

Network Biology

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2021

Daifeng Wang

daifeng.wang@wisc.edu

Goals for lecture

- Biological networks
- Challenges of integrating high-throughput assays
- Connecting relevant genes/proteins with interaction networks
- ResponseNet algorithm
- Evaluating pathway predictions
- Classes of signaling pathway prediction methods

High-throughput screening

- Which genes are involved in which cellular processes?
- Hit: gene that affects the phenotype
- Phenotypes include:
 - Growth rate
 - Cell death
 - Cell size
 - Intensity of some reporter
 - Many others

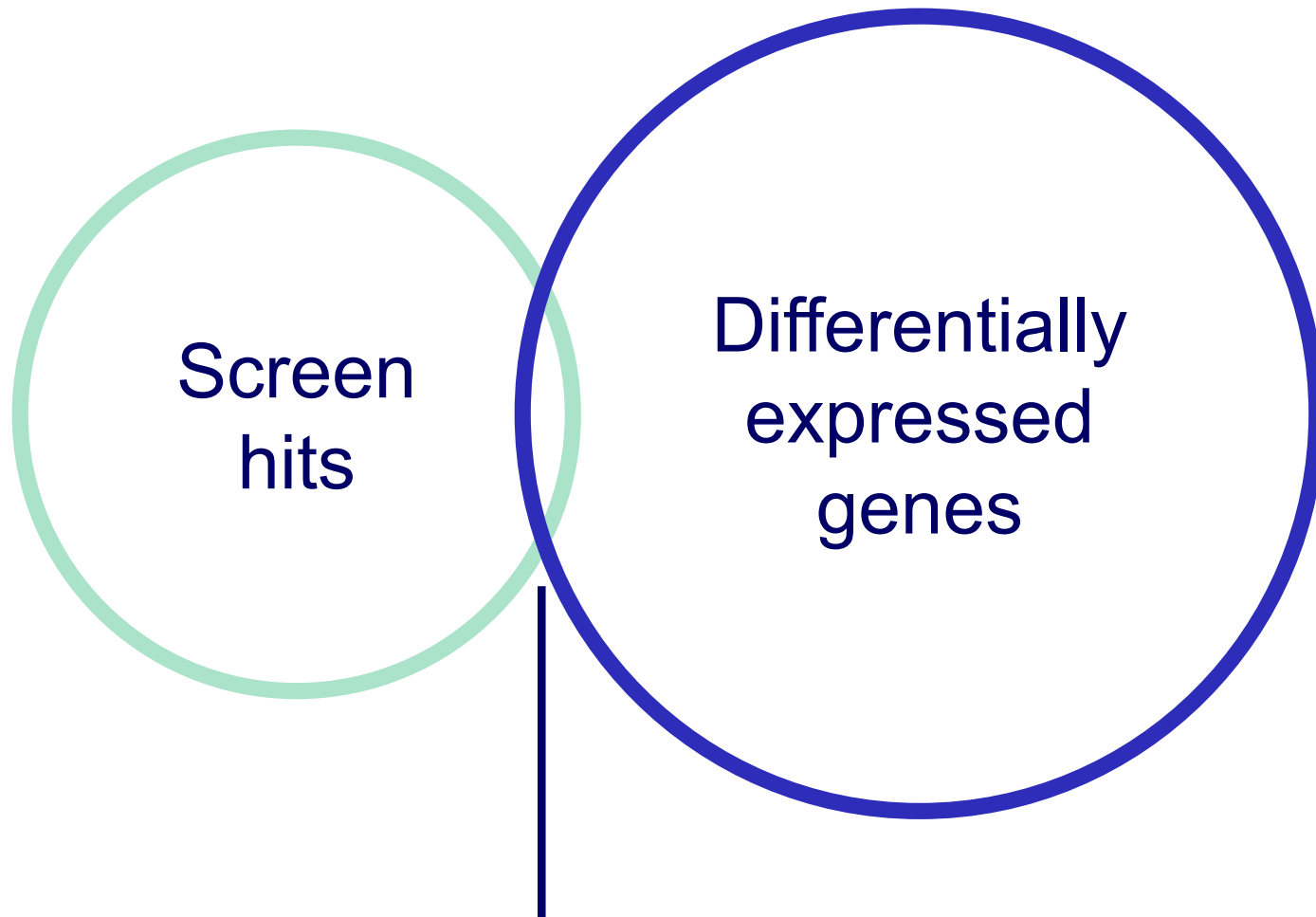
Types of screens

- Genetic screening
 - Test genes individually or in parallel
 - Knockout, knockdown (RNA interference), overexpression, CRISPR/Cas genome editing
- Chemical screening
 - Which genes are affected by a stimulus?

Differentially expressed genes

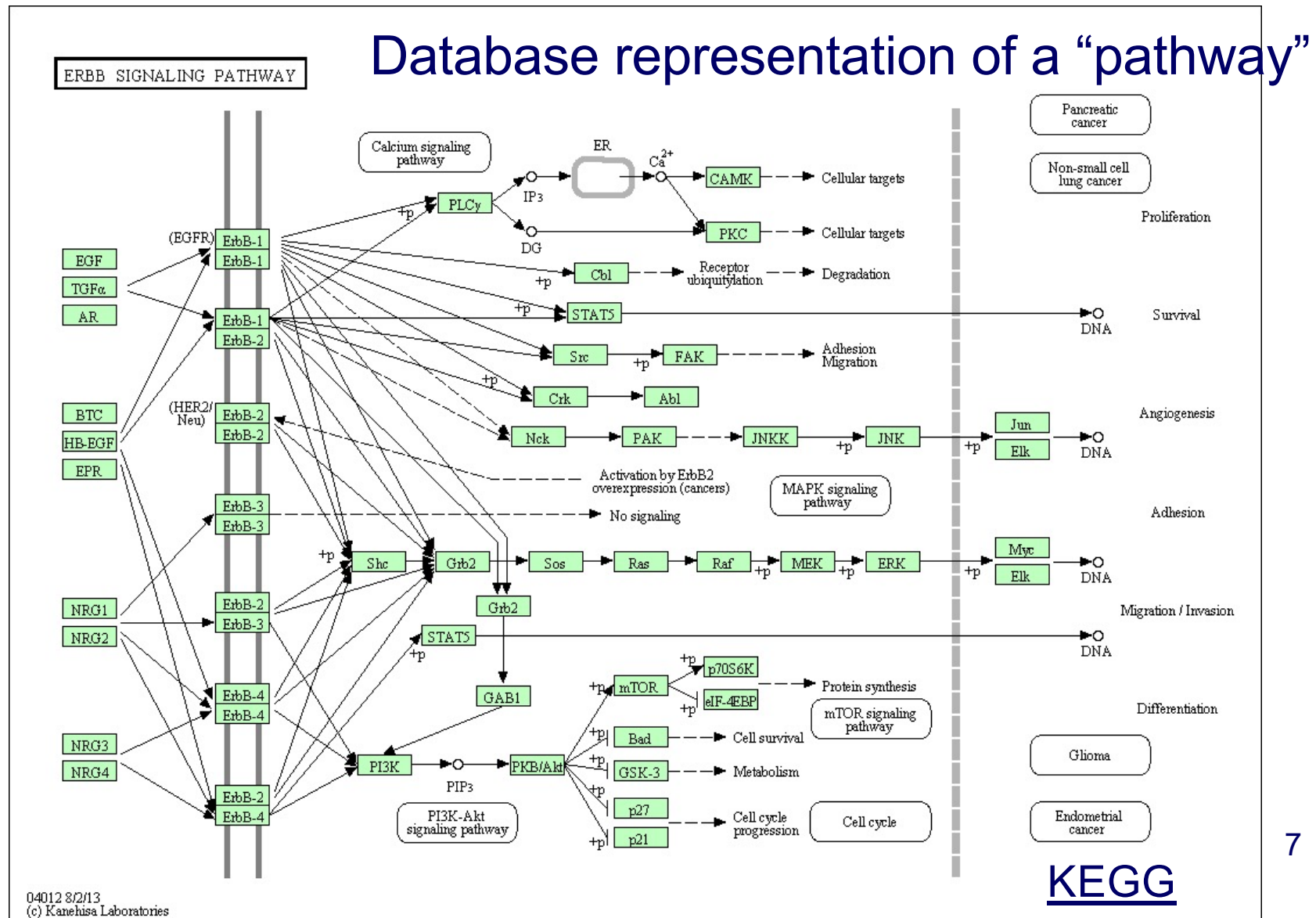
- Compare mRNA transcript levels between control and treatment conditions
- Genes whose expression changes significantly are also involved in the cellular process
- Alternatively, differential protein abundance or phosphorylation

Interpreting screens



Very few genes detected in both

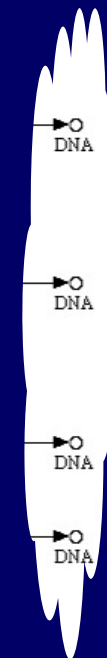
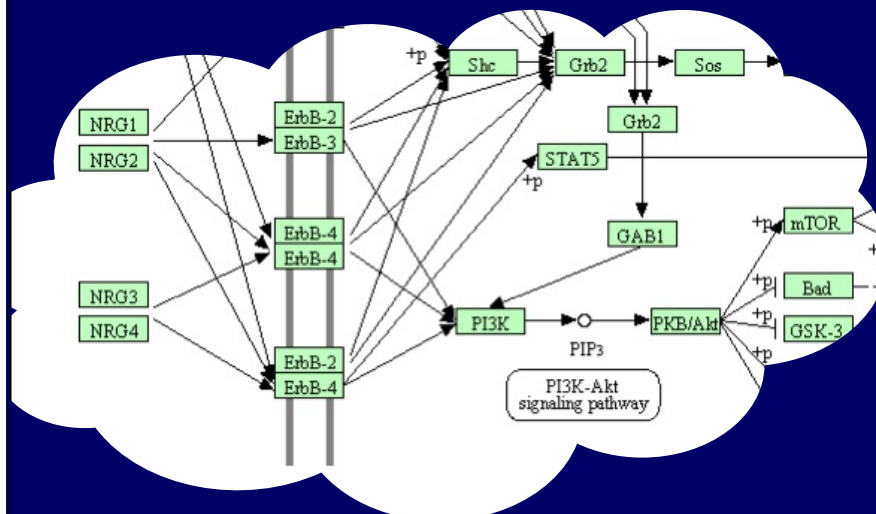
Assays reveal different parts of a cellular process



