
100
500
0.1 Distance Cutoff
0.05 Distance Cutoff

Figure 5

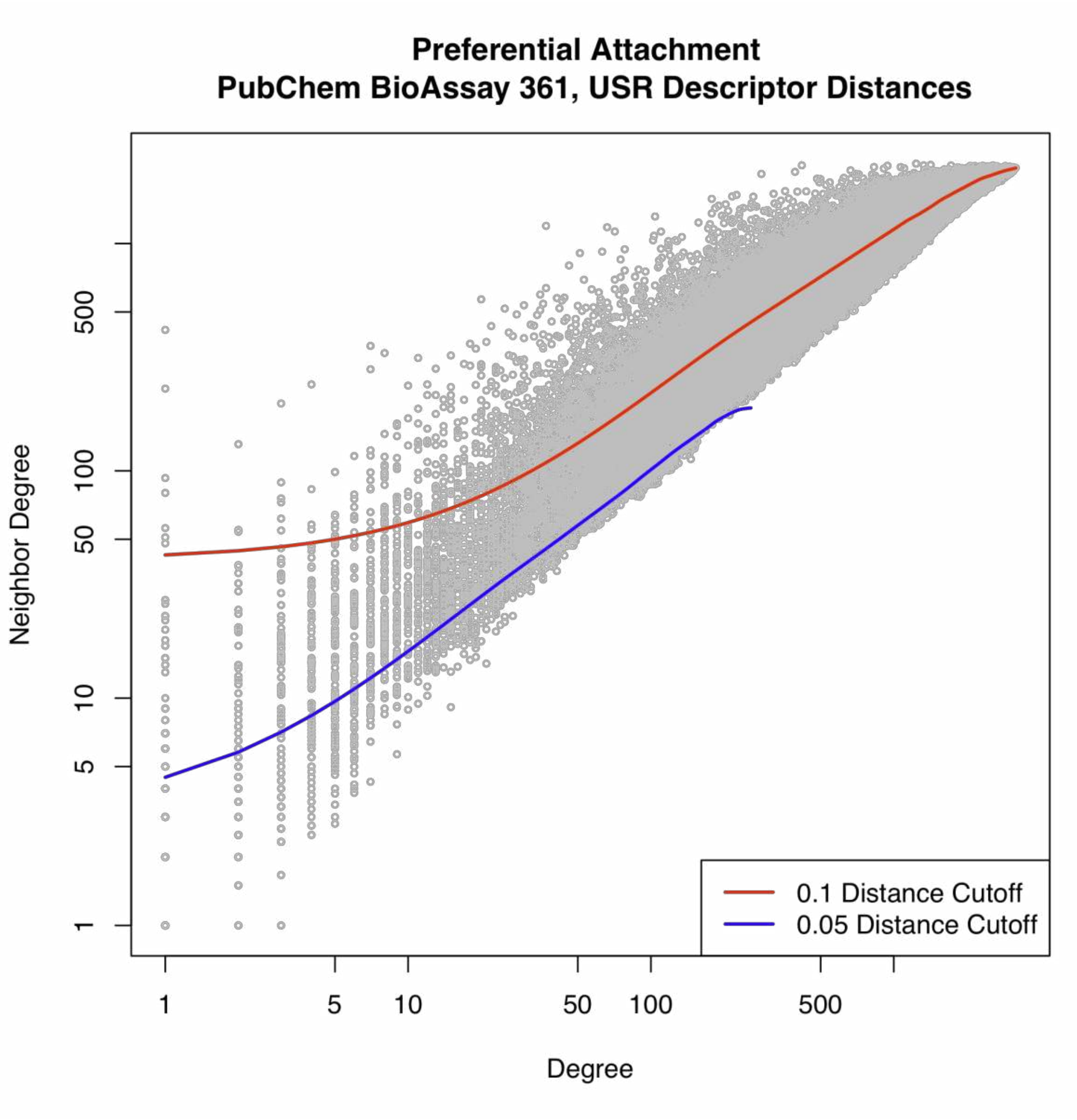

Preferential Attachment
PubChem BioAssay 361, USR Descriptor Distances
Neighbor Degree
1
5
10
50
100
500
Degree
1
5
10
50
100
500
0.1 Distance Cutoff
0.05 Distance Cutoff

