***Title*: Systems Chemical Biology and the Semantic Web: what they mean for the future of drug discovery research**

Authors: David J. Wild[1], Ying Ding[2], Amit P. Sheth[3], Lee Harland[4], Eric M. Gifford[5], Michael S. Lajiness[6]

[1]Indiana University School of Informatics and Computing, Bloomington, IN
[2]Indiana University School of Library and Information Science, Bloomington, IN
[3]Wright State University Department of Computer Science and Engineering, Dayton, OH
[4]ConnectedDiscovery Ltd., Deal, Kent, U.K.
[5]Pfizer Global Research and Development, Groton, CT
[6]Eli Lilly, Indianapolis, IN

Corresponding Author: Wild,. D.J. (djwild@indiana.edu)

Keywords: Semantic Web, Chemogenomics, Systems Chemical Biology

Teaser:

Systems chemical biology, the integration of chemistry, biology and computation to understand the way small molecules affect biological systems as a whole, and related fields such as chemogenomics, are central to emerging new paradigms of drug discovery such as drug repurposing and personalized medicine. Recent Semantic Web technologies such as RDF and SPARQL are technical enablers of systems chemical biology, allowing the deployment of advanced algorithms for searching and mining large-scale large integrated datasets. In this paper, we demonstrate how these technologies together can change the way that drug discovery gets done.

Traditionally, drug discovery paradigms involve identifying a protein target that is implicated in disease processes, and then identifying one or more chemical compounds that can safely interfere with these targets, either by activation (agonism) or inhibition (antagonism), that are then prioritized and further tested for safety and ultimately in clinical trials. Recent failures to bring the expected numbers of new drugs to market, along with increasing post-market drug withdrawals, have resulted in a questioning of this methodology, in particular that the "reductionist" approach is too simplistic, and is unable to properly assess risk of *in vivo* efficacy and safety problems.

Rather than reducing a complex system to simplistic models, the emerging field of *chemogenomics* seeks to build holistic models around the effects of compounds on multiple biological targets and pathways. Recent work in this area has mostly focused on identifying and predicting different aspects of small molecule-protein interactions, such as the use of chemical similarity as a probe of protein function [REF:Soichet]; the prediction of off-target effects of drugs using network methods [REF:Xie, Chang]; repurposing of known drugs for new targets [REF:Kinnings], drug-target interaction networks for exploring the Kinome [REF: Metz]; mapping of assay networks onto biological networks to relate compounds and targets [REF: chen] and using drug side-effect profiles to predict new biological targets [REF:campillos]. Whilst it is sometimes possible to have the luxury of a full matrix of experimental results of compounds against protein targets [REF:Bioinformatics-submitted, Metz], most work has focused on computational prediction based on available data. Although in its early stages as a research discipline, chemogenomics has demonstrated some early successes, including successful prediction of new targets for known drugs that are later experimentally verified [REF Soichet, Kinnings]. Chemogenomics is limited in that it only considers the relationships of chemical compounds and genes (along with their target proteins). A wider approach has been proposed that involves analyzing networks of many kinds of data including compounds, targets, genes, diseases, side-effects, metabolic pathways, and so on, with the purpose of investigating the complex systematic effects of drugs and other chemical compounds on biological systems. This field is tentatively termed *systems chemical biology* [REF:Oprea] although the term *chemical systems biology* has also been used [REF Xie]. In particular, realizing this approach requires a high level of integration of chemical and biological databases and of new kinds of computational tools to use these integrated databases. Lack of such integrated data sources hampers research in chemogenomics and systems chemical biology, and makes it difficult to replicate published research on other datasets [REF WildEODD]

The case has been made previously (including in this journal) for the large-scale integration of heterogeneous datasets, and that this integration must be *semantic*, i.e. there must be a shared understanding of meaning of and accessibility to tools across the datasets [REF Slater, Guha, EODD]. Such integration is a necessary precursor to systems chemical biology, particularly given the diversity of large public datasets now available describing chemical and biological entities and the relationships between them (PubChem, ChemSpider, UniProt, ChEMBL, KEGG, to name a few). However, such integration does not effect systems chemical biology: new kinds of algorithm and tools are needed that use these integrated sets, and new methodologies are needed to map these algorithms and tools to real drug discovery problems. In fact, then, a *stack* of capabilities, based on integrated data, is needed (as depicted in Figure 1).

