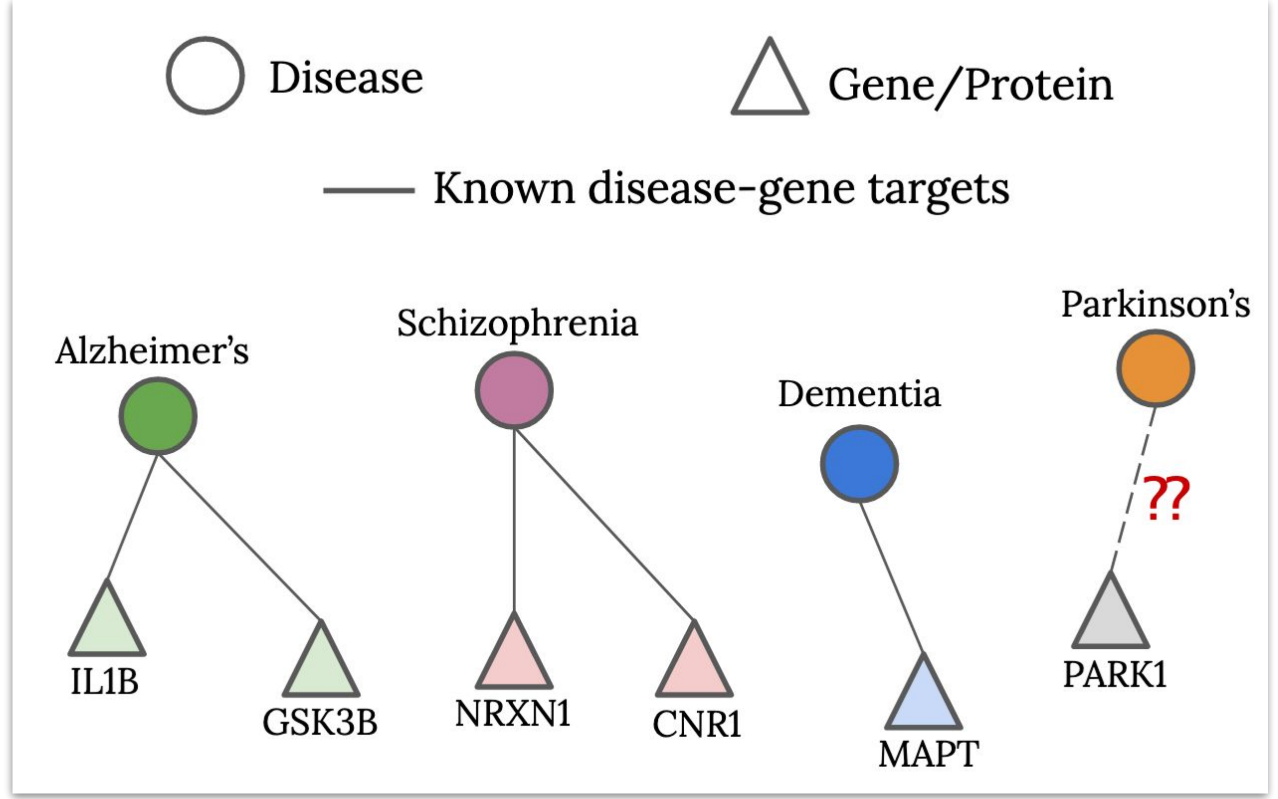
Relation-weighted Link Prediction for Disease Gene Identification

Srivamshi Pittala* William Koehler Jonathan Deans
Daniel Salinas Martin Bringmann Katharina Sophia Volz Berk Kapicioglu*

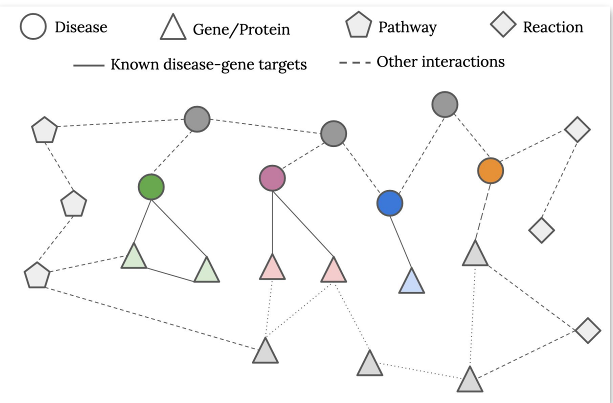
Disease Gene Identification

- Given a set of known targets for diseases, we aim to identify novel target genes for known and new diseases



Our Framework

- We formulate the problem as a link prediction task, where the goal is to predict new links between disease and gene nodes of a knowledge graph

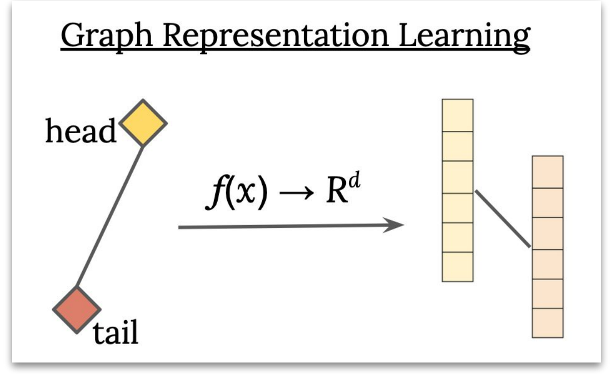


Build a heterogeneous knowledge graph by merging graphs that characterize other biological interactions involving disease and gene nodes.

