# Exploring the Role of Small Molecules in Biological Systems Using Network Approaches

Sourav Das[†] and Rajarshi Guha[‡]
[†]St. Jude Children's Research Hospital
262 Danny Thomas Place, Memphis, TN 38105
[‡]National Center for Advancing Translational Science
9800 Medical Center Drive Rockville, MD 20850

**Abstract**

A nice abstract

## 1 The Role of Networks in Drug Discovery

Omics technologies are enabling us today to elucidate and interrogate complex relationships between proteins, genetic material and small molecules at an unprecedented speed and scale. Interrogation of relationships reveal common mechanism of action. This may result in a therapeutic candidate finding application in a seemingly unrelated disease that share a previously unknown mechanism of action. This type of relationship elucidation may also prevent unwanted side-effects if interrogation would reveal an unexpected link to a vital pathway that ought to remain unperturbed. The process of interrogation starts with experiments aimed at elucidating molecular mechanisms and pathways such as yeast two-hybrid techniques, protein and small molecule mass spectrometry, DNA microarrays and copy number variation studies. On the chemistry side, experiments may include high throughput screening of millions of chemical compounds against biological targets of therapeutic interest. These experiments result in pairwise interaction relationships from which the functional big picture biological network can then be elucidated [**?**]. Community structure algorithm [**?**], for example, allows one to discover a set of sparsely interconnected subsets of vertices, the subsets in turn however being densely connected. In the biological context, this allows connecting a set of biological cycles and smaller pathways into a big picture biological network from which inferences on cross-reactivity can be drawn.

## 2 Handling Small Molecules in R

Though small molecule are represented in text based formats (SMILES, SDF, MOL2, InChI, etc.), the R environment does not support parsing and subsequent manipulation of small molecule structures. The traditional approach has been to compute numerical features for small molecules and perform cheminformatics related manipulation outside R and import the results of such operations into the R workspace. An alternative approach is to integrate cheminformatics toolkits such as the CDK [**?**], RDKit or Indigo into the R environment. Currently there are two packages that support this - `ChemmineR` [**?**] and `rcdk` [**?**]. In this section we briefly review the functionality provided by the `rcdk` package, which is an idiomatic R wrapper around the CDK Java library, primarily focused on manipulating molecular structures.

Working with small molecule structure data can be broadly grouped into three tasks - input/output, manipulation & modification of structures and computations on molecular structures. The `rcdk` package supports input of all chemical file formats supported by the CDK, including SMILES, MOL MOL2, SDF and PDB formats. When used with the `rinchi`, input and output of the InChI format is also supported. Chemical structure files can be read locally or over the internet. In addition to generic file loading, helper functions are available for the SMILES format, given its ubiquitous use as a structure exchange format.

Molecular structures are loaded in as references to Java objects. That is, they are not native R data structures and thus can only be manipulated using methods from the `rJava` package. While inconvenient, the low level details are hidden from view and the `rcdk` package provides an idiomatic R interface to various CDK methods that operate on molecules, bonds and atoms. For example, given a SMILES string one can count the number of aromatic atoms using the following R code

```
mol <- parse.smiles('CNCc1ccccc1')[[1]]
length(which(sapply(get.atoms(mol),
    function(atom)
        is.aromatic(atom))))
```

Other methods support operations on bonds, identifying substructures and retrieving atom or bond properties.

From an analytic point of view, functions that perform computations on molecular structures are probably the most useful. Such computations can range from evaluating molecular descriptors [**?**] and generating fingerprints to computing a variety of similarities and so on. The `rcdk` package provides a simple interface to descriptor and fingerprint calculation. Importantly, the results from these functions can easily be employed in a network based approach when coupled with `igraph`. As an example, consider a similarity network constructed for 100 molecules, using a Tanimoto

A cmmer

cutoff of 0.75

```
library(rcdk)
library(igraph)
library(fingerprint)
mols <- load.molecules('data/mipe-std.smi')
fps <- lapply(mols, get.fingerprint)
smat <- fp.sim.matrix(fps)
smat[ smat < 0.75 ] <- 0
g <- graph.adjacency(smat, mode='undirected',
    weight=TRUE, diag=FALSE)
```

