Applying Semantic and Network Methods in AOP Knowledge Discovery

David Wild, Ph.D

Associate Professor, Indiana University School of Informatics & Computing
Founder & CEO, Data2Discovery Inc

Adverse Outcome Pathways: From Research to Regulation
NIH Workshop, Bethesda MD, September 3-5 2014

Email djwild@indiana.edu | Web http://djwild.info

Purpose of this presentation and discussion

A view of what is possible when we bring together the emerging science of AOP's, and state of the art in the computational techniques of data science, semantic technologies and network science

For technical details, see SOT presentation at http://djwild.info

The Usborne Book of the Future, 1979

Semantic technologies and AOP's – a new opportunity

- Our understanding of the effects of chemicals on our body is moving from a reductionist approach to a system, network approach

- The impacts of a chemical on the body are complex
  - Multiple targets, pathways
  - Indirect cascade effects
  - Phenotype and genotype dependent

- Semantic technologies fit this model well, as a way to handle big, complex, networked data sets from multiple sources
  - Applications in drug discovery, safety and chemical toxicity

New “big” data approaches going mainstream in science

- NoSQL
  - Good for large amounts of **simple** or unstructured data
  - Very lightweight data structures e.g. tagging

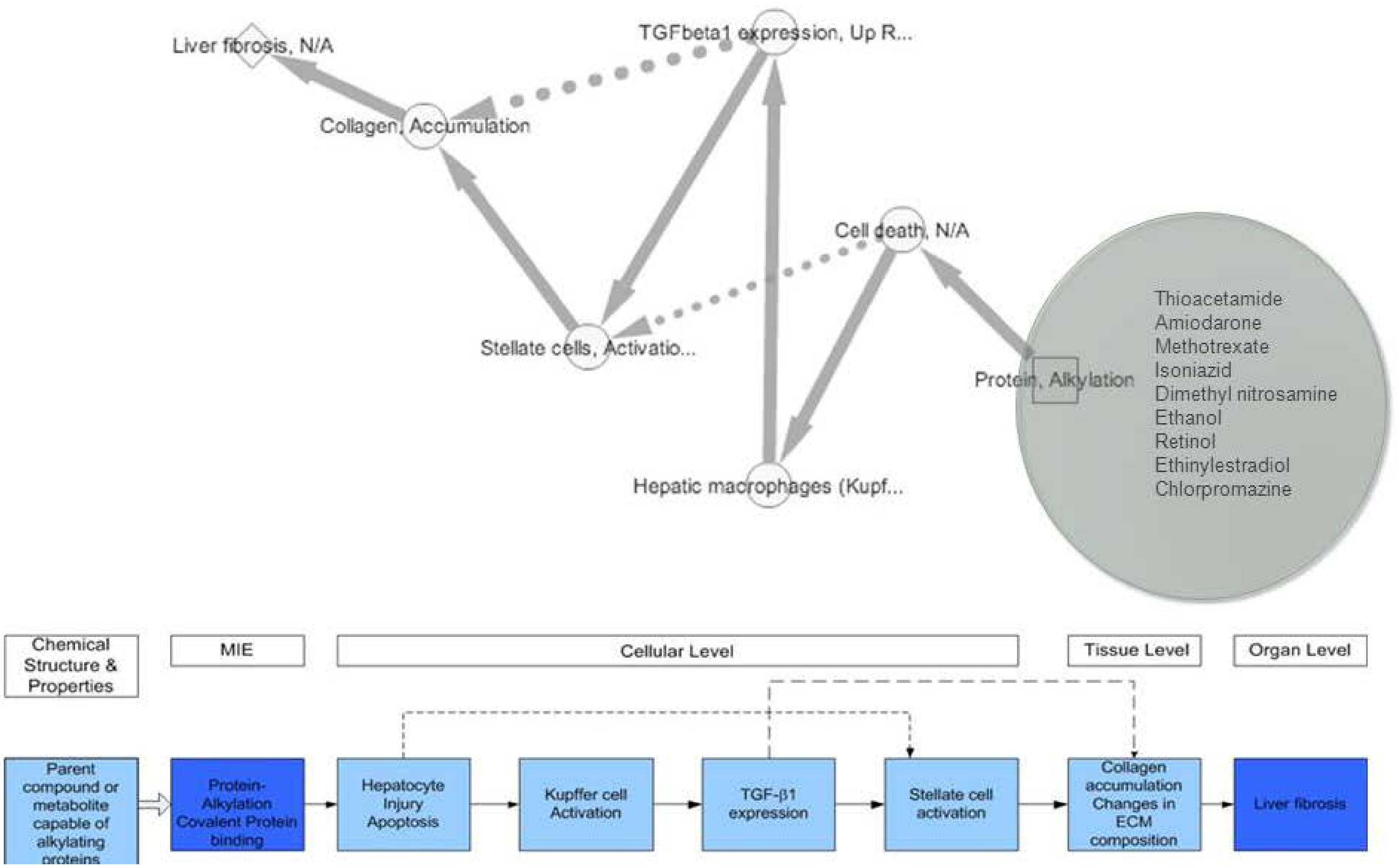
- Semantic technologies
  - Good for large amounts of **complex** data
  - Represents data as networks rather than tables
  - Highly flexible in incorporating and linking many different kinds of data
  - Ontologies apply meaning to the data and relationships
  - Identified by Gartner as one of the top technology trends impacting information infrastructure in 2013: http://www.gartner.com/newsroom/id/2359715
  - Now heavily used internally Google, Facebook, etc
  - Increasingly applied in scientific domains

