Présentation NETBIO

Vincent Rocher

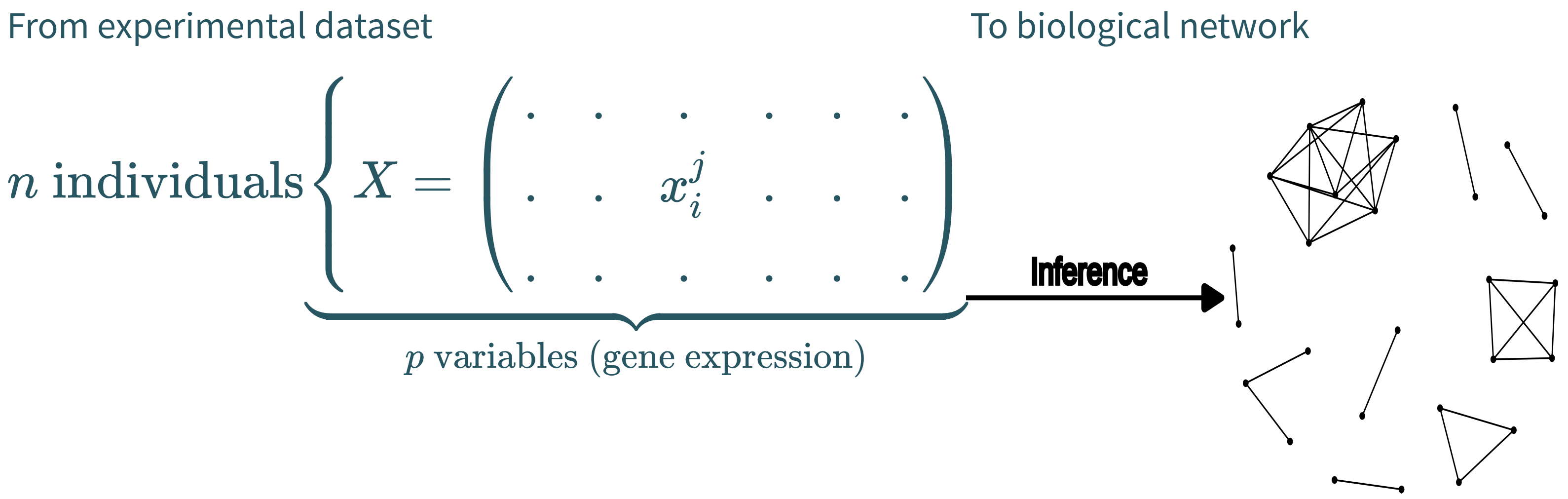
vincent.rocher@inrae.fr

15 Nov, 2023

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

Gene regulatory network (GRN) inference



From experimental dataset

To biological network

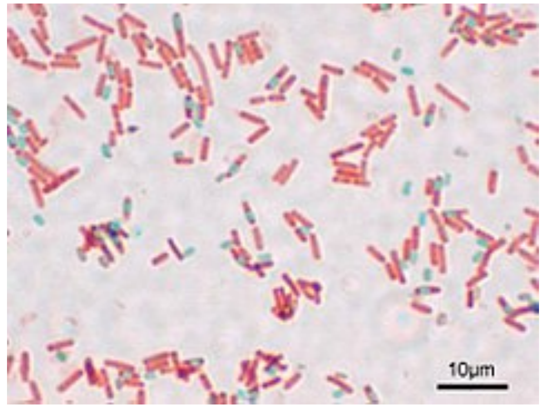
$$n \text{ individuals} \left\{ \underbrace{X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & x_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}}_{p \text{ variables (gene expression)}} \right. \xrightarrow{\text{Inference}}$$

Benchmarks:

- (Marbach et al. 2012): Wisdom of crowds for robust gene network inference
- (Chen and Mar 2018): Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data
- (Saint-Antoine and Singh 2023): Benchmarking Gene Regulatory Network Inference Methods on Simulated and Experimental Data

- Many statistical methods exist to infer networks from gene expression
- Benchmarks of these tools exist but:
  - Datasets: simulated or from incomplete databases
  - Usually simple edge evaluations (ROC/PR curves)
- More in-depth evaluation (global structure, modules, …)
- Work with a complete and manually curated network

Reconstruction of the GRN of *B. subtilis*



source: Wikipedia

Experimental dataset

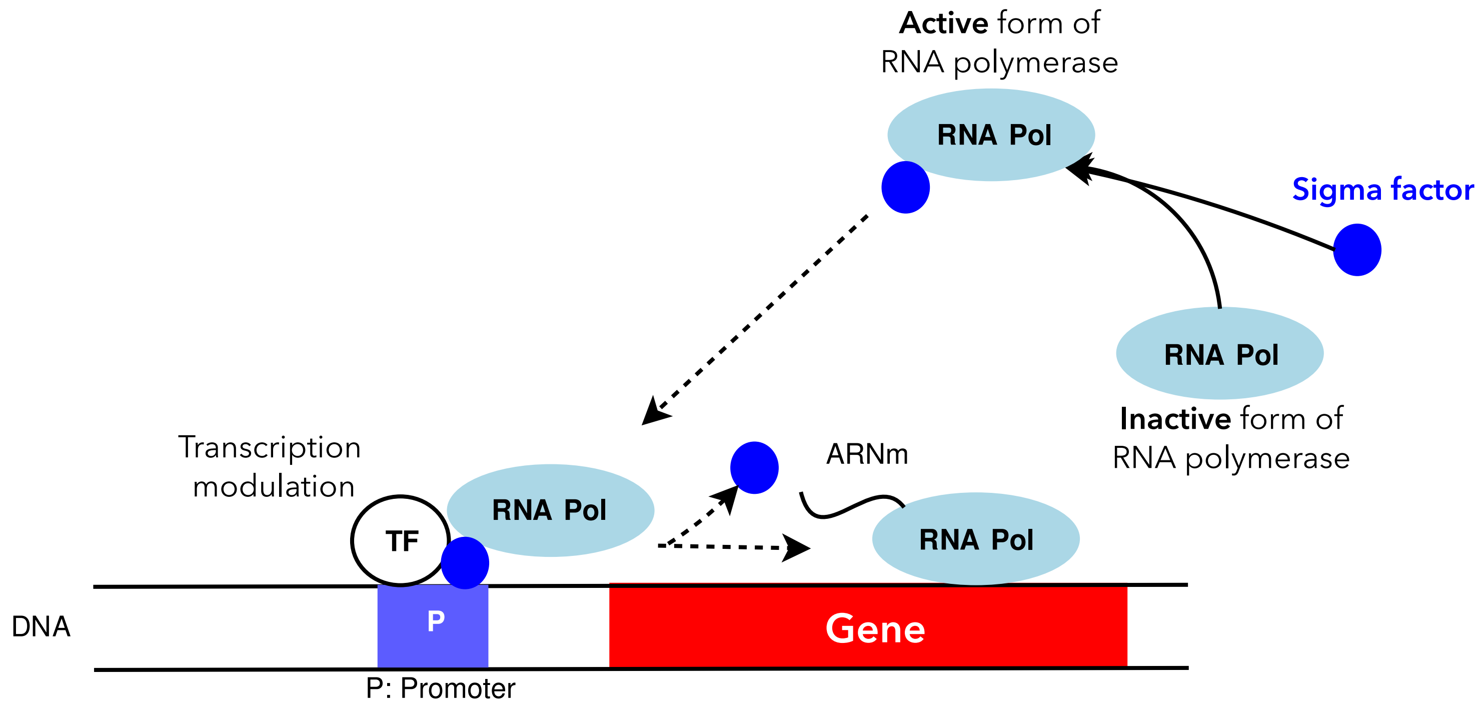
- Expression profiles of ~3900 genes for 269 experiments from (Nicolas et al. 2012)

Biological network

- Full gene regulatory and metabolic network from (Faria et al. 2016):
  - Genetic regulations (full GRN)
  - Metabolic pathways
  - Metabolic effectors acting on genes

