Efficient Network Inference using a Linear Programming Approach

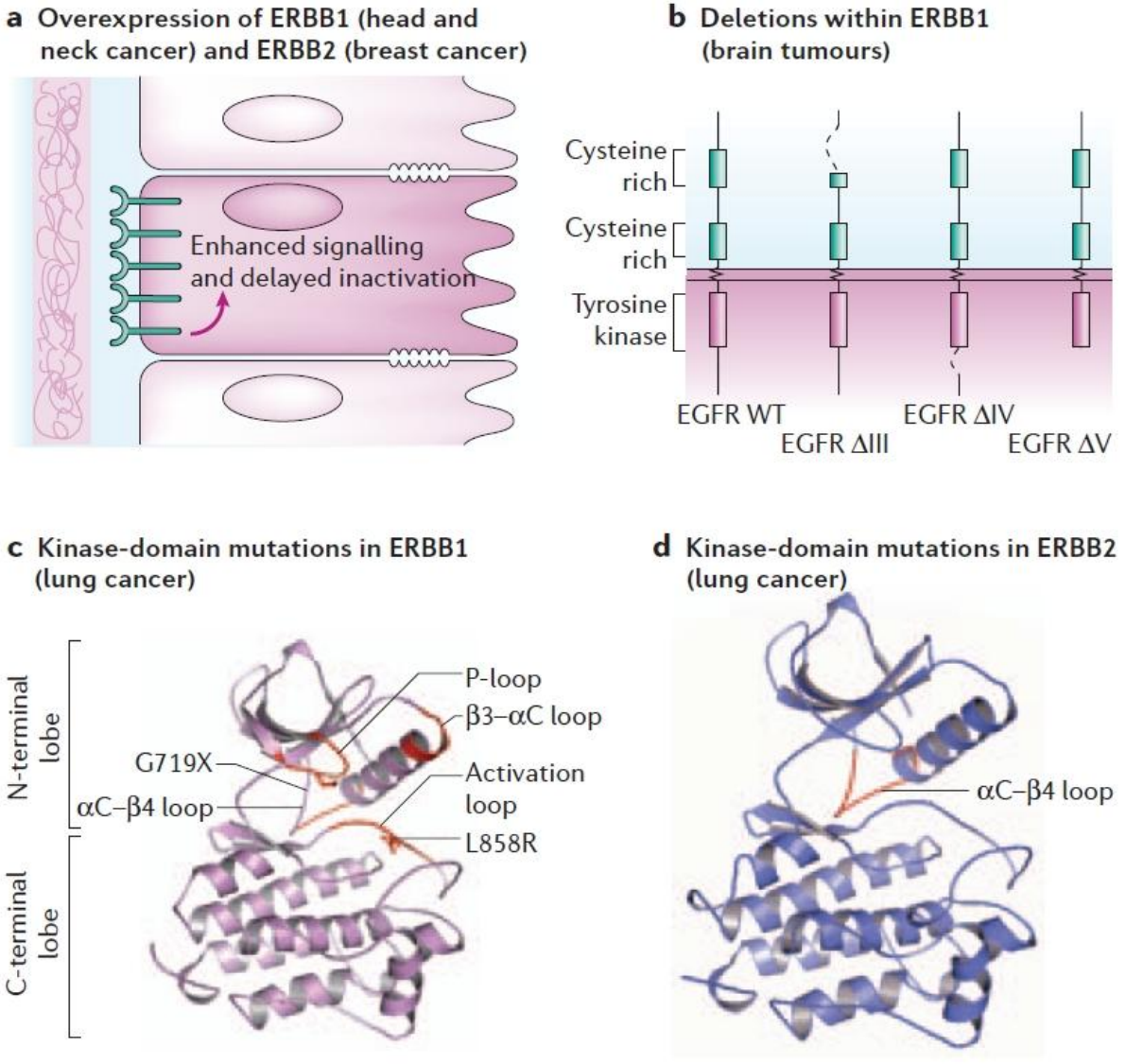
Bettina Knapp, Johanna Mazur, Lars Kaderali

Institute for Medical Informatics and Biometry,
Medical Faculty Carl Gustav Carus,
Dresden University of Technology