Assays reveal different parts of a cellular process

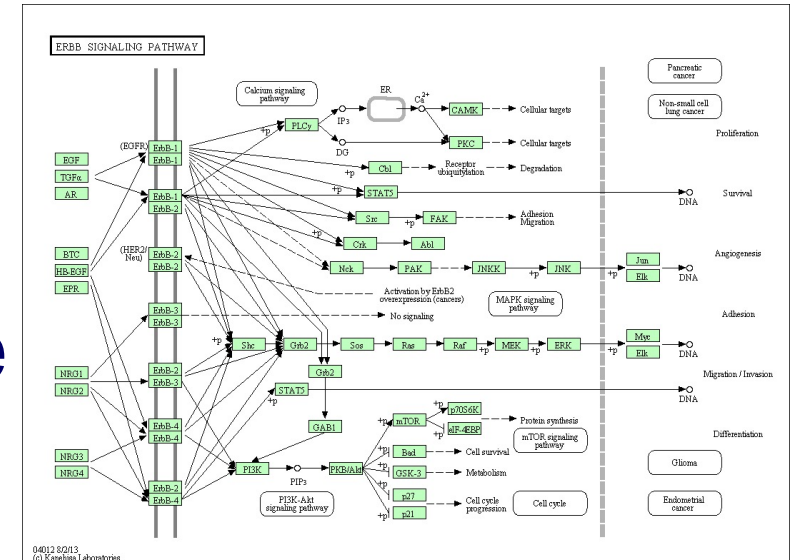
Differentially expressed genes

Genetic screen hits



Pathways connect the disjoint gene lists

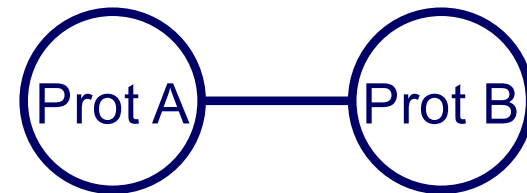
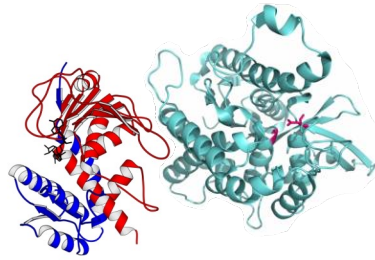
- Can't rely on pathway databases
- High-quality, low coverage
- Instead learn condition-specific pathways computationally
- Combine data with generic physical interaction networks



Physical interactions

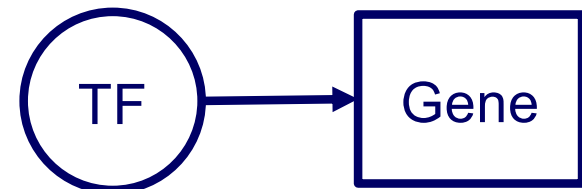
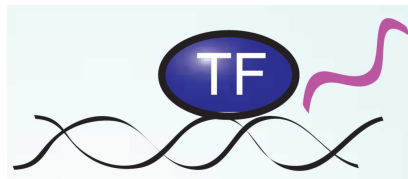
- Protein-protein interactions (PPI)

Appling Graz



- Metabolic
- Protein-DNA (transcription factor-gene)

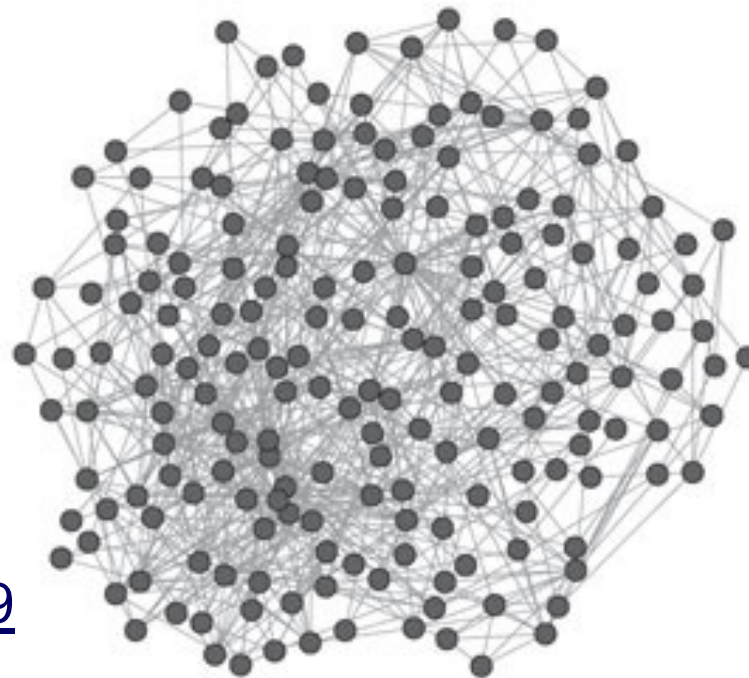
Yeger-Lotem2009



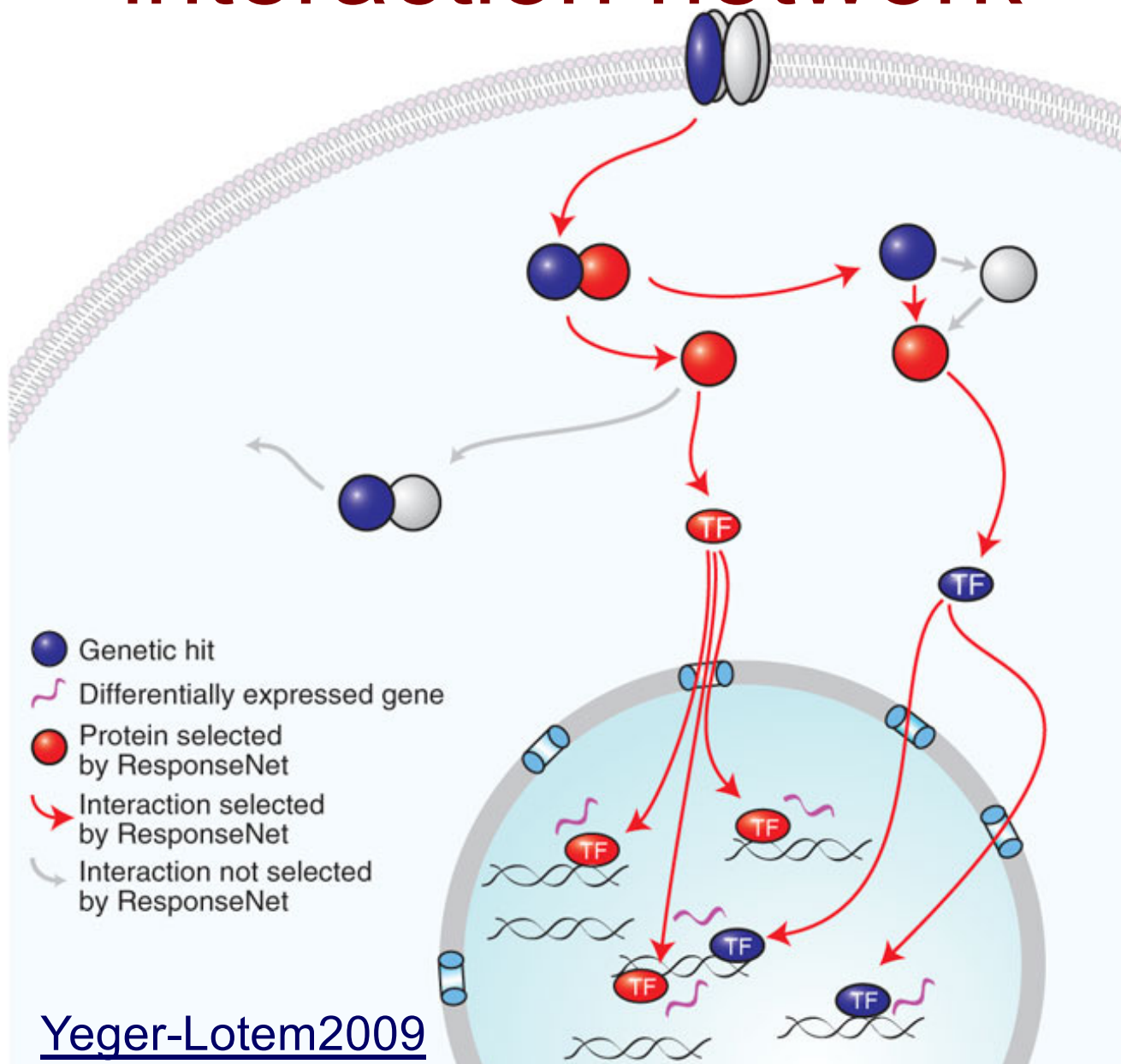
- Genes and proteins are different node types

Hairball networks

- Networks are highly connected
- Can't use naïve strategy to connect screen hits and differentially expressed genes



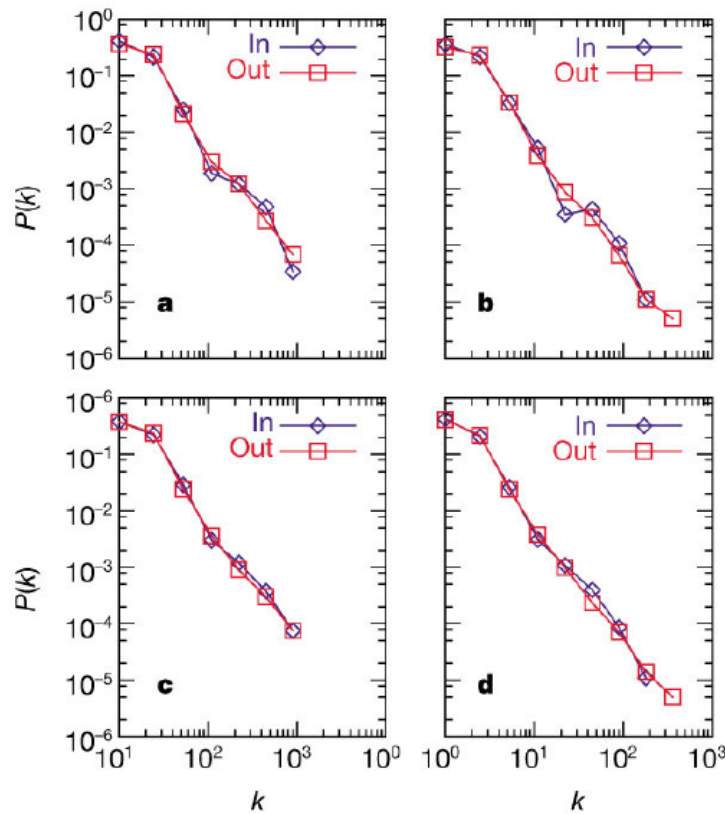
Identify connections within an interaction network



Biological Network Properties

- Degree: number of neighbors of a node
- Power law degree distribution
 - Most nodes have low degrees
 - Few highly connected nodes (hubs)
- Robust to random attacks
 - e.g., structure resilient to mutations
 - Mutations in hubs can damage the network
- Modular organization
 - High clustering coefficient (short paths)
 - Efficient signal propagation

