Knowledge discovery and machine learning in computational toxicology: Applications using ComptoxAI

Joseph D. Romano, PhD

USC R-TEN Executive Committee meeting

January 27, 2022

@jdromano2

Penn
University of Pennsylvania

Center of Excellence in Environmental Toxicology

Penn Institute for Biomedical Informatics

Toxicology

- Study of the adverse effects of chemicals on living organisms
  - Environmental toxicology focuses on environmental exposures
  - Occupational toxicology focuses on workplace exposures
  - Can also focus on toxic effects of pharmaceutical compounds
- Predictive toxicology: Use of *computational* and *statistical* techniques to predict (previously unobserved) toxic effects of specific chemicals

The role of AI in toxicology

- AI should help basic toxicologists by:
  - **Predicting** new associations between chemicals and endpoints of toxicity
  - **Explaining** mechanisms that may underlie those predictions
- It does not replace experimental validation; rather, it helps us to focus our time and effort

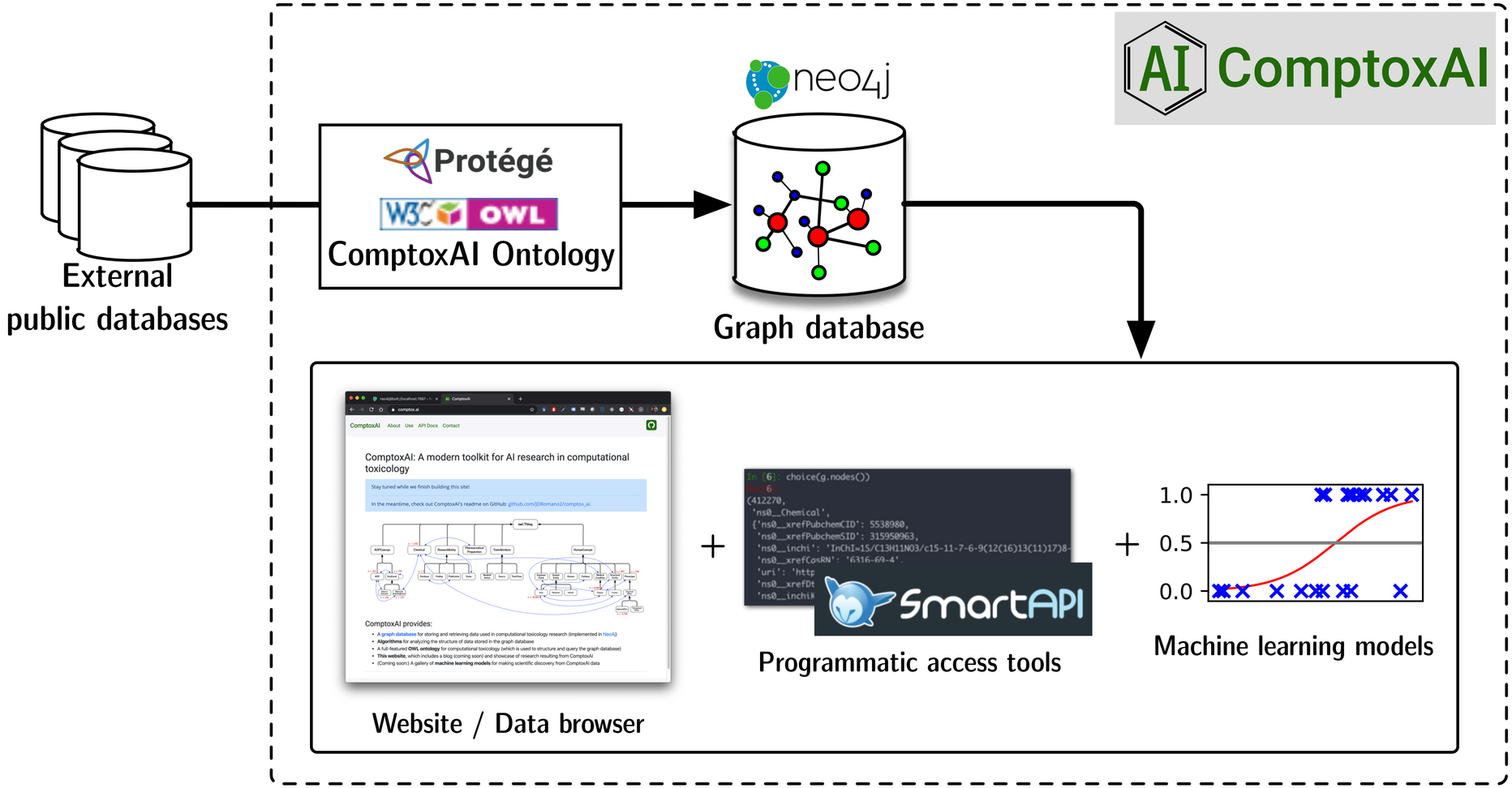
Outline

- ComptoxAI overview
- ComptoxAI: Data access and information retrieval
- ComptoxAI: GraphML to improve QSAR models

Outline

- ComptoxAI overview
- ComptoxAI: Data access and information retrieval
- ComptoxAI: GraphML to improve QSAR models

ComptoxAI



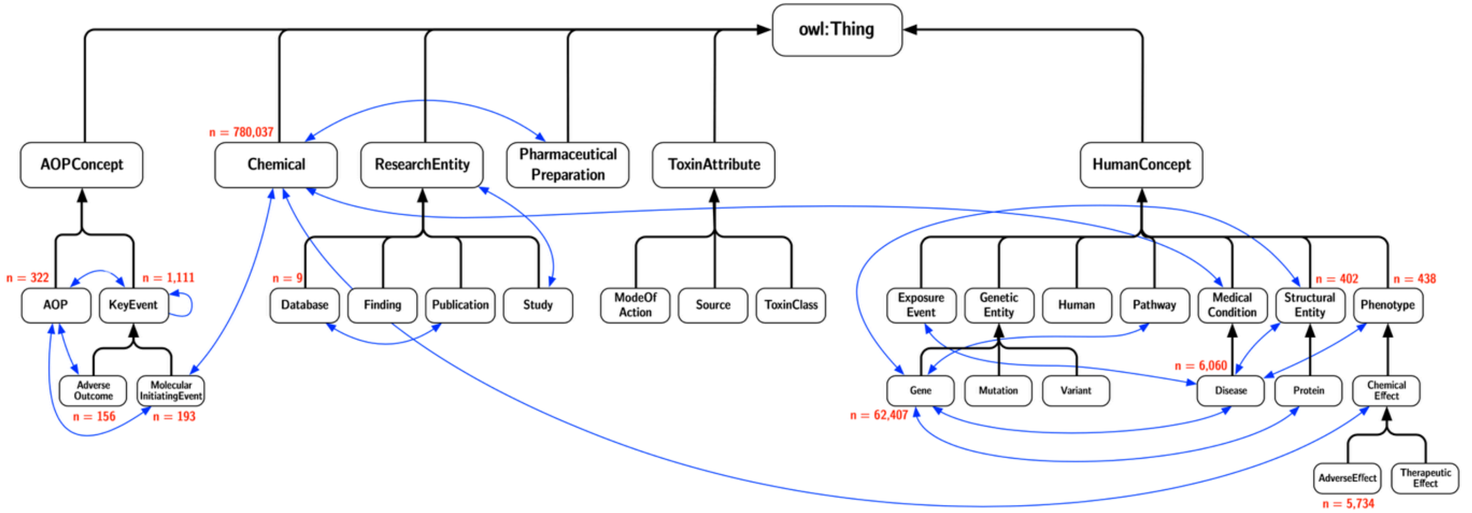
https://comptox.ai

Data vs. Knowledge

- **Data**:
  - Raw observations
  - Often (usually?) quantitative
  - E.g., specific gene expression measurements
- **Knowledge**:
  - Meaningful understanding of phenomena
  - Often results from analysis of many points of data
  - Typically leverages an idealized view of the world
  - E.g., "*Chemical [X] upregulates expression of gene [Y]*"