Basel, 08.09.2011

ERBB signaling pathway

Multiple pathways to oncogenesis

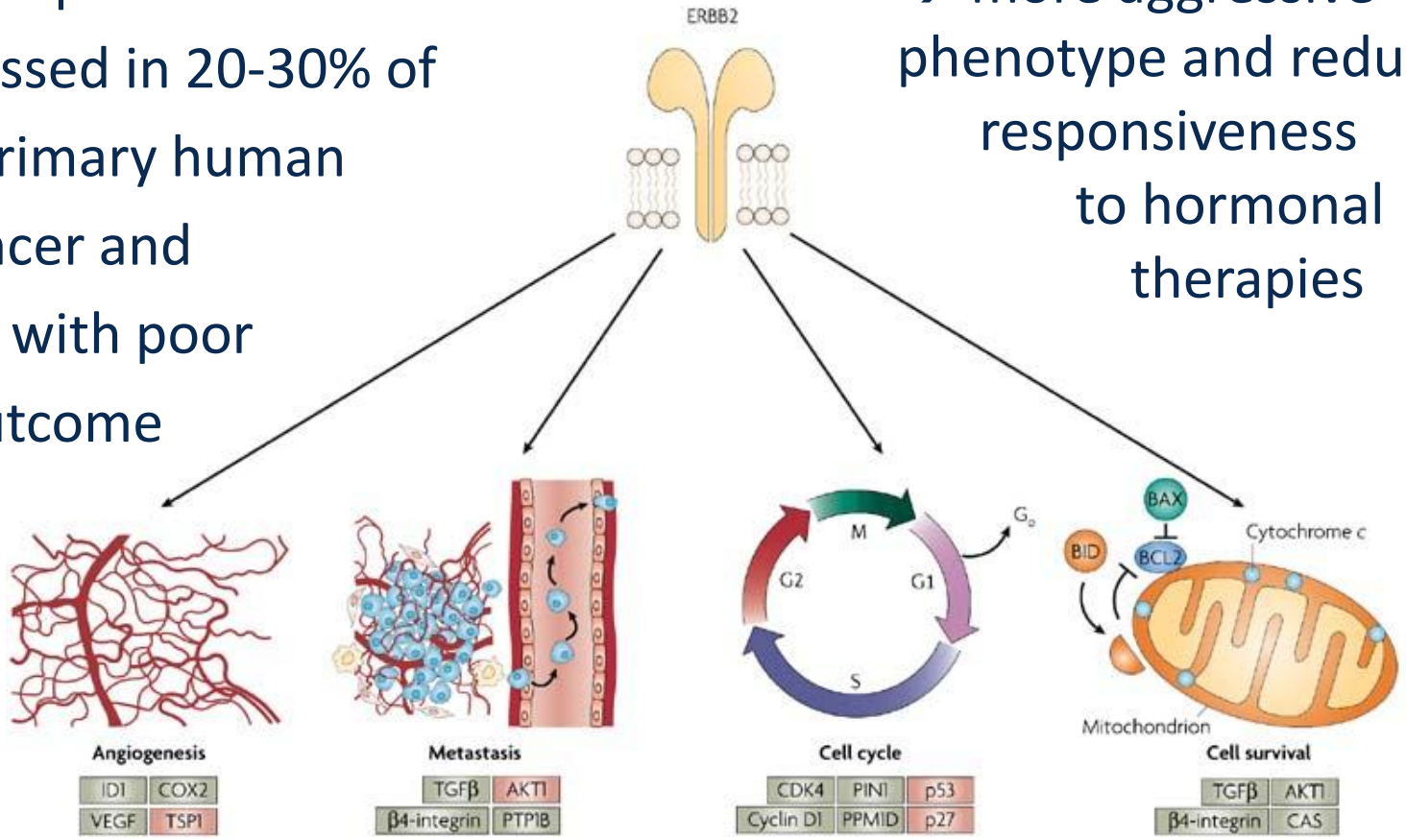


Citri, Yarden; *Nature Reviews, Molecular Cell Biology, 2006*

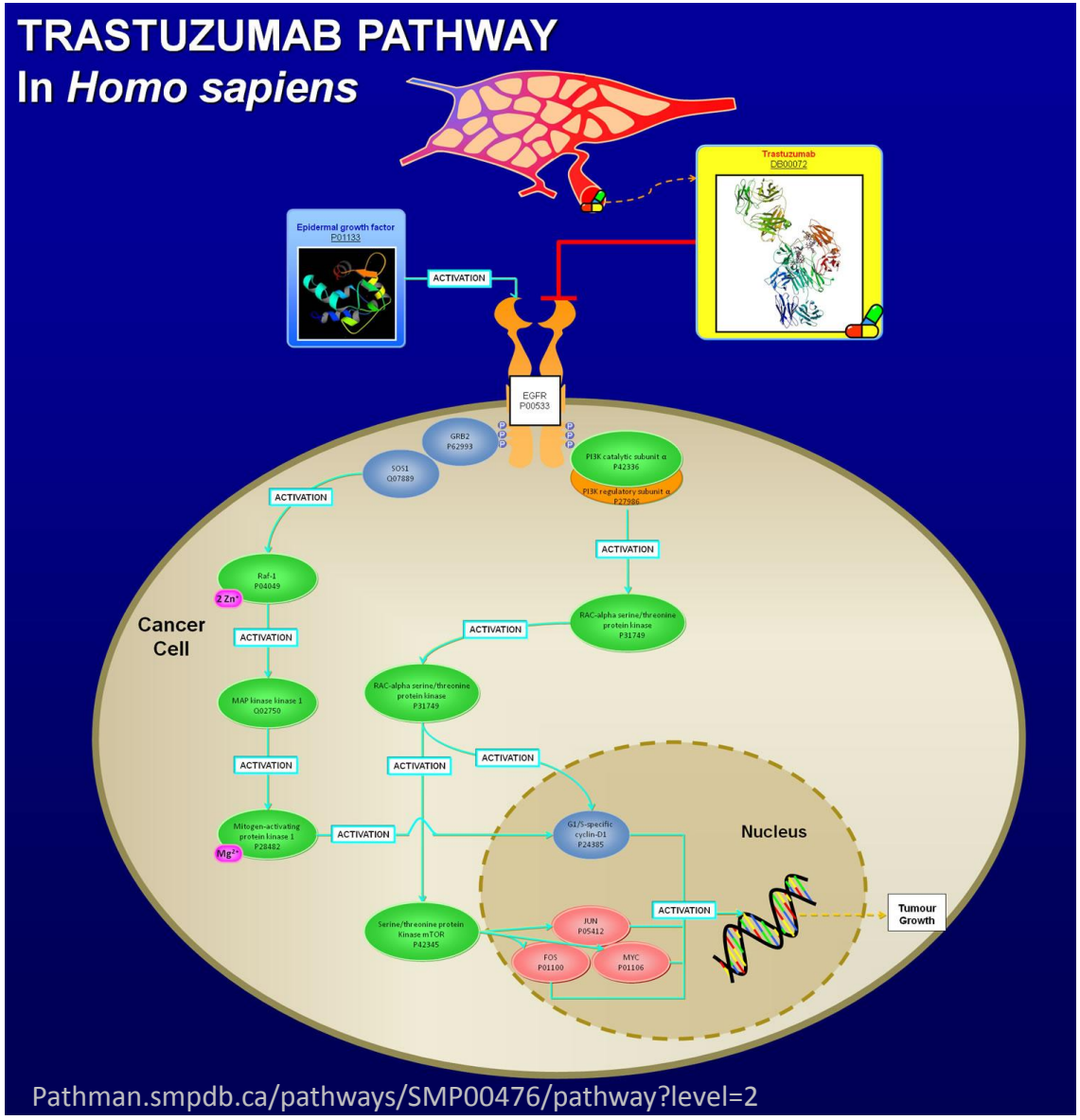
ERBB signaling pathway

ERBB2 is amplified and overexpressed in 20-30% of cases of primary human breast cancer and correlates with poor patient outcome

→ more aggressive phenotype and reduced responsiveness to hormonal therapies



Ursini-Siegel et al., 2007 Nature Reviews | Cancer



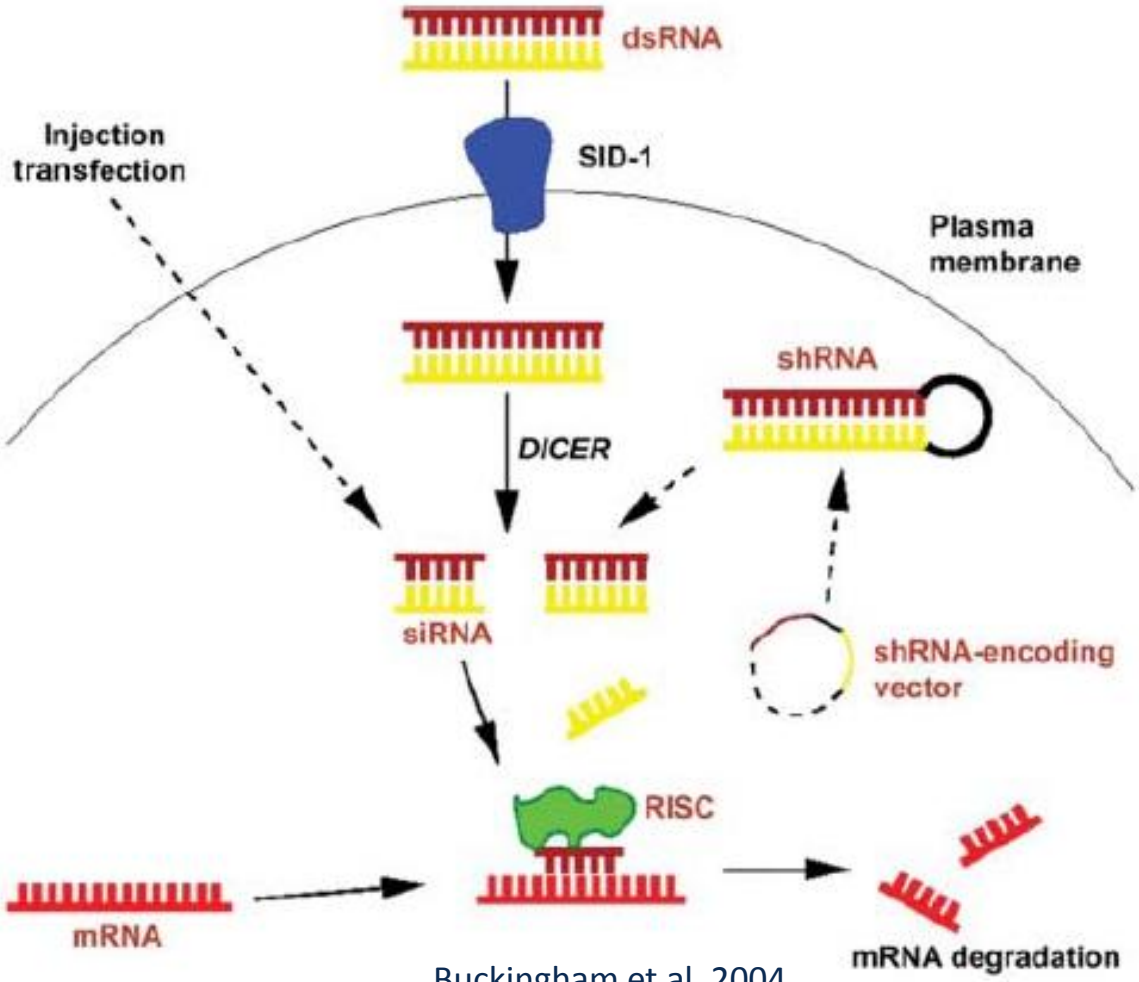
Pathman.smpdb.ca/pathways/SMP00476/pathway?level=2