Power law degree distribution

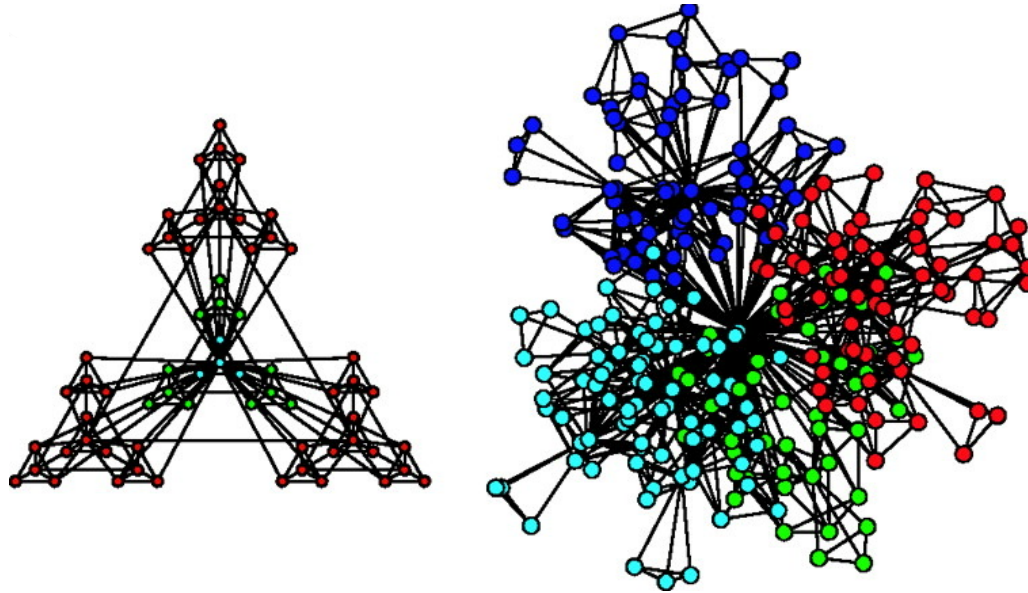


H. Jeong et al., Nature, 407 (2000)

- Probability of finding a highly connected node decreases exponentially with K

$$P(K) \sim K^{-\gamma}$$

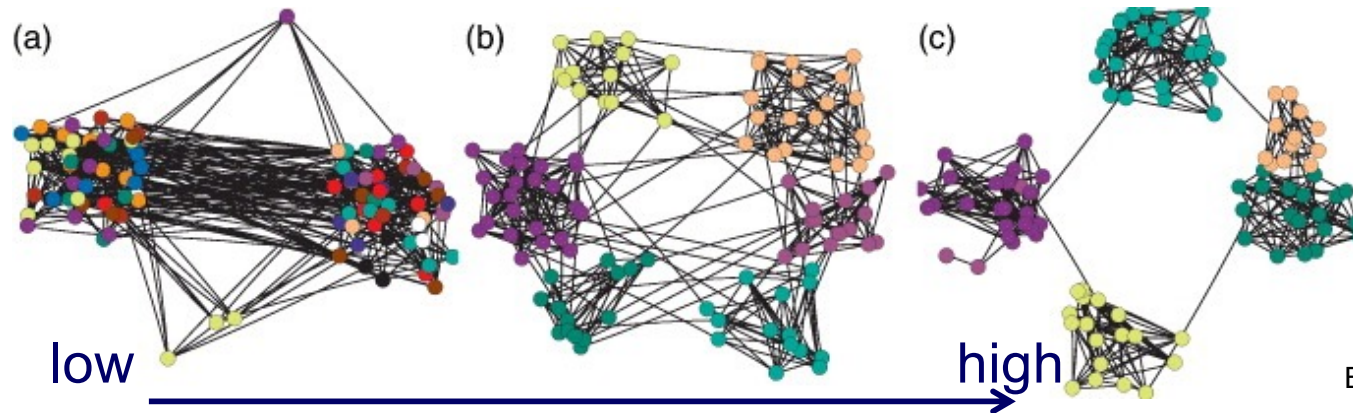
Modularity



E. Ravasz et al., Science 297, 1551 -1555 (2002)

- Small highly connected cohesive clusters that combine to form larger units
- Communication between clusters through hubs
- Hierarchical modularity overlaps with known metabolic functions

Measurement of Modularity



Brede, Europhysics Letters, 2010.

Modularity Q : measurement on strength of network division

$$Q = \frac{1}{2m} \sum_{i,j} \left(w_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

$\frac{1}{2m}$: normalization
 m : total number of edges
 w_{ij} : edge weight between nodes i and j
 $\frac{k_i k_j}{2m} = p_{ij}$: expected edge weight that would go between i and j
 $\delta_{\sigma_i \sigma_j}$: sum over nodes within a group (module)



Clustering goal: assign each node a module to maximize “modularity” as an objective function (module is a group of highly connected nodes)

Newman, PNAS, 2006.

Clustering coefficient

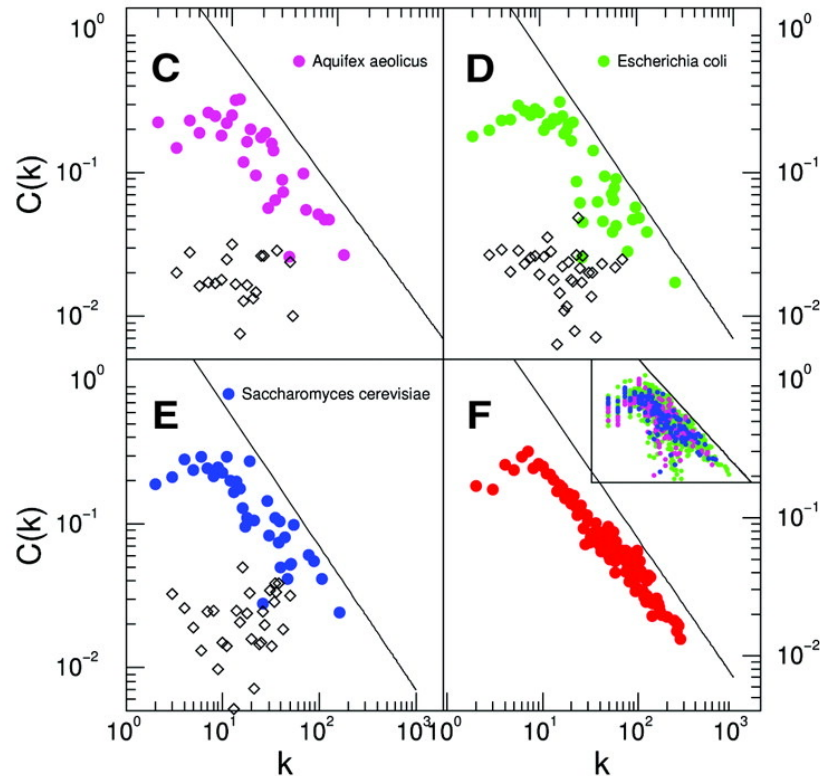
Measures the average probability that two neighbors of a node are connected

$$C_I = \frac{n_I}{\binom{k}{2}} = \frac{2n_I}{k \cdot (k-1)}$$

n_I : # edges between node I 's neighbors

k : # of neighbors of I

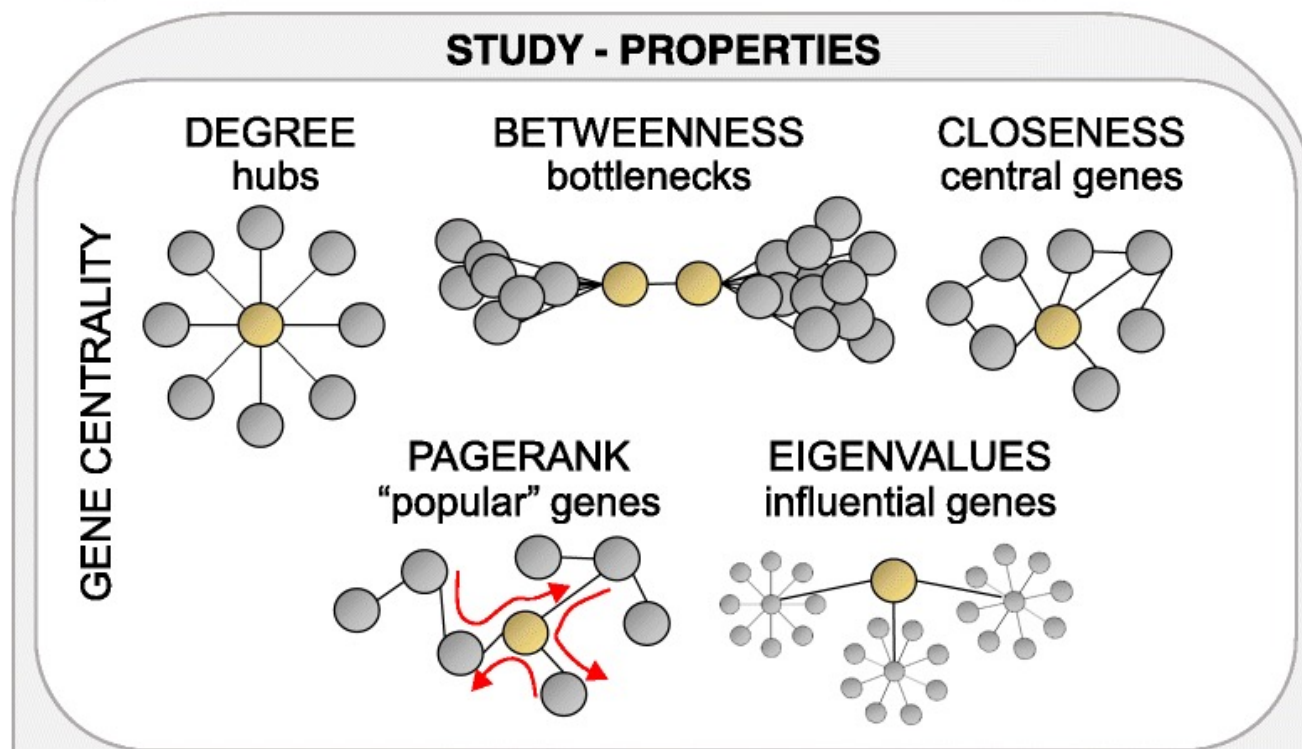
Clustering coefficient



- High degree nodes $\rightarrow$ low clustering coefficient CC
- Network's modularity $\rightarrow$ CC averaged over all nodes
- Metabolic networks have high intrinsic modularity