ComptoxAI: A modern toolkit for AI research in computational toxicology

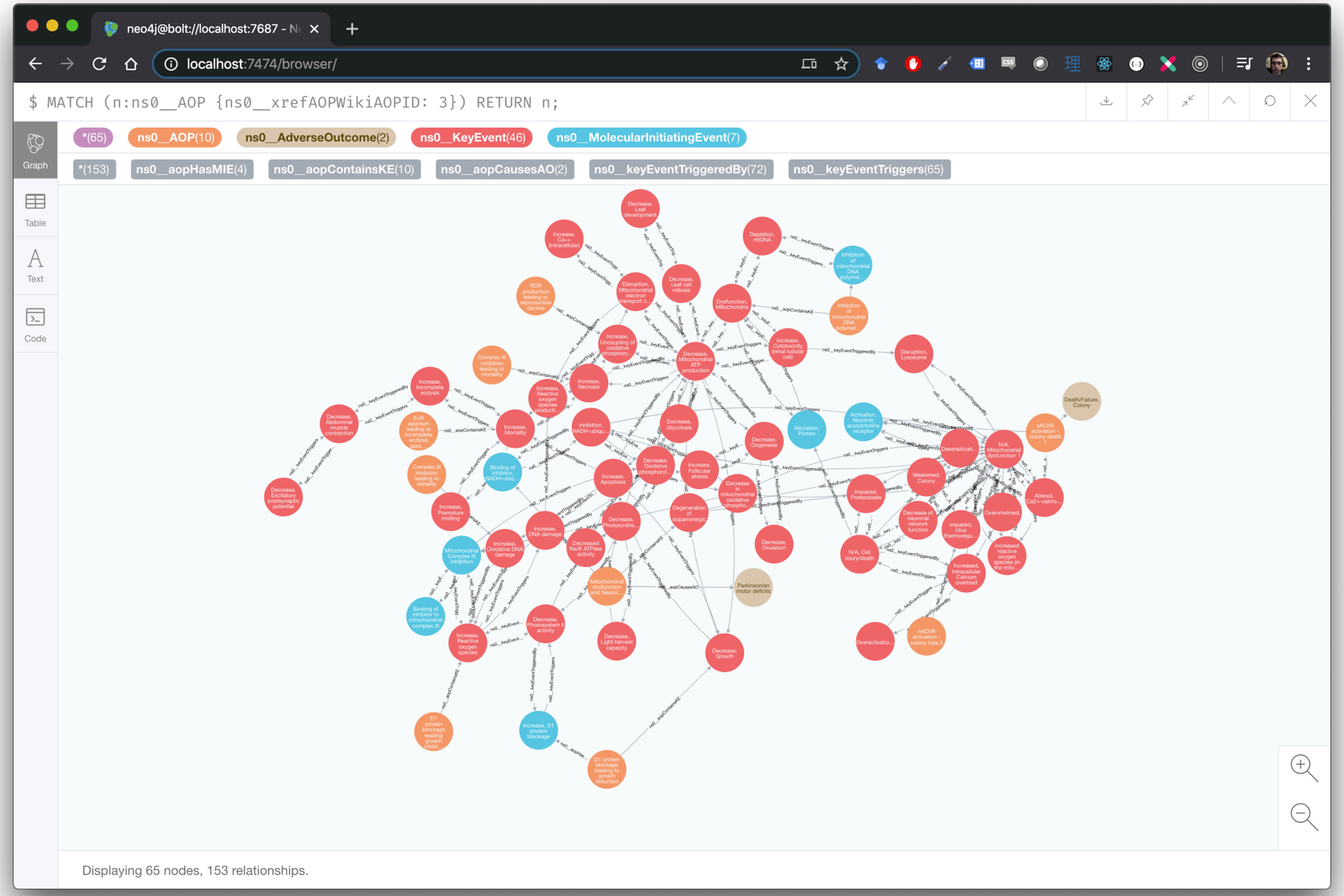
ComptoxAI



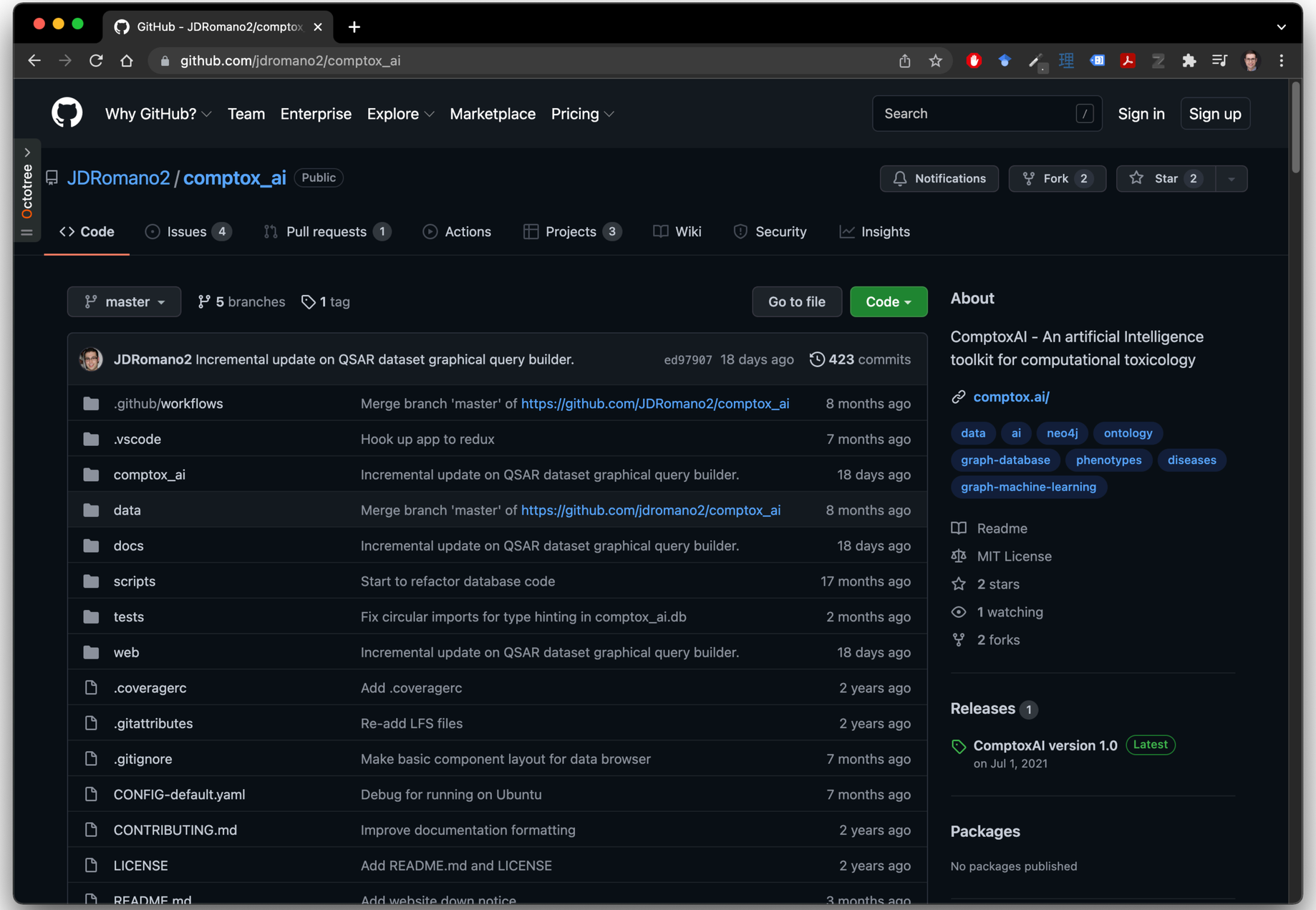
Browse ComptoxAI data

ComptoxAI provides:

- A **graph database** for storing and retrieving data used in computational toxicology research (implemented in Neo4j)
- **Algorithms** for analyzing the structure of data stored in the graph database
- A full-featured **OWL ontology** for computational toxicology (which is used to structure and query the graph database)
- **This website**, which includes a blog and showcase of research resulting from ComptoxAI
- (Coming soon:) A gallery of **machine learning models** for making scientific discoveries from ComptoxAI data


neo4j@bolt://localhost:7687 - Ne
localhost:7474/browser/
$ MATCH (n:ns0__AOP {ns0__xrefAOPWikiAOPID: 3}) RETURN n;
Graph
Table
Text
Code
*(65)
ns0__AOP(10)
ns0__AdverseOutcome(2)
ns0__KeyEvent(46)
ns0__MolecularInitiatingEvent(7)
*(153)
ns0__aopHasMIE(4)
ns0__aopContainsKE(10)
ns0__aopCausesAO(2)
ns0__keyEventTriggeredBy(72)
ns0__keyEventTriggers(65)
Displaying 65 nodes, 153 relationships.

Entity type	n
Chemical	780,037
Gene	62,407
Pathway	4,570
Key Event	1,111
Chemical List	311
Adverse Outcome Pathway	280
Molecular Initiating Event	193
Adverse Outcome	156
Assay	68


GitHub - JDRomano2/comptox
github.com/jdromano2/comptox_ai
Why GitHub? Team Enterprise Explore Marketplace Pricing
Search
Sign in
Sign up
Octotree
JDRomano2 / comptox_ai Public
Notifications
Fork 2
Star 2
Code
Issues 4
Pull requests 1
Actions
Projects 3
Wiki
Security
Insights
master
5 branches
1 tag
Go to file
Code
JDRomano2 Incremental update on QSAR dataset graphical query builder.
ed97907 18 days ago
423 commits
.github/workflows Merge branch 'master' of https://github.com/JDRomano2/comptox_ai 8 months ago
.vscode Hook up app to redux 7 months ago
comptox_ai Incremental update on QSAR dataset graphical query builder. 18 days ago
data Merge branch 'master' of https://github.com/jdromano2/comptox_ai 8 months ago
docs Incremental update on QSAR dataset graphical query builder. 18 days ago
scripts Start to refactor database code 17 months ago
tests Fix circular imports for type hinting in comptox_ai.db 2 months ago
web Incremental update on QSAR dataset graphical query builder. 18 days ago
.coveragerc Add .coveragerc 2 years ago
.gitattributes Re-add LFS files 2 years ago
.gitignore Make basic component layout for data browser 7 months ago
CONFIG-default.yaml Debug for running on Ubuntu 7 months ago
CONTRIBUTING.md Improve documentation formatting 2 years ago
LICENSE Add README.md and LICENSE 2 years ago
README.md Add website down notice 3 months ago
About
ComptoxAI - An artificial Intelligence toolkit for computational toxicology
comptox.ai/
data ai neo4j ontology graph-database phenotypes diseases graph-machine-learning
Readme
MIT License
2 stars
1 watching
2 forks
Releases 1
ComptoxAI version 1.0 Latest
on Jul 1, 2021
Packages
No packages published

2 paradigms for discovery:

- **Directly inferring relationships from network structure**
  - Stitching together chains of relationships originally fragmented across multiple sources
  - Inspecting higher-order network architecture
- **Predicting new, unobserved relationships**
  - Machine learning (especially graph machine learning)

Outline

- ComptoxAI overview
- ComptoxAI: Data access and information retrieval
- ComptoxAI: GraphML to improve QSAR models

ComptoxAI: Data Interfaces

- Data browser / dataset generator tools on website
- Direct access to graph database (local or remote)
- Web API (Programmatic access to data)
- Python package (Access data and construct machine learning models from the Python programming language)

ComptoxAI interactive data portal

From this page, you can search for individual entities (nodes) in ComptoxAI's graph database. When you select a query result, adjacent nodes (related data elements) are loaded and displayed below.

For detailed usage instructions, please see this page.

Nodes

Search LOAD EXAMPLE QUERY

Node Type: Gene

Node Field: Gene Symbol

Value: CYP2E1

SEARCH CLEAR FORM

Search Results

CLEAR NODE SEARCH RESULTS

Node details: COPY JSON

G Gene cytochrome P450 family 2 subfamily E member 1

External Identifiers:

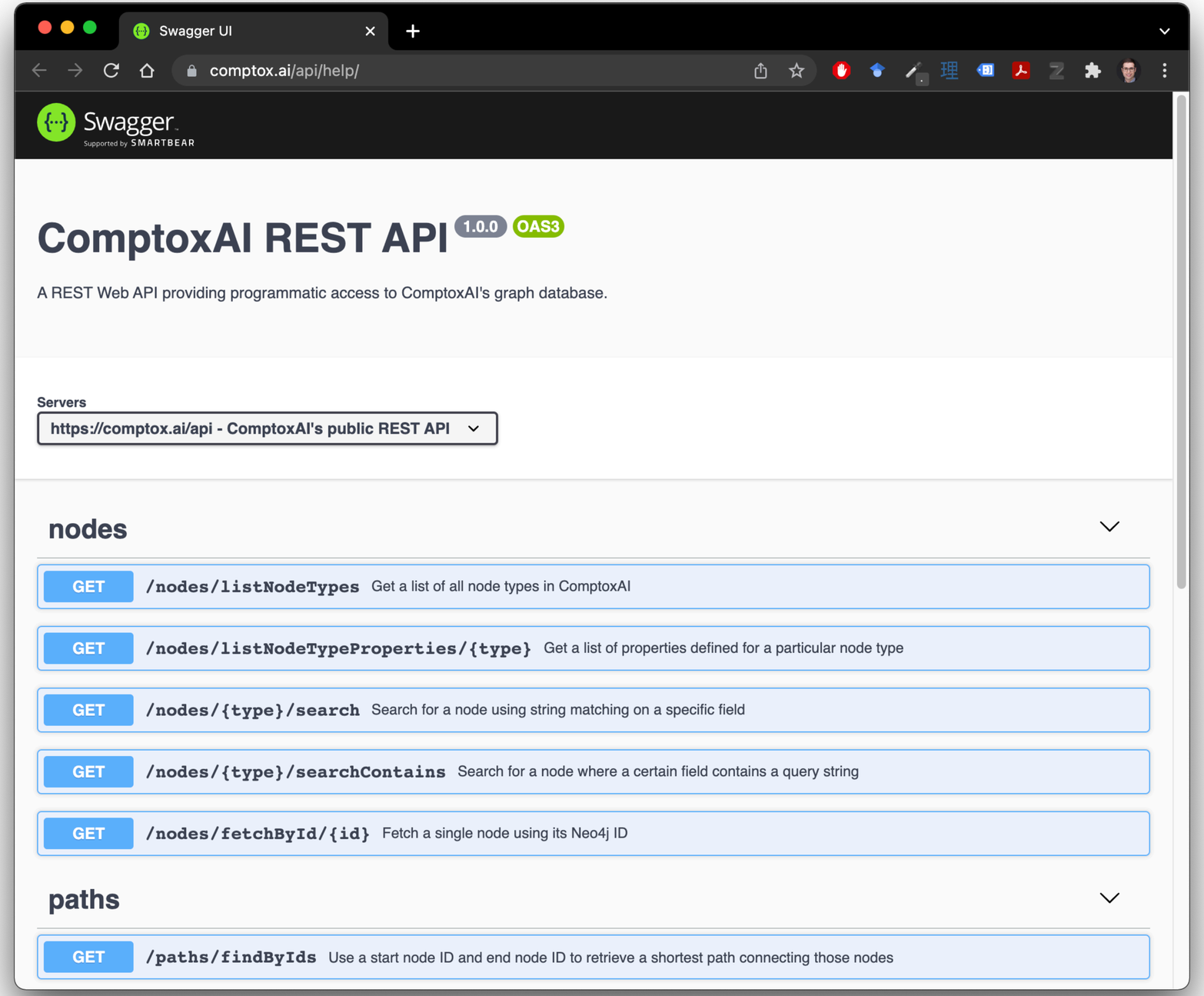
Database	Identifier
OMIM ID	124040
HGNC ID	2631
Ensembl ID	ENSG00000130649
NCBI Gene ID	1571

Other node features:

Feature name	Value
typeOfGene	protein-coding
geneSymbol	CYP2E1

Ontology IRI: `http://jdr.bio/ontologies/comptox.owl#gene_cyp2e1`

LOAD RELATIONSHIPS | PATH START NODE | PATH END NODE


Swagger UI
comptox.ai/api/help/
Swagger
Supported by SMARTBEAR
ComptoxAI REST API
1.0.0
OAS3
A REST Web API providing programmatic access to ComptoxAI's graph database.
Servers
https://comptox.ai/api - ComptoxAI's public REST API
nodes
GET /nodes/listNodeTypes Get a list of all node types in ComptoxAI
GET /nodes/listNodeTypeProperties/{type} Get a list of properties defined for a particular node type
GET /nodes/{type}/search Search for a node using string matching on a specific field
GET /nodes/{type}/searchContains Search for a node where a certain field contains a query string
GET /nodes/fetchById/{id} Fetch a single node using its Neo4j ID
paths
GET /paths/findByIds Use a start node ID and end node ID to retrieve a shortest path connecting those nodes