Value proposition

- Semantic and network technologies could aid researchers in building AOP’s and knowledge around AOP’s
  - Predicting associations between compounds, targets and end points
  - Testing hypothesis
  - “Auto suggestion” of AOP associations

- Semantic and network technologies could help us apply established AOP’s in problems like toxicity prediction
  - Profiling compounds across toxic end-points using computational representations of AOP’s

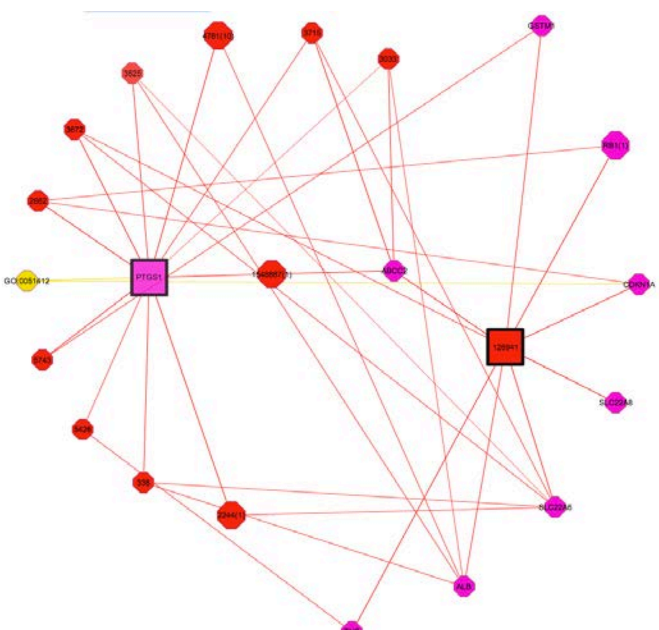
Example – Liver Fibrosis

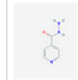
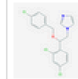
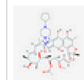
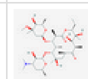
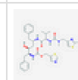
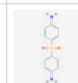
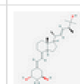
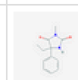