Figure 6

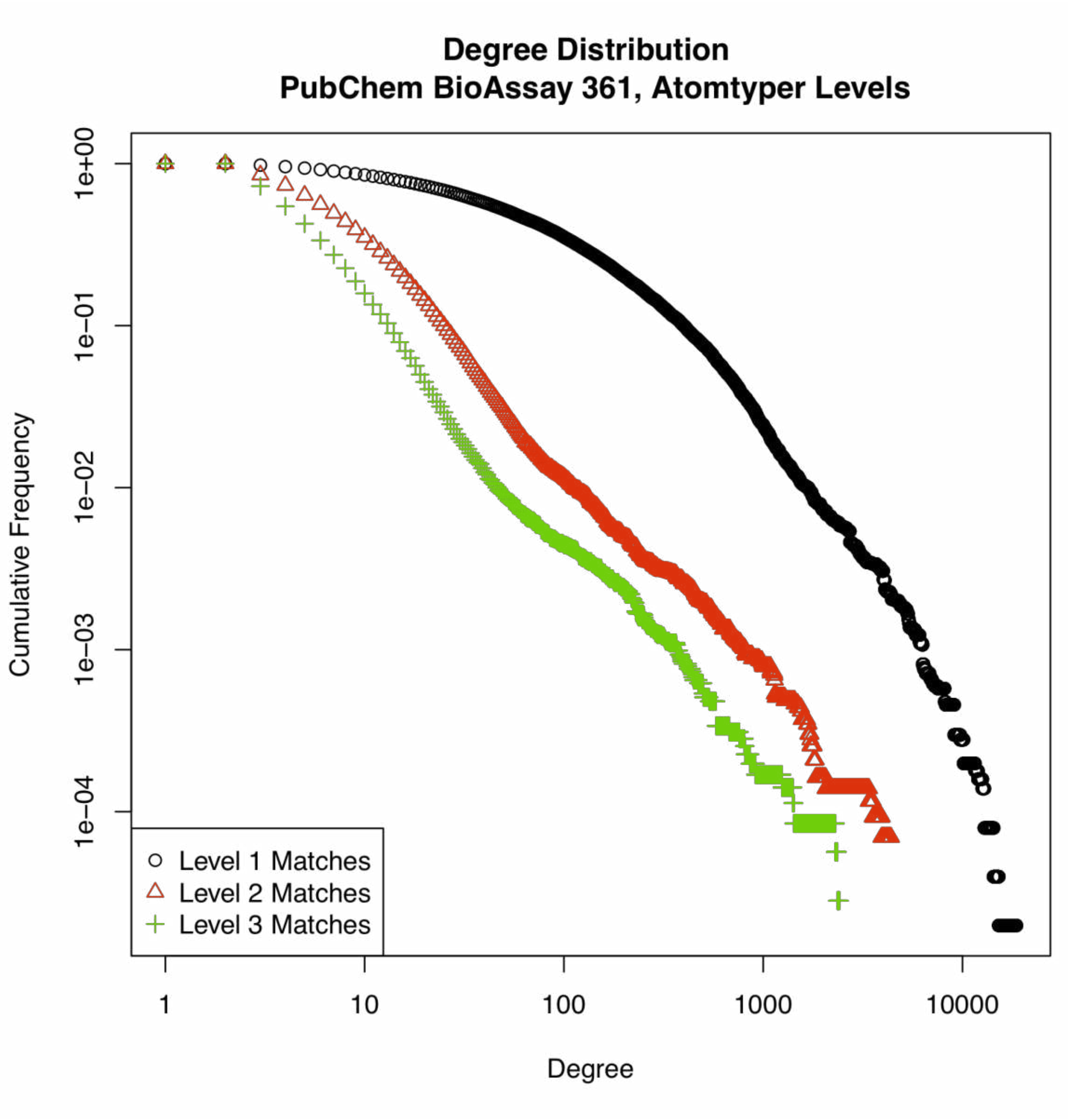

Degree Distribution
PubChem BioAssay 361, Atomtyper Levels
Cumulative Frequency
1e+00
1e−01
1e−02
1e−03
1e−04
Degree
1
10
100
1000
10000
○ Level 1 Matches
△ Level 2 Matches
+ Level 3 Matches

