Knowledge-driven AI enables mechanistic predictive toxicology

Presented by Joseph D. Romano, PhD

November 2, 2023

DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Center of Excellence in Environmental Toxicology

Penn
University of Pennsylvania

Penn Institute for Biomedical Informatics

The Romano Lab @ Penn

Joe Romano, PhD
Assistant Professor of Informatics and Pharmacology

Chloé Paris
PhD Student (GCB)

Yunchae Kim, MS
Software Developer

Estefania Morales
Summer Undergraduate Internship Program (SUIP) Fellow

Not Pictured:

Kevin Shen
Undergraduate Researcher

Tom Pan
Undergraduate Researcher

We study...

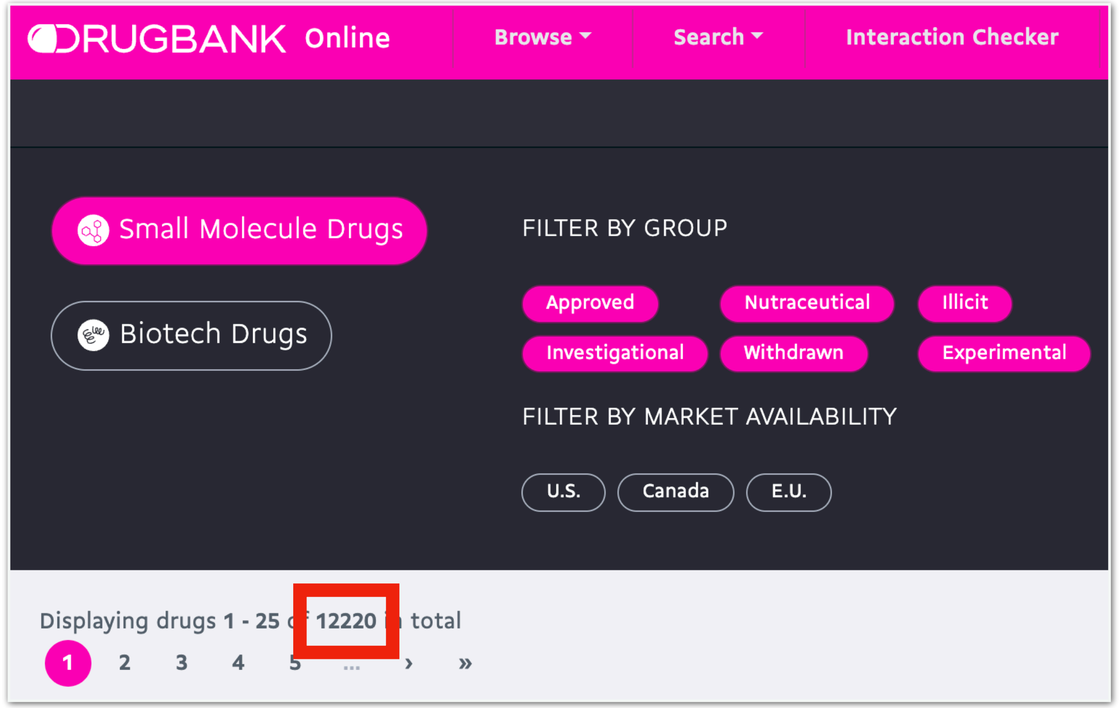
- Computational toxicology
  - Designing interpretable AI models to predict toxicity
  - Discovering *mechanisms* linking specific chemicals to human diseases
- Autoimmunity
  - Biomarker discovery for immune-related adverse events (irAEs)
- Clinical informatics
  - How to computationally model and extract knowledge from electronic health record data
- Artificial Intelligence and Machine Learning
  - Multimodal graph machine learning
  - Automated machine learning
  - Large language models (LLMs) and their applications in biomedicine

Computational Toxicology and Artificial Intelligence

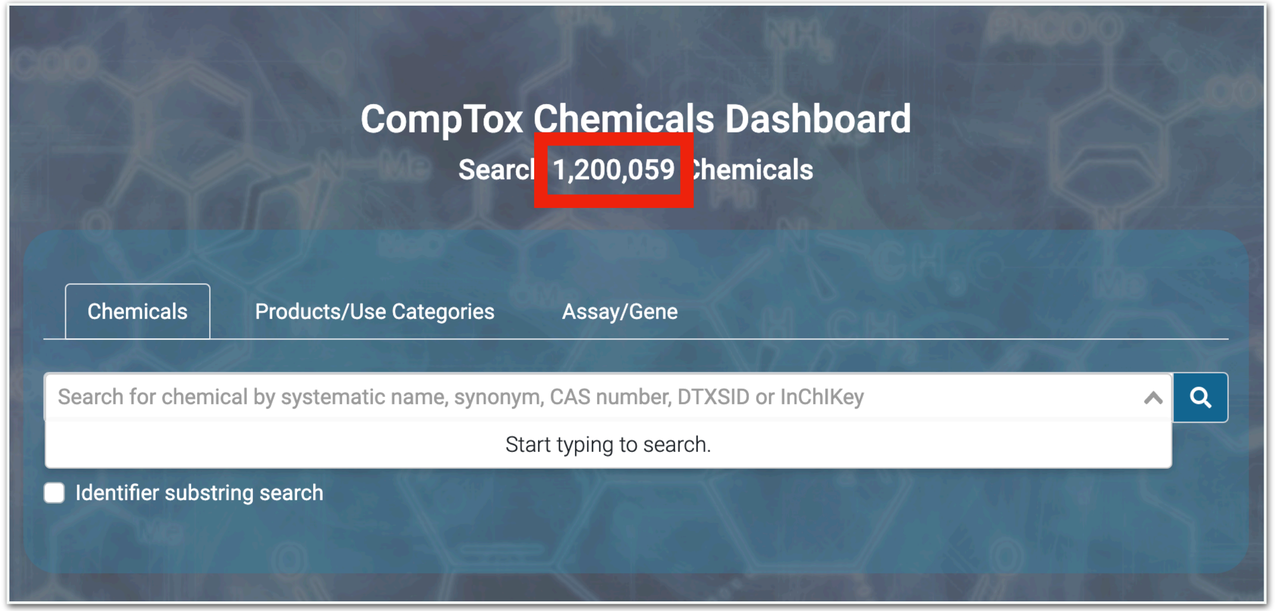
The role of AI in toxicology

- AI should help experimental 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 prioritize our time and effort

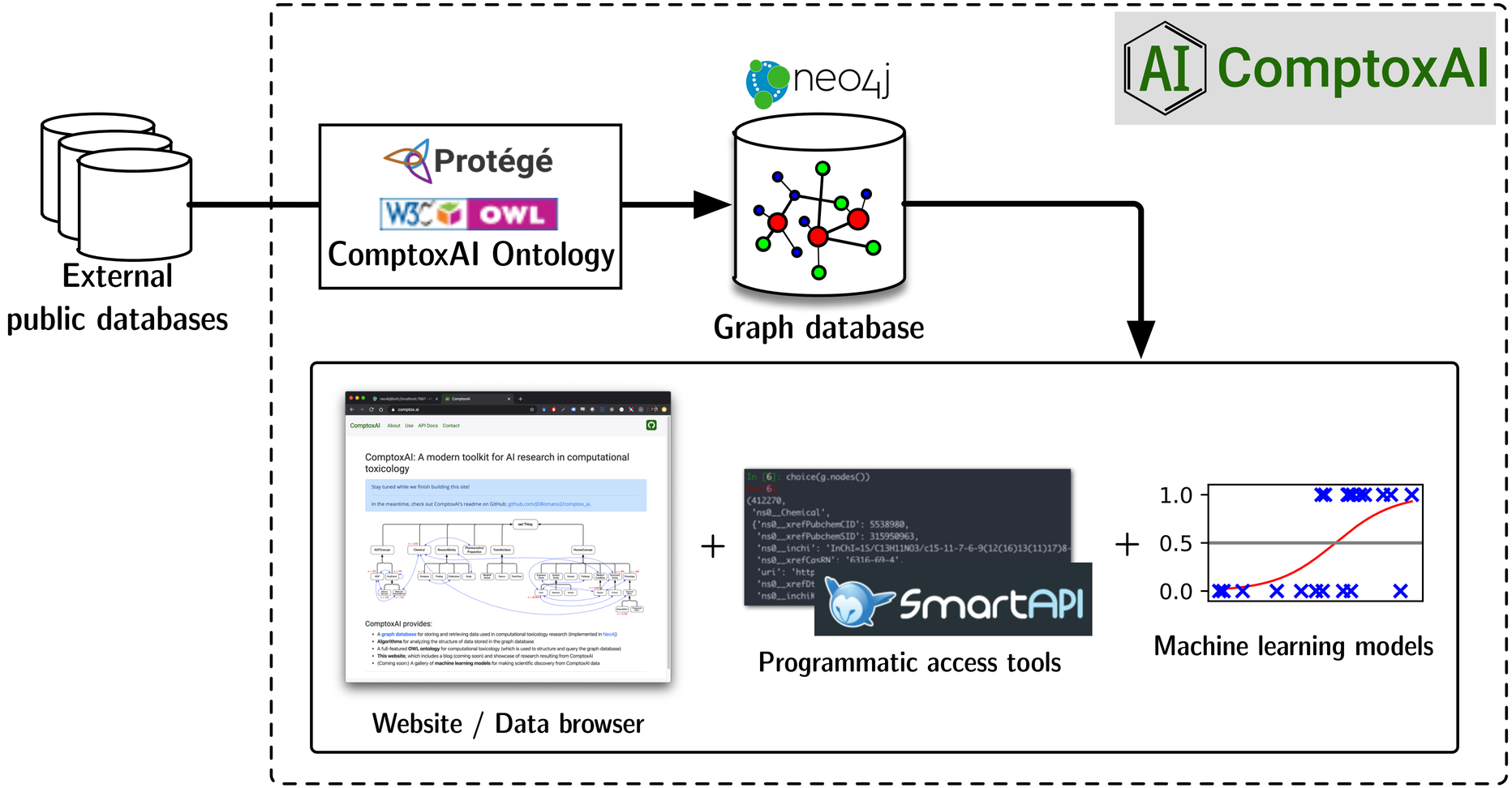
Drugs (incl. withdrawn, investigational…)



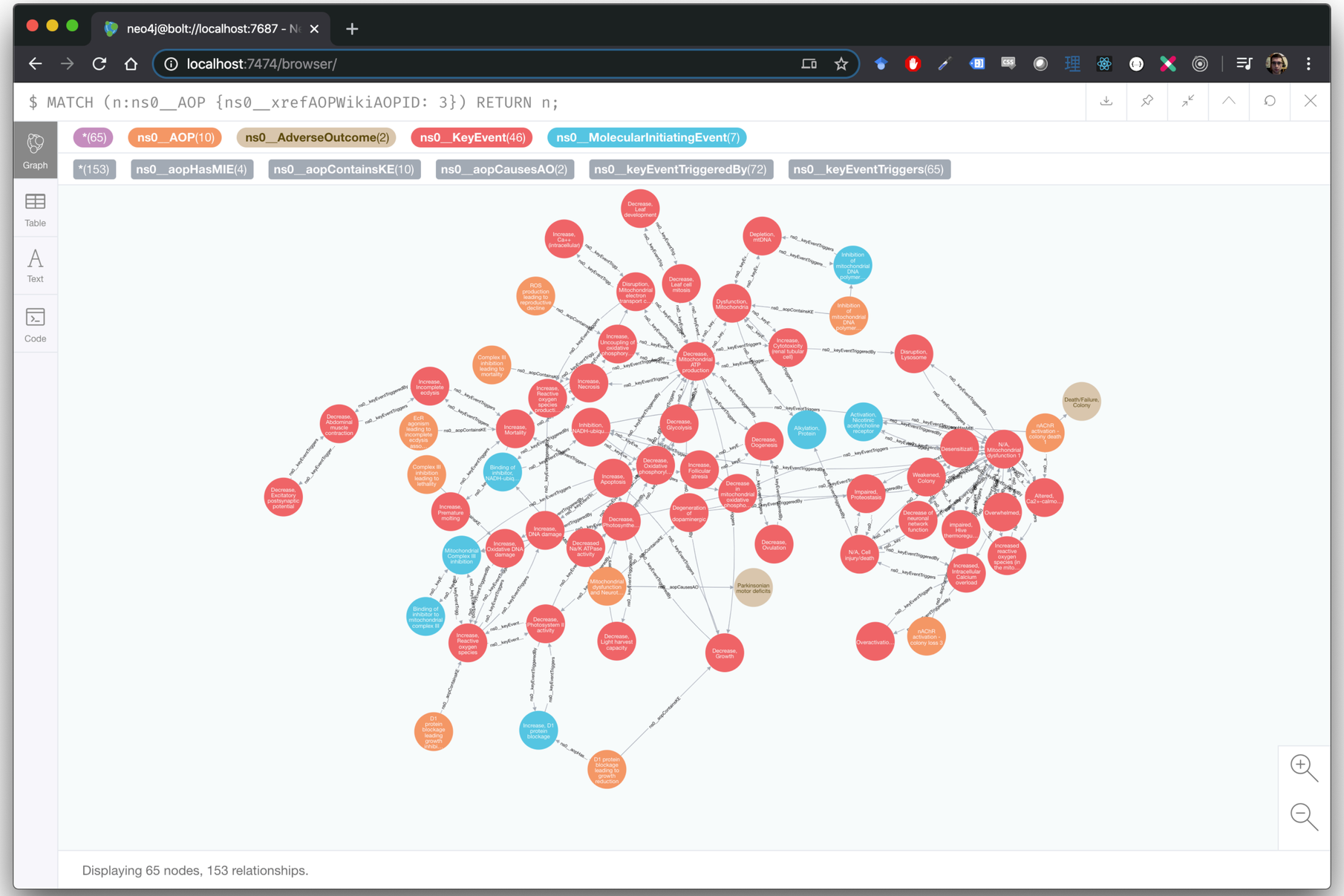
Environmental chemicals of toxicological concern