Challenges:
- Harmonizing information across datasets
- Processing data to remove redundancies

Learn abstract representations of diseases and genes that capture their interactions with each other and other biological entities.

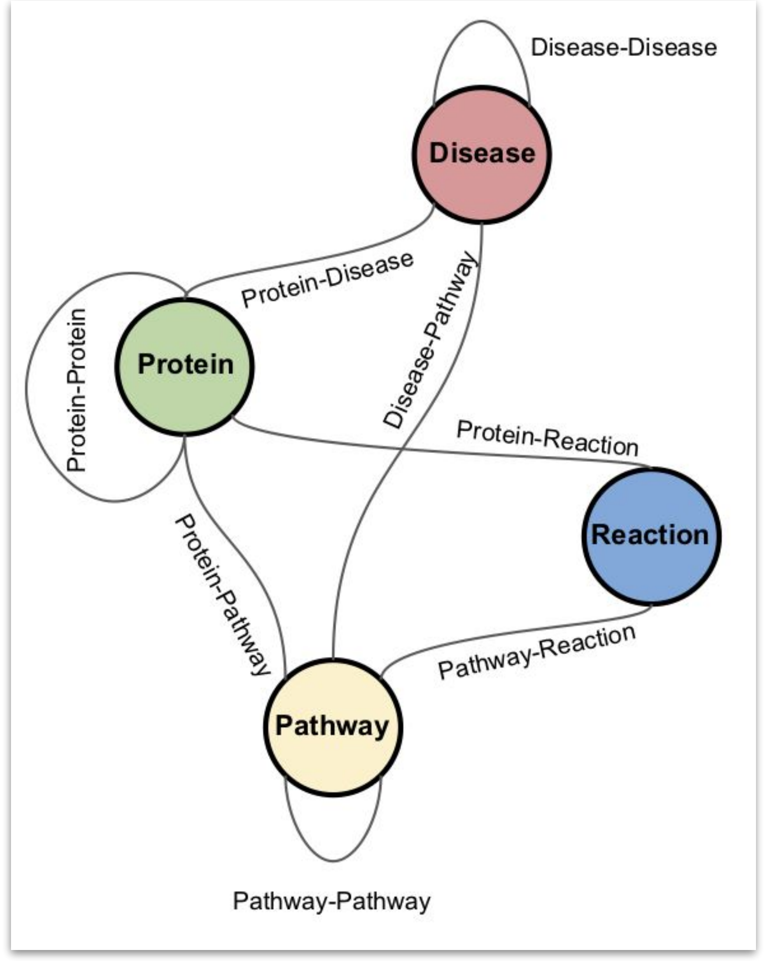
Challenges:
- Skewed distributions of nodes and edge types could bias the learning