Semantically integrated networks of data

Traditionally, data integration in pharmaceutical research has been achieved by developing schema that map relational database tables together within a single database management system (a tortuous manual process), by ad-hoc merging of data files to meet a particular immediate integration need, or by employing external vendor solutions often for organization-wide data integration. However, no widely-accessible, non-commercial technology has existed, until recently, for relatively straightforward integration of heterogeneous data sets *between organizations and data silos.* Three foundational components of what we now recognize as the Semantic Web, all recommendations developed by the World Wide Web Consortium, do now constitute a common core of technologies for such integration: *RDF* (Resource Description Framework), *OWL* (Web Ontology Language) and *SPARQL*. RDF is a simple language, implementable in a variety of formats (e.g. XML), that enables the representation of pairs of entities and the relationships between them (RDF triples). Because of their simple nature, RDF triples are extremely flexible in representing any kind of relationship between chemical or biological entities. This direct representation of relationships is key to capturing semantics of data, which has been missing in the popular relational model. Further, each RDF triple can be considered as two nodes of a network connected by an edge, and so in aggregate, a set of RDF triples describes a network of entities and relationships between them. OWL (Web Ontology Language) is used to represent ontologies, providing shared nomenclature or core vocabulary, and capturing a richer model of the domain using subclass relationships and constraints. SPARQL is a language for querying RDF triples, similar conceptually to the relational SQL language but allowing powerful integrative searching (i.e. involving multiple, heterogeneous sets linked by OWL ontologies). Recently, and key to the practical implementation of Semantic Web based resources, *triple stores* are also available for fast access and searching of reasonably large sizes of data in RDF. Since 2005, LOD (Linked Open Data) has become a significant force in open sharing of data, where in large corpuses of Semantic Web data are published as bubbles in the LOD cloud, and integrated with other datasets using a set of standard protocols. These technologies have only recently reached a point of maturity where they are practically effective, as demonstrated below, leading to the Semantic Web unfortunately being rejected prematurely in some quarters as “not up to the job” of practical integration.

There are now many successful demonstrations and deployments of Semantic Web technologies biological applications both in the public sphere and in industry [REF: Cheung, Ruttenberg, ChenHua, ChoiJoo, WildEODD] although there are clearly significant research challenges ahead [REF:Dumontier1]. A 2007 *BMC Bioinformatics* paper made the case for the use of the Semantic Web in translational medical research, giving examples of its successful use [REF: Ruttenberg]: this has since become the all-time most viewed article in the journal. Escalating importance is being given to the use of such methods in pharmaceutical research, as exemplified by the recent large EU grant given to the OpenPHACTS initiative specifically for the development of Semantic Web methods for drug discovery (see www.openphacts.org) and the active and growing membership of the W3C Semantic Web in Health Care and Life Sciences Interest Group (SWHCLS) (see http://www.w3.org/2001/sw/hcls/). Bio2RDF [REF:Bio2RDF] and a rapidly-growing biological component of the LOD cloud index around 5 billion triples of biological data. A subset of the LOD cloud relevant to drug discovery is the Linked Open Drug Data project [Samwald]. A recent special issue of the *Journal of Cheminformatics* included papers describing reviewed the current uses of RDF in chemistry and cheminformatics [REF Willighagen] and demonstrated its use in the LODD [REF:Samwald], providing open toxicology data [REF:Jeliazkova], creating an open QSAR framework [REF:Chepelev], semantic text mining of journal articles [Hawizy], and in describing chemical structure and reaction data [RAF:Chepelev2].

A variety of triple-store technologies are now available for practical implementation. Examples of demonstrated scalabilities of current systems are maintained on the W3C Consortium (see www.w3.org/wiki/LargeTripleStores). Experiments demonstrate the ability to store and search tens of billions of RDF triples in real time, with current implementations easily able to store and search several hundred million triples on a small server implementation: however it remains to be seen how well current systems scale in production environments.