Figure 7

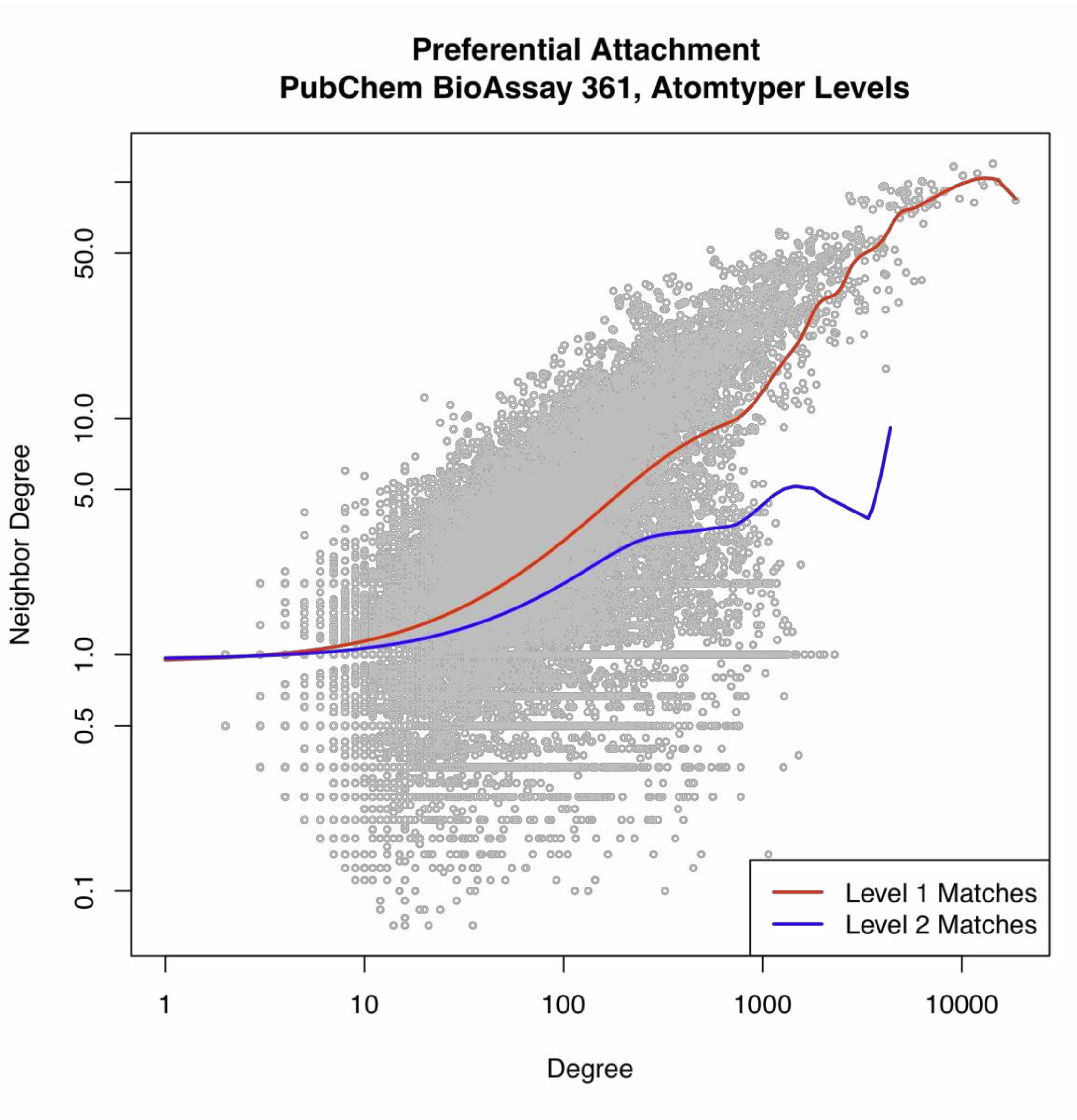

Preferential Attachment
PubChem BioAssay 361, Atomtyper Levels
Neighbor Degree
50.0
10.0
5.0
1.0
0.5
0.1
1
10
100
1000
10000
Degree
Level 1 Matches
Level 2 Matches

