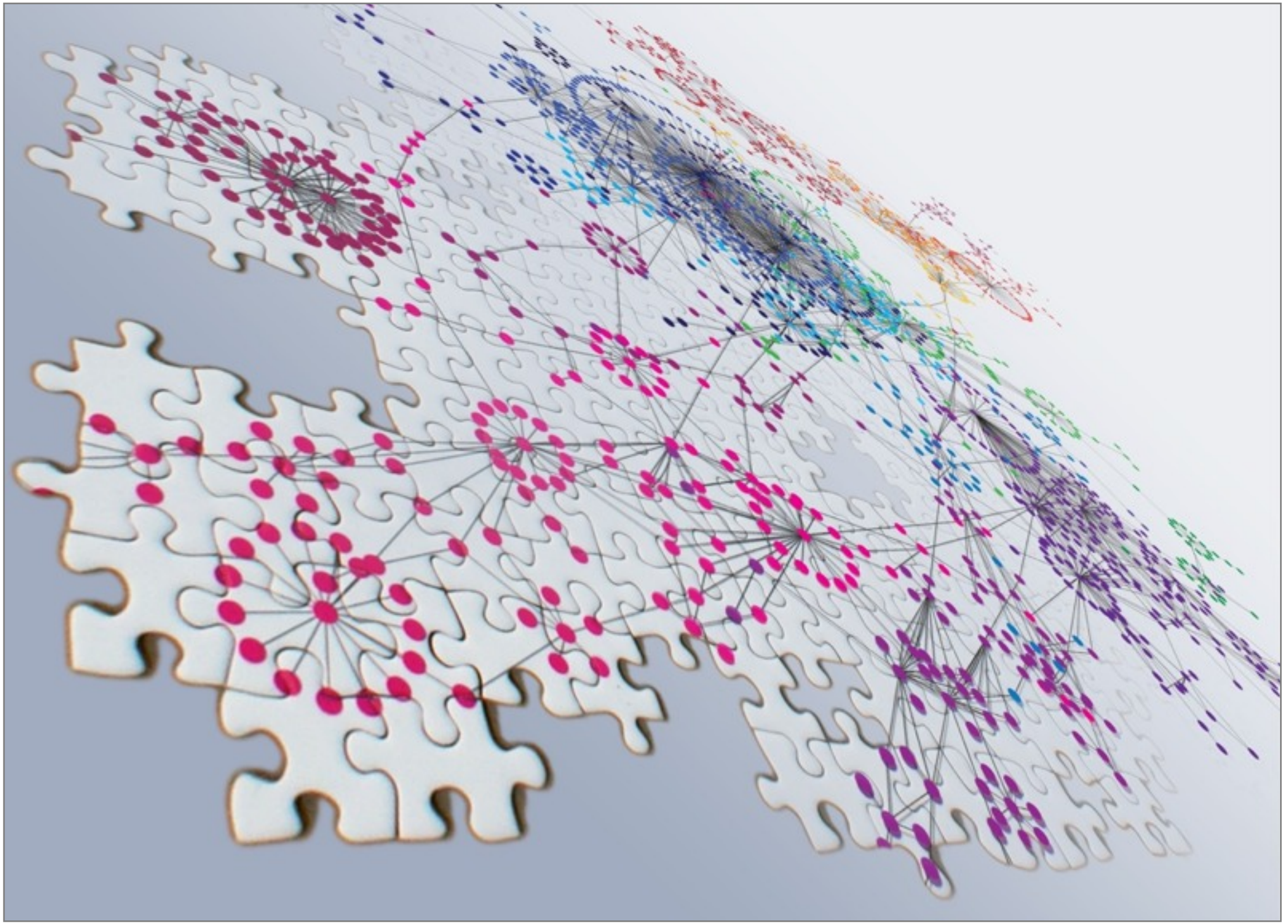
Robust inference of gene regulatory networks using bootstrapping

