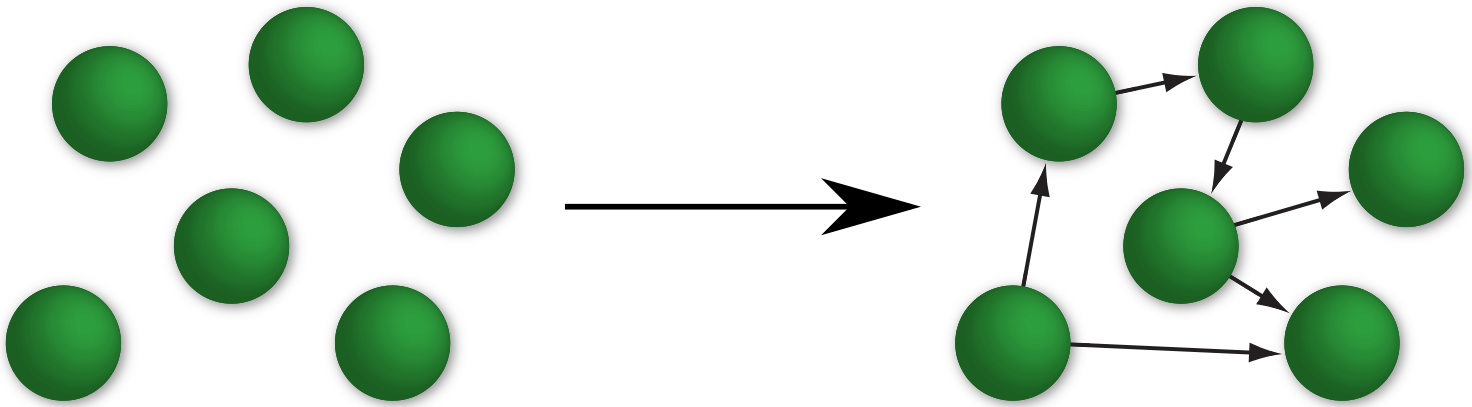
Gene regulatory network inference from expression and genetic data using tree-based methods

Vân Anh Huynh-Thu

STATSEQ meeting
March 28th, 2013

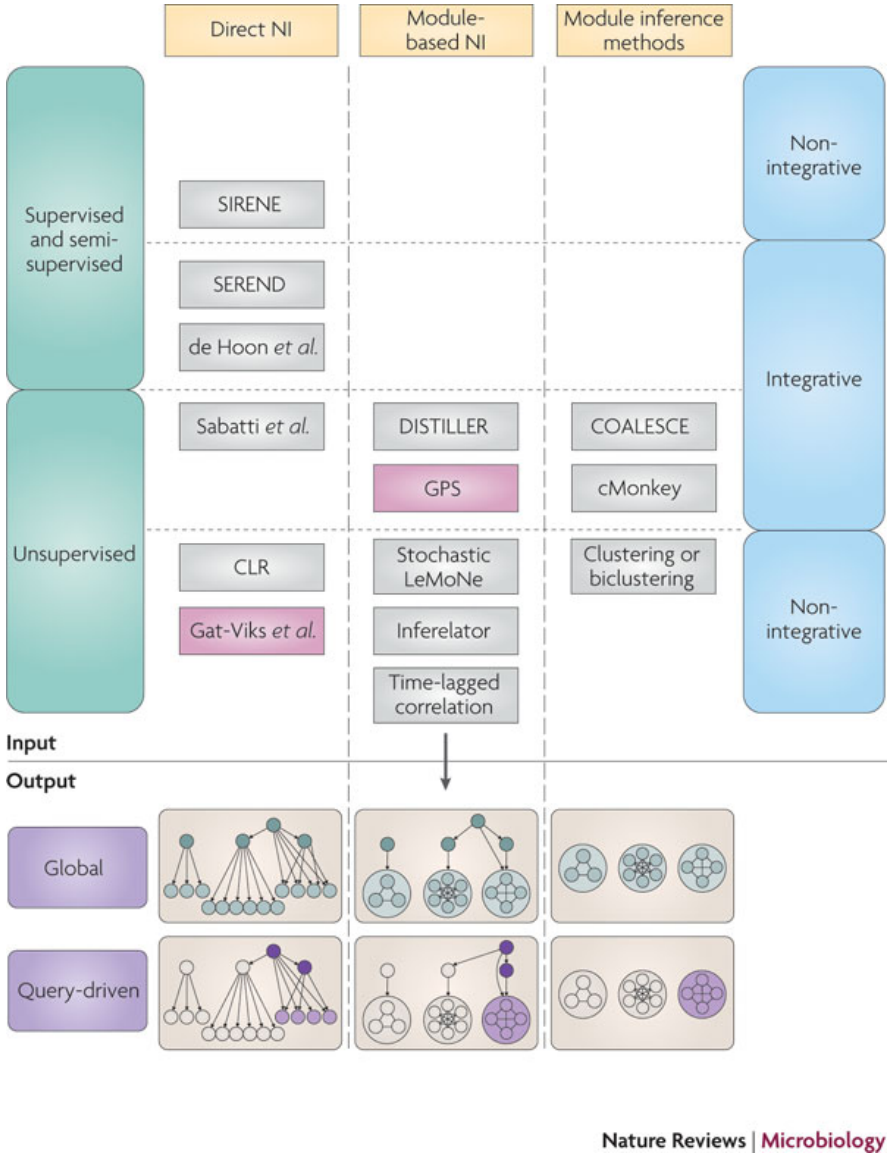
Inferring regulatory networks is a challenging problem