Source: AOP Wiki

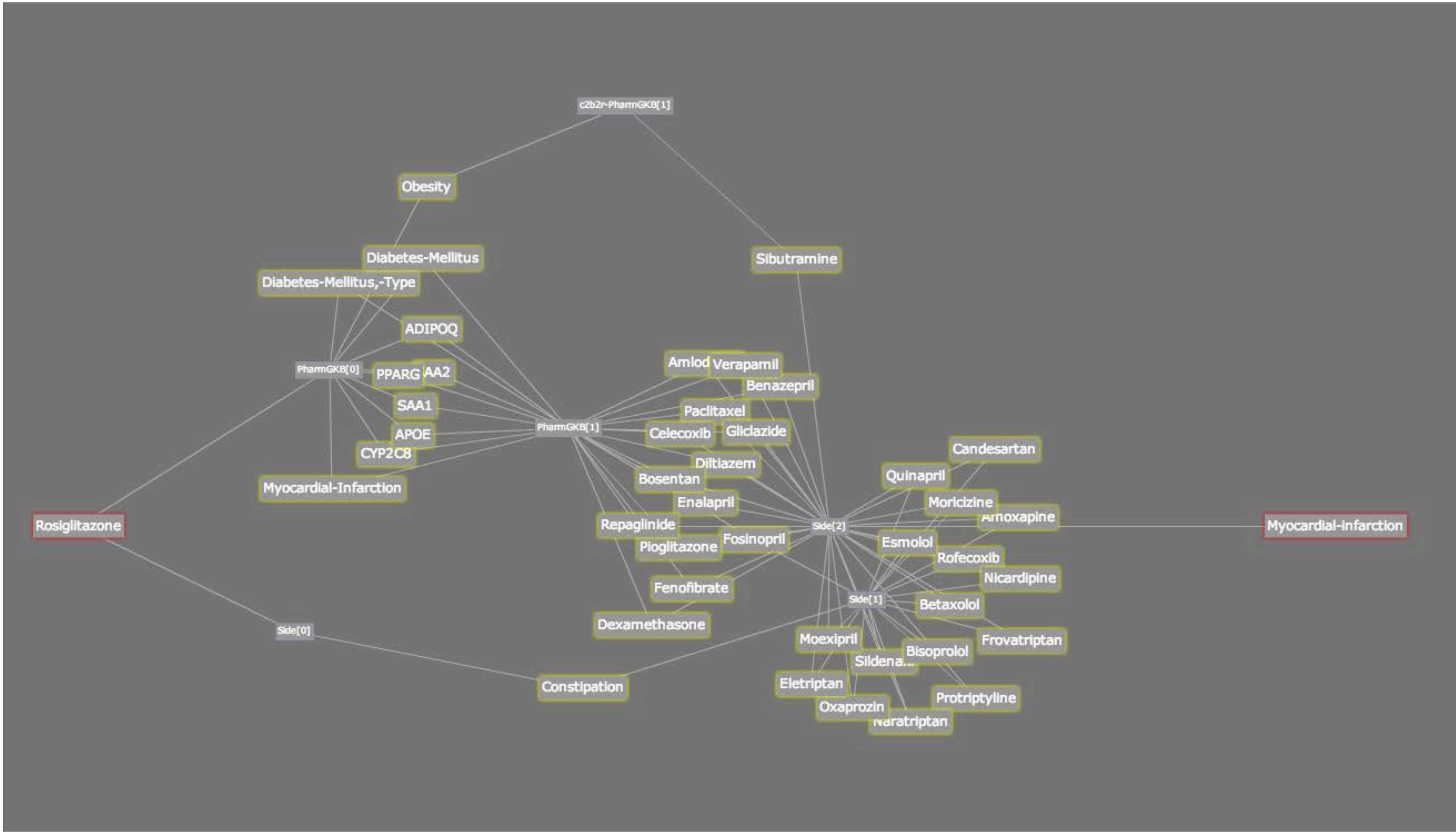
target	p value	score	type	chemohub
CYP2C9	3e-04	597.54	predicted	see paths
CYP2A6	0.002	266.26	predicted	see paths
CYP1A2	0.0027	228.29	predicted	see paths
HMOX1	0.0029	219.85	predicted	see paths
CYP2B6	0.0034	202.05	predicted	see paths
CYP3A4	0.0038	190.21	approved interaction	see paths
CYP17A1	0.0091	118.54	predicted	see paths
GSTA2	0.0432	43.14	approved expression	see paths
HMOX2	0.0744	28.28	approved expression	see paths
ABAT	0.1524	14.75	approved interaction	see paths
HPRT1	0.1983	11.16	approved expression	see paths
PLOD2	0.4368	3.81	approved expression	see paths

