A Nanopublication Framework for Biological Networks using Cytoscape.js

James P. McCusker,[1,3] Rui Yan,[1] Kusum Solanki,[2] John Erickson,[1]
Cynthia Chang,[1]
Michel Dumontier,[4] Jonathan S. Dordick,[2] Deborah L. McGuinness[1]

***Abstract*—We leverage semantic technologies and Cytoscape.js to create a provenance-aware, probabilistic analysis platform for systems biology and evaluate its usefulness in discovering links between drugs and diseases. In our efforts to create a systematic approach to discovering new uses for existing drugs, we have developed Repurposing Drugs with Semantics (ReDrugS). ReDrugS is a data curation and publication framework that accepts data from nearly any database containing biological or chemical entity interactions and produces visualizations using Cytoscape.js. A semantic web service API is provided that enables search, traversal, and provides composite probabilities for the resulting graph of biological entities using the SADI web service framework and Nanopublications. We show how associations between a postive control, topiramate, allows us to independently reconstruct a positive control of epilepsy and migraine, and potential consequences on bone health.**

[1]Department of Computer Science, [2]Department of Chemical & Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY http://www.rpi.edu
[3]5AM Solutions, Inc, Rockville, MD http://5amsolutions.com
[4]Stanford University, Stanford, CA http://stanford.edu

I. Introduction

Drug repurposing can often lead to effective new treatments for diseases. The ReDrugs system we are developing can assist in this procedure through the integration of multiple systems biology, pharmacology, disease association, and gene expression databases into a coherent repository of individually-supported assertions that can each be assigned their own probabilistic value. We have developed an initial database that includes drug/protein, protein/protein, and protein/biological process associations that is providing us a view into how drugs have the effects that they do.

II. Methods

We deployed an instance of the RPI semantic web toolsuite, Prizms, at http://redrugs.tw.rpi.edu and the Comprehensive Knowledge Archive Network (CKAN) to http://data.melagrid.org to catalog the available datasets [1]. Cataloging is an ongoing process, but initial datasets were added to the catalog, initializing the Prizms conversion process. We were then able to use the Prizms infrastructure to generate RDF for publication to our SPARQL endpoint. We used the BigData RDF store with named graph and text indexing support enabled.

A. Inferring Probabilities

Two molecular biology and biochemistry experts, Michel Dumontier and Pascale Gaudet, assigned a score from low to high confidence of 1-3, evidence and/or technique associated with the interaction. The confidence measure was based on the comparative analyses of techniques [2], [3], and experience of the experts in reviewing data of this kind. The confidence assignment is based on a number of factors including degree of indirection in the assay, sensitivity and specificity of the approach, and reproducibility of results under different conditions. The confidence scores for both experts were encoded as classes of evidence, where each experimental method class was assigned two superclasses, one for each expert. This ontology was created from a spreadsheet and expanded to full inferences using Pellet [4]. At the same time, SPARQL-based reasoning is used to classify nanopublication assertions by their available evidence, and thereby assign a class of confidence codes to it.

B. SADI Web Service Interface

We developed four Semantic Automated Discovery and Integration (SADI) web services in Python[1] to support easy access to the nanopublications. We use SADI to provide a discoverable, consistent API that can be re-used in other applications or directly consumed by analytical tools.

The services perform these computational tasks that would otherwise be difficult to perform with SPARQL queries. The services return only one interaction for each triple (source, interaction type, target) but multiple, probabilities per interaction, and more than one interaction per interaction type. This is because the interaction may have been recorded in multiple databases, based on different experimental methods. To provide a single probability score for each triple, the interactions are combined. This is done to indicate that multiple experiments that produce the same results reinforce each other, and should therefore give a higher overall probability than would be indicated by taking their mean.

$$P\left(x_{1\dots n}\right) = \mathrm{CDF} \quad \sum_{i=1}^{n} \mathrm{CDF}^{-1}\left(P\left(x_i\right)\right)\Bigg)$$

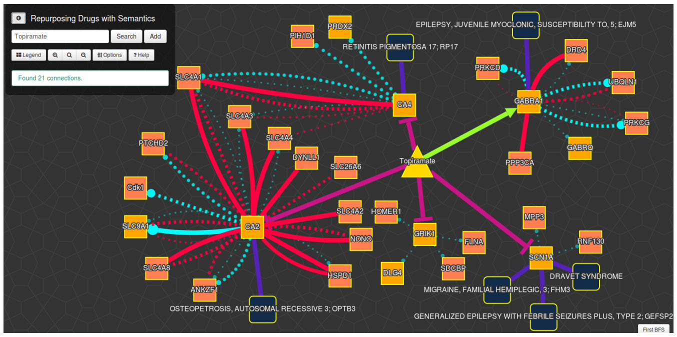
[1] For further information on developing web services in Python using SADI, see this tutorial: https://code.google.com/p/sadi/wiki/BuildingServicesInPython

C. User Interface

Users can search for biological entities and processes, which can then be autocompleted to specific entities that are in the ReDrugS graph. Users can then add those entities and processes to the displayed graph and retrieve upstream and downstream connections and link out to more details for every entity. Cytoscape.js is used as the main rendering and network visualization tool, and provides node and edge rendering, layout, and network analysis capabilities.