RDF triple stores are made significantly more useful by the employment of ontologies, usually in the OWL language, which structure the allowed content of the RDF triple statements. Without an ontology, links between heterogeneous sets are mostly limited to “same-as” statements (e.g. “compound in set X is the same as drug in set Y”) but with an ontology individual data fields can be mapped to higher level classes which may be described differently between sets (for example, to distinguish an IC_{50} from a percent inhibition). Attempts to create grand-scale ontologies (for instance to cover the whole of chemistry) have generally failed in the life sciences due to problems of complexity and fuzzy boundaries with other disciplines. However, there are several well-used ontologies available that have a wide scope, including the Gene Ontology. The most successful approaches to ontologies now seem to be to use existing ontologies where possible, and to build new ontologies for specific purposes to “fill the gaps” with proper linking to existing ontologies. This is facilitated by open ontology portals, most notably, for the life sciences, OBO (obofoundry.org) and NCBO BioPortal (bioportal.bioontology.org). The latter has well over 250 ontologies at the time of writing. For industry use, it is also valid to develop internal ontologies closely mapped to internal data sources, but externally linked to other public ontologies to promote integration between internal and external data.

Integrative tools and algorithms

When drug discovery data is represented in RDF format in a triple store, the most basic kind of searching is to use a *SPARQL endpoint*, i.e. an access point for searching the RDF data using the SPARQL language. This approximately maps to using SQL to search a relational database, but it is much more powerful (especially if an OWL ontology is employed), as it permits searches that span heterogeneous sets all in a single query. Inference is supported which can, for example, allow the use of a single general class of drug to be mapped into all its subclasses and variants. Demonstrated examples of this integrated searching include finding compounds with similar polypharmacology profiles to a known drug, suggesting multiple target inhibitors of MAP-Kinase, and the identification of metabolic pathways with multiple gene associations that map to a given side effect [REF Chem2Bio2RDF]. SPARQL has significant limits though: in particular, it is primarily a searching language, and thus does not provide access to advanced data mining algorithms. It is also a complex language for humans to learn, relegating its use to computing specialists more than end-user scientists. Use of ontology-supported graphical query formulation tools such as Cubee [REF Cubee] now make it significantly easier to give a scientist access to the more powerful capabilities of the Semantic Web without learning new languages.

The first generation of generic Semantic Web tools have been designed primarily for browsing, visualizing and searching of RDF data but are now being developed with more powerful tools such as hypothesis testing. Topbraid (www.topquadrant.com) is a series of tools for the integration of existing internal data sources into RDF-based formats, and for utilizing these integrated data. IO Informatics (www.io-informatics.com) produces a suite of software designed specifically for life science users for integrating heterogenous data into RDF format, and then visualizing and searching this data in a variety of ways. Franz Allegrograph (www.franz.com) combines a cloud-enabled RDF triple store (which it is claimed can handle over 300 billion triples) with tools for SPARQL searching of the data, visualization, and limited reasoning capabilities. The RDFScape project (www.bioinformatics.org/rdfscape) adds Semantic Web features to the free Cytoscape network visualization tool, allowing it to query, visualize and reason on ontologies represented in OWL or RDF within Cytoscape. A number of free generic RDF graph visualization tools are available including RDF Gravity (semweb.salzburgresearch.at/apps/rdf-gravity), SIMILE (similie.mit.edu) and Triple Map (www.triplemap.com).

When dealing with networks of data, it is useful to be able to apply graph mining algorithms such as breadth/depth first search and shortest path finding. Methods for handling graph theoretic querying [REF:Anyanwu2] and semantic association finding [REF:Anyanwu3] have been previously described and an algorithm for computing semantic associations has been recently applied to RDF drug discovery data [REF:He]. This allows multiple shortest or otherwise meaningful paths between any two entities in a network to be identified. This has been implemented, along with the BioLDA algorithm for literature mining [REF:Wang] into a prototype association search tool that shows, for any pair of entities, the network paths between them that have the highest level of literature support (as measured by KL-divergence). This has

shown promise for suggesting gene associations that can account for a drug's side effects or interactions with a disease. An example of this is given in Figure 3, which shows the gene-based associations between one of the drugs of the thiazolinedione class, *Rosiglitazone* (Avandia), and the side effect of *Myocardial Infarction*. This is significant as Rosiglitazone has been found to have rare but serious cardiac side-effects, and thus this provides a mechanism for suggesting potential gene actors in the process. This association-finding tool is now being implemented in a variety of systems at Pfizer.