Figure 8

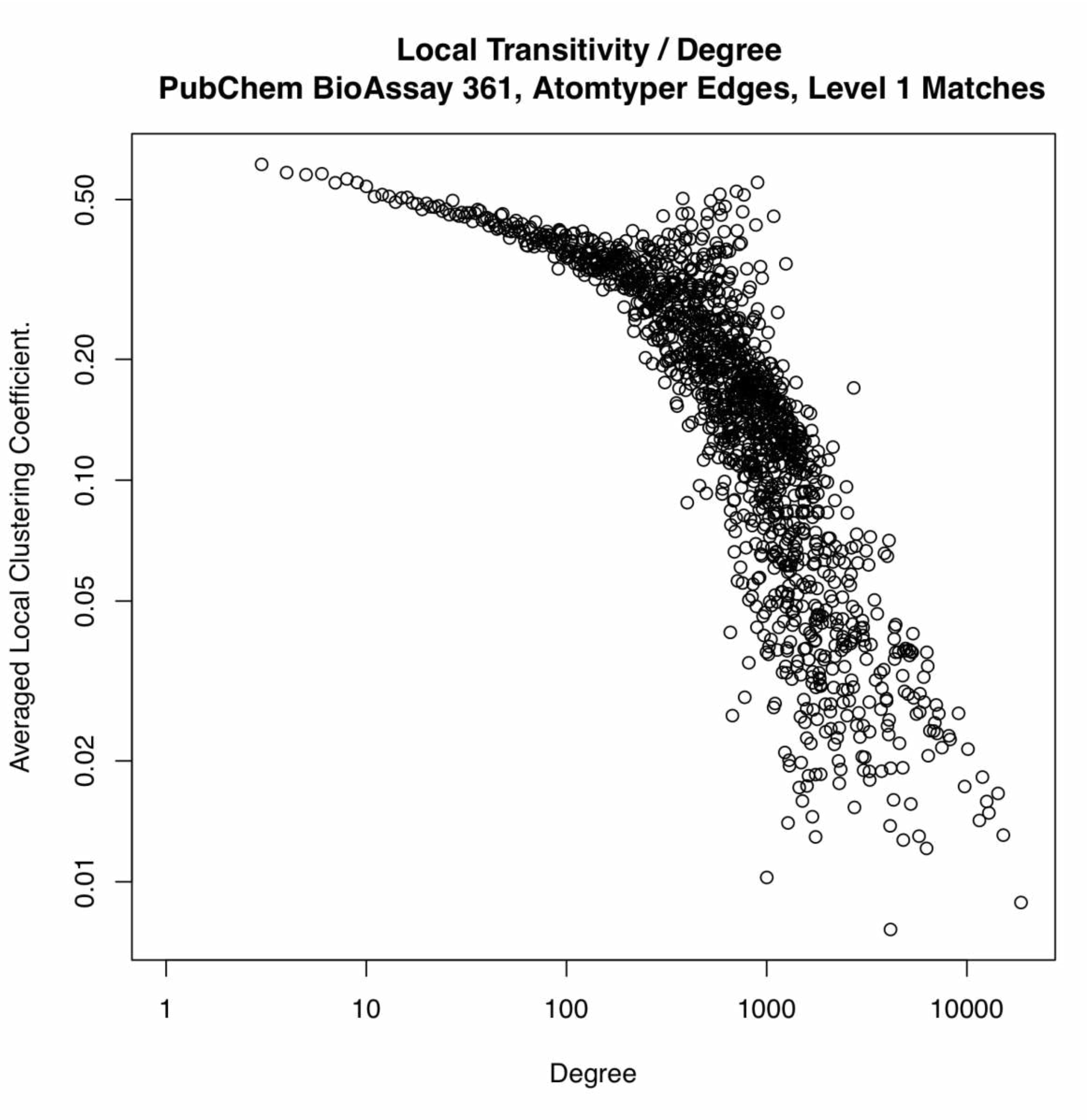


Figure 9

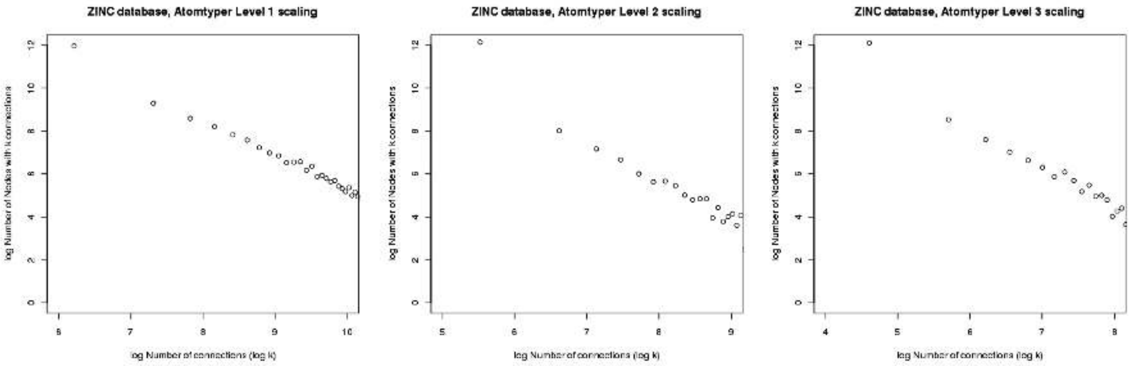


Figure 10

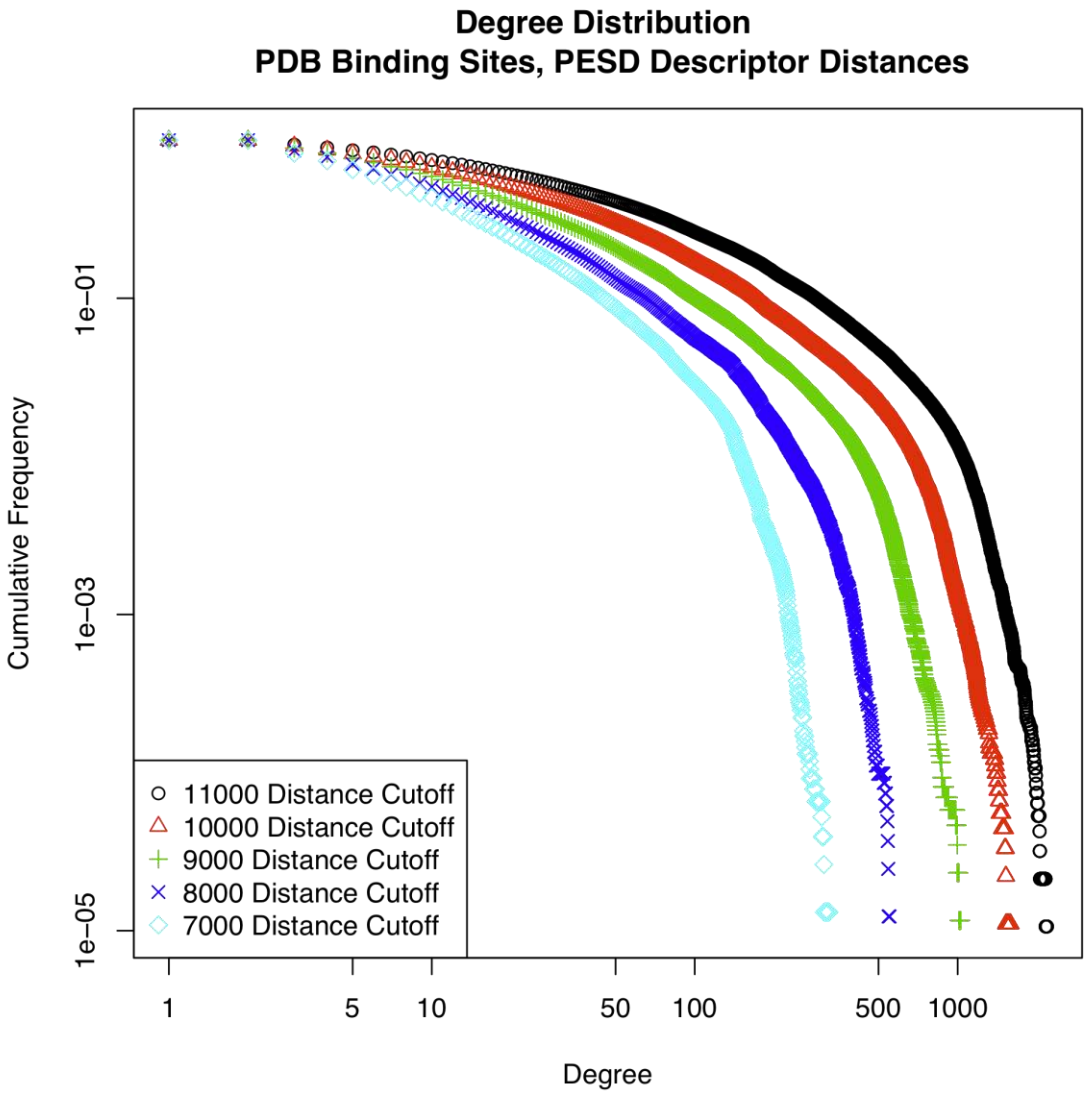

Degree Distribution
PDB Binding Sites, PESD Descriptor Distances

Figure 11