III. Evaluation

In order to evaluate this knowledge base, we developed a demonstration web interface[2] based on the Cytoscape.js[3]. It lets users enter biological entity names, and as the user types, the text is resolved to a list of entities to be selected. After that, the entity is submitted to all three SADI services via a basic JavaScript SADI client.[4] The resulting interactions and nodes are added to the Cytoscape.js graph, which can be laid out according to a number of algorithms. Users are also able to select nodes and populate upstream or downstream connections. An example of this is shown in Figure 1. This figure was obtained by putting "Topiramate" as a query in the search box, which returned all of the biological entities that topiramate is directly associated with. We then expanded the network downstream to see what biological entities are affected by topiramate's targets.


Repurposing Drugs with Semantics
Search
Add
Legend
Options
? Help
PIH1D1
PRDX2
RETINITIS PIGMENTOSA 17; RP17
EPILEPSY, JUVENILE MYOCLONIC, SUSCEPTIBILITY TO, 5; EJM5
DRD4
PRKCD
UBQLN1
SLC4A1
CA4
GABRA1
PRKCG
SLC4A3
SLC4A4
DYNLL1
PTCHD2
SLC26A6
Topiramate
PPP3CA
GABRQ
Cdk1
SLC4A2
HOMER1
MPP3
SLC9A1
CA2
NONO
GRIK1
FLNA
RNF130
SCN1A
SLC4A8
HSPD1
DLG4
SDCBP
DRAVET SYNDROME
ANKZF1
MIGRAINE, FAMILIAL HEMIPLEGIC, 3; FHM3
OSTEOPETROSIS, AUTOSOMAL RECESSIVE 3; OPTB3
GENERALIZED EPILEPSY WITH FEBRILE SEIZURES PLUS, TYPE 2; GEFSP2
First BFS

Topiramate

Found 21 connections.

Fig. 1. The ReDrugS user interface allows users to build networks of drugs, proteins, and diseases based on provenance-driven data from iRefIndex, DrugBank, UniProt Gene Ontology Annotations, and Online Mendelian Inheritance in Man (OMIM). Users can select entities and add entities that affect or are affected by the selected entities. They can also search for entities by name (here Topiramate was used).

IV. Discussion and Future Work

We are able to successfully navigate a protein-drug-disease interaction graph that is a consensus of 16 diverse sources, to infer prior probabilities for more than three million individual assertions using their provenance and experts' confidence in different experimental methods and to find drug/disease associations that are not directly expressed by any one database.

We plan to add further data sources, especially those that provide direct experimental results that predict protien (or gene)/disease associations like the Gene Expression Atlas (in progress) [5]. Further, we are very interested in integrating the newest version of the Connectivity Map dataset [6], as it provides gene expression signature similarities for a large number of chemical and genetic perturbations. Finally, as we develop new hypotheses about potential new drug effects, we plan to test them using a new three-dimensional cellular microarray to perform high throughput drug screening [7] with reference samples.

V. Conclusion

We have developed a framework for collecting, searching, analyzing, and visualizing important components of biological systems. We were able to build this by converting existing databases into a common nanopublication structure that uses the provenance of the database records to determine the quality of any given piece of information through the methods used to provide it. We use the Semantic Automated Discovery and Integration framework to provide simple access to data, and can visualize results using an existing interaction graph tool. The resulting application makes it easy to search for biological entities and see how they interact. We have already found some hypotheses of proteins through which drugs influence disease conditions. We plan to expand the loaded set of data with protein/disease associations as well as gene expression profiles, and will be using ReDrugS to produce prospective testable hypotheses.

Acknowledgments

A special thanks to Pascale Gaudet, who, with Michel Dumontier, evaluated the experimental methods and evidence codes listed in the Protein/Protein Interaction Ontology and Gene Ontology.

References

[1] J. P. McCusker, T. Lebo, M. Krauthammer, and D. L. McGuinness, "Next Generation Cancer Data Discovery, Access, and Integration Using Prizms and Nanopublications," in *Data Integration in the Life Sciences*. Springer, 2013, pp. 105–112.

[2] J. C. Obenauer and M. B. Yaffe, "Computational prediction of protein-protein interactions," in *Protein-Protein Interactions*. Springer, 2004, pp. 445–467.

[3] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein–protein interaction data?" *Journal of molecular biology*, vol. 327, no. 5, pp. 919–923, 2003.

[4] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical owl-dl reasoner," *Web Semantics: science, services and agents on the World Wide Web*, vol. 5, no. 2, pp. 51–53, 2007.

[5] R. Petryszak, T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, and et al., "Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments," *Nucleic Acids Research*, vol. 42, no. D1, p. D926–D932, Jan 2014. [Online]. Available: http://dx.doi.org/10.1093/nar/gkt1270

[6] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[7] M.-Y. Lee, R. A. Kumar, S. M. Sukumaran, M. G. Hogg, D. S. Clark, and J. S. Dordick, "Three-dimensional cellular microarray for high-throughput toxicology assays," *Proceedings of the National Academy of Sciences*, vol. 105, no. 1, pp. 59–63, 2008.