unknown network

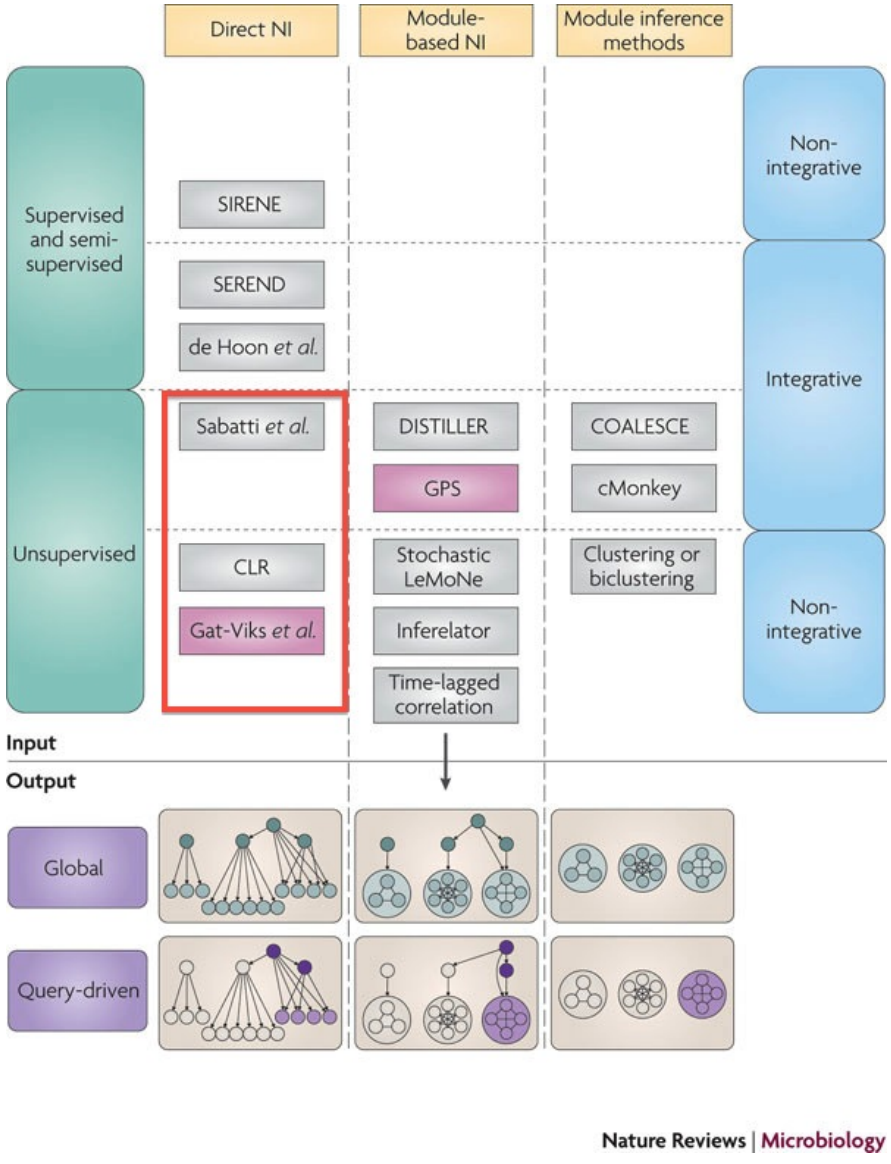
inferred network

Regulatory network inference methods



(De Smet & Marchal, *Nature Reviews Microbiology*, 2010)

Regulatory network inference methods

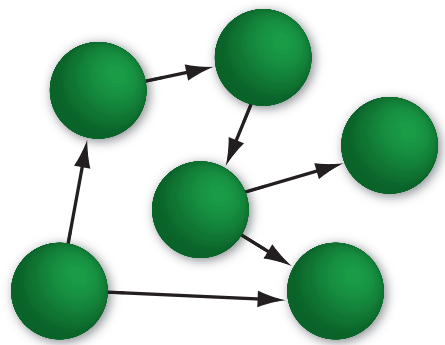


(De Smet & Marchal, *Nature Reviews Microbiology*, 2010)

Network inference with GENIE3

GENIE3 and time series

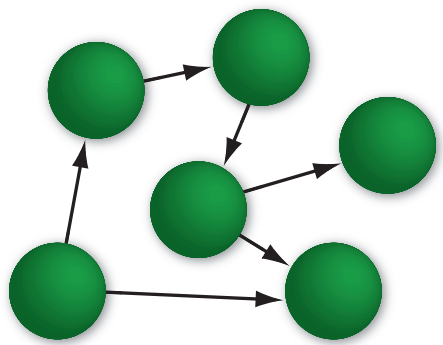
GENIE3 and systems genetics



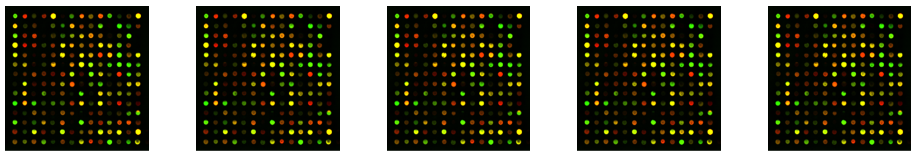
Network inference with GENIE3

GENIE3 and time series

GENIE3 and systems genetics



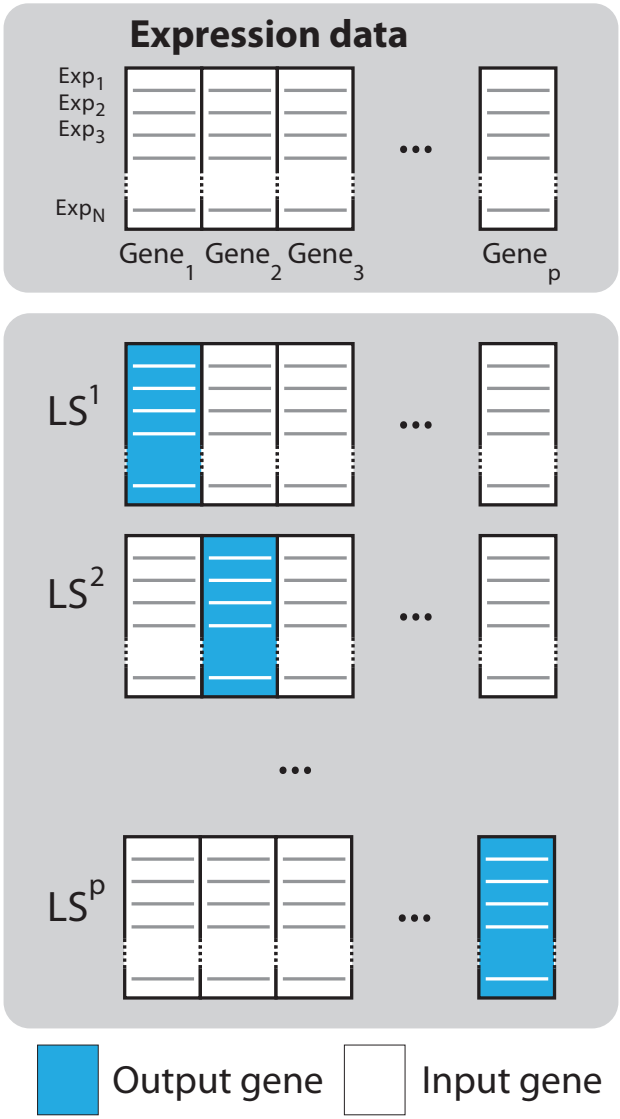
The goal of GENIE3 is to learn a weight for each edge



$\Downarrow$

		Target gene			
		gene 1	**gene 2**	$\cdots$	**gene p**
Regulating gene	**gene 1**	-	0.05	$\cdots$	0.56
	gene 2	0.19	-	$\cdots$	0.03
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	gene p	0.11	0.42	$\cdots$	-

Inference is decomposed into p sub-problems



Sub-problem i

$=$

Find the regulators of gene i

Supervised learning consists in extracting knowledge from input-output pairs

X_1	X_2	$\cdots$	X_m	Y
-0.61	0.41	$\cdots$	0.51	0.56
-2.3	0.1	$\cdots$	-0.21	0.43
0.33	-0.45	$\cdots$	0.3	-0.16
0.23	0.87	$\cdots$	0.09	0.71
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
-0.69	-0.61	$\cdots$	0.02	-0.75