ComptoxAI



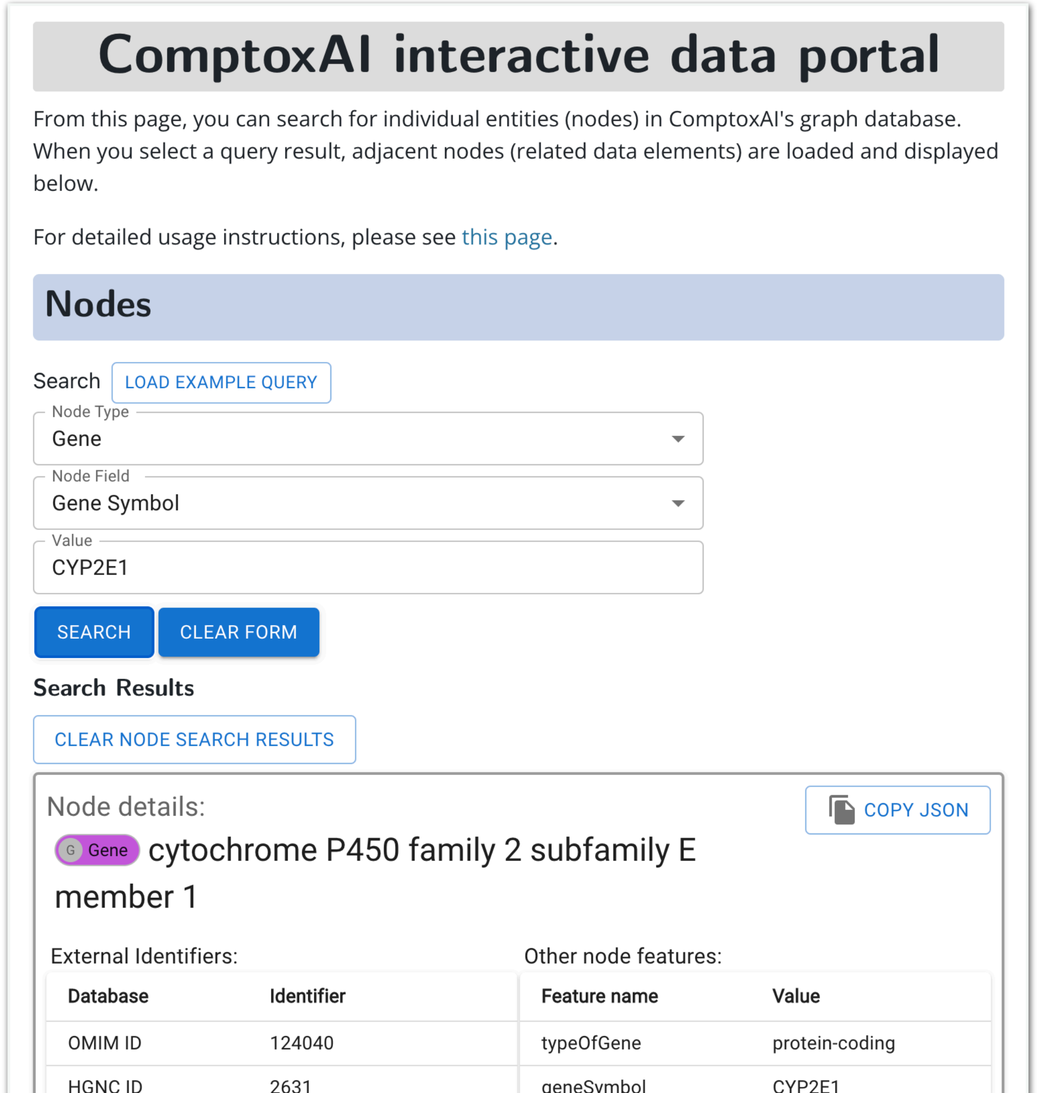
https://comptox.ai


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.

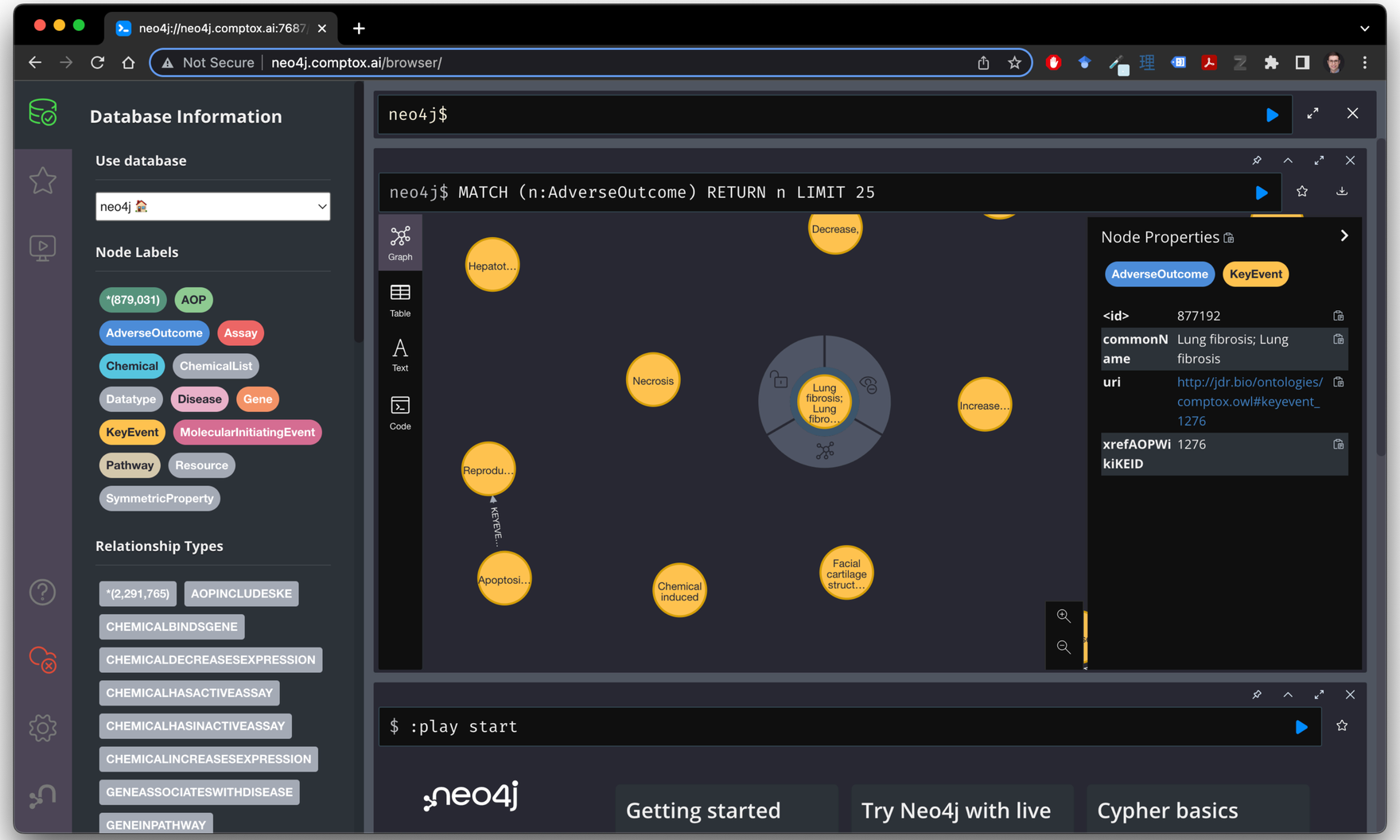
ComptoxAI: Data Interfaces

- 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)