Isoniazid



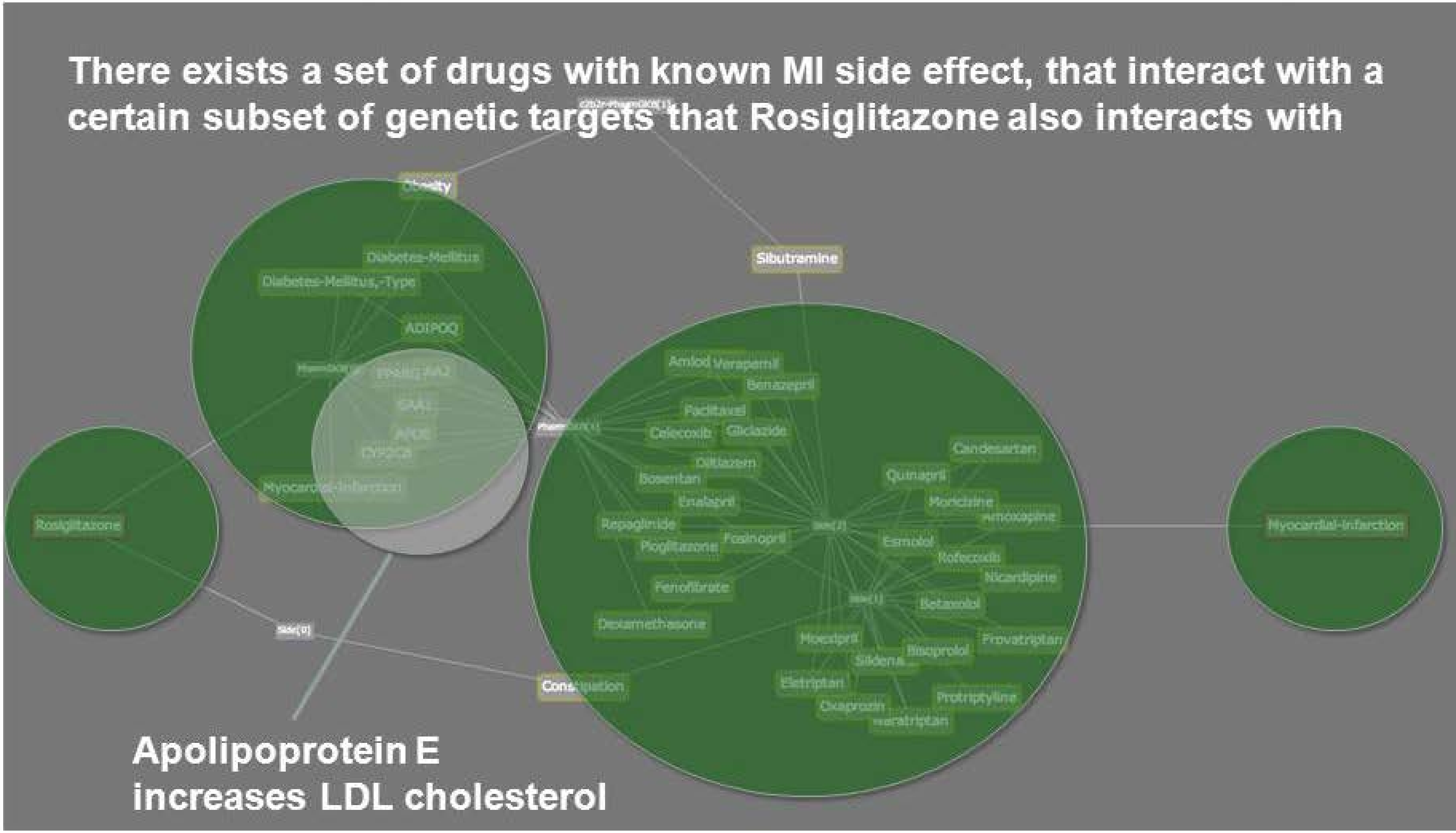
PubChem CID	structure	Drug Name	Similarity	Related Diseases	ATC
3767		Isoniazid	1	Tuberculosis	J04AC01
3198		Econazole	0.915		D01AC03 G01AF05
5362440		Indinavir	0.894	HIV	J05AE02
6323497		Rifapentine	0.885	Tuberculosis	J04AB05
5342		Sulfinpyrazone	0.873		M04AB02
4189		Miconazole	0.841		A01AB09 A07AC01 D01AC02 G01AF04 J02AB01 S02AA13
12560		Erythromycin	0.819		D10AF02 J01FA01 S01AA17
392622		Ritonavir	0.809	HIV Viral infection	J05AE03
2955		Dapsone	0.806	Leprosy	J04BA02
5281104		Paricalcitol	0.8	Hyperparathyroidism	A11CC07
4060		Mephenytoin	0.793	Epilepsy	N03AB04
55283		Itraconazole	0.783		J02AC02
5472		Ticlopidine	0.768	Stroke	B01AC05