Specificities in bacteria

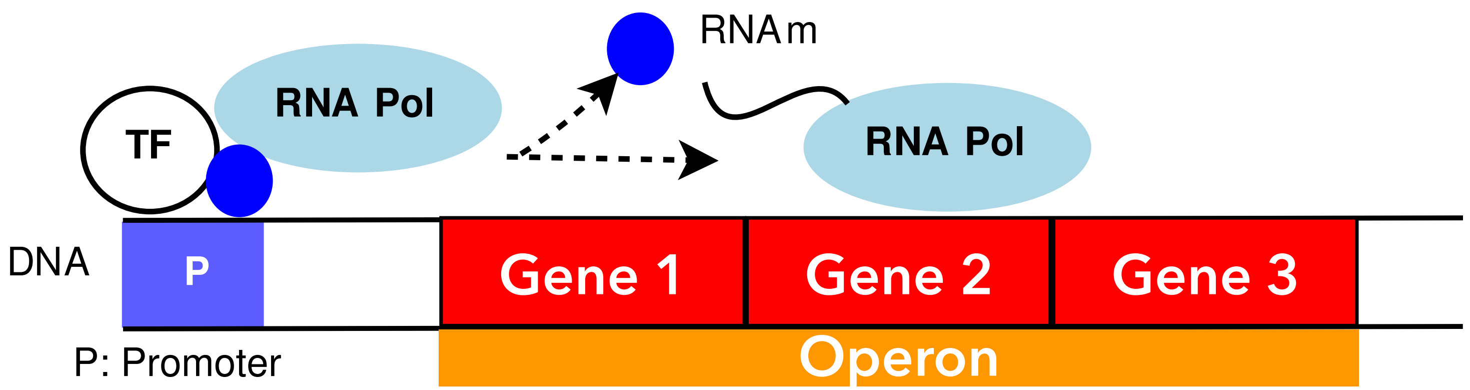
Transcription regulations



(from Goelzer (2010))

- Genes are regulated by sigma factors as well as by transcription factors (TF)

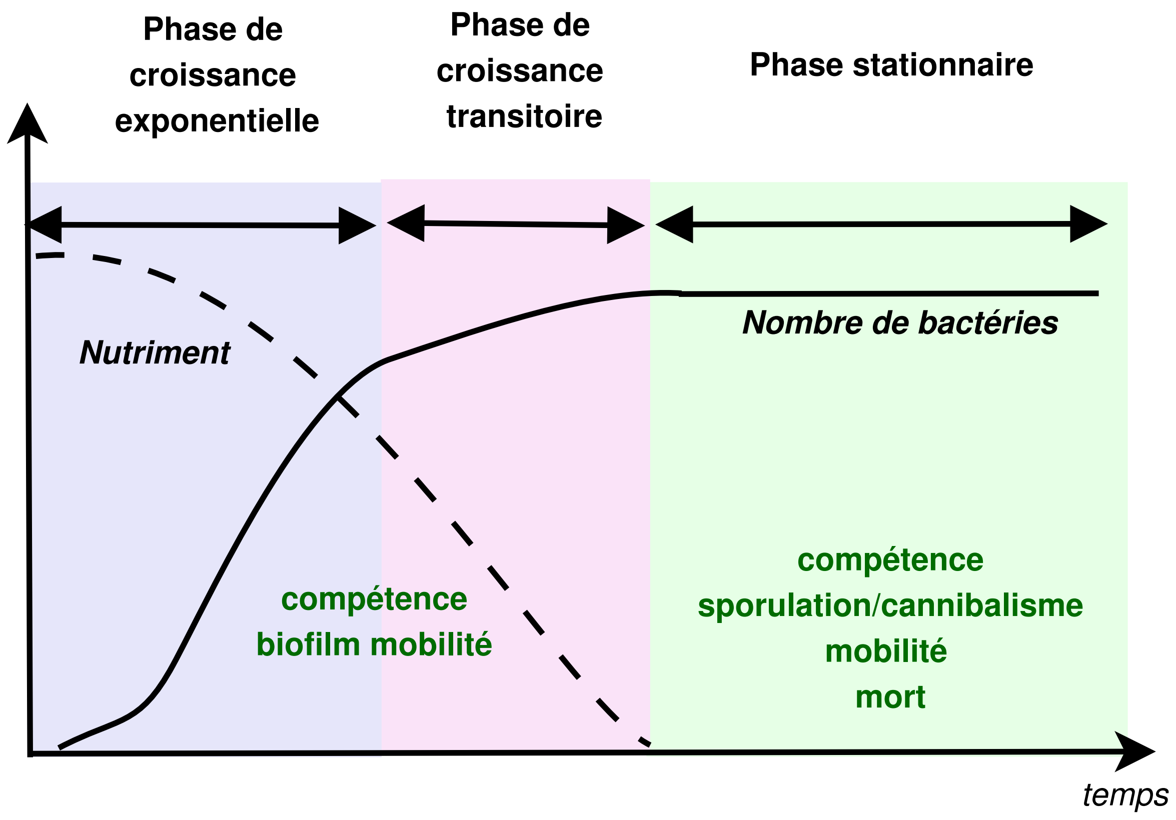
Operons



- Genes are transcribed simultaneously in transcription unit (operon)

Specificities in bacteria

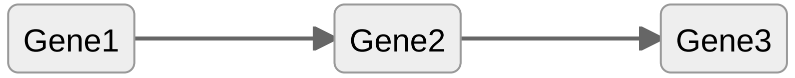
Growth phases



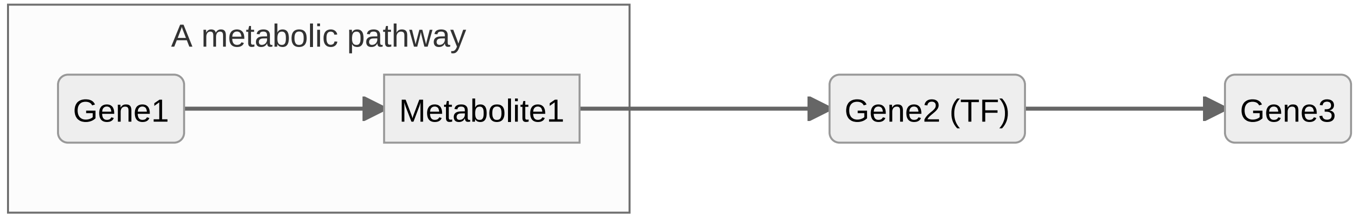
(from Goelzer (2010))

Regulations

Gene only regulation (sporulation):



Effector based regulation (Exponential phase):



- 83% of the regulators have an effector
- BUT effector are hidden in GRN

Need to integrate effectors → known (Faria et al. 2016)

Scientific questions

- Benchmarking:
  - How do GRN inference behave compared to a fully reconstructed biological network?
  - Real scale genome-wide regulation analysis
  - Improve evaluation to understand which biological signal the methods capture