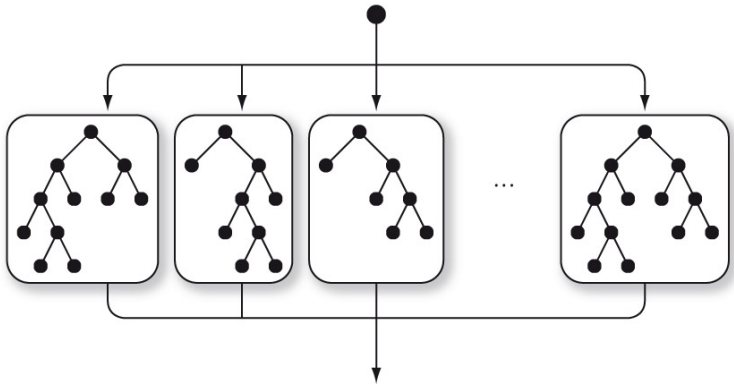
Model

Feature selection/ranking

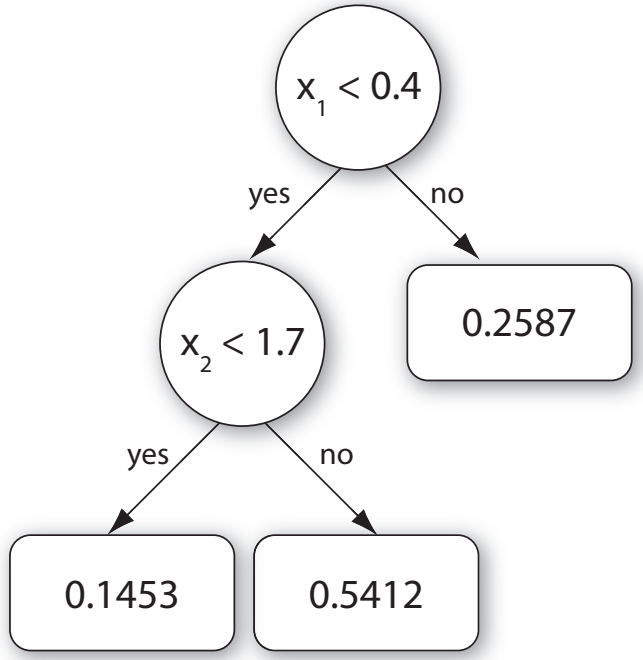
$$\hat{Y} = f(X_1, X_2, \ldots, X_m)$$

feat.	score
X_5	0.248
X_8	0.122
X_2	0.082
$\cdots$	$\cdots$
X_3	0.011

A tree-based ensemble method is used for supervised learning



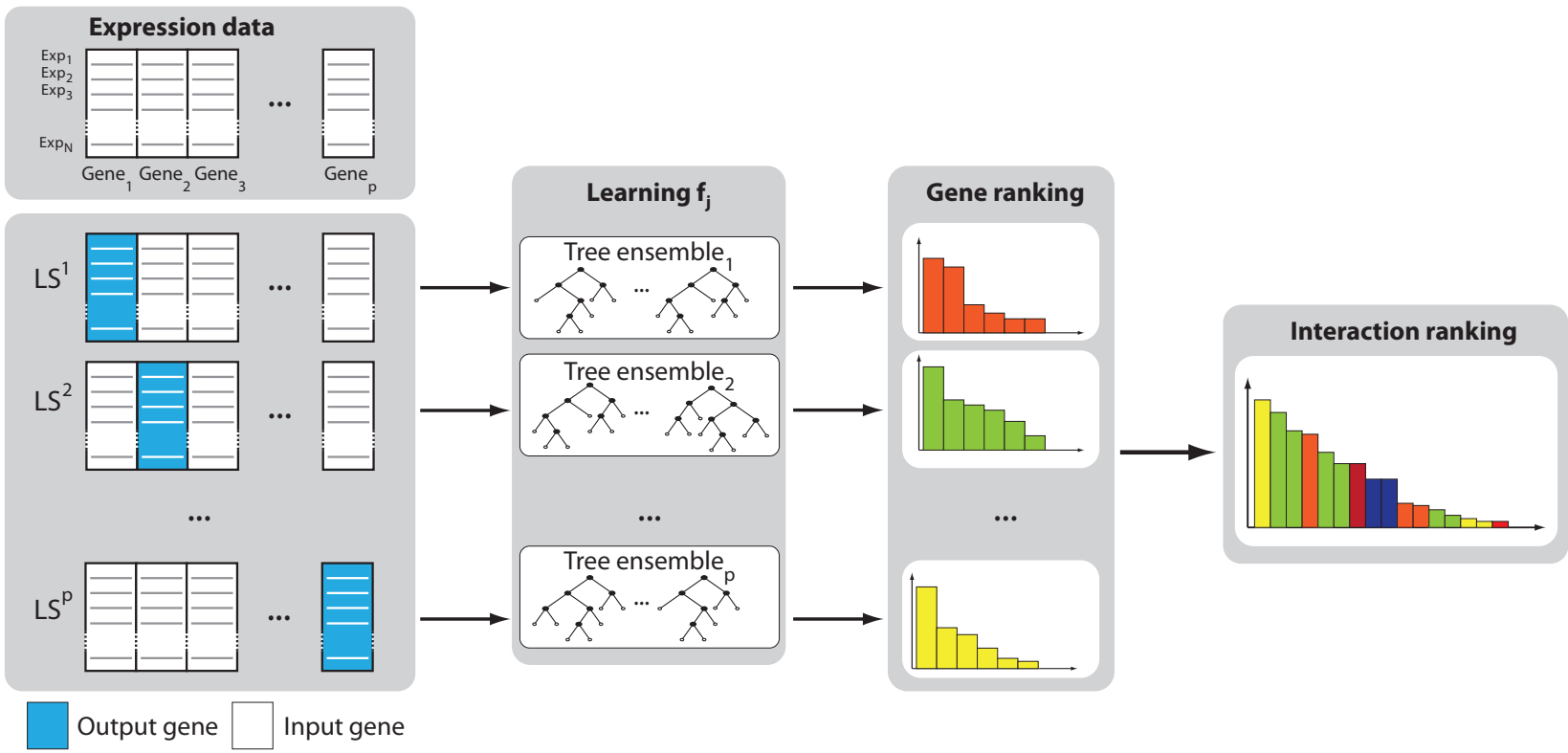
Bagging
Random Forests
Extra-Trees
...



Each interior node tests an input variable.

Each leaf node contains a predicted output value.

GENIE3: GEne Network Inference with Ensemble of trees



GENIE3 was the overall best performer in both challenges

DREAM4

5 synthetic networks
of 100 genes

Multifactorial
steady-state data

Rank	Team	Overall score
1	**GENIE3**	**37.428**
2	Team 549	28.165
3	Team 498	27.053
4	Team 395	26.139
5	Team 425	25.905
...	...	...

DREAM5

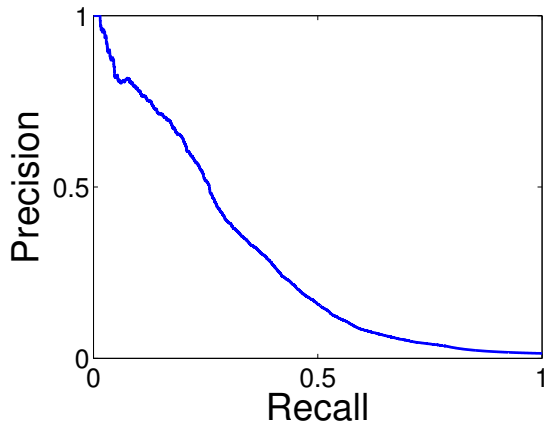
1 artificial and
2 real networks
(*E. coli* and *S. cerevisiae*)

Rank	Team	Overall score
1	**GENIE3**	**40.279**
2	Team 543	34.023
3	Team 776	31.099
4	Team 862	28.747
5	Team 548	22.711
...	...	...

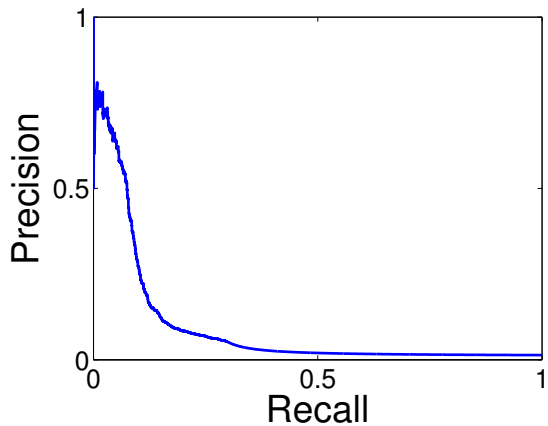
(Huynh-Thu *et al.*, PLoS ONE, 2010)
(Marbach *et al.*, Nature Methods, 2012)

Methods yield more accurate predictions for the artificial network than for the real networks