Association graph search – finding evidence paths



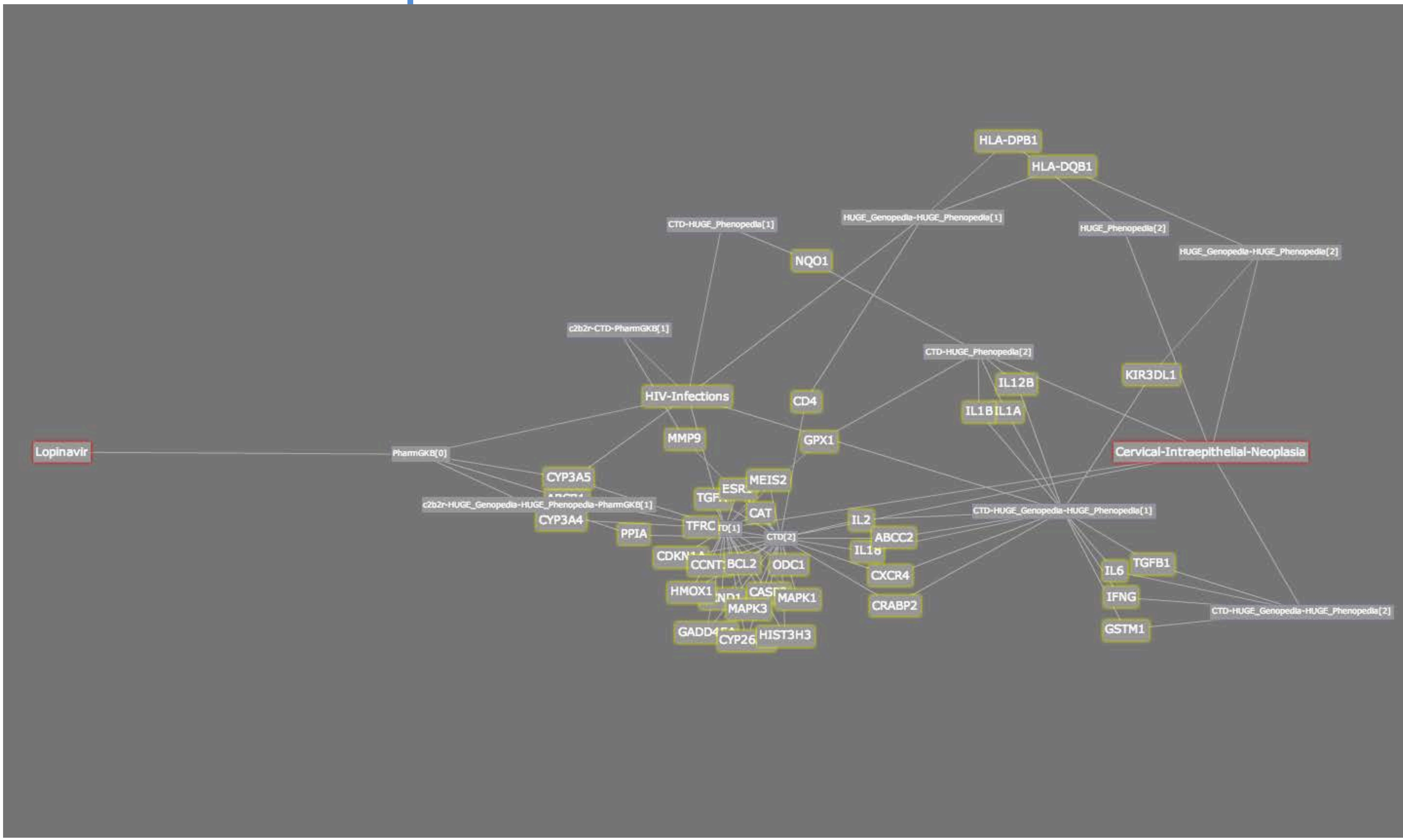
He, B., Tang, J., Ding, Y., Wang, H., Sun, Y., Shin, J.H., Chen, B., Moorthy, G., Qiu, J., Desai, P., Wild, D.J., **Mining relational paths in biomedical data** *PloS One*, 2011, e27506.

