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

Perelman School of Medicine, University of Pennsylvania
Penn Institute for Biomedical Informatics

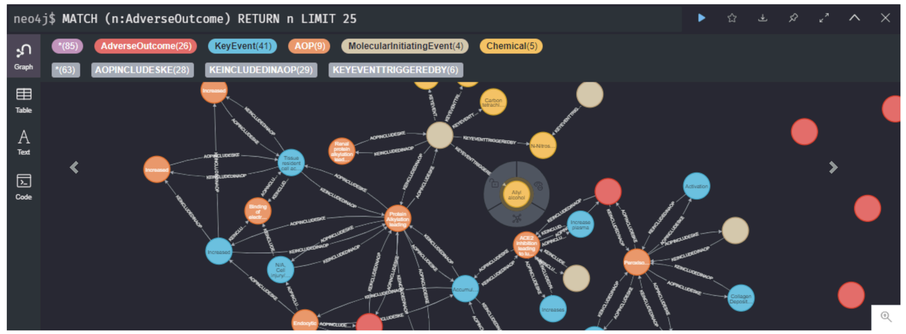
Joseph D. Romano*, Yun Hao*, Jason H. Moore
Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA
**These authors contributed equally*

Source code: https://github.com/EpistasisLab/qsar-gnn
ComptoxAI: https://comptox.ai

Paper: https://doi.org/10.1142/9789811250477_0018

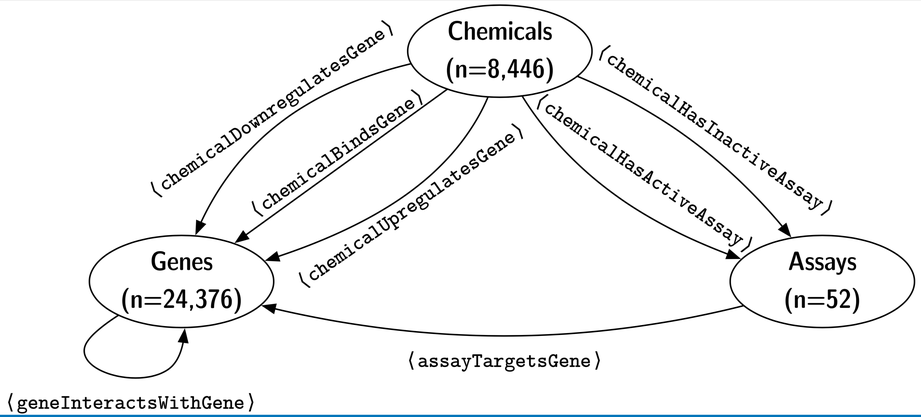
- Quantitative Structure-Activity Relationship (QSAR) modeling is the most prevalent method for *in silico* toxicity prediction.
- The disappointing performance and low interpretability of existing QSAR models call for new methodological innovation in the field.
- We introduce a GNN-based approach that aggregate data from ComptoxAI, and evaluate it on data from 52 Tox21 toxicity assays to show that it significantly outperforms existing methods.

ComptoxAI is a new graph database containing diverse entity and relationship types that pertain to translational mechanisms of toxicity

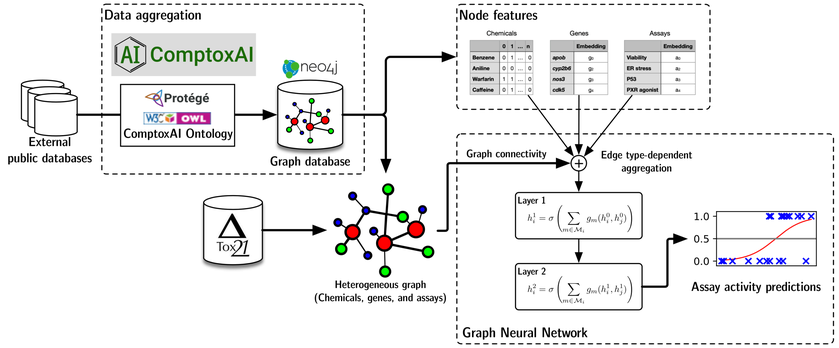


- Example view of ComptoxAI graph database

Subgraph scheme for assay outcome prediction

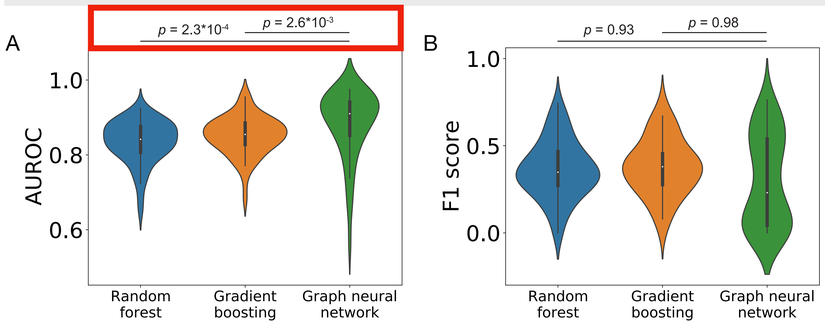


Overview of the graph machine learning approach



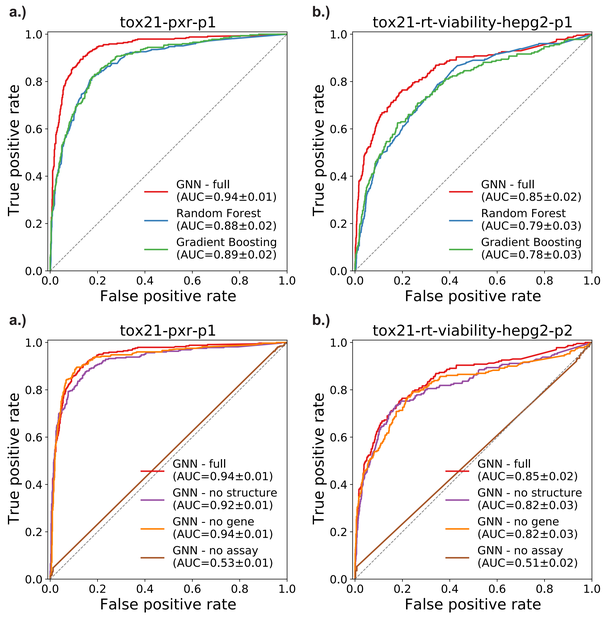
- Graph convolutional network architecture
  - Node representation
    - Chemical nodes: 166 bits MACCS fingerprint
    - Assay and gene nodes: single-valued feature optimized during model training
  - GCN layer
    - 2 hidden layers connected by leaky ReLU, softmax applied to output of the 2nd
    - Each layer is defined as an edge-wise aggregation of adjacent nodes
  - Optimization: minimizing binary cross-entropy loss with Adam optimizer

GNN model significantly outperforms baseline QSAR models



- Dataset
  - 52 assays and their accompanying chemical screening data from Tox21
  - 80%/20% train/test split on the label chemicals

GNN achieves better performance with the added context of network relationships between chemicals, assays, and genes



- **Our GNN models are highly interpretable**
  - Highest weighted assay for HepG2 viability prediction: Caspase 3/7 and Shh antagonist (both induce apoptosis)
- **Our GNN approach is robust to sources of bias**
  - The graph incorporate biological knowledge that can fill in gaps left by incomplete or inaccurate data

Funding: K99-LM013646 (PI: Joseph Romano), R01-LM010098, R01-LM012601, R01-AI116794, UL1-TR001878, UC4-DK112217 (PI: Jason Moore), T32-ES019851, and P30-ES013508 (PI: Trevor Penning).