Large-Scale Analysis of Disease Pathways in the Human Interactome

Marinka Zitnik

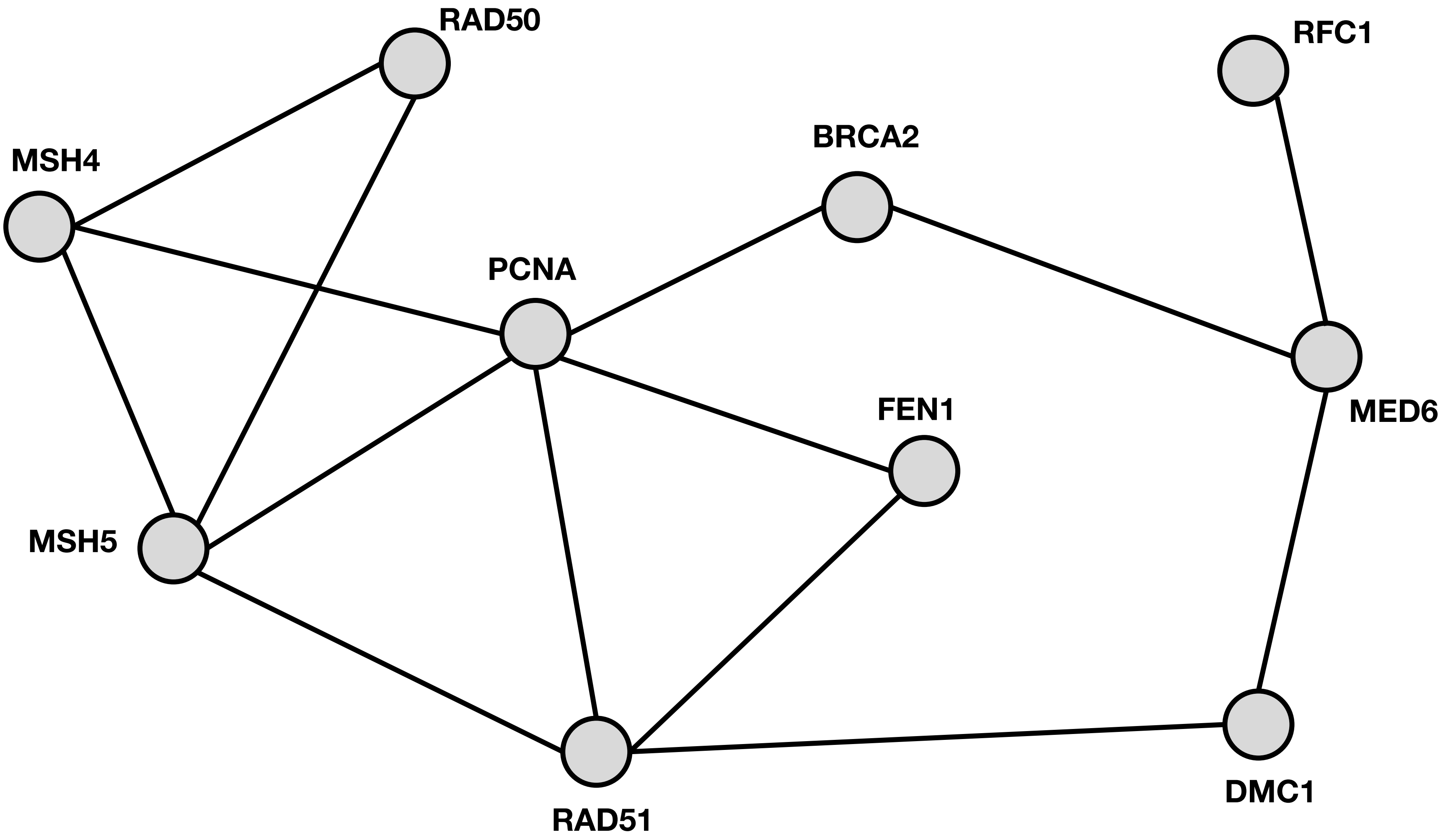
Joint work with Monica Agrawal and Jure Leskovec