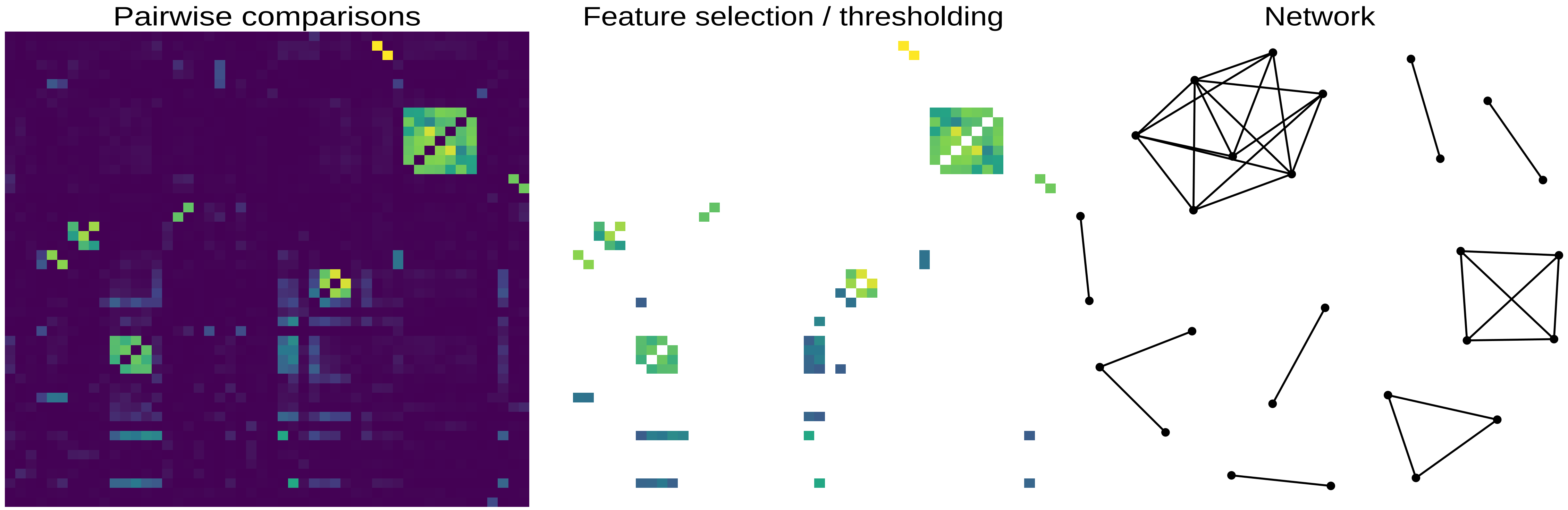
→ Integrate metabolic knowledge

- Traditional GRN inference methods
- Network evaluation methods through published benchmarks

Statistical methods for GRN inference

- **Relevance networks**: Simplest methods using **correlation** (Correlation / Mutual information) metrics
- **Gaussian Graphical Model (GGM)**: Remove indirect relationship by using **partial correlation**
- **Random forest (RF) methods**: Generalize regression problems
- Bayesian network methods: Introduce causality to GRN inference
- **Deep learning**

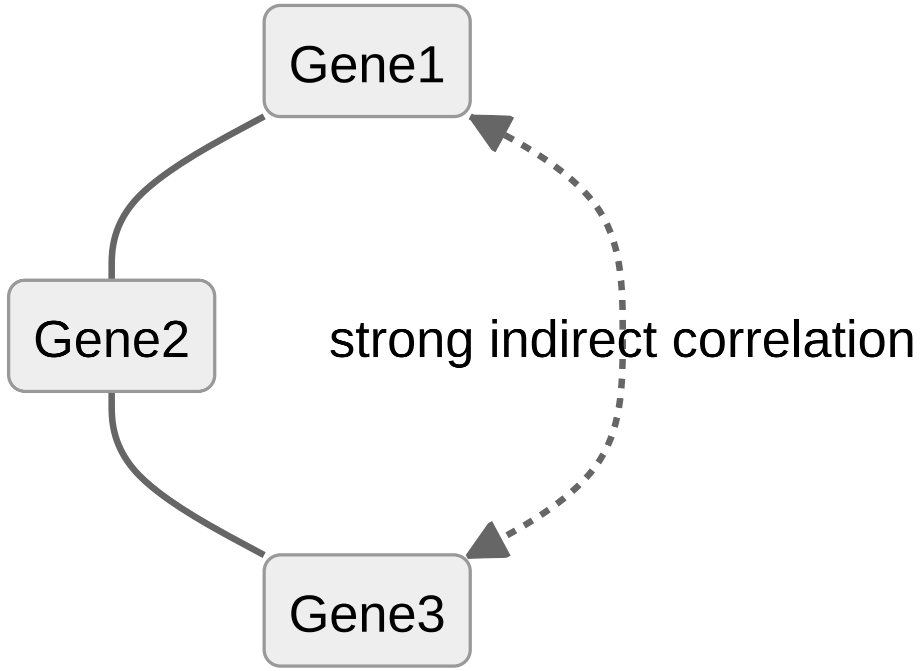
Relevance network: the simplest way



- Correlation (Butte and Kohane 1999): WGCNA (Langfelder and Horvath 2008)
- Mutual Information: ARACNE (Butte and Kohane 2000), minet (Meyer, Lafitte, and Bontempi 2008), CLR (Faith et al. 2007), PIDC (Chan, Stumpf, and Babtie 2017)

Gaussian Graphical Model (GGM)

Indirect relationship between Gene1 and Gene3:



- For $(X^j)_{j=1,\dots,p} \sim N(0,\Sigma)$ (gene expressions):
  - $\Rightarrow$ using partial correlations: edge between j and j' $\Leftrightarrow \mathrm{Cor}(X^j, X^{j'} | (X^k)_{k \neq j,j'}) \neq 0$
- **Graphical Gaussian Models (GGM):** $\mathrm{Cor}(X^j, X^{j'} | (X^k)_{k \neq j,j'}) \neq 0 \Leftrightarrow \beta_{jj'} \neq 0$ in the regression models $X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$
  - need to incorporate regularization or selection (Lasso) in these models (J. Friedman, Hastie, and Tibshirani 2008; Meinshausen and Bühlmann 2006)
  - Implemented in huge (Jiang et al. 2021), `glasso` (Jerome Friedman, Hastie, and Tibshirani 2019), `the Inferelator` (Skok Gibbs et al. 2022)

But computationally expensive (when p is large):

- $\Rightarrow$ Partial correlations between triplets of genes: **(Reverter and Chan 2008)** (`PCIT`): $\mathrm{Cor}(X^j, X^{j'} | X^k)_{k \neq j,j'}$

Random forest

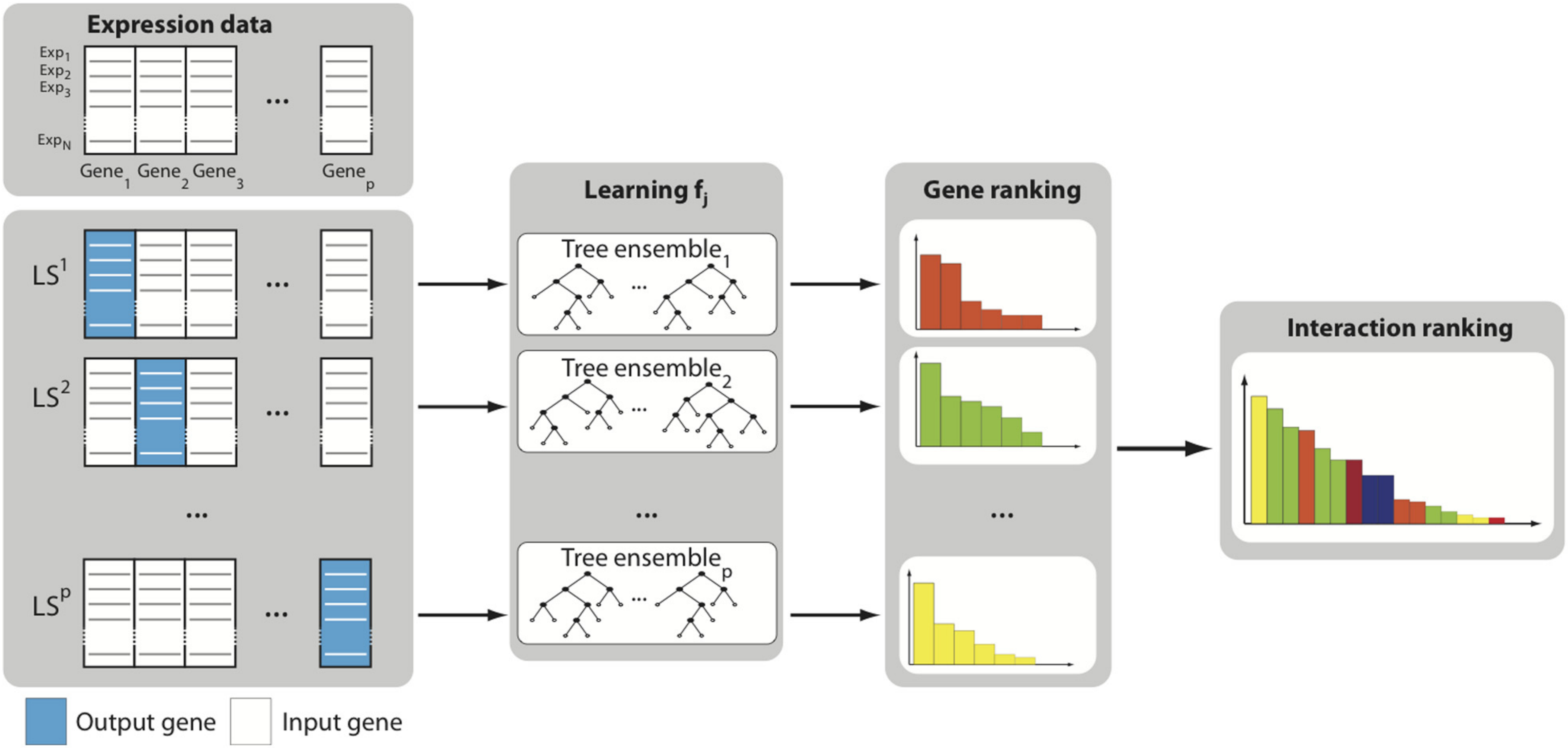
- **Graphical Gaussian Models (GGM)**: Gaussian assumption + fit of p linear regressions:
  - $\forall j = 1, \dots, p, \qquad X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$
  - **Problems**: Only linear dependencies, restricted to Gaussian case