Network centralities

Topological importance of a node



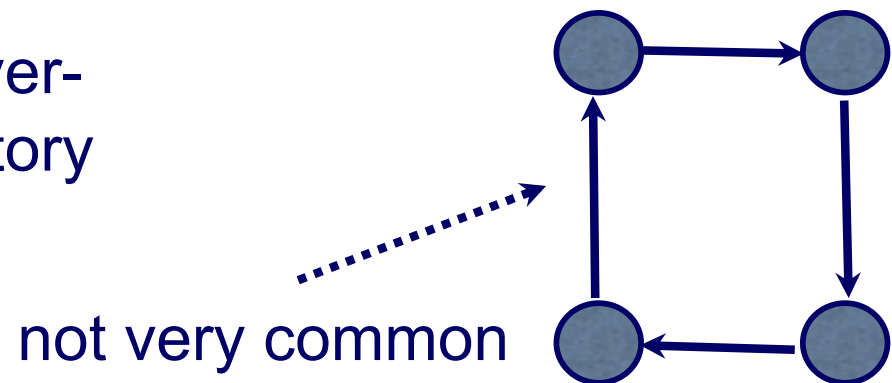
G. Iacono et al., Genome Biology 20 (2019)

Network problems

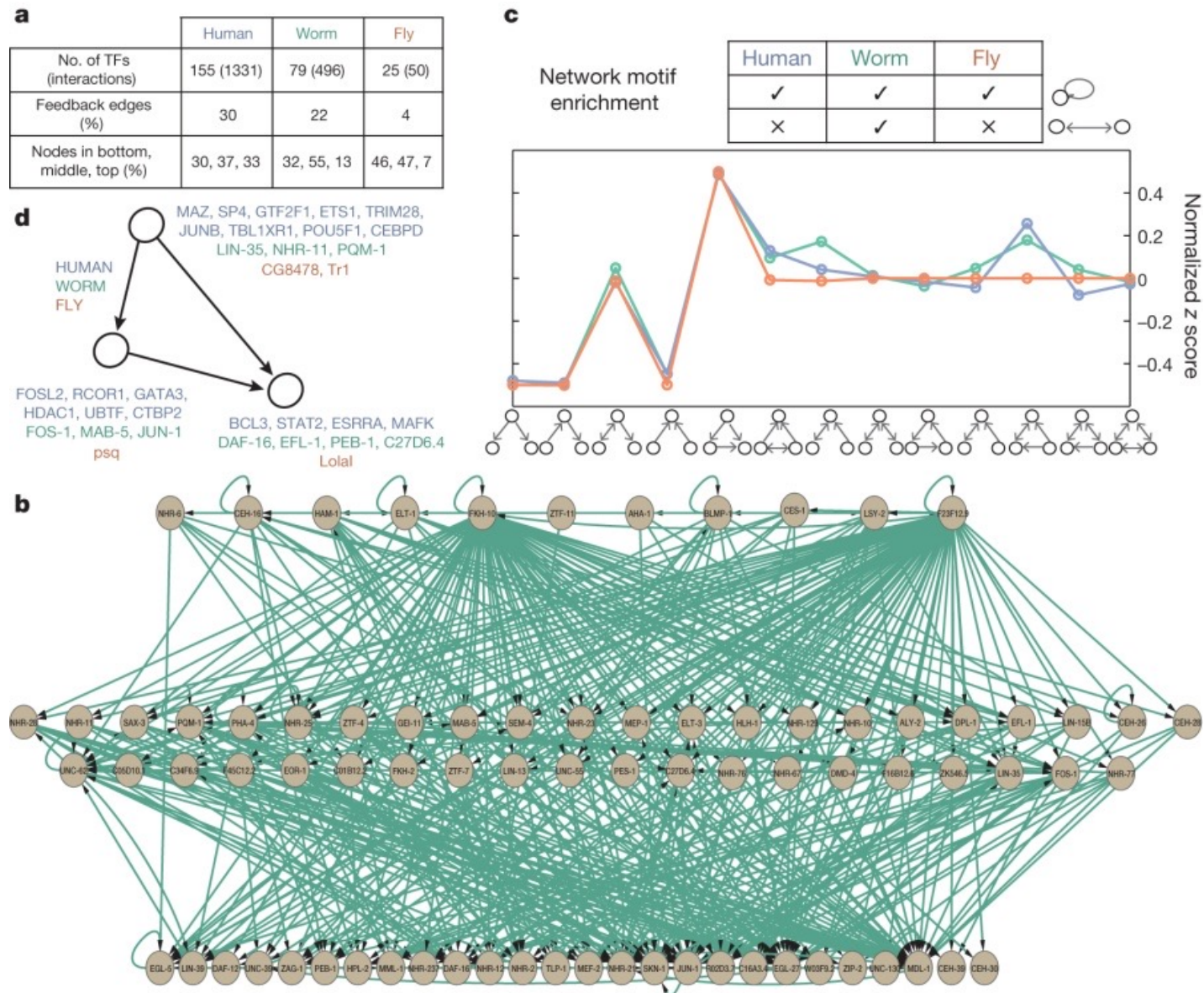
- Network inference
 - Infer network structure
- Motif finding
 - Identify common subgraph topologies
- Pathway or module detection
 - Identify subgraphs of genes that perform the same function or active in same condition
- Network comparison, alignment, querying
- Conserved modules
 - Identify modules that are shared in networks of multiple species/conditions

Network motifs

- Problem: Find subgraph topologies that are statistically more frequent than expected
- Brute force approach
 - Count all topologies of subgraphs of size m
 - Randomize graph (retain degree distribution) and count again
 - Output topologies that are over/under represented



Gene regulatory network motifs



Network modules

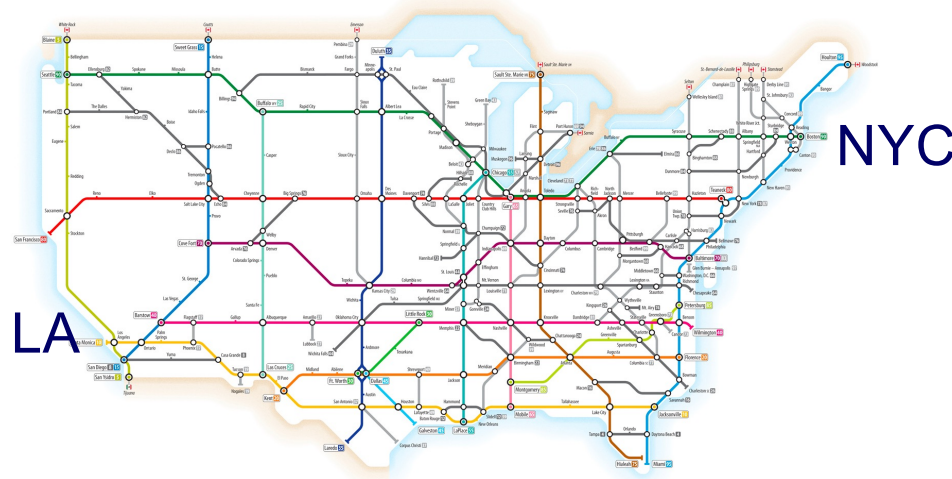
- Modules: dense (highly-connected) subgraphs (e.g., large cliques or partially incomplete cliques)
- Problem: Identify the component modules of a network
- Difficulty: definition of module is not precise
 - Hierarchical networks have modules at multiple scales
 - At what scale to define modules?

How to define a computational “pathway”

- **Given:**
 - Partially directed network of known physical interactions (e.g. PPI, kinase-substrate, TF-gene)
 - Scores on source nodes
 - Scores on target nodes
- **Do:**
 - Return directed paths in the network connecting sources to targets

Network flow problem

- Finding an optimal route by minimizing transportation costs from LA to NYC
 - $c_{i,j}$, the cost between City i and City j
 - $f_{i,j} = 1$ if in route, $= 0$ if not
 - $\operatorname{argmin}_f \sum_{i,j} c_{i,j} * f_{i,j}$ s.t. constraints