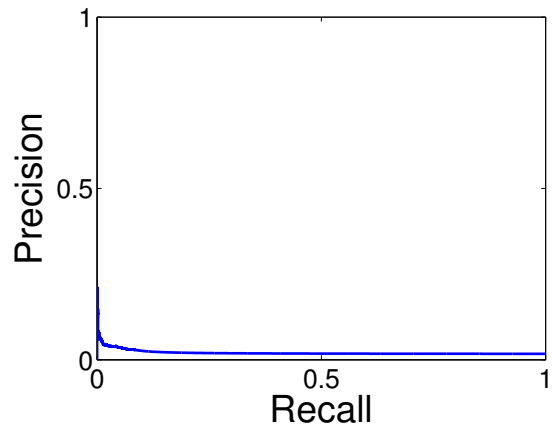
In silico



E. coli



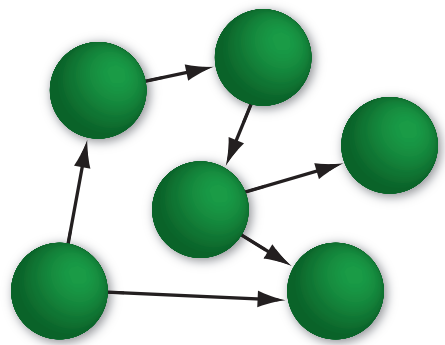
S. Cerevisiae



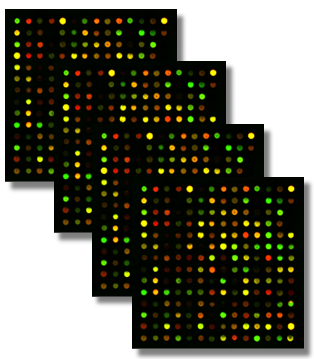
Network inference with GENIE3

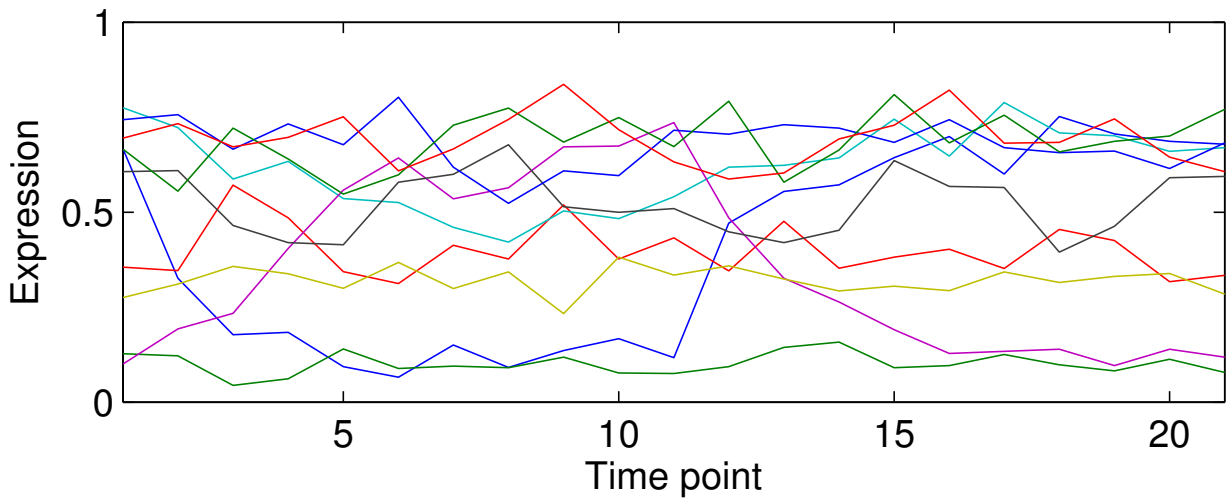
GENIE3 and time series

GENIE3 and systems genetics



Time series of gene expressions





GENIE3 with time series data

GENIE3-time:

Weight of $g_i \rightarrow g_j$ is the importance of expression of g_i at time t for the prediction of expression of g_j at time $t + h$.

Learning sample:

Inputs				Output
$g_1(t_1)$	$g_2(t_1)$	$\cdots$	$g_p(t_1)$	$g_j(t_1 + h)$
$g_1(t_2)$	$g_2(t_2)$	$\cdots$	$g_p(t_2)$	$g_j(t_2 + h)$
$g_1(t_3)$	$g_2(t_3)$	$\cdots$	$g_p(t_3)$	$g_j(t_3 + h)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

GENIE3 with time series + steady-state data

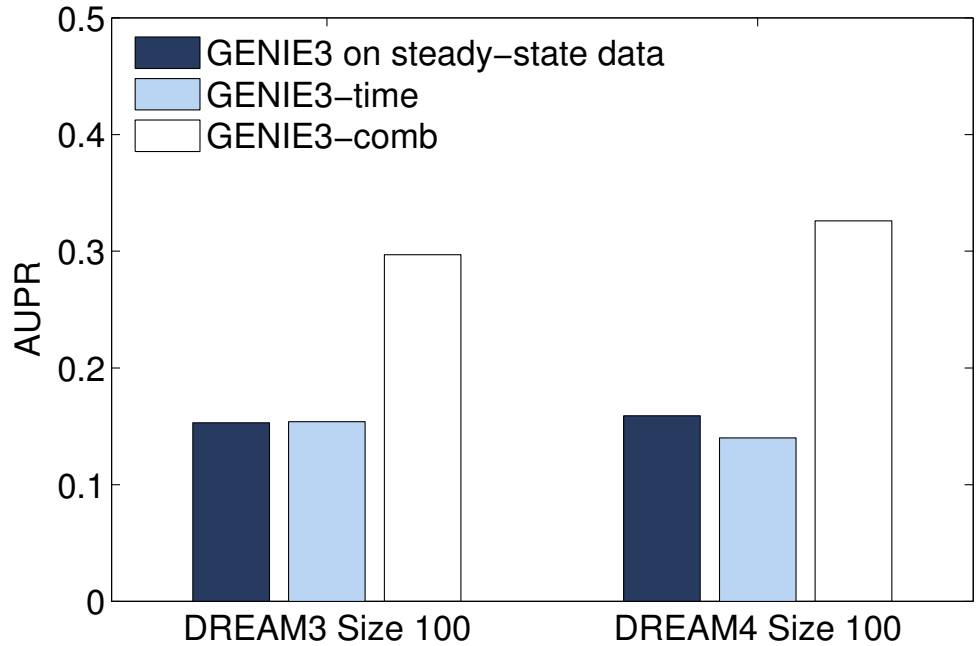
GENIE3-comb:

Learn a **single** model from both datasets by concatenating them.

Learning sample:

Inputs				Output
$g_1(exp_1)$	$g_2(exp_1)$	$\cdots$	$g_p(exp_1)$	$g_j(exp_1)$
$g_1(exp_2)$	$g_2(exp_2)$	$\cdots$	$g_p(exp_2)$	$g_j(exp_2)$
$g_1(exp_3)$	$g_2(exp_3)$	$\cdots$	$g_p(exp_3)$	$g_j(exp_3)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$g_1(t_1)$	$g_2(t_1)$	$\cdots$	$g_p(t_1)$	$g_j(t_1+h)$
$g_1(t_2)$	$g_2(t_2)$	$\cdots$	$g_p(t_2)$	$g_j(t_2+h)$
$g_1(t_3)$	$g_2(t_3)$	$\cdots$	$g_p(t_3)$	$g_j(t_3+h)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Integrating both types of data improves the predictions



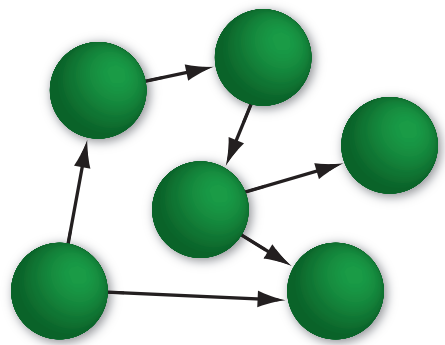
GENIE3-comb would have been ranked:

- 2nd on DREAM3 (5 artificial networks of 100 genes)
- 3rd on DREAM4 (5 artificial networks of 100 genes)

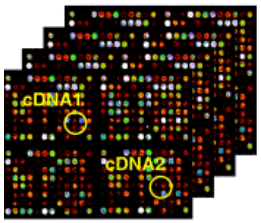
Network inference with GENIE3

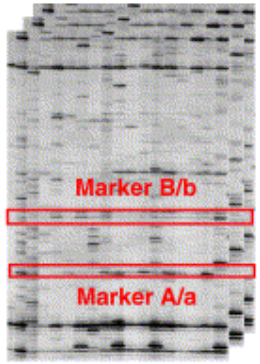
GENIE3 and time series

GENIE3 and systems genetics



How to incorporate genetic data into GENIE3?