Graph-theoretic analysis can also be used to predict new associations based on an existing graph. Eli Lilly has employed a tool called Chemogenomic Explorer, based on a previous profiling tool [REF WENDI] that uses a rule-based inference engine to suggest potential disease associations for new compound [REF CE]. Based on manually curated rules, "evidence paths" (chains of linked RDF statements) linking compounds and genes are created which then in aggregate represent a cluster of independent or semi-independent evidence linking a compound to a disease. Such "evidence clustering" may be important as a way of mitigating the risks of errors in data, as well as the known propensity for individual pieces of published medical research to be later proved incorrect [REF:Ioannidis]. Probabilistic methods can also be applied to networks to provide a quantitative measure of association between any two entities based on the semantics and topology of the network. An ongoing project at Indiana University is investigating the use of such methods for the prediction of "missing links" in networks, and also as a virtual screening method. Already-published methods, such as the SEA analysis [REF:Soichet], may also be used for this purpose.

RDF also offers the possibility of encoding data in scholarly publications, and then applying algorithms to mine this data. In recent work [REF:Wang], a database of recent PubMed abstracts (for the last 4 years) was analyzed to identify *Bioterms*, i.e terms that can be associated with chemical and biological entities that already exist in OWL ontologies (compounds, drugs, genes, etc). These Bioterms constitute an RDF association that can be mined. A *Latent Dirichlet Allocation* algorithm was used to identify latent topics in the PubMed literature based on these terms, which are then used to create a measure of distance between entities (via topics) known as KL-divergence.

Knowledge discovery processes & biomedical insights

"The proof of the pudding is in the eating" and thus significant research effort must be put into the evaluation of these new integrative tools and processes in real drug discovery efforts. Thus, as well as the "horizontal" effort needed to develop a wide range of tools and algorithms, "vertical" efforts are needed to discover how these new approaches can complement existing computational approaches (such as docking, QSAR, sequence similarity searching and ligand-based virtual screening) to accelerate the discovery of new drugs for specific therapeutic purposes, and to identify the key pieces of knowledge (biomedical insights) necessary to understand disease processes. One can imagine a convergence of tools into "integrative virtual screens" that fuse and balance a variety of virtual screening methods (including network-based methods), but also specific and perhaps even unique combinations of tools being applied for

individual drug discovery problems. At the time of writing, very little research has been carried out at this tier of the stack – i.e. how the tools can be best mapped to real drug discovery problems – although work in related areas, such as data fusion and virtual screening should help.

What this means, and what needs to be done

We believe the work we describe here constitutes a first step in realizing systems chemical biology, that is in providing a progressive framework for the development of integrated data resources, algorithms and tools, and knowledge discovery processes that combine systems chemical biology with more traditional approaches. Data integration efforts using RDF are well underway both in the public sphere and internally in pharmaceutical organizations, but care does need to be taken that these efforts are at least sufficiently aligned that mapping entities between repositories is straightforward (for example, by maintaining PubChem identifiers for internal repository compounds that are also externally available). Key to this are collaborative efforts such as W3 SWHCLS, the Pistoia Alliance, and OpenPHACTS, along with publicly available open resources such as Chem2Bio2RDF and Bio2RDF. Critical also is the separation of tools from data: historically many tools and algorithms have been developed to work on specific datasets or repositories, and are not easily extensible to other sets. This must be addressed: in particular, tools should not be critically dependent on a specific ontological mapping in a set.

Addressing quality is an essential step, and one that is fraught with numerous complexities. Example questions that demonstrate this challenge are: is a PubChem BioAssay IC_{50} result comparable with one in CheMBL or from an internal assay? Is an experimental result always more significant than a predicted result or an association extracted from a journal article? What happens when we get so many links between things that we can't separate the signal from the noise? Ultimately, we are constrained by the data sources available: we have a choice which datasets to include or exclude, and we have methods (such as provenance tracking) for keeping track of the history of a piece of data, but we are bound by the quality of whatever data we choose to use. Quality should thus be addressed primarily at the tool level, allowing users to select which datasets they are comfortable using, and understanding the caveats in doing so. There is a case in some instances for using only very limited datasets of known quality, and at other times using all available data. Ideally it will be possible to make such quality determinations within tools and environments in a meaningful way. There is also a need for research into the use of multiple semi-independent evidence paths found in networks of data as a way of "building consensus" that mitigates quality issues in specific data sources (which in turn makes a case for improved provenance tracking in Semantic Web implementations [REF Sahoo,Sahoo2]).

