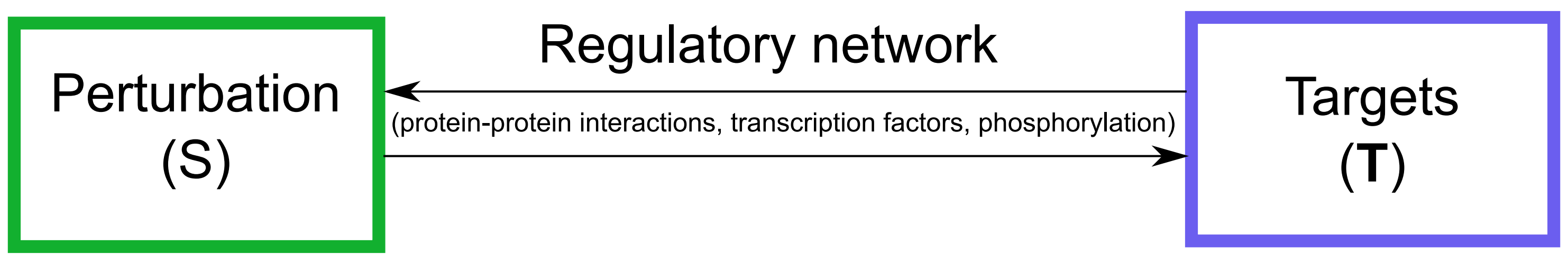
Predicting causes and effects in regulatory networks

Marc Hulsman, Marcel J.T. Reinders
Delft Bioinformatics Lab, Faculty of Electical Engineering, Mathematics and Computer Science
Delft University of Technology

Problem description

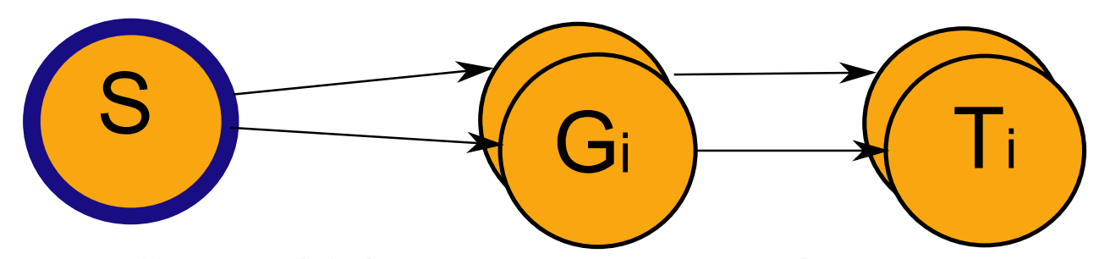
Connecting regulators with the genes that they regulate is difficult. While various large scale data sources are nowadays available (expression, literature, chip-chip, protein interactions, etc.), each of these is in itself an incomplete view of the regulatory network.

Our goal is to integrate them. We build a network that is able to predict how a perturbation in a gene S will affect other genes **T**.

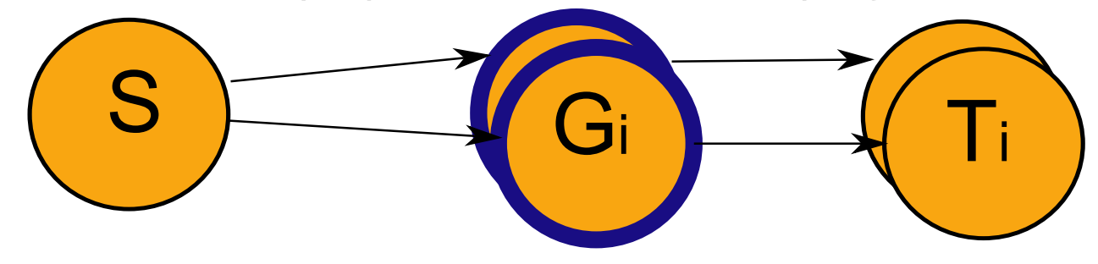

Perturbation (S)
Regulatory network
(protein-protein interactions, transcription factors, phosphorylation)
Targets (**T**)

Using this predictor, we want to be able to:

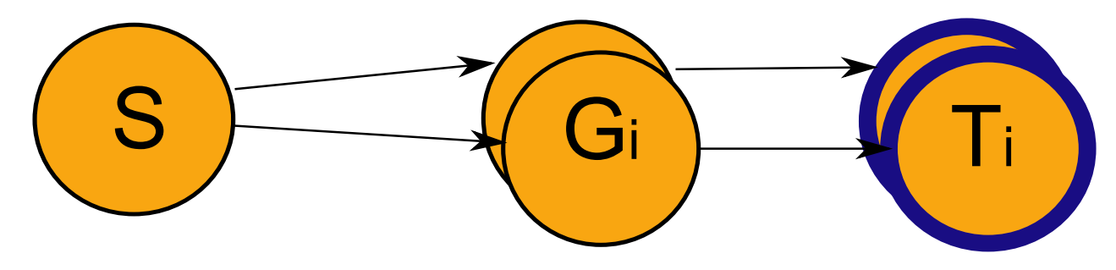
- given measured effects (T_i), determine most likely causal perturbation (S)



- identify possible unmeasured effects (G_i) in the network (e.g. protein state changes)

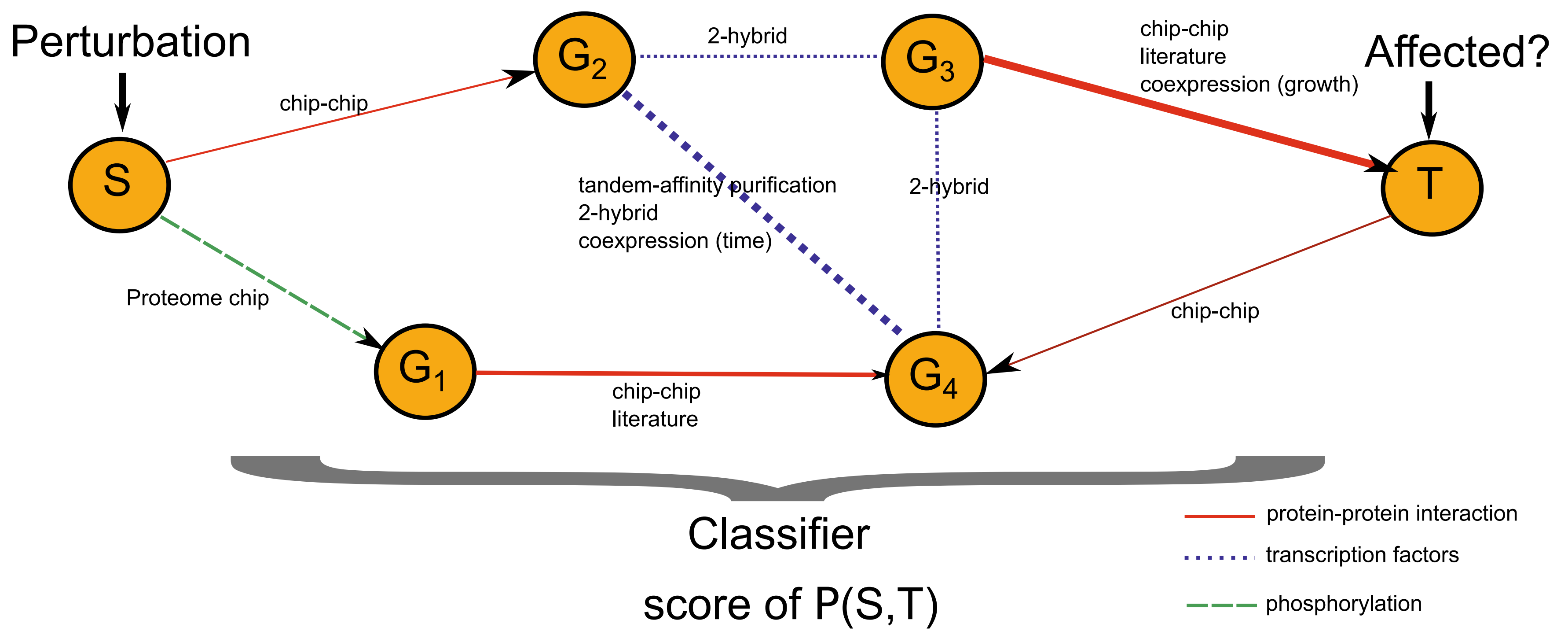


- find genes which form a robust group of markers (T_i) for a certain perturbation (S)