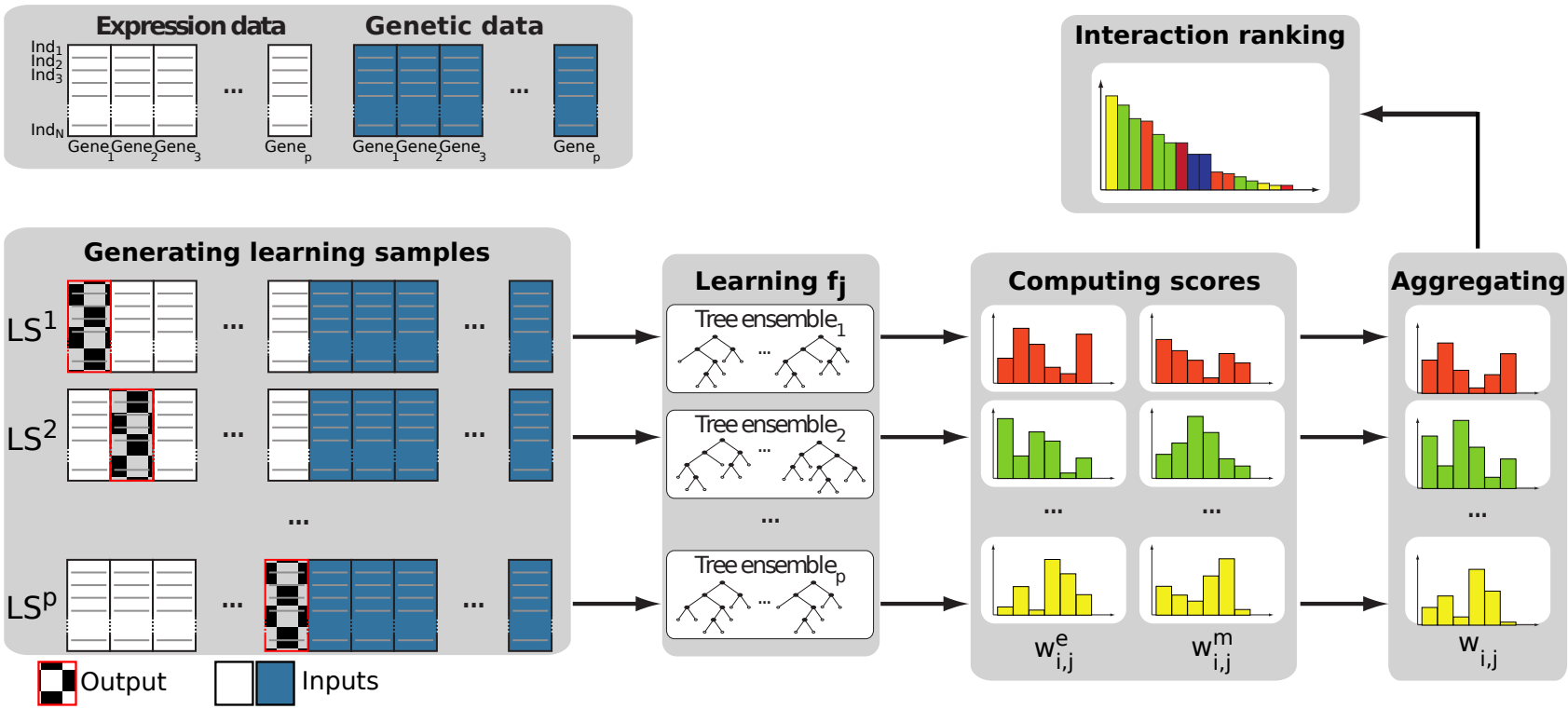


$\Downarrow$

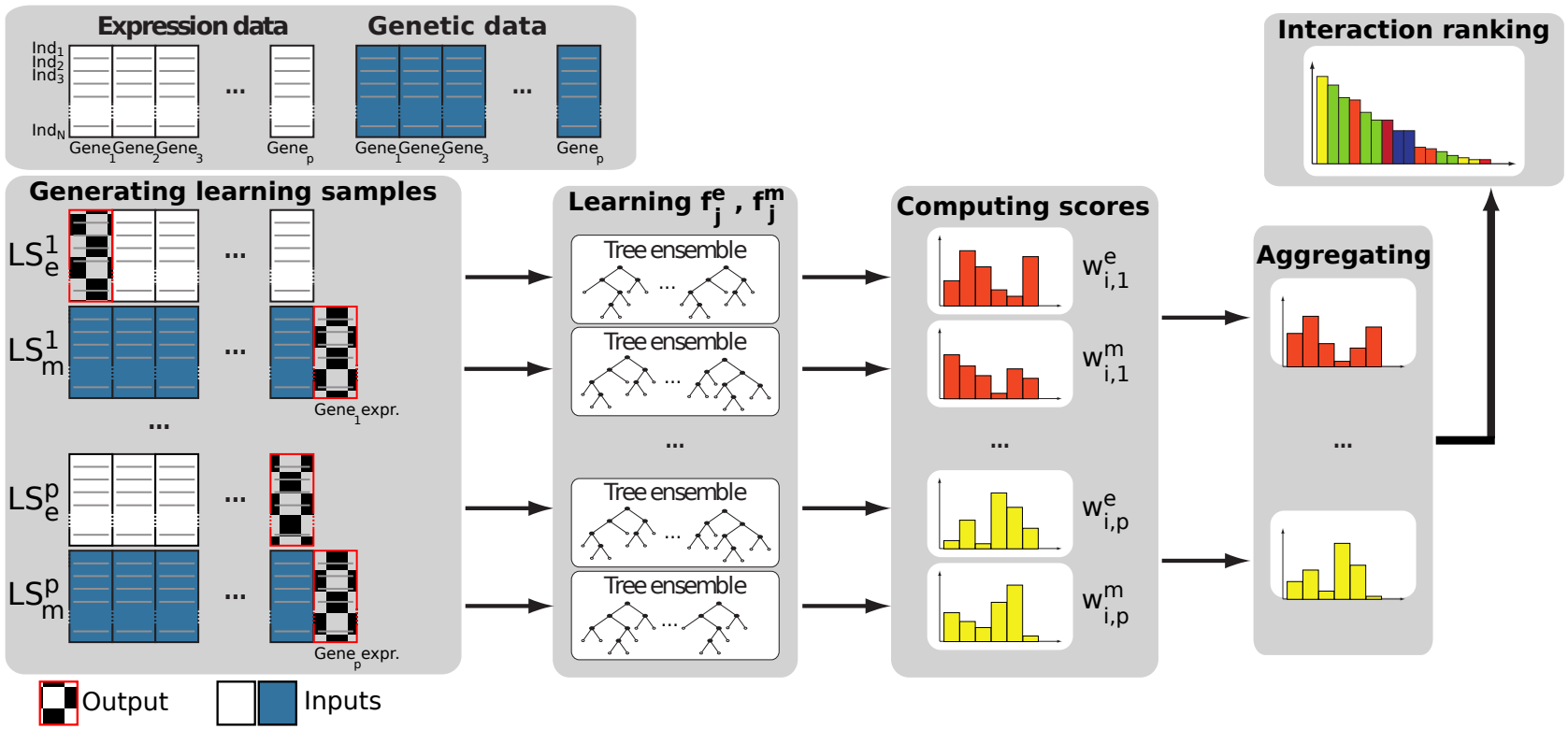
		Target gene			
		gene 1	**gene 2**	$\cdots$	**gene p**
Regulating gene	**gene 1**	-	0.05	$\cdots$	0.56
	gene 2	0.19	-	$\cdots$	0.03
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	gene p	0.11	0.42	$\cdots$	-