[2] http://lod.melagrid.org/redrugs

[3] http://cytoscape.github.io/cytoscape.js

[4] https://sadi.googlecode.com/svn/trunk/javascript/sadi.js

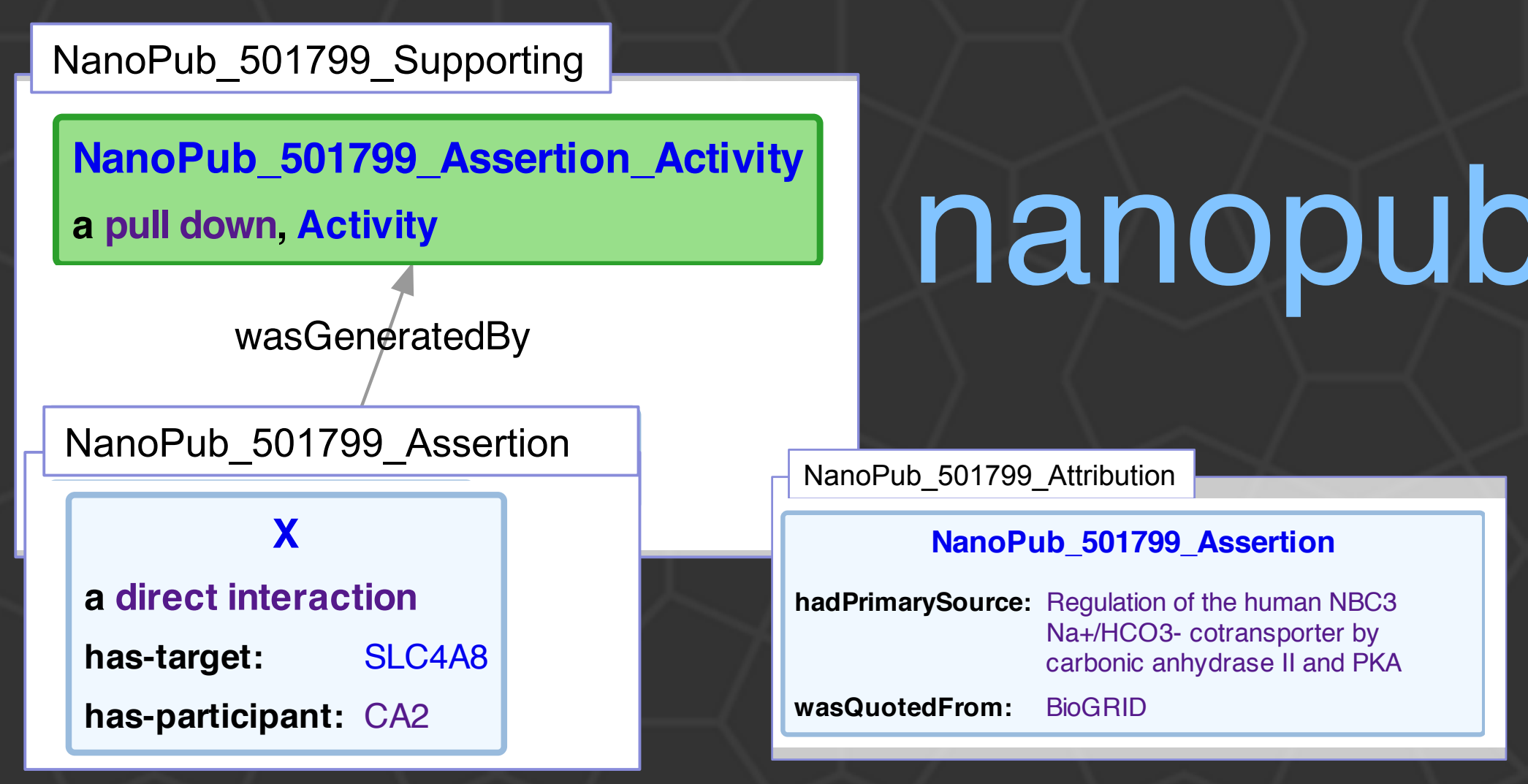
A Nanopublication Framework for Biological Networks using Cytoscape.js

James P. McCusker,[1,2] Rui Yan,[1] Kusum Solanki,[1] John Erickson,[1] Cynthia Chang,[1]
Michel Dumontier,[3] Jonathan Dordick,[1] and Deborah McGuinness[1]

[1]Rensselaer Polytechnic Institute, Troy, NY
[2]5AM Solutions, Inc., Rockville, MD
[3]Stanford University, Stanford, CA

Availability: http://redrugs.tw.rpi.edu

We leverage semantic technologies and Cytoscape.js to create a provenance-aware, probabilistic analysis platform for systems biology and evaluate its usefulness in discovering links between drugs and diseases. A number of databases have been developed that serve as a patchwork across the landscape of systems biology, each focused on different experimental methods, many species, and a wide diversity of inclusion criteria. Systems biology has been used in the past to generate hypotheses for drug effects, but has become fragmented under the large number of disparate and disconnected databases. In our efforts to create a systematic approach to discovering new uses for existing drugs, we have developed Repurposing Drugs with Semantics (ReDrugS). ReDrugS is a data curation and publication framework that can take data from nearly any database containing biological or chemical entity interactions and display it using Cytoscape.js. ReDrugS is able to infer probability of the assertions based on its provenance using experimental methods and data sources. A semantic web service API is provided that can search, traverse, and provide composite probabilities for the resulting graph of biological entities using the SADI web service framework and Nanopublications. We show how associations between a postive control, topiramate, allows us to independently reconstruct a positive control of epilepsy and migraine, and potential consequences on bone health. Future work will incorporate additional protein/disease associations, enabling hypothesis generation on indirect drug targets, and leading to testing the resulting hypotheses using high throughput drug screening.