Supervisor: Daniel Marbach



Biological circuits

Dynamics

Modification

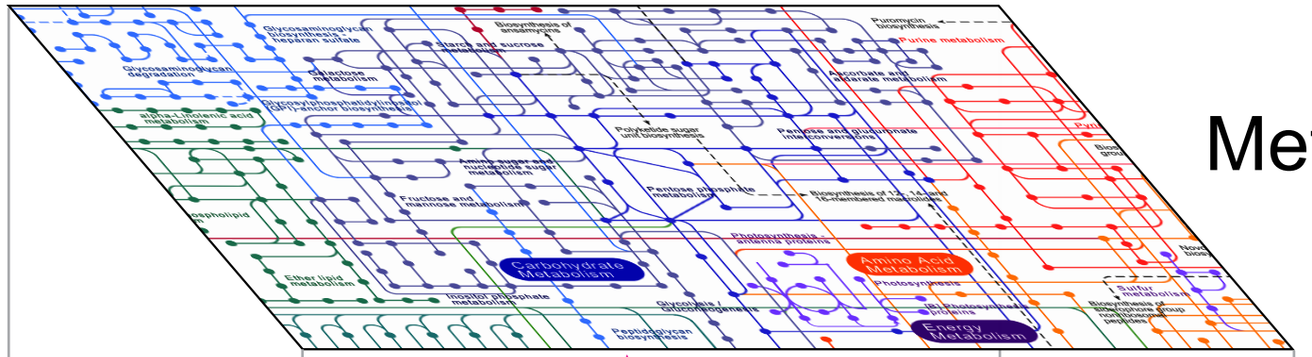
Proteins

Translation

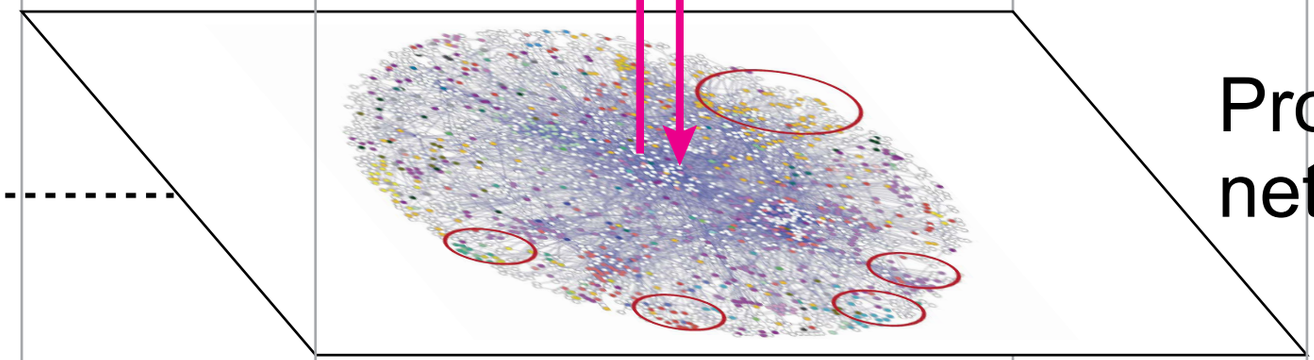
RNA

Transcription

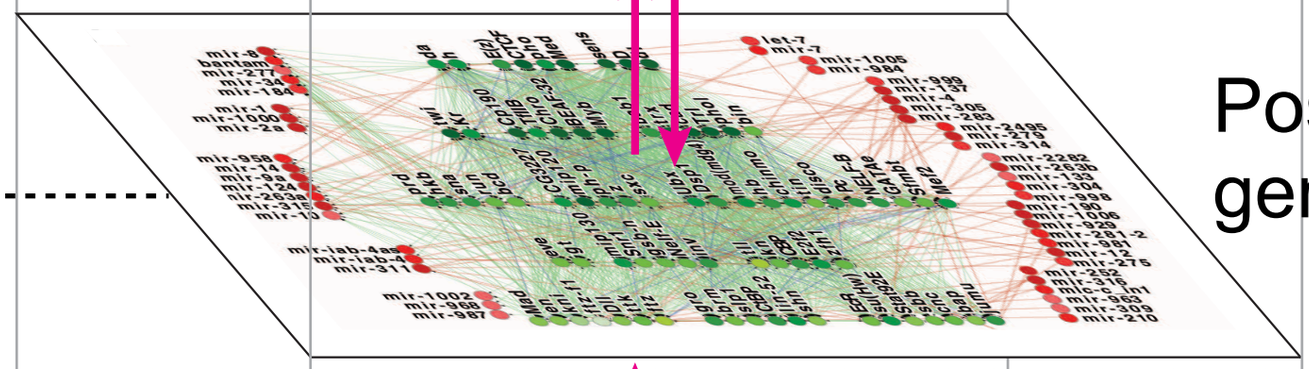
Genome



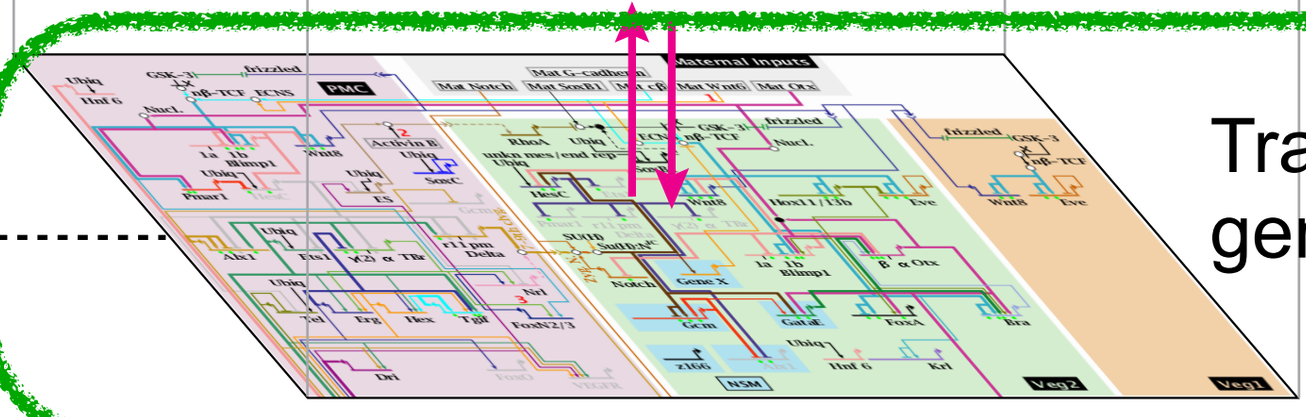
Metabolic networks



Protein & signaling networks

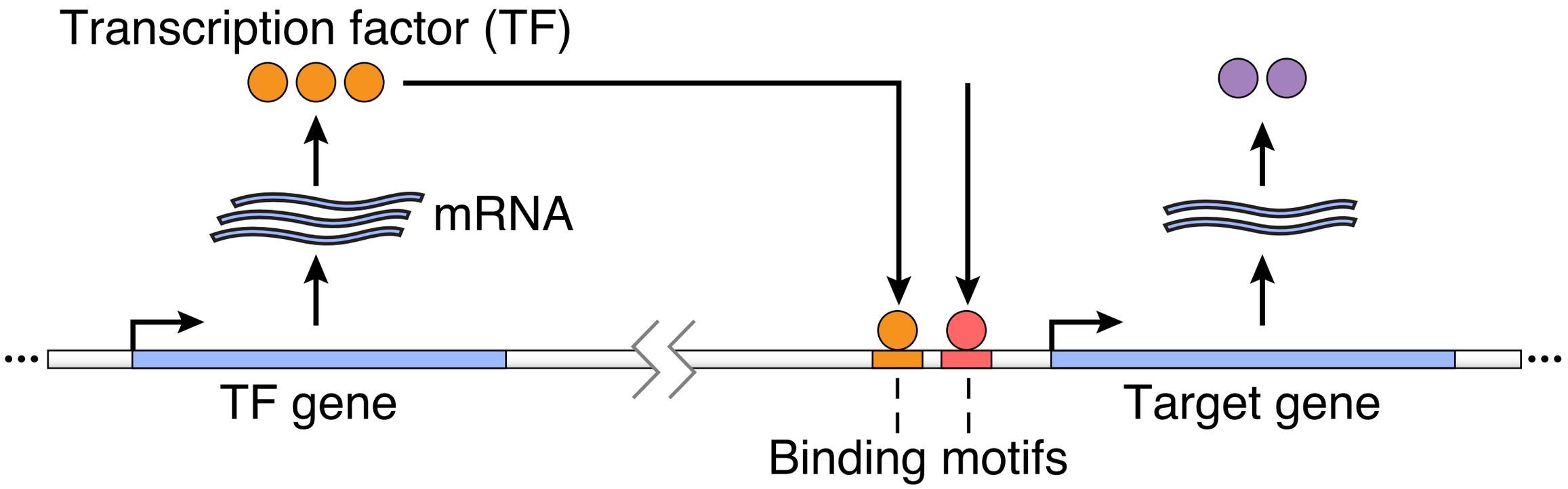


Post-transcriptional gene regulation



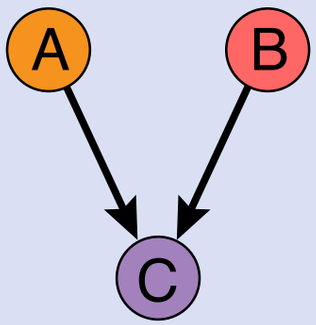
Transcriptional gene regulation