Approach

Various (high-througput) genome-scale data sources, containing information on different types of cellular networks, are used to determine the connectivity between source and target gene:

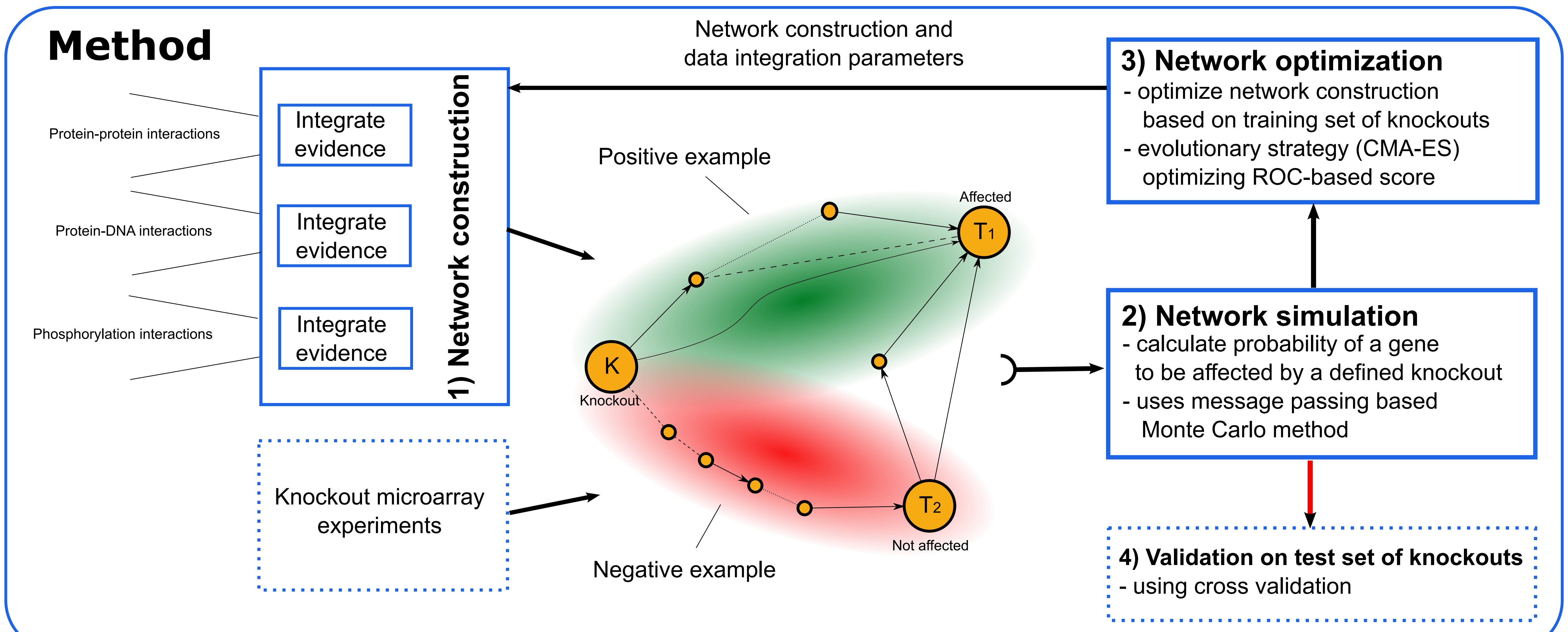


Classifier
score of P(S,T)

We train a predictive rule which takes the connectivity data between S and T into account to determine a score for P(S,T), describing how likely it is that a perturbation of S will affect T.

This rule is trained by using knockout microarray expression data, for which we both know the perturbed gene as well the genes that show an effect.