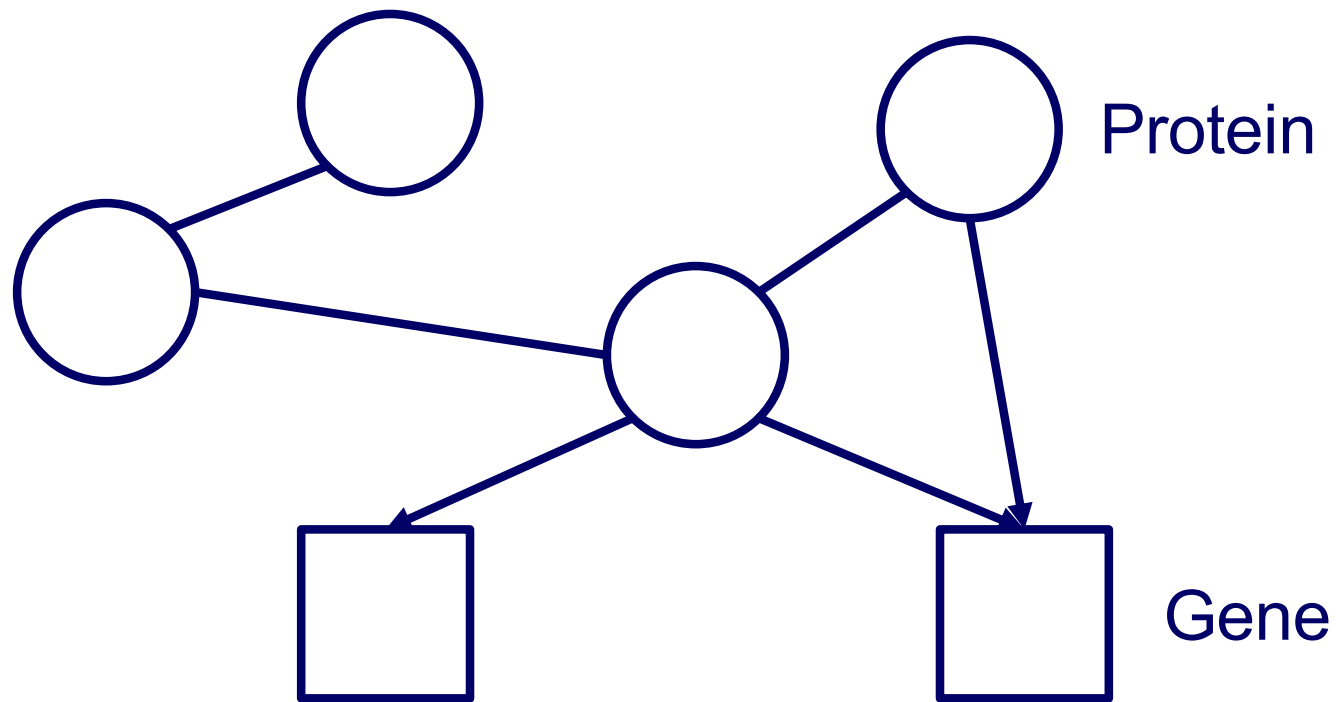


1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----

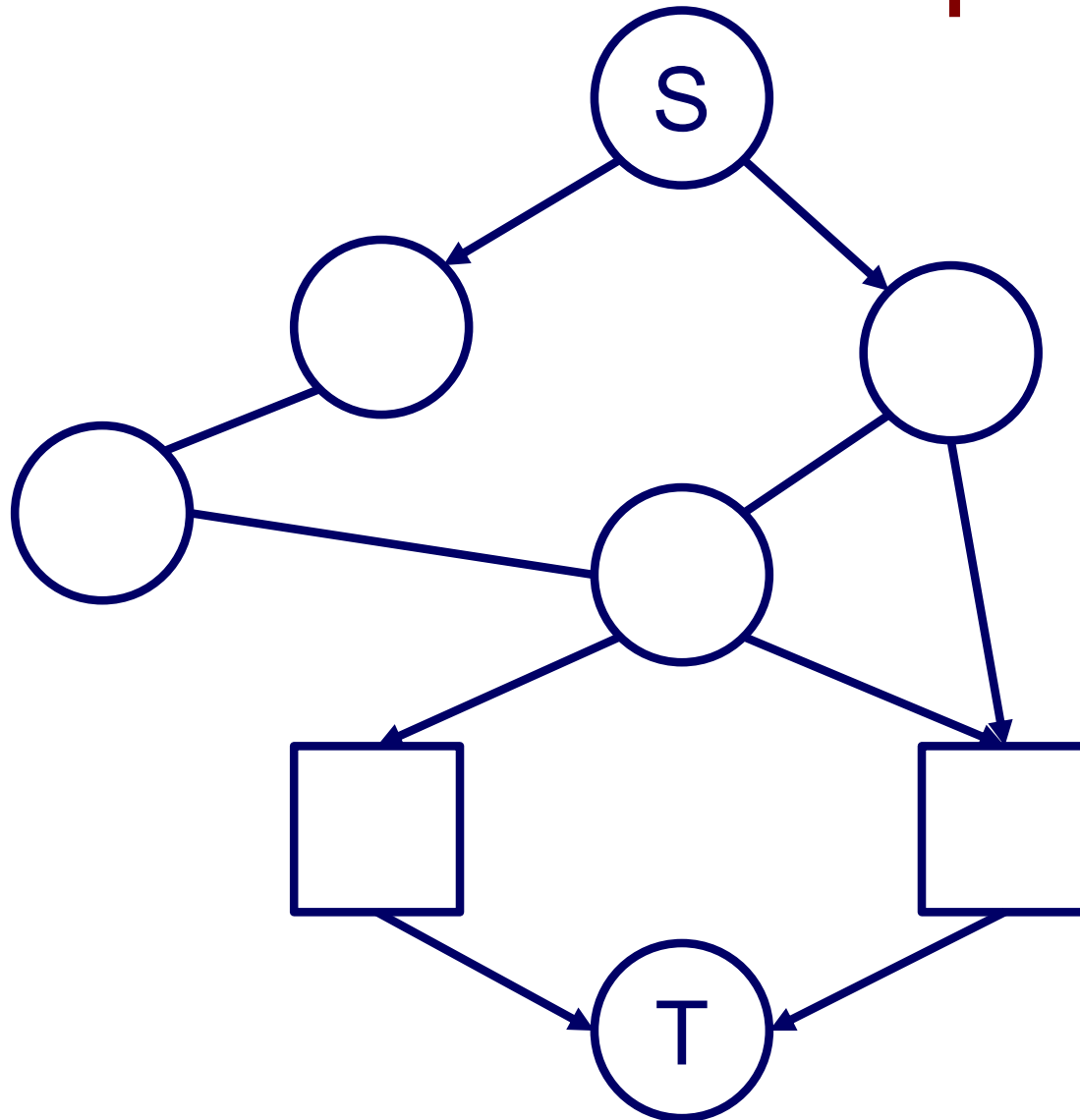
ResponseNet optimization goals

- Connect screen hits and differentially expressed genes
- Recover sparse connections
- Identify intermediate proteins missed by the screens
- Prefer high-confidence interactions
- Minimum cost flow formulation can meet these objectives

Construct the interaction network

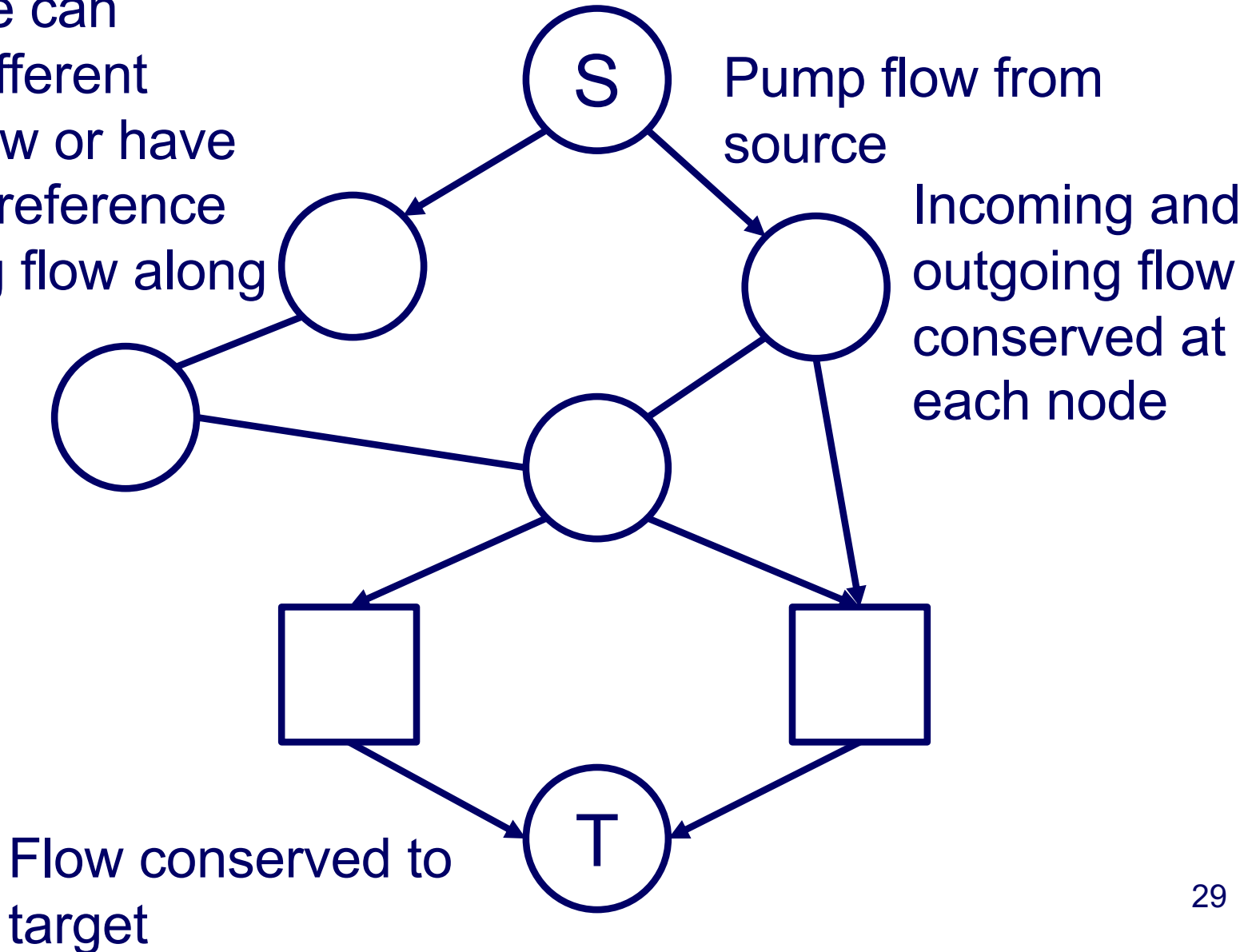


Transform to a flow problem



Max flow on graphs

Each edge can tolerate different level of flow or have different preference of sending flow along that edge



Weighting interactions

- Probability-like confidence of the interaction

Proteins

	MP2K1_HUMAN	Homo sapiens	Temporarily not available for viewing in Netility.
	MK01_HUMAN	Homo sapiens	Temporarily not available for viewing in Netility.