Transcriptional gene regulation



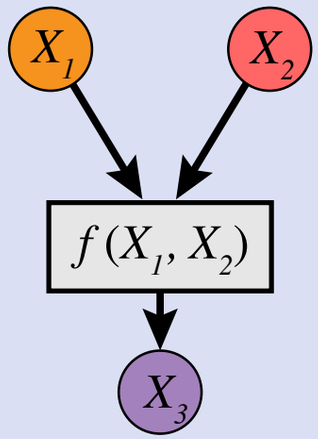
- **Modeling**

Structure



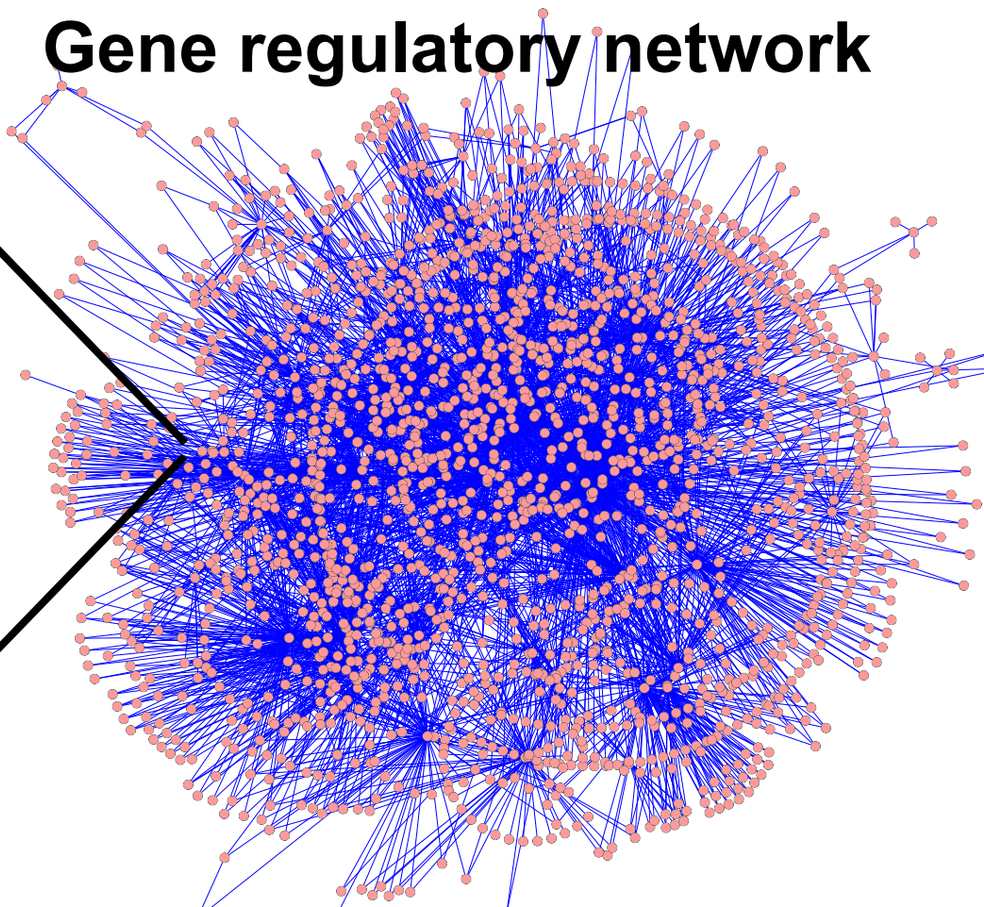
Who are the regulators?

Function



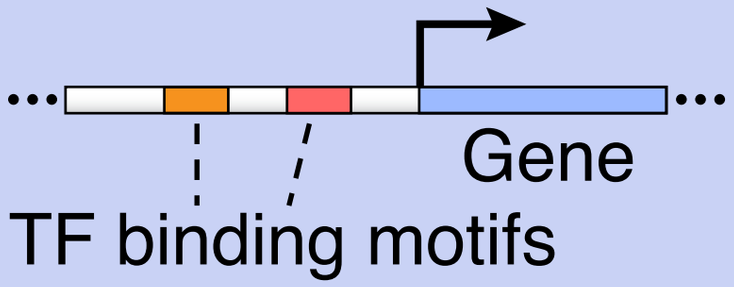
How do they determine expression?

Gene regulatory network

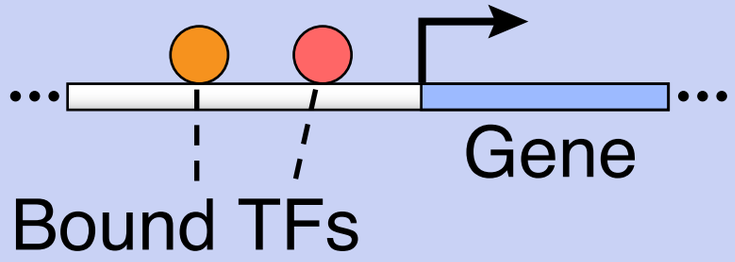


Strategies for reconstructing gene networks

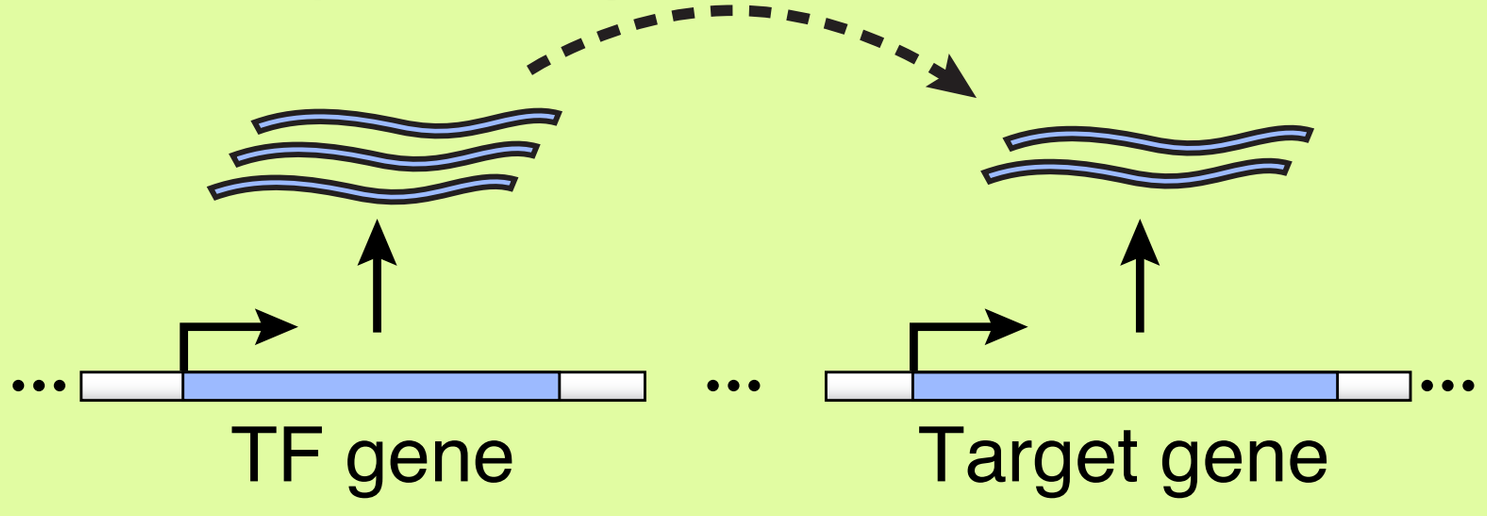
Identification of regulatory motifs



TF-binding data (ChIP-seq)