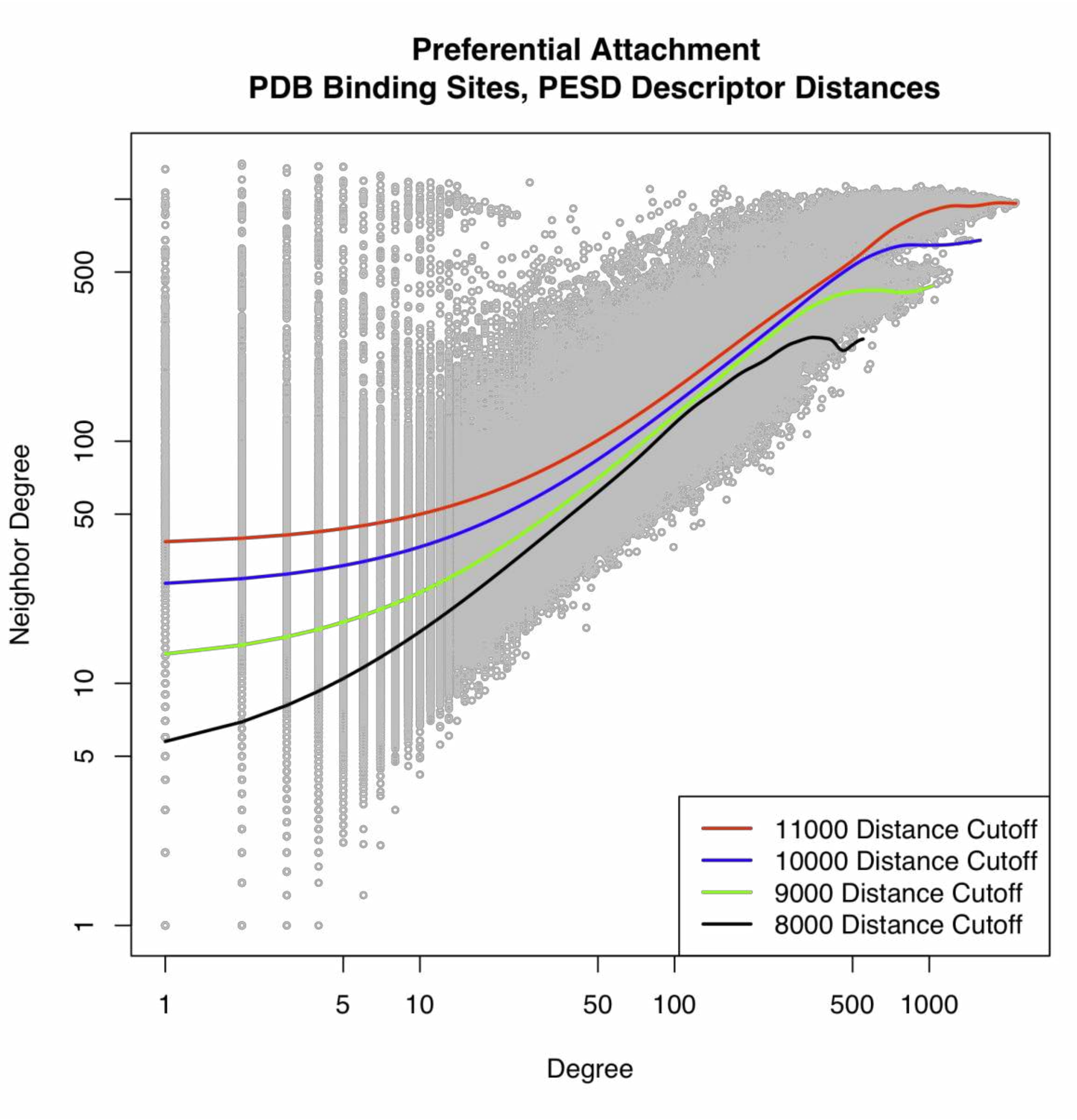

Preferential Attachment
PDB Binding Sites, PESD Descriptor Distances
Neighbor Degree
1
5
10
50
100
500
Degree
1
5
10
50
100
500
1000
11000 Distance Cutoff
10000 Distance Cutoff
9000 Distance Cutoff
8000 Distance Cutoff

