Ensemble Learning Based Gene Regulatory Network Inference

Sergio Peignier[a], Baptiste Sorin, Federica Calevro

Univ Lyon, INSA Lyon, INRAE
BF2I, UMR0203, F-69621
Lyon, France

[a] sergio.peignier@insa-lyon.fr

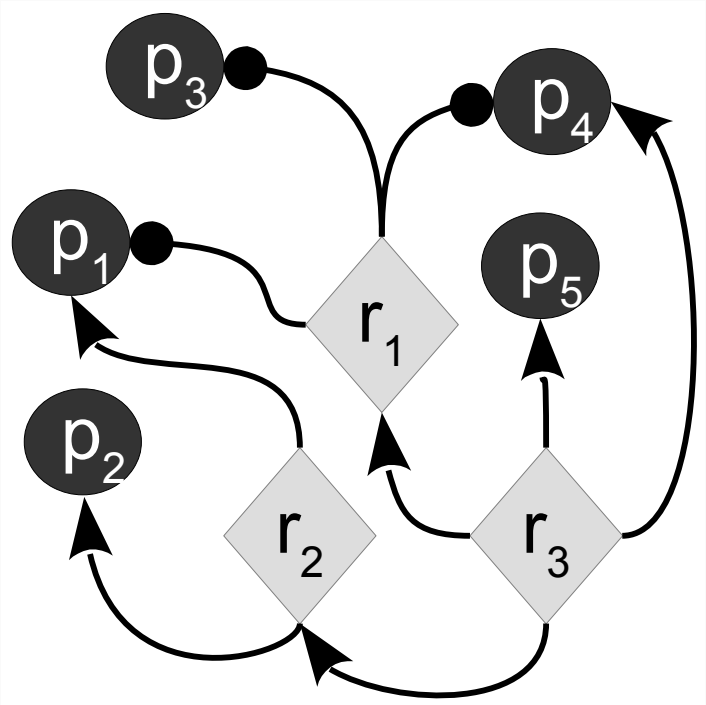
Gene Regulatory Networks (GRNs)

Central Dogma of Molecular Biology

$$Gene \xrightarrow{\text{Transcription}} mRNA \xrightarrow{\text{Translation}} Protein$$

Definition

Set of **interacting molecular regulators** (e.g. transcription factors) **controlling** the **creation** of **gene products** (e.g. mRNA, proteins).



Important biological role:

- Adaptation
- Versatility
- Differentiation
- Morphogenesis ...

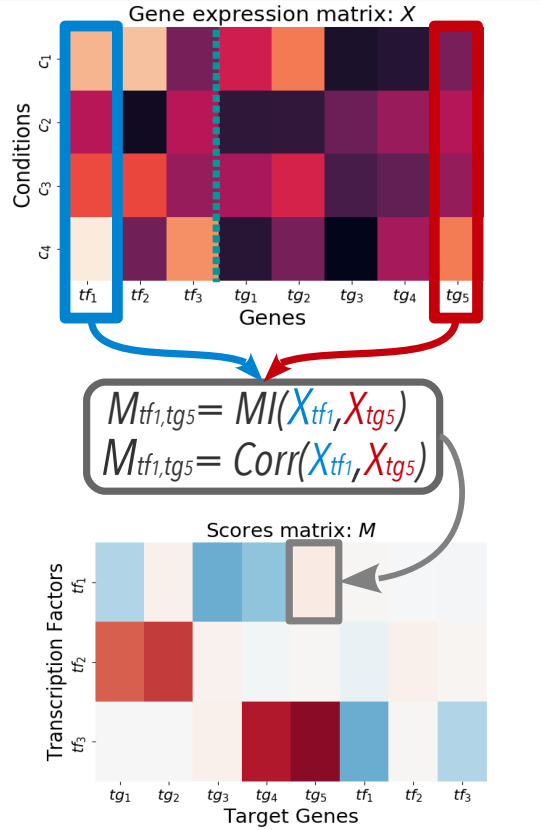
Data-driven GRN Inference

General principle

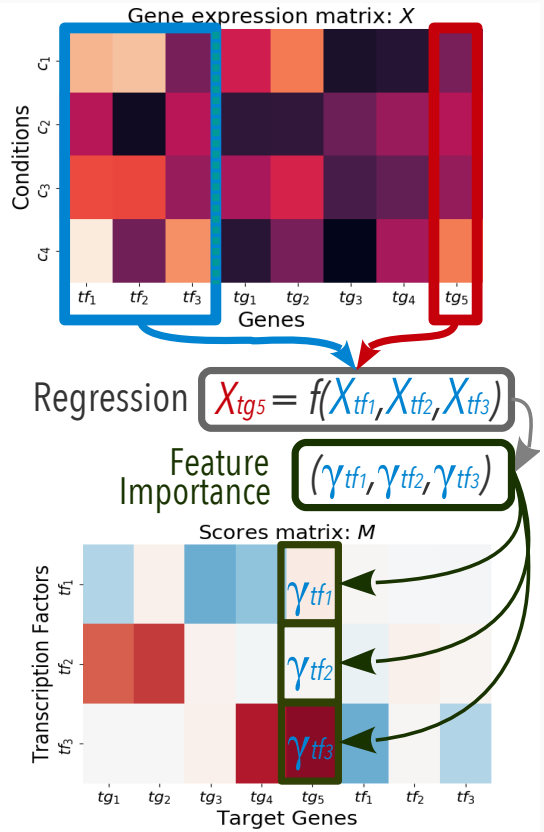
- Based on high-throughput gene expression data.
- Score possible links between:
  - Regulators, i.e. Transcription Factors (TFs)
  - Target Genes (TGs)
- Select most promising links.