# 3 Linking Small Molecules to Targets, Pathways and Diseases

Since small molecules are fundamentally weighted networks (of atoms and bonds), a variety of graph algorithms can be applied to them to derive invariants (also known as topological descriptors [**?**]). However, small molecules interact with protein targets and the targets themselves interact with each other. Such explicit interactions lend themselves naturally to network representations. More generally, observed or computed relationships between small molecules, protein or gene targets and diseases allow one to develop network representations which can then be visualized and quantified to support integrative analyses of multiple data types. In the following subsections we highlight applications of this approach and where possible provide examples of R code to generate and analyze such networks.

## 3.1 Drug-target networks

## 3.2 Disease networks

## 3.3 SAR networks

Network representations have been used to characterize structure-activity relationship (SAR) datasets, most notably in the context of activity cliffs [**?**] - compounds that are structurally similar, but show very different activities. In the approach described by Guha and Van Drie [**?**], a Structure Activity Landscape Index (SALI) value is computed for each pair of molecules in a dataset. The resultant matrix of SALI values is then visualized in a network form, where compounds (nodes) are connected by an edge if their SALI value lies above a user specified threshold (larger SALI values imply a bigger activity cliff). An example of such a network is shown in Figure 1. The represen-

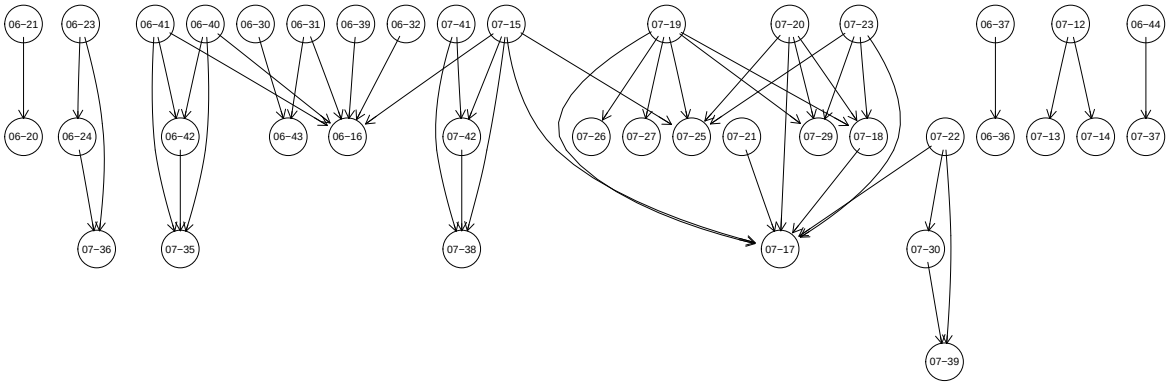


Figure 1: A SALI network constructed from XXX. See Ref. [**?**] for further details

tation allows one to rapidly zoom in on compound pairs that exhibit activity cliffs. In addition, the network representation was also employed as a way to quantify the ability of a predictive model to correctly predict activity cliffs [**?**].

Network similarity graphs (NSG's) are an alternative network visualization of SAR data described by Wawer et al [**?**] that employs a different characterization of activity cliffs. This approach is designed to multiplex information by combining connectivity (representing similarity relationships) with node color (potency) and size (discontinuity score [**?**]). This approach has been implemented in the SARANEA tool [**?**] and extended [**?**] the NSG concept to include mechanism of action information.

Networks have been applied to a number of other SAR related problems. For example, Webb et al [**?**] proposed a network approach to the interpretation of arbitrary machine learning models. Hanser et al [**?**] constructed a network model to represent SAR knowledge data, resulting in a "hypothesis network". A key feature of this approach is the ability to integrate disparate data types and build predictive models on top of the network. Krein and Sukumar [**?**] describe an analysis of chemical spaces using a graph representation and compare chemical spaces using network metrics.