$\Rightarrow$ Two extensions of GENIE3

GENIE3-SG-joint: one predictive model is learned for each target gene



GENIE3-SG-sep: two predictive models are learned for each target gene



Weight aggregation

In both GENIE3-SG procedures, we have for each target gene g_j:

$w^e_{i,j}$ = importance of the **expression** of g_i

$w^m_{i,j}$ = importance of the **marker** of g_i

Aggregation 1:

$$w_{i,j} = w^e_{i,j} + w^m_{i,j}$$

$\sim g_i \rightarrow g_j$ if *either* marker or expression of g_i are predictive of expression of g_j

Aggregation 2:

$$w_{i,j} = w^e_{i,j} \times w^m_{i,j}$$

$\sim g_i \rightarrow g_j$ if *both* marker and expression of g_i are predictive of expression of g_j

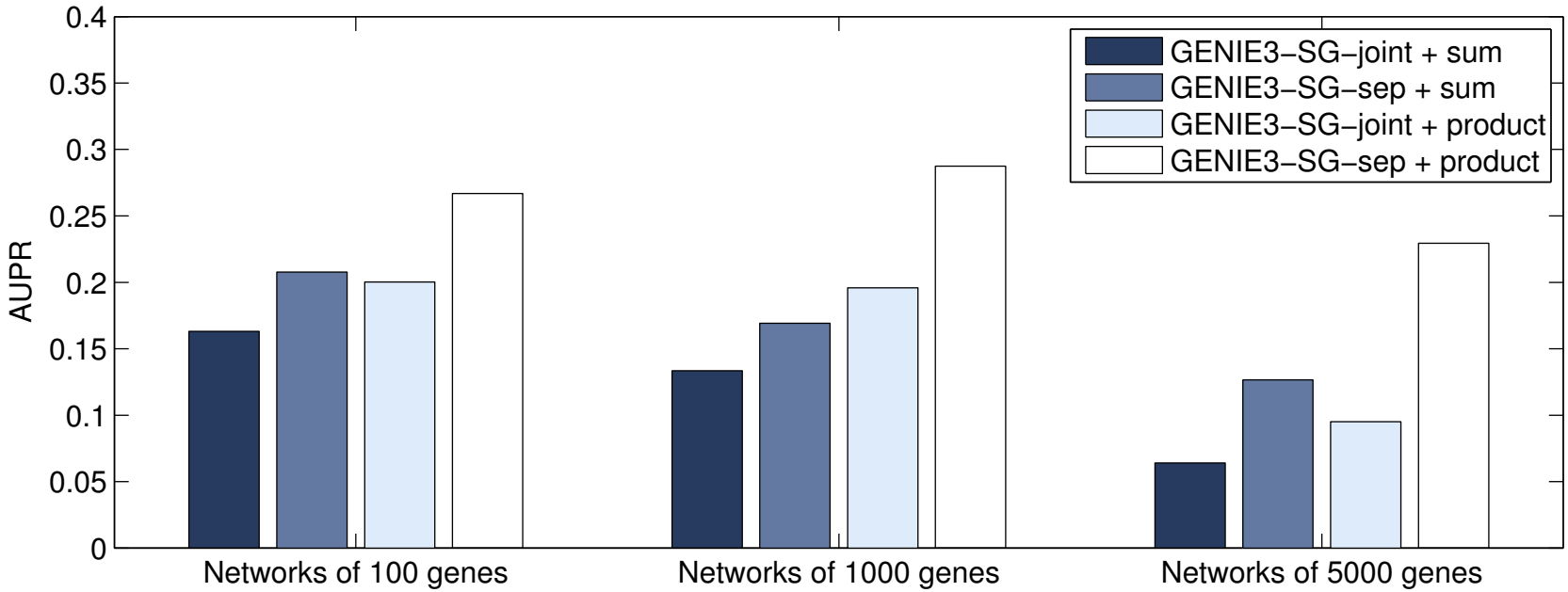
STATSEQ datasets

Networks of 100, 1000, and 5000 genes.

For each network, 8 datasets have been generated:

Configuration	Marker Distance	Heritability	Population Size
1	$\sim \mathcal{N}(5, 1)$	High	300
2	$\sim \mathcal{N}(5, 1)$	High	900
3	$\sim \mathcal{N}(5, 1)$	Low	300
4	$\sim \mathcal{N}(5, 1)$	Low	900
5	$\sim \mathcal{N}(1, 0.1)$	High	300
6	$\sim \mathcal{N}(1, 0.1)$	High	900
7	$\sim \mathcal{N}(1, 0.1)$	Low	300
8	$\sim \mathcal{N}(1, 0.1)$	Low	900