Once we can apply validated integrative tools and algorithms freely on data of our choosing from the full breadth of available information, the problem becomes one of what the right questions are to ask of the data, and how to interpret and follow up on the results. This can only be done by the practical application of these methods in real drug discovery problems. Ideally,

research efforts will occur in academia (and perhaps in pre-competitive collaboration between industry and academia) so that effective integrative methodologies for drug repurposing, can be publicly validated.

In the near future, emerging patient-level datasets, including those derived from Electronic Medical Records (EMRs), Next Generation Sequencing / Genome-Wide Association Search (GWAS) tools, and metagenomics will massively increase the available data and will thus introduce issues of scale that will need to be addressed both at the triple-store and algorithm level. However, these sets will also provide the opportunity to gain understanding of how individuals will respond to drugs, rather than the body as a generic entity. Research is needed at the interface of these datasets with existing chemical, biological, and pharmacological sets, to provide a public corpus of data that in aggregate will form a biomedical map that bridges the molecular and clinical – "from molecule to man".

If we assume that successful discovery on new safe, effective drugs is going to require that we step beyond the "lock and key" model of drug and protein target to understand the much greater complexities of how drugs interact with the body, realizing the emerging disciplines of chemogenomics and systems chemical biology through enabling integrative technologies (such as the Semantic Web) is going to be a critical foundation to success in 21st century drug discovery. Promising efforts are already underway, but there is much more basic research and industry-academia collaboration required to accelerate progress in these fields.

References

[Soichet] Keiser, M.J. *et al*. (2009) Predicting new molecular targets for known drugs. *Nature*, 462,175-181.

[Xie] Xie, L. *et al*. (2011) Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* Web publication date Feb 1, 2011.

[Chang] Chang, R.L. *et al*. (2010) Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model. *PLoS Comput. Biol.*, 6, 9

[Kinnings] Kinnings S.L. *et al*. (2010) Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.,* 5, 7.

[Metz] Metz, J.T. *et al.* (2011) Navigating the Kinome, *Nat. Chem. Biol.* Web publication date Feb 20, 2011.

[Chen] Chen, B. *et al.* (2009) PubChem as a source of polypharmacology. *J. Chem, Inf. Model*. 49, 2044-2055.

[Campillos] Campillos, M. *et al*. (2008) Drug Target Identification using Side-Effect Similarity. *Science*. 321(5886) 263-266.

[Oprea] Oprea, T.I. *et al*. (2007) Systems Chemical Biology. *Nat. Chem. Biol.*, 3, 447-450.

[WildEODD] Wild, D.J. (2009) Mining large heterogeneous datasets in drug discovery, *Expert Opinion on Drug Discovery*, 4(10), 995-1004.

[Slater] Slater, T. *et al*. (2008) Beyond Data Integration. *Drug Discovery Today*, 13, 584-589

[Guha] Guha, R. *et al*. (2010) Advances in Cheminformatics Methodologies and Infrastructure to Support the Data Mining of Large, Heterogeneous Chemical Datasets. *Current Computer-Aided Drug Design*. 6, 50-67.

[Cheung] Cheung, K.H. *et al*. (2011) Semantic Web for Health Care and Life Sciences: a review of the state of the art. *Briefings in Bioinformatics*, 10, 111-113

[Ruttenberg] Ruttenberg, A. *et al*. (2007) Advancing translational research with the Semantic Web, *BMC Bioinformatics,* 8(3), 1-16

[ChenHua] Chen, H. and Xie, G. (2010) The use of web ontology languages and other semantic web tools in drug discovery, *Expert Opinion on Drug Discovery*, 5(5), 413-423

[ChoiJoo] Choi, J. *et al*. (2010) A Semantic Web Ontology for Small Molecules and Their Biological Targets, *J. Chem. Inf. Model.,* 50(5), 732-741.

[Dumontier1] Dumontier, R. (2010) Building an effective Semantic Web for Health Care and the Life Sciences. *Semantic Web - Interoperability, Usability, Applicability*. Special Issue: Vision Statements.

[Bio2RDF] Belleau F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*. Oct;41(5):706-16.

[Chem2Bio2RDF] Chen, B. *et al.* (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255.