LELAND STANFORD JUNIOR UNIVERSITY
ORGANIZED 1891

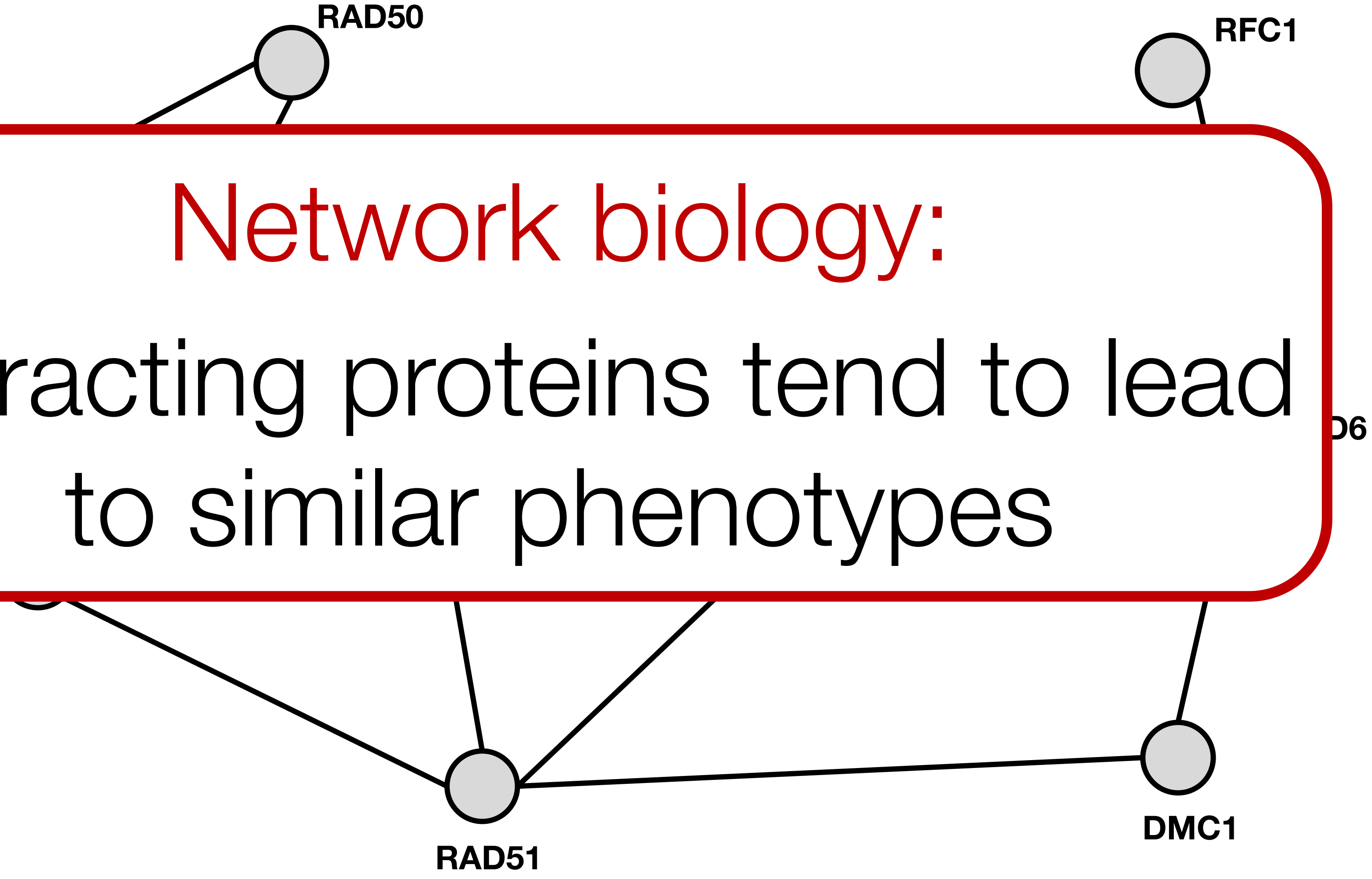
STANFORD INFOLAB

CHAN ZUCKERBERG BIOHUB

Human Interactome



Human Interactome



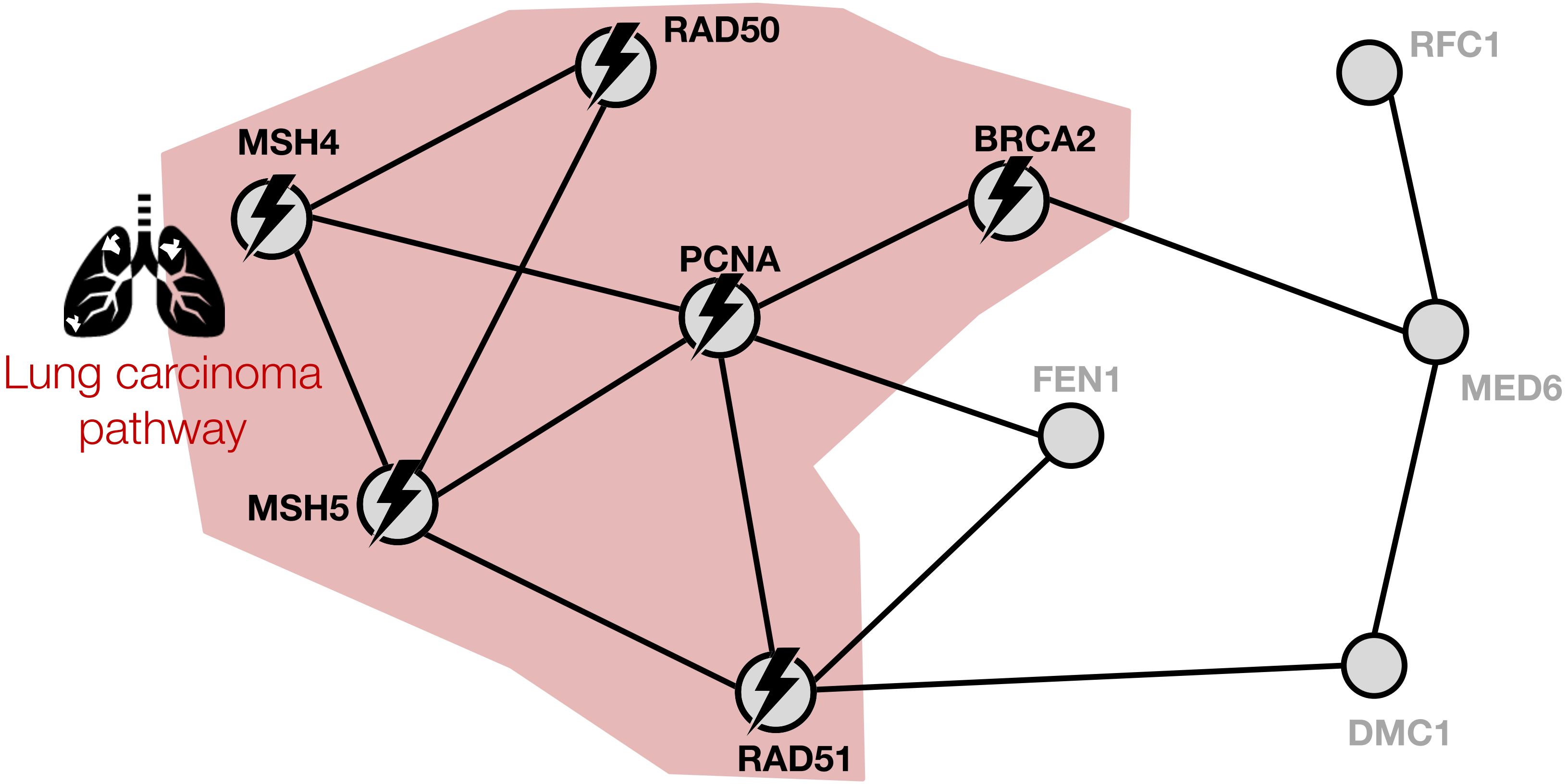
Network biology:

Interacting proteins tend to lead to similar phenotypes

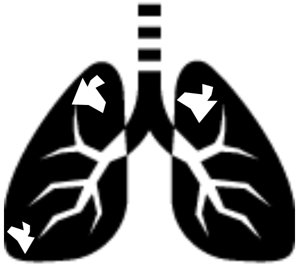
[Menche et al., Science 2015, Costanzo et al., Science 2016]

Disease Pathways

- Pathway: Subnetwork of interacting proteins associated with a disease



This Work:
Research Question



What is the protein interaction network structure of disease pathways?

Disease Pathway Dataset

- Protein-protein interaction (PPI) network culled from 15 knowledge databases:
  - 350k physical interactions, e.g., metabolic enzyme-coupled interactions, signaling interactions, protein complexes
  - All protein-coding human genes (21k)
- Protein-disease associations:
  - 21k associations split among 519 Mendelian and complex diseases
- Disease categories, e.g., cancers (68), nervous system diseases (44), cardiovascular diseases (33), immune system diseases (21)

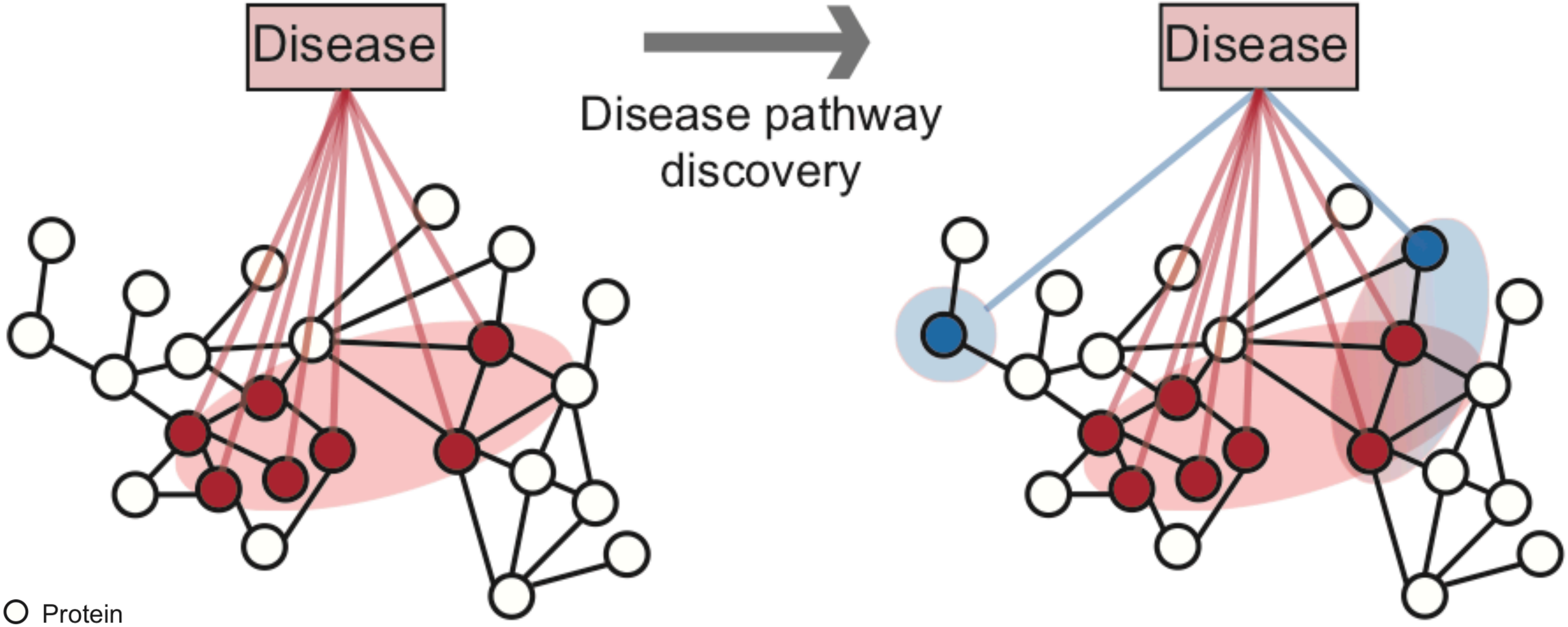
- Pros: Experimentally validated data, comprehensive analysis