Method


Network construction and data integration parameters

Protein-protein interactions
Protein-DNA interactions
Phosphorylation interactions

Integrate evidence
Integrate evidence
Integrate evidence

1) Network construction

Knockout microarray experiments

3) Network optimization
- optimize network construction based on training set of knockouts
- evolutionary strategy (CMA-ES) optimizing ROC-based score

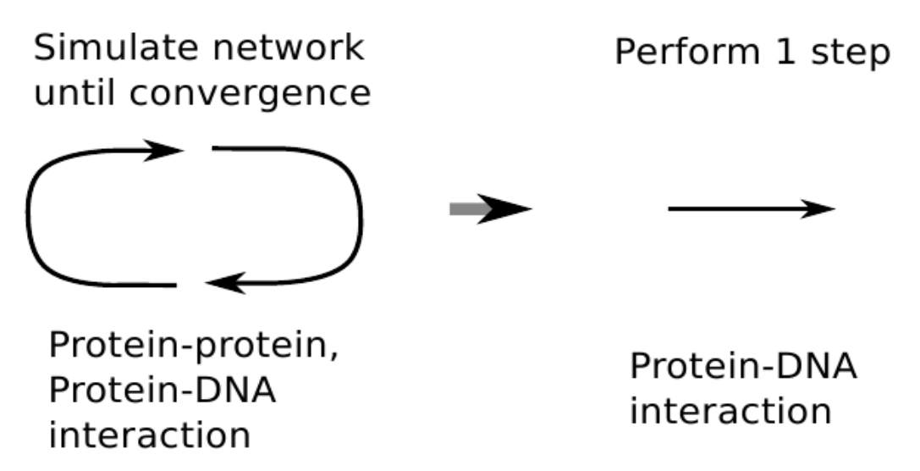
2) Network simulation
- calculate probability of a gene to be affected by a defined knockout
- uses message passing based Monte Carlo method

4) Validation on test set of knockouts
- using cross validation

Validation

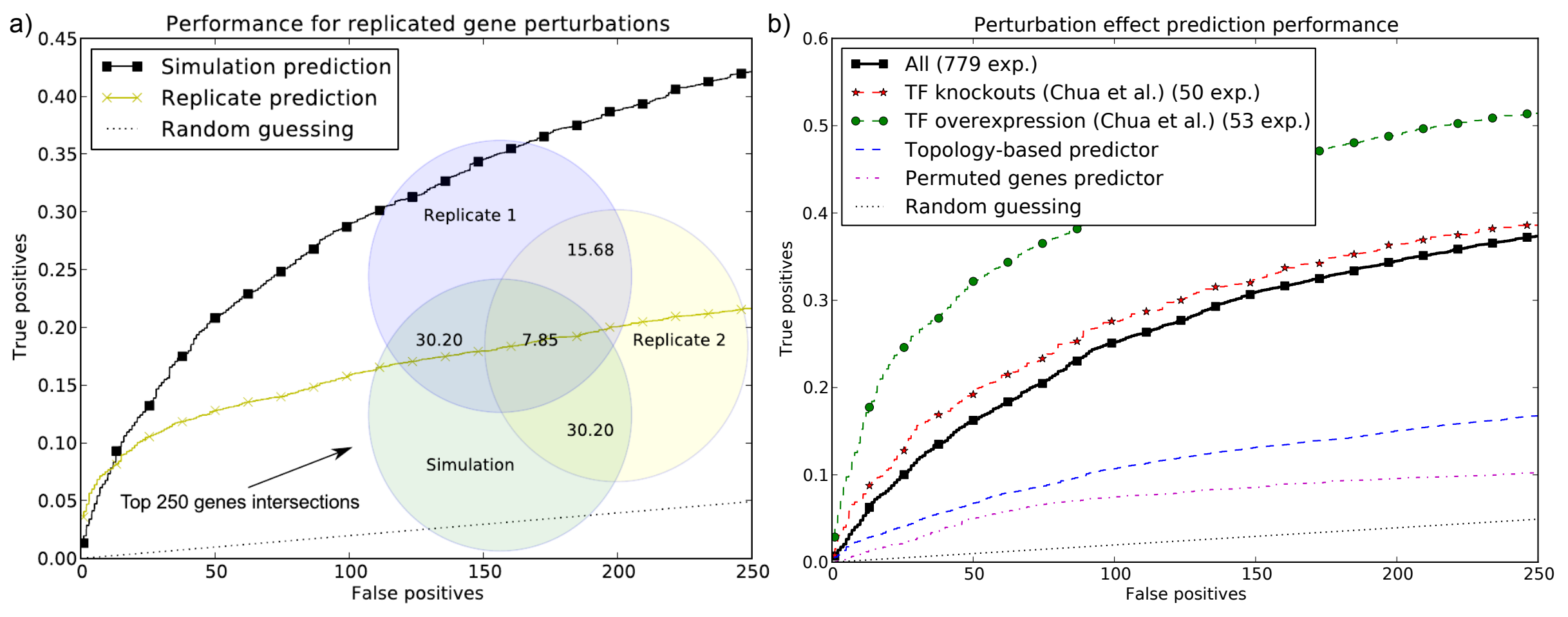
We perform in silico-validatioin. We leave a set of gene perturbation microarray experiments out of the data that is used for making predictions. Next, we check if the predictions that we make are validated by this left-out data