## 3.4 Assay networks

Most high throughput screening campaigns tend to involve multiple assays. These usually include a primary screen followed by one or more secondary screens as well as counter-screens. When data from such campaigns are deposited into public databases such as Pubchem, the temporal sequence of the assays is not always maintained. As a result it can be difficult to track the progression of chemical matter through the series of assays. In addition, a view of the complete assay sequence can allow one to detect commonalities between programs and better compare the performance of

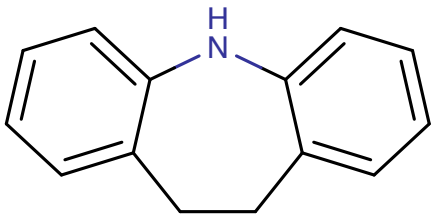


Figure 2: An example of a scaffold. In this case, this scaffold is common to many tricyclic antidepressants

screening workflows. Calhoun et al [**?**] describe an approach to reconstructing screening workflows (also termed workflow graphs) from screening datasets. They employed four heuristic rules to construct graphs from the data as well as using Pubchem Bioassay meta data and text mining to identify assays from a given project. These assay networks can be visualized using an online tool located at http://swami.wustl.edu/flow. Swamidass et al [**?**] describe an approach to constructing "assay networks" based on Pubchem data, where assays are connected when they exhibit common active (non-promiscuous) compounds. In this approach connected assays are correlated since they exhibit compounds with similar activities. This method leads to connections between apparently unrelated screen allowing one to identify compounds that exhibit polypharmacological behavior (i.e., activity against multiple targets) and thus are candidates for repurposing. By also labeling screens as phenotypic or biochemical, this approach allows one to propose target candidates based on connectivity, thus serving as one approach to target deconvolution.

## 3.5 Scaffold networks

The notion of a scaffold is common place in medicinal chemistry and is usually considered to be a ring-containing substructure pruned of sidechains that represents the core structure of a series of molecules. An example is the tricyclic core common to a number of antidepressants (e.g., amitriptylene and imipramine) shown in Figure 2. Scaffolds allow one to summarize a collection of molecules (such as in patent claims) and one can usually assume that compounds with the same scaffold will share a common synthetic pathway or even a common mechanism of action.

While a number of approaches have been described to generate scaffolds [**?**, **?**, **?**], a few approaches have considered hierarchical decomposition of scaffolds. Examples include the HierS method [**?**] and the Scaffold Tree method [**?**]. The latter generates a series of scaffolds based on an iterative removal of rings. The output of this method is a series of scaffolds arranged in a tree-like structure whose root node represents the simplest ring obtainable from the starting scaffold. Two scaffolds are connected if there exists a parent–child relationship between the two. Scaffold trees