Prediction Task

● Known (seed) disease protein

● Predicted disease protein

— Predicted protein-disease association



Disease

Disease pathway discovery

Disease

○ Protein

● Disease protein

— Protein-protein interaction

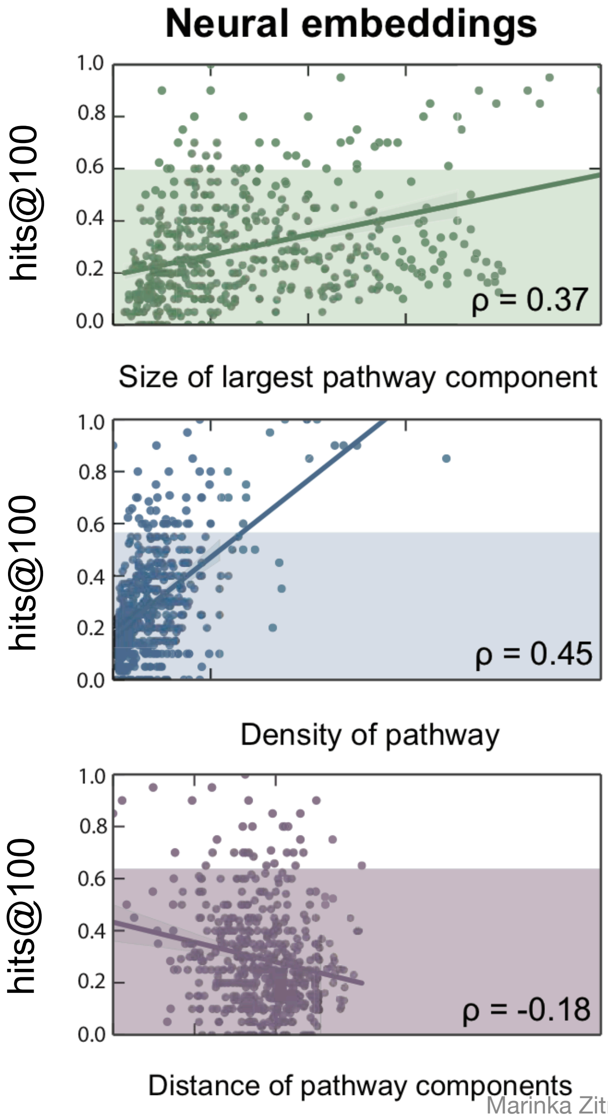
— Protein-disease association

⬬ Pathway component

Methods and Setup

- 5 methods: neural embeddings, matrix completion, neighbor scoring, diffusion, connectivity significance
  - Get a score for each node: probability that protein is associated with a disease
- For each disease:
  - Train the method using training proteins
  - Predict disease proteins in test test

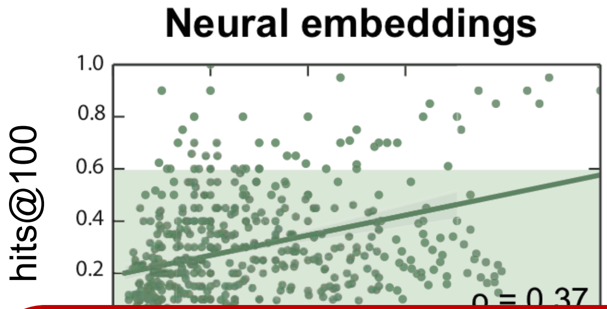
Prediction Results



- Best performers:
  - Random walks $\text{hits@100} = 0.36$
  - Neural embeddings $\text{hits@100} = 0.30$
- Worst performer:
  - Neighbor scoring $\text{hits@100} = 0.24$

Full results for all methods in the paper.

Prediction Results



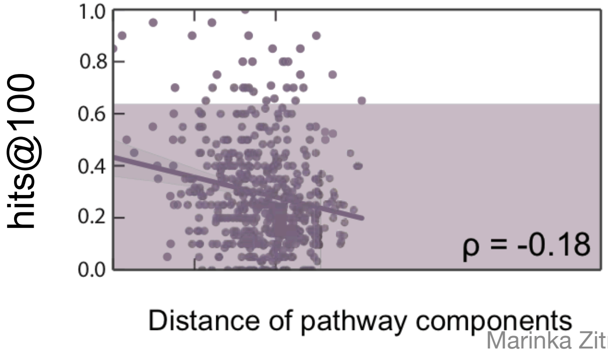
- Best performers:
  - Random walks
    hits@100 = 0.36

Limited success of current methods

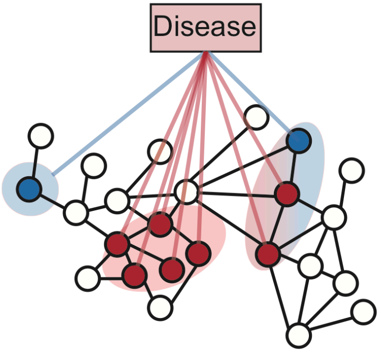
Failure cases not well understood

hi
0.2
0.0
ρ = 0.45
Density of pathway

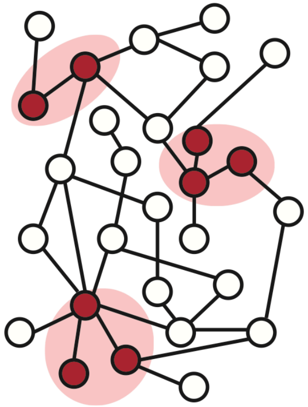
- Worst performer:
  - Neighbor scoring
    $\mathrm{hits@100} = 0.24$



Full results for all methods in the paper.



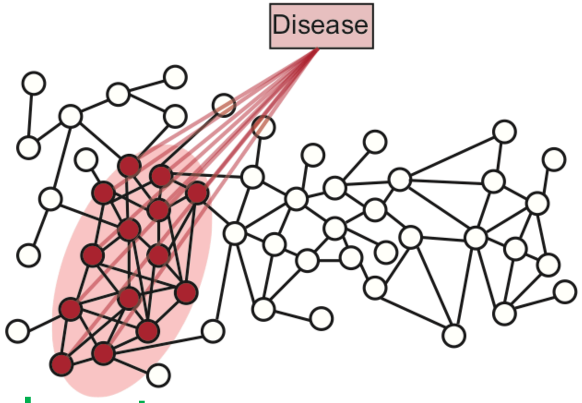
How can we explain failure cases of **disease pathway prediction**?



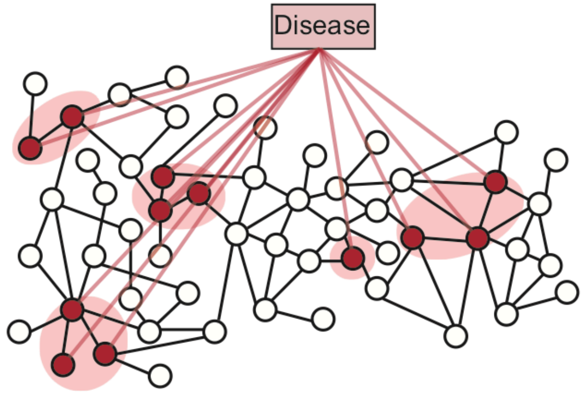
What is the **network structure** of disease pathways?

Competing Views

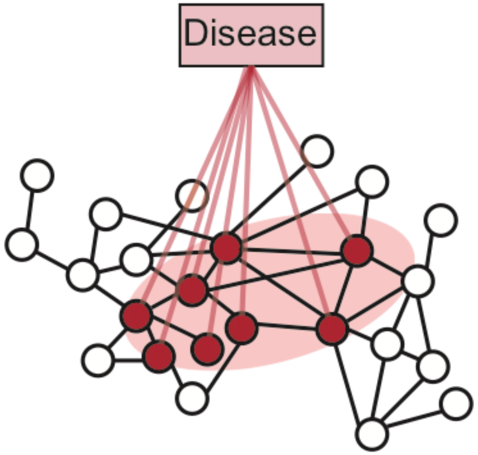
1. **Current:** Traditional network clusters
   - Well connected internally
   - Localized in the PPI net
   - Few edges pointing outside



2. **Our work:** Multi-regional objects
   - Loosely interlinked
   - Distributed in the PPI net
   - Many edges pointing outside
   - Higher-order connectivity

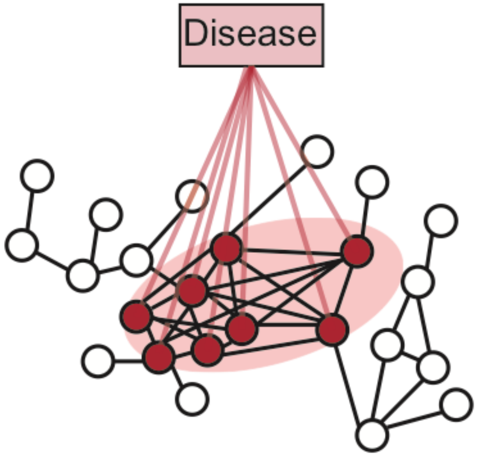


Are Pathways Well Interlinked?