The prediction method can handle path constraints, e.g. only considering effect propagations that end in a protein-DNA interaction (i.e. expression-changing). This allows for training/validation using microarray results.


Simulate network until convergence
Perform 1 step
Protein-protein, Protein-DNA interaction
Protein-DNA interaction

Results

Using 10-fold cross-validation, the algorithm was validated on unseen gene perturbation microarray experiments.

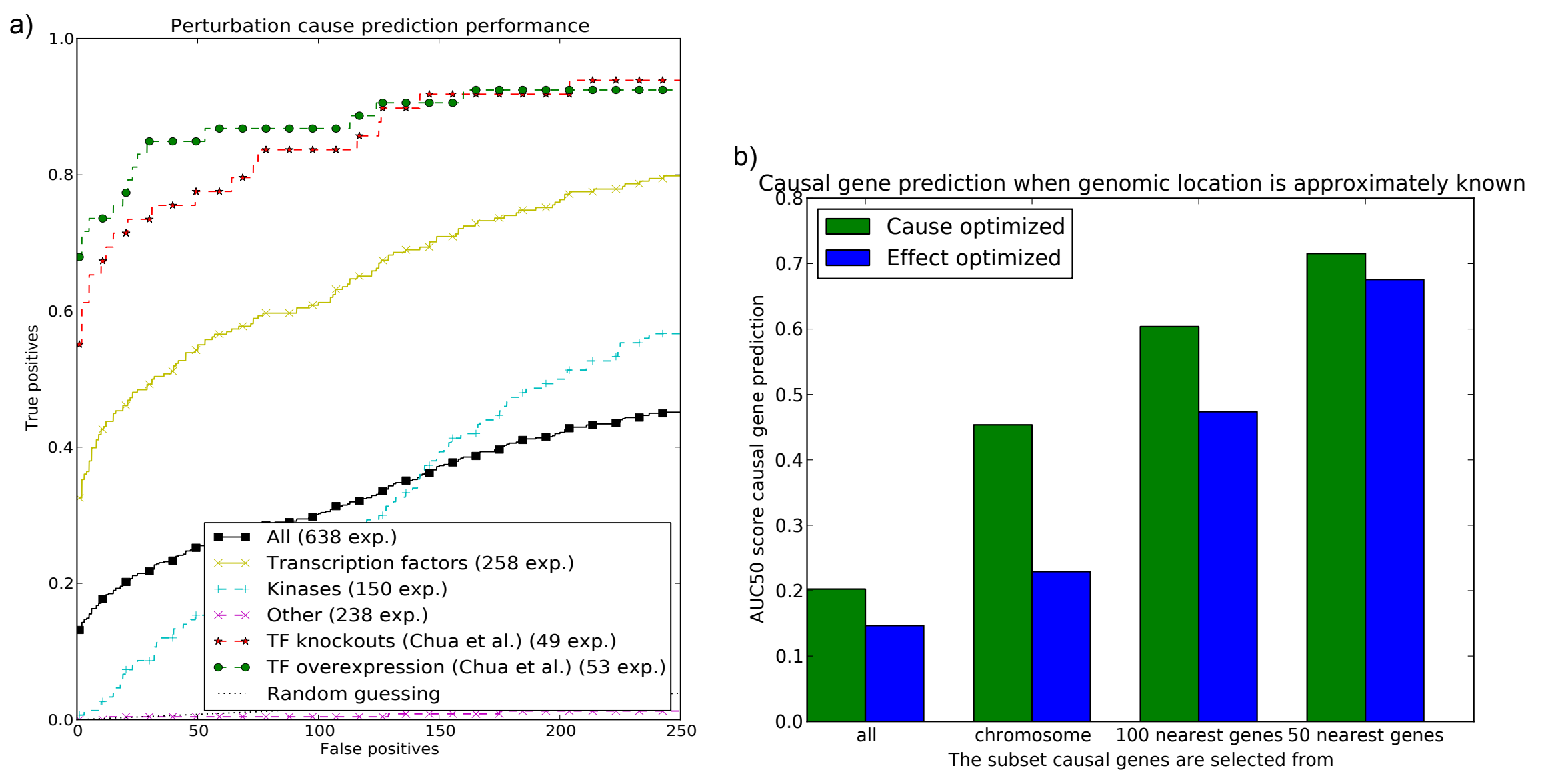


Perturbation effect prediction

a) A significant fraction of the measured effects can be predicted, from among all genes in the yeast network. Surprisingly, the network simulator is a much better predictor of effects, compared to replicate perturbations (i.e perturbations of the same gene obtained from different datasets) being used to predict the effects of each other.