- Just fit p regressions:
  - $X^j = F_j \left(\{X^{j'}\}_{j' \neq j} \right) + \epsilon_j$
  - Where F_j : random forest, gradient boosting

- Implementations: `GENIE3` (Huynh-Thu et al. 2010), `GRNBoost2` (Moerman et al. 2019)

BUT Direct interpretation of parameters is lost

How to select important predictors in Random Forest ?

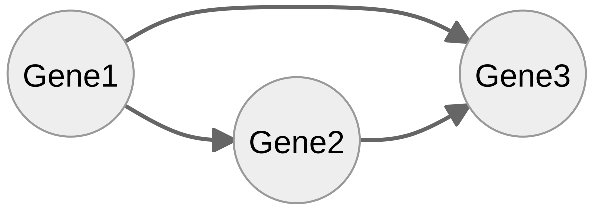


(Huynh-Thu et al. 2010): GENIE3 procedure.

Using deep learning for GRN inference

Analogy with linear structural equation model (SEM): $X = A^T X + Z$:

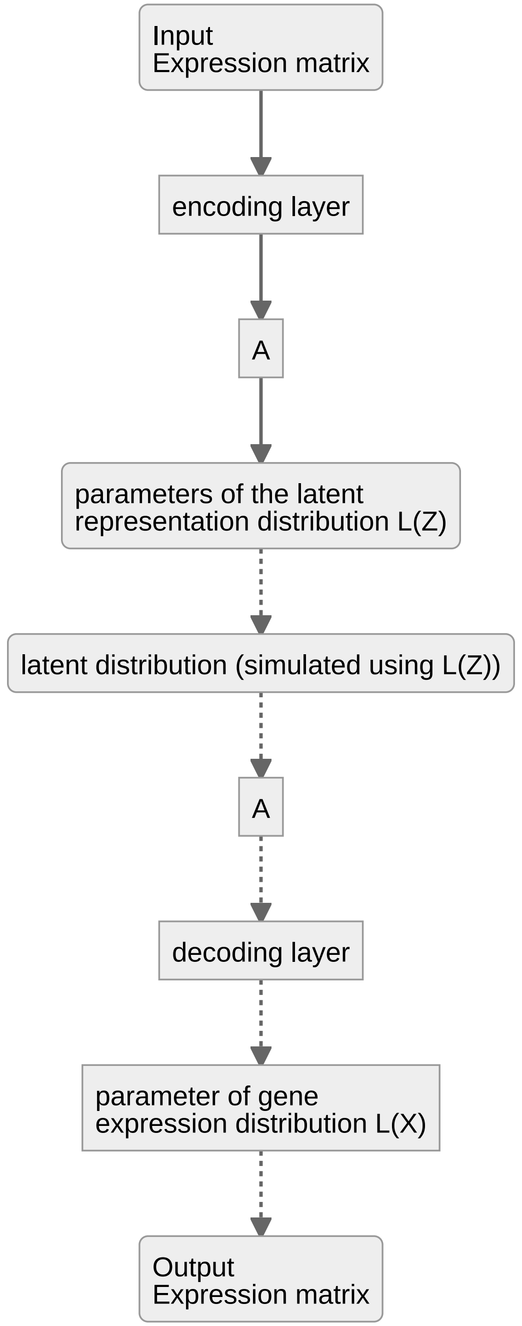
Here: GRN is a DAG:



… A: its adjacency matrix:

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

- A is a weight matrix in an Variational Auto-Encoder
  - Encoder: $\mathcal{L}(Z) = (I - A^\top)X$ ($\mathcal{L}(Z)$:= parameters of the distribution of Z)
  - Decoder: $\mathcal{L}(X) = (I - A^\top)^{-1} Z$


Input
Expression matrix

encoding layer

A

parameters of the latent
representation distribution L(Z)

latent distribution (simulated using L(Z))

A

decoding layer

parameter of gene
expression distribution L(X)

Output
Expression matrix

Unsupervised deep learning GRN using a variational autoencoder (Yu et al. 2019; Shu et al. 2021)

Evaluation of GRN inference methods

- (**Marbach et al. 2012**): Wisdom of crowds for robust gene network inference
- (**Chen and Mar 2018**): Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data
- (**Saint-Antoine and Singh 2023**): Benchmarking Gene Regulatory Network Inference Methods on Simulated and Experimental Data

Input dataset & Ground Truth network

- **Input Dataset**:
  - Gene Expression datasets: Microarray (Marbach et al. 2012) or Single-cell (Chen and Mar 2018; Saint-Antoine and Singh 2023)
  - Simulated expression dataset : GeneNetWeaver (Chen and Mar 2018; Marbach et al. 2012) and in house (Saint-Antoine and Singh 2023)

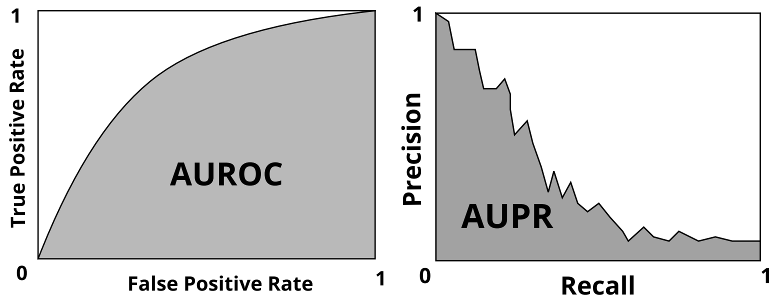
- **Ground Truth network**:
  - From databases (RegulonDB, STRING) + experiments (ChIP-seq)
  - Simulated networks (GeneNetWeaver + in house)

But:

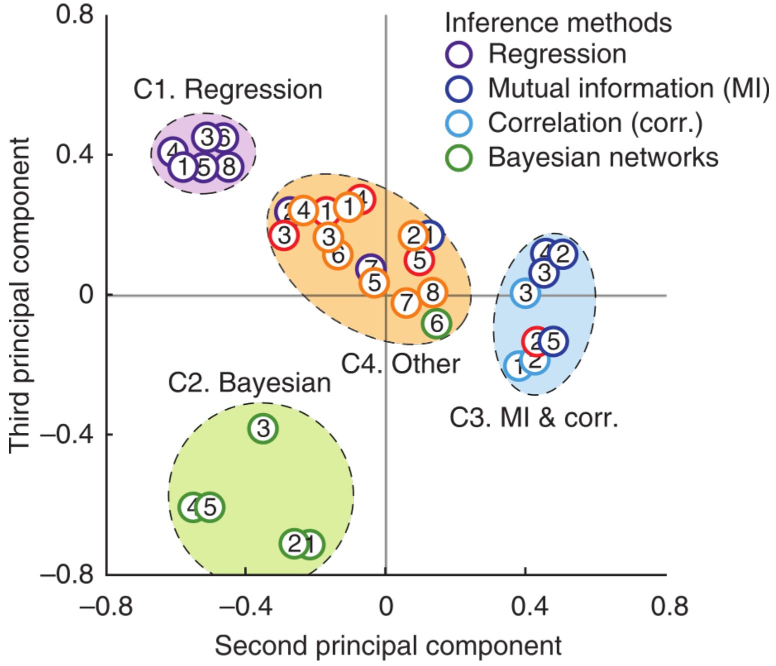
- Small set of genes: ~100/1000 genes for ~100/10k samples
- Very few edges: between 100 to 600 edges (except (Marbach et al. 2012) with ~4k edges)