Curl

```
curl -X GET "https://comptox.ai/api/nodes/listNodeTypeProperties/Chemical" -H  "accept: */*"
```

Request URL

```
https://comptox.ai/api/nodes/listNodeTypeProperties/Chemical
```

Server response

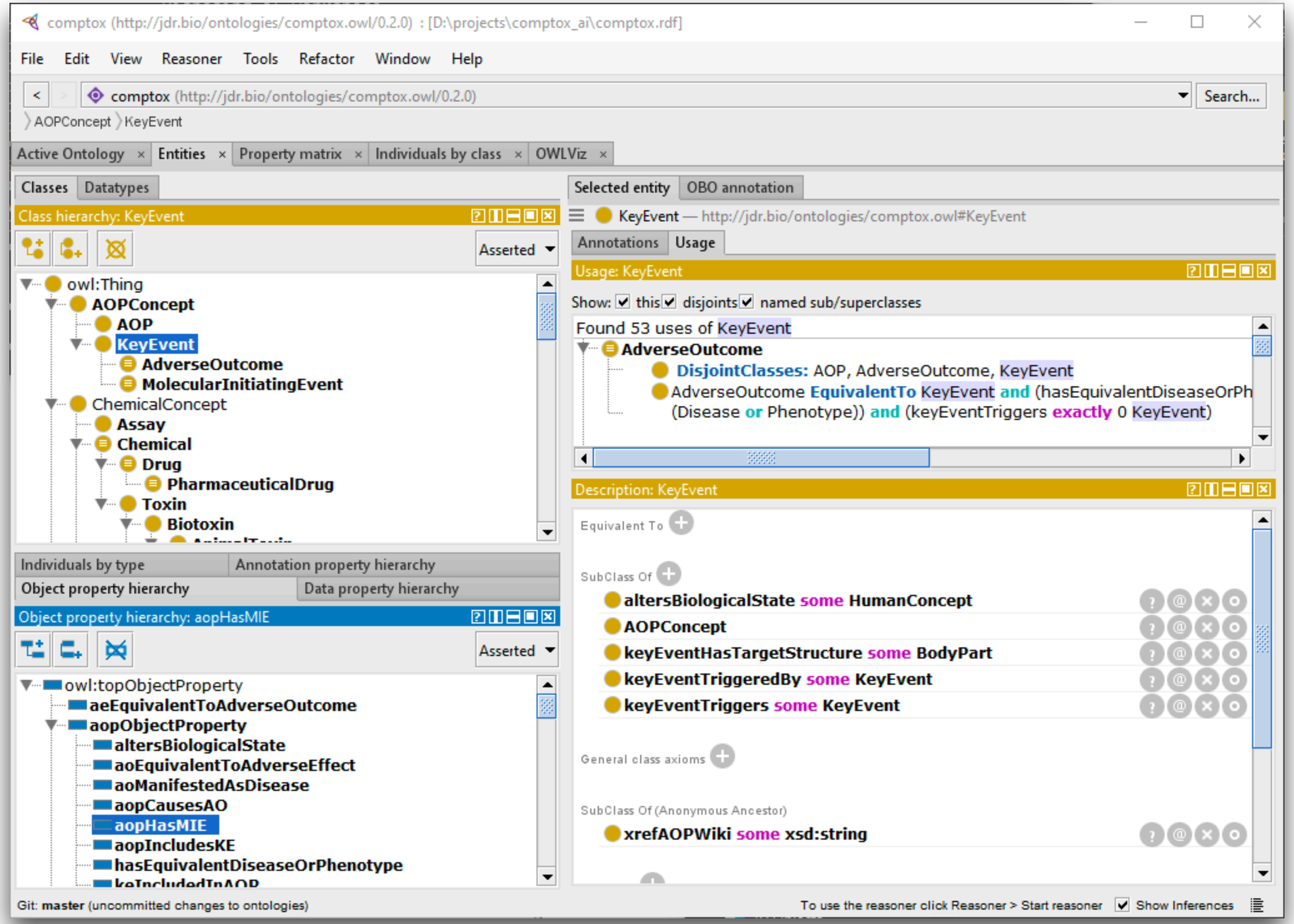
Code	Details
200	Response body


```
[
  {
    "property": "commonName",
    "type": "STRING"
  },
  {
    "property": "maccs",
    "type": "LIST"
  },
  {
    "property": "xrefMeSH",
    "type": "STRING"
  },
  {
    "property": "xrefDrugbank",
    "type": "STRING"
  },
  {
    "property": "xrefPubchemSID",
    "type": "STRING"
  },
  {
    "property": "xrefDTXSID",
    "type": "STRING"
  }
```


Download

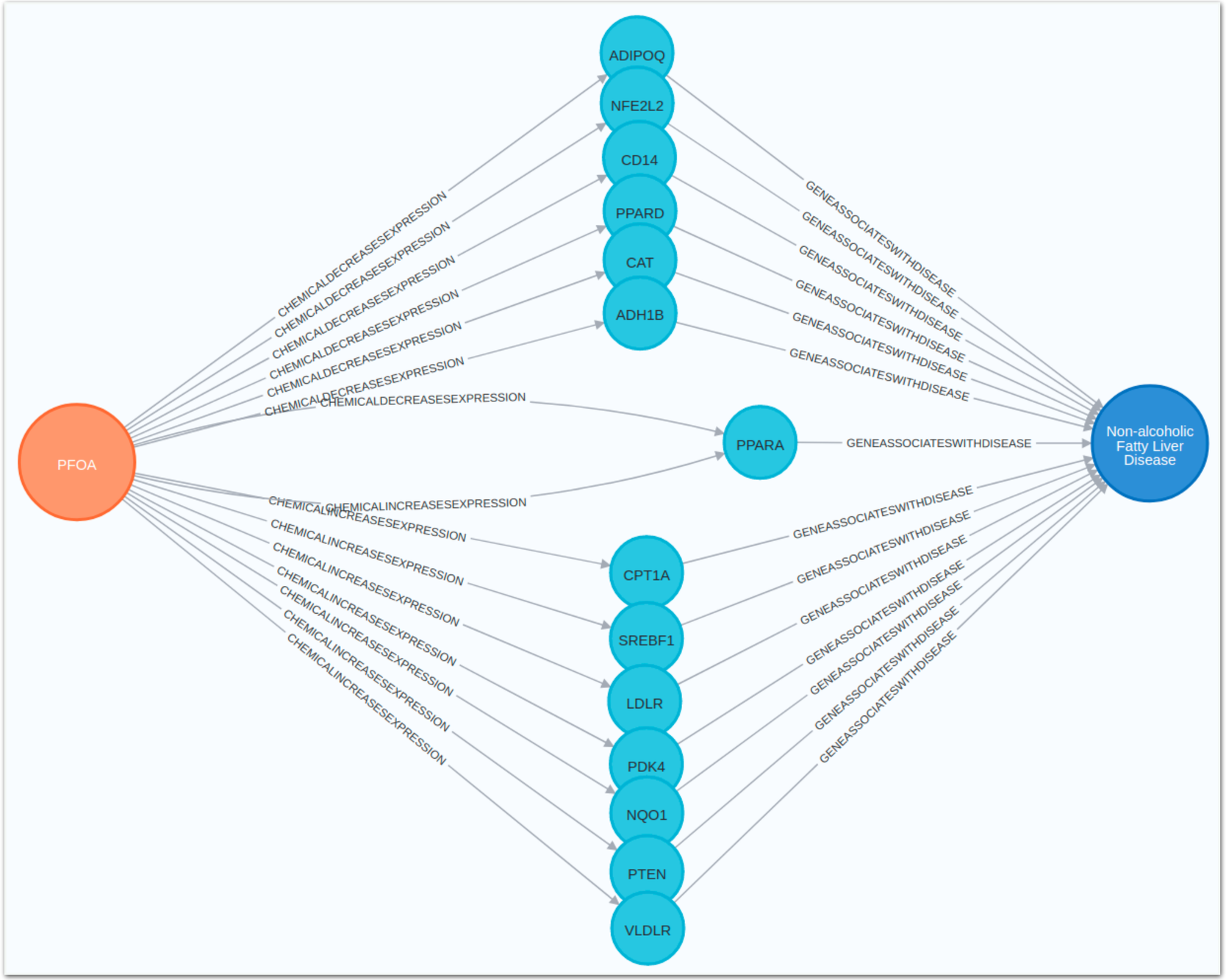
Response headers

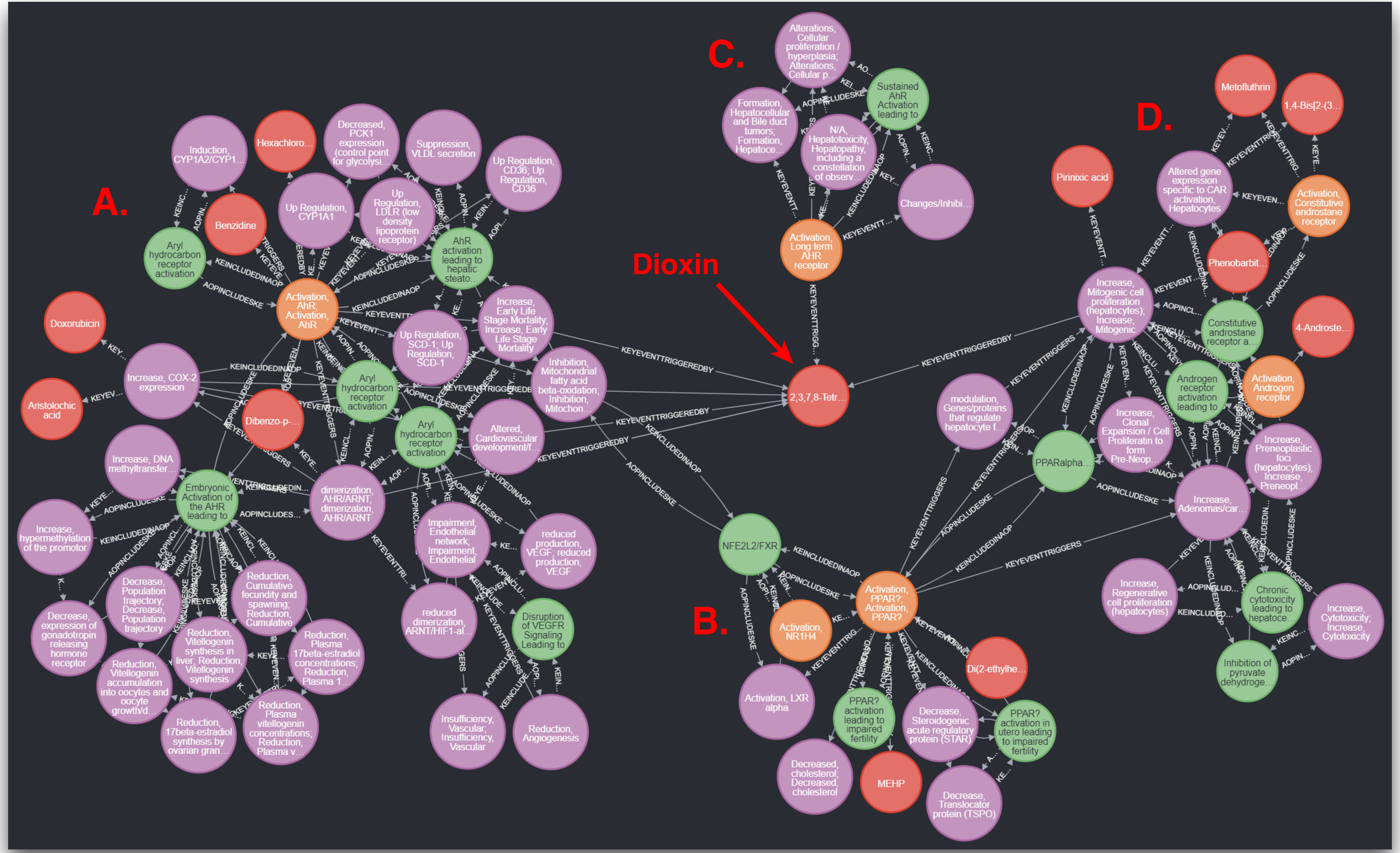
```
access-control-allow-credentials: true
access-control-allow-headers: Origin,X-Requested-With,Content-Type,Accept,Authorization
access-control-allow-methods: GET,HEAD,OPTIONS,POST,PUT,DELETE
access-control-allow-origin: *
connection: keep-alive
content-length: 372
content-type: application/json; charset=utf-8
```