b) Similarly, a predictor based just on high quality network topology (shortest paths) also performs worse compared to the network simulation. That the network topology still plays a significant role is however shown by the low performance of the *permuted* network simulation, for which the genes were permuted across the network.

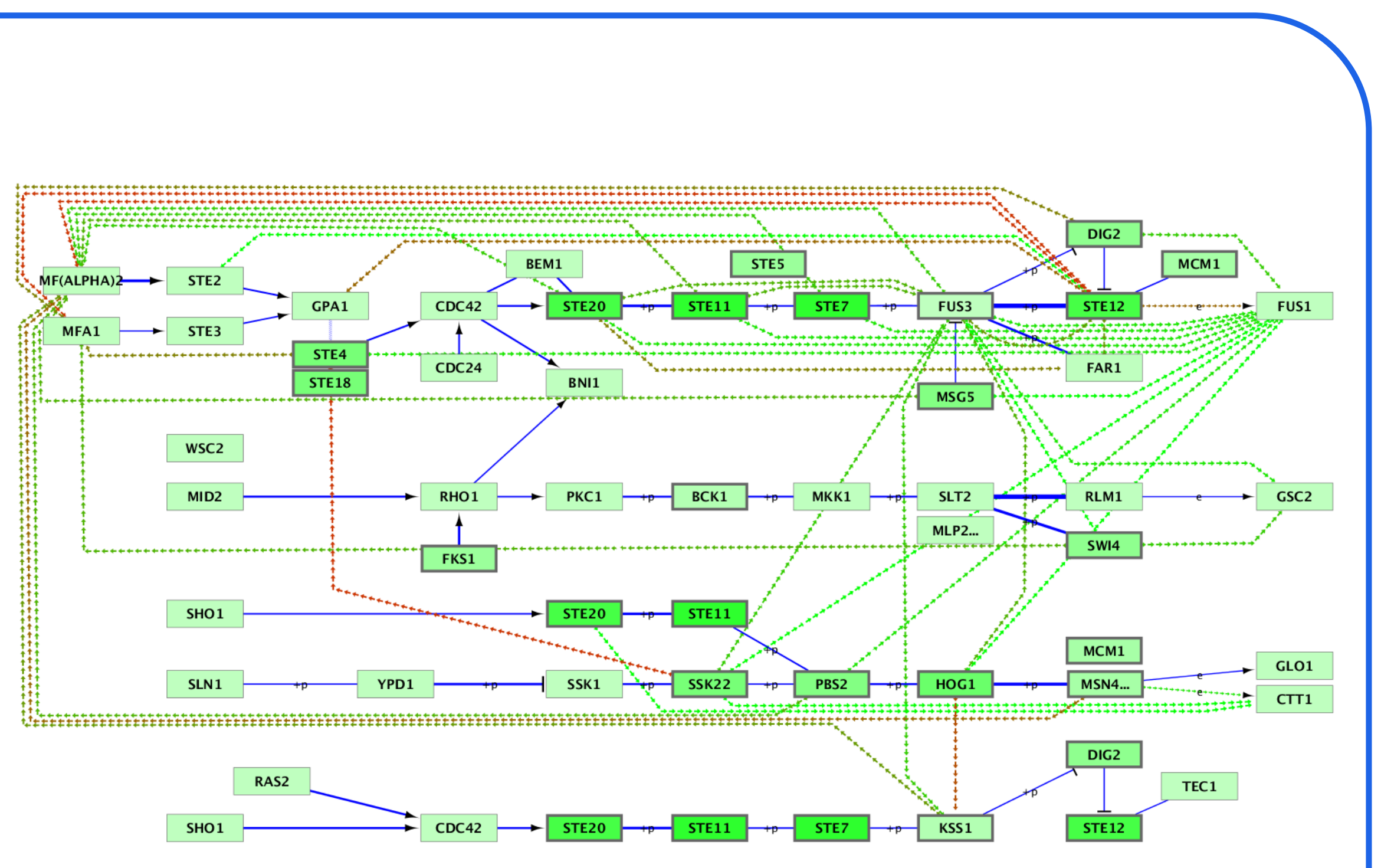
Part of the non-predicted effects are due to indirect effects such as stress response. For instance, the most often affected gene in the knockout measurements is HSP30 (a stress response gene) . This is corroborated by gene overexpression effects being easier to predict then gene knockout effects.