Evaluation

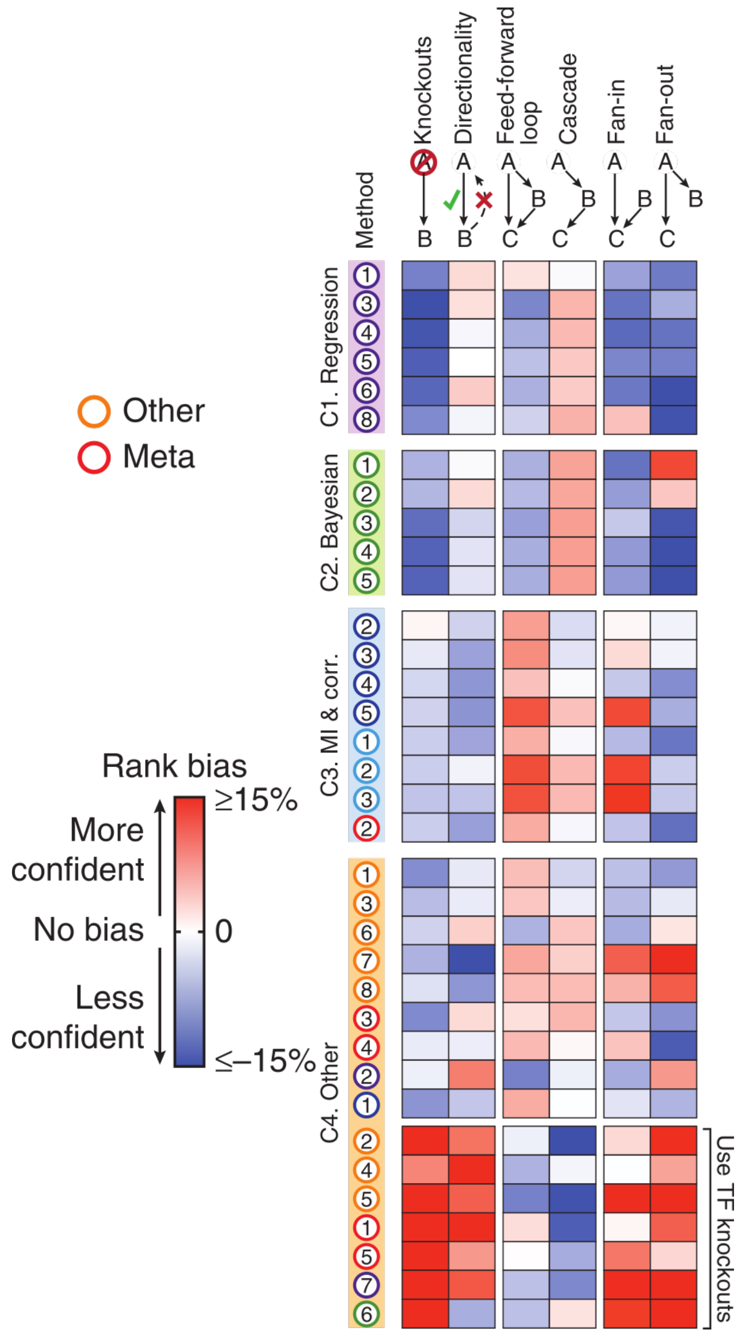
- Edge evaluation using machine learning approaches (ROC/PR)
- Network motifs analysis (using edge ranking)
- PCA representation of inference methods



(Saint-Antoine and Singh 2023)



(Marbach et al. 2012)

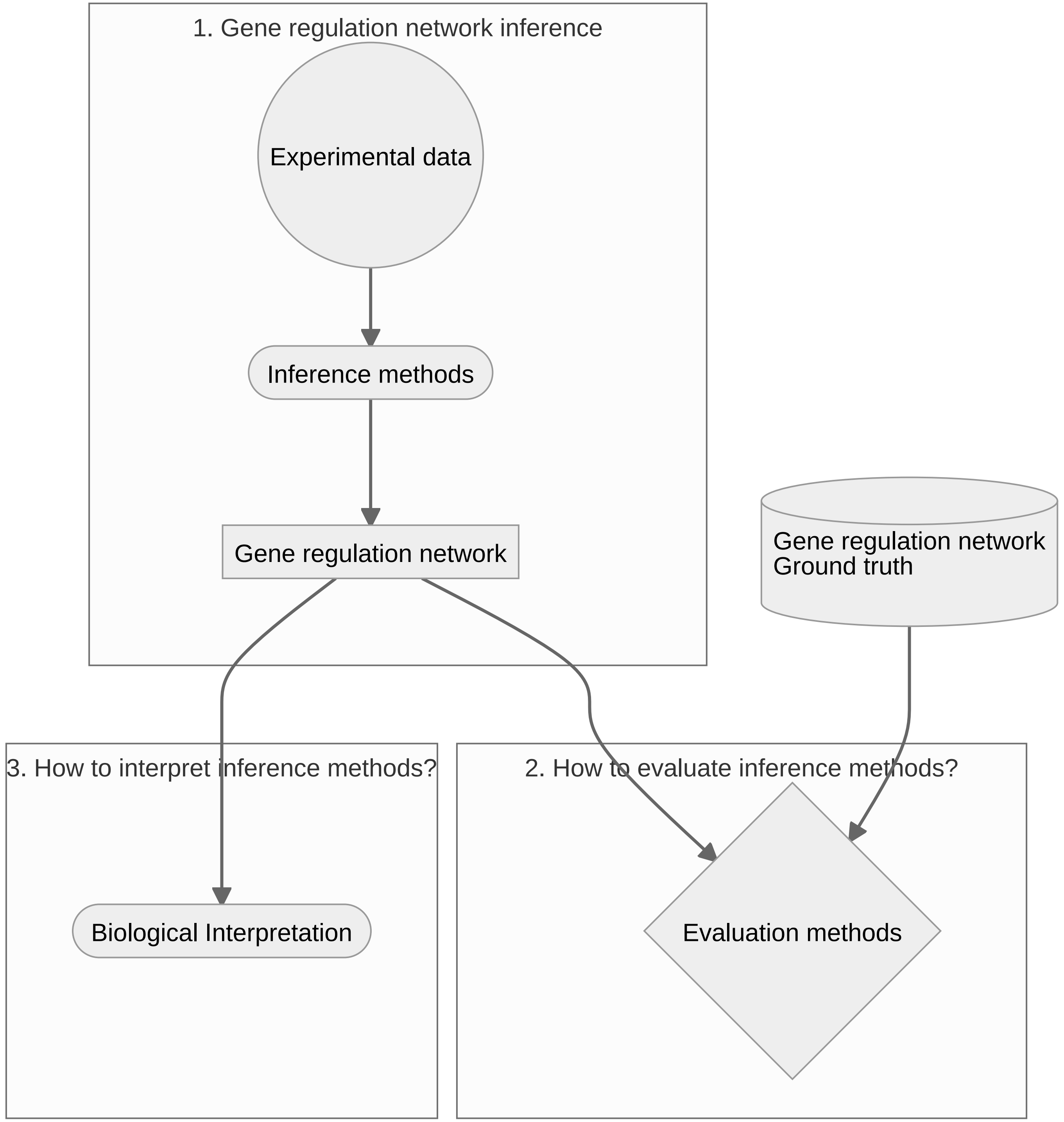


(Marbach et al. 2012)