[Willighagen] Willighagen, E.L. and Brändle, M.P. (2011) Resource description framework technologies in chemistry. *J. Cheminf.*, 3:15

[Samwald] Samwald. M. *et al*. (2011) Linked open drug data for pharmaceutical research and development. *J. Cheminf.*, 3:19

[Jaliazkova] Jeliazkova, N. and Jeliazkov, V. (2011) AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminf.*, 3:18.

[Chepelev] Chepelev L.L. and Dumontier, M. (2011) Semantic Web integration of Cheminformatics resources with the SADI framework. *J. Cheminf.*, 3:16

[Hawizy] Hawizy *et al.* (2011) ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminf.*, 3:17.

[Chepelev2] Chepelev L.L. and Dumontier, M. (2011) Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. *J. Cheminf.*, 3:20

[Anyanwu2] Anyanwu, K. *et al*. (2007) SPARQ2L: towards support for subgraph extraction queries in RDF databases. *Proceedings of the 16th international conference on World Wide Web*. ACM, New York.

[Anyanwu3] Anyanwu, K. and Sheth, A. (2003) ρ-Queries: enabling querying for semantic associations on the semantic web. *Proceedings of the 12th international conference on World Wide Web*. ACM, New York.

[He] He, B. *et al.* (2011) Mining relational paths in in relational biomedical data *PloS One*, Submitted May 2011.

[Wang] Wang, H. *et al*. (2011) Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One*, 6 (3), e17243

[WENDI] Zhu, Q. *et al*. (2010) WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminf.*, 2, 6.

[CE] Zhu, Q. *et al*. (2011) Semantic inference using Chemogenomics Data for Drug Discovery. *BMC Bioinformatics*,12, 256.

[Ioannidis] Ioannidis, J. (2005) Why most published research findings are false. *PLoS Med.* 2(8), e124.

[Sahoo] Sahoo, S.S. *et al.* (2008) Semantic provenance for eScience: managing the deluge of scientific data. *Internet Computing.* 12(4) 46-54.

[Sahoo2] Sahoo, S.S. *et al.* (2010) Provenance Context Entitiy (PaCE): Scalable Provenance Tracking for Scientific RDF Data. *Lecture Notes in Computer Science*. 6187, 461-470, 2010.