Data-driven GRN Inference Families

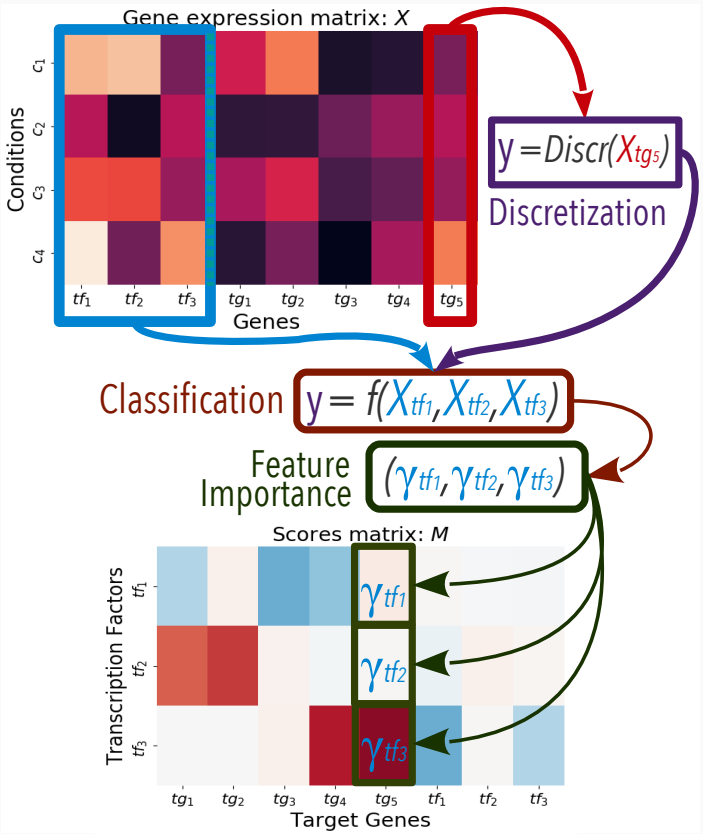
Correlation | Mutual Information



Regression Methods



Classification Methods



Ensembles of GRN inference methods

"Community" Ensemble of **35 methods** [Marbach et al. 2012]

Advantages:

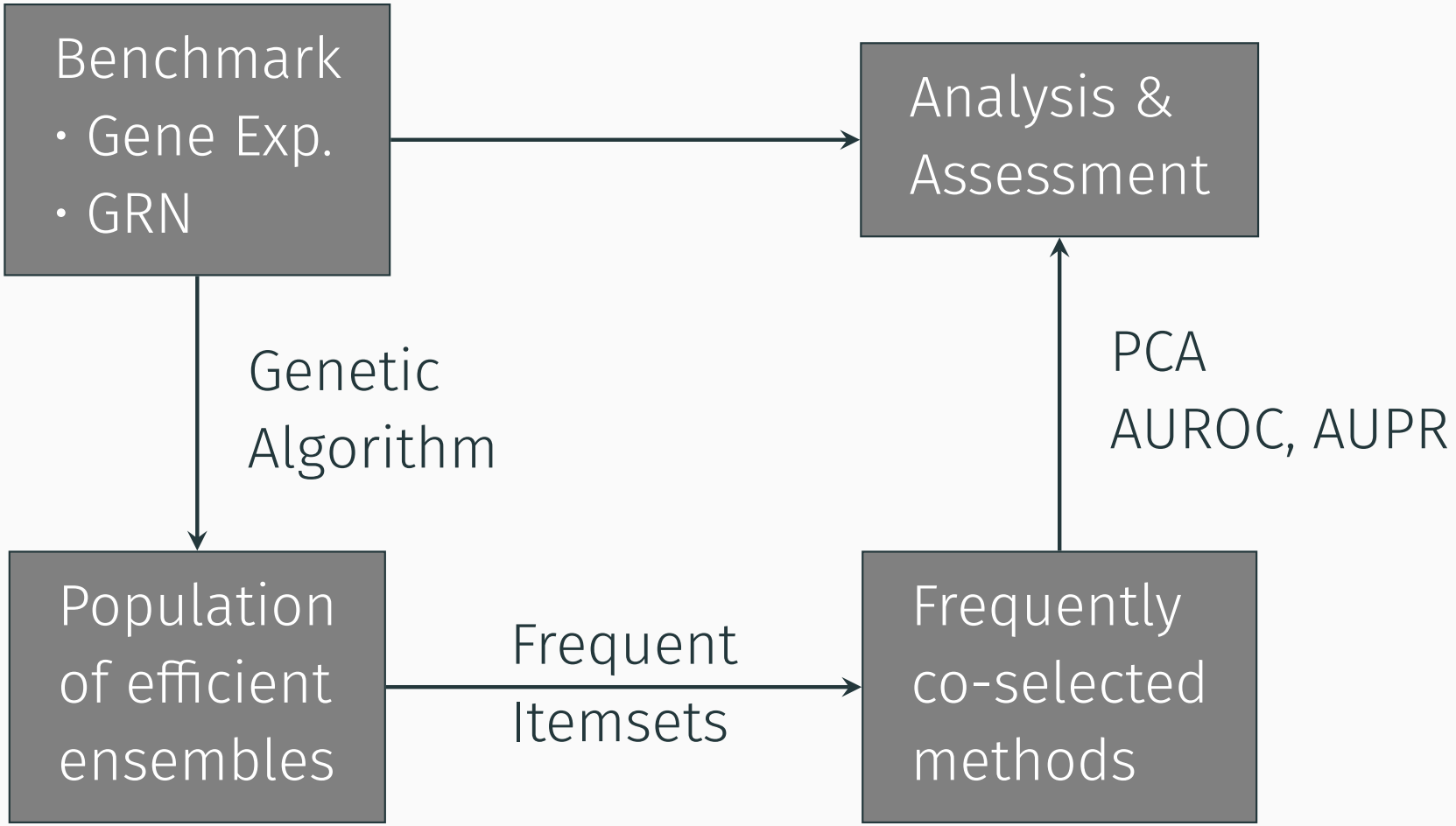
- **Better results** than base learners
- **Robust** across datasets

Drawbacks:

- **Higher runtimes**
- **More computer resources**
- Some methods could be **detrimental**

Objective

Design **small**, **robust** and **efficient ensembles** of **GRN inference** methods