Main conclusions

- Best performances:
  - Random forest (Microarray); Relevance Network (Single-cell)
  - poor performances in single-cell data
- method-specific edges are observed
- → ensemble methods improve performances (Marbach et al. 2012)
- Better performance using proteomics rather than transcriptomics (simulated data) (Saint-Antoine and Singh 2023)

Work plan: automation and reproducibility



1. Gene regulation network inference

Experimental data

Inference methods

Gene regulation network

Gene regulation network Ground truth

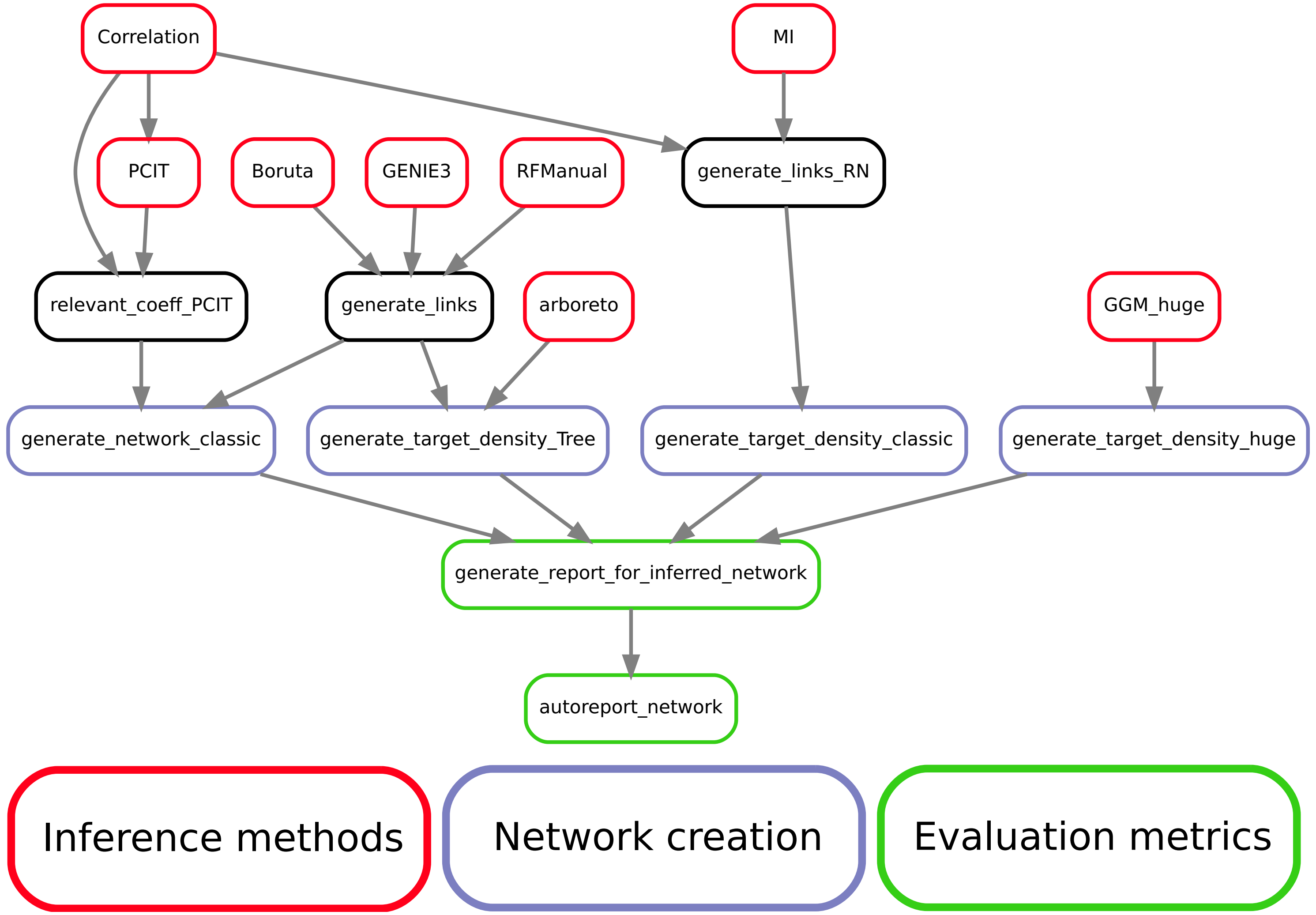
3. How to interpret inference methods?