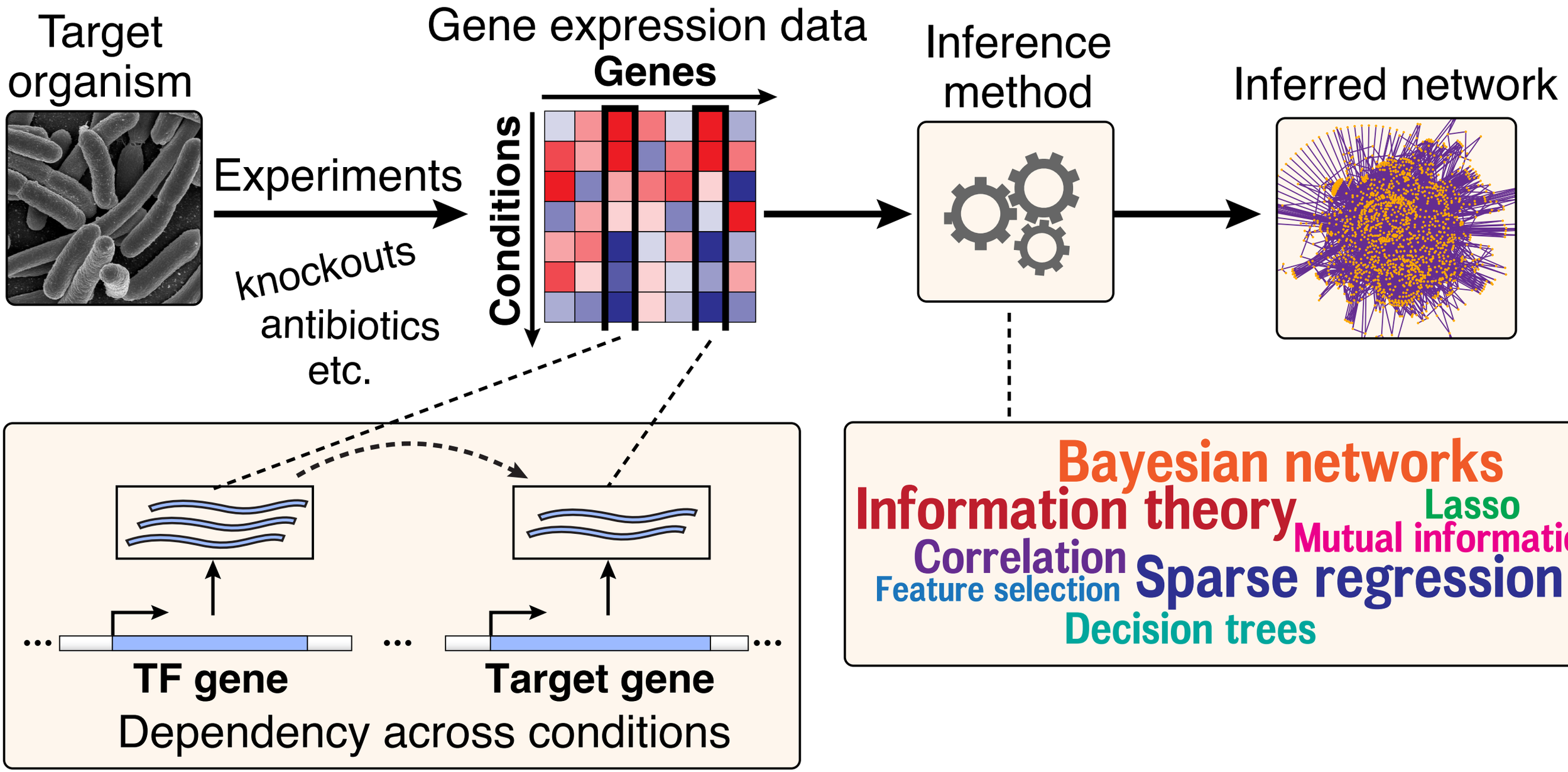
Expression-based network inference



Integrative

Combine different data sources

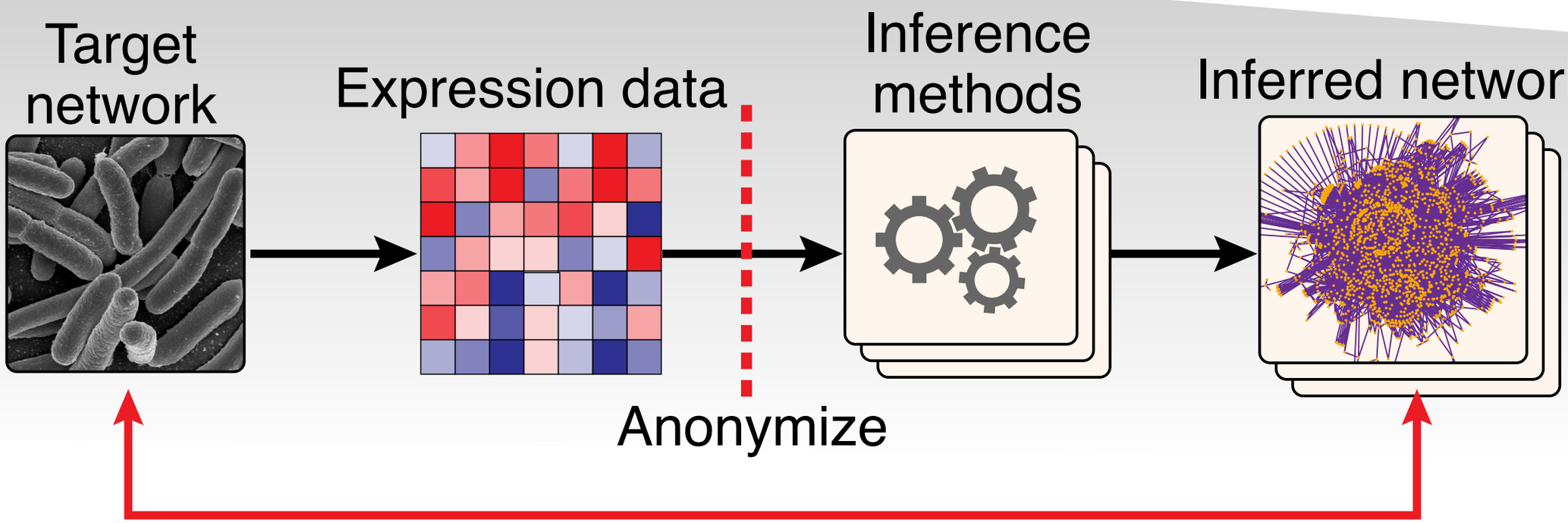
Expression-based gene network inference



- **Hundreds of methods have been proposed**
- **Are they any good?**

DREAM network inference challenge

Develop benchmarks → Assess methods → Characterize classes → Community predictions → Biological insight



Edition	#Teams
DREAM3 (2008)	29
DREAM4 (2009)	33
DREAM5 (2010)	29
DREAM8 (2013)	planned

DREAM network inference challenge

Develop benchmarks → Assess methods → Characterize classes → Community predictions → Biological insight