Benchmark Evaluation Setup

- **Evaluation protocol** designed by **[Marbach et al. 2012]**
- **Standard evaluation measures**: **AUROC** and **AUPR**
- DREAM5 **Benchmark datasets**

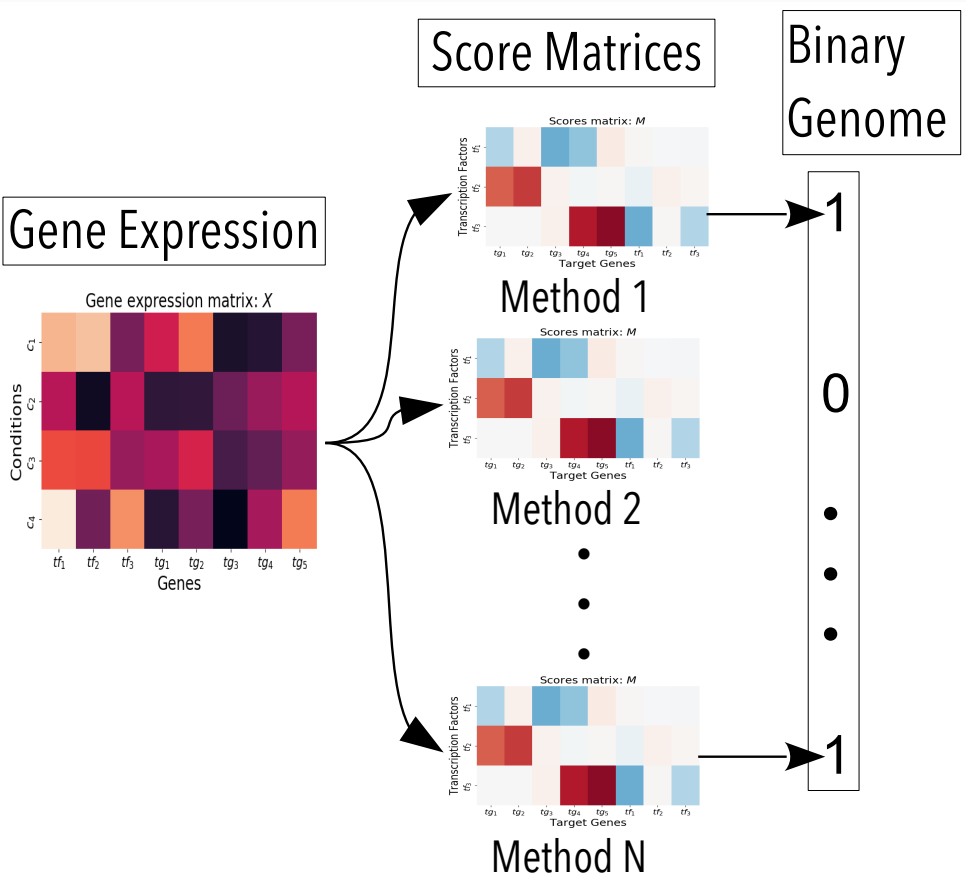
Dataset	Data	# cond.	# genes	# TFs	# Links	$\frac{\#TrueLinks}{\#Links}$
InSilico	Simulated	805	1,643	195	4,012	0.014
S.aureus	Microarray	160	2,810	99	515	0.028
E.coli	Microarray	805	4,511	334	2,066	0.013
S.cerevisiae	Microarray	536	5,950	333	3,940	0.017

GRN Inference Base Methods

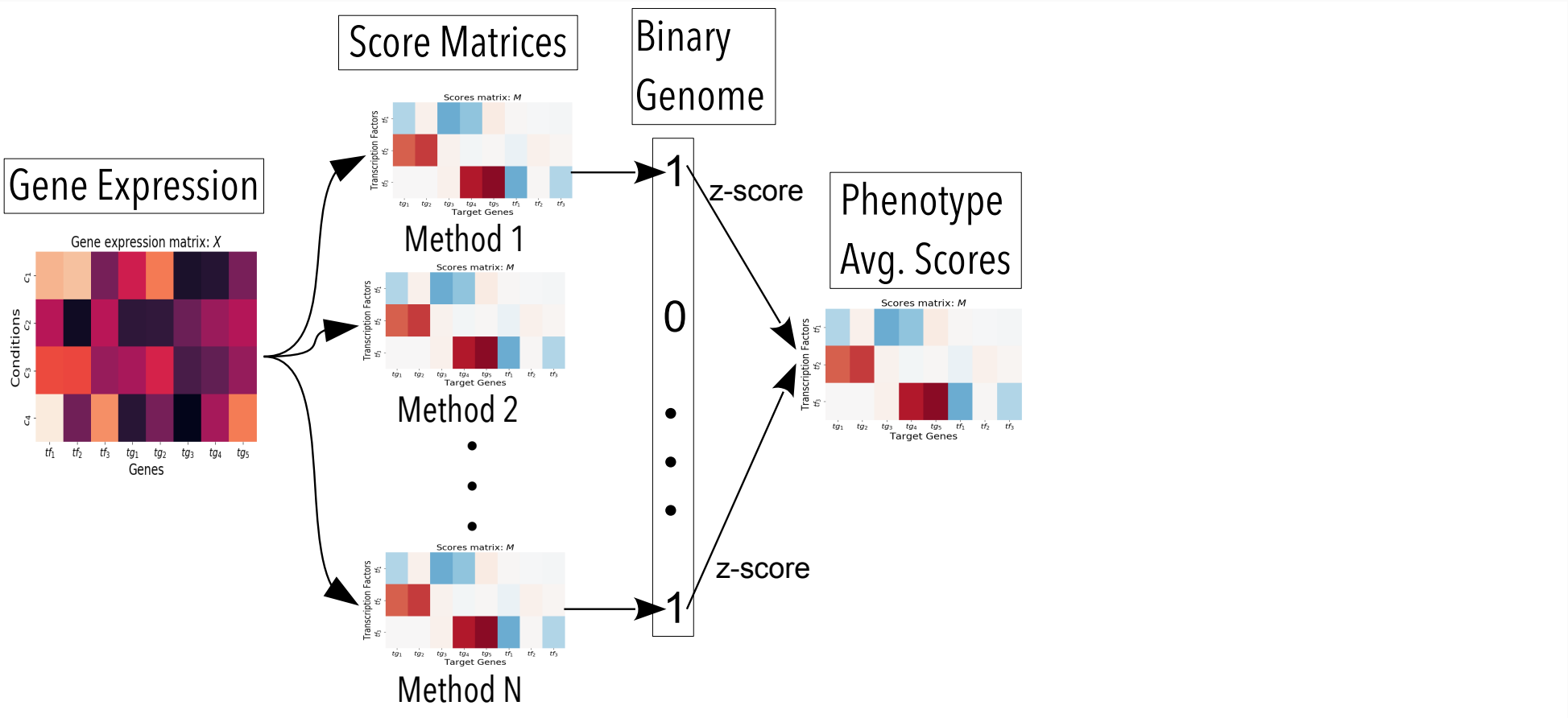
GRN Inference Methods: 17 methods from the **GReNaDIne** library based on **SciPy** and **Scikit-learn** methods:

- **Pearson** corr. (Pcorr), **Spearman** corr. (Scorr), **Kendall-tau** (Ktau) **Mutual Information** (MI).
- Linear **SVM**, **Random Forest**, **Gradient Boosting**, **AdaBoost** regressors and classifiers (SVMr,SVMc,RFr,RFc,GBr,GBc,ABr,ABc)
- Linear Regressors Stability: **TIGRESS** and **Stability Randomized Lasso** (SRLr)
- **BayesianRidge** Regression (BRSr)

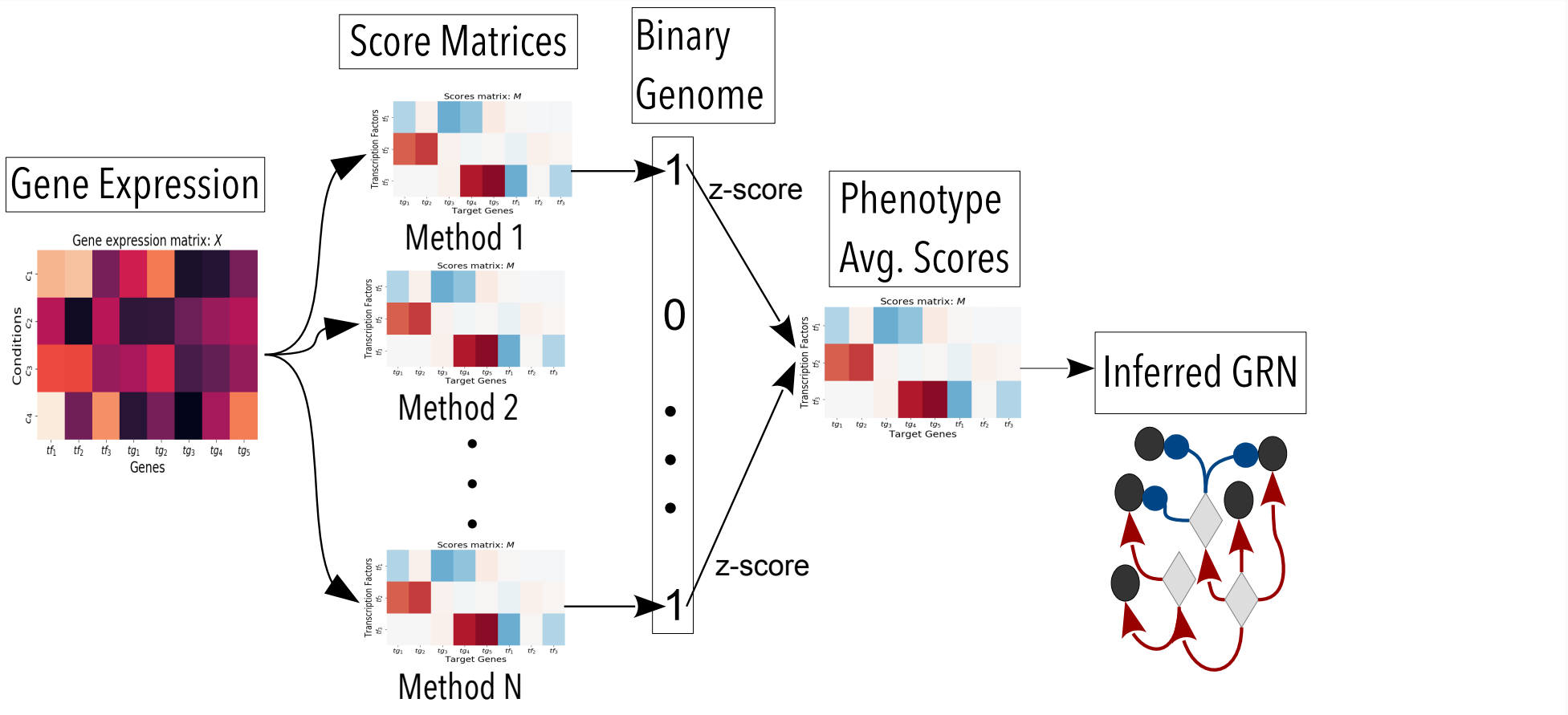
Genotype → Phenotype → Fitness



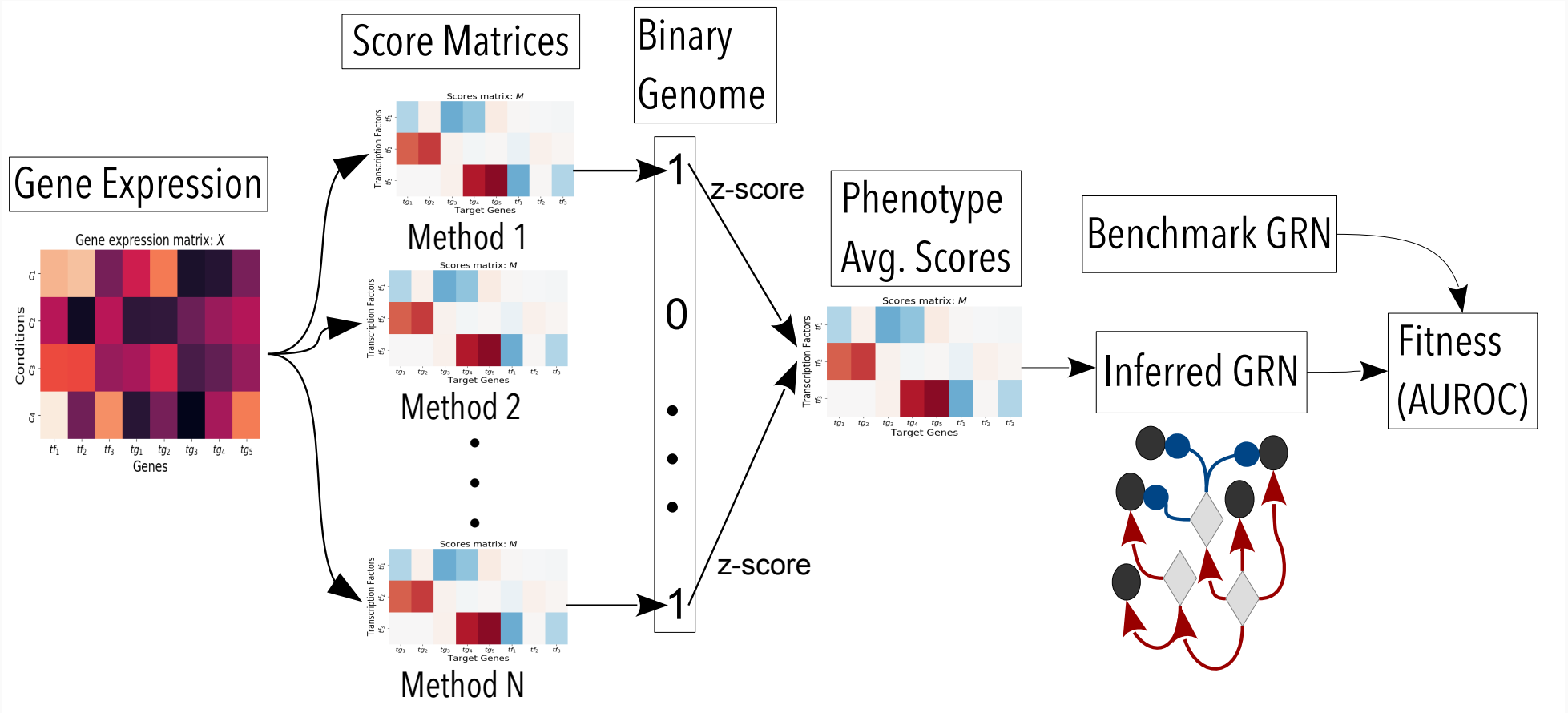
Genotype → Phenotype → Fitness



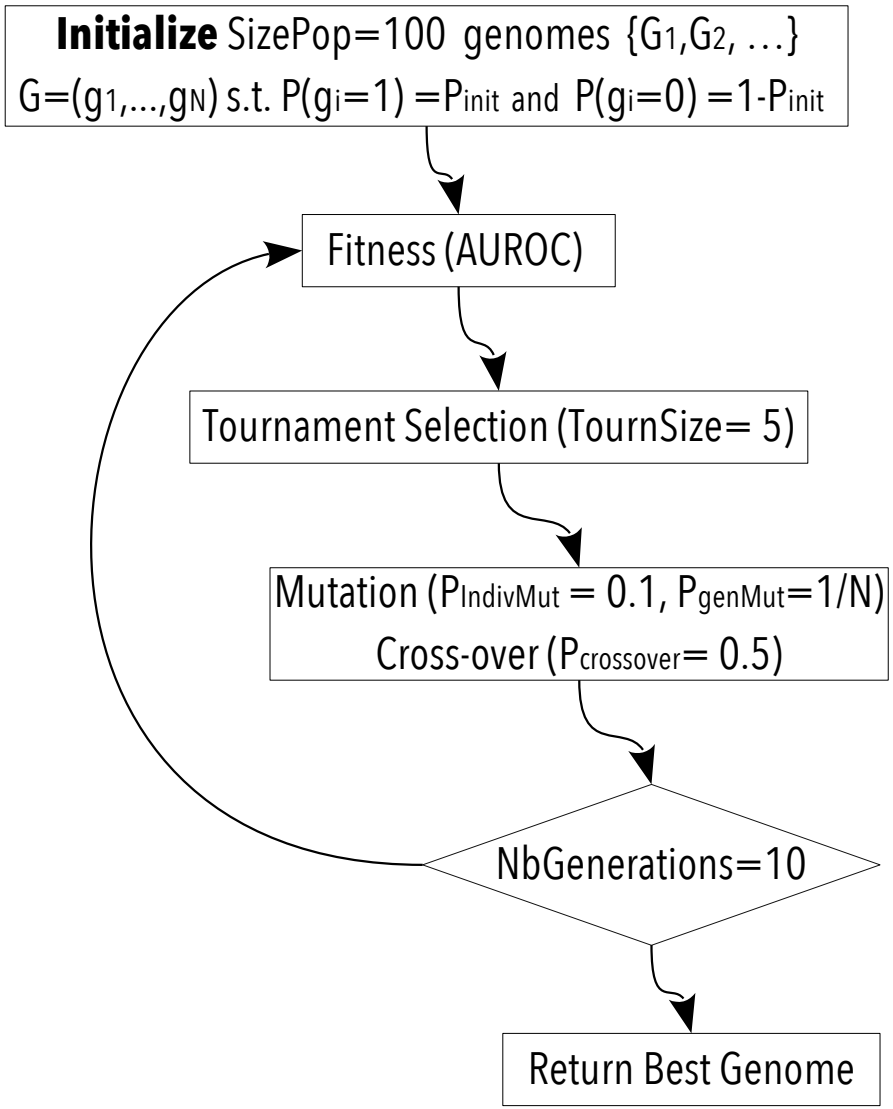
Genotype → Phenotype → Fitness



Genotype → Phenotype → Fitness