Figure 12

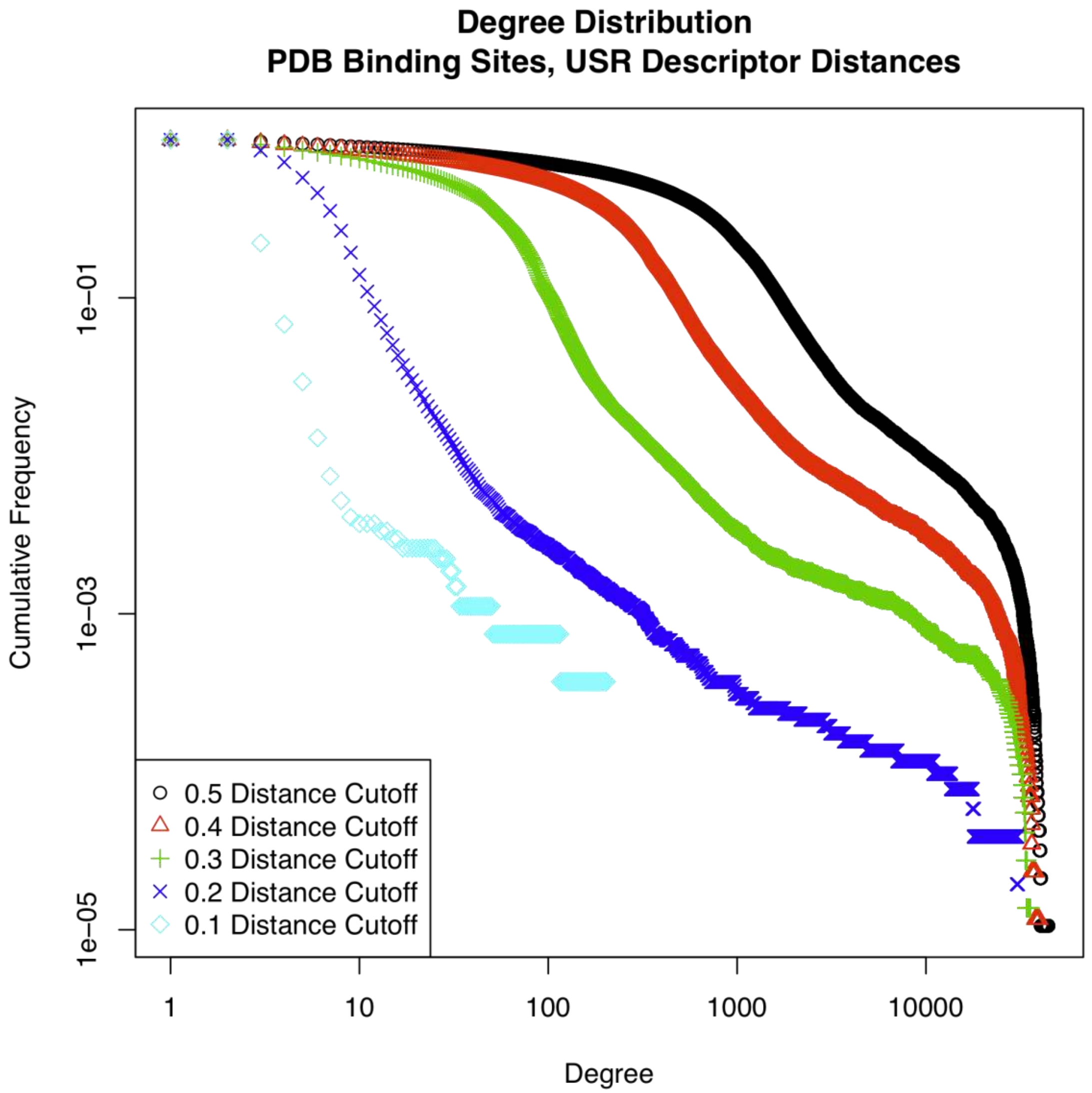

Degree Distribution
PDB Binding Sites, USR Descriptor Distances
Cumulative Frequency
1e−01
1e−03
1e−05
1
10
100
1000
10000
Degree
○ 0.5 Distance Cutoff
△ 0.4 Distance Cutoff
+ 0.3 Distance Cutoff
× 0.2 Distance Cutoff
◇ 0.1 Distance Cutoff