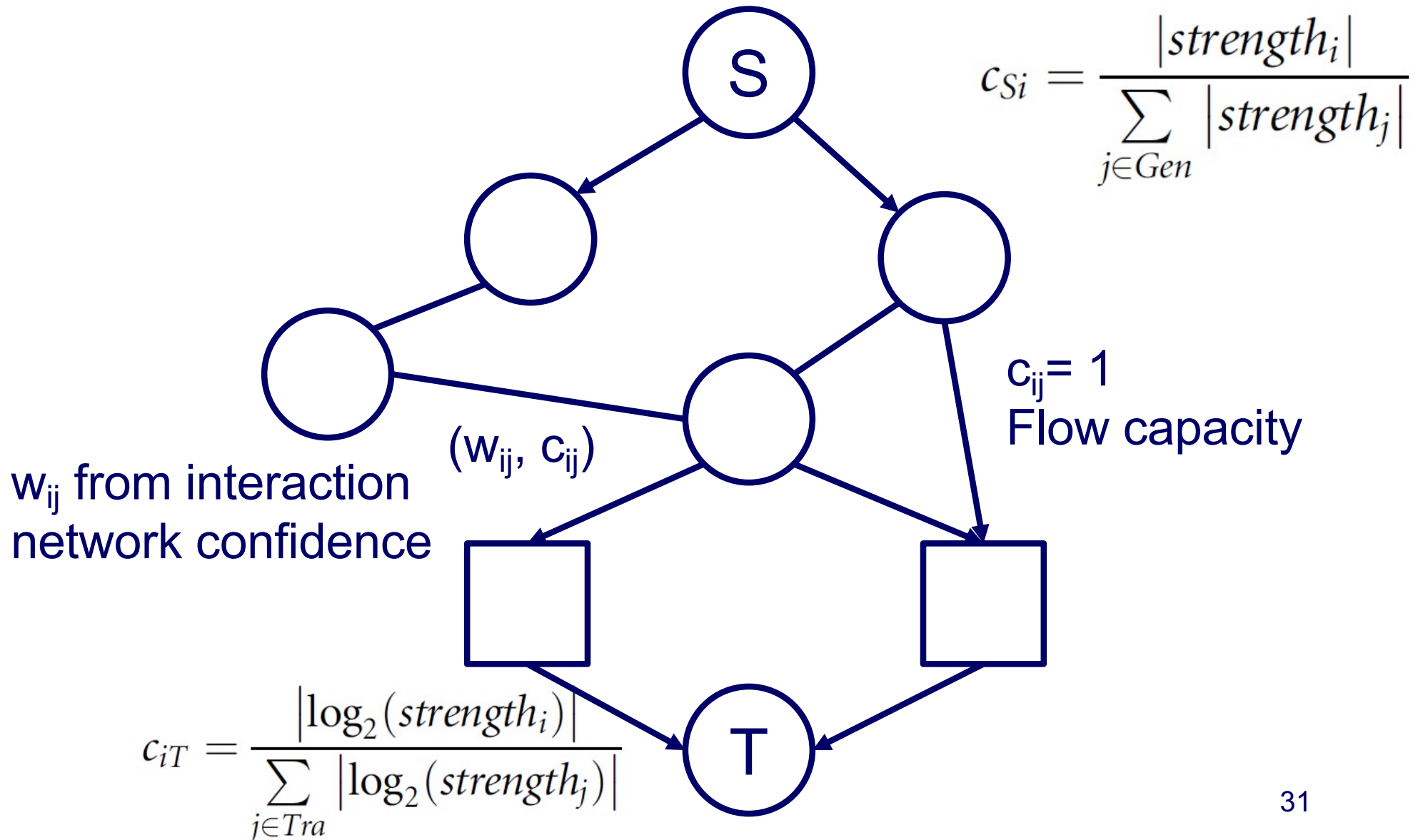
Evidence

Source DB ↕	Source ID ↕	Interaction Type ↕	PSI MI Code ↕	PubMed ID ↕	Detection Type ↕	PSI MI Code ↕
biogrid	857930	direct interaction	MI:0407	12788955	enzymatic study	MI:0415
ophid	17231	aggregation	MI:0191	11352917	confirmational text mining	MI:0024
ophid	17231	aggregation	MI:0191	15657099	deglycosylase assay	MI:1006
ophid	17234	aggregation	MI:0191	11352917	confirmational text mining	MI:0024
ophid	17234	aggregation	MI:0191	15657099	deglycosylase assay	MI:1006
biogrid	259225	direct interaction	MI:0407	12697810	t7 phage display	MI:0108
intact	EBI-8279991 	phosphorylation reaction	MI:0217	23241949	biosensor	MI:0968

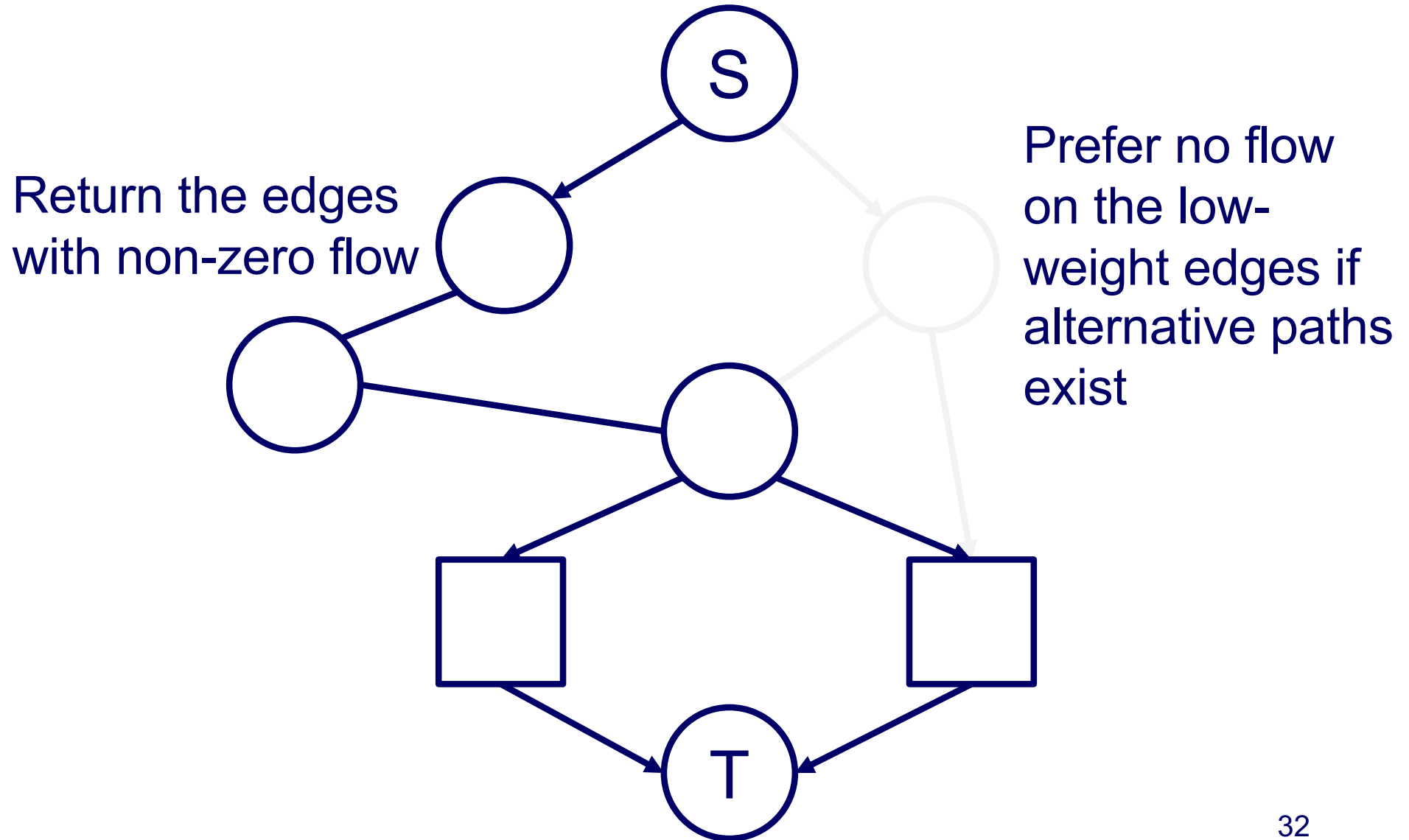
[iRefWeb](#)

- Example evidence: edge score of 1.0
- 16 distinct publications supporting the edge 30

Weights and capacities on edges



Find the minimum cost flow



Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right)$$

Positive flow on
an edge incurs a
cost

Flow on an
edge

Cost is greater for
low-weight edges

Parameter
controlling the
amount of flow from
the source

Formal minimum cost flow

$$\min_f \left(\left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right) \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

Flow coming in to a node
equals flow leaving the
node

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

Flow leaving the
source equals flow
entering the target

Formal minimum cost flow

$$\min_f \left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

Flow is non-negative and does not exceed edge capacity

$$0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'$$

Formal minimum cost flow

$$\min_f \left(\left(\sum_{i \in V', j \in V'} -\log(w_{ij}) * f_{ij} \right) - \left(\gamma * \sum_{i \in Gen} f_{Si} \right) \right)$$

Subject to:

$$\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}$$

$$\sum_{i \in Gen} f_{Si} - \sum_{i \in Tra} f_{iT} = 0$$

$$0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'$$

Linear programming

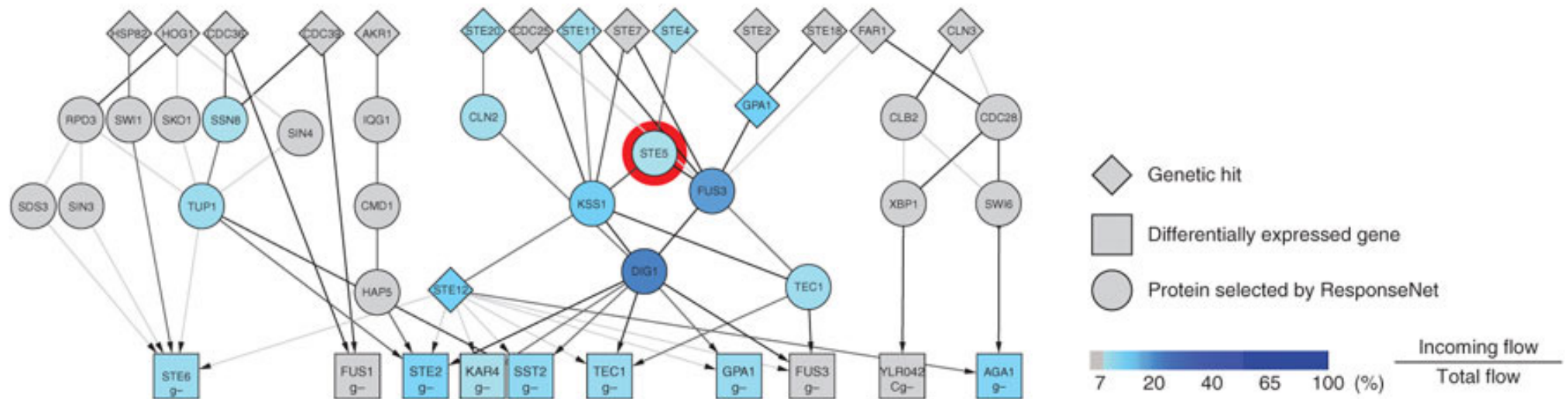
- Optimization problem is a linear program
- Canonical form

$$\begin{array}{ll}\text{maximize} & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & \mathbf{Ax} \leq \mathbf{b} \\ \text{and} & \mathbf{x} \geq 0\end{array}$$

[Wikipedia](#)

- Polynomial time complexity
- Many off-the-shelf solvers
- Practical Optimization: A Gentle Introduction
 - Introduction to linear programming
 - Simplex method
 - Network flow

ResponseNet pathways



- Identifies pathway members that are neither hits nor differentially expressed
- Ste5 recovered when *STE5* deletion is the perturbation