NanoPub_501799_Supporting
NanoPub_501799_Assertion_Activity
a pull down, Activity
wasGeneratedBy
NanoPub_501799_Assertion
X
a direct interaction
has-target: SLC4A8
has-participant: CA2
NanoPub_501799_Attribution
NanoPub_501799_Assertion
hadPrimarySource: Regulation of the human NBC3 Na+/HCO3- cotransporter by carbonic anhydrase II and PKA
wasQuotedFrom: BioGRID

nanopub

+

Class: GeneratedBy_MI_0096
EquivalentClass:
wasGeneratedBy some 'pull down'
SubClassOf: Confidence2

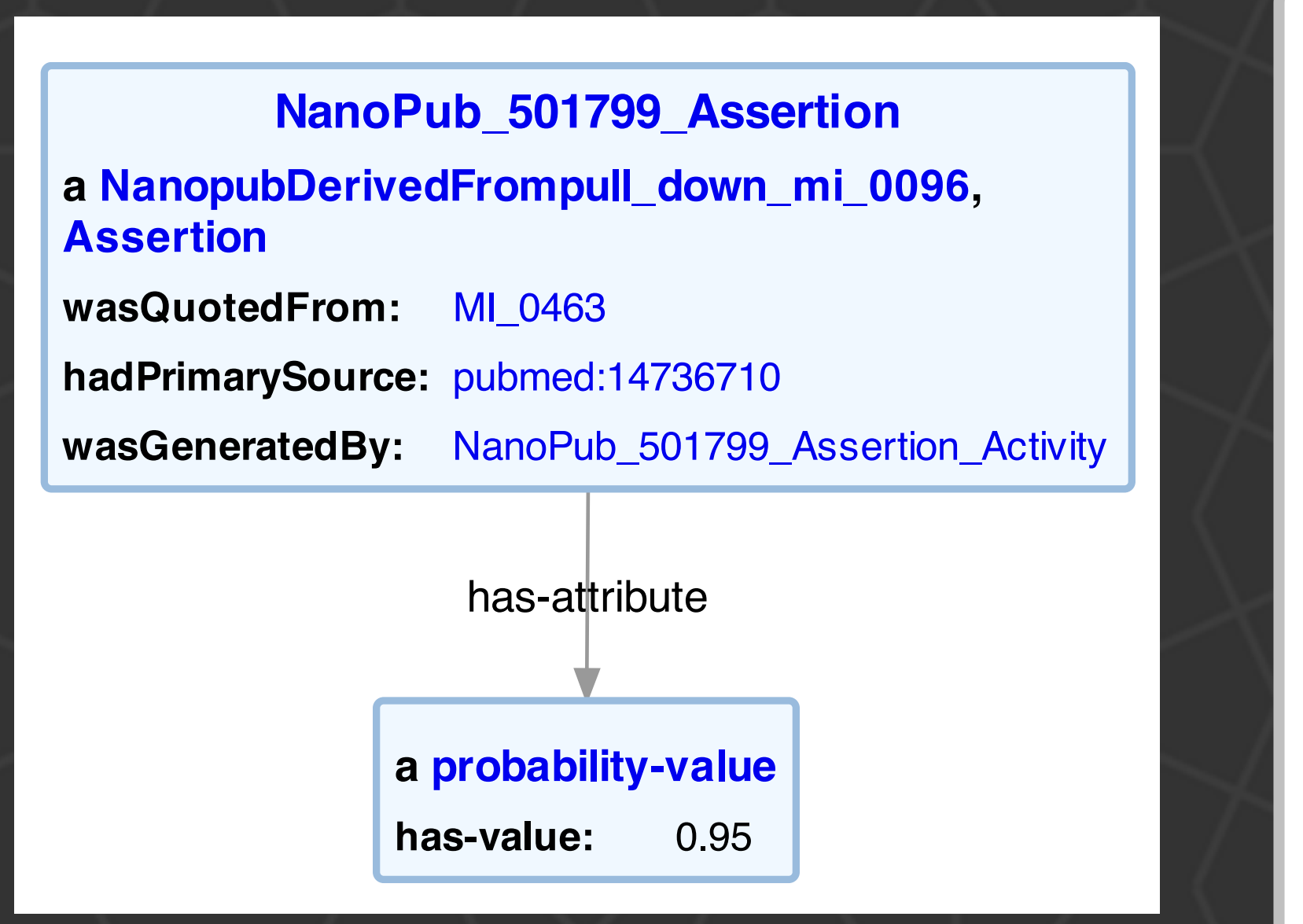
ontology

Class: Confidence2
EquivalentClass: 'has attribute' min 1
('probability value' and
('has value' value 0.95))