Figure 13

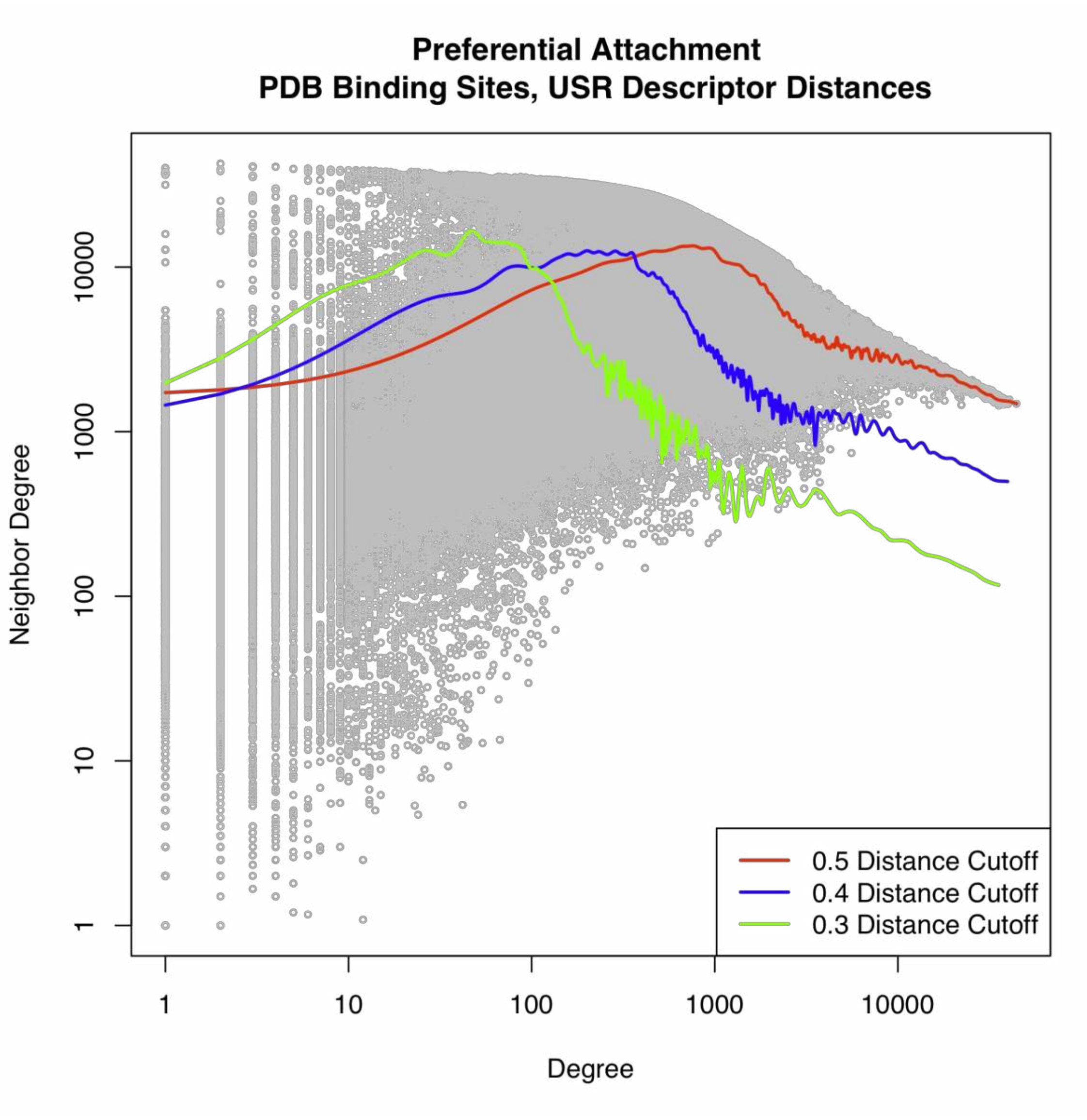

Preferential Attachment
PDB Binding Sites, USR Descriptor Distances
Neighbor Degree
1
10
100
1000
10000
Degree
1
10
100
1000
10000
0.5 Distance Cutoff
0.4 Distance Cutoff
0.3 Distance Cutoff