Predict novel disease-gene links and validate experimentally

Our Knowledge Graph

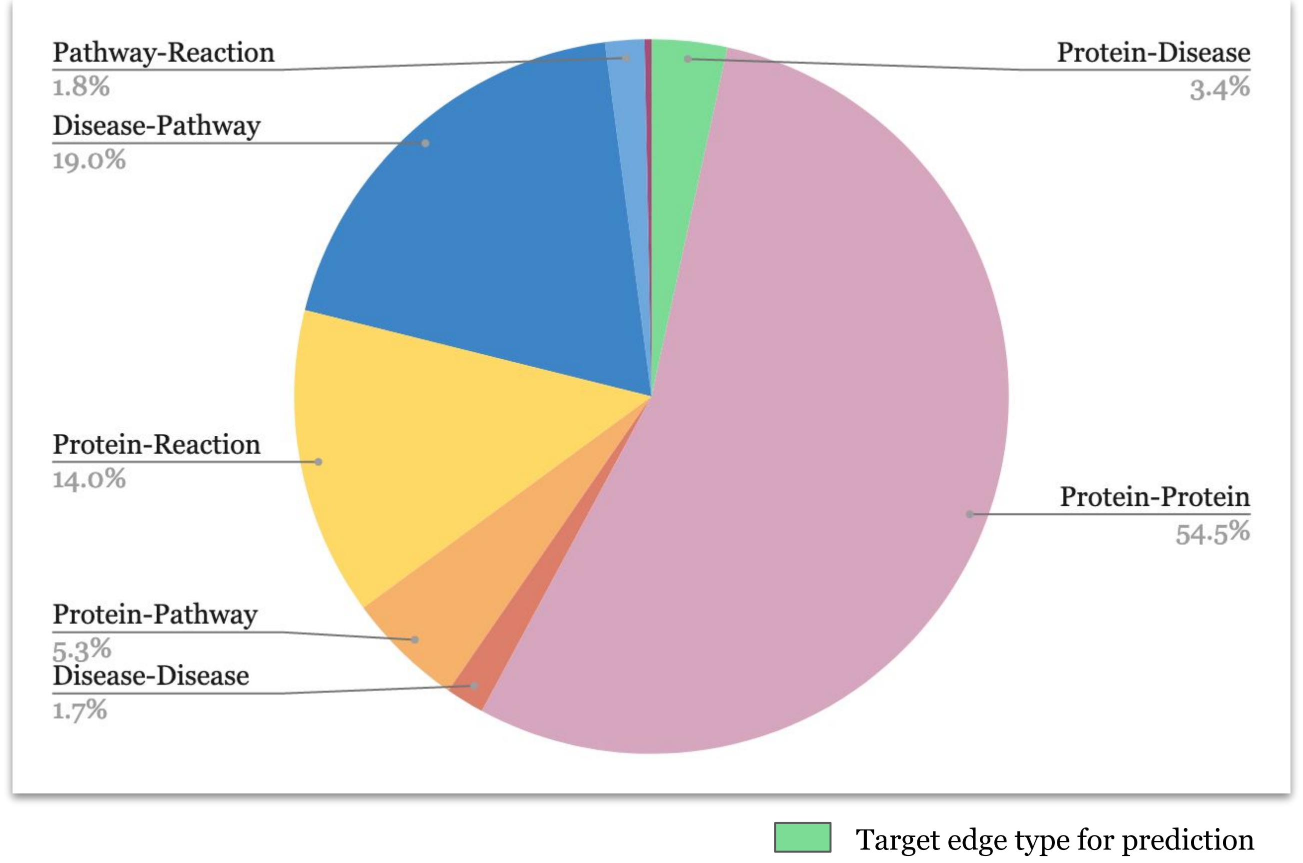
- Transformed a disease-gene graph, DisGeNET into DoidGeNET, a non-redundant version
- Processed and combined 4 different biomedical databases, resulting in a heterogeneous knowledge graph consisting of 8 edge types and 4 node types



Meta-representation of the knowledge graph

- The ***protein-disease*** edge type formed only 3.41% of the total edges in the graph

Edge Type Distribution



Relation-weighted Link Prediction

- We propose a learnable relation-specific weight to adjust for the skewed distributions
- We demonstrate our proposal on RotatE, a state-of-the-art method for link prediction
- In the max-margin objective function, the relation-specific distance is scaled by a weight w_r
- The loss function can be written as the following:

$$L = -\log \sigma\left(\gamma - w_r * d_r\left(\mathbf{h}, \mathbf{t}\right)\right) - \sum_{i=1}^{n} p\left(h_i^{'}, r, t_i^{'}\right) \log \sigma\left(w_r * d_r\left(\mathbf{h}_i^{'}, \mathbf{t}_i^{'}\right) - \gamma\right)$$

d_r Relation-specific distance function w_r Relation-specific weight
σ Sigmoid function γ Margin h Head r Relation t Tail

- Training is carried out by optimizing for w_r along with rest of the hyper-parameters

Experimental Results

- It helps to augment the graph with layers representing different biological processes

Variant	hit@30	Mean Rank	Mean Percentile
DG	0.189	4995.65	72.77
DG + ST	0.287	2029.74	88.94
DG + ST + DG_uc	0.353	1467.84	91.64
DG + ST + DG_uc + DO	0.363	1256.69	92.84
DG + ST + DG_uc + DO + RT	0.375	1186.81	93.32

DG: DoidGeNET; ST: STRING; DG_uc: DG uncurated; DO: Disease Ontology; RT: Reactome

- It helps to weigh the edge types in a heterogeneous graph

Variant	hit@30	Mean Rank	Mean Percentile
Original	0.368	1298.44	92.70
Our Method	0.375	1186.81	93.32

- Our method outperforms other state-of-the-art methods

Method	hit@30	Mean Rank	Mean Percentile
Random Walk	0.007	4597.91	72.78
Direct Neighborhood scoring	0.250	3339.61	80.24
DIAMOnD	0.336	NA	NA
Our Method	0.375	1186.81	93.32

- Our method retrospectively identifies more targets in trials than Open Targets

Measured overlap between top 50 predictions and Trialtrove

	Parkinson's	Crohn's	Schizophrenia
Open Targets	9	10	7
Our Method	14	22	10

Future Work

- Experimental validation of novel disease-gene predictions
- Augment our knowledge graph with additional layers