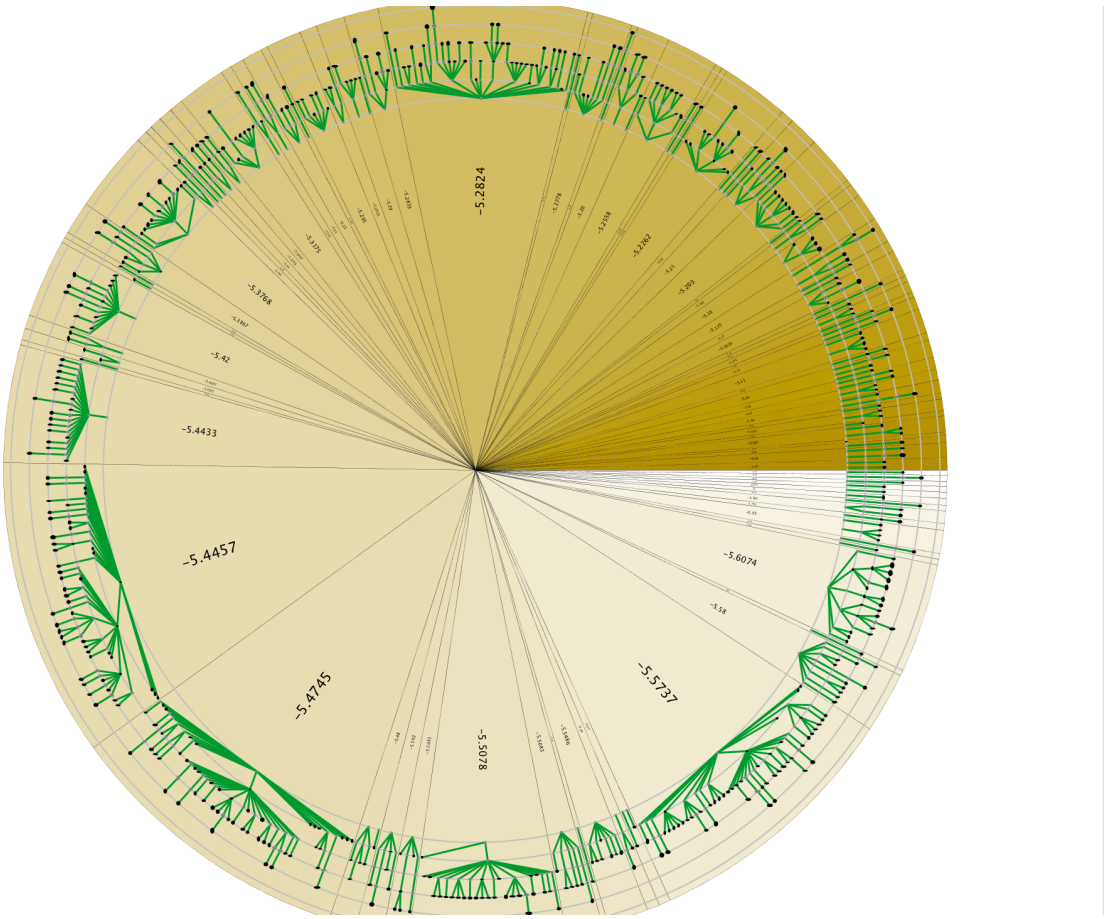


Figure 3: An example of a scaffold tree displayed as a radial network, generated from a set of pyruvate kinase inhibitors.

generated by this method lend themselves naturally to network visualizations. Scaffold Hunter [**?**] is an example of a tool that employs a network representation of scaffold trees. Figure 3 shows an example of a scaffold tree generated from a set of 602 pyruvate kinase inhibitors. The graph is colored based on $\log AC_{50}$ and the data is obtained from Pubchem AID 361. This visualization lends itself naturally to adding overlays such as molecular properties either on leaf nodes or else as aggregated values on non-leaf nodes. The scaffold tree has been applied in a number of studies [**?**, **?**] and the structure of the tree can be used as a guide to exploring chemical space, focusing on regions (as represented by a scaffold) where there is minimal activity (a "biological hole") or high activity.

Varin et al [**?**] describe "scaffold networks" which are similar in nature to scaffold trees but instead of a tree structure where each scaffold has a single parent, a DAG is constructed by considering all scaffolds at a given level of the hierarchy. An example of a scaffold network generated from Alosetron is shown in Figure 4, where the blue structures are those that would be generated using the Scaffold Tree algorithm and the green structures would be the extra scaffolds generated

using the Scaffold Network algorithm. Using this approach, together with the compound set enrichment method [**?**] the authors were able to efficiently identify active scaffold and active molecules

### 3.6 Dynamic networks

# 4 R as a Platform for Computational Drug Discovery

# 5 Discussion

- Networks on the chemistry side tend to be more for visualization
- In many cases the network does not necessarily lead to nove output
- A key utility of network approaches is the ability to integrate different data types such as small molecules and proteins
- Networks represent an approach to representing multiple relationships in a single construct - multiple properties of a relationship between two entities can be encoded by edge properties (width, color, length). Similarly multiple properties of a given entity can be encoded in node properties (size, shape, color)