No single best performer

35 inference methods

ID	Synopsis	Reference
REGRESSION: TFs are selected by TG specific (1) sparse linear regression and combined by (2) stability selection.		
1	(1) Lasso; (2) the regularization parameter is chosen so that 5 TFs are selected per TG in each bootstrap sample.	39[a]
2	(1) Steady state and time series data are combined by group… bootstrapping.	40[a]
3	Combination of lasso (TESLA toolbox) … gression models learned from RJMCMC simulations.	41[a]
4	(1) Lasso; (2) boots…	42
5	(1) Lasso; (2) area u…	42
6	Application of the Las… dard parameters.	43
7	Lasso, regression mo… combined by the maximum regularization parameter value selecting a given edge for the first time.	42[a]
8	Linear regression determines the contribution of a TF to the expression of the TG.	44[a,b]
MUTUAL INFORMATION: Interactions are ranked based on variants of mutual information (MI).		
1	CLR, for a given TF_x-TG_y edge, an MI re-scoring scheme reflects TG_y and TF_x in the distribution of all TGs and … respectively.	8[a,b]
2	MI is co…	45[a,b]
3	ARACN… …essing inequali…	6[a,b]
4	Topolog… …ON-PC)/Bayesian local causal discovery (BLCD).	46[a]
5	Topology is estimated by MI and Pearson's correlation. The direction of edges is determined by HITON-PC/BLCD.	46[a]
CORRELATION: Interac…		
1	Absolute value c…	45
2	Signed value of …	45[a,b]
3	Signed value of …	45[a,b]
BAYESIAN NETWORKS optimize posterior probabilities by different heuristic searches.		
1	Simulated annealing (catnet R package), aggregation of three runs.	47
2	Simulated annealing (catnet R package)…	47
3	Max-Min Pa…	48
4	Markov blan…	49
5	Markov boun… sets.	50
6	Models TF K… data and time series by dynamic Bayesian networks (Infer.NET toolbox).	51[a]
OTHER APPROACHES: Network inference by heterogeneous and novel methods.		
1	Genie3: A random forest is trained to predict TG expression. Putative TFs are selected as tree nodes if they consistently reduce the variance of the TG.	28[a]
2	TF-TG co-dependencies are detected by the non-linear correlation coefficient η^2 (two-way ANOVA). TF KO/OE data receive increased weights.	29[a]
3	TFs are selected maximizing the conditional entropy for a … represented as Boolean vectors with probabilities, avoiding discretizatio…	52[a]
4	Putative TFs are preselected from TF… TFs are then tested by iterative Bayesian Model…	53
5	A Gaussian noise… changes significantly in TF KO/OE measurem…	54
6	After scaling, TGs … correlation. A neural network was trained (genetic algorithm) and para… propagation).	55[a]
7	Data were discretiz… Gaussian mixture models and clustering (Ckmeans); Detects interactions by generalized logical network modeling (χ^2 test).	56[a]
8	The χ^2 test was applied to evaluate the probability of a simultaneous shift in TF and TG expression in TF KO/OE experiments.	56[a]
META PREDICTORS combine (1) several approaches by calculating (2) aggregate scores.		
1	(1) Calculates z-scores for TG in TF KO data, applies time-lagged CLR for time series, and linear ODE models constrain… (2) resampling.	57[a]
2	(1) Pearson's cor…	58
3	(1) Calculates TG… co-deviation analysis; (2) weig…	59[a]
4	(1) CLR filtered b… es, and TF KO/OE data; (2) combination by z-scores.	60
5	(1) Pearson's correlation, differential expression (limma, Gauss tail), and time series analysis (maSigPro); (2) Naïve Bayes.	61[a]