Biological Interpretation

2. How to evaluate inference methods?

Evaluation methods

Reproducible GRN inference and evaluation



Correlation

MI

PCIT

Boruta

GENIE3

RFManual

generate_links_RN

relevant_coeff_PCIT

generate_links

arboreto

GGM_huge

generate_network_classic

generate_target_density_Tree

generate_target_density_classic

generate_target_density_huge

generate_report_for_inferred_network

autoreport_network

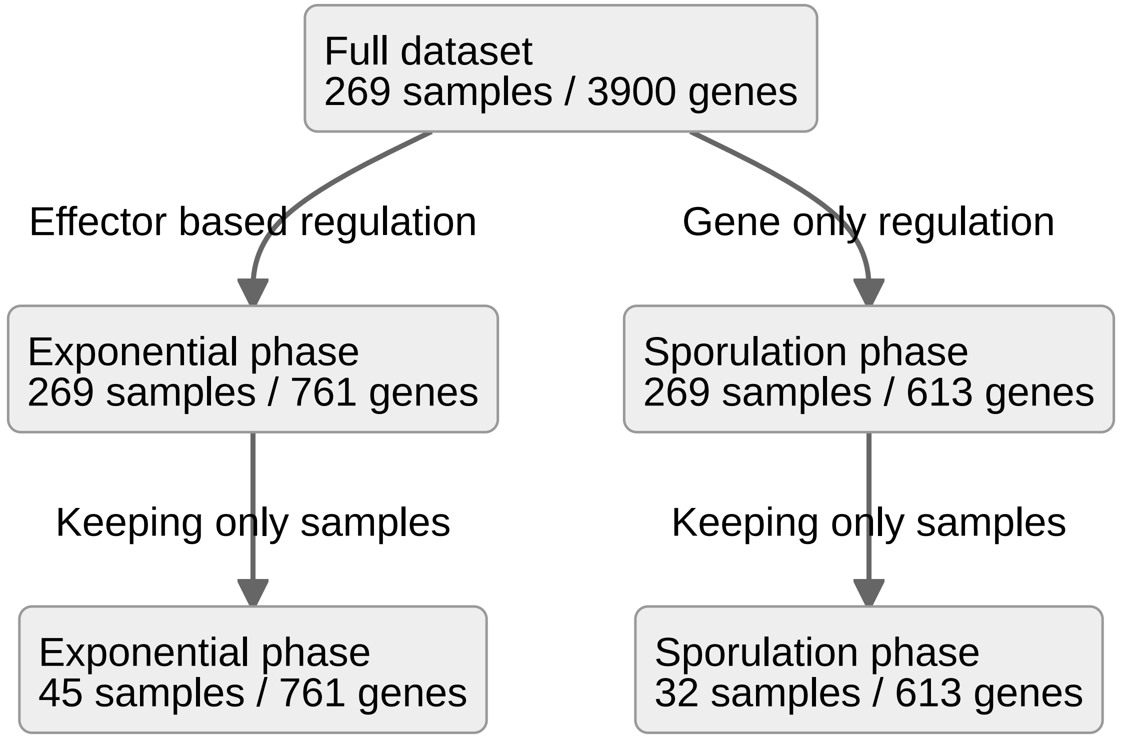
Inference methods

Network creation

Evaluation metrics

Snakemake (Mölder et al. 2021) pipeline

Used training datasets



Full dataset
269 samples / 3900 genes

Effector based regulation

Gene only regulation

Exponential phase
269 samples / 761 genes

Sporulation phase
269 samples / 613 genes

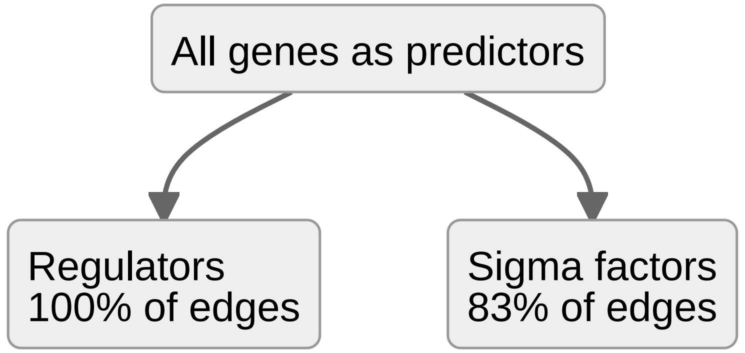
Keeping only samples

Keeping only samples

Exponential phase
45 samples / 761 genes

Sporulation phase
32 samples / 613 genes

Using prior to improve inference

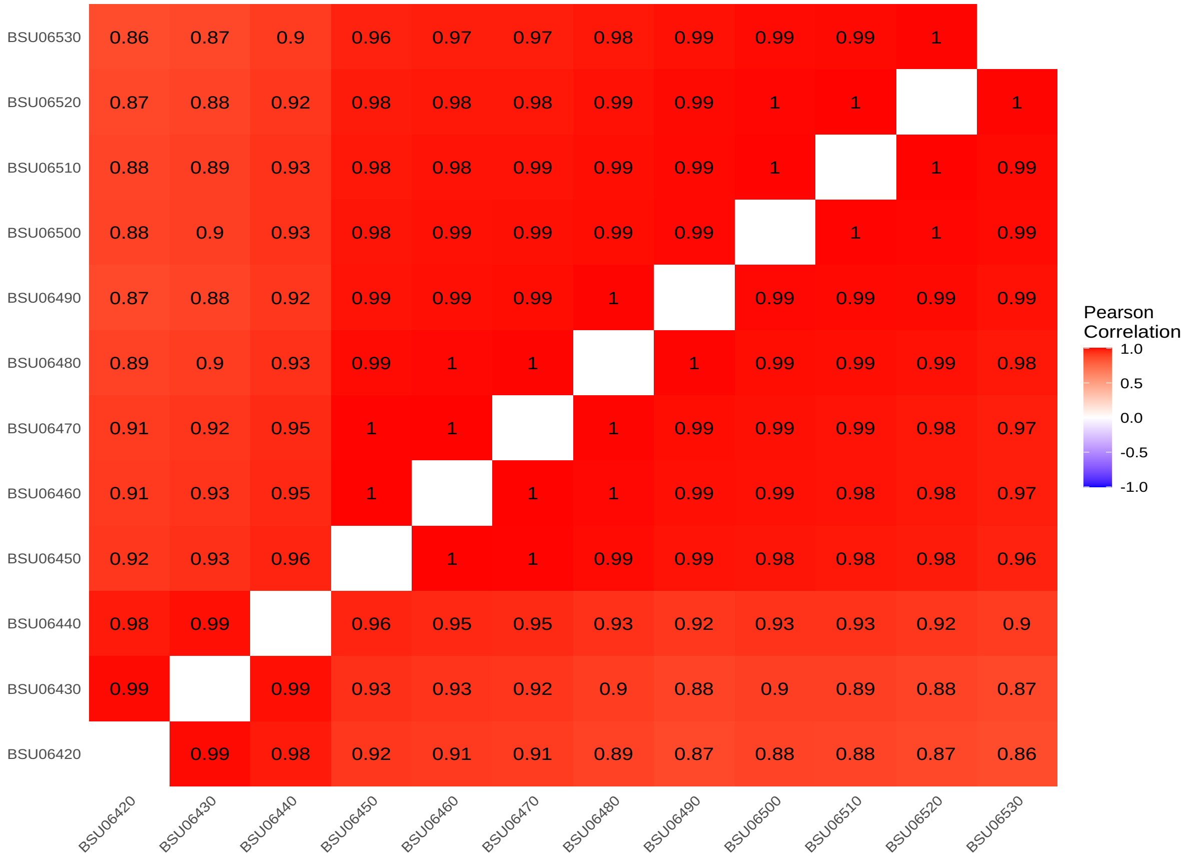


All genes as predictors

Regulators
100% of edges

Sigma factors
83% of edges

Highly correlated genes in operons



	BSU06420	BSU06430	BSU06440	BSU06450	BSU06460	BSU06470	BSU06480	BSU06490	BSU06500	BSU06510	BSU06520	BSU06530
BSU06530	0.86	0.87	0.9	0.96	0.97	0.97	0.98	0.99	0.99	0.99	1	
BSU06520	0.87	0.88	0.92	0.98	0.98	0.98	0.99	0.99	1	1		1
BSU06510	0.88	0.89	0.93	0.98	0.98	0.99	0.99	0.99	1		1	0.99
BSU06500	0.88	0.9	0.93	0.98	0.99	0.99	0.99	0.99		1	1	0.99
BSU06490	0.87	0.88	0.92	0.99	0.99	0.99	1		0.99	0.99	0.99	0.99
BSU06480	0.89	0.9	0.93	0.99	1	1		1	0.99	0.99	0.99	0.98
BSU06470	0.91	0.92	0.95	1	1		1	0.99	0.99	0.99	0.98	0.97
BSU06460	0.91	0.93	0.95	1		1	1	0.99	0.99	0.98	0.98	0.97
BSU06450	0.92	0.93	0.96		1	1	0.99	0.99	0.98	0.98	0.98	0.96
BSU06440	0.98	0.99		0.96	0.95	0.95	0.93	0.92	0.93	0.93	0.92	0.9
BSU06430	0.99		0.99	0.93	0.93	0.92	0.9	0.88	0.9	0.89	0.88	0.87
BSU06420		0.99	0.98	0.92	0.91	0.91	0.89	0.87	0.88	0.88	0.87	0.86

purine nucleotide synthesis operon

Merge genes in operons (work in progress)

- But: some annotated operons contain low correlated genes (Nicolas et al. 2012)
- Solution: Automatically define operons using adjacency-constrained hierarchical clustering (`adjclust` Ambroise et al. (2019))

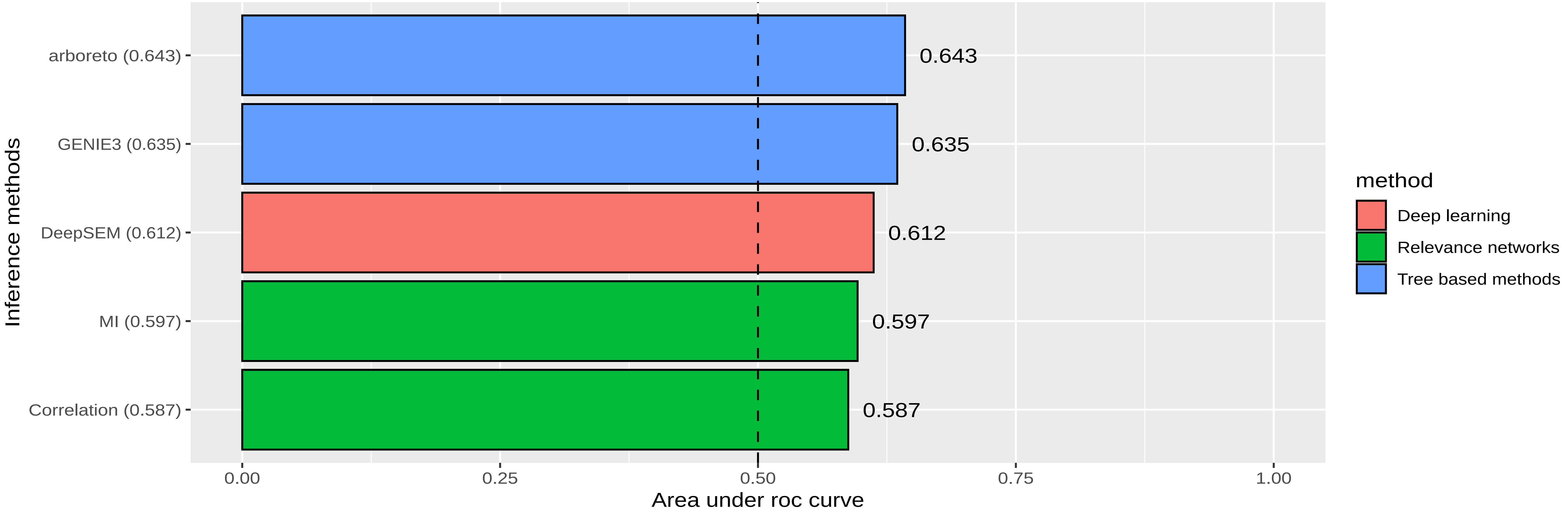
How to evaluate and interpret an inferred network ?