- Treatment of breast cancer patients with trastuzumab, a neutralizing monoclonal ERBB2 antibody, prolongs the disease-free survival and improves the clinical outcome for breast cancer patients
- However, at least 2/3 of the patients are *de novo* resitant, mechanisms are poorly understood

→ New targets are necessary

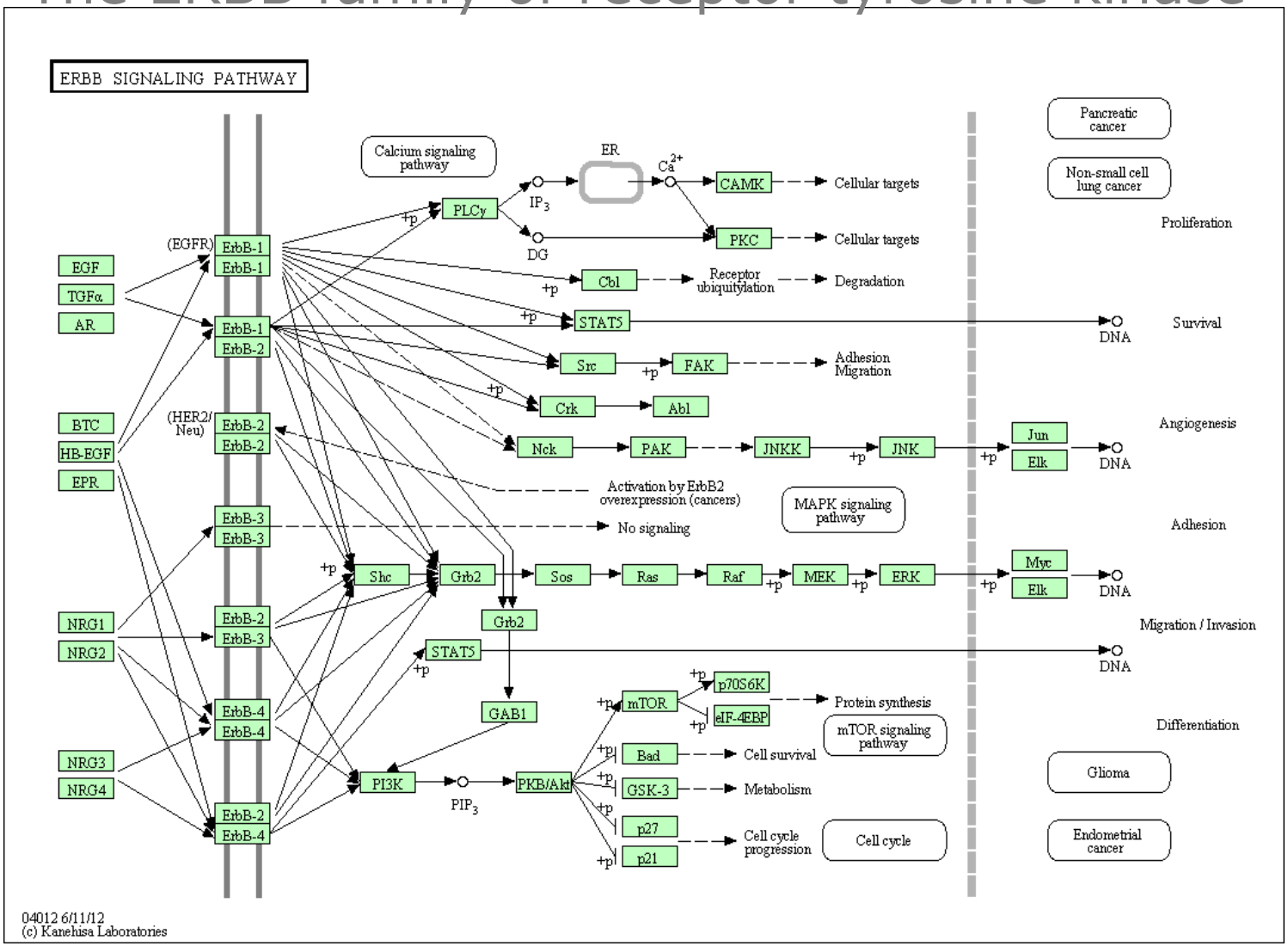
RNA interference (RNAi)

- Knockdown experiments to elucidate gene function and gene involvement in biological processes
- Identification of hit genes which play a role in a certain disease



Buckingham et al. 2004

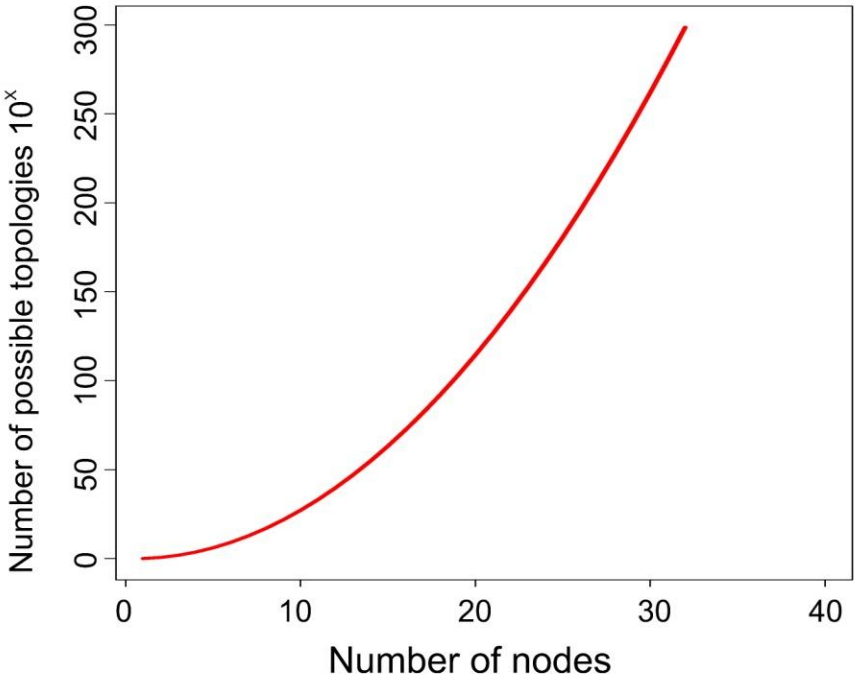
The ERBB family of receptor tyrosine kinase



Network Inference

Problem: for a directed graph with n nodes there are $2^{n(n-1)}$ possible network topologies:

i.e. for 17 nodes: $7.59*10^{81}$ possible networks.

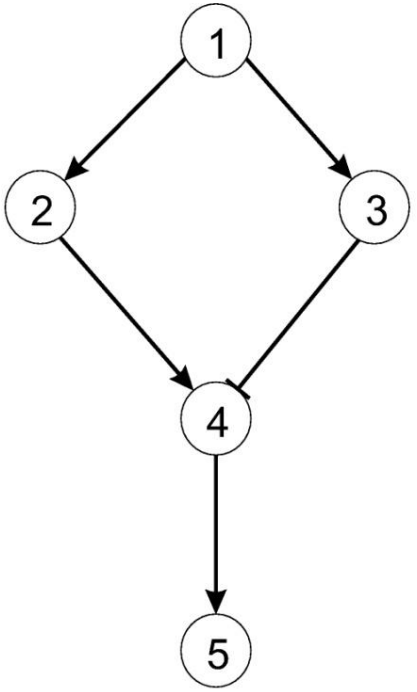


Network Inference

Solution: use linear programming (LP) → can be solved efficiently even for large-scale problems.