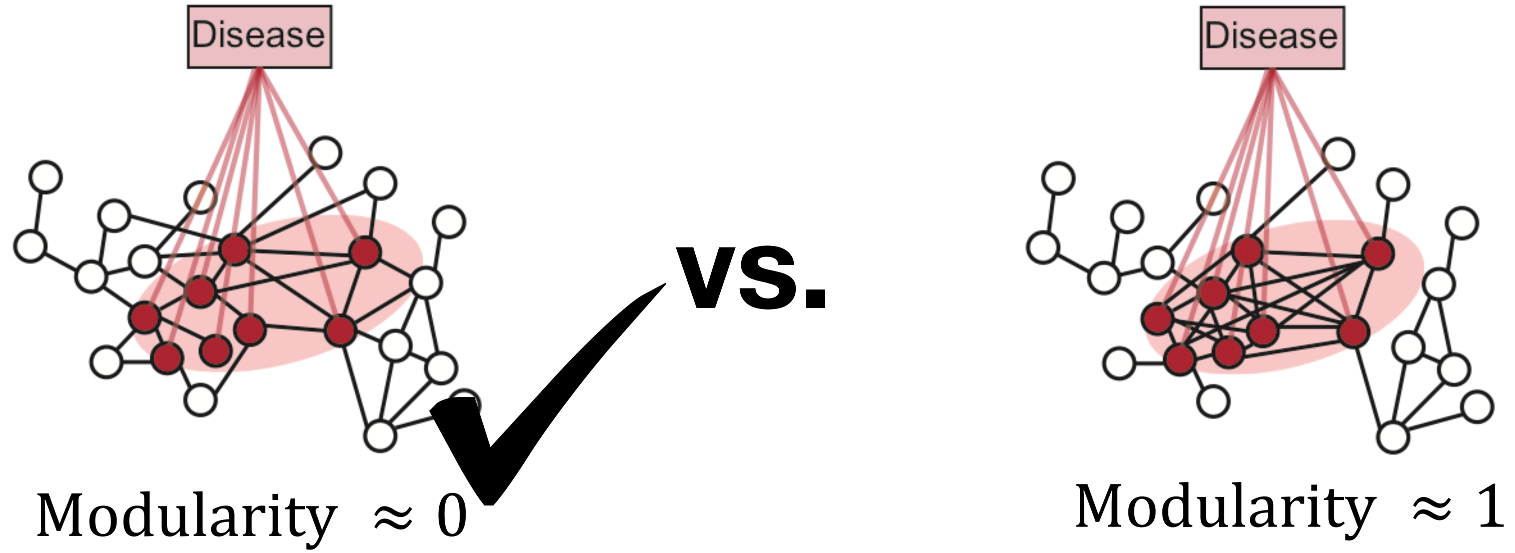
Modularity ≈ 0

vs.



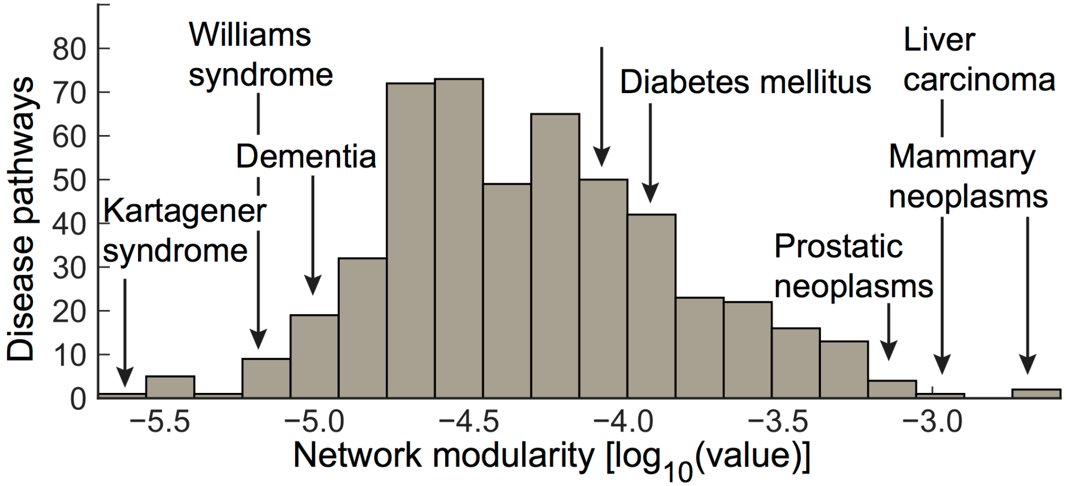
Modularity ≈ 1

Are Pathways Well Interlinked?



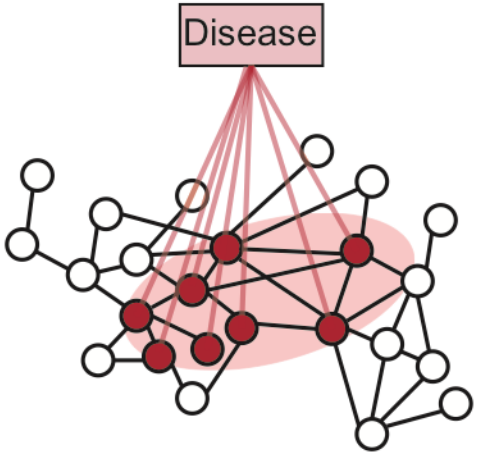
$\text{Modularity} \approx 0$

$\text{Modularity} \approx 1$



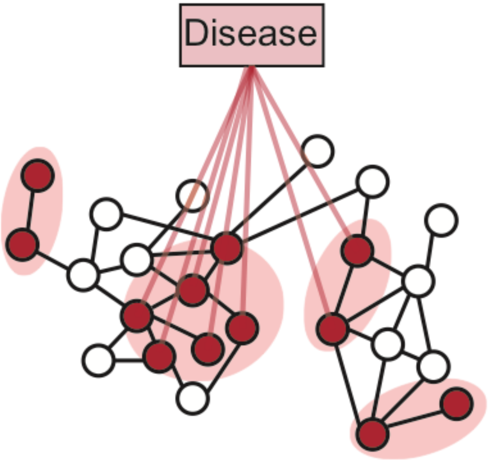
- No! - Pathways are embedded within PPI net
- Modularity: Interactions within the pathway minus the expected interactions

Are Pathways Connected?



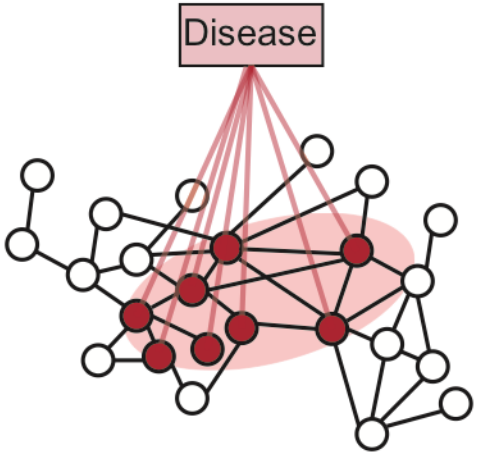
$\text{Pathway components} = 1$

vs.

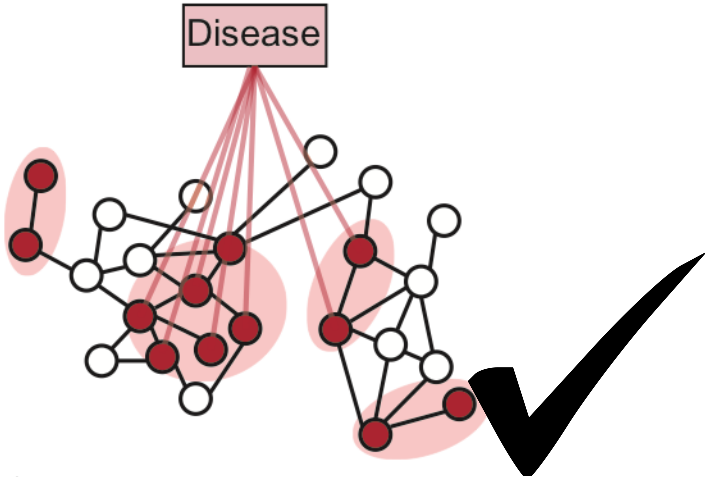


$\text{Pathway components} = 4$

Are Pathways Connected?