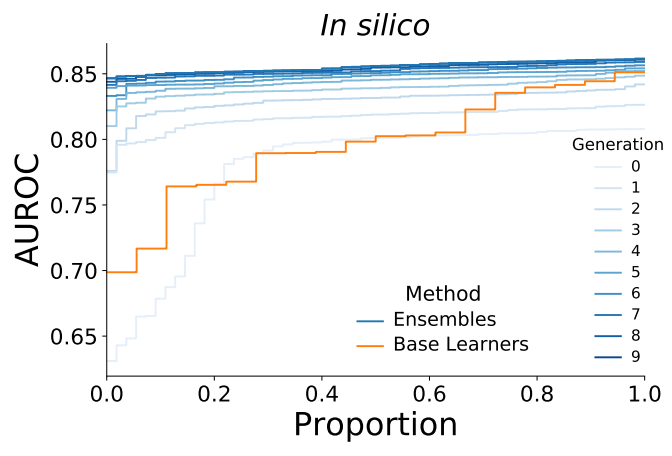
Efficient Ensemble Candidates Population Evolution

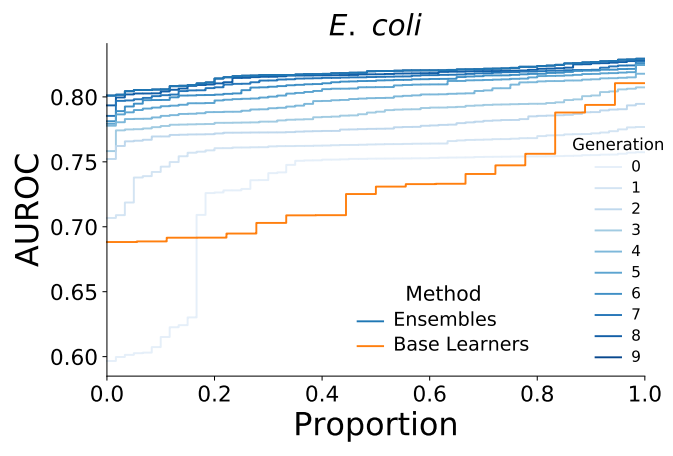


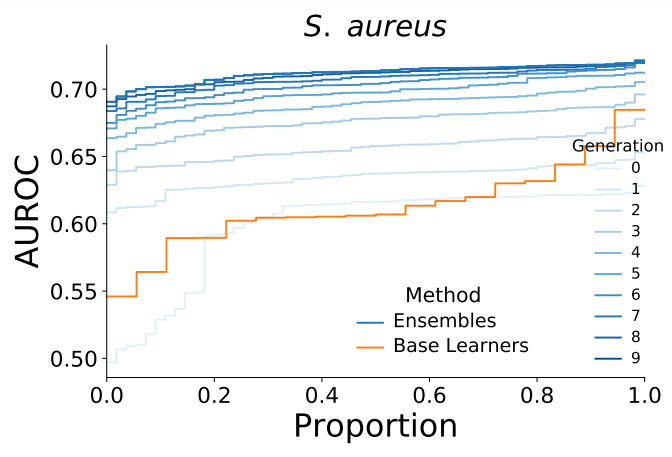
Run **10 indep. populations** for:

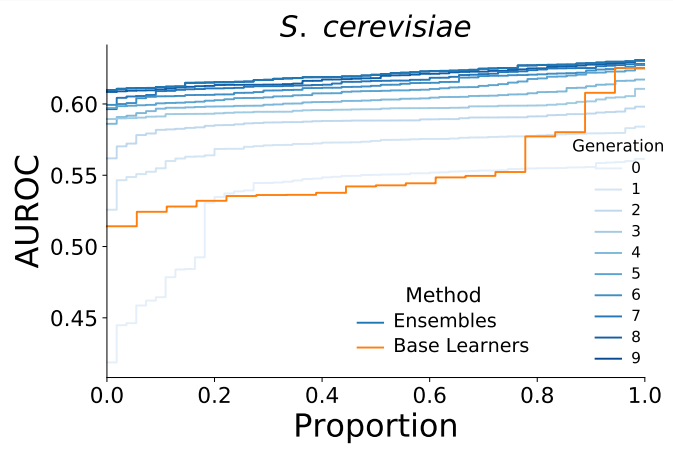
- **Each dataset**
- $P_{Init} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$