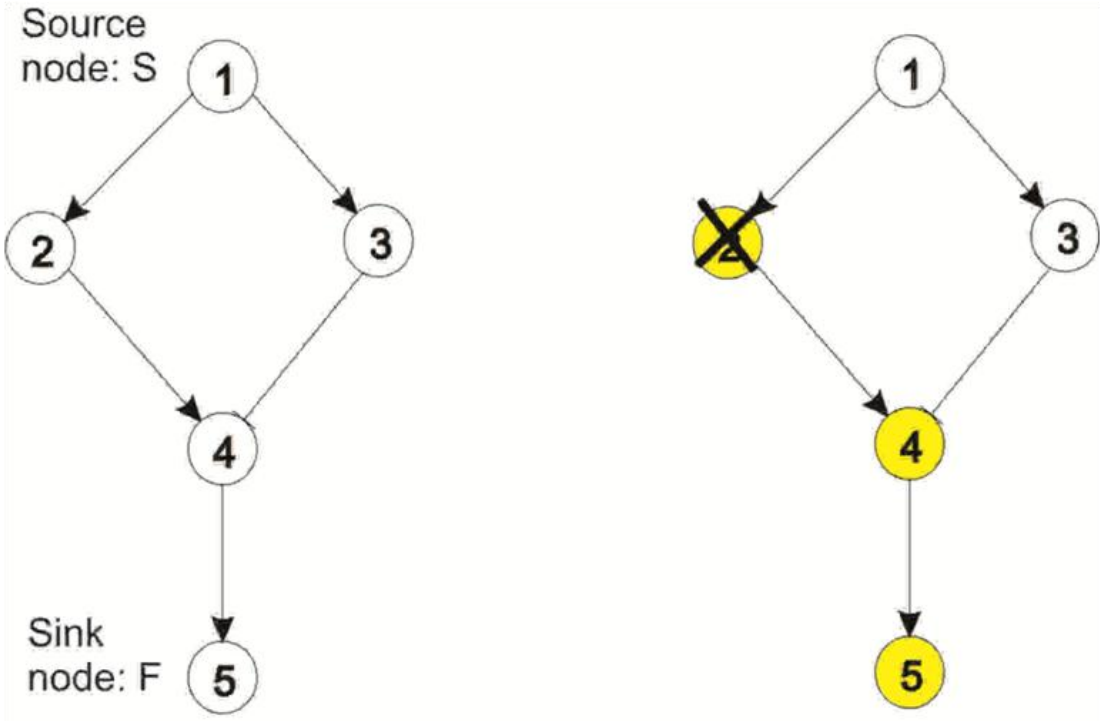
Knockdown of	Effect on		
	Gene 1	Gene 2	Gene 3
Gene 1	0	0	0
Gene 2	1	0	1
Gene 3	1	1	0
Gene 2 & 3	1	0	0

$$\min \; z(x) := c^T x$$
$$s.t. \; Ax \leq b$$
$$x \geq 0$$



LP Model assumptions

- Information flow starts at a source node *S* and ends at a sink node *F.*
- Each perturbation effect is propagated along the network.



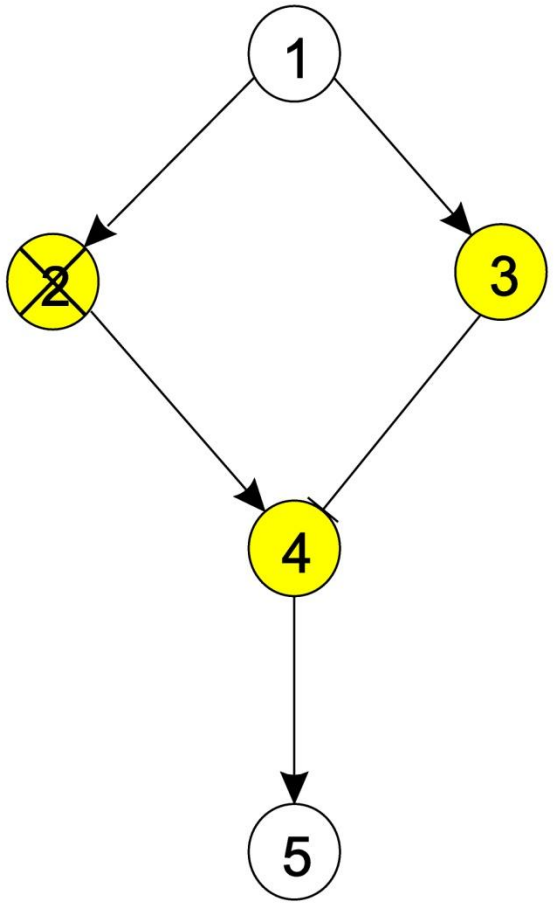
LP Model constraints

If $x_{ik} \geq \delta_i$ and $b_{ik} = 1$:

$$w_i^0 + \sum_{j \neq i} (w_{ji}^+ - w_{ji}^-) x_{jk} \geq \delta_i$$

If $x_{ik} < \delta_i$ and $b_{ik} = 1$:

$$w_i^0 + \sum_{j \neq i} (w_{ji}^+ - w_{ji}^-) x_{jk} \leq 0 + \xi_l$$



LP-SF Model

$$\min \; z(w_{ji}^{+}, w_{ji}^{-}, w_{i}^{0}, \xi_l) := \left(\sum_{i,j} (w_{ji}^{+} + w_{ji}^{-}) + \sum_{i} w_{i}^{0} + \frac{1}{\lambda} \sum_{l} \xi_l \right)$$

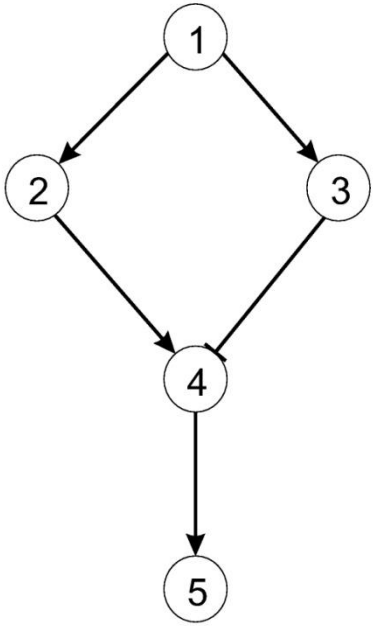
$$s.t.$$

$$\text{if } x_{ik} \geq \delta_i \text{ and } b_{ik} = 1: \quad w_i^0 + \sum_{j \neq i} (w_{ji}^{+} - w_{ji}^{-}) x_{jk} \geq \delta_i$$

$$\text{if } x_{ik} < \delta_i \text{ and } b_{ik} = 1: \quad w_i^0 + \sum_{j \neq i} (w_{ji}^{+} - w_{ji}^{-}) x_{jk} \leq 0 + \xi_l$$

Determine penalty parameter λ :

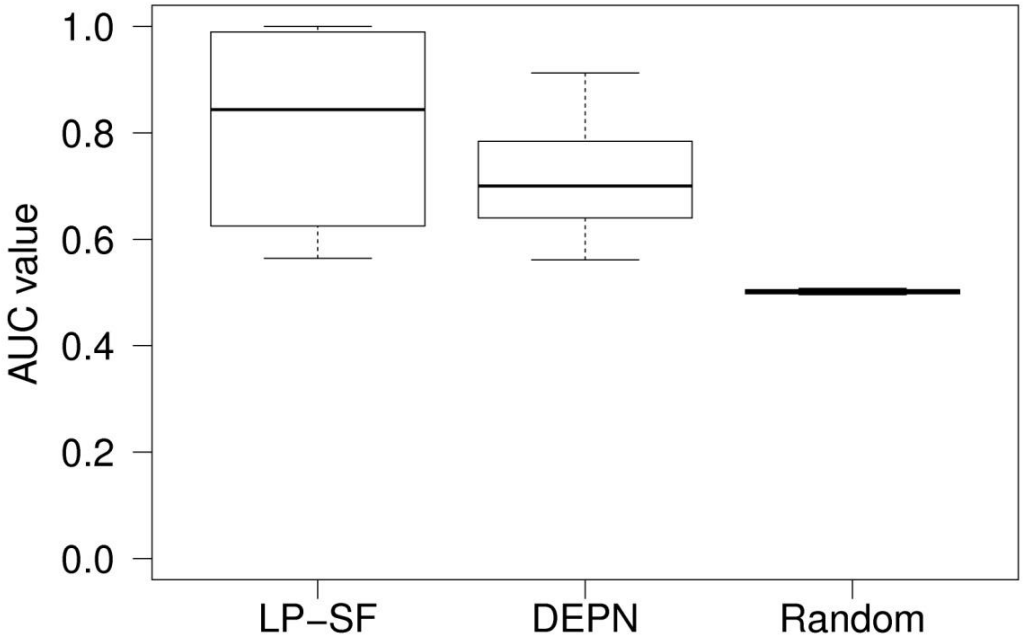
- Leave-one-out cross-validation and mean squared error.
- 10-fold stratified cross-validation for larger networks.