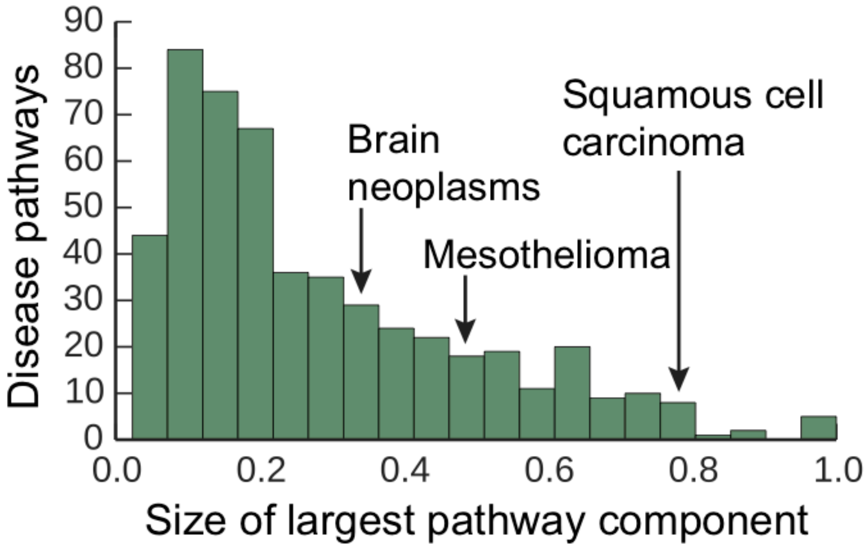


vs.



Pathway components = 1

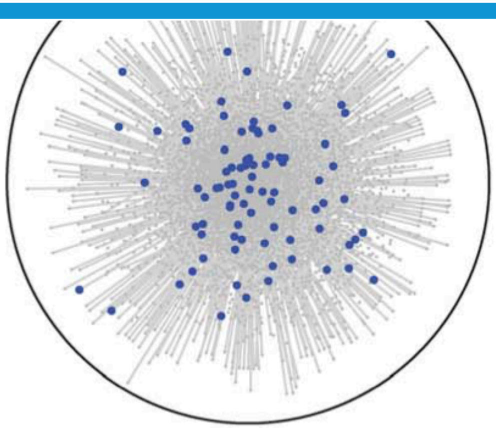
Pathway components = 4



No! - Pathways have fragmented PPI structure:

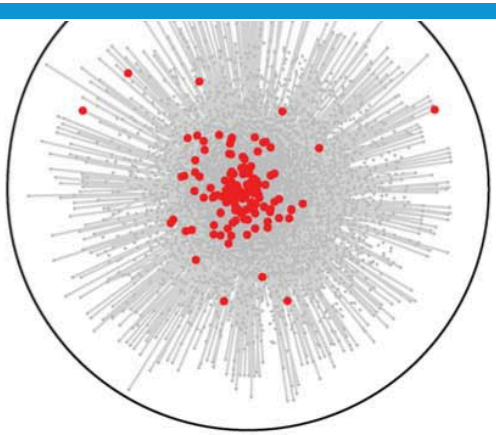
- 16 pathway components
- 10% of pathways have 60+% proteins in the largest component

Do Pathways Localize in Net?



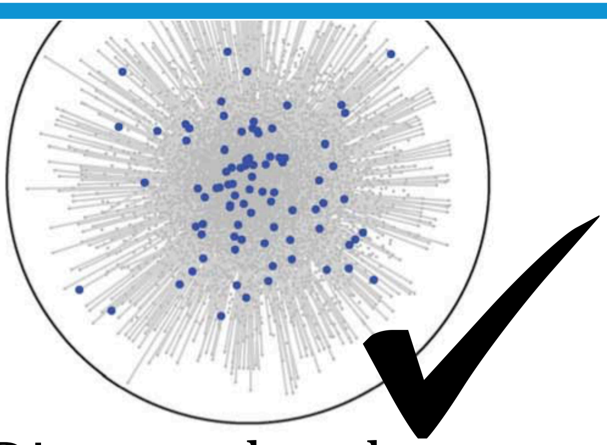
Dispersed pathway

vs.



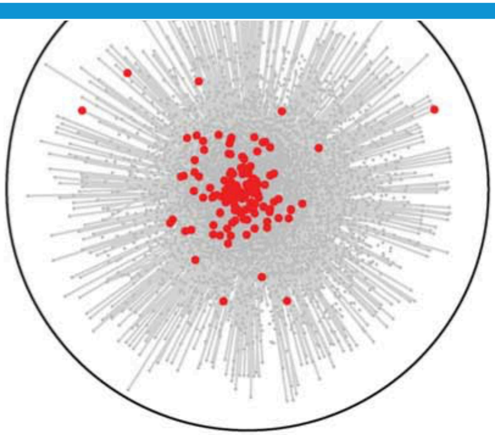
Localized pathway

Do Pathways Localize in Net?

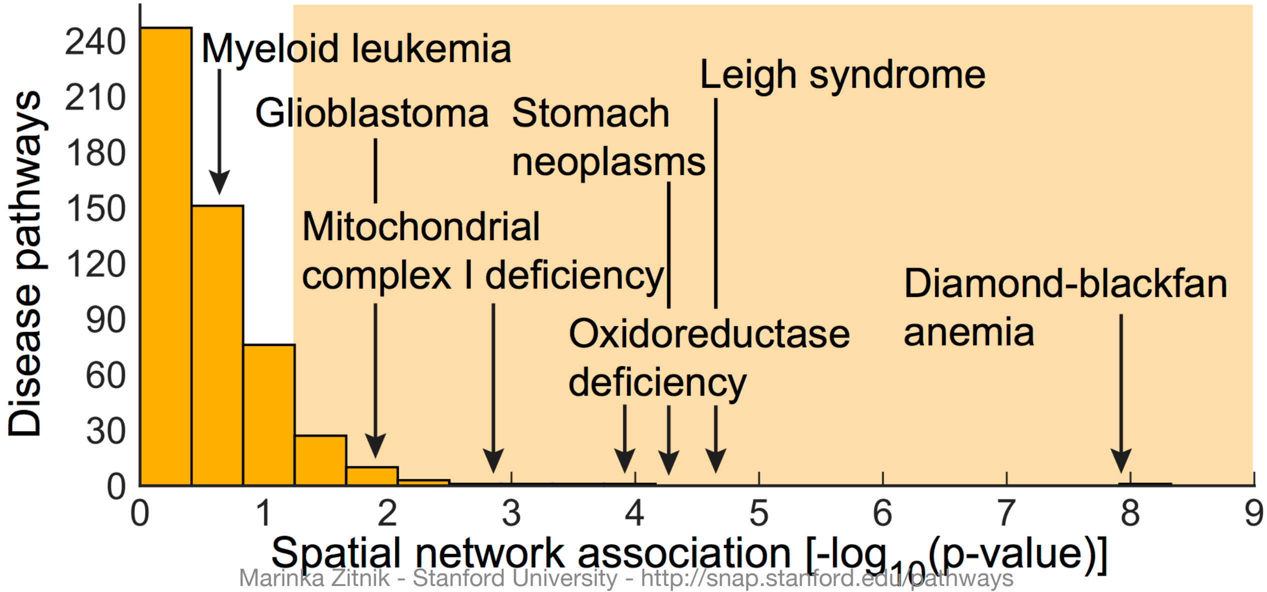


Dispersed pathway

vs.