Methods comparison

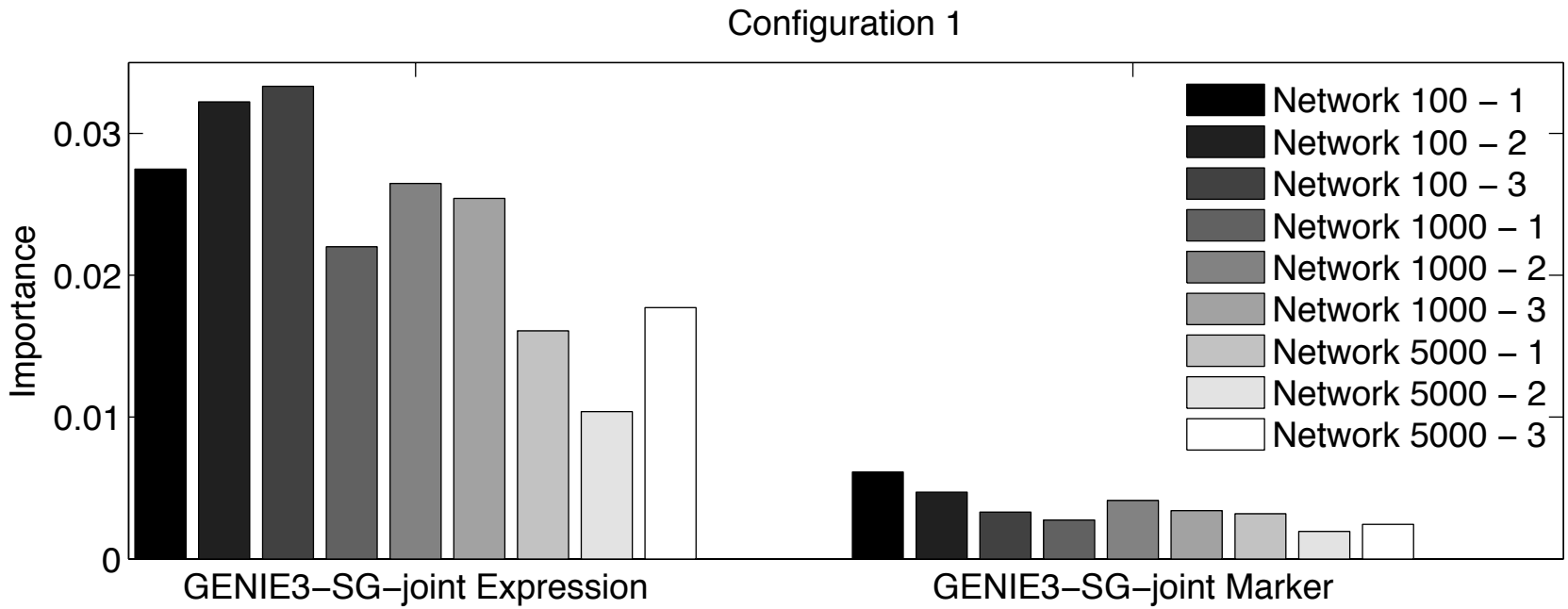


GENIE3-SG-sep performs better than GENIE3-SG-joint

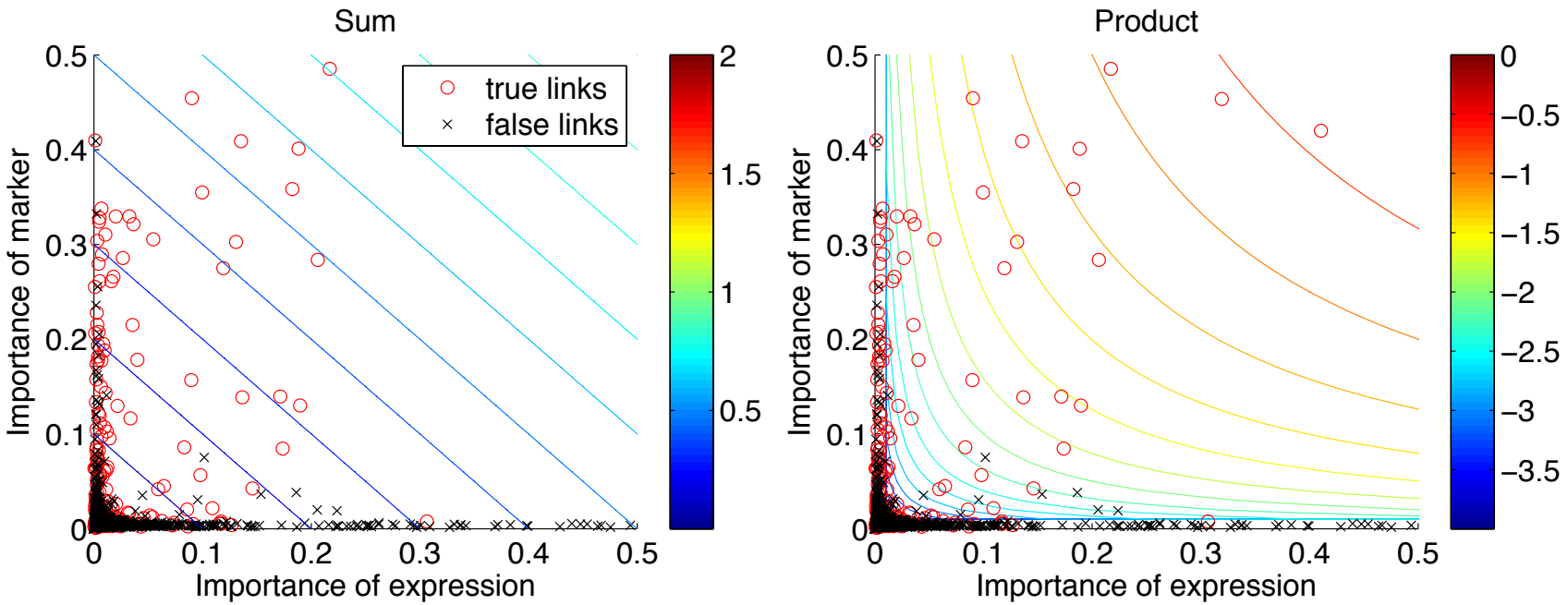
Product performs better than sum

GENIE3-SG-joint: tree-based methods have a positive bias for continuous variables (expression)

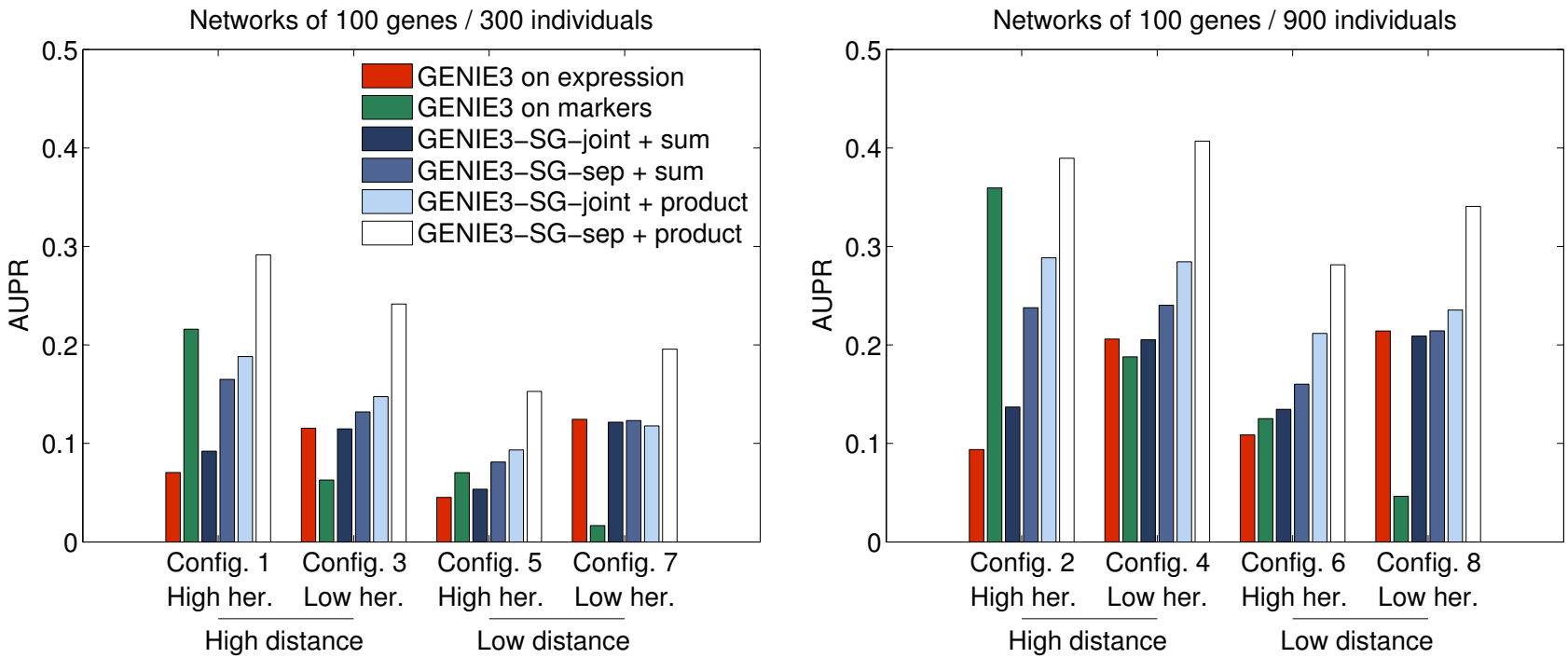
Configuration 1: High heritability and distance between markers.