comptox (http://jdr.bio/ontologies/comptox.owl/0.2.0) : [D:\projects\comptox_ai\comptox.rdf]
File Edit View Reasoner Tools Refactor Window Help
comptox (http://jdr.bio/ontologies/comptox.owl/0.2.0)
Search...
AOPConcept KeyEvent
Active Ontology Entities Property matrix Individuals by class OWLViz
Classes Datatypes
Class hierarchy: KeyEvent
Asserted
owl:Thing
AOPConcept
AOP
KeyEvent
AdverseOutcome
MolecularInitiatingEvent
ChemicalConcept
Assay
Chemical
Drug
PharmaceuticalDrug
Toxin
Biotoxin
Individuals by type
Annotation property hierarchy
Object property hierarchy
Data property hierarchy
Object property hierarchy: aopHasMIE
Asserted
owl:topObjectProperty
aeEquivalentToAdverseOutcome
aopObjectProperty
altersBiologicalState
aoEquivalentToAdverseEffect
aoManifestedAsDisease
aopCausesAO
aopHasMIE
aopIncludesKE
hasEquivalentDiseaseOrPhenotype
Selected entity
OBO annotation
KeyEvent — http://jdr.bio/ontologies/comptox.owl#KeyEvent
Annotations
Usage
Usage: KeyEvent
Show: this disjoints named sub/superclasses
Found 53 uses of KeyEvent
AdverseOutcome
DisjointClasses: AOP, AdverseOutcome, KeyEvent
AdverseOutcome EquivalentTo KeyEvent and (hasEquivalentDiseaseOrPh
(Disease or Phenotype)) and (keyEventTriggers exactly 0 KeyEvent)
Description: KeyEvent
Equivalent To
SubClass Of
altersBiologicalState some HumanConcept
AOPConcept
keyEventHasTargetStructure some BodyPart
keyEventTriggeredBy some KeyEvent
keyEventTriggers some KeyEvent
General class axioms
SubClass Of (Anonymous Ancestor)
xrefAOPWiki some xsd:string
Git: master (uncommitted changes to ontologies)
To use the reasoner click Reasoner > Start reasoner
Show Inferences

IR Tools

- "**Shortest Path**" - Identifies the most direct mechanistic routes linking two (or more) entities
- "**Expand Network**" - Shows an entity in the context of a network of nearby 'neighbor' entities
- "**QSAR Dataset Generator**" - Dynamically builds tabular datasets for predicting a toxic endpoint using fingerprints for a list of chemicals


PFOA
ADIPOQ
NFE2L2
CD14
PPARD
CAT
ADH1B
PPARA
CPT1A
SREBF1
LDLR
PDK4
NQO1
PTEN
VLDLR
Non-alcoholic Fatty Liver Disease
CHEMICALDECREASESEXPRESSION
CHEMICALINCREASESEXPRESSION
GENEASSOCIATESWITHDISEASE


A.
B.
C.
D.
Dioxin
2,3,7,8-Tetr...
NFE2L2/FXR
Activation, PPAR?; Activation, PPAR?
Activation, NR1H4
Activation, LXR alpha
Activation, AhR; Activation, AhR
Aryl hydrocarbon receptor activation
Benzidine
Induction, CYP1A2/CYP1...
Hexachloro...
Up Regulation, CYP1A1
Decreased, PCK1 expression (control point for glycolysi...
Suppression, VLDL secretion
Up Regulation, LDLR (low density lipoprotein receptor)
Up Regulation, CD36; Up Regulation, CD36
AhR activation leading to hepatic steato...
Doxorubicin
Aristolochic acid
Increase, COX-2 expression
Dibenzo-p-...
Up Regulation, SCD-1; Up Regulation, SCD-1
Increase, Early Life Stage Mortality; Increase, Early Life Stage Mortality
Inhibition, Mitochondrial fatty acid beta-oxidation; Inhibition, Mitochon...
Aryl hydrocarbon receptor activation
Aryl hydrocarbon receptor activation
Altered, Cardiovascular development/f...
Increase, DNA methyltransfer...
Embryonic Activation of the AHR leading to
dimerization, AHR/ARNT; dimerization, AHR/ARNT
Increase, hypermethylation of the promotor
Decrease, expression of gonadotropin releasing hormone receptor
Decrease, Population trajectory; Decrease, Population trajectory
Reduction, Vitellogenin accumulation into oocytes and oocyte growth/d...
Reduction, Vitellogenin synthesis in liver; Reduction, Vitellogenin synthesis
Reduction, 17beta-estradiol synthesis by ovarian gran...
Reduction, Cumulative fecundity and spawning; Reduction, Cumulative
Reduction, Plasma 17beta-estradiol concentrations; Reduction, Plasma 1...
Reduction, Plasma vitellogenin concentrations; Reduction, Plasma v...
Impairment, Endothelial network; Impairment, Endothelial
reduced production, VEGF; reduced production, VEGF
reduced dimerization, ARNT/HIF1-al...
Disruption of VEGFR Signaling Leading to
Insufficiency, Vascular; Insufficiency, Vascular
Reduction, Angiogenesis
Alterations, Cellular proliferation / hyperplasia; Alterations, Cellular p...
Formation, Hepatocellular and Bile duct tumors, Formation, Hepatoce...
N/A, Hepatotoxicity, Hepatopathy, including a constellation of observ...
Sustained AhR Activation leading to
Changes/Inhibi...
Activation, Long term AHR receptor
Pirinixic acid
Metofluthrin
1,4-Bis[2-(3...
Altered gene expression specific to CAR activation, Hepatocytes
Activation, Constitutive androstane receptor
Phenobarbit...
Increase, Mitogenic cell proliferation (hepatocytes); Increase, Mitogenic
Constitutive androstane receptor a...
4-Androste...
modulation, Genes/proteins that regulate hepatocyte f...
Androgen receptor activation leading to
Activation, Androgen receptor
Increase, Clonal Expansion / Cell Proliferatin to form Pre-Neop...
PPARalpha...
Increase, Preneoplastic foci (hepatocytes); Increase, Preneopl...
Increase, Adenomas/car...
Increase, Regenerative cell proliferation (hepatocytes)
Chronic cytotoxicity leading to hepatoce...
Increase, Cytotoxicity; Increase, Cytotoxicity
Inhibition of pyruvate dehydroge...
Di(2-ethylhe...
PPAR? activation leading to impaired fertility
PPAR? activation in utero leading to impaired fertility
Decrease, Steroidogenic acute regulatory protein (STAR)
Decreased, cholesterol; Decreased, cholesterol
MEHP
Decrease, Translocator protein (TSPO)
KEYEVENTTRIGGEREDBY
KEINCLUDEDINAOP
AOPINCLUDESKE
KEYEVENTTRIGGERS

Important Caveats

- Advanced users might get more mileage (e.g., by constructing graph queries by hand)
  - New ‘entry-level’ features will be continuously in development!
- Information retrieval is limited to what we already know and what is already in the source databases

Outline

- ComptoxAI overview
- ComptoxAI: Data access and information retrieval
- ComptoxAI: GraphML to improve QSAR models

› Pac Symp Biocomput. 2022;27:187-198.

Improving QSAR Modeling for Predictive Toxicology using Publicly Aggregated Semantic Graph Data and Graph Neural Networks

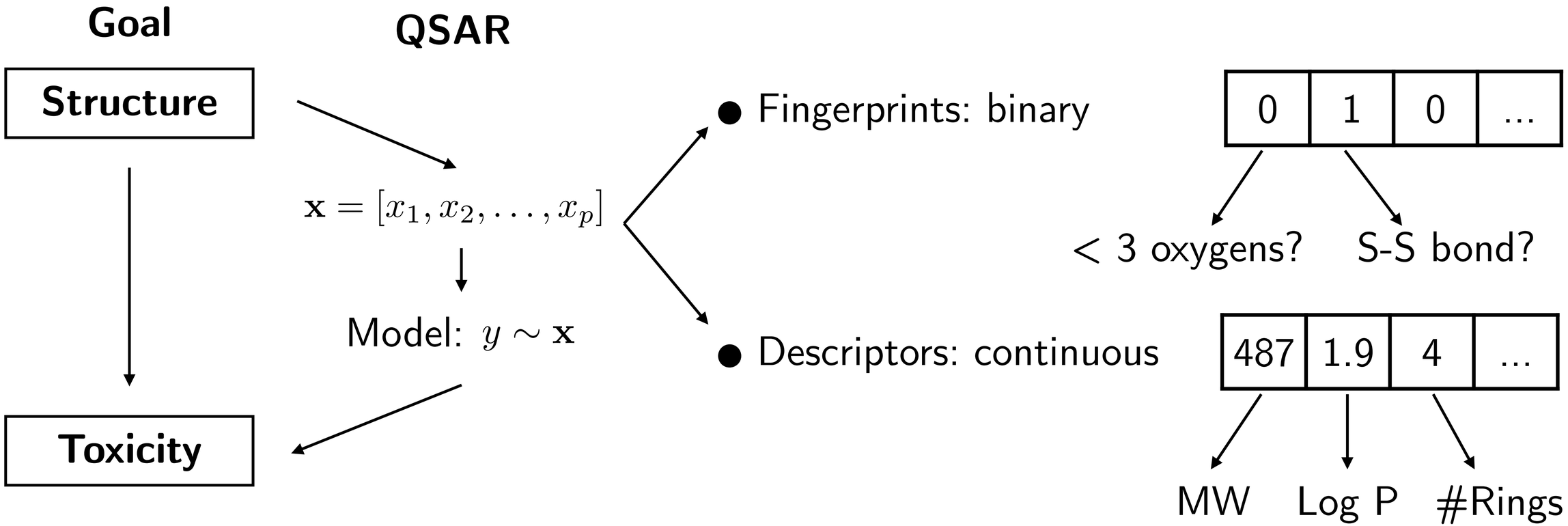
Joseph D Romano [1], Yun Hao, Jason H Moore