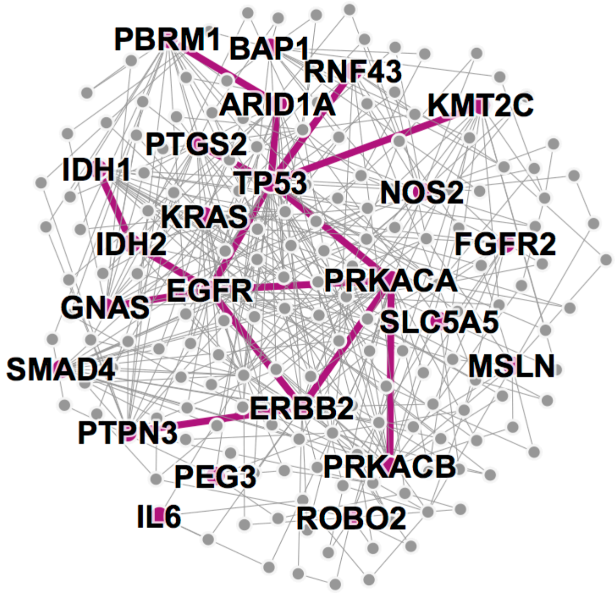
Localized pathway



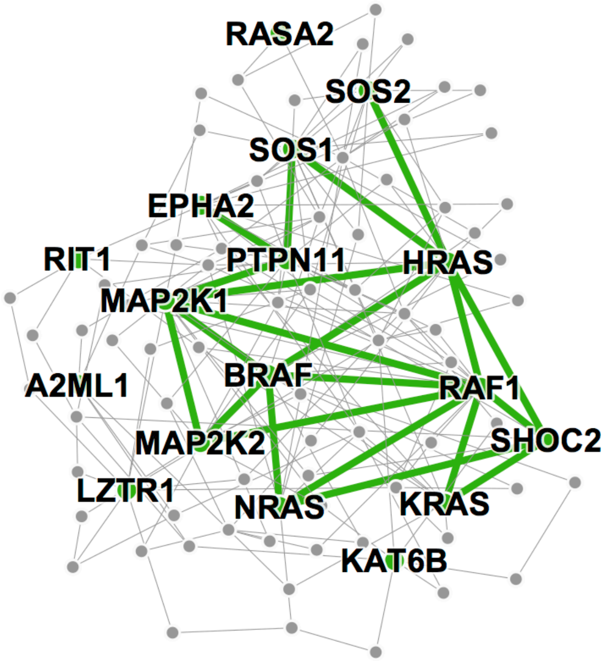
Do Pathways Localize in Net?

Disease pathways are weakly embedded in the PPI network, e.g.:

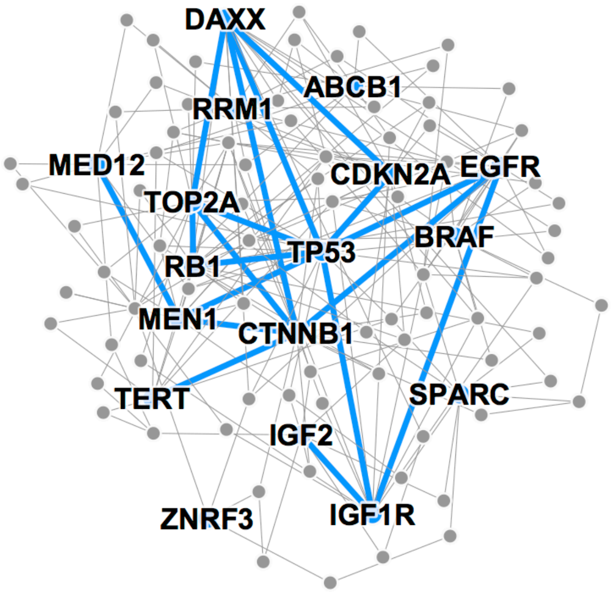
Cholangiocarcinoma



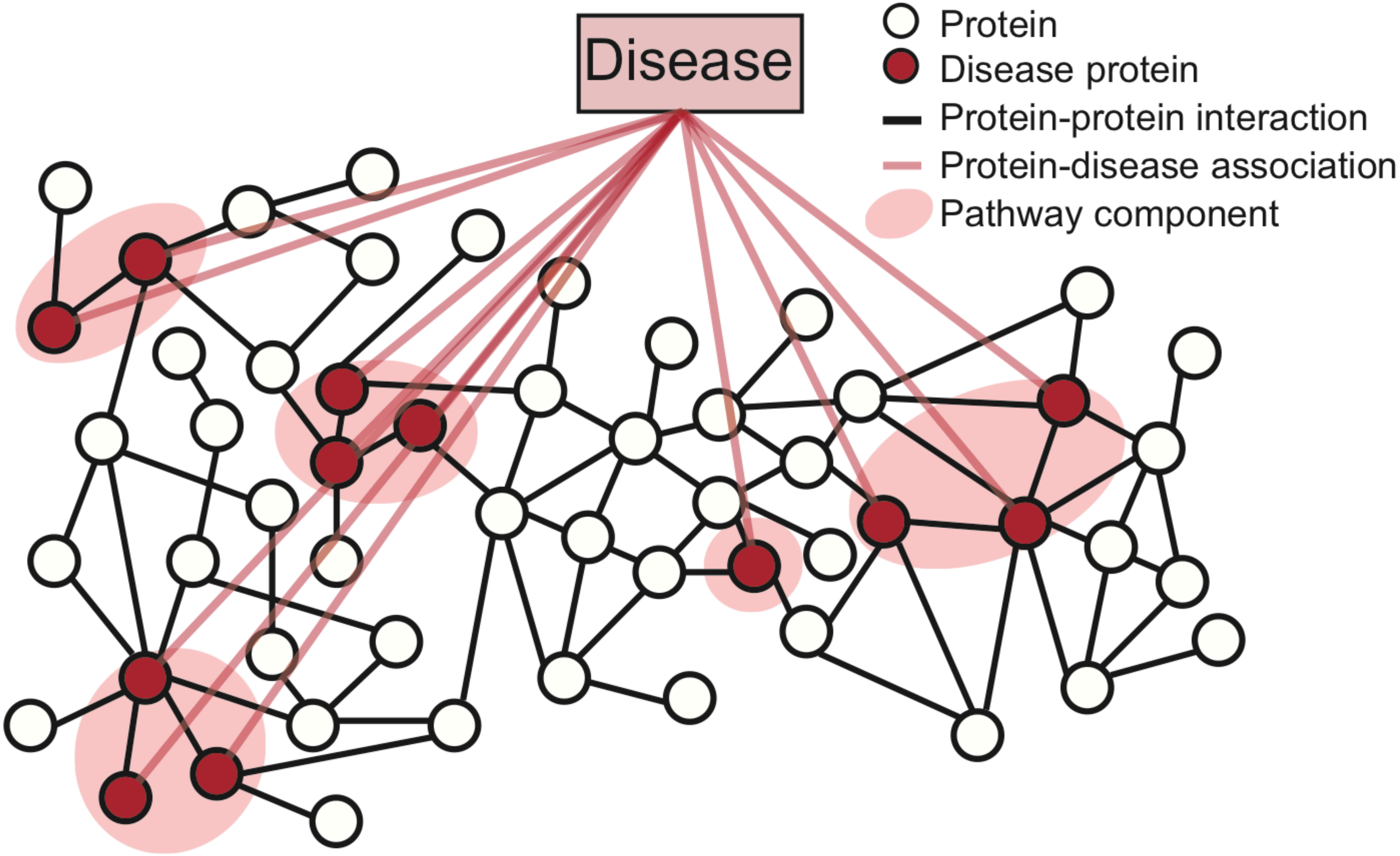
Noonan syndrome



Adrenal cortex carcinoma

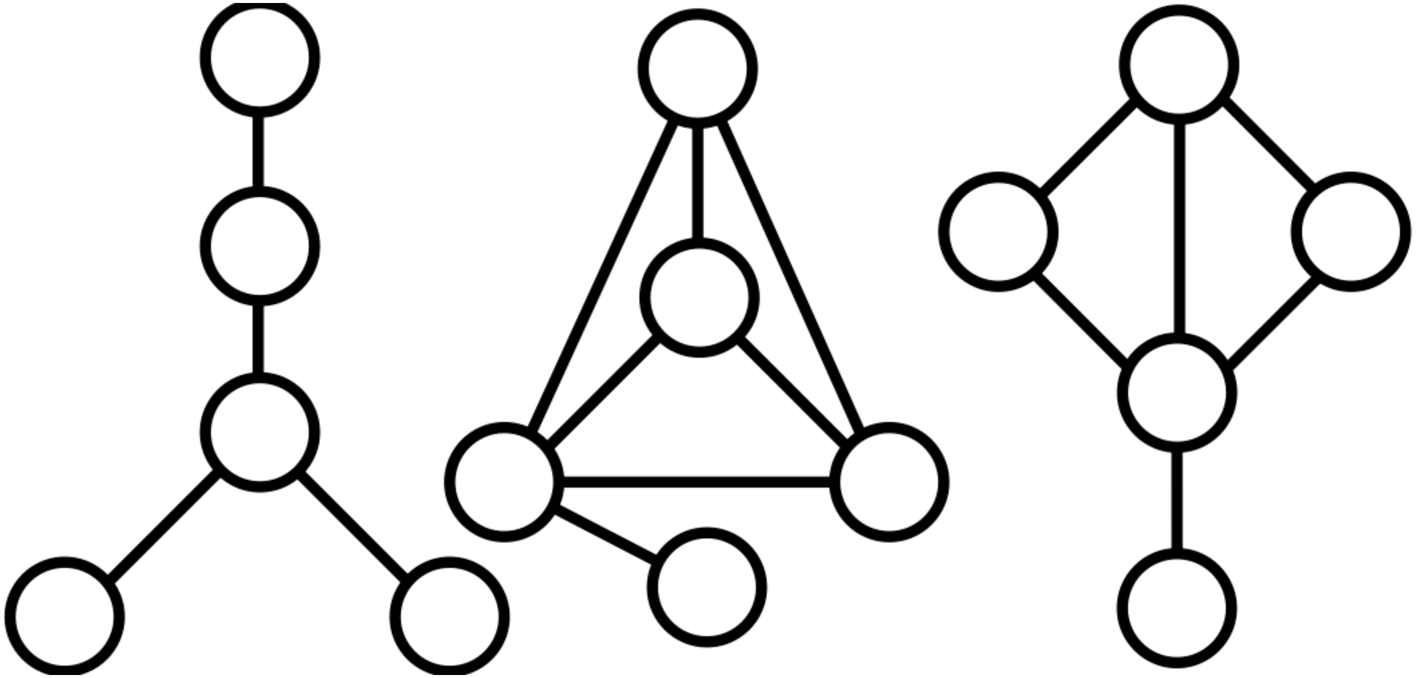


Pathways are Multi-Regional!



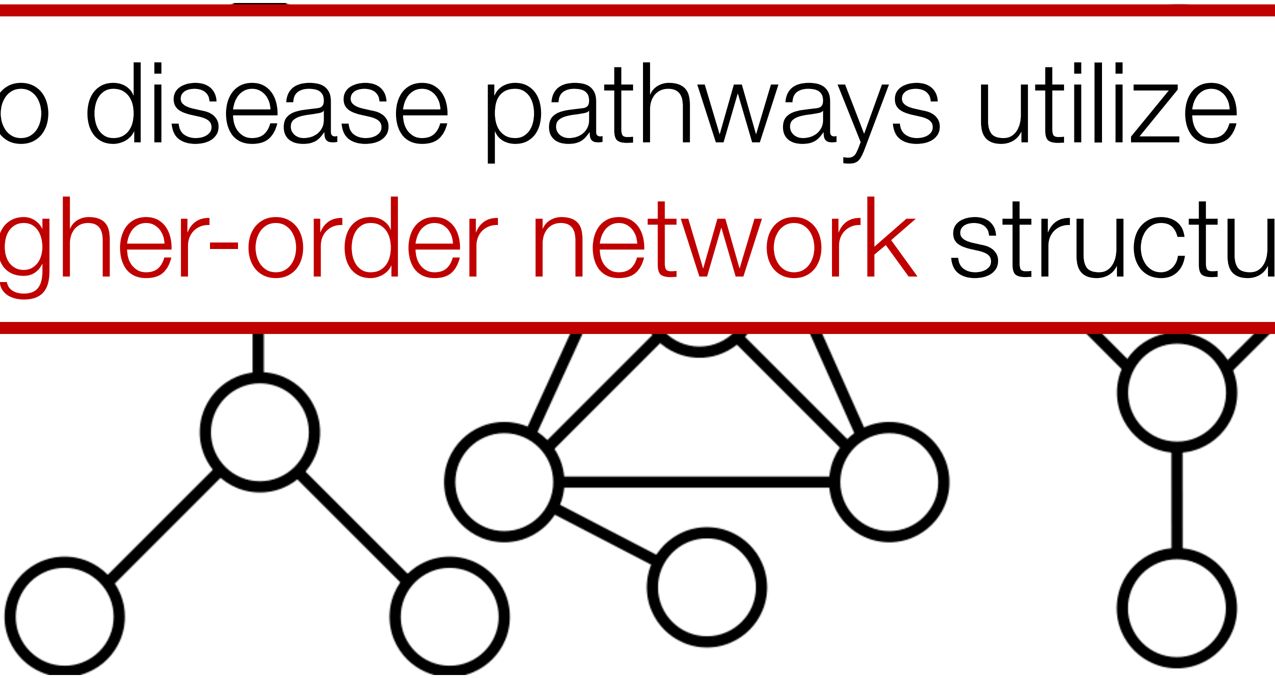
How To Proceed?

- Network motifs: Higher-order network structures



How To Proceed?

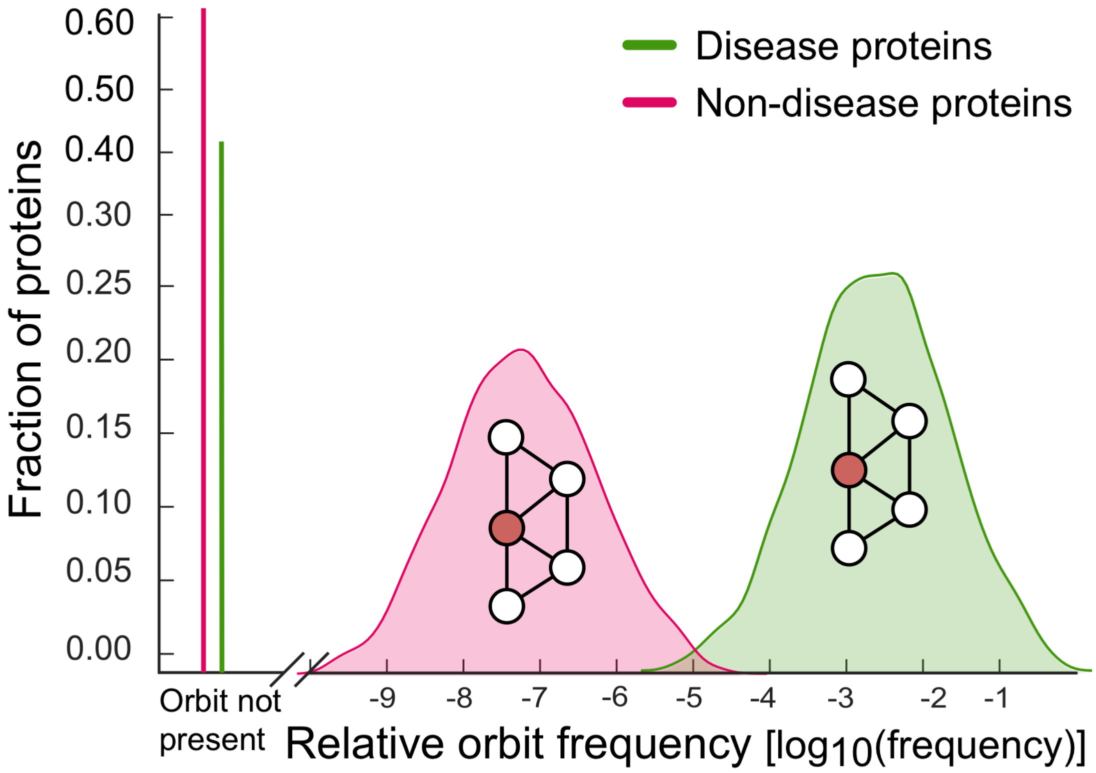
- Network motifs: Higher-order network structures