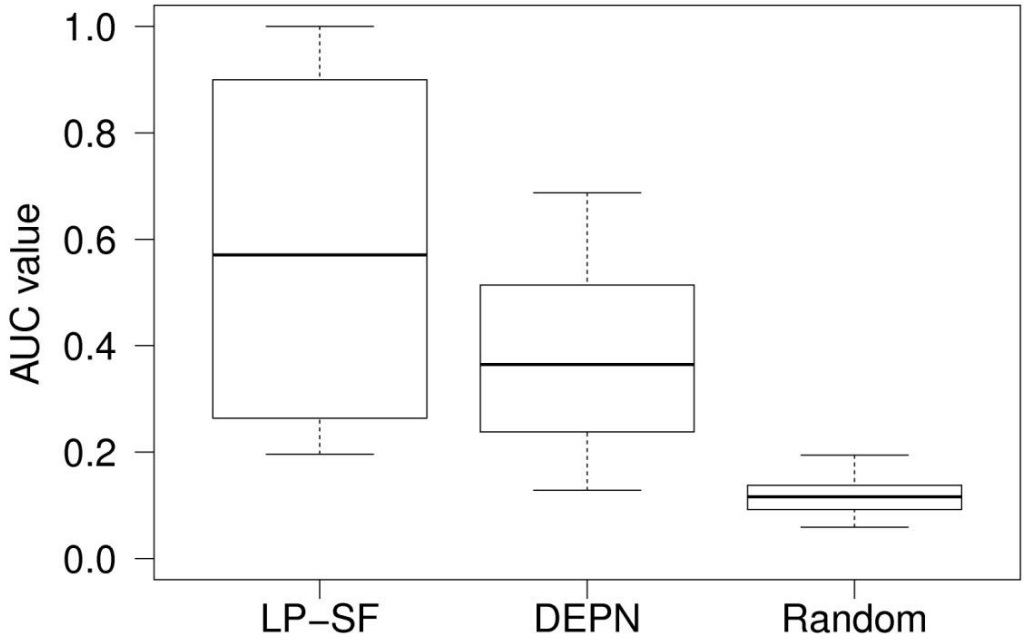
Simulations

- Activity of a node is computed from two normal distributions.
- Simulated single, double (randomly chosen) and one experiment without any knockdown.
- Computation of area under the ROC (AU-ROC) and area under the precision-to-recall curve (AU-PR).
- Different standard deviations to simulate noise and simulations for missing data.
- Using networks extracted from KEGG:
  - 10 sub-networks randomly selected with 10 nodes.
  - 5 random sub-networks of larger size.
- Comparison with an approach published by Froehlich et al. in 2009: Deterministic Effects Propagation Networks (DEPNs).