Affiliations + expand

PMID: 34890148 PMCID: PMC8714189

Free PMC article

QSAR

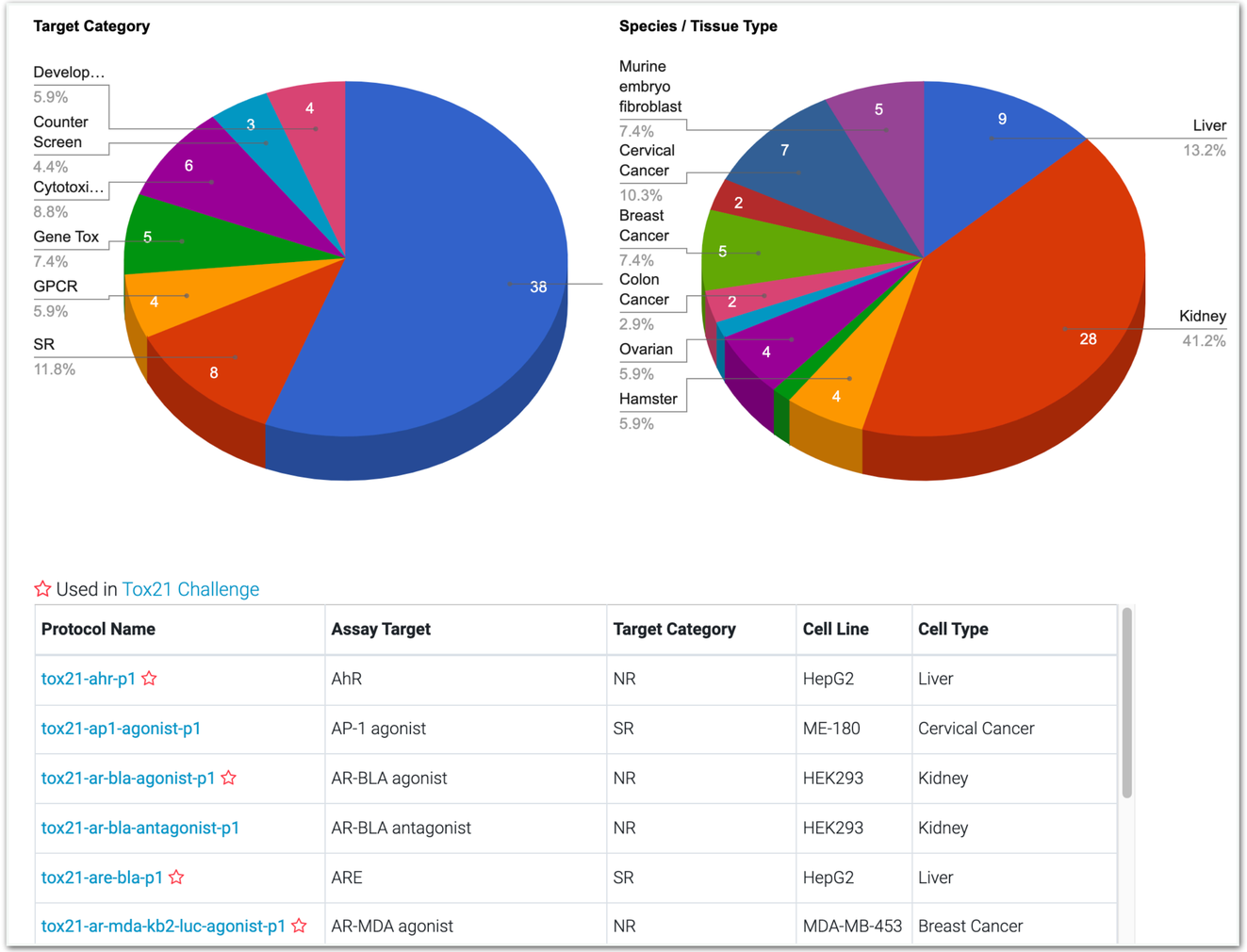


- QSAR: Quantitative Structure-Activity Relationship

Slide credit: Yun Hao

Tox21 screening dataset

- Tox21: "Toxicology in the 21st Century" dataset for high-throughput chemical screening
  - ~60 specific toxicology-focused biochemical assays
  - ~8,000 chemicals evaluated on those assays

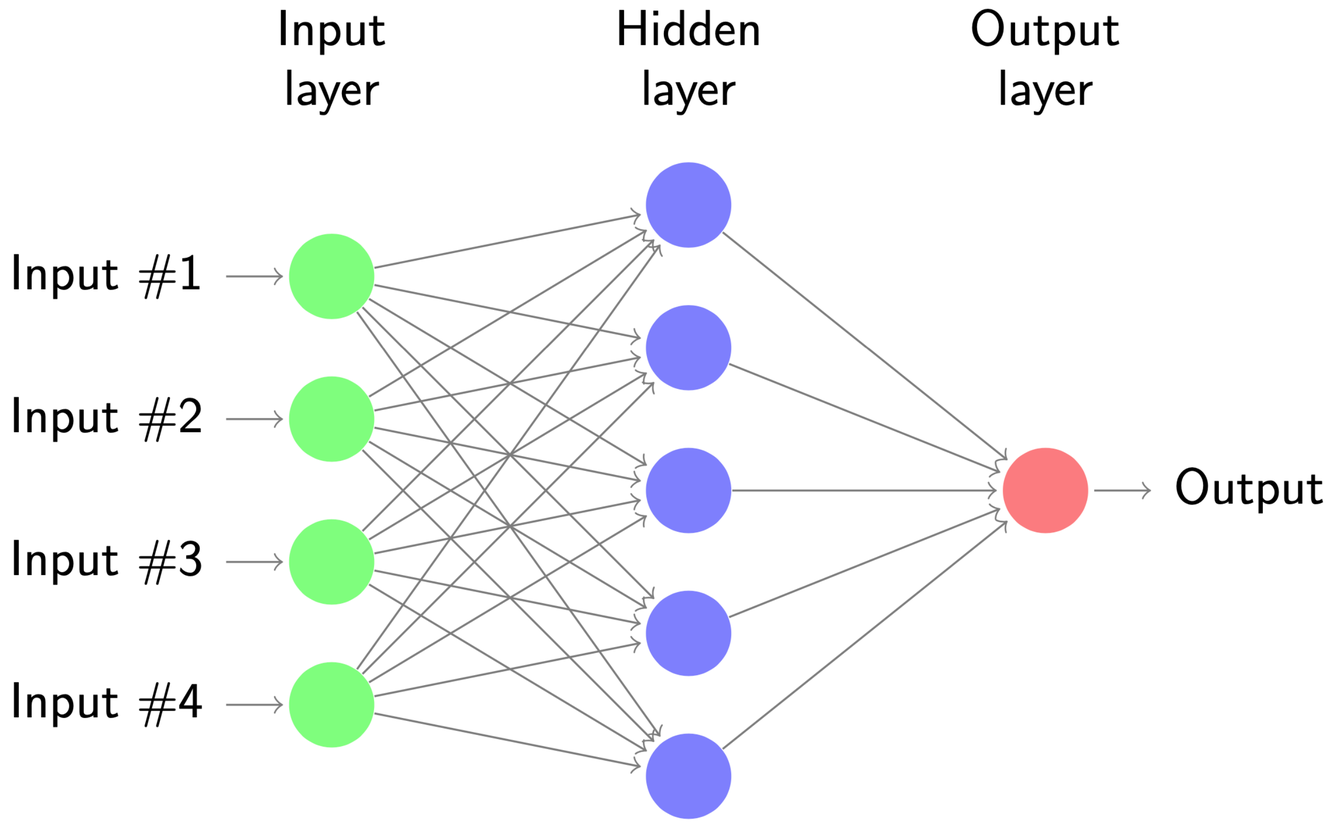


☆ Used in Tox21 Challenge

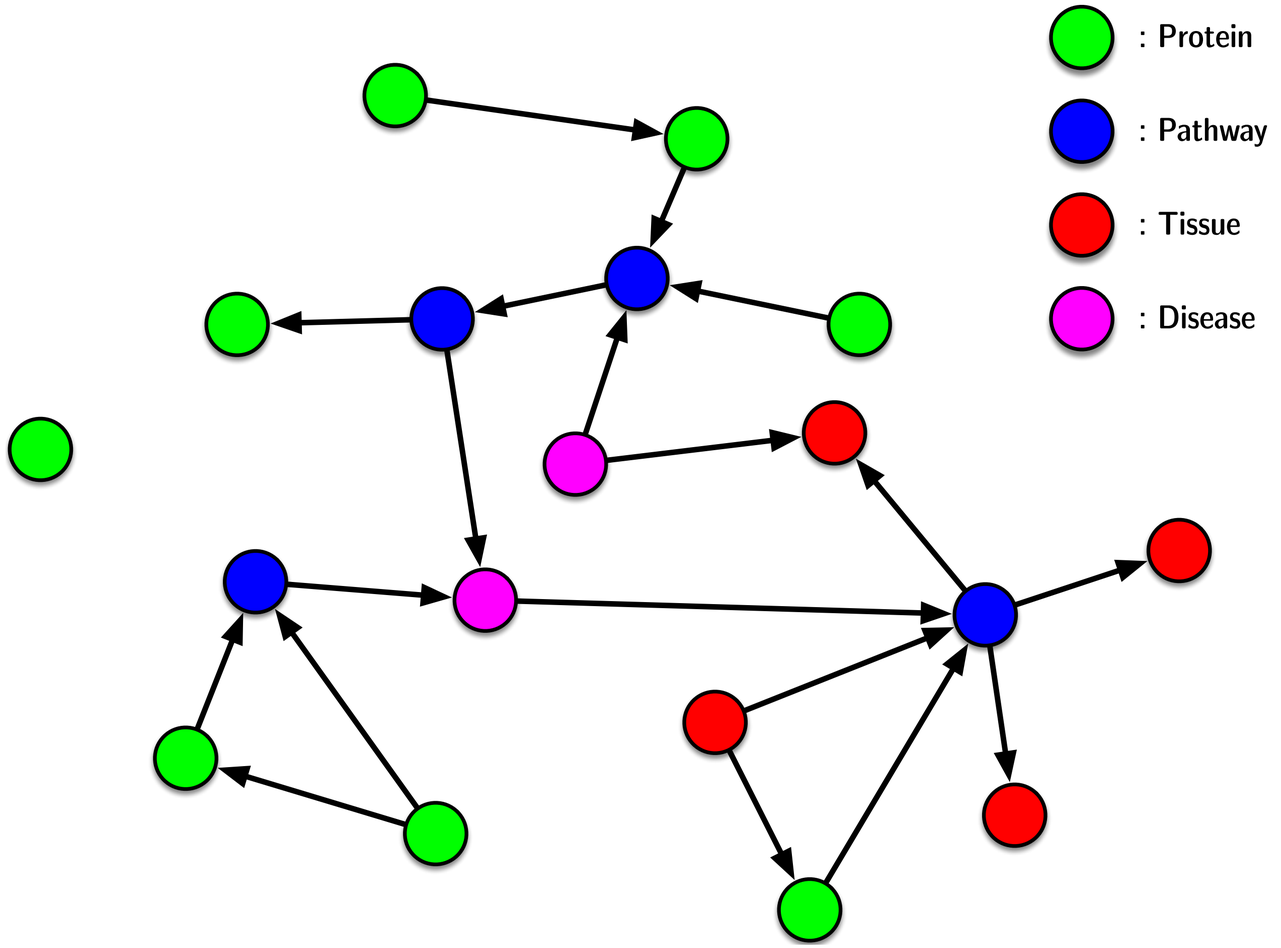
Protocol Name	Assay Target	Target Category	Cell Line	Cell Type
tox21-ahr-p1 ☆	AhR	NR	HepG2	Liver
tox21-ap1-agonist-p1	AP-1 agonist	SR	ME-180	Cervical Cancer
tox21-ar-bla-agonist-p1 ☆	AR-BLA agonist	NR	HEK293	Kidney
tox21-ar-bla-antagonist-p1	AR-BLA antagonist	NR	HEK293	Kidney
tox21-are-bla-p1 ☆	ARE	SR	HepG2	Liver
tox21-ar-mda-kb2-luc-agonist-p1 ☆	AR-MDA agonist	NR	MDA-MB-453	Breast Cancer

(Artificial) Neural Networks

- Consist of nodes organized into layers, which are usually stacked
- Deep learning —> NN with tens or hundreds of layers