Association graph search – finding evidence paths



He, B., Tang, J., Ding, Y., Wang, H., Sun, Y., Shin, J.H., Chen, B., Moorthy, G., Qiu, J., Desai, P., Wild, D.J., **Mining relational paths in biomedical data** *PloS One*, 2011, e27506.

Lopinavir – Cervical Cancer



He, B., Tang, J., Ding, Y., Wang, H., Sun, Y., Shin, J.H., Chen, B., Moorthy, G., Qiu, J., Desai, P., Wild, D.J., **Mining relational paths in biomedical data** *PloS One*, 2011, e27506.

Application – Profiling Adverse Events

	hERG	Atrial Fib	Long QT	Tachycardia	Bradycardia	Cardiotoxicity	Cardiac failure	Cardiomyop.	Myocard. Inf.	Tachycardia	Cardiotoxicity	Long QT	Atrial Fib	hERG	Long QT	Tachycardia	Cardiotoxicity	Bradycardia	Tachycardia	Tachycardia	Cardiotoxicity	Tachycardia	Atrial Fib
A10366245																							
A10366585																							
A49585949																							
A48480494																							
Aspirin																							
Rosuvastatin																							
Pioglitazone																							
AGGREGATE																							

Why is semantic data powerful?

- Breaking down data and domain silos
  - Chemistry – biology – toxicology – adverse event - endpoint
  - Molecular – patient
  - Public – commercial – proprietary

- Easy to repurpose existing and harvest new data
  - RDF format is standard
  - Separation of the data from the structure of the data

- Semantic networks -> biological networks
  - Systems chemical biology / network biology
  - Move away from naïve drug/target or target/endpoint
  - Hugely powerful algorithms in networking community
  - Prediction, hypothesis testing, interpretation

Proposal: Semantic Toolkit for AOPs

New biomedical insights

Integrated knowledge discovery processes

Integrative tools & algorithms

Accessible networks of semantically integrated data

Semantic Toolset for AOP's

Compound-endpoint prediction and association finding

Comprehensive RDF triple store and ontologies for AOP's

Comprehensive semantic store for AOPs

- Contains all public data of relevance, from compound to organism. As a start…
  - **OnTop***: PubMed, GO, KEGG, MeSH, NCI, UniProt, Entrex Gene, NCBO, CTD, ACToR, ToxRefDB, ToxMiner, ToxCat
  - **Chem2Bio2RDF/Chem2Bio2OWL**: 52 public datasets relating to compounds, genes, pathways, diseases and side effects
  - Other relevant sets – e.g. FDAERS, social media
- Ontologically mapped to concepts in AOPWiki
- SPARQL endpoint for searching