Figure 14

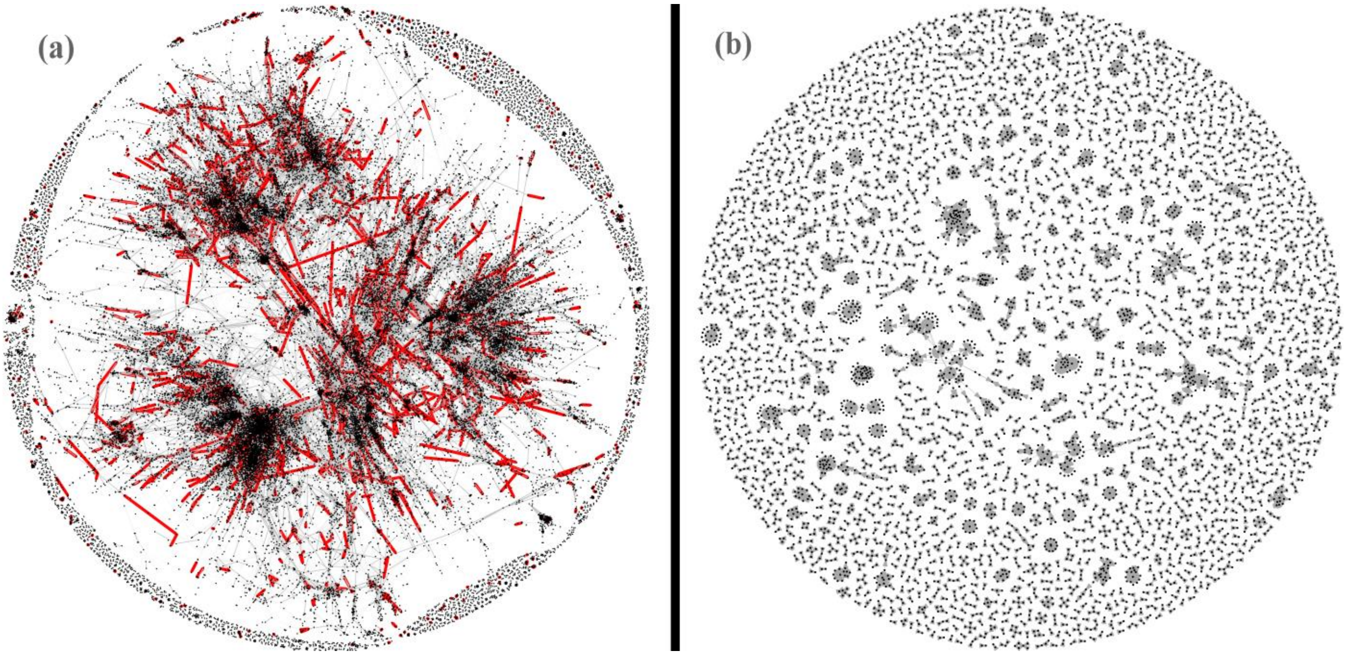


Figure 15

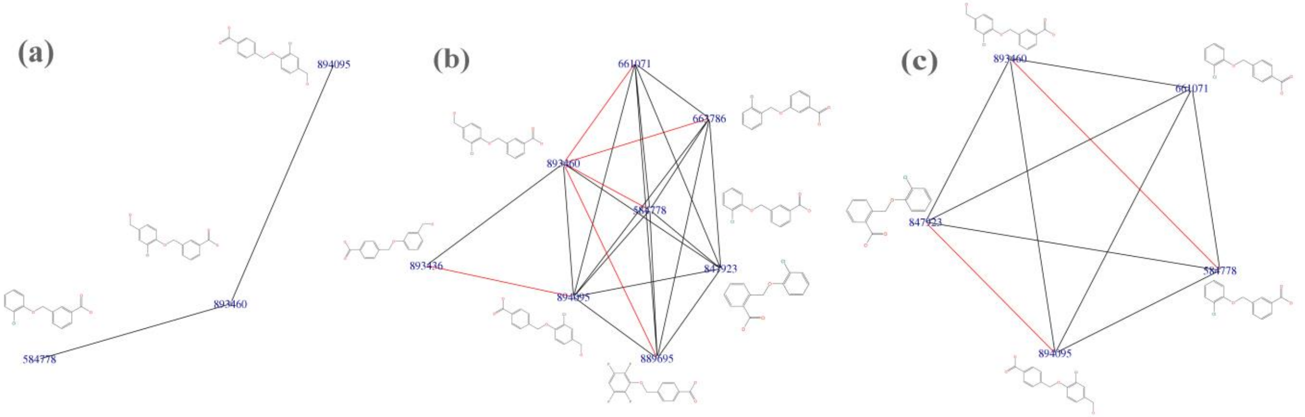
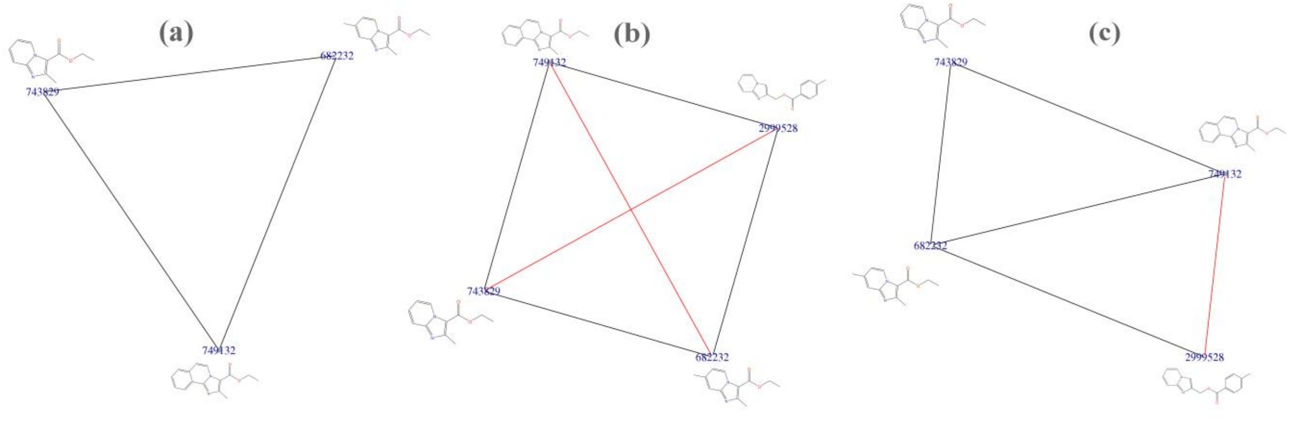


Figure 16


(a)
743829
682232
749132
(b)
749132
2999528
743829
682232
(c)
743829
749132
682232
2999528