Figure legends

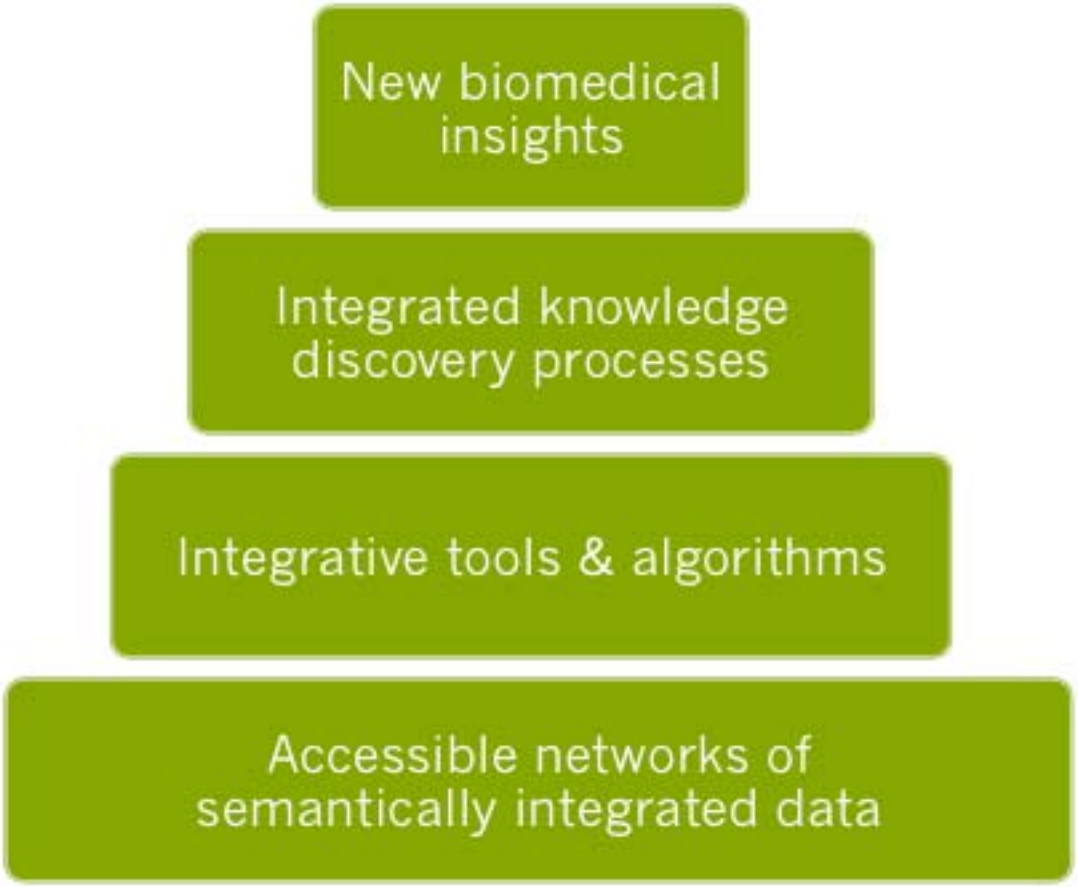

Figure 1: Stack of integrative capabilities required to realize the aims of systems chemical biology and chemogenomics.

```
PREFIX c2b2r: <http://chem2bio2rdf.org/chem2bio2rdf.owl#>
PREFIX bp: <http://www.biopax.org/release/biopax-level3.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select distinct ?target
from <http://chem2bio2rdf.org/owl#>
where
{
?chemical rdfs:label ?drugName ;
          c2b2r:hasInteraction ?interaction .
?interaction c2b2r:hasTarget [bp:name ?target];
             c2b2r:drugTarget true .

FILTER (str(?drugName)="Troglitazone")
}
```

Figure 2: Example Chem2Bio2RDF SPARQL query to identify all targets of a dug across all datasets

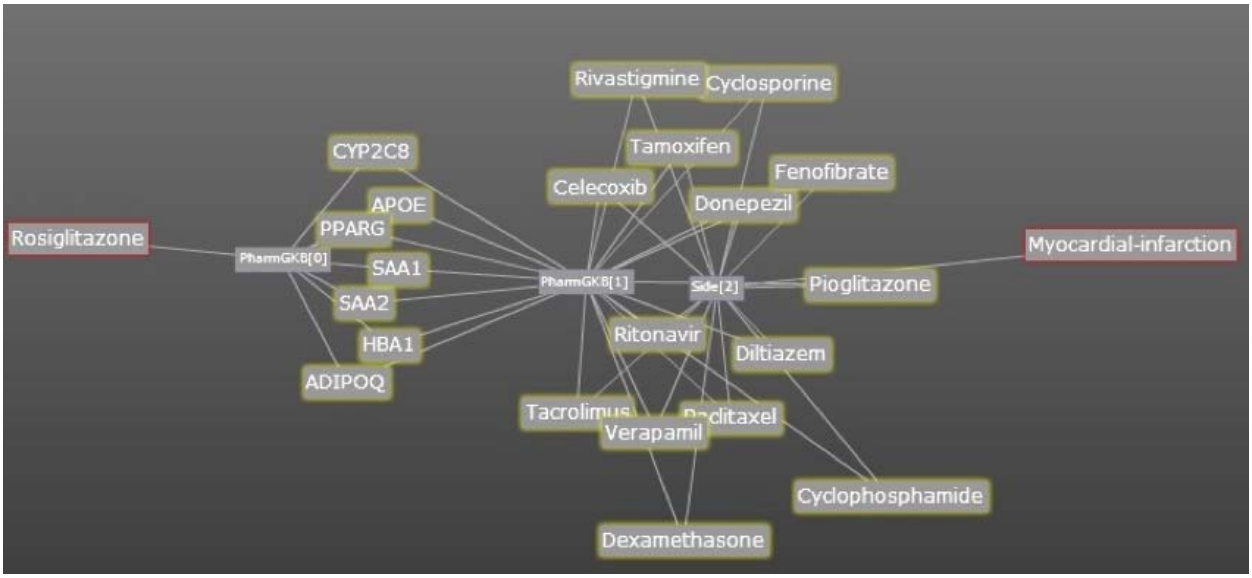

Figure 3. Constrained association search between Myocardial Infarction and Rosiglitazone showing ranked paths up to 3 edges in length that (i) contain a gene and (ii) are ranked highly by KL-divergence showing literature support