Perturbation cause prediction

a) Based on observed effects, the method can predict the most likely causal genes. We find that regulatory causal genes are much easier to predict compared to non-regulatory genes, likely due to the fact that only regulatory effects are encoded in the network topology.

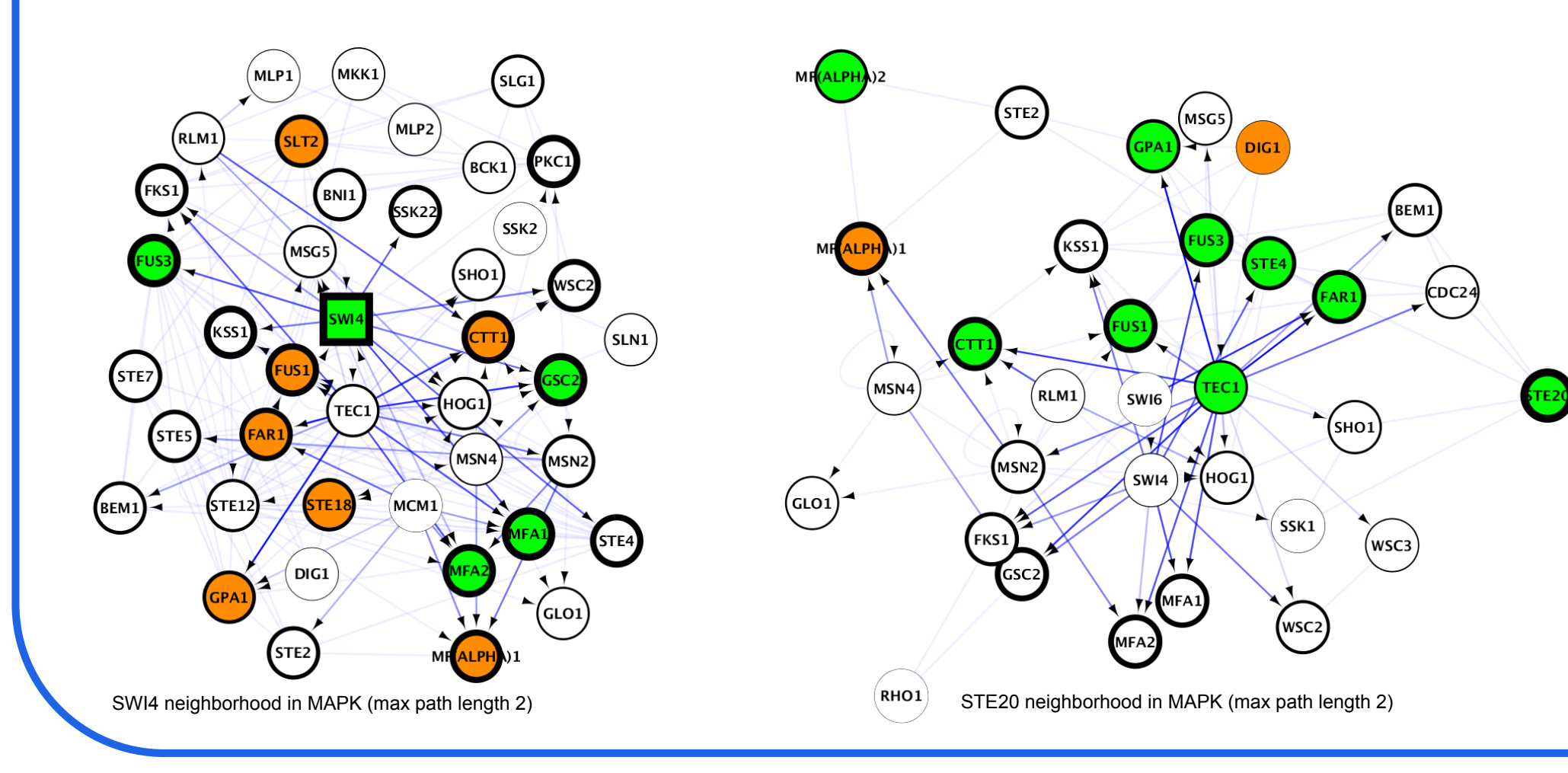
b) A common usage pattern for these type of predictions will be one in which causal genes have to be determined from among a certain subselection of genes (e.g. genes affected by mutations). In these cases, performance can become significantly higher. Also, in this plot we show the difference in performance for a network optimized for causal prediction, and a network optimized for effect prediction.



Effect prediction within the yeast MAPK pathway

Effect predictions in the KEGG MAPK pathway. Blue lines indicate pathway edges, and their thickness indicates regulatory influence (inferred by the algorithm). Arrow lines show predictions (i.e. ranking in top 250) that were validated (p-value < 0.001), with the color indicating the number of false positives ranking higher within the MAPK pathway (ranging from 0 to 8, 69% <= 3).

Note that the algorithm predicts *feed-forward*, *feedback* and even *cross-talk* effects accurately. This is possible by simulating the pathway not on its own, but within its complete cellular context.



SWI4 neighborhood in MAPK (max path length 2)
STE20 neighborhood in MAPK (max path length 2)

Dense neighborhoods