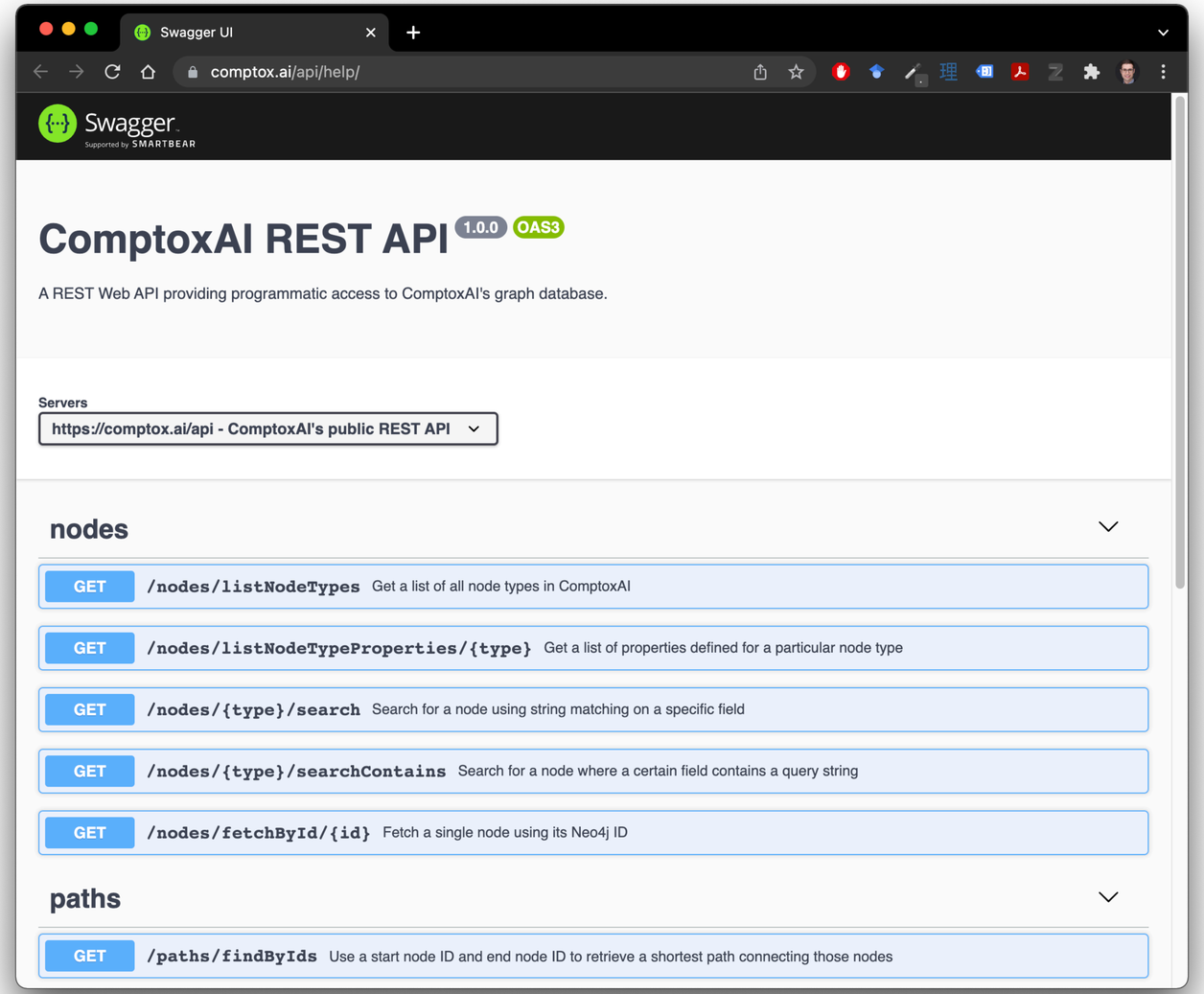
Simplified data browser



Graph database: Neo4j browser



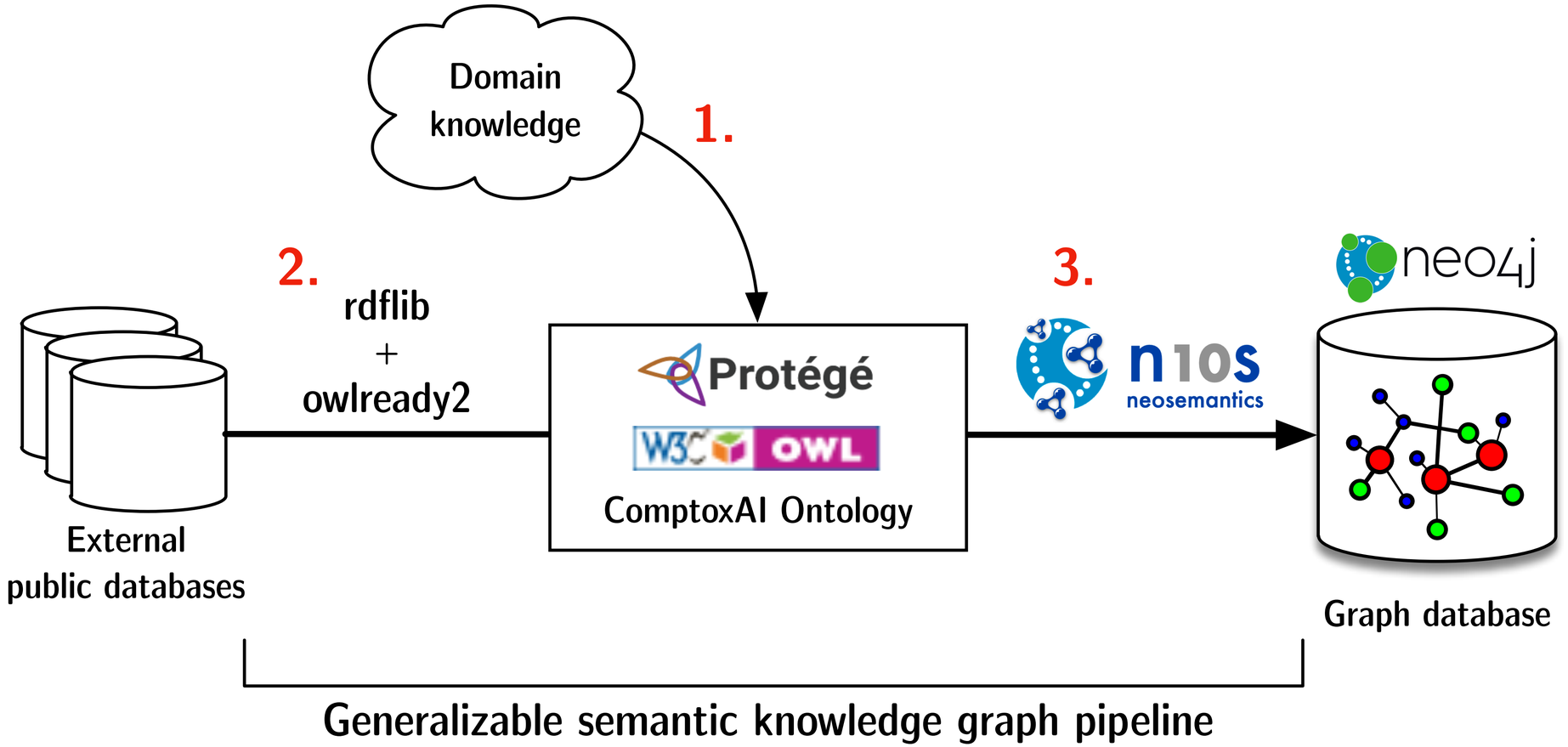
Web API


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

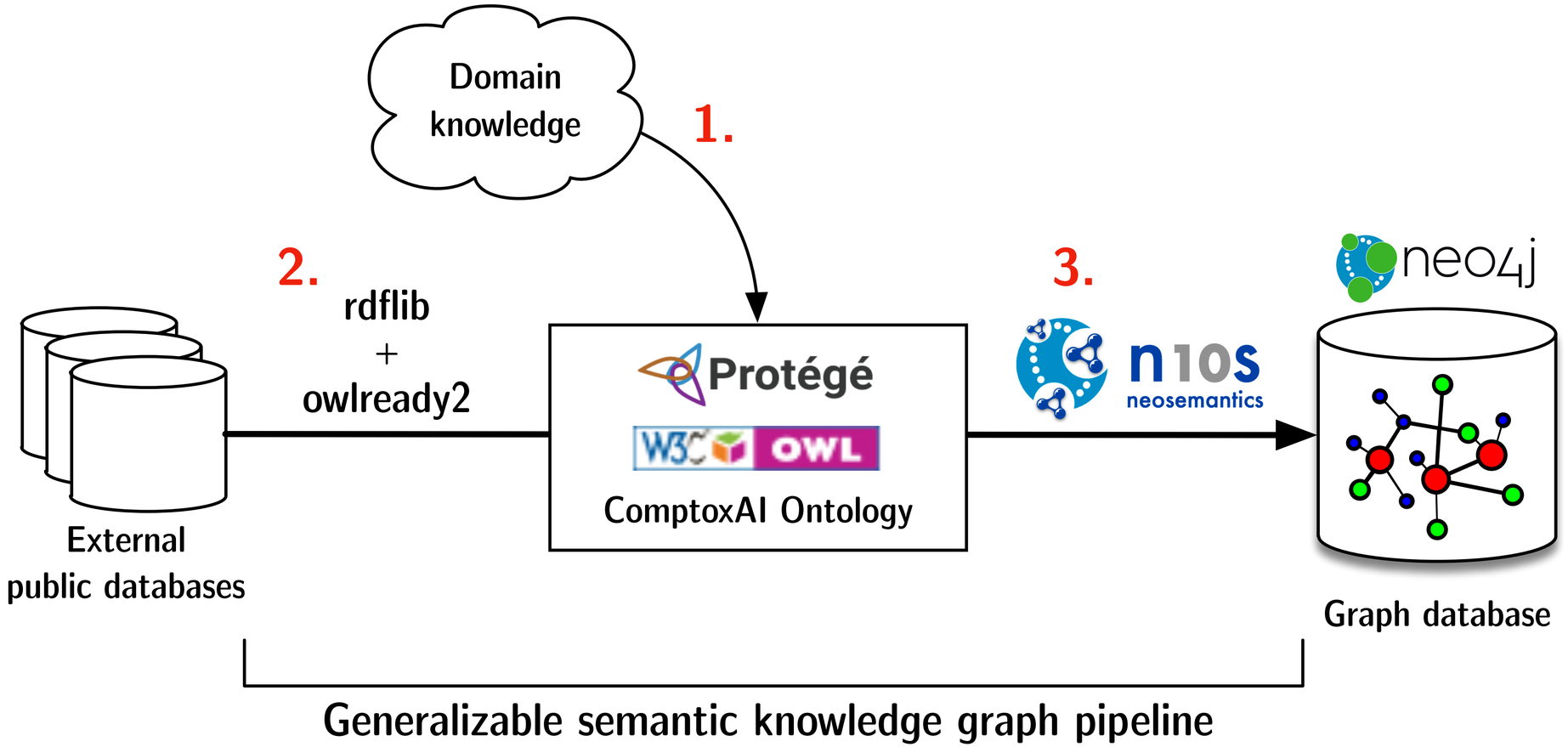
Python package

```
Python 3.7.4 (default, Aug 13 2019, 15:17:50)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.13.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: from comptox_ai.db import GraphDB
RDFLib Version: 5.0.0

In [2]: db = GraphDB(hostname="neo4j.comptox.ai")

In [3]: result = db.run_cypher("MATCH (c:Chemical {commonName: 'PFOA'}) RETURN c;")

In [4]: print(result)
[{'c': {'commonName': 'PFOA', 'maccs': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0], 'xrefMeSH': 'MESH:C023036', 'xr
efPubchemSID': '315676838', 'xrefDTXSID': 'DTXSID8031865', 'xrefPubchemCID': '9554', 'xrefCasRN': '335-67-1', 'uri
': 'http://jdr.bio/ontologies/comptox.owl#chemical_dtxsid8031865'}}]
```

Fragmented data —> Knowledge graph



Fragmented data —> Knowledge graph



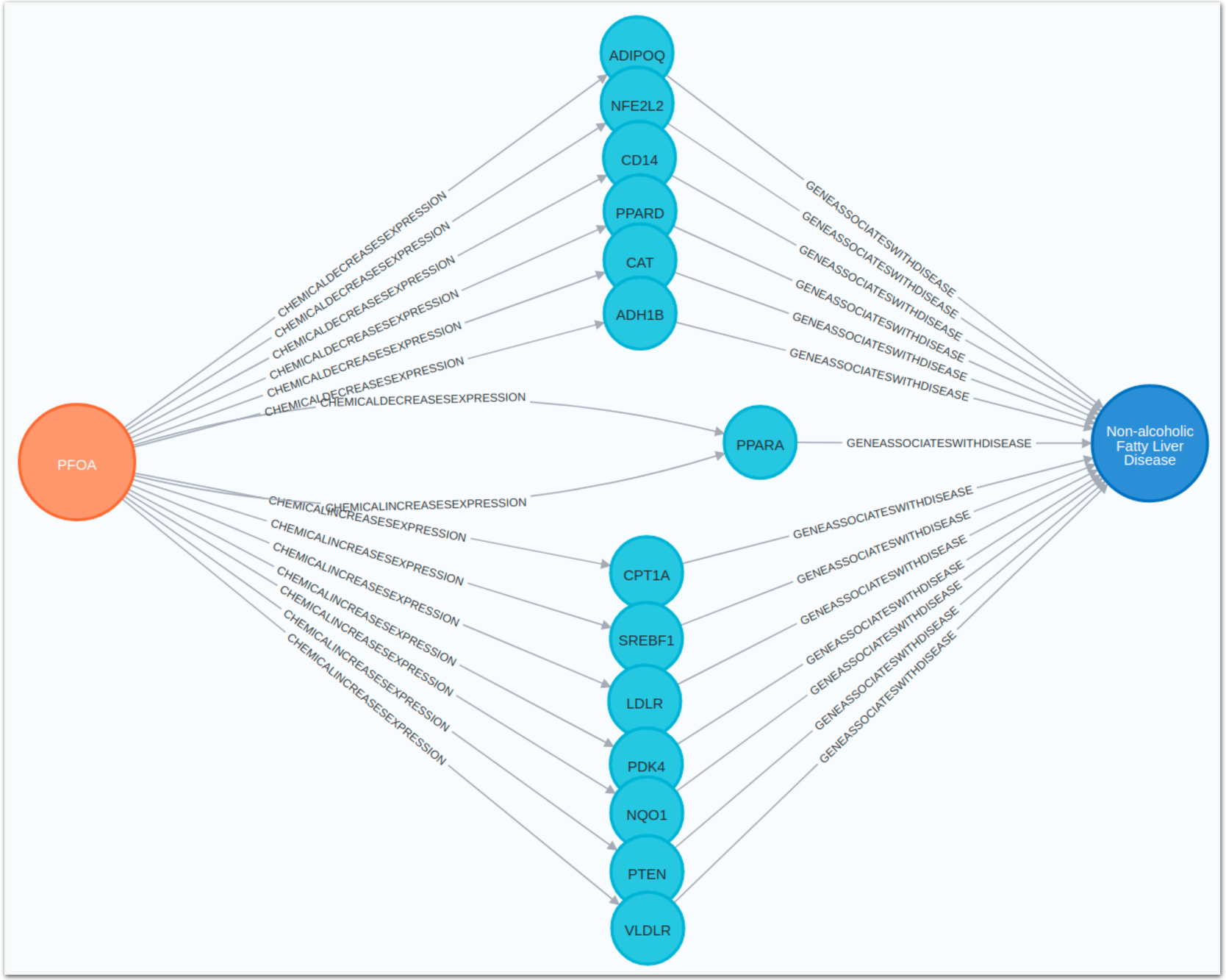
Now used in ComptoxAI, AlzKB, and other KGs under development!

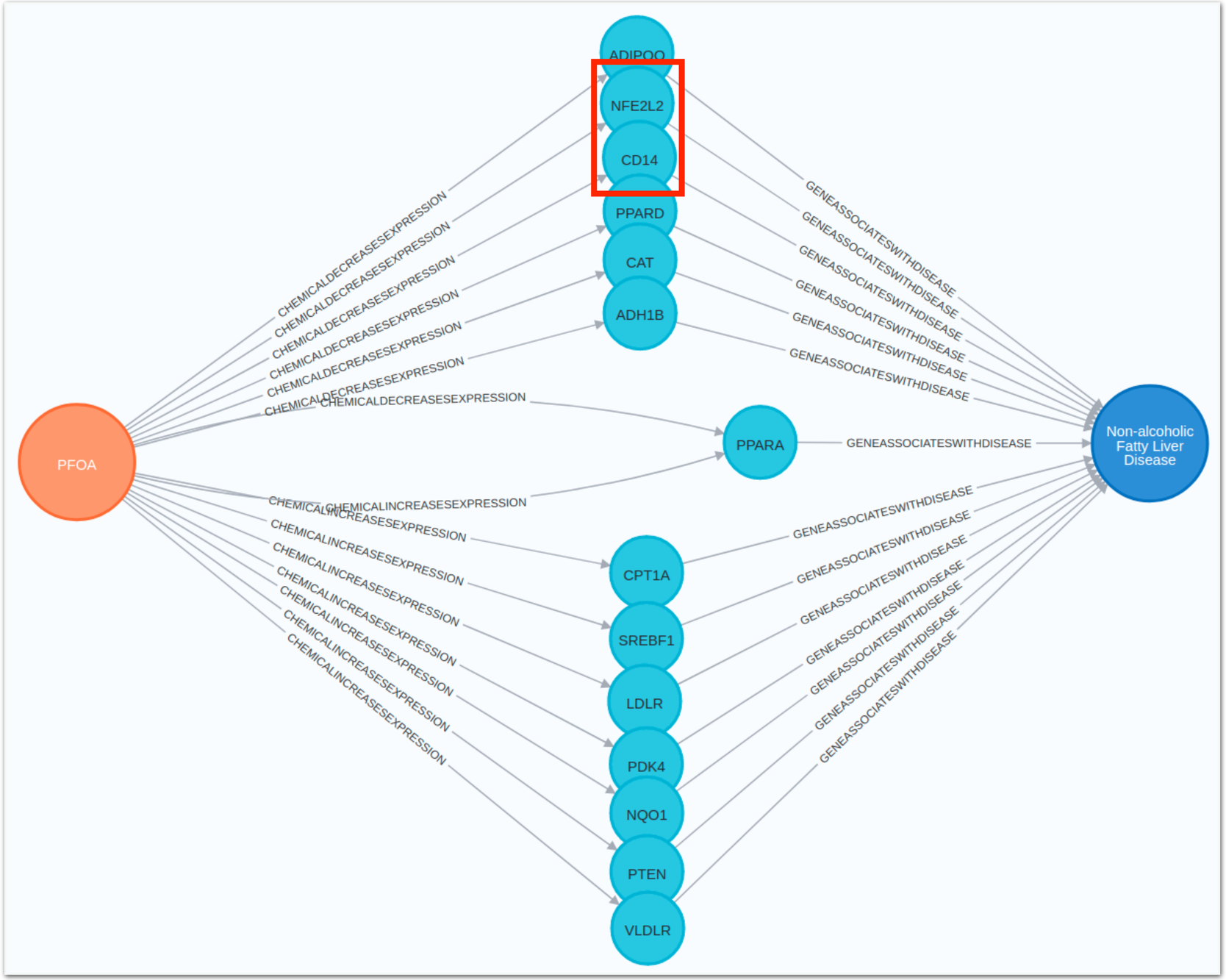
2 paradigms for discovery:

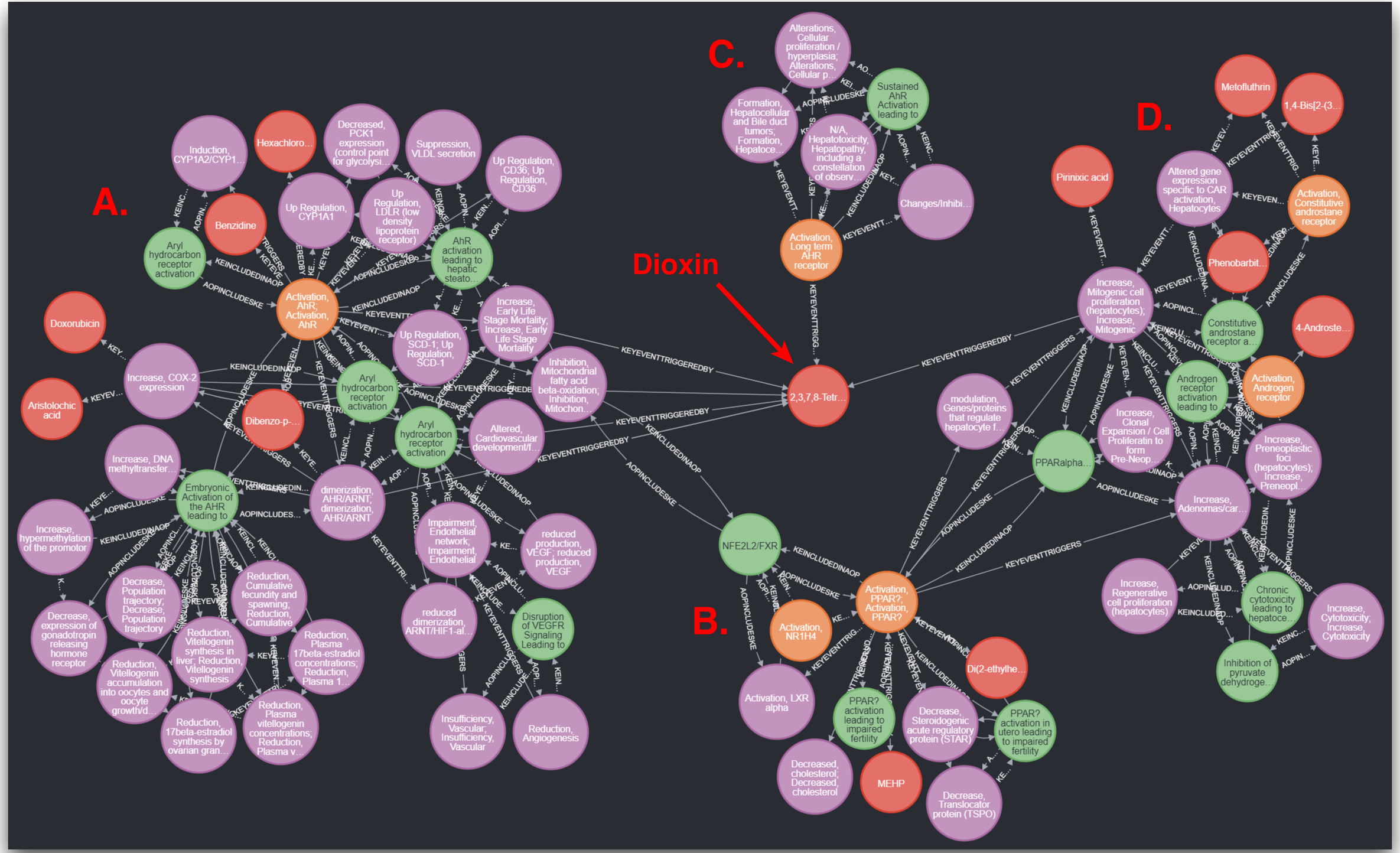
- **Directly inferring relationships from network structure**
  - Stitching together chains of relationships originally fragmented across multiple sources
  - Inspecting higher-order network architecture
- **Predicting new, unobserved relationships**
  - Machine learning (especially graph machine learning)

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)


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


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

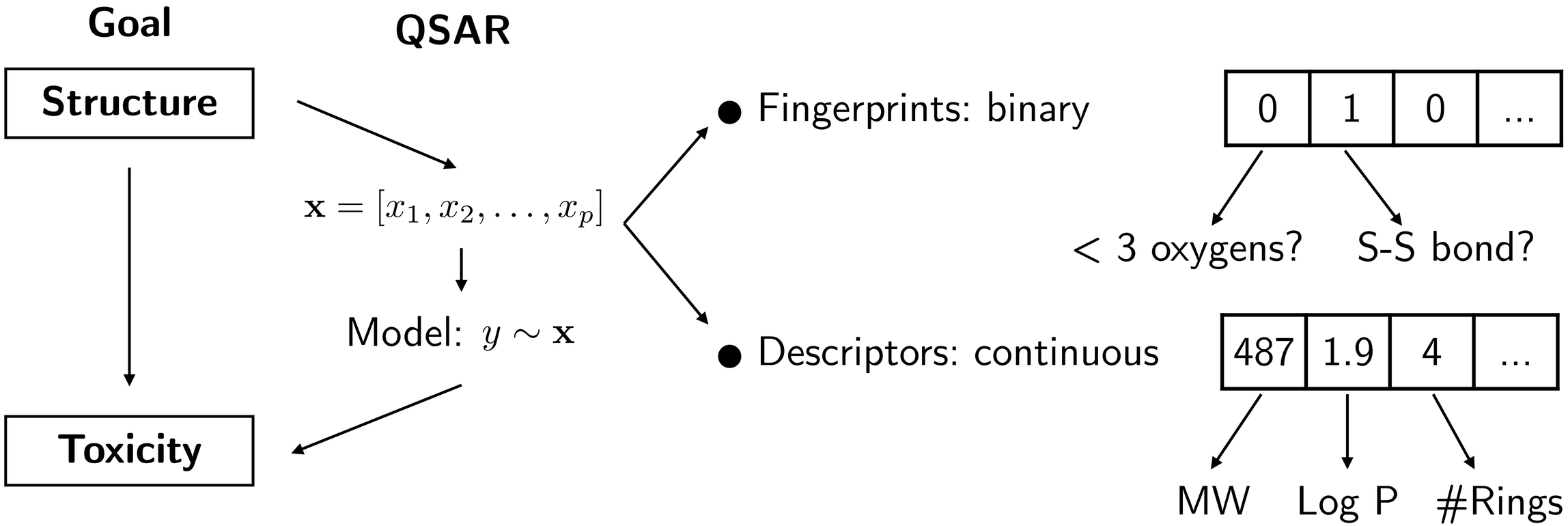
2 paradigms for discovery:

- **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)

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)