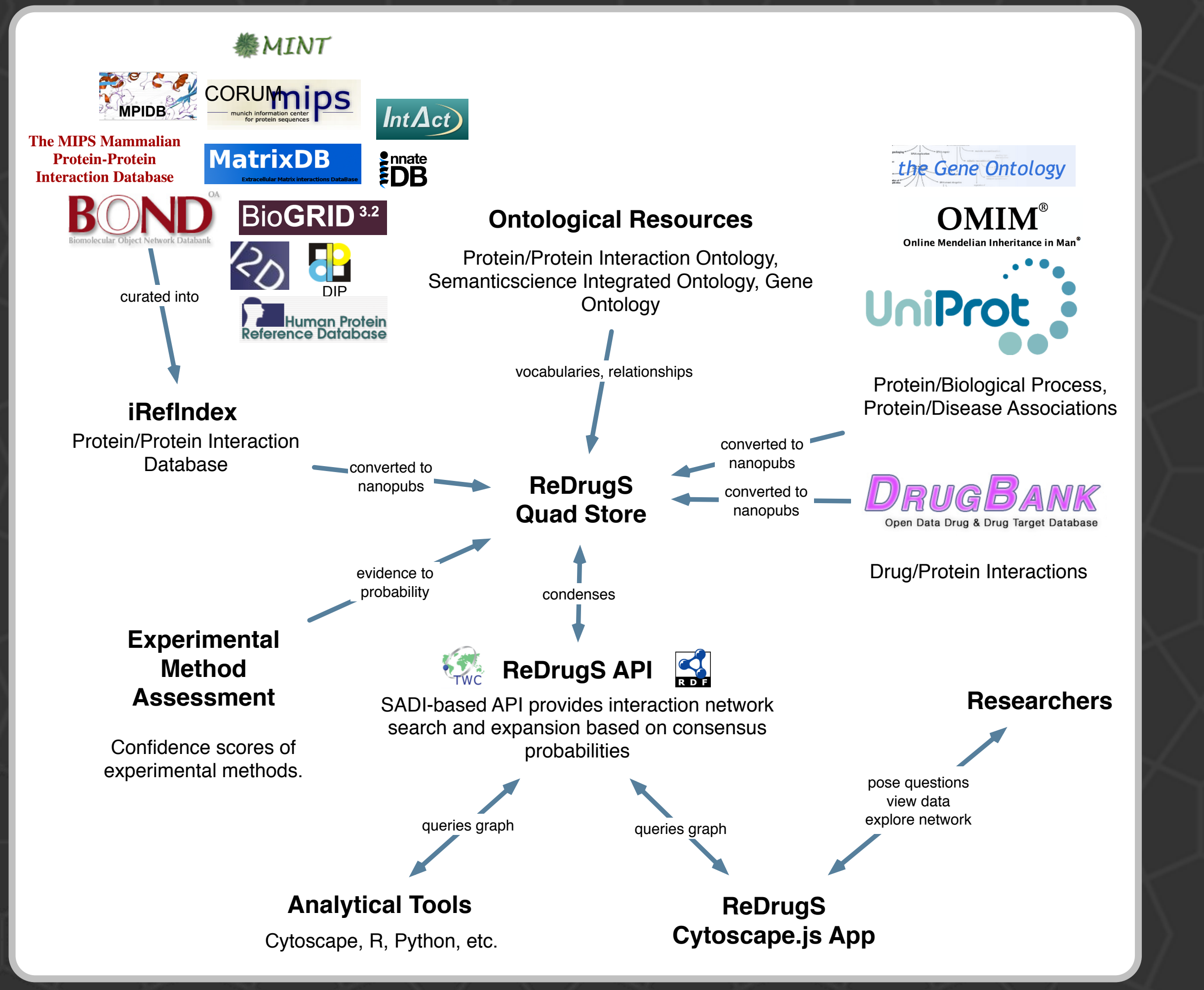
+

reasoner

⇒


NanoPub_501799_Assertion
a NanopubDerivedFrompull_down_mi_0096, Assertion
wasQuotedFrom: MI_0463
hadPrimarySource: pubmed:14736710
wasGeneratedBy: NanoPub_501799_Assertion_Activity
has-attribute
a probability-value
has-value: 0.95