1. Evaluate **network topology** instead of edges:
   - Clustering comparisons
   - Distances with graph kernels (`graphkernels` Sugiyama et al. (2018))

2. Use biological annotation:
   - Motif / pathway evaluation using edge ranking (adaptation of Marbach et al. (2012))
   - Precision/Recall by regulation types
   - Functional enrichment analysis of modules in networks

First result:

evaluation of edges (AUROC) on the complete dataset



ROC curves and AUROC for 3 methods + variations in comparison with the real network

Conclusions

- Bad results on AUROC:
  - On the complete dataset
  - dataset contains highly correlated variables (operons)

Perspectives

- How do methods work on expression regulated pathways (*e.g.*, sporulation)?
- More evaluation criteria
- Modify inference methods to integrate metabolomics information (effectors)

References

Ambroise, Christophe, Alia Dehman, Pierre Neuvial, Guillem Rigaill, and Nathalie Vialaneix. 2019. “Adjacency-Constrained Hierarchical Clustering of a Band Similarity Matrix with Application to Genomics.” *Algorithms for Molecular Biology* 14: 22. https://doi.org/10.1186/s13015-019-0157-4.

Butte, A. J., and I. S. Kohane. 1999. “Unsupervised knowledge discovery in medical databases using relevance networks.” *Proc AMIA Symp*, 711–15.

———. 2000. “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.” *Pac Symp Biocomput*, 418–29.

Chan, Thalia E, Michael PH Stumpf, and Ann C Babtie. 2017. “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures.” *Cell Systems* 5 (3): 251–67.

Chen, S., and J. C. Mar. 2018. “Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data.” *BMC Bioinformatics* 19 (1): 232.

Faith, Jeremiah J, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. 2007. “Large-Scale Mapping and Validation of Escherichia Coli Transcriptional Regulation from a Compendium of Expression Profiles.” *PLoS Biology* 5 (1): e8.

Faria, José P, Ross Overbeek, Ronald C Taylor, Neal Conrad, Veronika Vonstein, Anne Goelzer, Vincent Fromion, Miguel Rocha, Isabel Rocha, and Christopher S Henry. 2016. “Reconstruction of the Regulatory Network for Bacillus Subtilis and Reconciliation with Gene Expression Data.” *Frontiers in Microbiology* 7: 275.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2019. *Glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. https://CRAN.R-project.org/package=glasso.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. “Sparse inverse covariance estimation with the graphical lasso.” *Biostatistics* 9 (3): 432–41.

Goelzer, Anne. 2010. “Emergence de Structures Modulaires Dans Les Régulations Des Systèmes Biologiques : Théorie Et Applications à Bacillus Subtilis.” PhD thesis. http://www.theses.fr/2010ECDL0030/document.

Huynh-Thu, V. A., A. Irrthum, L. Wehenkel, and P. Geurts. 2010. “Inferring regulatory networks from expression data using tree-based methods.” *PLoS One* 5 (9).

Jiang, Haoming, Xinyu Fei, Han Liu, Kathryn Roeder, John Lafferty, Larry Wasserman, Xingguo Li, and Tuo Zhao. 2021. *Huge: High-Dimensional Undirected Graph Estimation*. https://CRAN.R-project.org/package=huge.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An r Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9 (1): 1–13.

Marbach, D., J. C. Costello, R. ffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, et al. 2012. “Wisdom of crowds for robust gene network inference.” *Nat Methods* 9 (8): 796–804.

Meinshausen, Nicolai, and Peter Bühlmann. 2006. “High-dimensional graphs and variable selection with the Lasso.” *The Annals of Statistics* 34 (3): 1436–62. https://doi.org/10.1214/009053606000000281.

Meyer, Patrick E, Frederic Lafitte, and Gianluca Bontempi. 2008. “Minet: AR/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information.” *BMC Bioinformatics* 9: 1–10.

Moerman, Thomas, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. 2019. “GRNBoost2 and Arboreto: Efficient and Scalable Inference of Gene Regulatory Networks.” *Bioinformatics* 35 (12): 2159–61.

Mölder, F, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, et al. 2021. “Sustainable Data Analysis with Snakemake [Version 2; Peer Review: 2 Approved].” *F1000Research* 10 (33). https://doi.org/10.12688/f1000research.29032.2.

Nicolas, Pierre, Ulrike Mäder, Etienne Dervyn, Tatiana Rochat, Aurélie Leduc, Nathalie Pigeonneau, Elena Bidnenko, et al. 2012. “Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus Subtilis.” *Science* 335 (6072): 1103–6. https://doi.org/10.1126/science.1206848.

Reverter, Antonio, and Eva K. F. Chan. 2008. “Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks.” *Bioinformatics* 24 (21): 2491–97.

Saint-Antoine, Michael, and Abhyudai Singh. 2023. “Benchmarking Gene Regulatory Network Inference Methods on Simulated and Experimental Data.” *bioRxiv*. https://doi.org/10.1101/2023.05.12.540581.

Shu, Hantao, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. 2021. “Modeling Gene Regulatory Networks Using Neural Network Architectures.” *Nature Computational Science* 1 (7): 491–501.

Skok Gibbs, Claudia, Christopher A Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, et al. 2022. “High-Performance Single-Cell Gene Regulatory Network Inference at Scale: The Inferelator 3.0.” *Bioinformatics* 38 (9): 2519–28.

Sugiyama, Mahito, M Elisabetta Ghisu, Felipe Llinares-López, and Karsten Borgwardt. 2018. “Graphkernels: R and Python Packages for Graph Comparison.” *Bioinformatics* 34 (3): 530–32.

Yu, Yue, Jie Chen, Tian Gao, and Mo Yu. 2019. “DAG-GNN: DAG Structure Learning with Graph Neural Networks.” In *International Conference on Machine Learning*, 7154–63. PMLR.

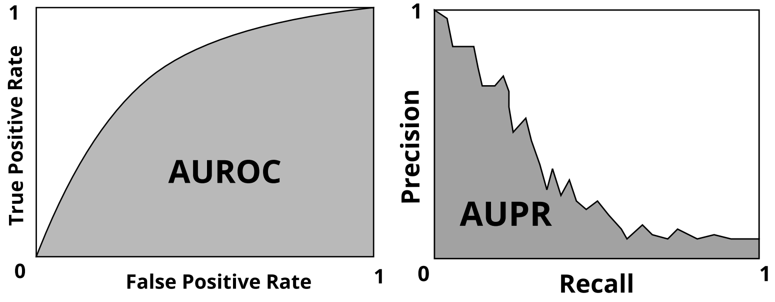
Vision d’ensemble des métriques envisagées

Question: Comment évaluer et interpréter un réseau inféré ?

Évaluation des arêtes

We apply the classic evaluation methods with the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN):

	Edge Exists	Edge Does NOT Exist
Edge Predicted	True Positive (TP)	False Positive (FP)
Edge NOT Predicted	False Negative (FN)	True Negative (TN)



(Saint-Antoine and Singh 2023)

Comparison between the inferred network and the Ground truth :

- $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$ for ROC curves.
- $Precision = \frac{TP}{TP+FP}$ and $Recall = TPR$ for PR curves.

Area Under Curves (AUC) Allows to visualize how the method behaves when you vary its threshold.