Results: Genetic Algorithm







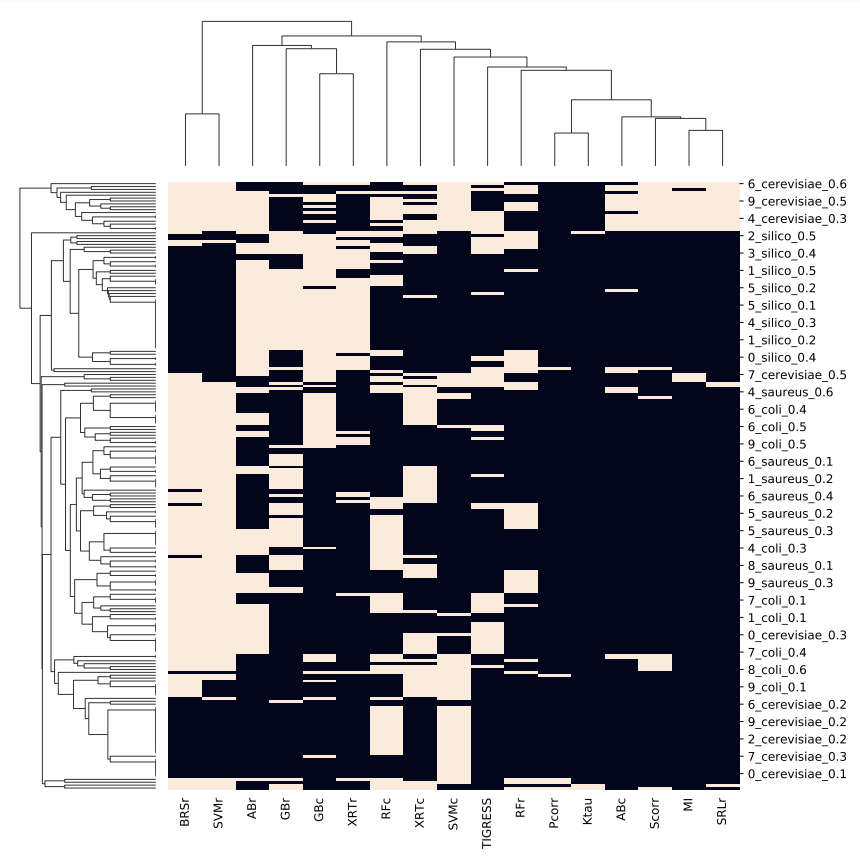


- 10 generations were **sufficient**.
- **Last generation ensembles** are **better** than the **best single-method**

Maximal Frequent Itemsets

Goal : Detect the subsets of frequently co-selected methods.

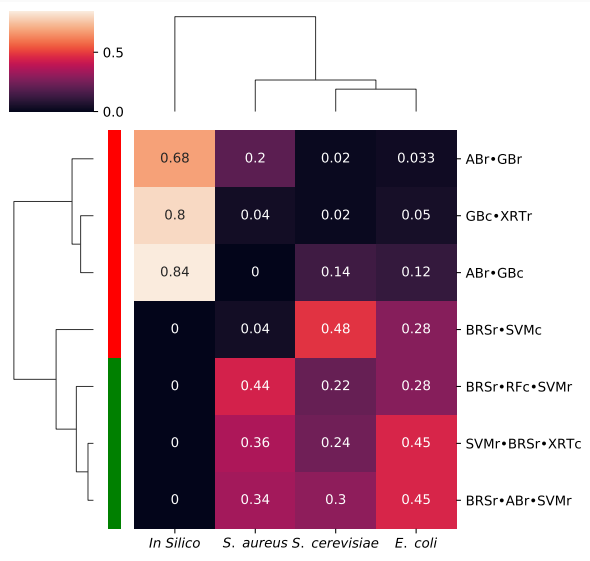
- GRN Inference **method** $\rightarrow$ **Item**
- **Individual** using methods $\mathcal{M}_t \rightarrow$ **itemset/transaction**
- **Population** $\rightarrow$ **transactions** dataset $\mathcal{T} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T\}$

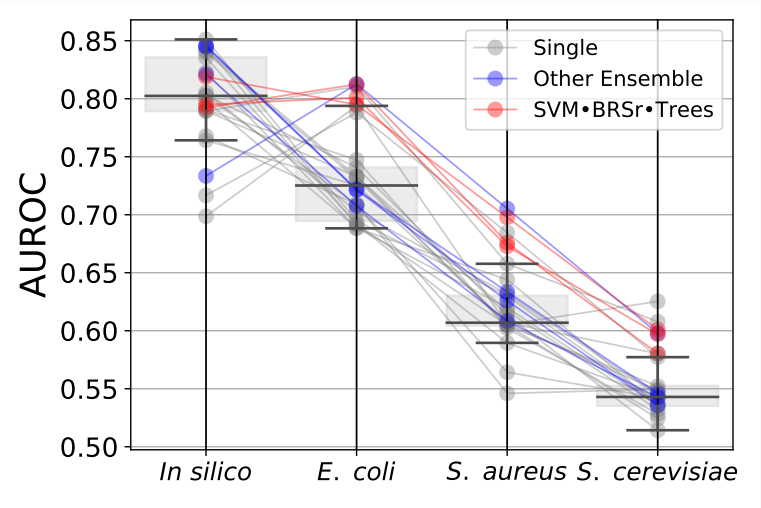


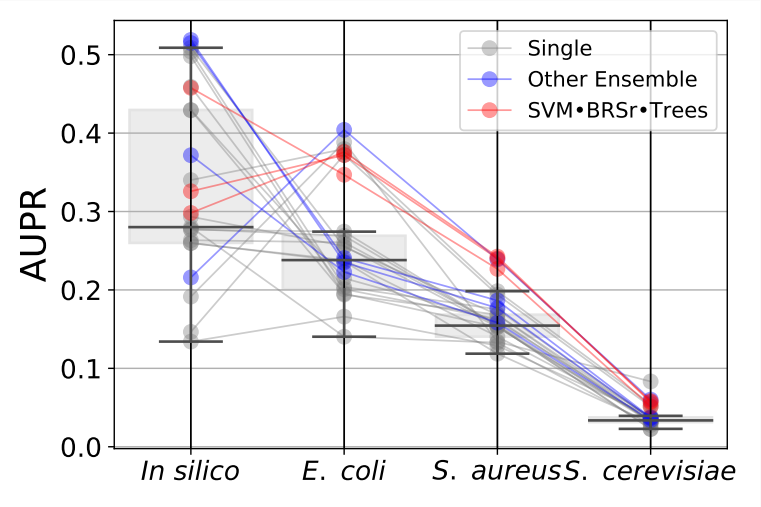
- **Support** of Itemset $\mathcal{M}$
$supp(\mathcal{M}, \mathcal{T}) = \frac{|\{\mathcal{M}_t \in \mathcal{T} \mid \mathcal{M} \subseteq \mathcal{M}_t\}|}{|\mathcal{T}|}$
- $\mathcal{M}$ is **frequent itemset** if
$supp(\mathcal{M}, \mathcal{T}) > MinSupp$
- $\mathcal{M}$ is **Maximal frequent itemset** if
$\nexists \mathcal{M}' \in \mathcal{T}$ s.t. $\mathcal{M} \subset \mathcal{M}'$ and
$supp(\mathcal{M}', \mathcal{T}) > MinSupp$
- In practice **FP-max** algorithm was used with $MinSupp = 0.2$