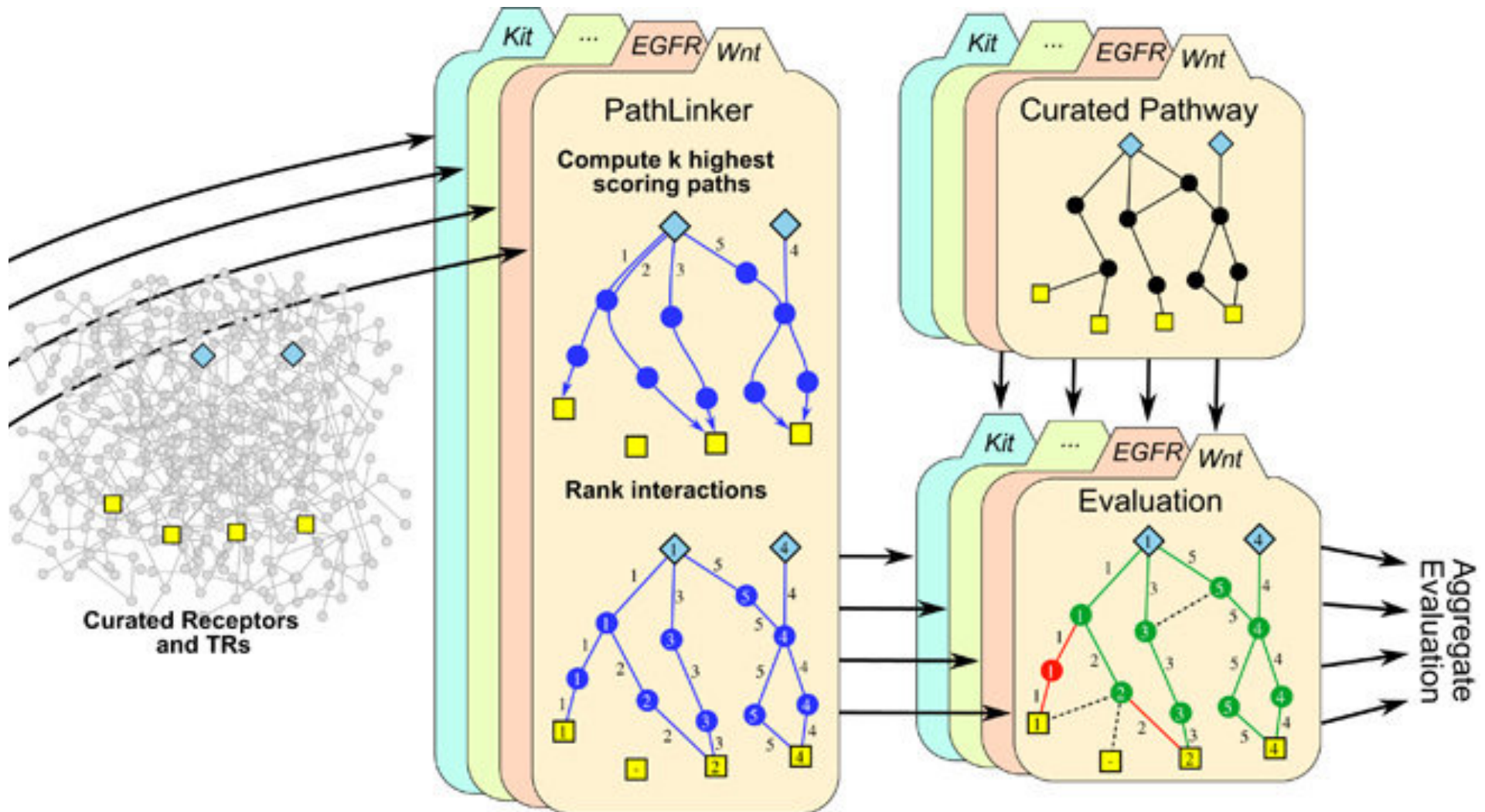
ResponseNet summary

- Advantages
 - Computationally efficient
 - Integrates multiple types of data
 - Incorporates interaction confidence
 - Identifies biologically plausible networks
- Disadvantages
 - Direction of flow is not biologically meaningful
 - Path length not considered
 - Requires sources and targets
 - Dependent on completeness and quality of input network

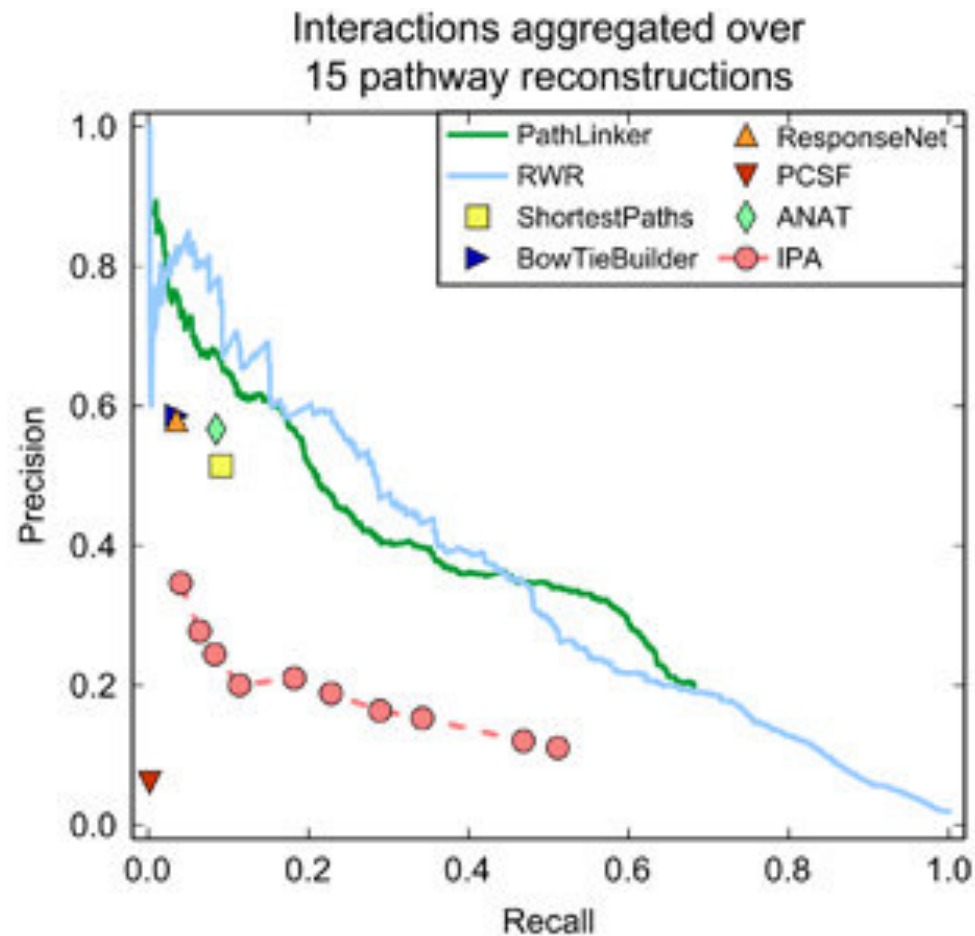
Evaluating pathway predictions

- Unlike PIQ, we don't have a complete gold standard available for evaluation
- Can simulate “gold standard” pathways from a network
- Compare relative performance of multiple methods on independent data

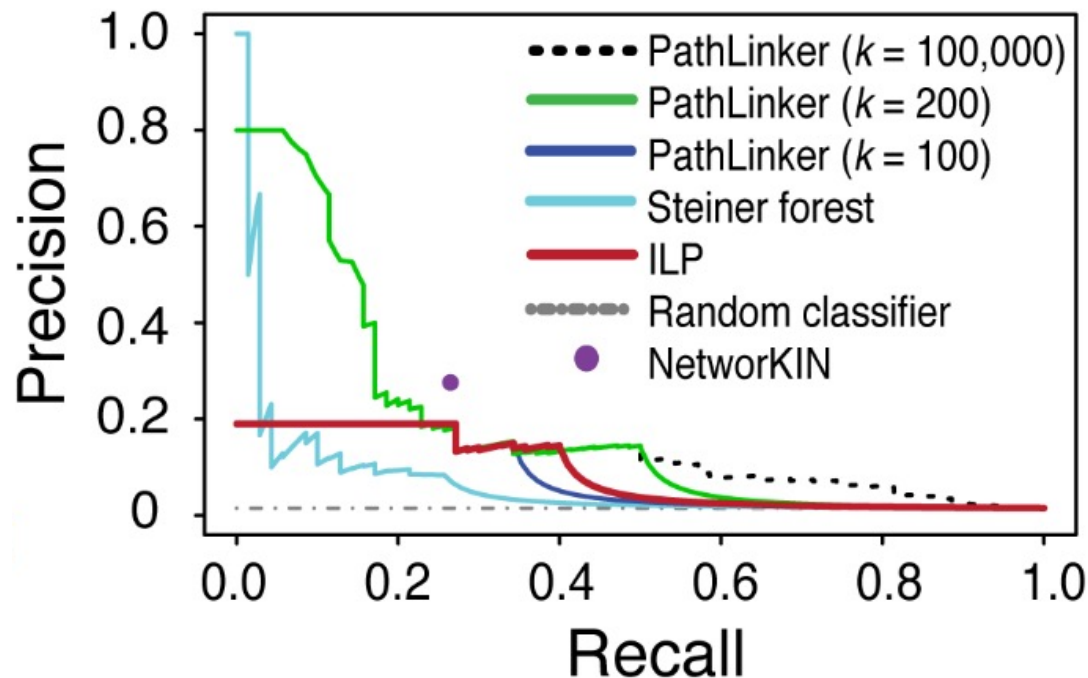
Evaluating pathway predictions



Evaluating pathway predictions



Evaluating pathway predictions



[MacGilvray2018](#)

- PR curves can evaluate node or edge recovery but not the global pathway structure

Evaluation beyond pathway databases

- Natural language processing can also help semi-automated evaluation

- Literome

PMID: 14611643

WNK1, the kinase mutated in an inherited high-blood-pressure syndrome, is a novel PKB (protein kinase B)/Akt substrate.

... that PKB mediates the ... of WNK1 at ... (details)