probability

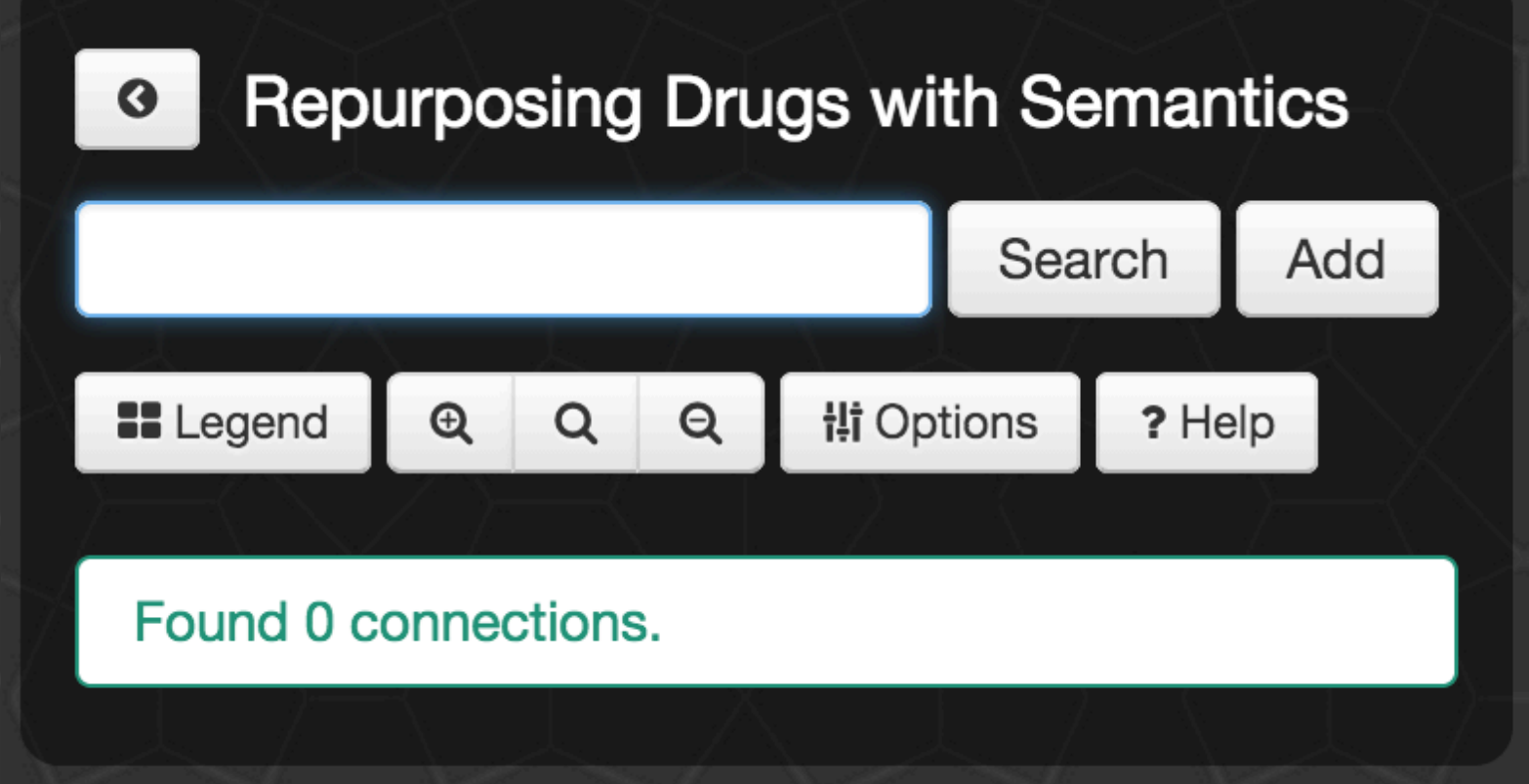

MINT
MPIDB
CORUM
mips
IntAct
The MIPS Mammalian Protein-Protein Interaction Database
MatrixDB
InnateDB
BOND
BioGRID 3.2
DIP
Human Protein Reference Database
curated into
iRefIndex
Protein/Protein Interaction Database
Ontological Resources
Protein/Protein Interaction Ontology, Semanticscience Integrated Ontology, Gene Ontology
vocabularies, relationships
the Gene Ontology
OMIM
Online Mendelian Inheritance in Man
UniProt
Protein/Biological Process, Protein/Disease Associations
converted to nanopubs
converted to nanopubs
converted to nanopubs
ReDrugS Quad Store
DrugBank
Open Data Drug & Drug Target Database
Drug/Protein Interactions
evidence to probability
condenses
Experimental Method Assessment
Confidence scores of experimental methods.
ReDrugS API
SADI-based API provides interaction network search and expansion based on consensus probabilities
Researchers
pose questions
view data
explore network
queries graph
queries graph
Analytical Tools
Cytoscape, R, Python, etc.
ReDrugS Cytoscape.js App

Flow of data through the ReDrugS pipeline

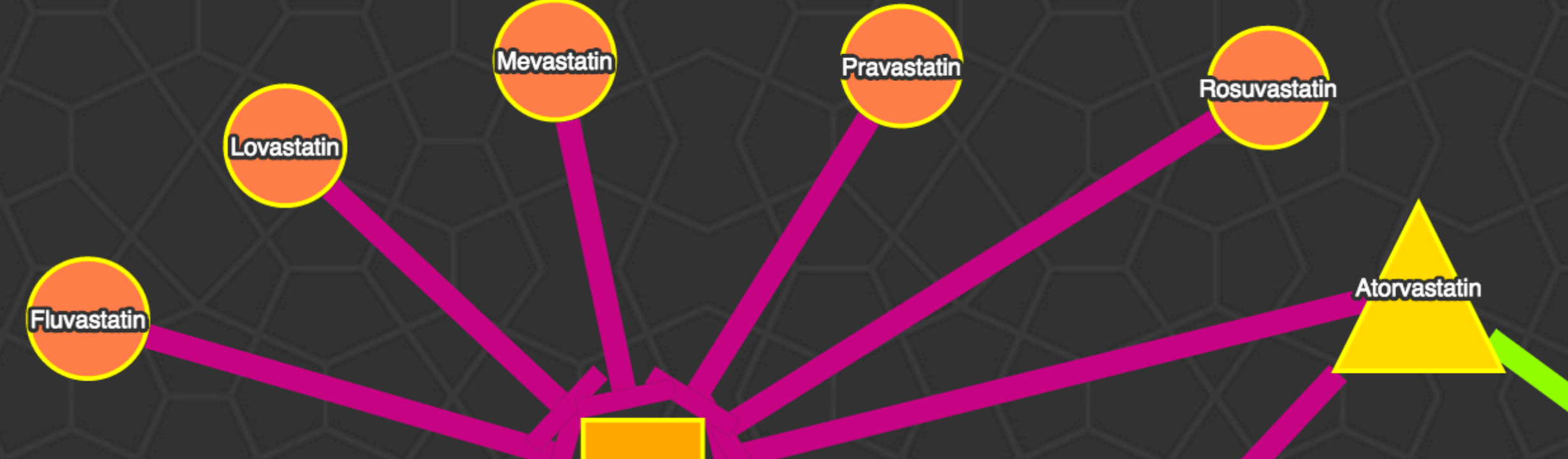
Different databases can provide the same assertions. This might be experimental replication! We model this with composite z-scores:

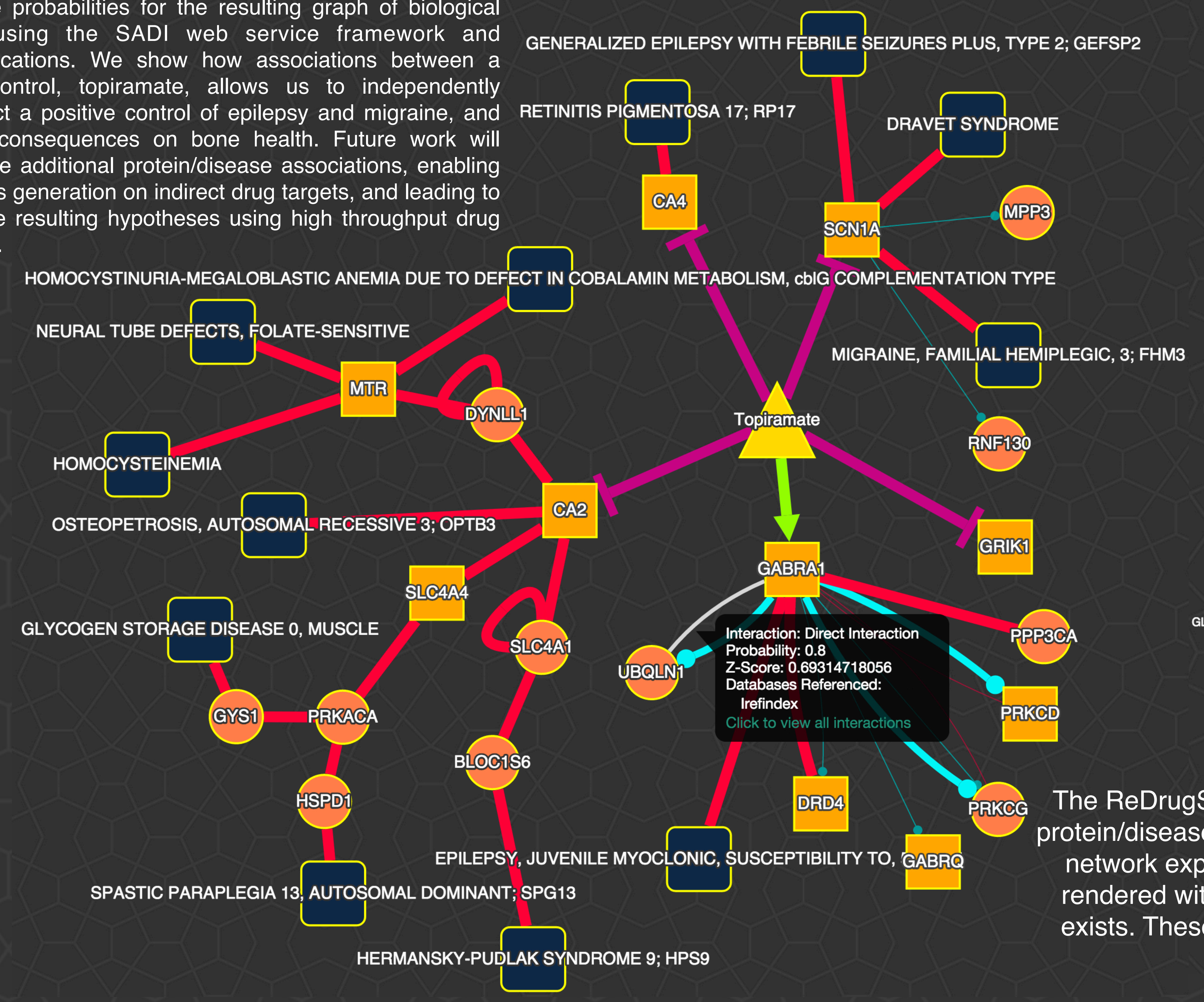
$$P(x) = F\left(\sum_{i=1}^{n} F^{-1}(p_i)\right)$$

F(x): Cumulative Distribution Function
(converts z-scores to probabilities)