Results: Maximal Frequent Itemsets

Support



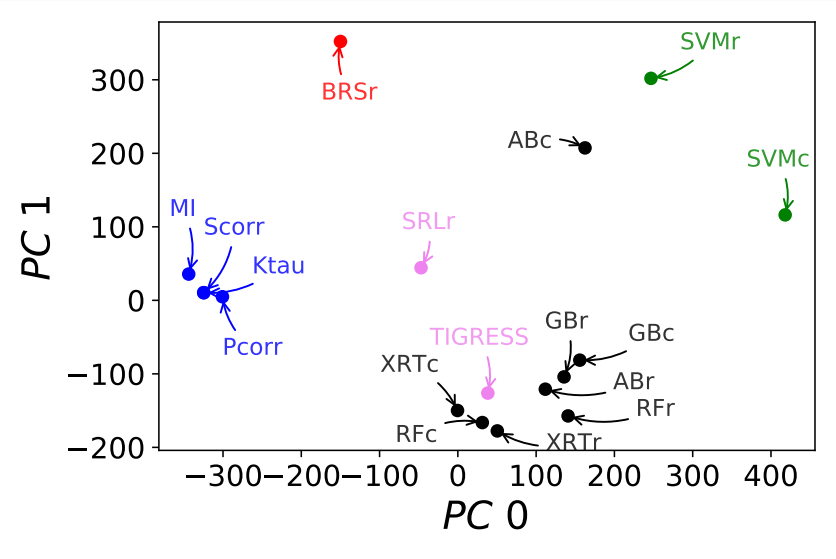




- **Tree-based ensembles**
  - Selected in simulated datase
  - Best results for Simulated dataset
  - Mediocre results for real datasets
- **BRSr•SVMr•Trees**
  - Mainkly selected in real datasets
  - Best results in real datasets
  - Decent results in simulated dataset

Base learners diversity exploration

PCA-based **base learners relatedness** evaluation [Marbach et al. 2012]



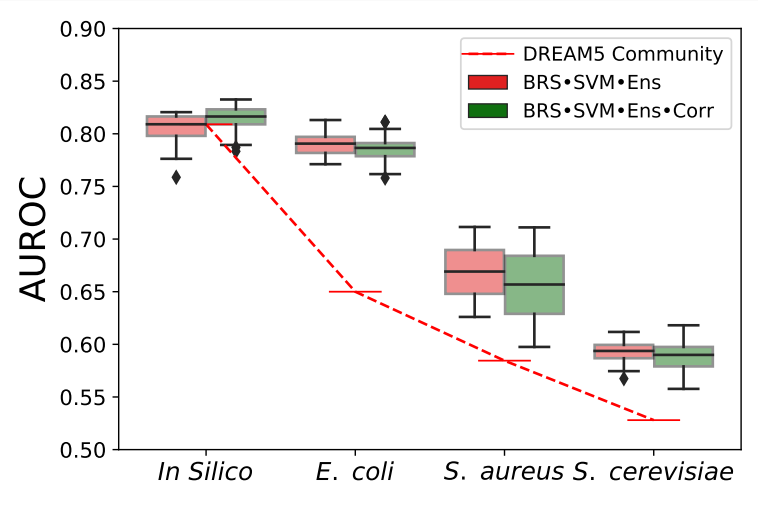
- **Close methods** are likely to share the **same intrinsic biases**
- **Performance** and **robustness** increase with **base learners diversity**
- BRSr·SVMr·Trees $\rightarrow$ **diverse methods**

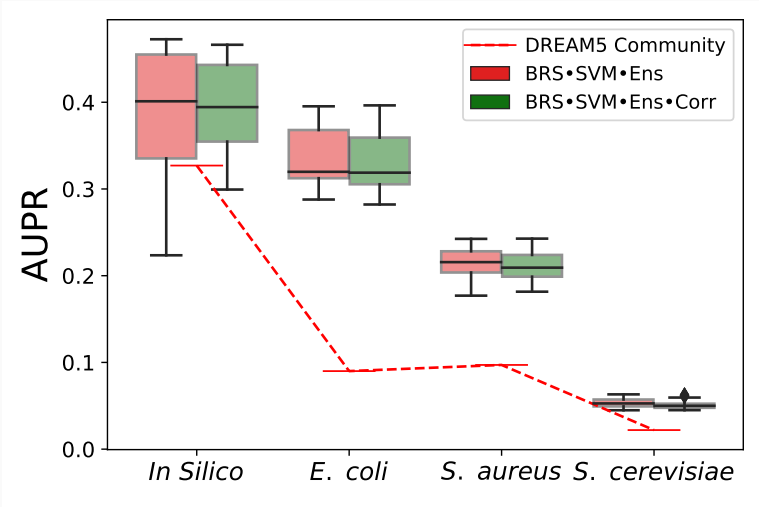
Promising combinations:

- SVM + BRSr + tree-based, TIGRESS or SRLr
- SVM + BRSr + tree-based, TIGRESS or SRLr + correlation method

Promising Enssembles Assessment

- **BRSr•SVM•Ens**: BRSr + SVM + tree-based, TIGRESS or SRLr
- **BRSr•SVM•Ens•Corr** : BRSr•SVM•Ens + correlation-based method





Results

- **Small** ensembles can **outperform large** DREAM5 community
- Adding Corr. method does not improve results (for these datasets)

Conclusion

Summary

- **Small** and **diverse** ensembles of methods could be **sufficient** to:
  - **Improve prediction quality**
  - **Improve Robustness** across datasets
  - **Limit computational requirements**
- **BRSr•SVM•Trees** $\rightarrow$ **efficient GRN inference ensemble**

Implementation

- **GReNaDIne Python package**:

Gene Regulatory Network Data-driven Inference

- **GitLab repository**:

gitlab.com/bf2i/grenadine

- **Documentation**:

grenadine.readthedocs.io