While pathway maps often suggest relatively simple interaction topologies, in reality, the number of interactions that have been measured among even a relatively limited number of genes is often rather large, complicating accurate perturbation effect predictions. This underlines the need for the use of multiple information sources to determine the regulatory activity of these interactions.

In the figures to the left, we show the neighborhood of two genes within the MAPK pathway. One is a transcription factor (SWI4), the other (STE20) is a kinase. The node color indicates if an effect has been observed in microarray studies, after perturbation of respectively SWI4 or STE20 (green: p-value < 0.001, orange: p-value < 0.05). The thickness of node borders indicate if the network simulation predicts an pertubation effect or not.

Only the highest quality interactions are shown (STRING score > 900), with their predicted regulatory activity being indicated by their color.

Discussion

Building genome-wide, simulation models of cellular networks is difficult, as learning the activity of each of the millions of possible interactions easily leads to large numbers of false positives.

We simplify the problem, by learning a more general rule on how to integrate various data sources, using them to determine the regulatory activity of the interactions, thereby enabling predictions of causes/effects in regulatory networks.

We hope to transfer this knowledge to other species, enabling us to make maximally use of the available information in inferring regulatory networks.

Kluyver CENTRE Kluyver Centre for Genomics of Industrial Fermentation
TU Delft Delft University of Technology