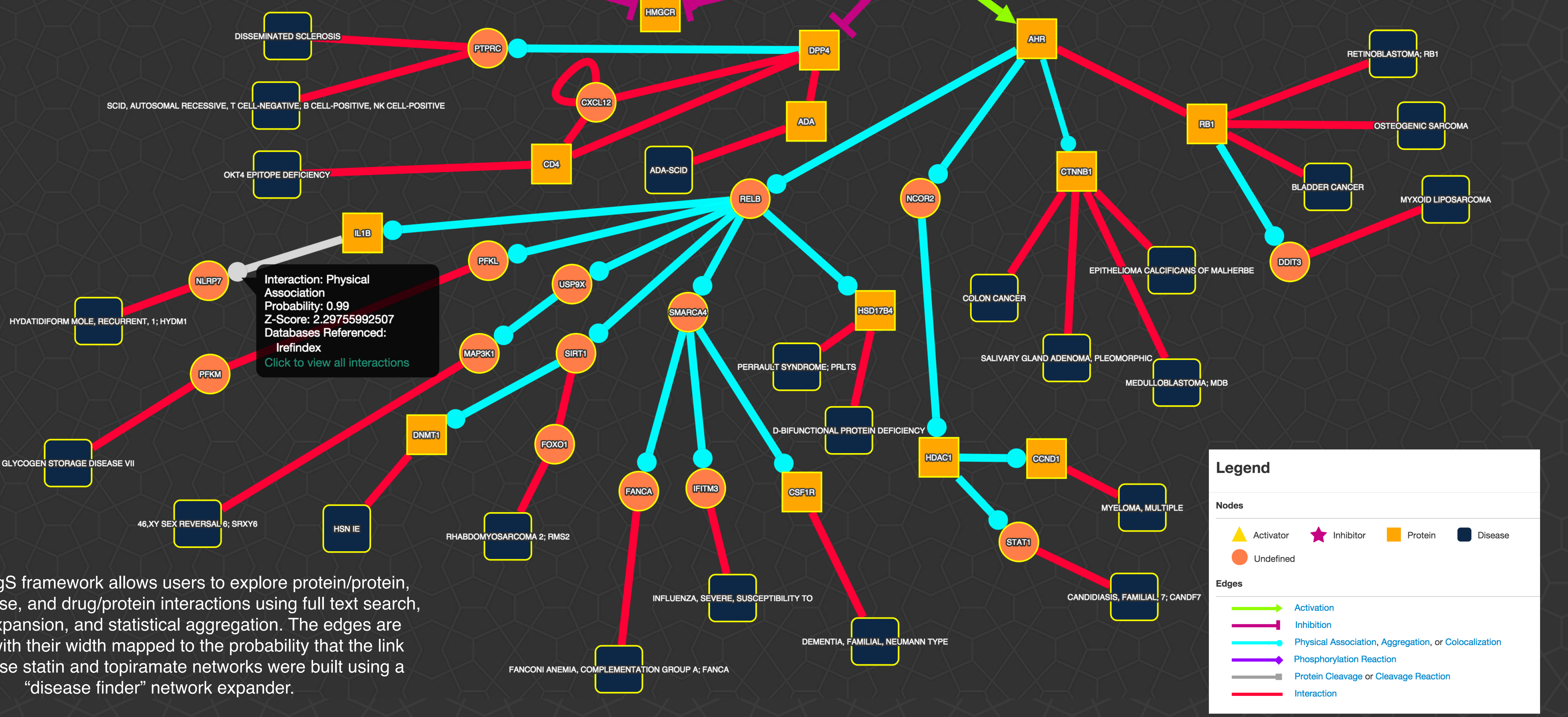


Found 0 connections.




GENERALIZED EPILEPSY WITH FEBRILE SEIZURES PLUS, TYPE 2; GEFSP2
RETINITIS PIGMENTOSA 17; RP17
DRAVET SYNDROME
CA4
SCN1A
MPP3
HOMOCYSTINURIA-MEGALOBLASTIC ANEMIA DUE TO DEFECT IN COBALAMIN METABOLISM, cblG COMPLEMENTATION TYPE
NEURAL TUBE DEFECTS, FOLATE-SENSITIVE
MIGRAINE, FAMILIAL HEMIPLEGIC, 3; FHM3
MTR
DYNLL1
Topiramate
RNF130
HOMOCYSTEINEMIA
CA2
OSTEOPETROSIS, AUTOSOMAL RECESSIVE 3; OPTB3
GRIK1
GABRA1
SLC4A4
GLYCOGEN STORAGE DISEASE 0, MUSCLE
SLC4A1
UBQLN1
PPP3CA
GYS1
PRKACA
PRKCD
BLOC1S6
DRD4
PRKCG
HSPD1
EPILEPSY, JUVENILE MYOCLONIC, SUSCEPTIBILITY TO,
GABRQ
SPASTIC PARAPLEGIA 13, AUTOSOMAL DOMINANT; SPG13
HERMANSKY-PUDLAK SYNDROME 9; HPS9

Interaction: Direct Interaction
Probability: 0.8
Z-Score: 0.69314718056
Databases Referenced:
Irefindex
Click to view all interactions


HMGCR
DISSEMINATED SCLEROSIS
PTPRC
DPP4
AHR
RETINOBLASTOMA; RB1
CXCL12
SCID, AUTOSOMAL RECESSIVE, T CELL-NEGATIVE, B CELL-POSITIVE, NK CELL-POSITIVE
ADA
RB1
OSTEOGENIC SARCOMA
OKT4 EPITOPE DEFICIENCY
CD4
ADA-SCID
CTNNB1
BLADDER CANCER
RELB
NCOR2
MYXOID LIPOSARCOMA
IL1B
PFKL
DDIT3
EPITHELIOMA CALCIFICANS OF MALHERBE
NLRP7
USP9X
COLON CANCER
HYDATIDIFORM MOLE, RECURRENT, 1; HYDM1
SMARCA4
HSD17B4
MAP3K1
SIRT1
SALIVARY GLAND ADENOMA, PLEOMORPHIC
PFKM
PERRAULT SYNDROME; PRLTS
MEDULLOBLASTOMA; MDB
DNMT1
FOXO1
D-BIFUNCTIONAL PROTEIN DEFICIENCY
HDAC1
CCND1
GLYCOGEN STORAGE DISEASE VII
FANCA
IFITM3
CSF1R
MYELOMA, MULTIPLE
STAT1
46,XY SEX REVERSAL 6; SRXY6
HSN IE
RHABDOMYOSARCOMA 2; RMS2
CANDIDIASIS, FAMILIAL, 7; CANDF7
INFLUENZA, SEVERE, SUSCEPTIBILITY TO
DEMENTIA, FAMILIAL, NEUMANN TYPE
FANCONI ANEMIA, COMPLEMENTATION GROUP A; FANCA

Interaction: Physical Association
Probability: 0.99
Z-Score: 2.29755992507
Databases Referenced:
Irefindex
Click to view all interactions

Legend

Nodes

▲ Activator ★ Inhibitor ■ Protein ■ Disease
● Undefined

Edges

→ Activation
⊣ Inhibition
● Physical Association, Aggregation, or Colocalization
◆ Phosphorylation Reaction
■ Protein Cleavage or Cleavage Reaction
— Interaction

The ReDrugS framework allows users to explore protein/protein, protein/disease, and drug/protein interactions using full text search, network expansion, and statistical aggregation. The edges are rendered with their width mapped to the probability that the link exists. These statin and topiramate networks were built using a "disease finder" network expander.