QSAR

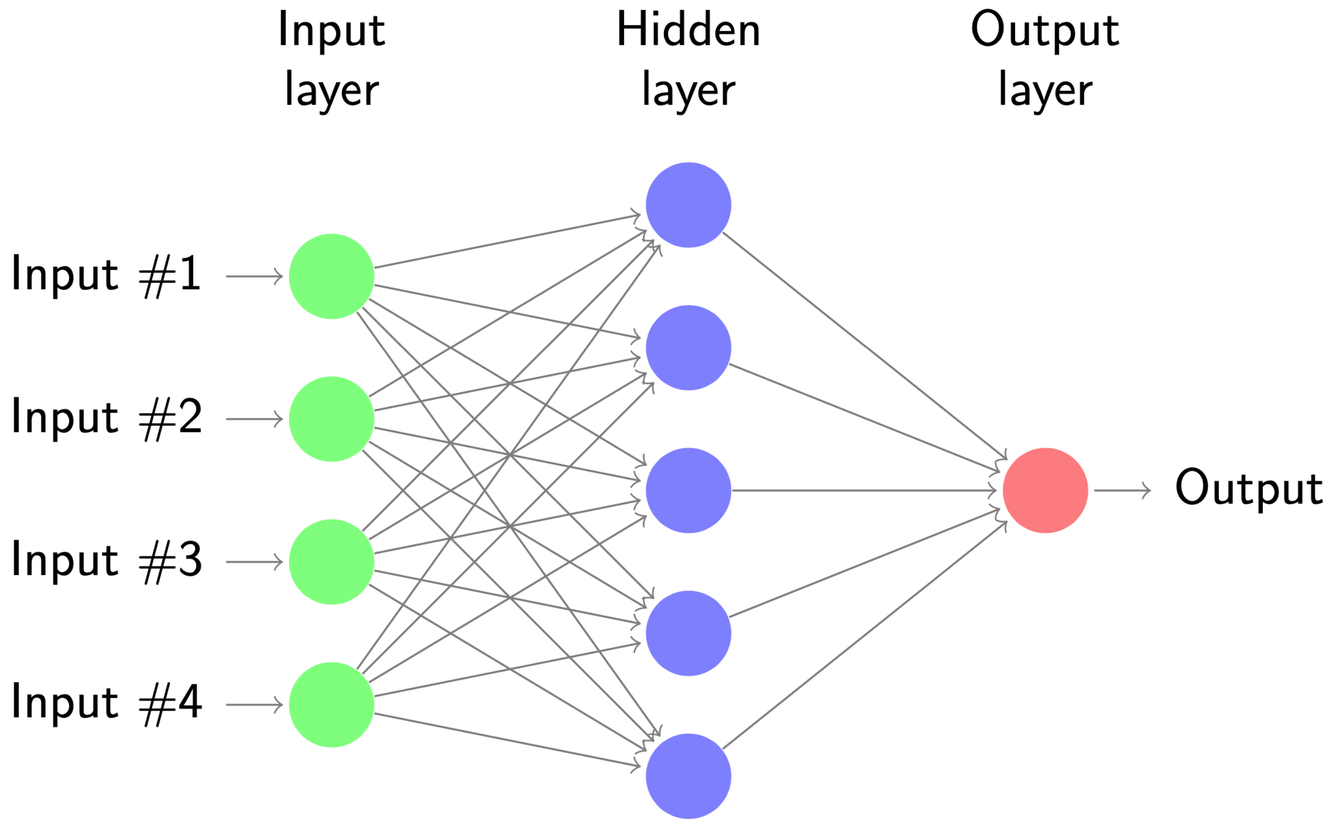


- QSAR: Quantitative Structure-Activity Relationship

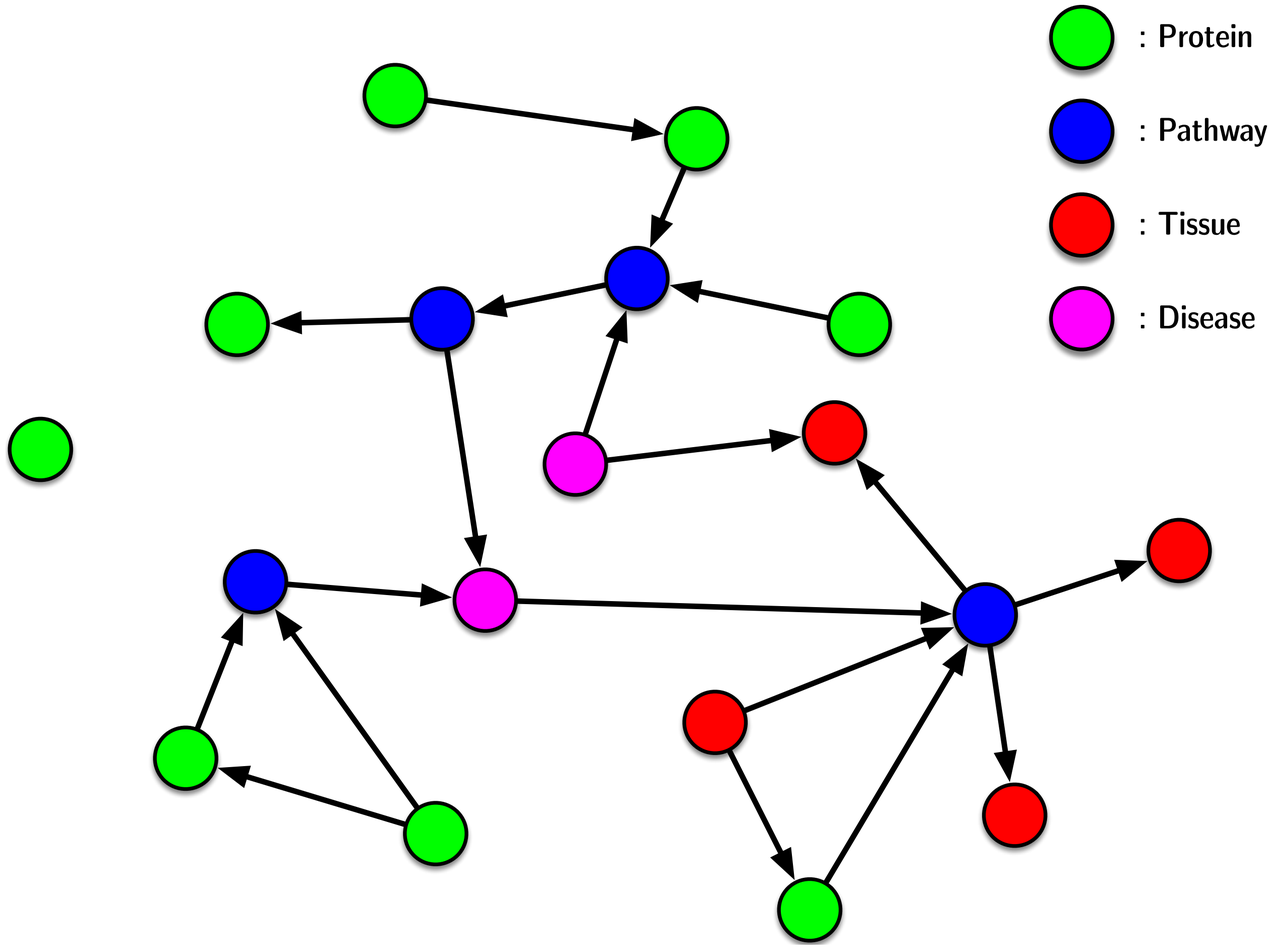
Slide credit: Yun Hao

(Artificial) Neural Networks

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

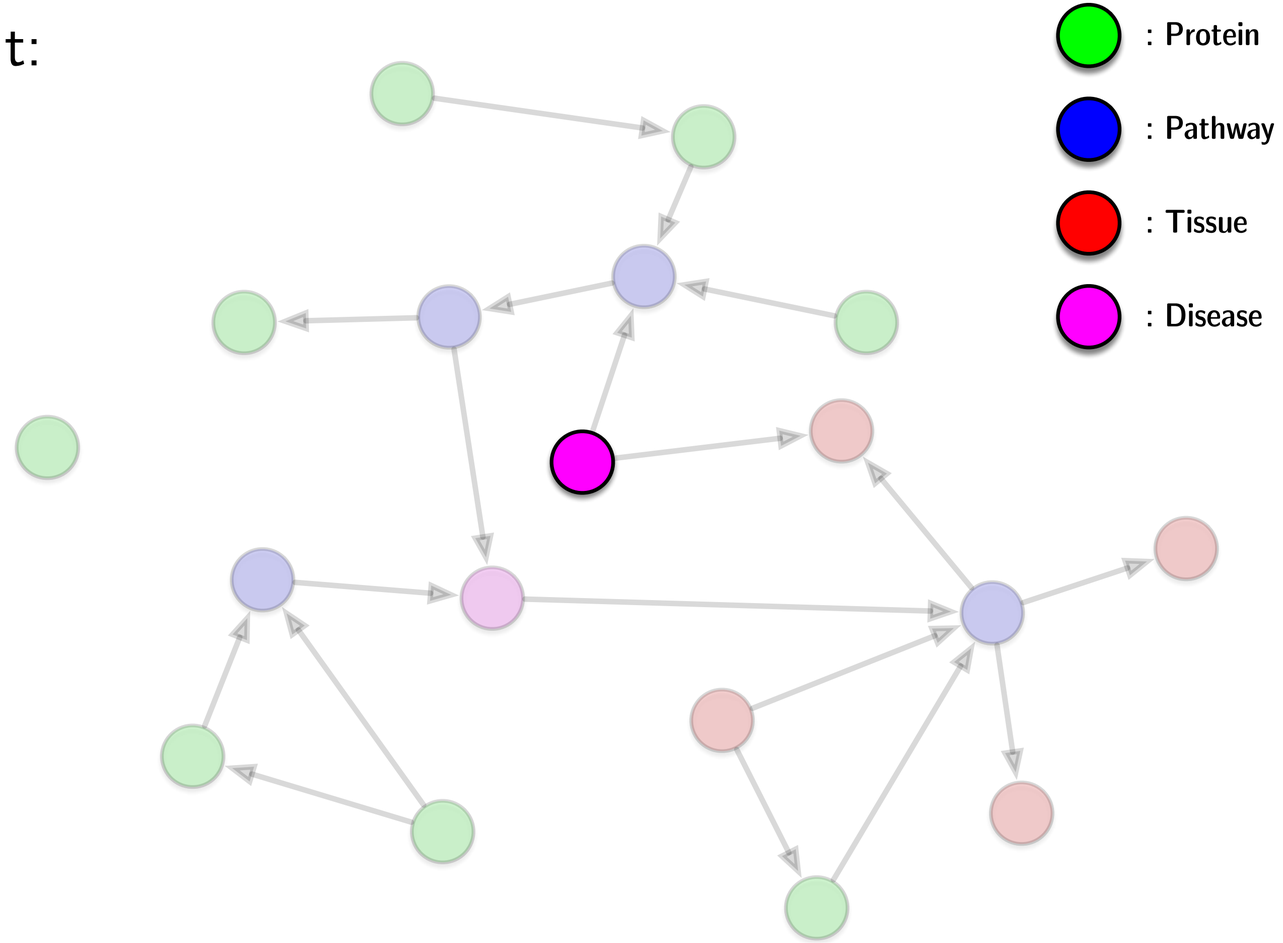


Graph Neural Networks



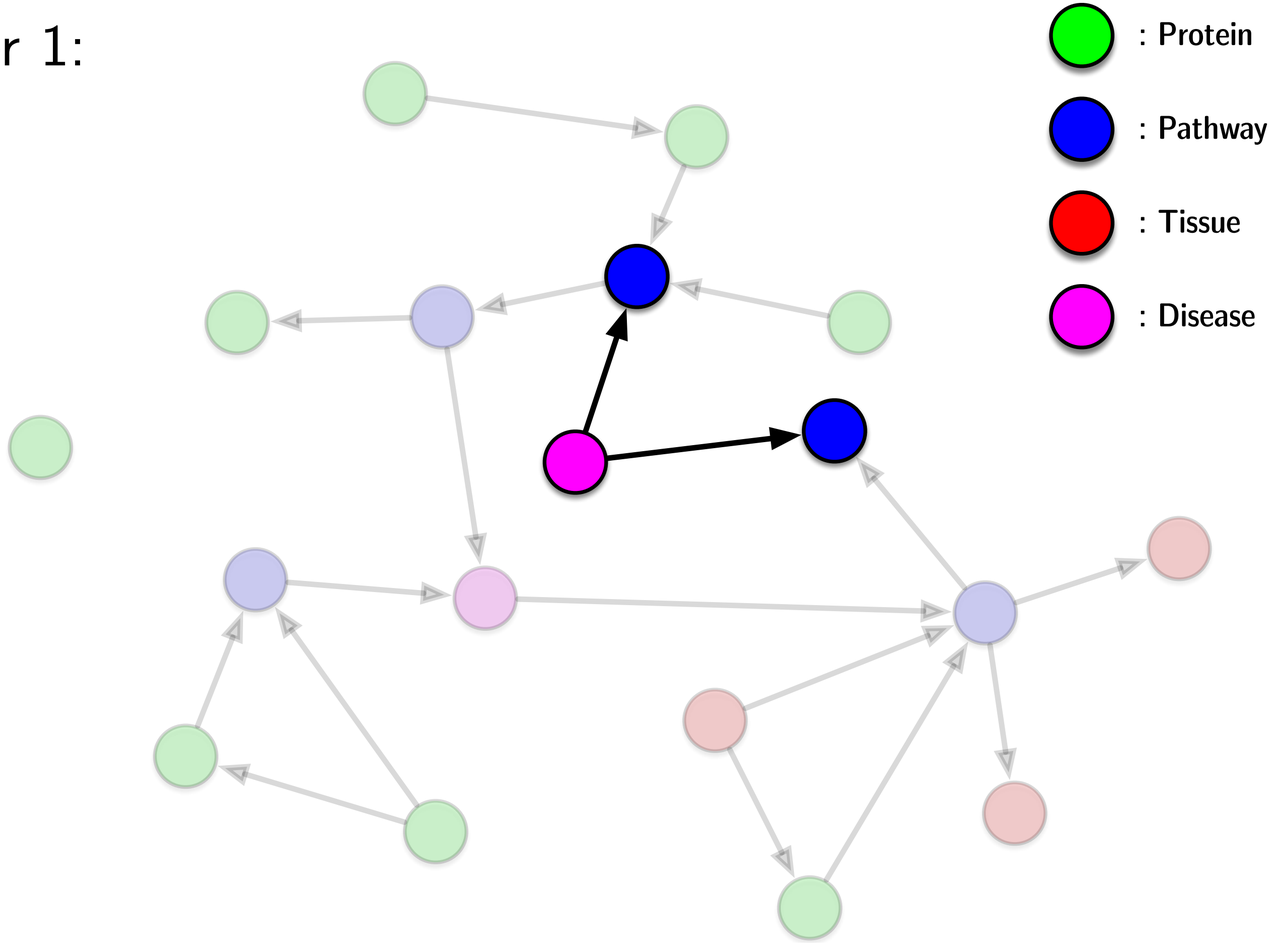
Graph Neural Networks

- Input:



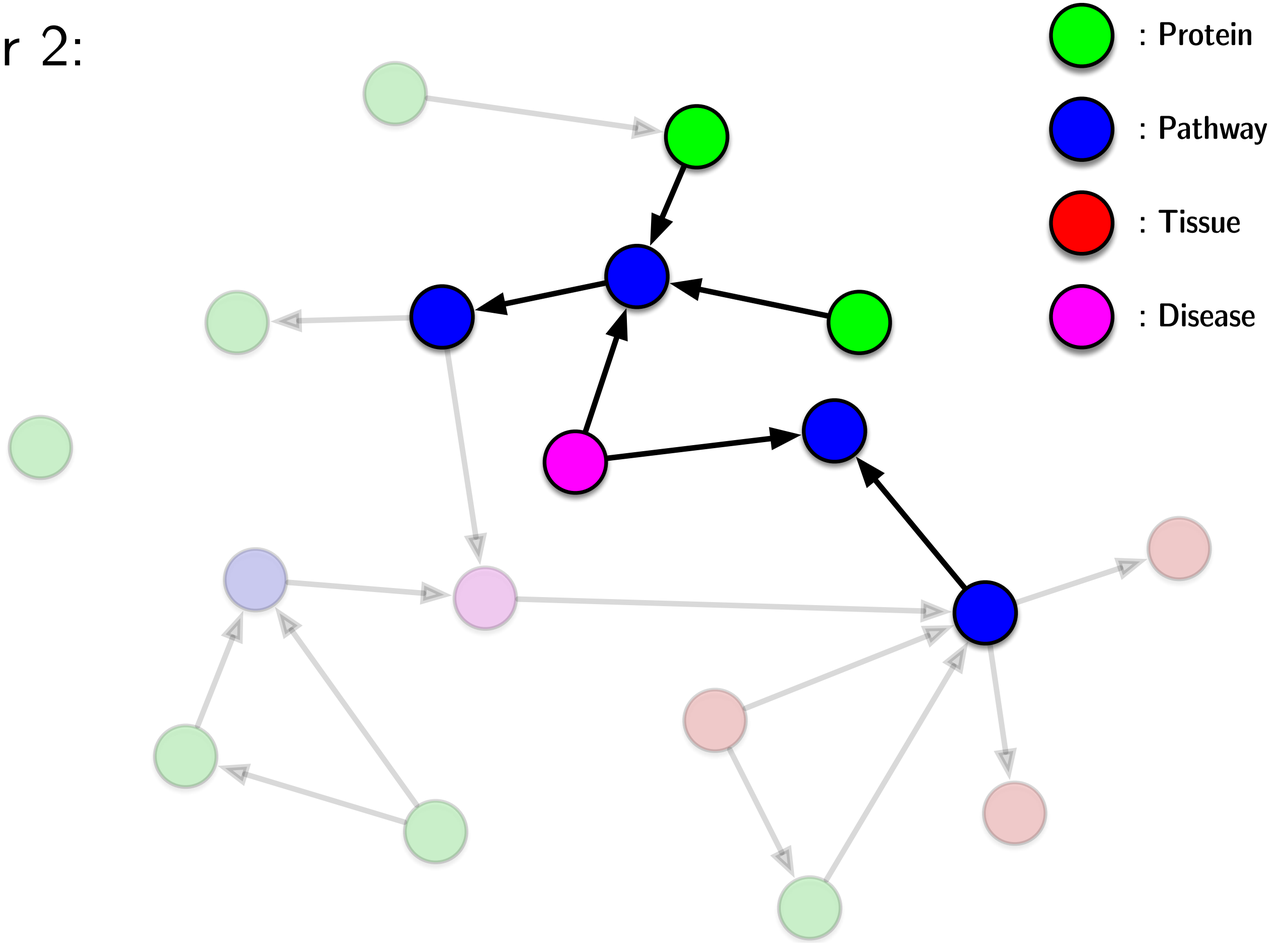
Graph Neural Networks

- Layer 1:



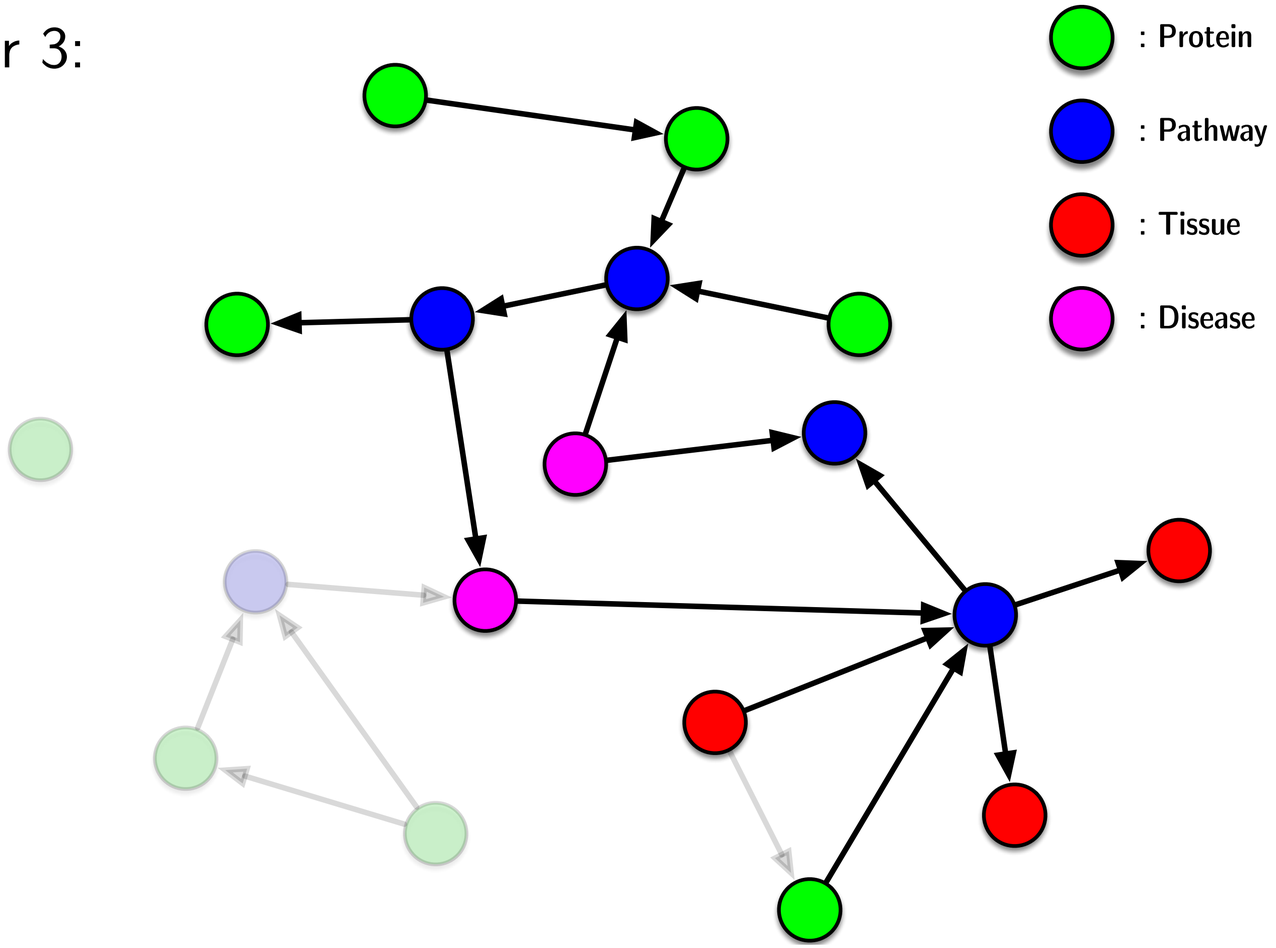
Graph Neural Networks

- Layer 2:

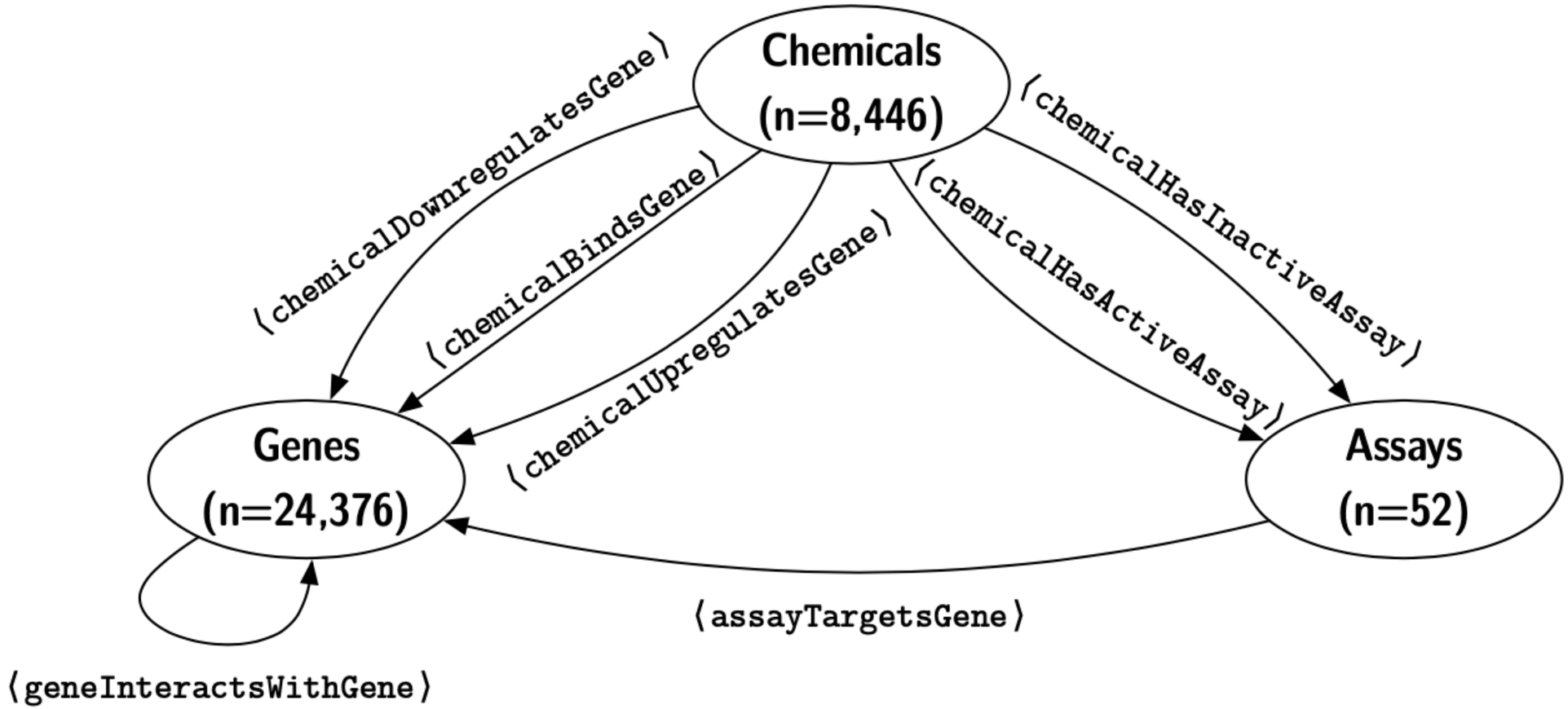


Graph Neural Networks

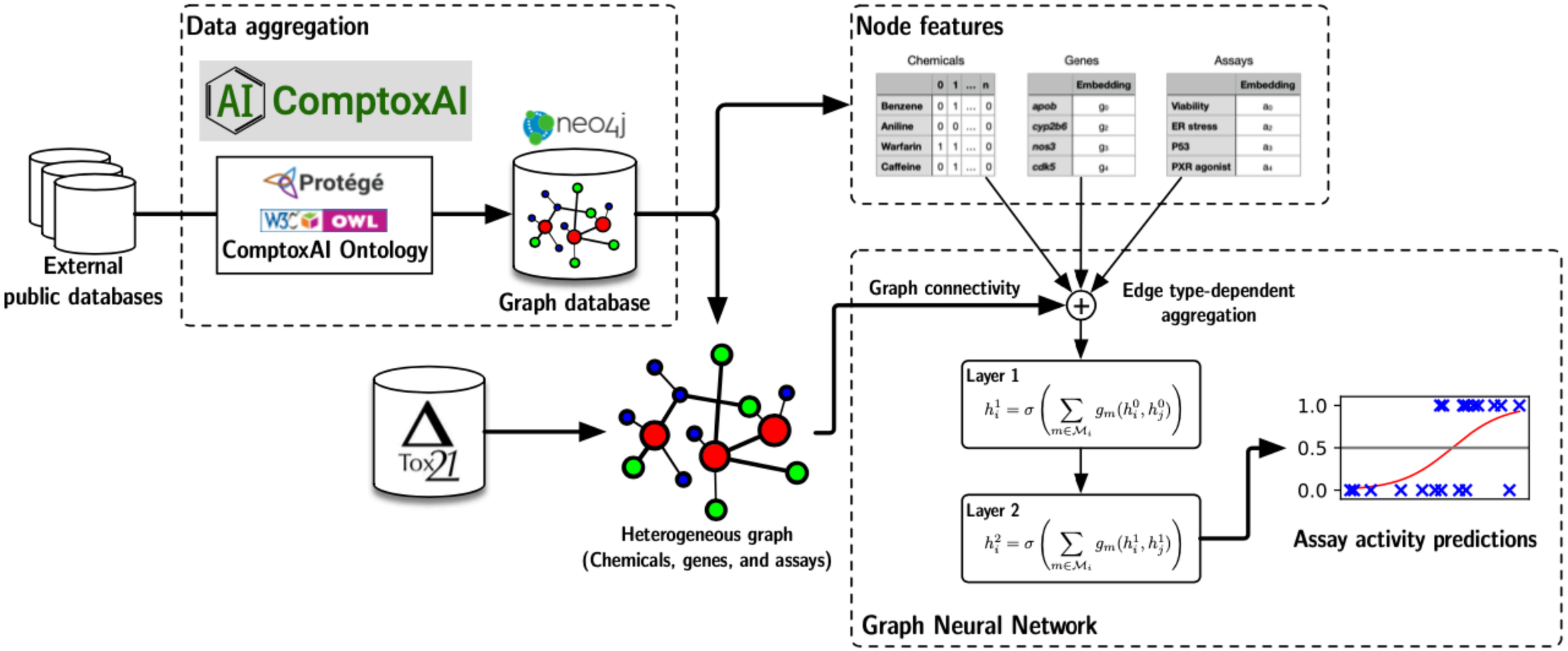
- Layer 3:

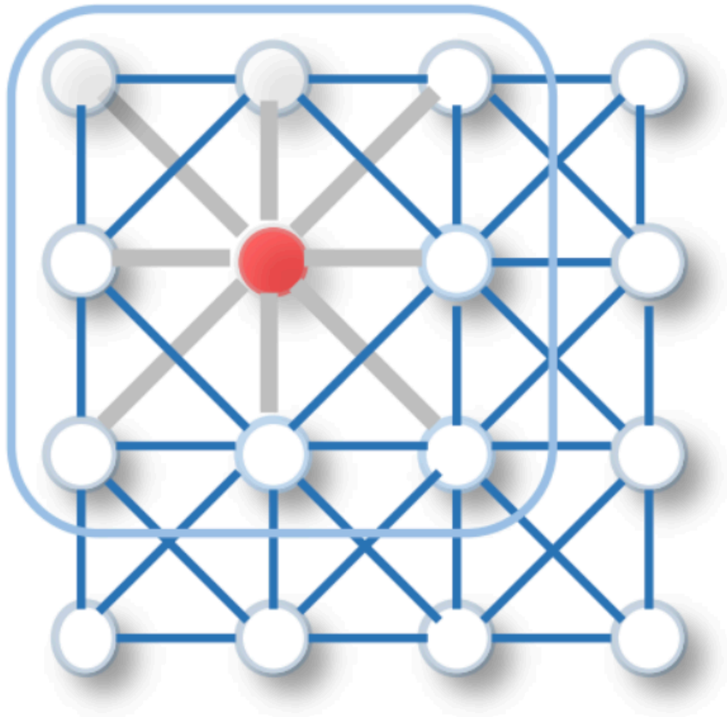


QSAR Subgraph

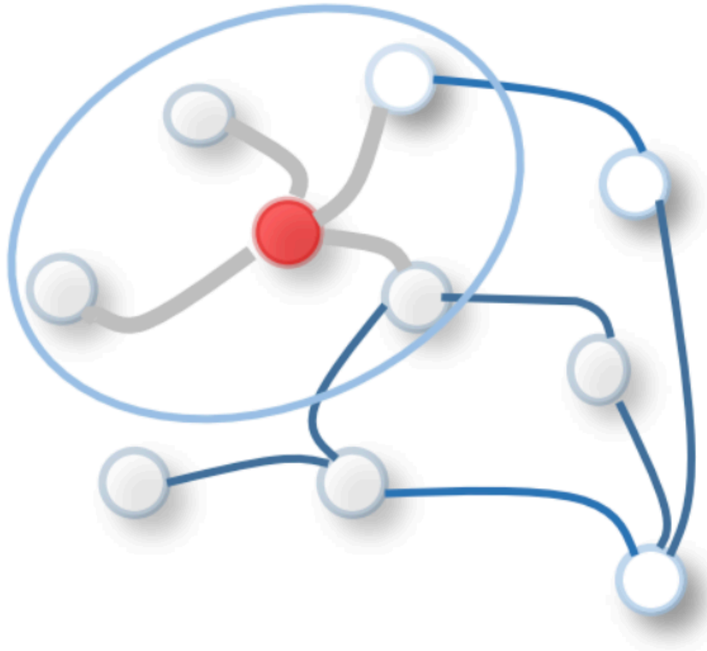


Predicting chemical toxicity with graph neural networks



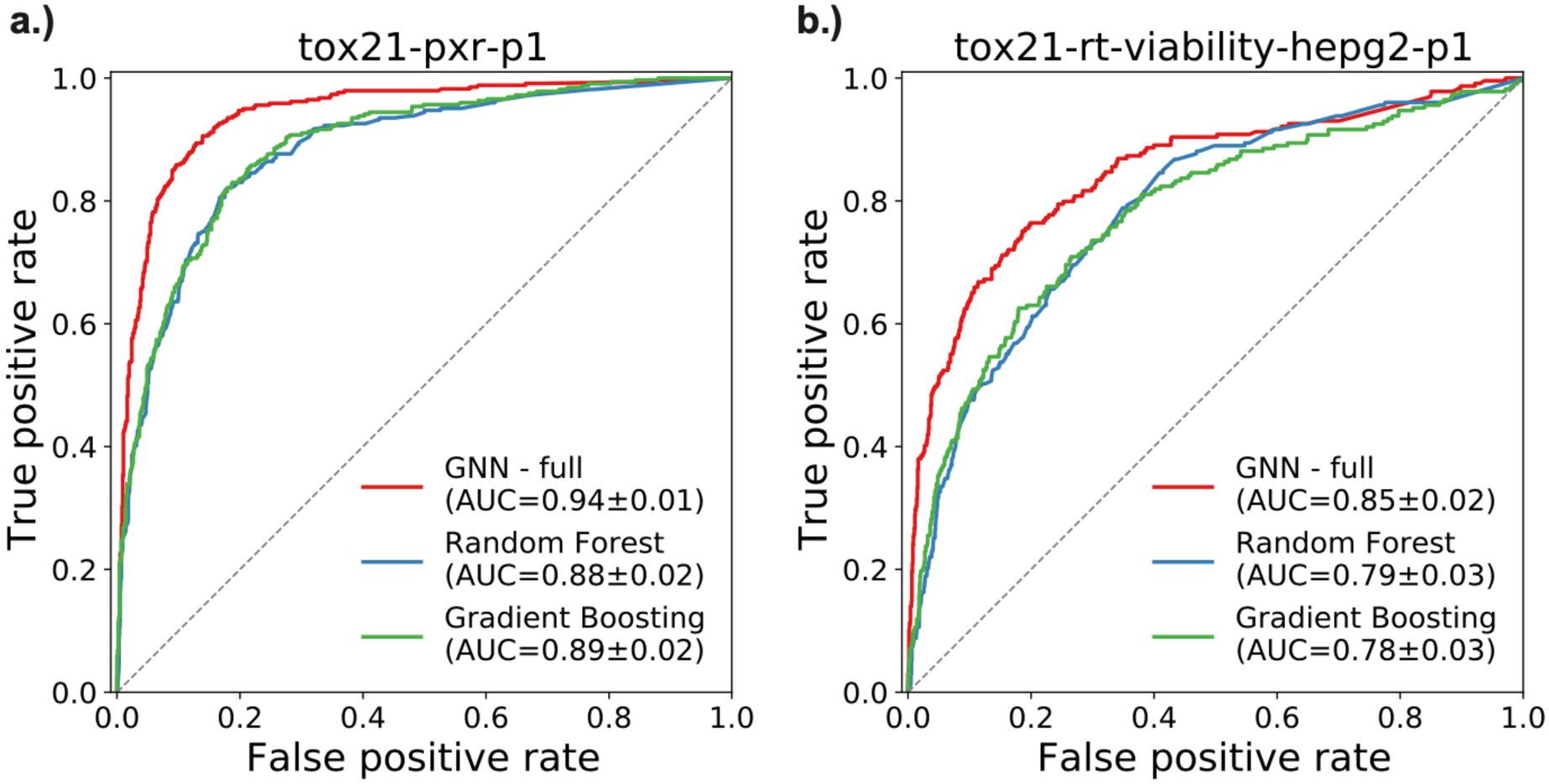


Convolutions on pixels (images)