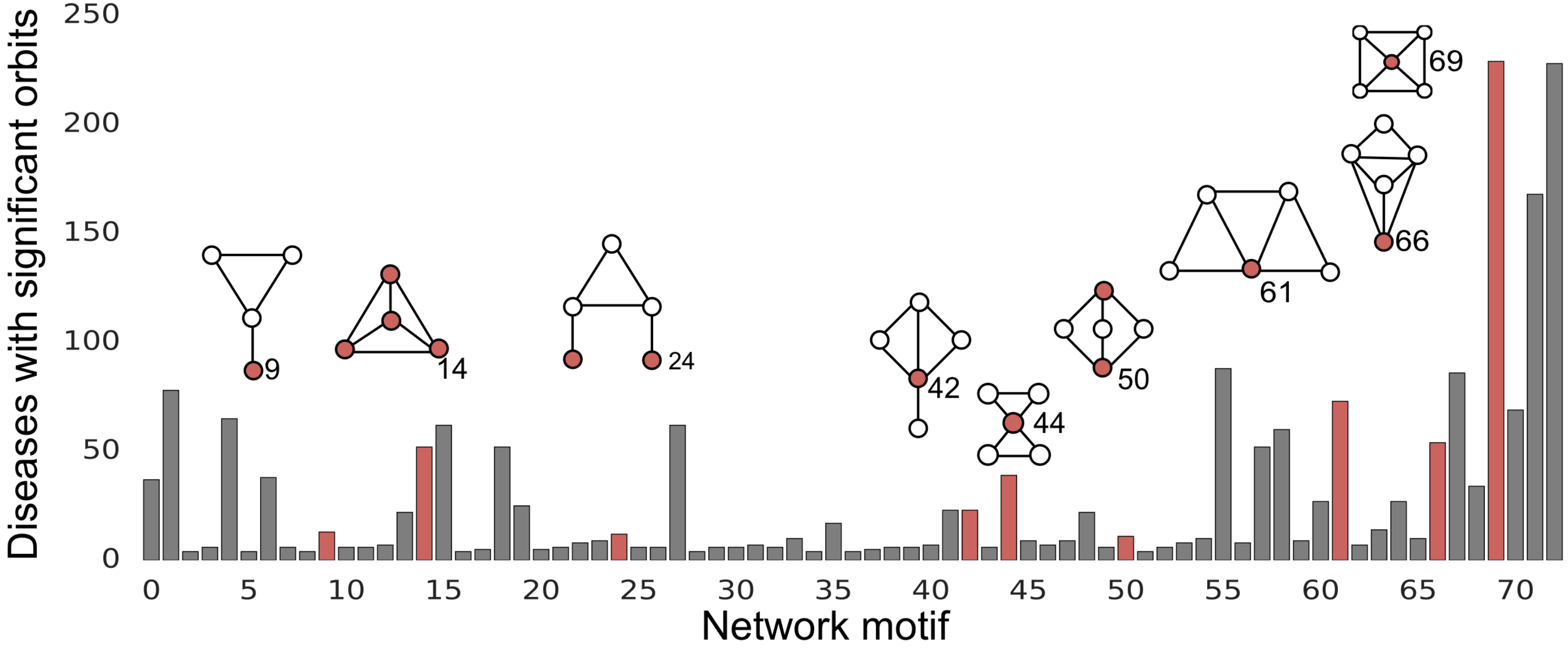
Do disease pathways utilize higher-order network structure?

Counting Network Structures

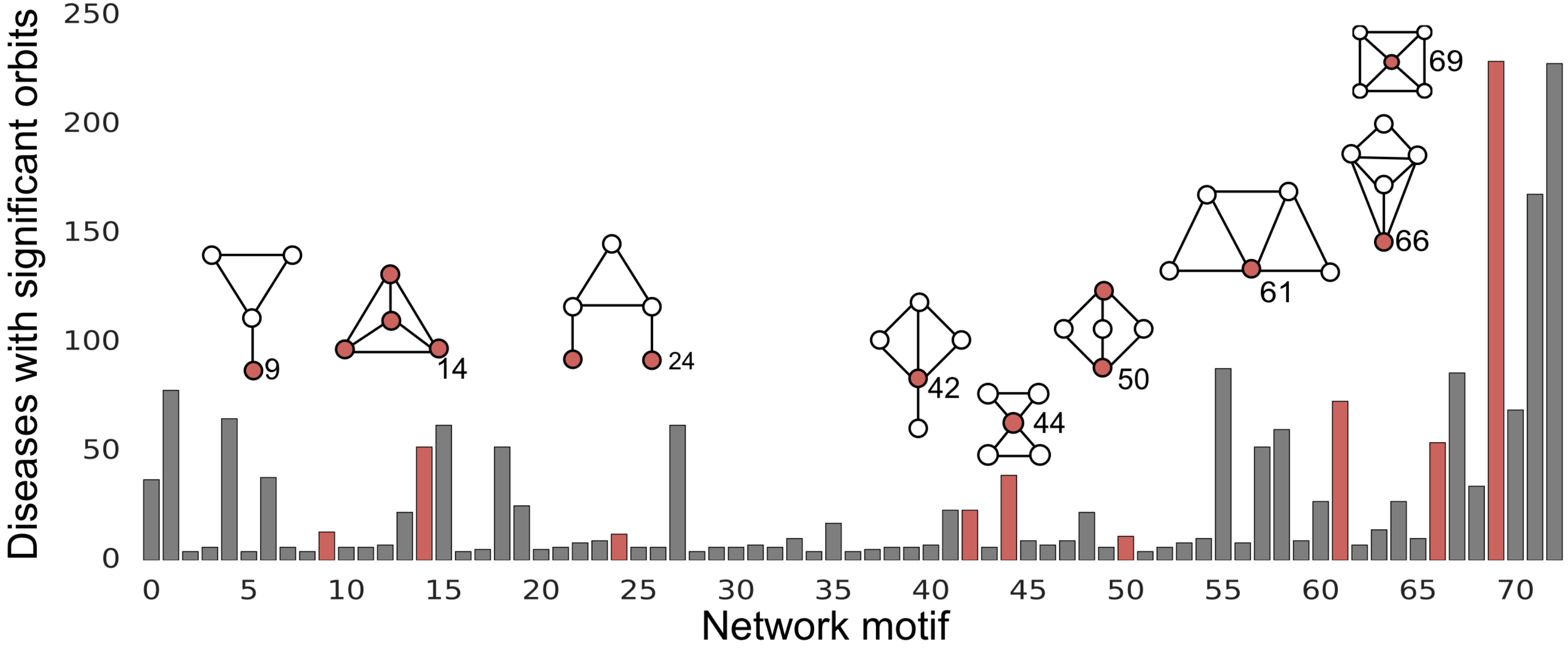
- 73 possible structures of size 2 to 5 nodes (edge → size-5 clique)



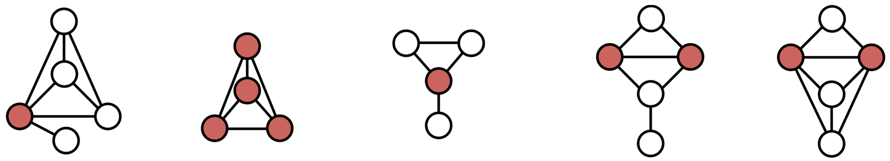
Are Network Motifs Abundant?



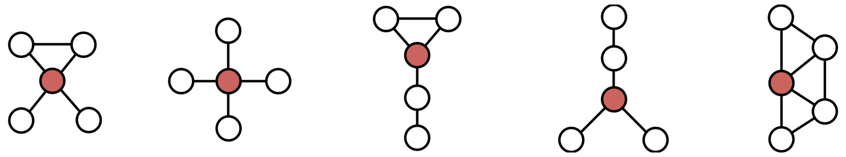
Are Network Motifs Abundant?



Cardiovascular diseases, e.g.,
Cardiomyopathy, Tachycardia

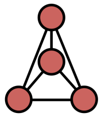


Cancers, e.g.,
Tumor of salivary gland, Thyroid carcinoma

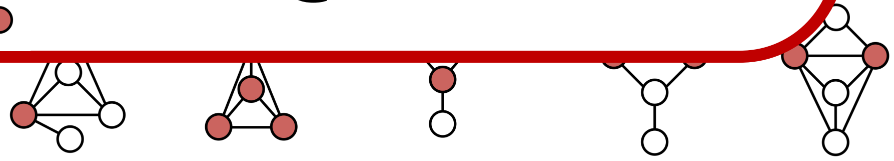


Are Network Motifs Abundant?

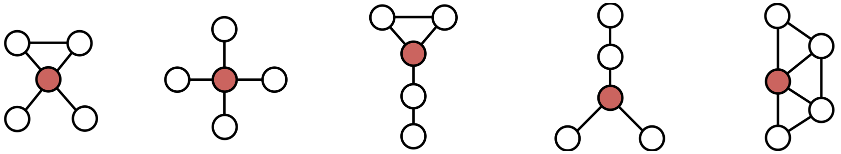
- Higher-order structures provide additional signal past edge connectivity
- Lead to better performance (11%, avg.)
- Example: Hearing loss:

$\text{hits@100} = 0.03$ →  → $\text{hits@100} = 0.77$

Cardiovascular diseases, e.g.,
Cardiomyopathy, Tachycardia



Cancers, e.g.,
Tumor of salivary gland, Thyroid carcinoma



Summary & Conclusions

- Current method assumptions not valid
- Propose **new prediction paradigm:**
  - Disease pathways are loosely interlinked
  - Multi-regional objects with regions distributed throughout the PPI network
  - Higher-order connectivity is important

snap.stanford.edu/pathways