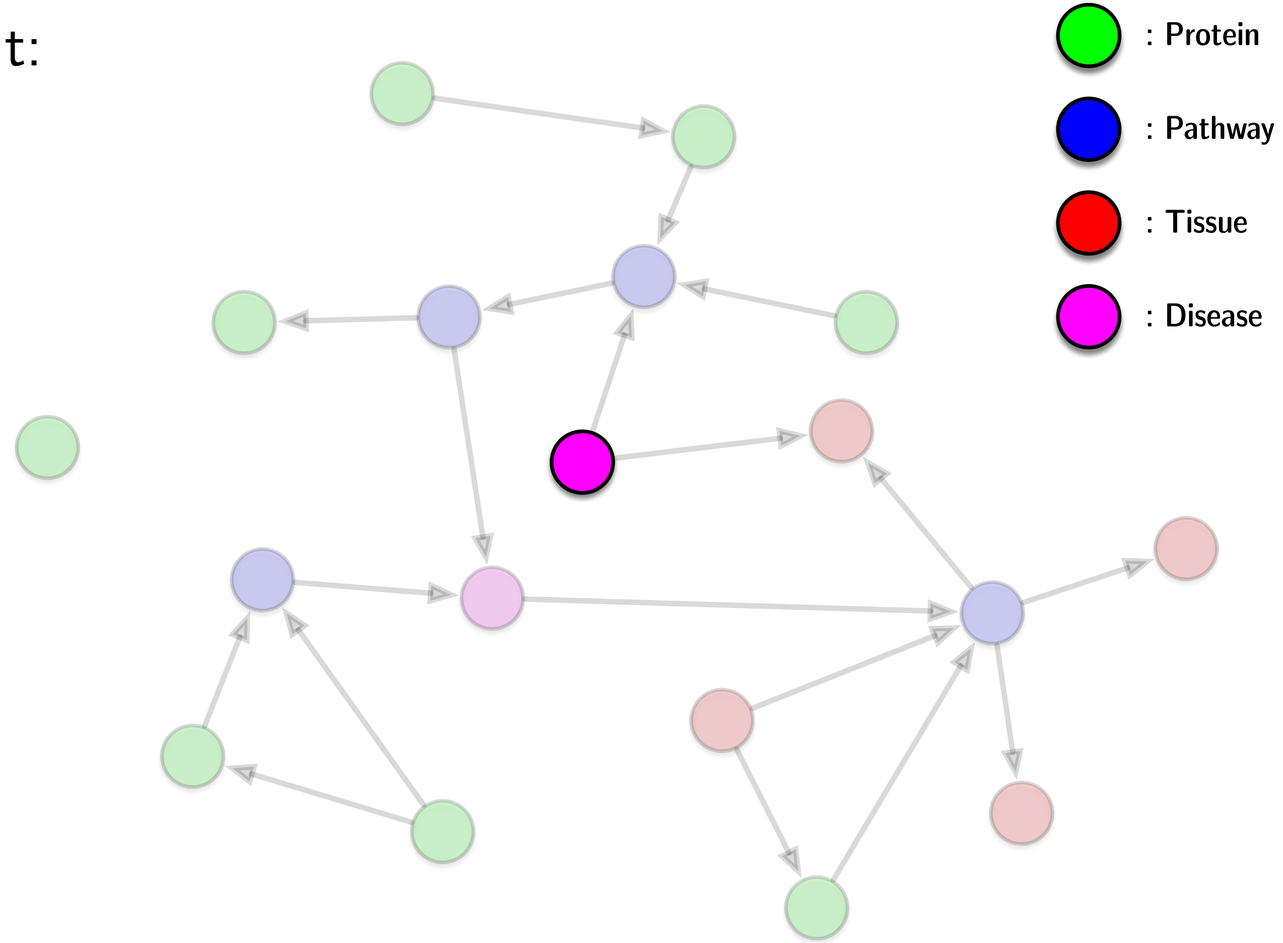


Graph Neural Networks



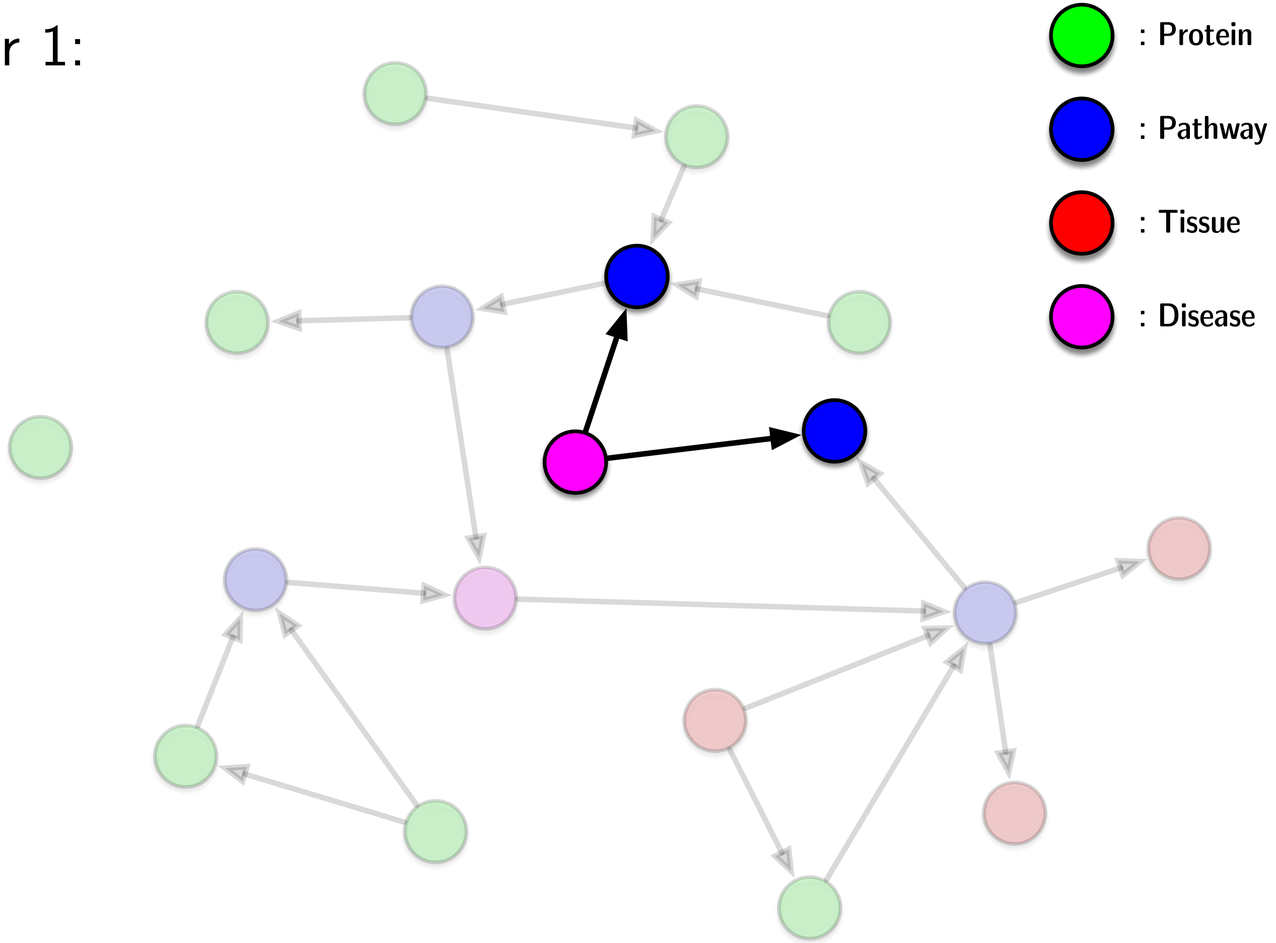
Graph Neural Networks

- Input:



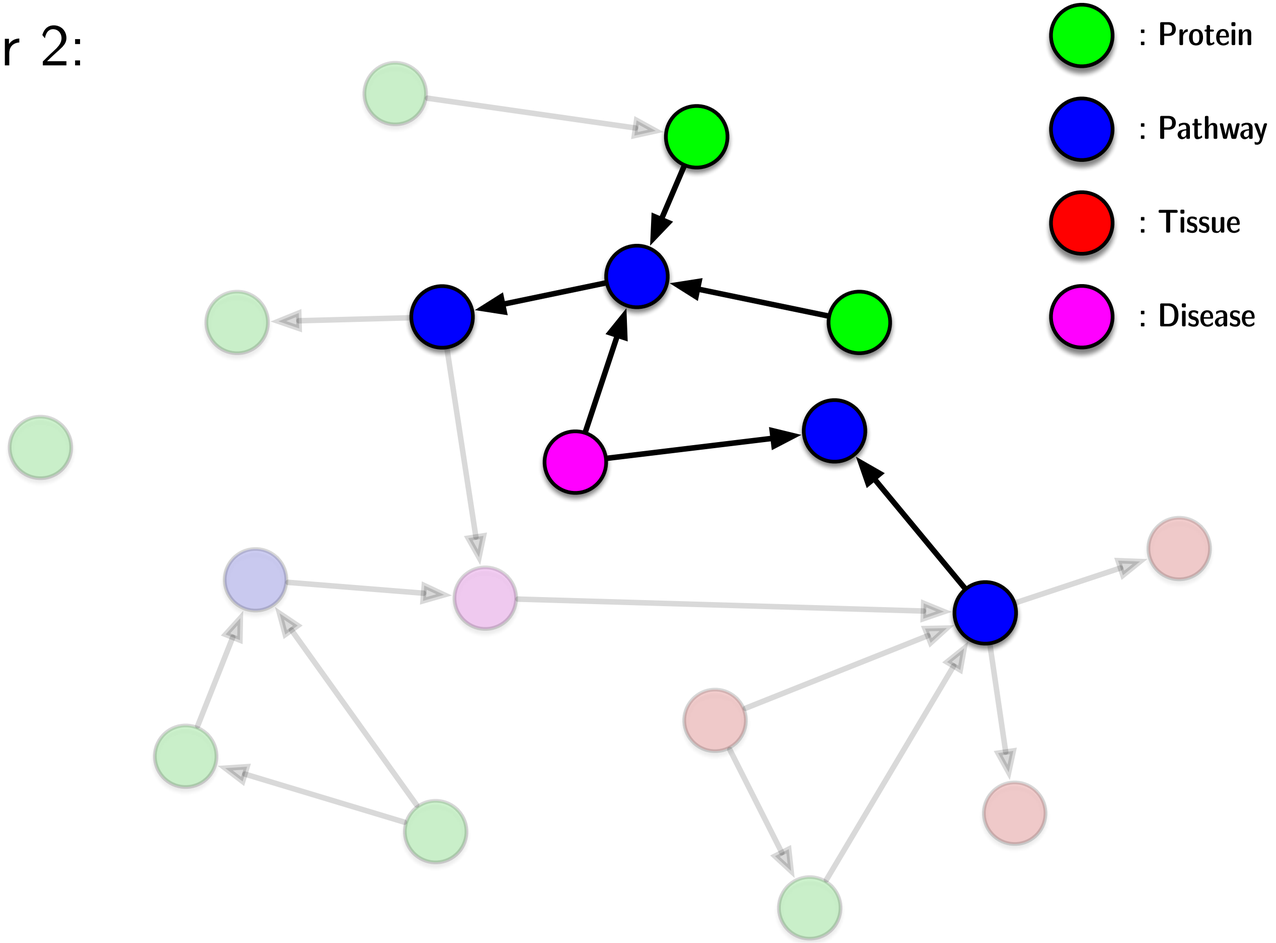
Graph Neural Networks

- Layer 1:



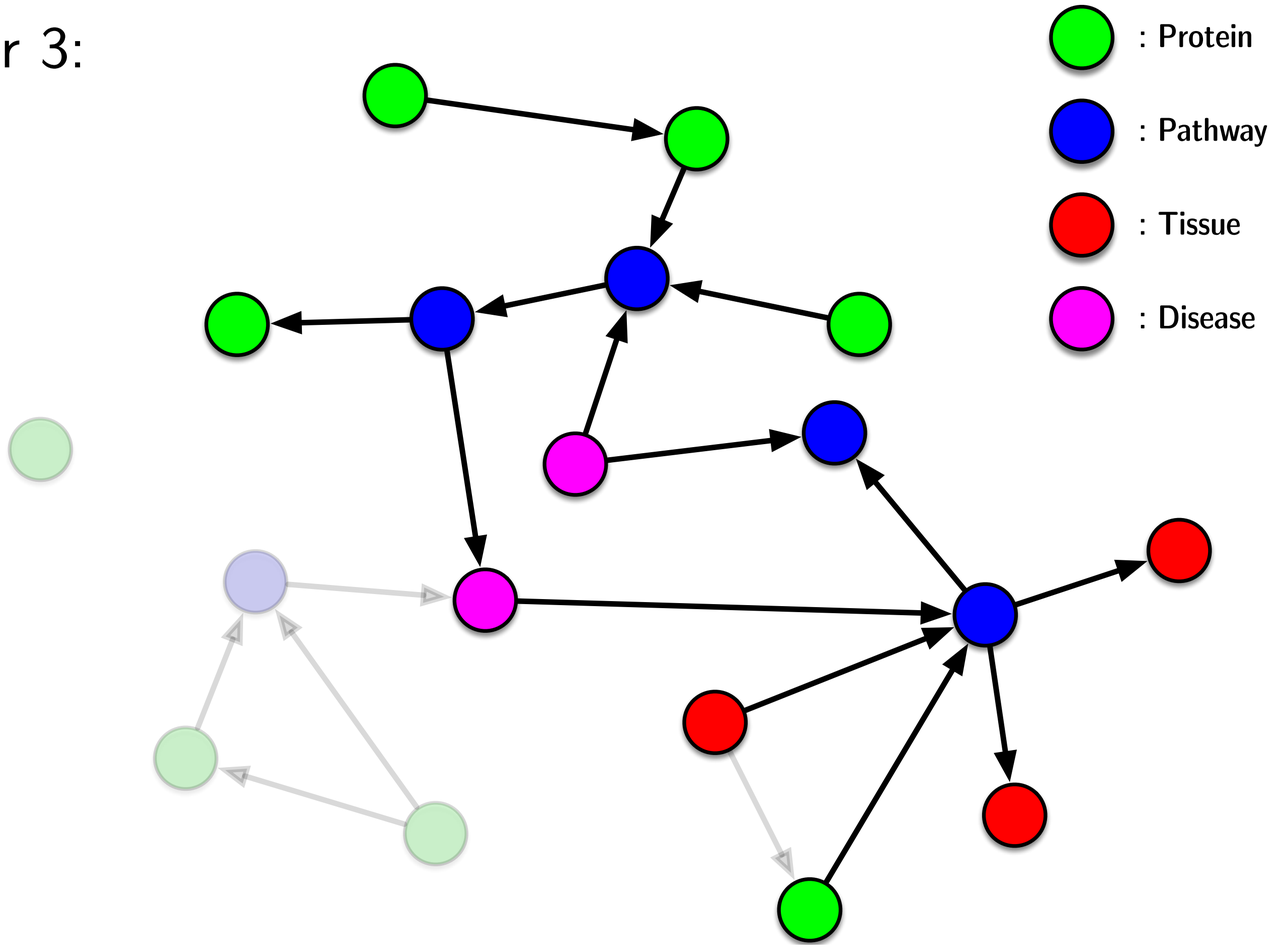
Graph Neural Networks

- Layer 2:

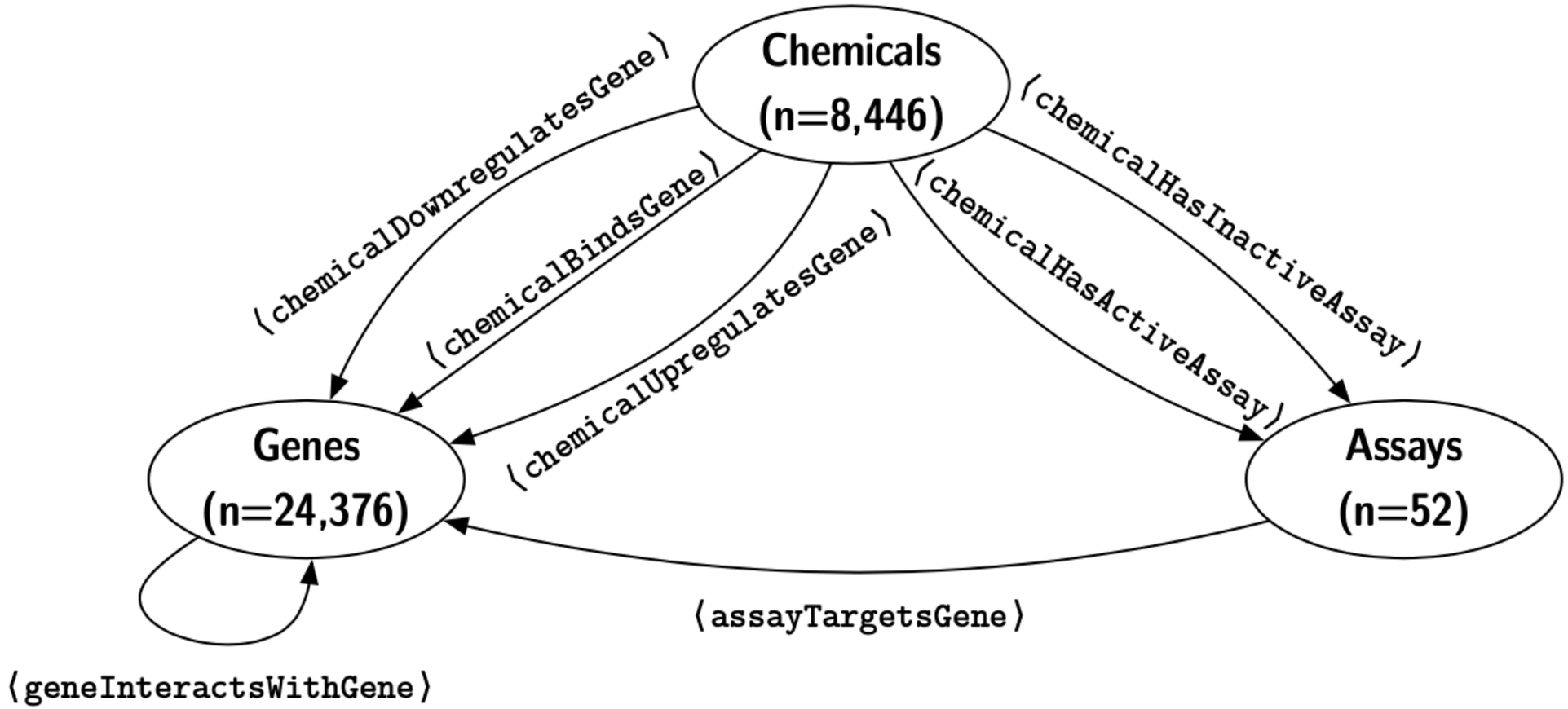


Graph Neural Networks

- Layer 3:



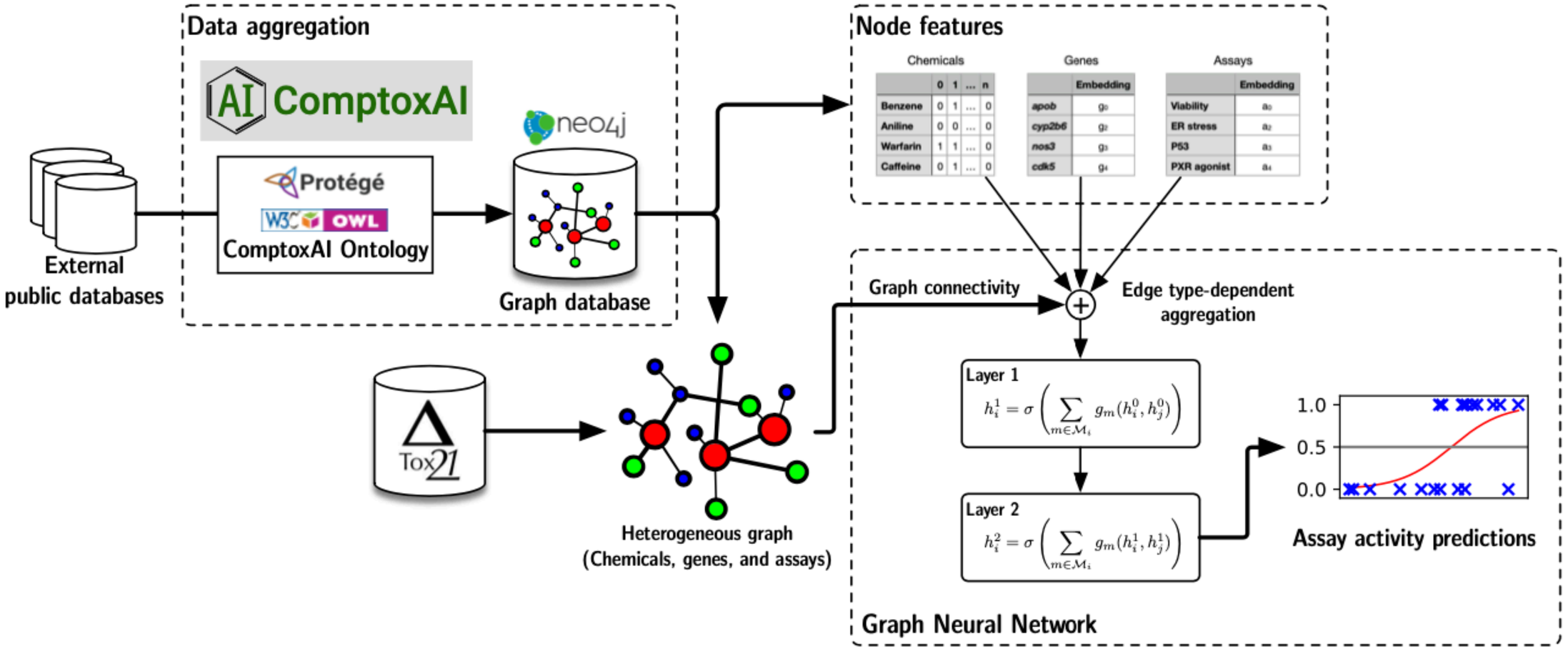
QSAR Subgraph



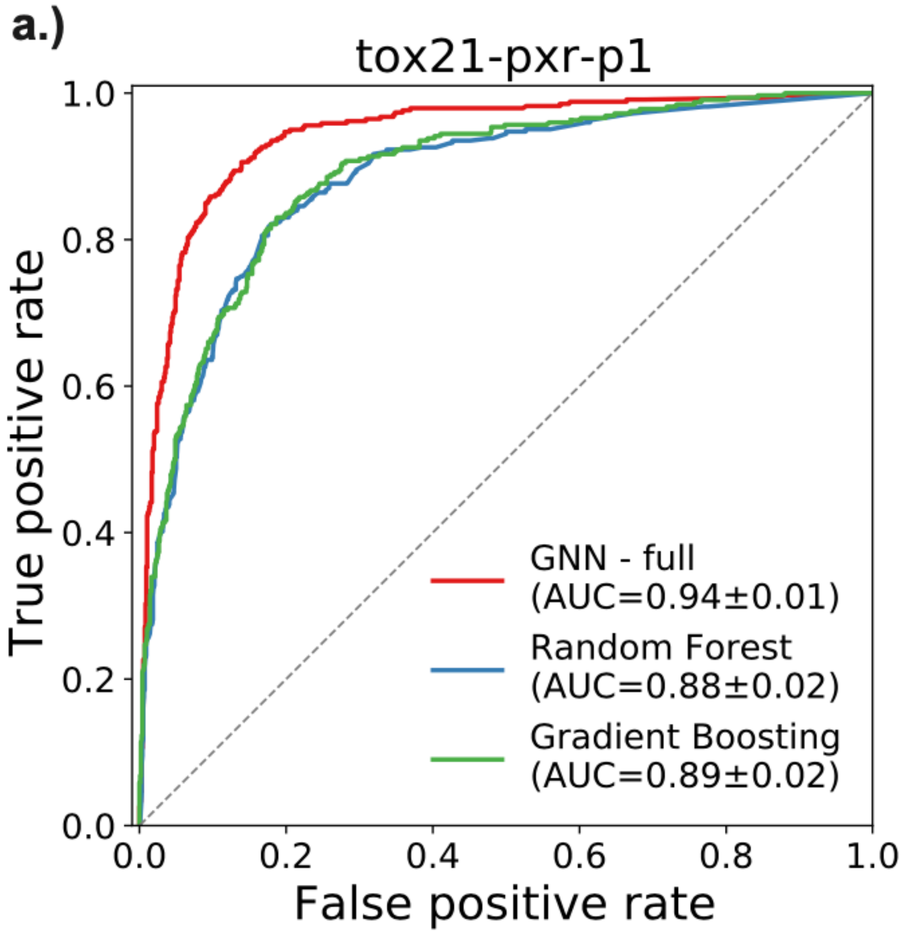
Node classification labeling algorithm

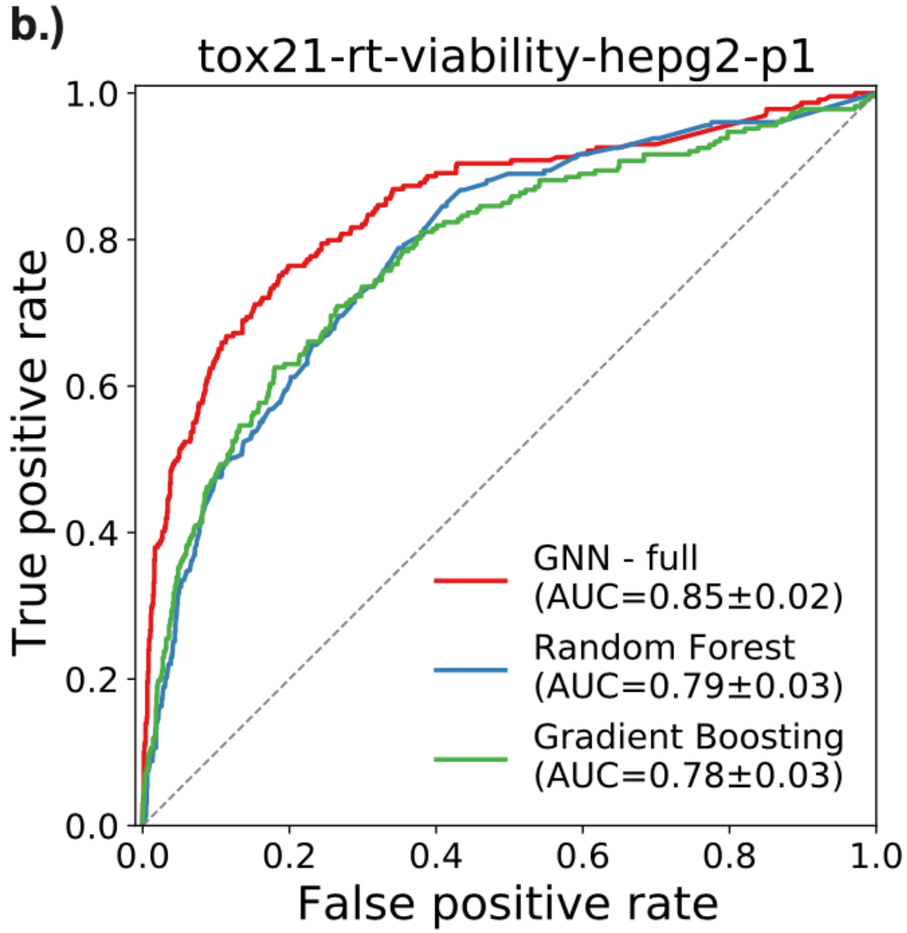
- To build a training dataset for a single assay:
  - Look at the edge linking each chemical to the assay of interest
    - If edge is “chemicalHasActiveAssay”, label the chemical “1”
    - If edge is “chemicalHasInactiveAssay”, label the chemical “0”
    - If there is no edge, don’t label the chemical
  - Remove the node (and incident edges) for the assay of interest to prevent information leakage