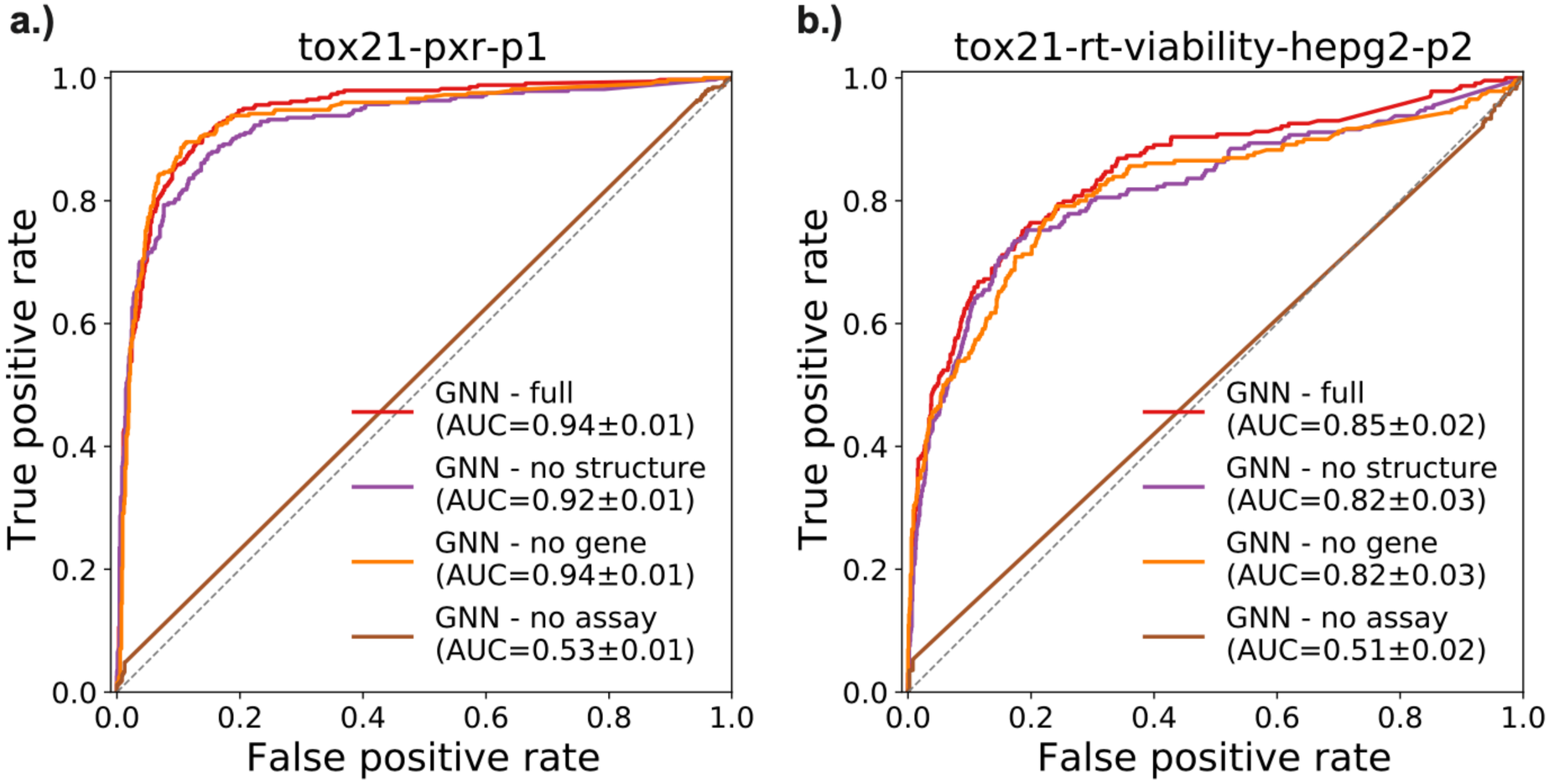


Convolutions on nodes (graphs)

Graph neural network performance

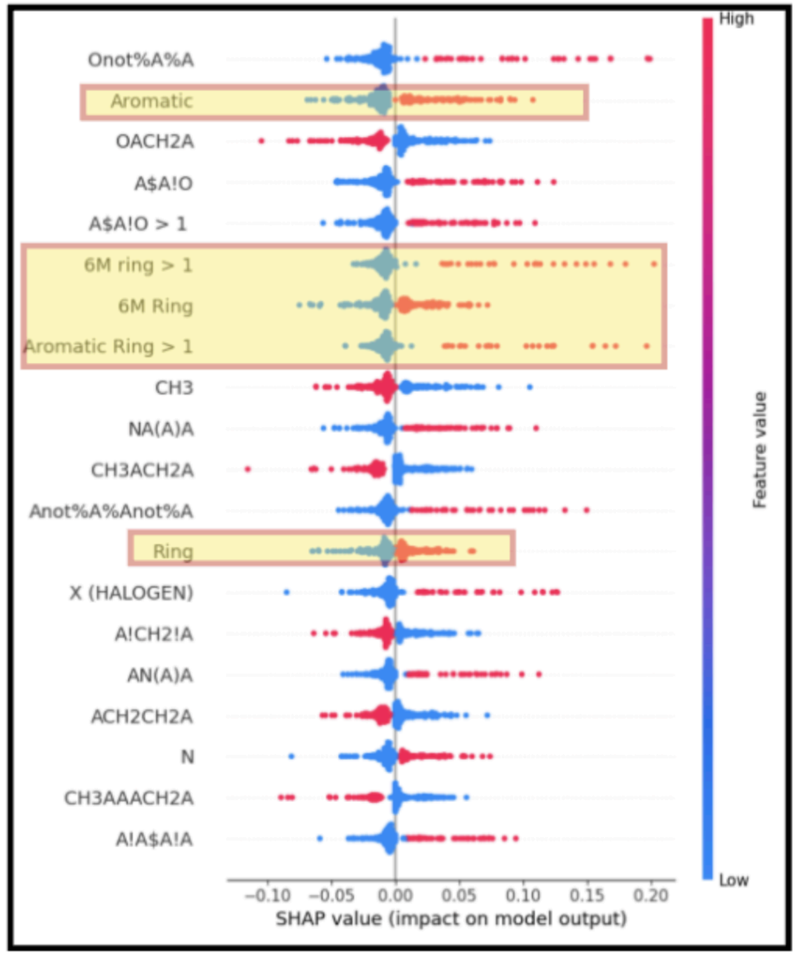


Why do the GNNs perform so much better?

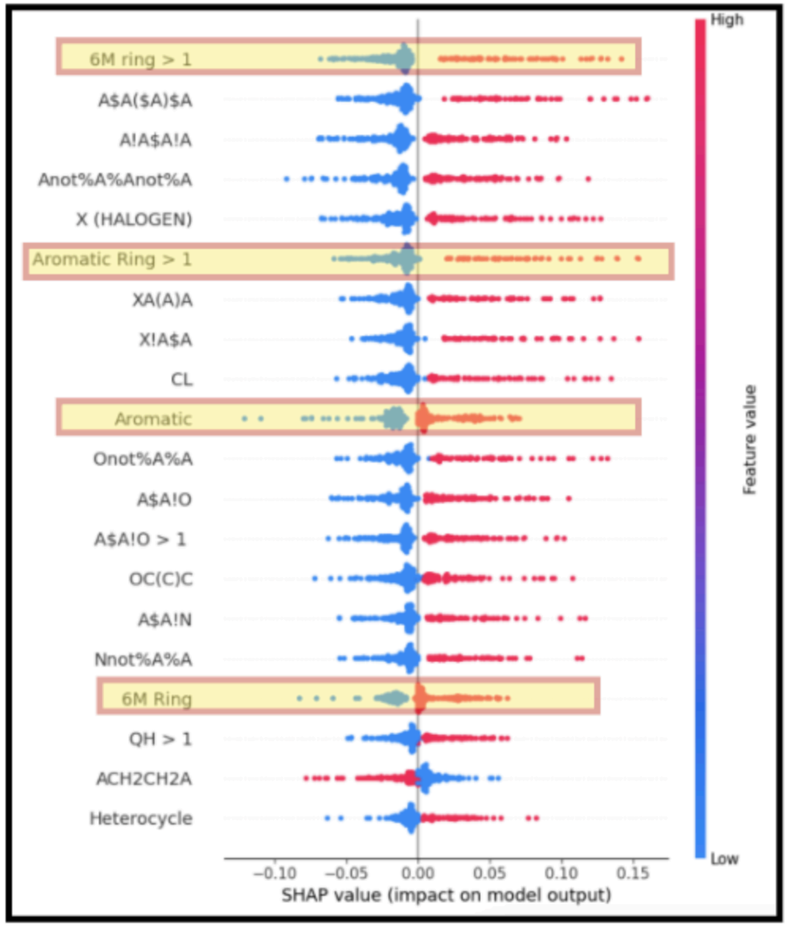


Interpretable Predictive Toxicology

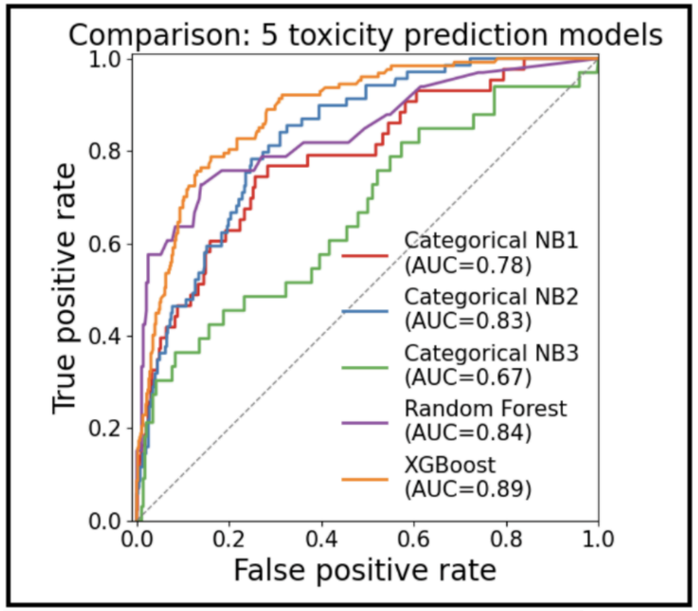
Food Contact Chemicals Database with Liver Assay



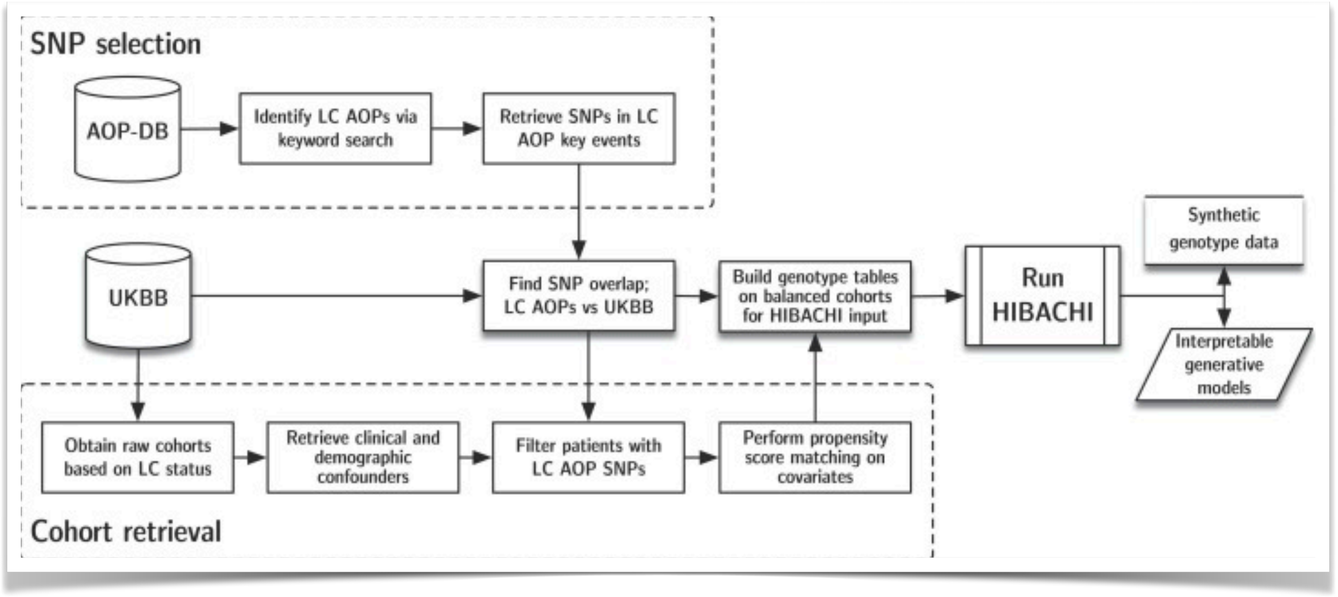
Hazardous Substances with Colon Cancer Assay



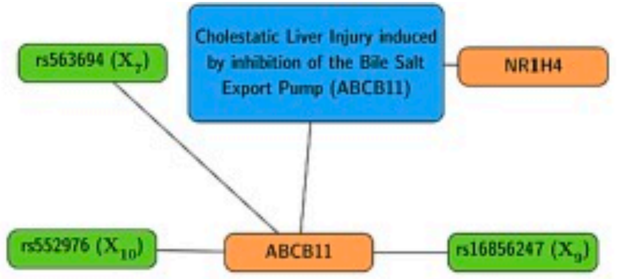
Estefania Morales
Summer Undergraduate Internship Program (SUIP) Fellow