Regression

Mutual information

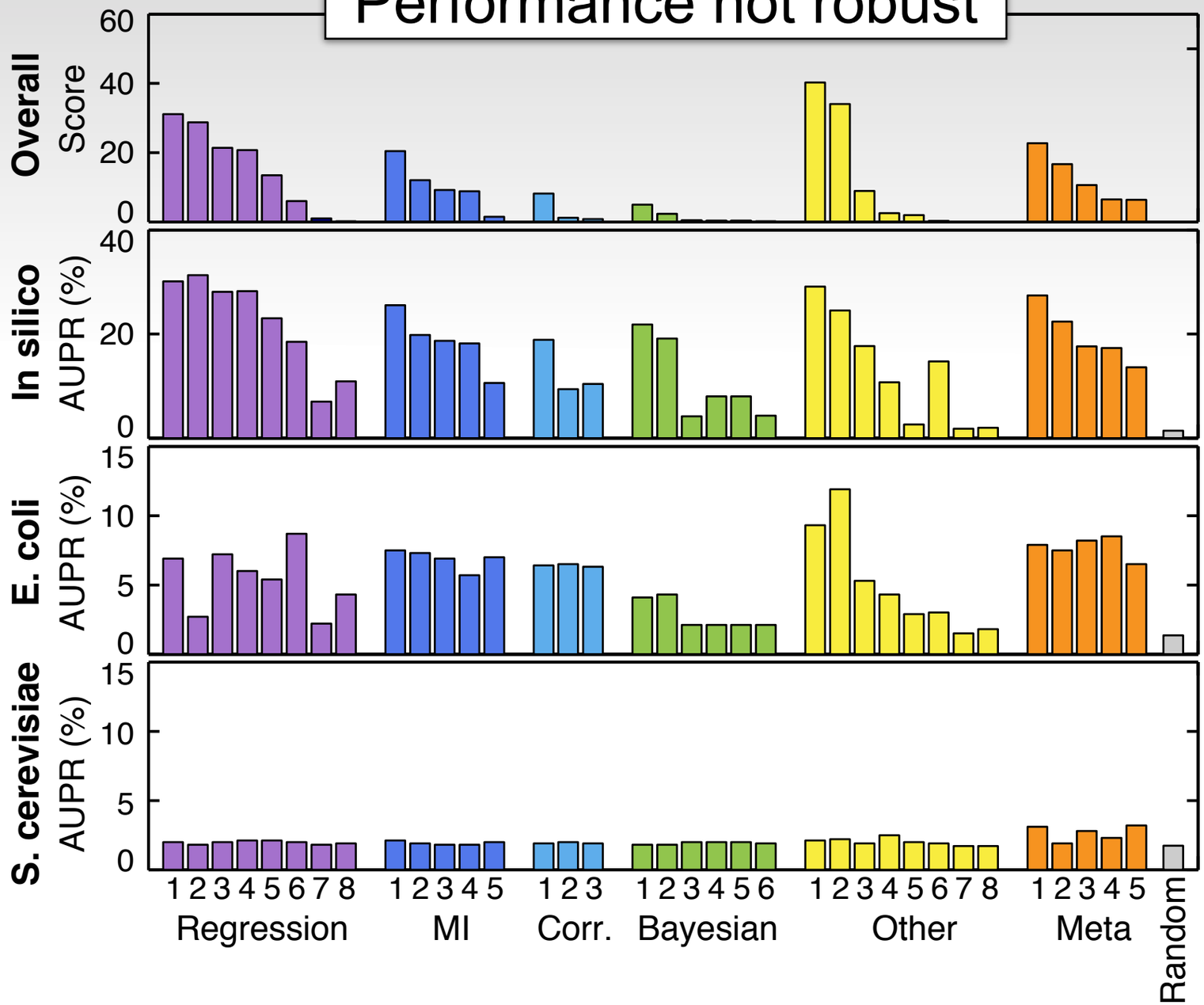
Correlation

Bayesian networks

Various other

Meta predictors

Performance not robust

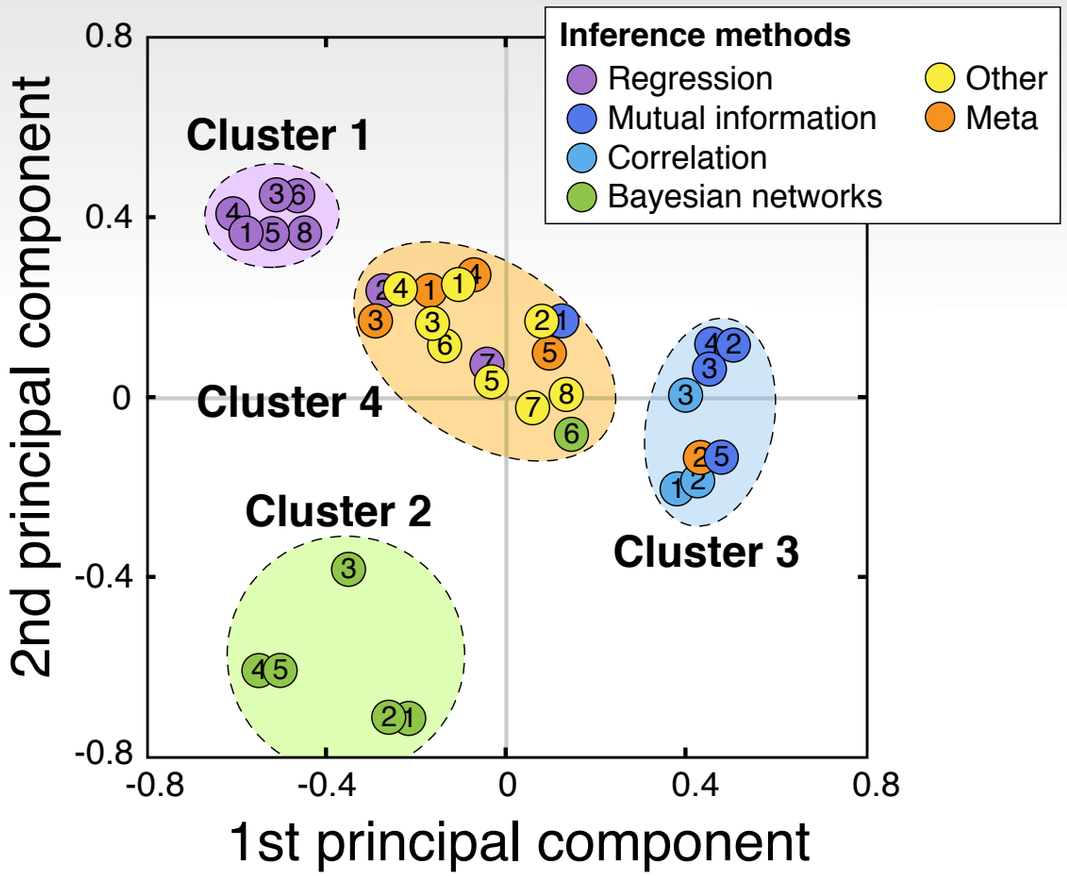


DREAM network inference challenge

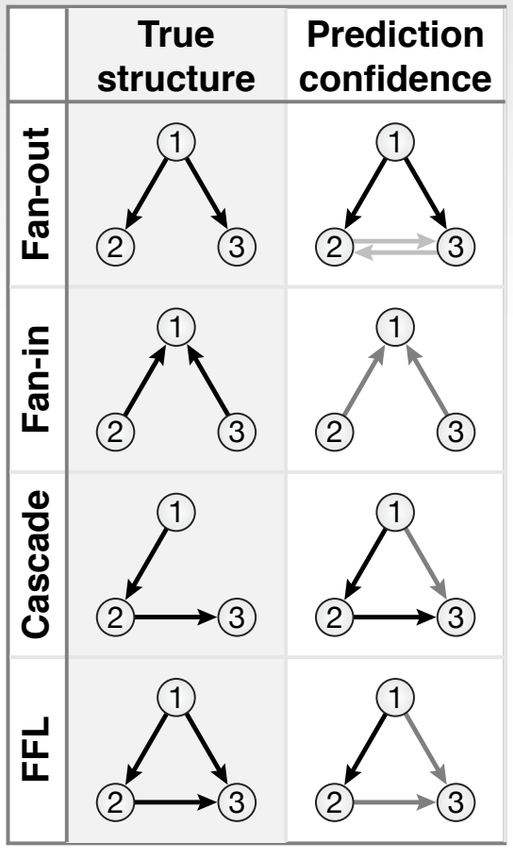
Develop benchmarks → Assess methods → Characterize classes → Community predictions → Biological insight