Results 10-node networks

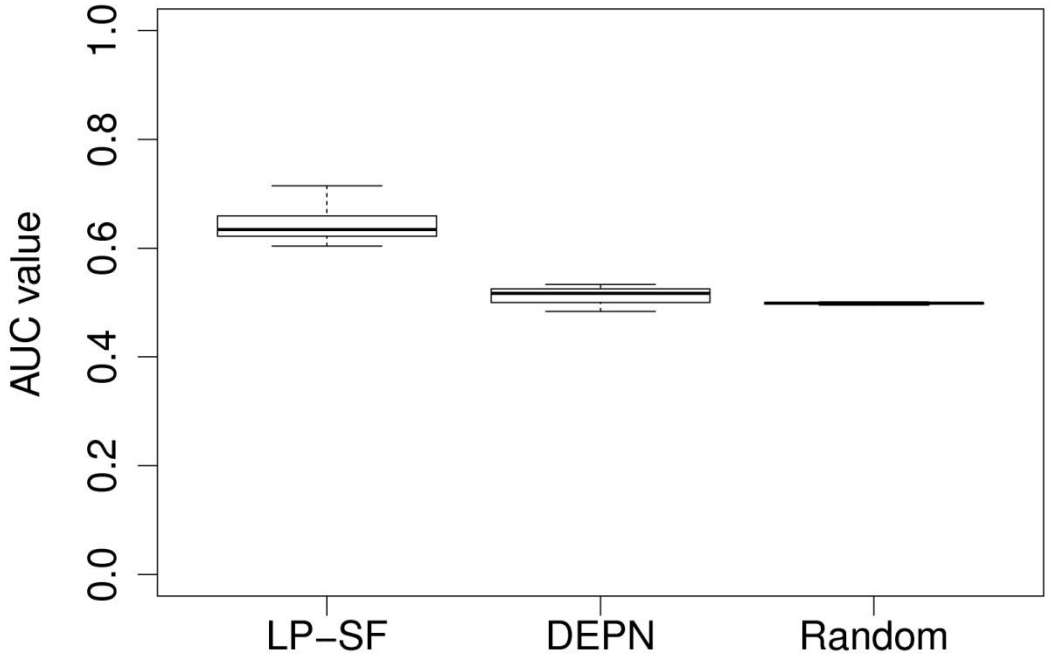


ROC

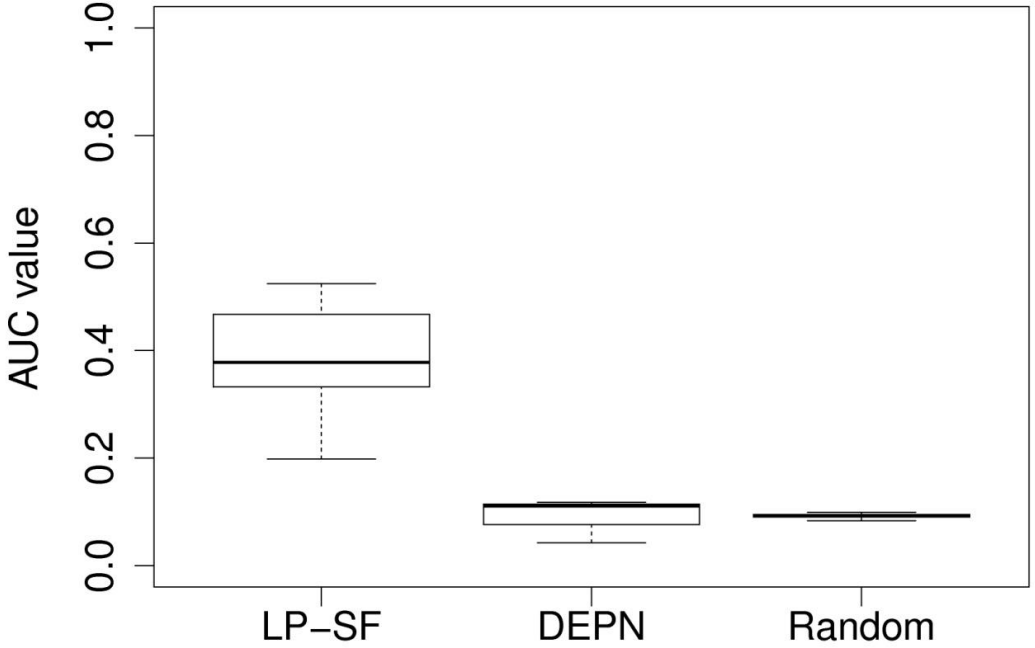


Precision-to-recall

Results large-scale networks

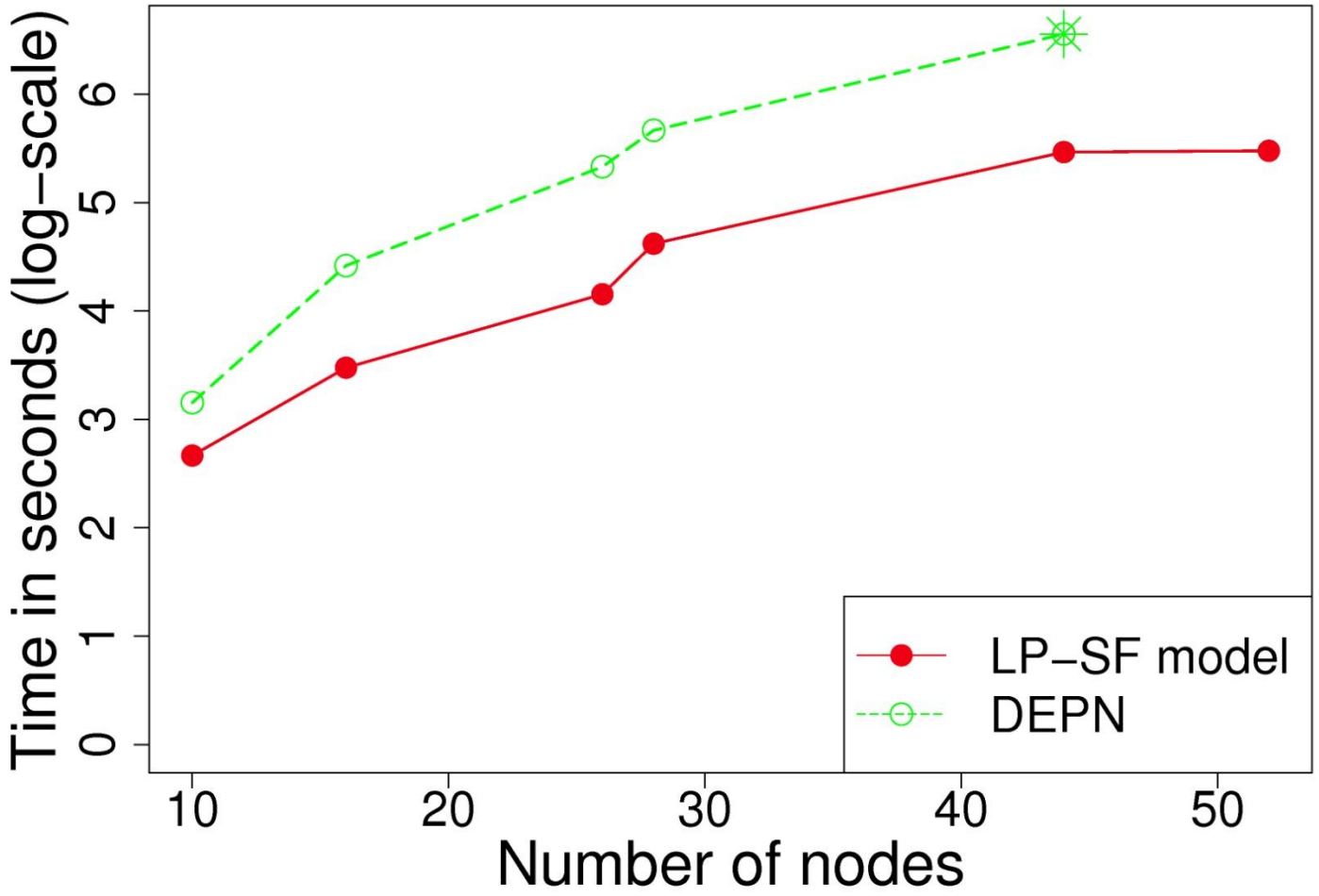


ROC

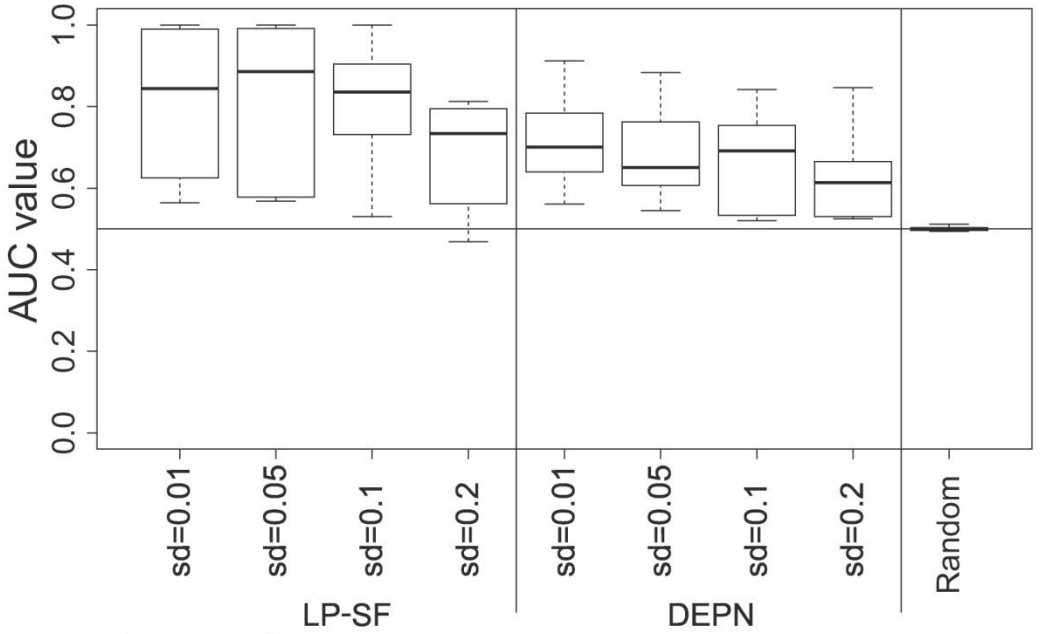


Precision-to-recall

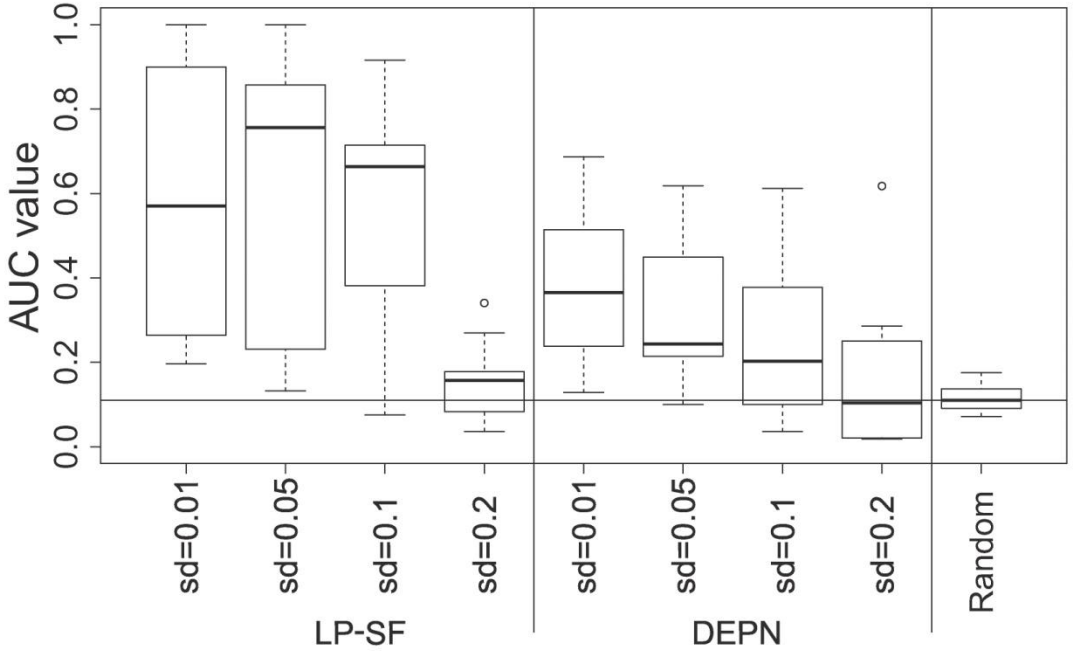
Run time


Time in seconds (log–scale)
0
1
2
3
4
5
6
10
20
30
40
50
Number of nodes
LP–SF model
DEPN