High values for both $w^e_{i,j}$ and $w^m_{i,j}$ are obtained only for true edges



Performances of expressions and markers vary depending on the configuration



DREAM5 Systems Genetics challenge

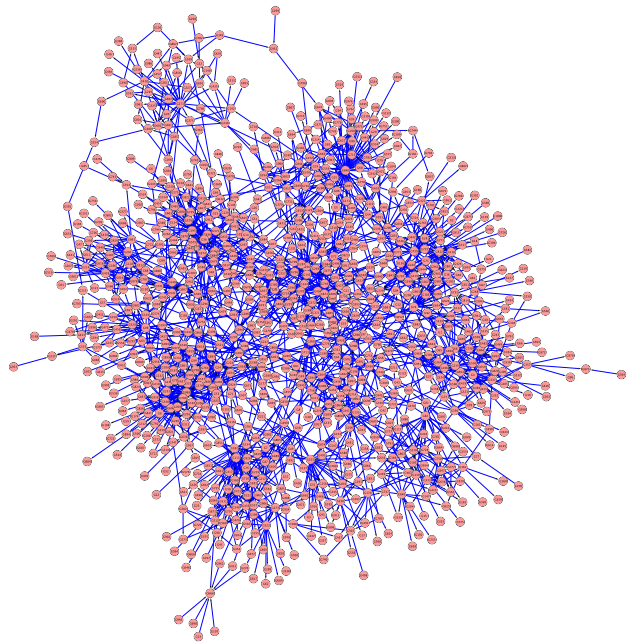
Inference of synthetic regulatory networks

1000 genes in each network

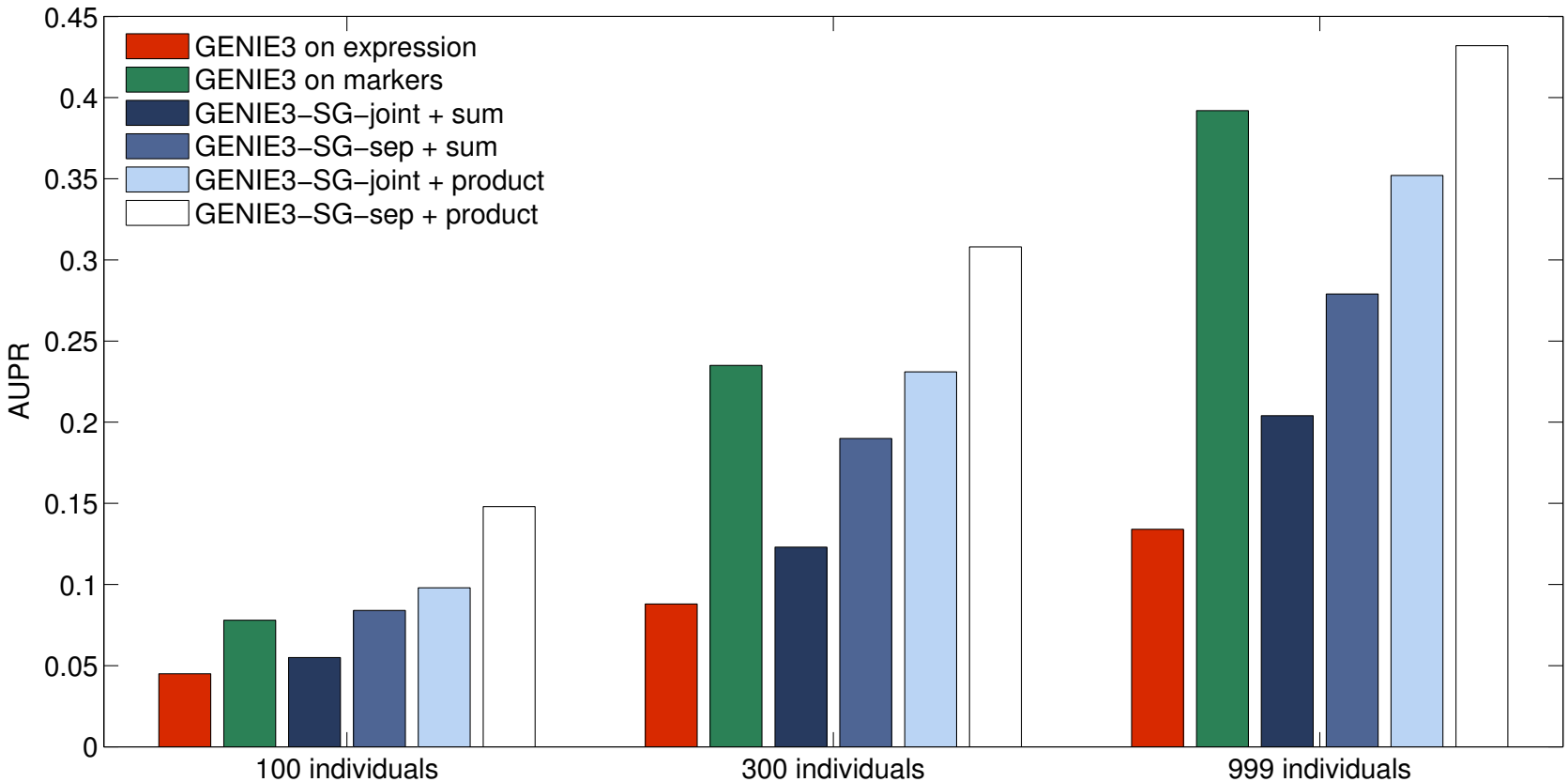
1 marker per gene

Populations of 100, 300, and 999 individuals

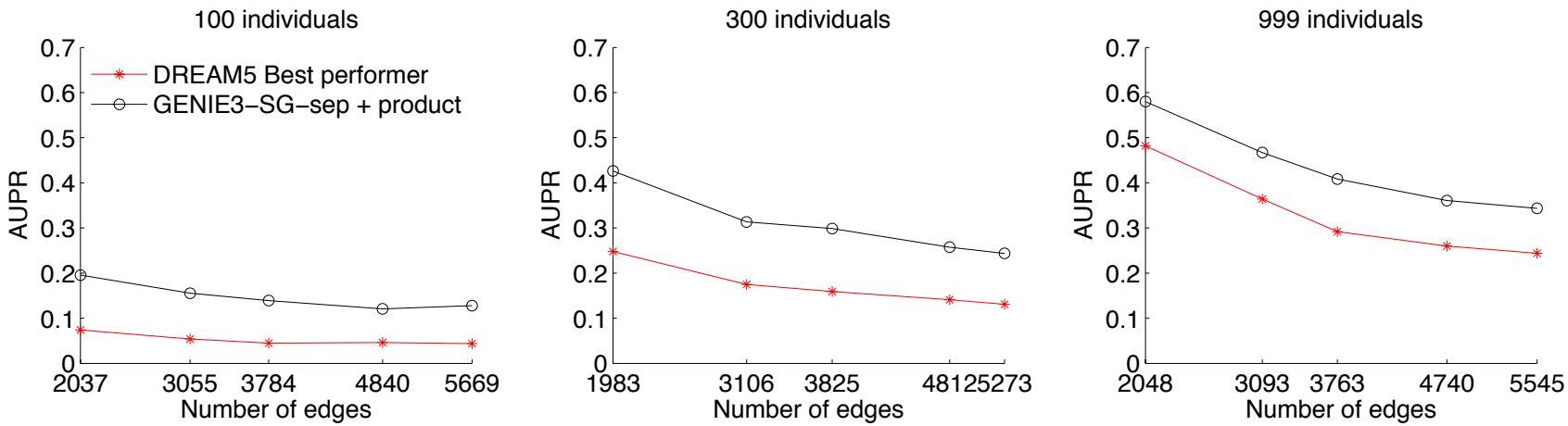
5 networks per population size



GENIE3-SG-sep is better than GENIE3-SG-joint and product is better than sum



GENIE3-SG-sep outperforms the methods of the best performer



Future works

Application to real genetical genomics datasets:
- Yeast: Brem and Kruglyak (2005)
- Drosophila: Aerts and Hassan (2011)
- **Other public datasets for validation?**

GENIE3 assumes one observed genetic marker for each gene.
How to deal with more realistic datasets (more or less than one marker per gene)?

Comparison of ranking techniques other than tree-based methods