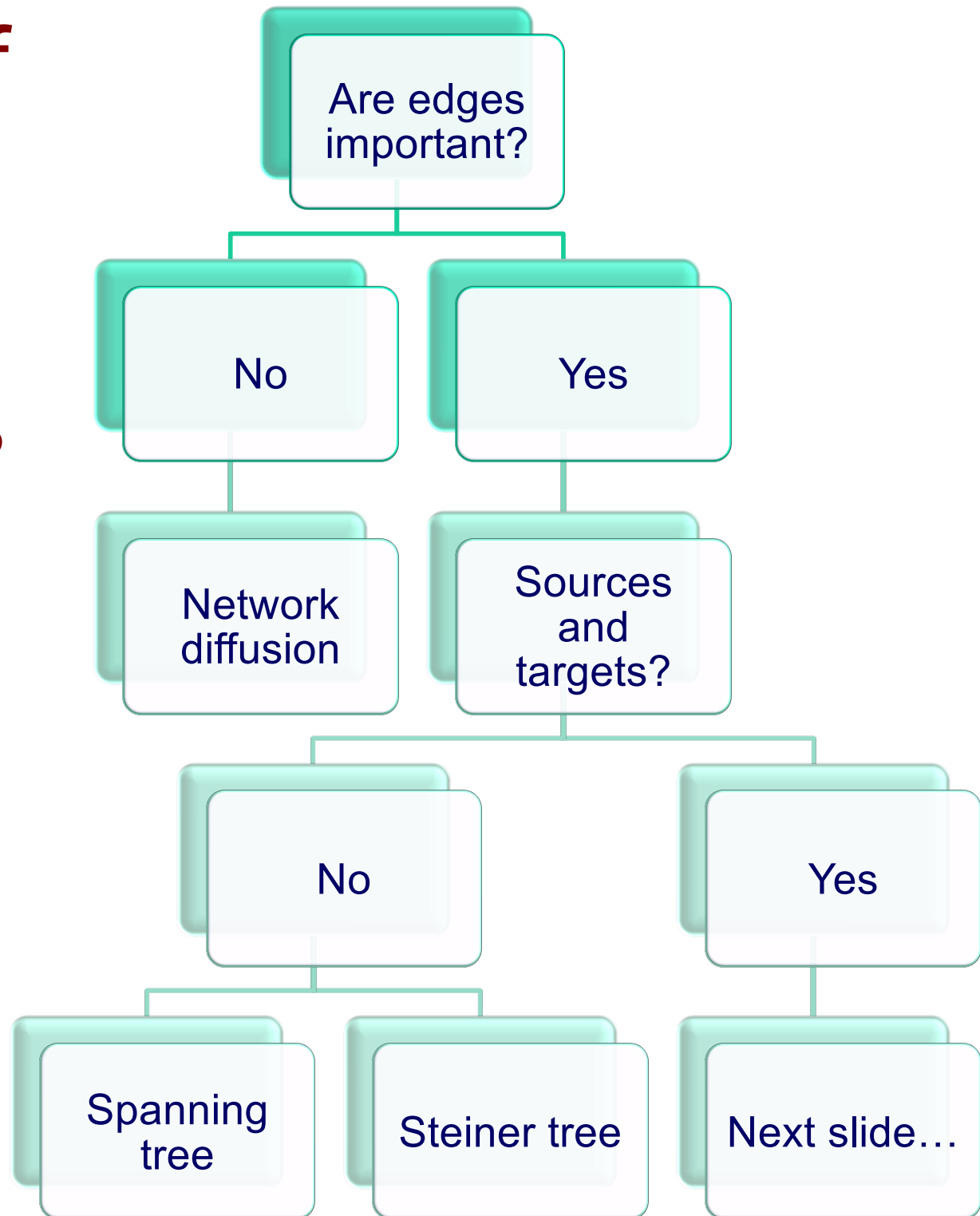
- Chilibot

- Our studies reveal a novel mechanism in which phosphorylation of STAT3 is mediated by a constitutively active JNK2 [MAPK9] isoform, JNK2 [MAPK9] $\hat{I}^{\pm}$. Ref: Oncogene, 2011, PMID: 20871632

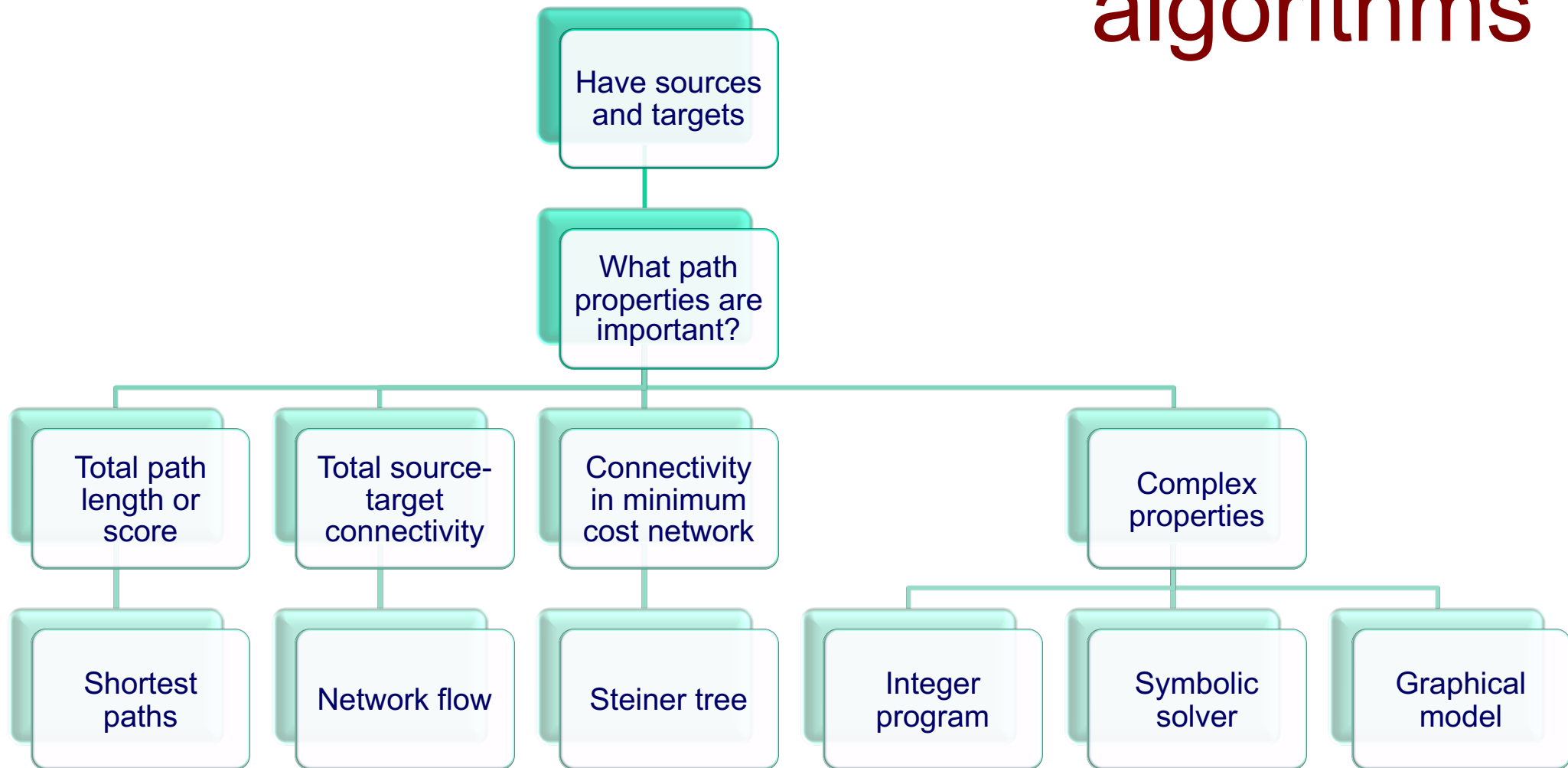
- iHOP

Akt1 ☆, but not Akt2, phosphorylates palladin ☆ at Ser507 in a domain that is critical for F-actin bundling. [2010]

Classes of pathway prediction algorithms



Classes of pathway prediction algorithms



Alternative pathway identification algorithms

- k-shortest paths
 - [Ruths2007](#)
 - [Shih2012](#)
- Random walks / network diffusion / circuits
 - [Tu2006](#)
 - eQTL electrical diagrams ([eQED](#))
 - [HotNet](#)
- Integer programs
 - Signaling-regulatory Pathway INference ([SPINE](#))
 - [Chasman2014](#)

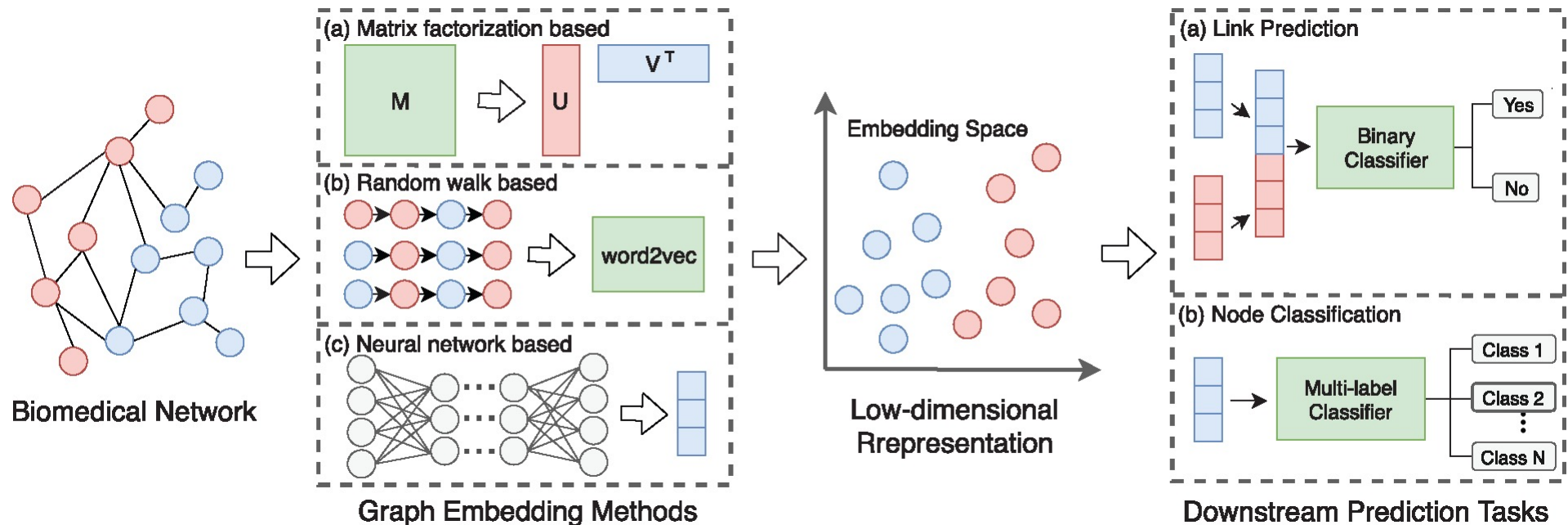
Alternative pathway identification algorithms

- Path-based objectives
 - Physical Network Models (PNM)
 - Maximum Edge Orientation (MEO)
 - Signaling and Dynamic Regulatory Events Miner (SDREM)
- Steiner tree
 - Prize-collecting Steiner forest (PCSF)
 - Belief propagation approximation (msgsteiner)
 - Omics Integrator implementation
- Hybrid approaches
 - PathLinker: random walk + shortest paths
 - ANAT: shortest paths + Steiner tree

Recent developments in pathway discovery

- Multi-task learning: jointly model several related biological conditions
 - ResponseNet extension: SAMNet
 - Steiner forest extension: Multi-PCSF
 - SDREM extension: MT-SDREM
- Temporal data
 - ResponseNet extension: TimeXNet
 - Steiner forest extension and ST-Steiner
 - Temporal Pathway Synthesizer

Graph embedding for biological networks



Condition-specific genes/proteins used as input

- Genetic screen hits (as causes or effects)
- Differentially expressed genes
- Transcription factors inferred from gene expression
- Proteomic changes (protein abundance or post-translational modifications)
- Kinases inferred from phosphorylation
- Genetic variants or DNA mutations
- Enzymes regulating metabolites
- Receptors or sensory proteins
- Protein interaction partners
- Pathway databases or other prior knowledge