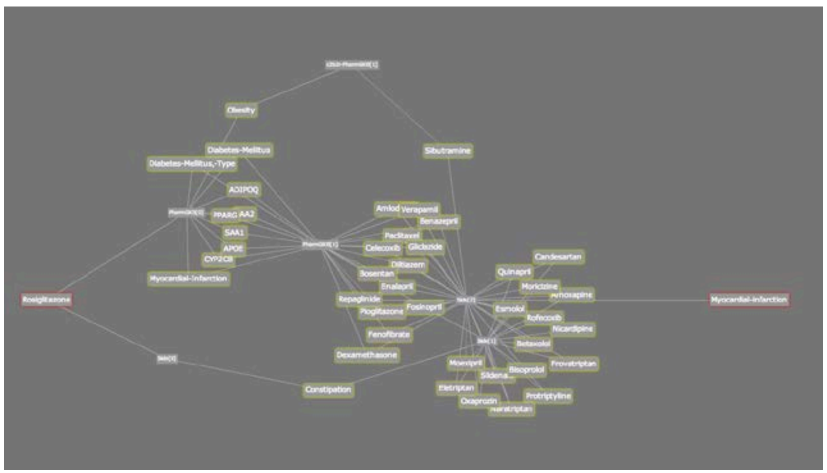
* Ontology for modeling adverse outcome pathways: semantic tools for systems tox. Imran Shah, EPA-NIEHS Advancing Environmental Health Data Sharing and Analysis: Finding a Common Language. June 25, 2013

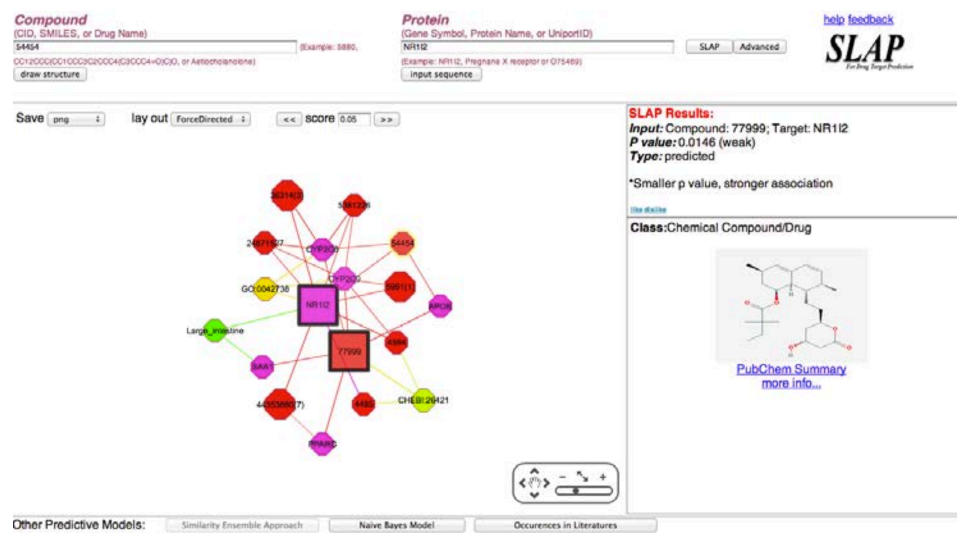
Compound-Endpoint prediction & association finding

- Predicting compound-endpoint associations with SLAP
  - Modified version of current compound-target algorithm
  - Association score and p-value
- Automatic generation of preliminary AOP networks
  - Using SLAP significant subnetwork between compound and endpoint
  - “starting point” for understanding potential AOPs
- Generation of literature supported association networks
  - More open-ended association finding and visualization
- Random-Walk methods
  - Most recent research at IU

Semantic Toolset for AOP's

- Open toolset made available in association with AOP KB and integrated with other tools
- AOP prediction, exploration, hypothesis testing & application





Summary

- Semantic technologies becoming mainstream for big / complex data problems; increasing applications in science
- IU and EPA have demonstrated applicability of semantic technologies in chemical, biological data and for AOP's
- AOP's map particularly well onto the semantic approach
- Huge potential is realized when network and predictive algorithms are applied – the "semantic stack"
- Direct opportunity to engage semantic technologies in the emerging AOP KB / AOPWiki projects