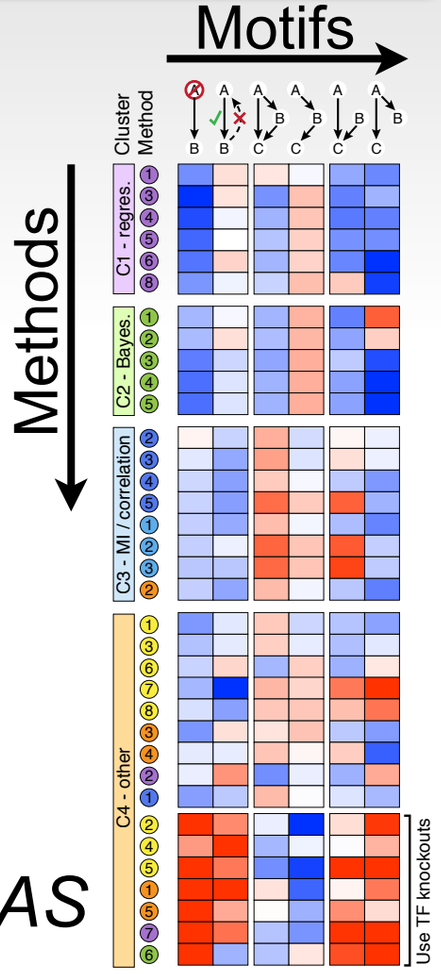
Complementary strengths & weaknesses

Similarity of predictions



Network motif biases





Marbach et al. 2010, *PNAS*

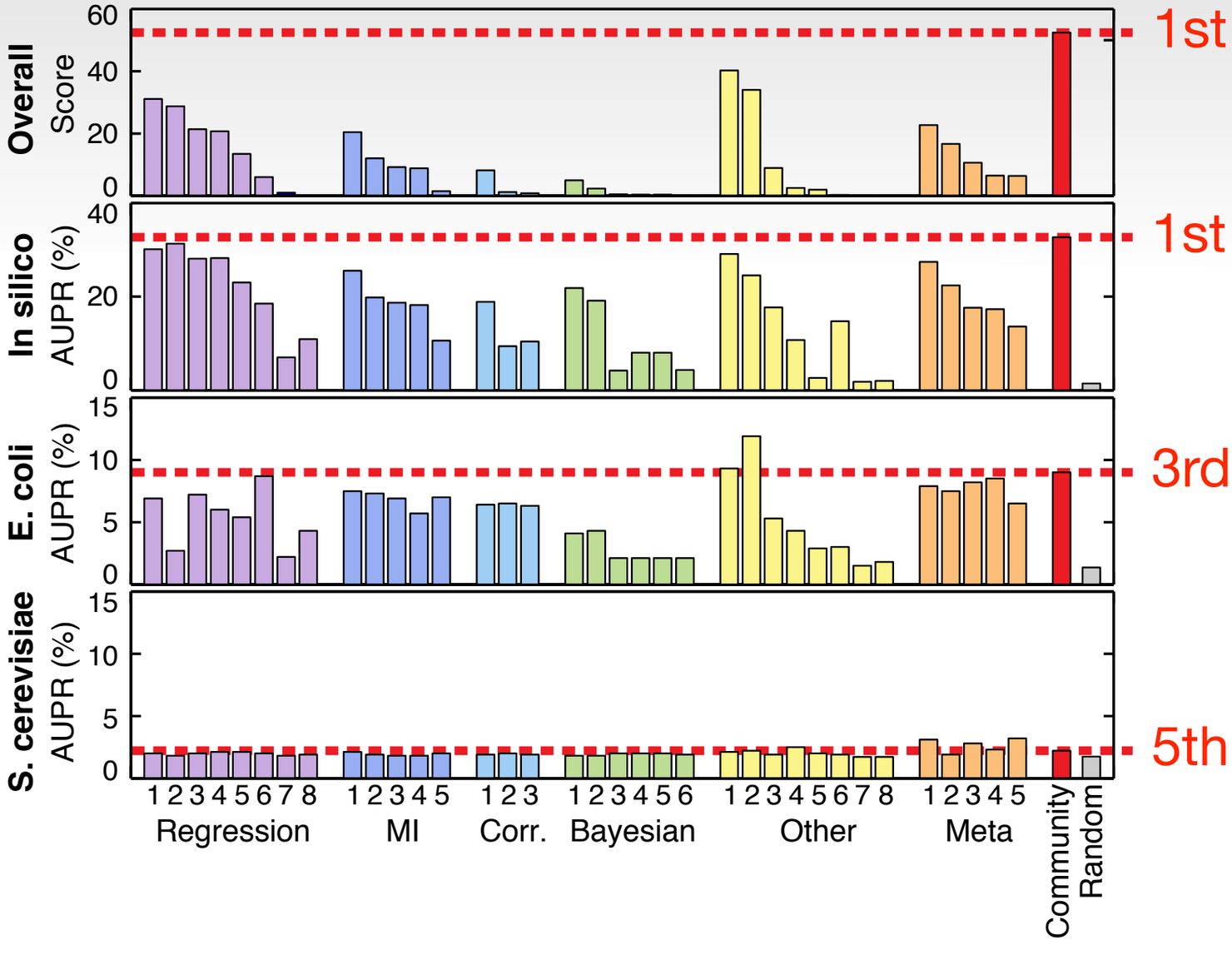
➡ **Can we integrate methods to form more accurate predictions?**

DREAM network inference challenge

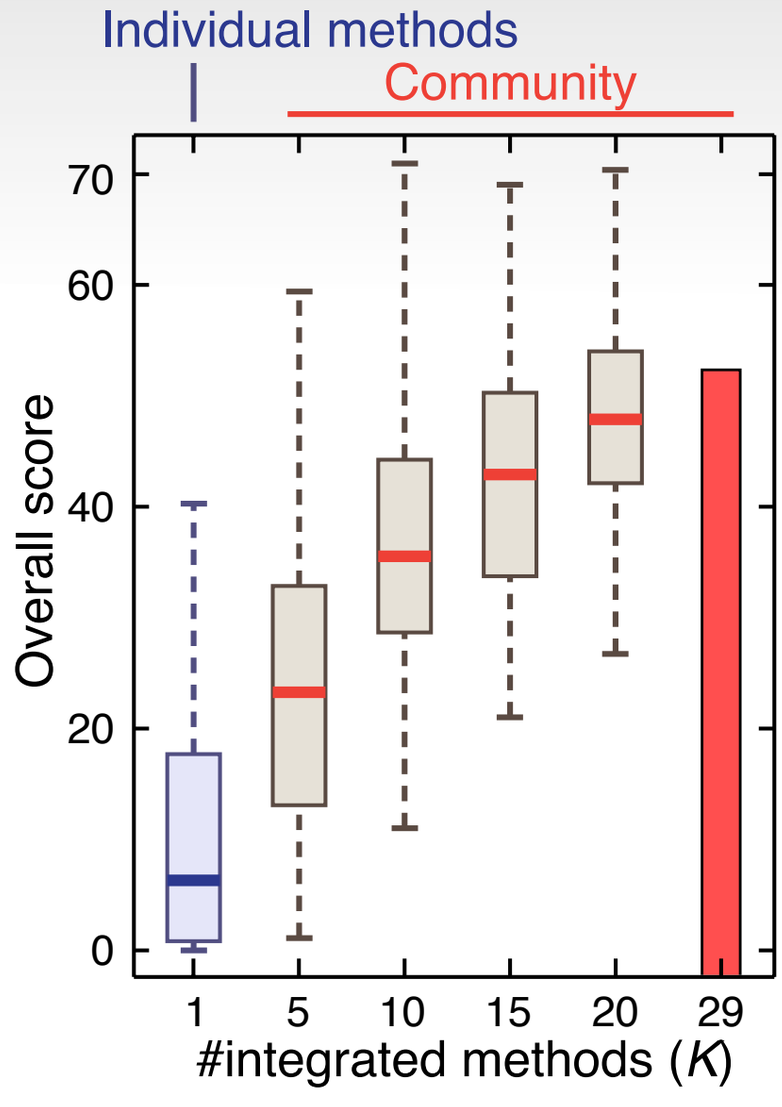
Develop benchmarks → Assess methods → Characterize classes → Community predictions → Biological insight