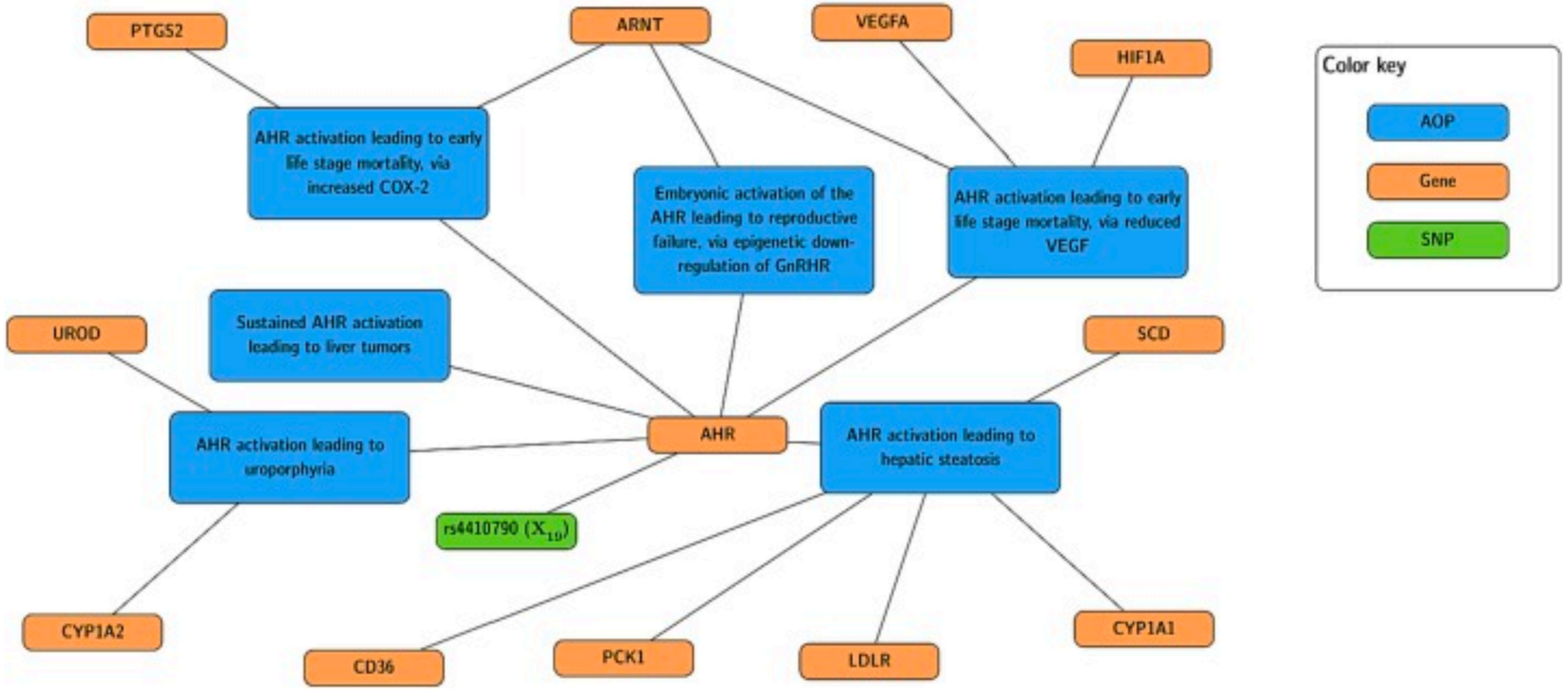


Observational data in exposomics research



Individual	Model	Fitness Score
1	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq ((\log_{X_{10}}(X_2) + X_{19}!) \bmod 2))))$	2.730
2	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq X_9!)))$	2.280
3	$((X_{10}\ \texttt{XOR}\ X_7)\ \texttt{AND}\ (X_{19} - (X_7 \neq X_{24})))$	2.244
4	$((X_{10}\ \texttt{XOR}\ X_{19})\ \texttt{AND}\ (X_{19} - X_7))$	2.204
5	$(X_{19} + (\neg X_6)) \bmod 2$	1.824
6	$X_9 \neq X_{19}$	1.254
7	$X_{10}!$	0.232



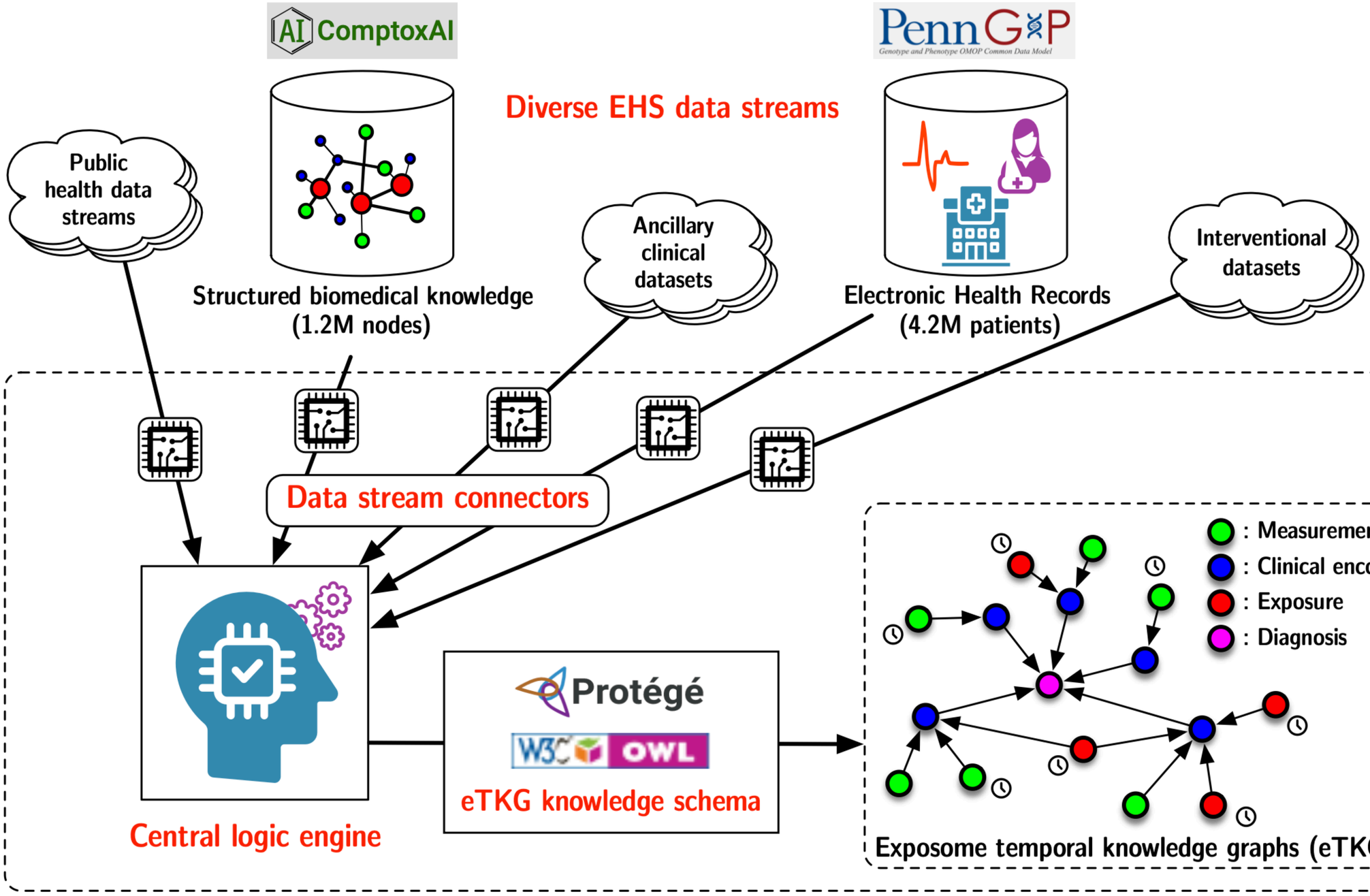


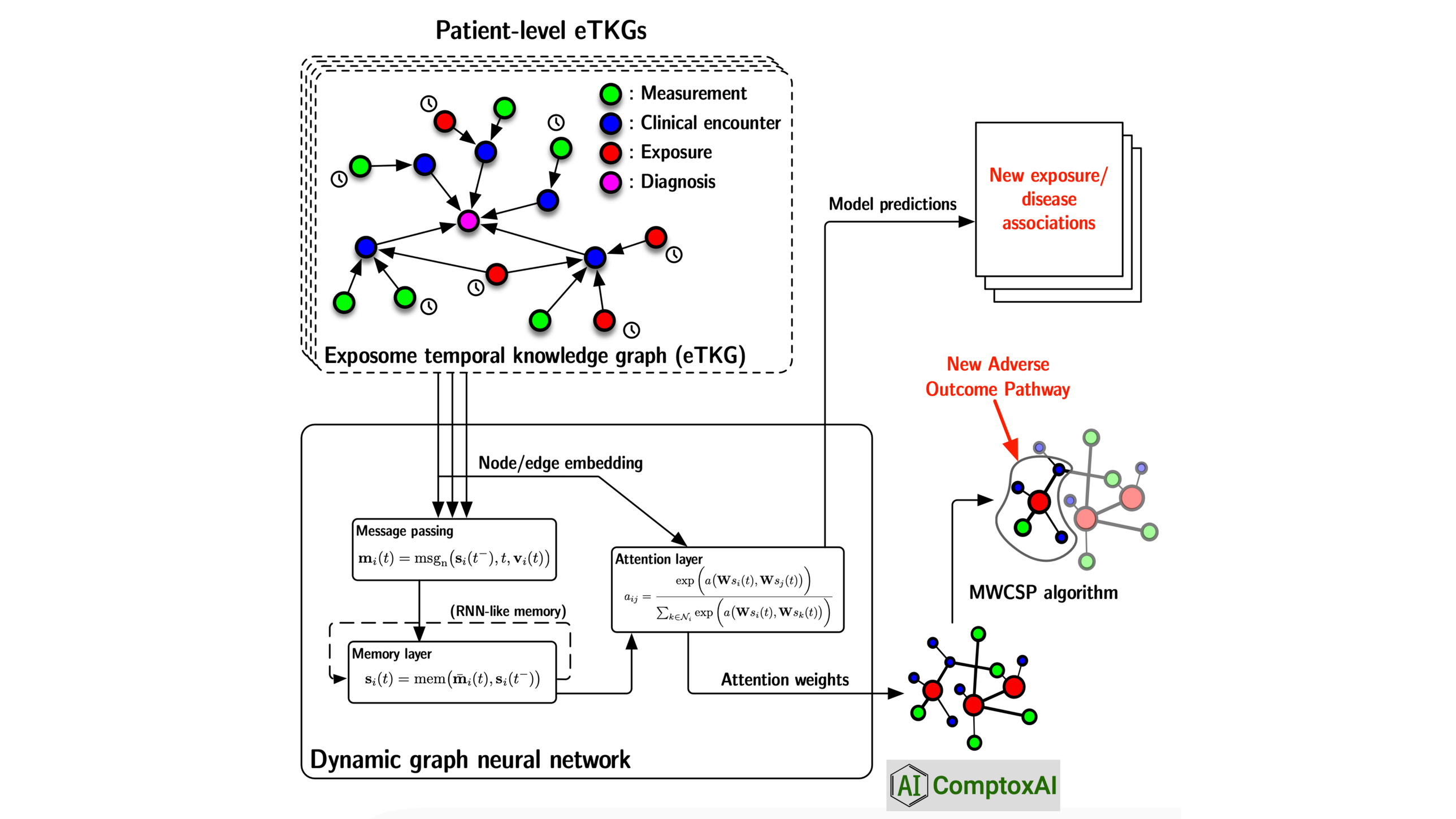
GxE interactions in PFAS-induced hepatocellular carcinoma

Chloé Paris

Gene Symbol	Function
TP53	Most common mutation found in HCC (plays important role in genomic stability)
DNAJB1	Fusion of DNAJB1 and PRKACA suggested as a possible marker for FLC
HMOX1	Suggested to play a role in HCC cell migration + its **metabolite bilirubin** has been associated with HCC

- Calculate the base **communicability score** between all PFAS-HCC node pairs in subgraph
  - Defined as a weighted sum of all possible “walks” linking a pair of nodes
- For each gene in the network:
  - Remove the gene from the network
  - Recalculate communicability scores for all PFAS-HCC pairs
  - Determine the change in communicability score
- Identify genes yielding the greatest mean change in communicability


AI ComptoxAI
PennGoP
Genotype and Phenotype OMOP Common Data Model
Diverse EHS data streams
Public health data streams
Structured biomedical knowledge (1.2M nodes)
Ancillary clinical datasets
Electronic Health Records (4.2M patients)
Interventional datasets
Data stream connectors
Central logic engine
Protégé
W3C OWL
eTKG knowledge schema
: Measurement
: Clinical encounter
: Exposure
: Diagnosis
Exposome temporal knowledge graphs (eTKGs)


Patient-level eTKGs
: Measurement
: Clinical encounter
: Exposure
: Diagnosis
Exposome temporal knowledge graph (eTKG)
Model predictions
New exposure/ disease associations
New Adverse Outcome Pathway
MWCSP algorithm
Node/edge embedding
Message passing
$\mathbf{m}_i(t) = \mathrm{msg}_n\left(\mathbf{s}_i(t^-), t, \mathbf{v}_i(t)\right)$
(RNN-like memory)
Memory layer
$\mathbf{s}_i(t) = \mathrm{mem}\left(\bar{\mathbf{m}}_i(t), \mathbf{s}_i(t^-)\right)$
Attention layer
$a_{ij} = \frac{\exp\left(a\left(\mathbf{W}s_i(t), \mathbf{W}s_j(t)\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(a\left(\mathbf{W}s_i(t), \mathbf{W}s_k(t)\right)\right)}$
Attention weights
Dynamic graph neural network
AI ComptoxAI

Acknowledgements

- The Romano Lab:
  - Chloé Paris
  - Estefania Morales
  - Kevin Shen
  - Tom Pan
- Jesse Goodrich, PhD (USC)
- Danielle Mowery, PhD
- Trevor Penning, PhD
- Li Shen, PhD
- Jason Moore, PhD (Cedars-Sinai)
- Yun Hao, PhD (Simons foundation)
- Grant support:
  - `K99/R00-LM013646` (PI: Romano)
  - Colton Informatics Fellowship
  - `P30-ES013508` (PI: Penning)

DEPARTMENT *of* BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

Center *of* Excellence *in* Environmental Toxicology

Penn Institute for Biomedical Informatics

See more at: romanolab.org