Noisy Data: 10 node-nw

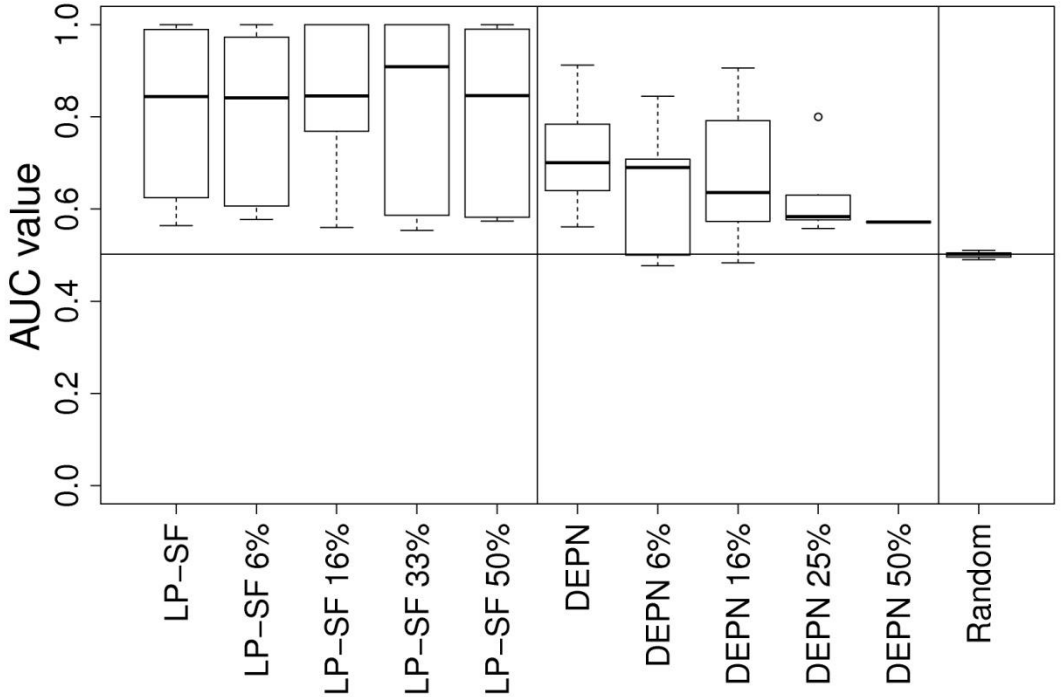


ROC

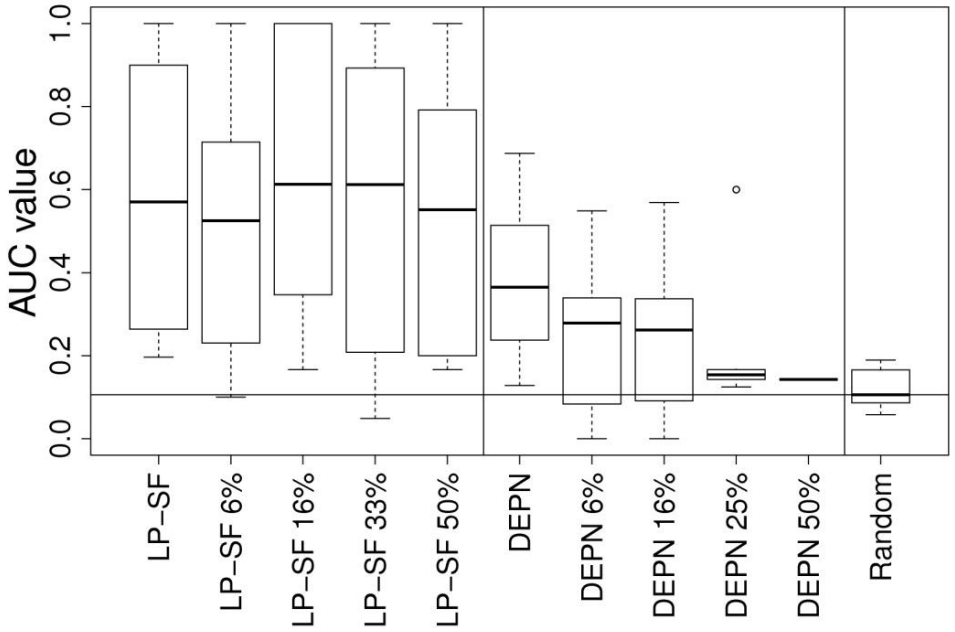


Precision-to-recall

Incomplete Data: 10 node-nw



ROC



Precision-to-recall

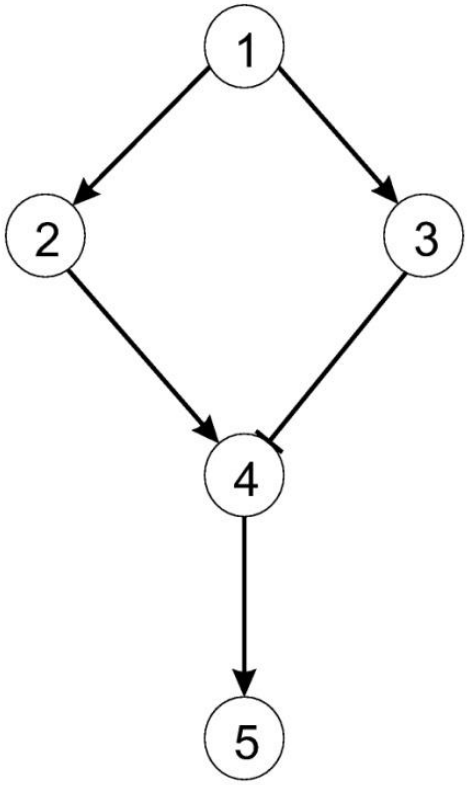
LP Model constraints if prior knowledge is available

If $i \in V \setminus S$:

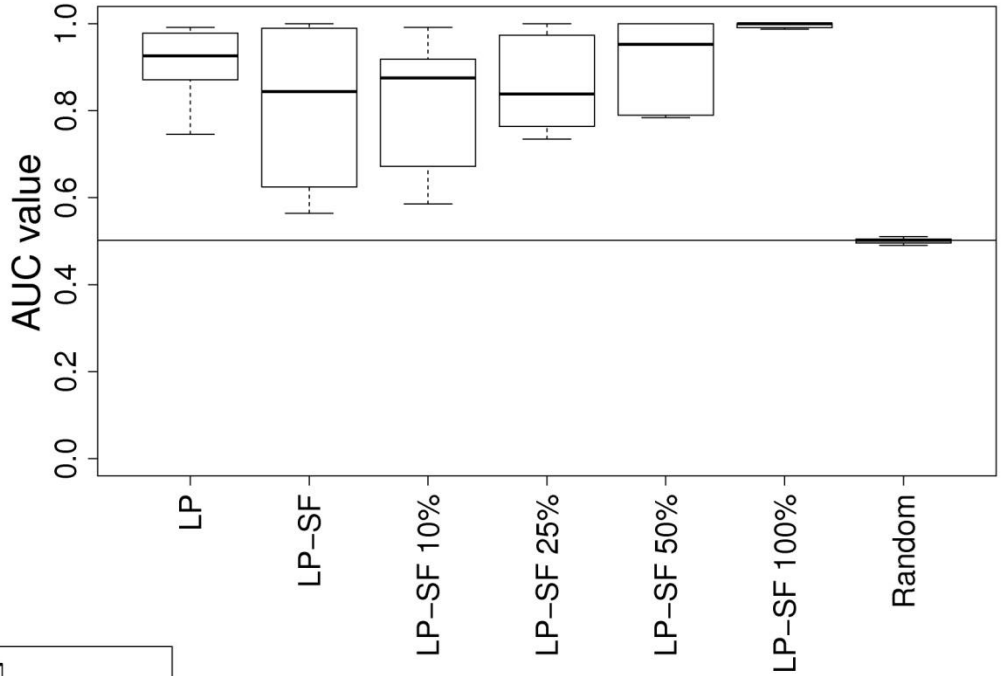
$$\sum_{j \in V, j \neq i} (w_{ji}^{+} + w_{ji}^{-}) \geq \delta_i$$