# References

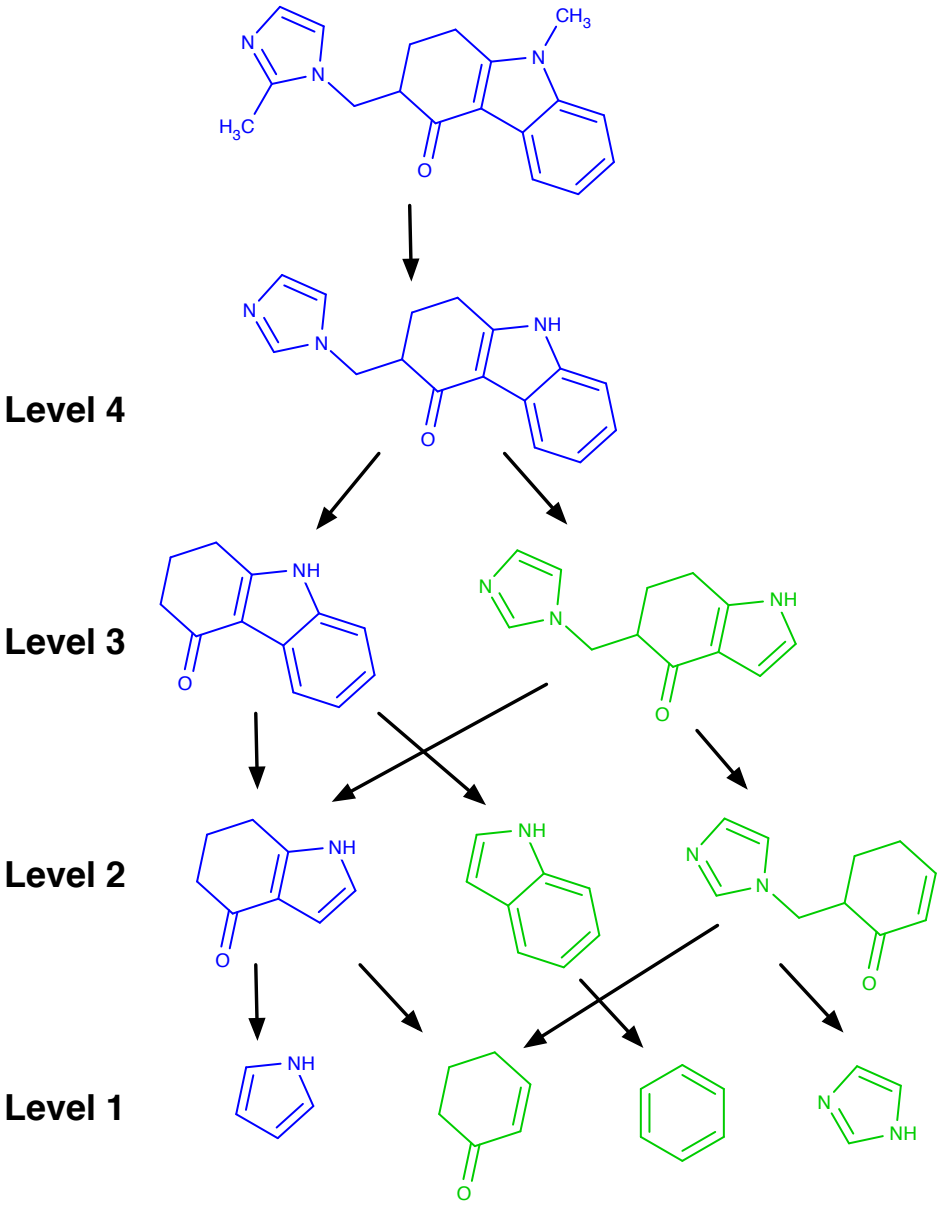


Figure 4: An example of the scaffolds generated using the Scaffold Tree and Scaffold Network algorithms, starting from Alosetron. Blue scaffolds are generated by both methods and green scaffolds are generated only by the Scaffold Network method. Modified from Figure 1 in Varin et al [**?**].