Wisdom of crowds

Community outperforms individual methods



Performance increases as more methods are added

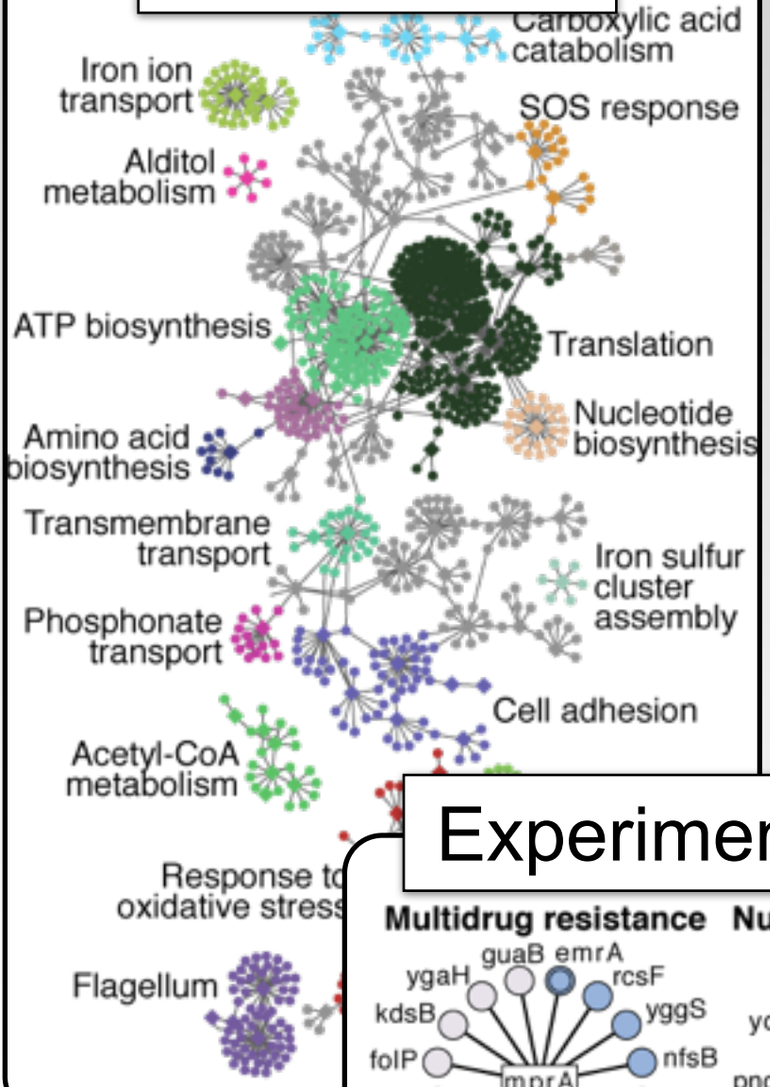


DREAM network inference challenge

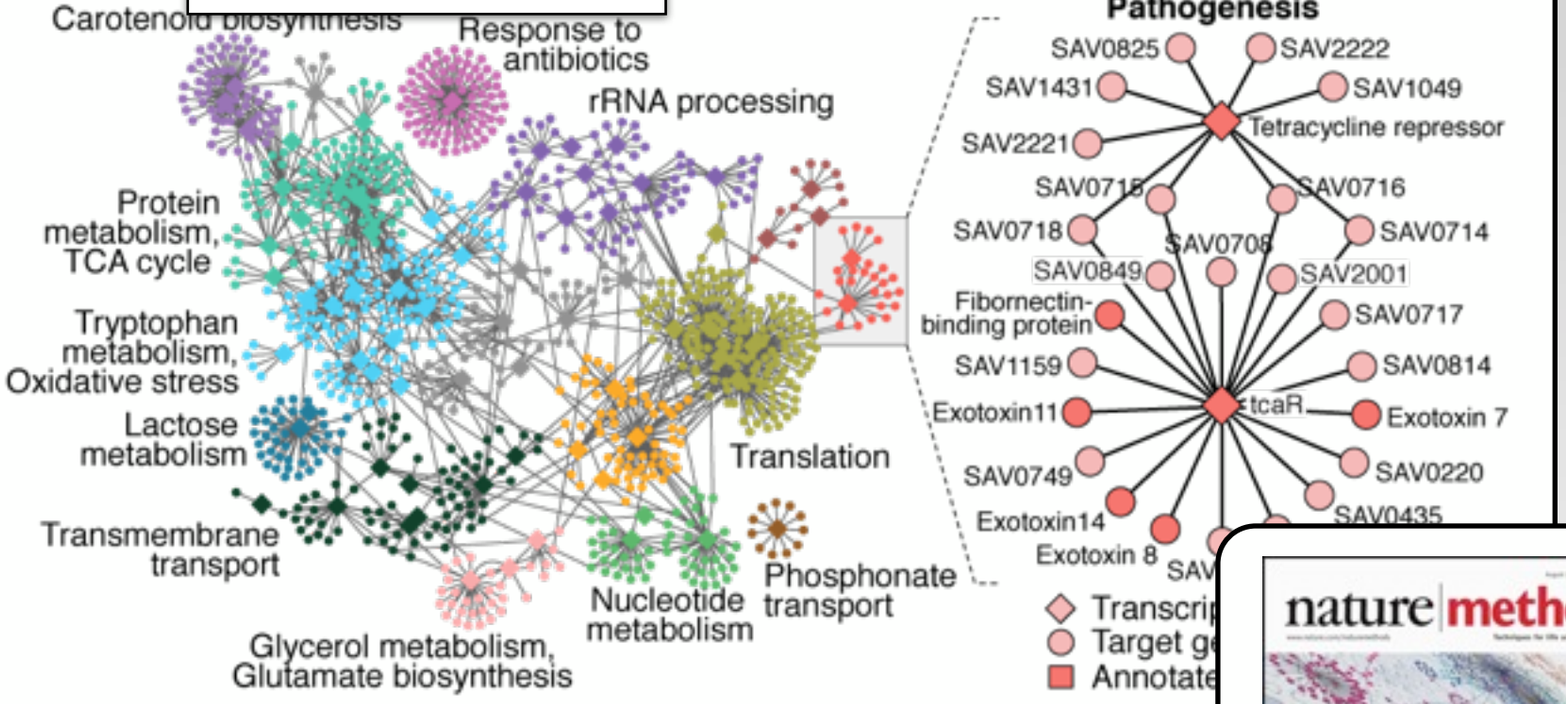
Develop benchmarks → Assess methods → Characterize classes → Community predictions → Biological insight

High-confidence networks

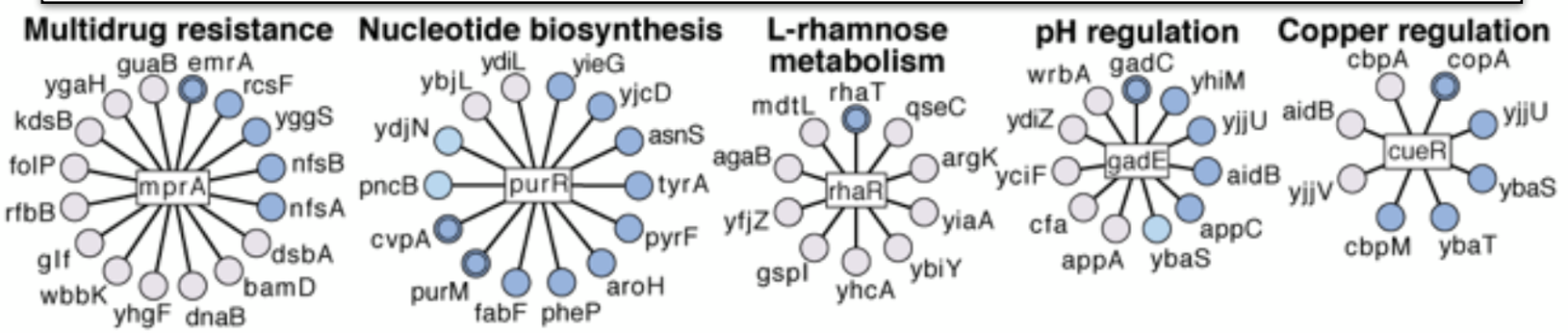
E. coli



S. aureus



Experimental validation of novel interactions



Marbach 2012, *Nature Methods*

The aims of this project

- **Understand different network inference approaches at a high level**
- **Are methods robust to variations in the data?**
  - Top 5 inference methods available on web platform (dream.broadinstitute.org)
  - Evaluate performance robustness on subsets of the expression compendia (data subsampling)
- **Can robustness be improved using bootstrapping?**
  - Run methods on subsets of the data and combine predictions
  - Do you obtain a new best-performing method in this way?