If $i \in V \setminus F$:

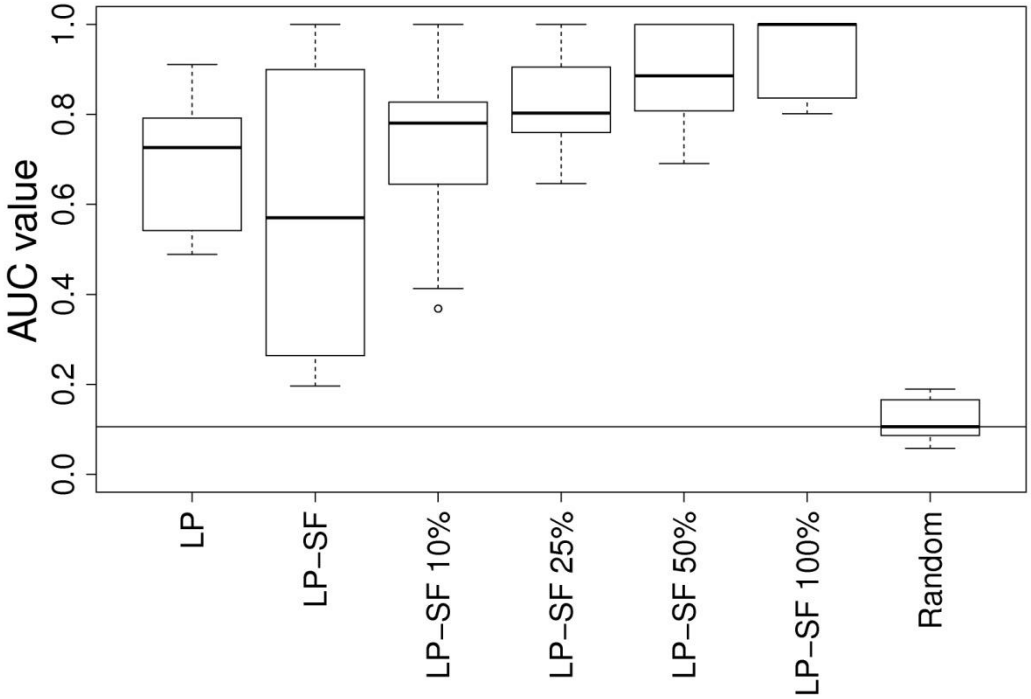
$$\sum_{j \in V, j \neq i} (w_{ij}^{+} + w_{ij}^{-}) \geq \delta_i$$



Prior knowledge integration: 10 node-nw



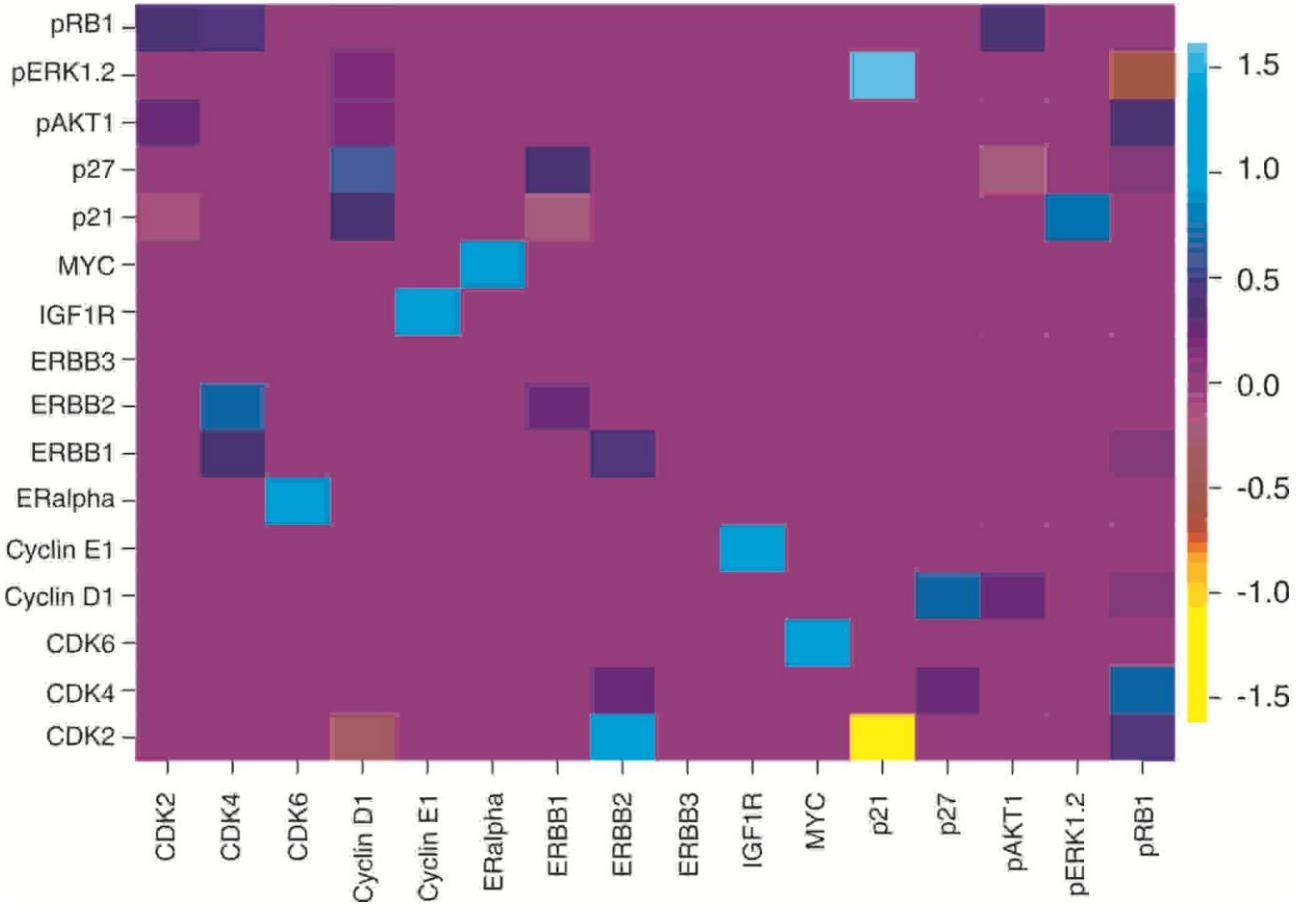
ROC



Precision-to-recall

ERBB signaling data

Normalized data is given for 16 genes of the ERBB signaling network (Froehlich et al., 2009): 16 kds (3 double kd).



ERBB signaling data

Evalutation of the results based on String database. Comparison with the DEPNs and random networks:

	LP model	DEPN	random
True positives	9	7	13.11
True negatives	72	73	59.11
False positives	15	14	27.9
False negatives	32	34	27.9
Specificity	0.83	0.84	0.68
Sensitivity	0.22	0.17	0.32
Precision	0.38	0.33	0.32
Accuracy	0.63	0.63	0.56

Summary

- Formulation as an LP allows an efficient computation.
- Model can include double (multiple) knockdowns.
- Inferred edges are activating and deactivating.
- Prior knowledge can be easily incorporated but is not essential for the network inference.

Problems and Open Questions

- Nonlinearities → topologies are not always connected
- Only steady-state data → loops not fully resolved

Thank you for your attention!

Bettina Knapp

Institute for Medical Informatics and Biometry,
Dresden University of Technology

Bettina.knapp@tu-dresden.de
http://141.76.248.53/homepages/knapp

Acknowledgments

- Prof. Dr. Lars Kaderali,

Medical Faculty Carl Gustav Carus, TU Dresden

- Dr. Johanna Mazur,

University Medical Center of the Johannes Gutenberg University Mainz