GNN Pipeline

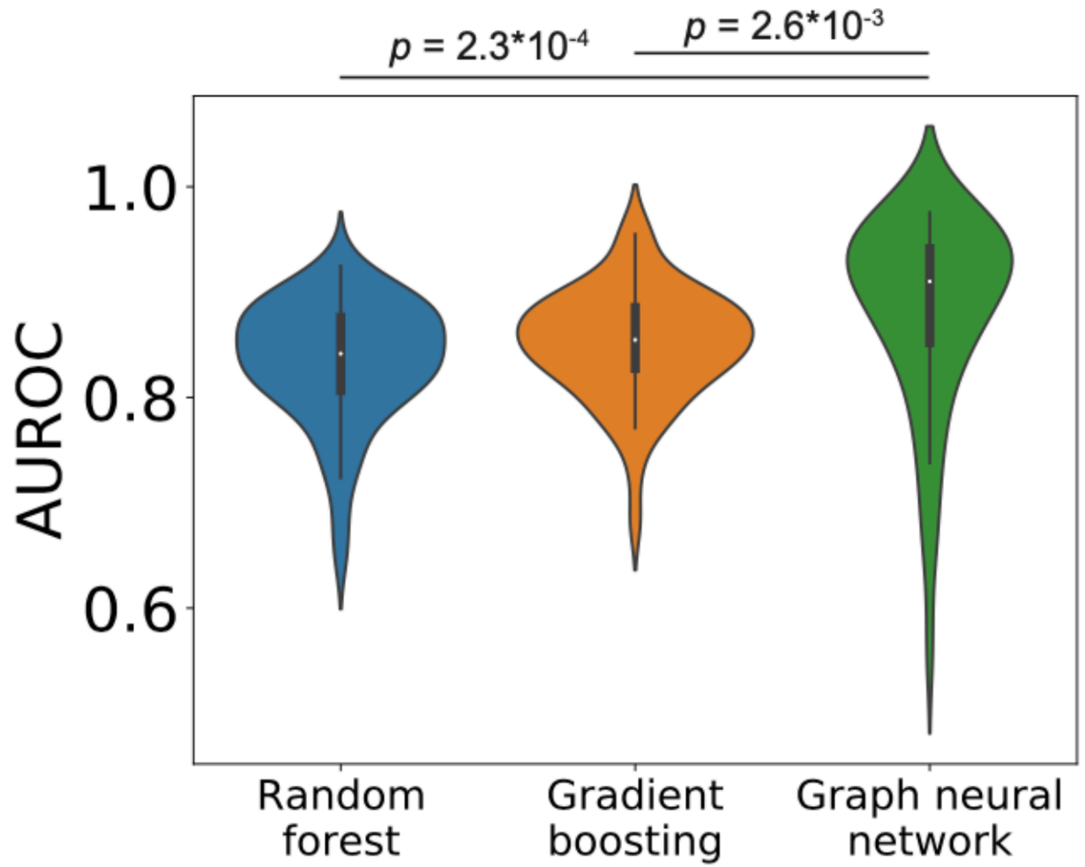


QSAR Performance

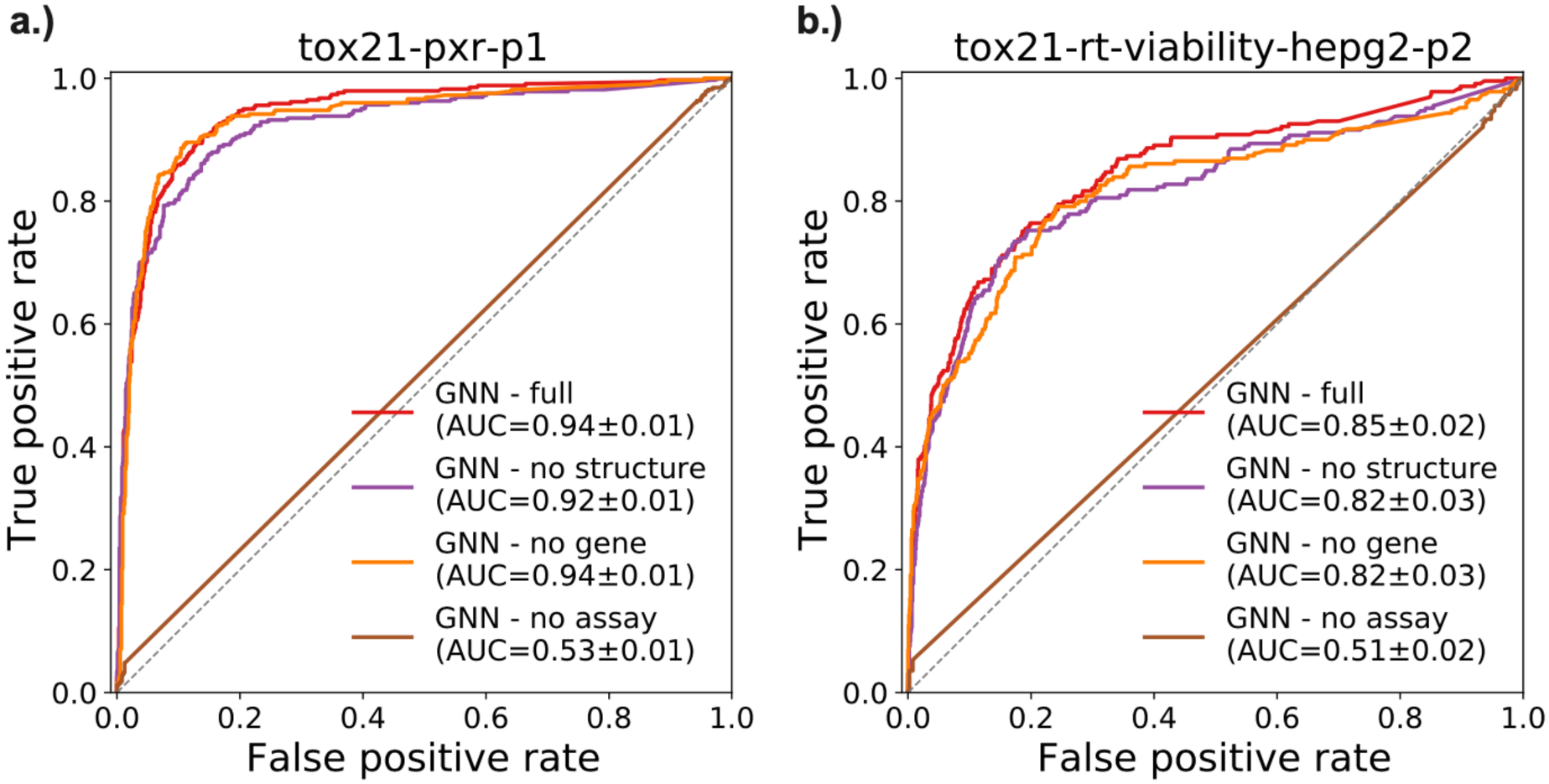




QSAR Performance



Why do the GNNs perform so much better?



Potential for model explainability

- Each relationship in the graph conveys semantic meaning based on node types and relationship types
- For any given assay's GNN, edge weights are proportional to their influence on the final prediction
- Example: *HepG2 cell viability* assay activity prediction
  - Top weighted "other" assays:
    - HepG2 Caspase-3/7 mediated cytotoxicity
    - NIH/3T3 Sonic hedgehog antagonism
  - The first makes obvious sense; is there a mechanistic explanation for the other?

Future work

- Expand on the concept types included in the subgraph (i.e., add diseases, pathways, cell types, etc.)
- Test continuous endpoints (IC_{50}, etc.)
- Evaluate more complex network architectures:
  - Link prediction models
  - Use regularization to better utilize information from non-Assay nodes (important for Graph ML in heterogeneous networks)
  - Deeper networks? May be useful as the network grows
- Develop easy-to-use graphical tools to lower the barrier for diverse user types
  - Use ontology reasoning to further improve explainability

- **Let me know if you want to use ComptoxAI in your research**! I'd be happy to give you a plug on the website.
  - joseph.romano@pennmedicine.upenn.edu
- I'm always happy to take suggestions, questions, and contributions (data, code, documentation, etc.)
- Check back in a few weeks for a more complete feature set including everything described in this talk (and more!)

Acknowledgements

- Jason Moore (Cedars-Sinai)
- Trevor Penning (Penn)
- Yun Hao (Penn)
- Li Shen (Penn)

- Grant funding:
  - K99-LM013646 (Romano)
  - T32-ES019851, P30-ES013508 (Penning)
  - R01-LM010098, R01-LM012601, R01-AI116794, UL1-TR001878, UC4-DK112217 (Moore)

Center of Excellence in Environmental Toxicology

Institute for Biomedical Informatics

Perelman School of Medicine
University of Pennsylvania

@jdromano2

Website: jdr.bio

Email: joseph.romano@pennmedicine.upenn.edu

Additional slides:

GCN Architecture details

Each layer of the network is defined as an edge-wise aggregation of adjacent nodes:

$$h_i^{(l)} = \sigma\left(\sum_{r\in\mathcal{R}} \rho_{j\in\mathcal{N}_i^r}\left(W_r^{(l-1)} h_j^{(l-1)} + W_0^{(l-1)} h_i^{(l-1)}\right)\right). \tag{A.1}$$

where h_i^l is the hidden representation of node i in layer l, $\mathcal{N}(i)$ is the set of immediate neighbors of node i, and σ is a nonlinear activation function (either softmax or leaky ReLU, as explained in **Appendix B**). ρ can be any differential 'reducer' function that combines messages passed from incident edges of a single type; in the case of this study we use summation. Since our graph contains relatively few edge types, regularization of the weight matrices W is not needed.

(See paper for more details)

Node Classification details

For classifying chemicals as active or inactive with regards to an assay of interest, we stack 2 GCN layers in the form given by (A.1), with a leaky ReLU activation between the two layers and softmax applied to the second layer's output. Since we only classify chemical nodes, we ignore outputs for all other node types (and for chemicals with undefined labels); labels are generated via **Algorithm 1** We train the network by minimizing binary cross-entropy between the network's softmax outputs and true activity values:

$$\mathcal{L} = -\sum_{i \in \mathcal{Y}} \ell(h_i^{(0)}) \cdot \ln h_i^{(2)} + (1 - \ell(h_i^{(0)})) \cdot \ln(1 - h_i^{(2)}). \tag{B.1}$$

where $\mathcal{Y}$ is the set of all labeled nodes, $\ell(h_i^{(0)})$ is the true label of node i, and $h_i^{(2)}$ is the final layer output of node i.

The relatively shallow architecture of the network allows us to optimize the model using the Adam algorithm applied to the entire training data set, but the model can be adapted to mini-batch training when appropriate or necessary.

(See paper for more details)