Boolean networks as a framework to model human preimplantation development

Mathieu Bolteau

Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

GT Bioss workshop: networks and biological models inference

Monday, July 3rd 2023

Motivations

Need to **better understand preimplantation development**
(especially cell fate transition)

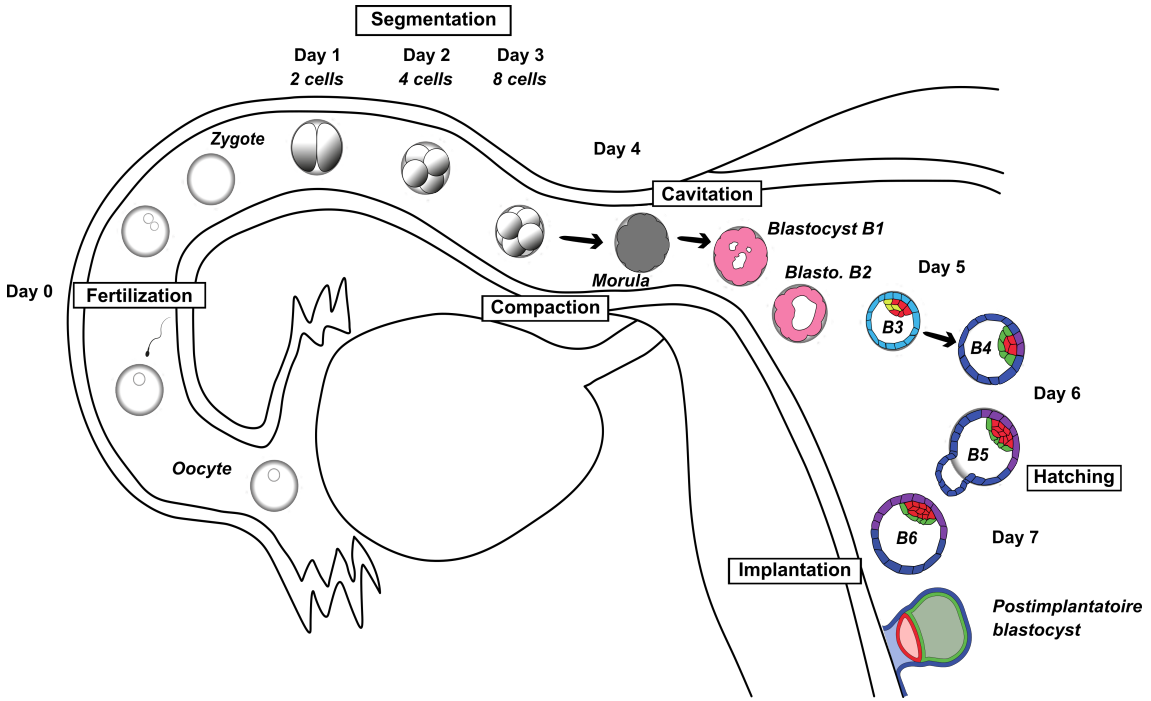
Motivations

Need to **better understand preimplantation development**
(especially cell fate transition)

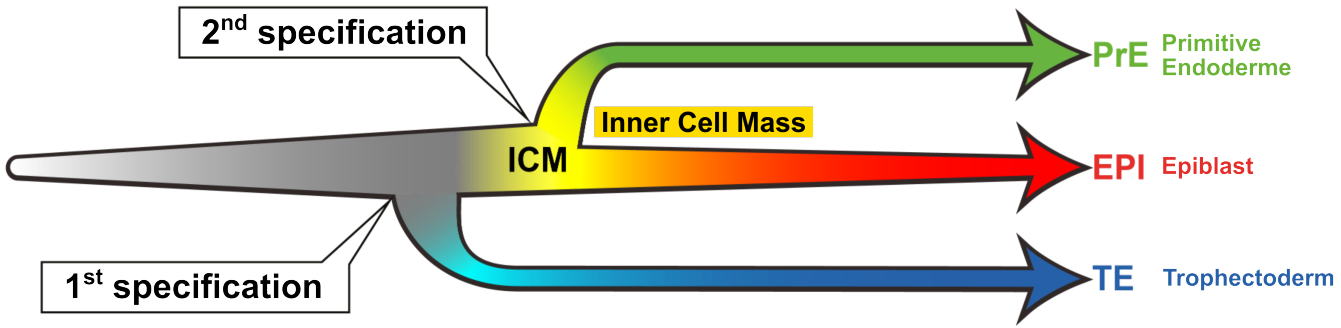
Research on human embryos is **limited** (experiments, law, ethics)

In silico predictive model of the cell fate transition
during the human preimplantation development

Human embryonic development



Adapted from Meistermann, Thesis (2020).



Adapted from Meistermann D.

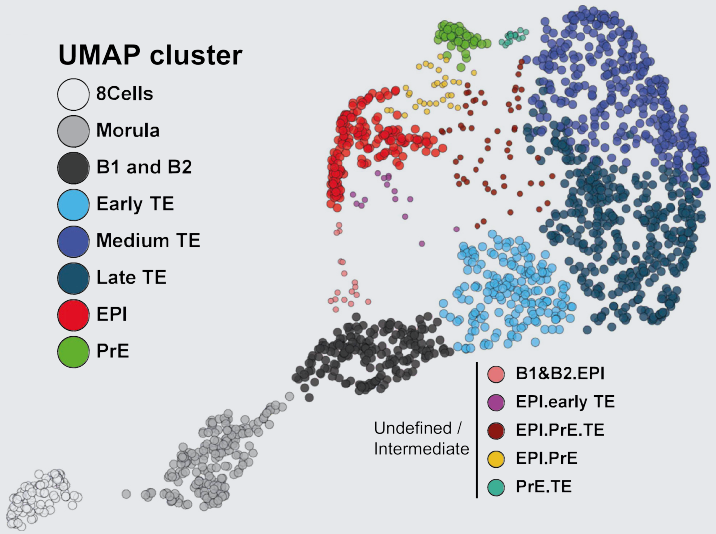
Background [Meistermann, et al. *Cell Stem Cell*, 2021]

scRNAseq data from multiple stage embryos

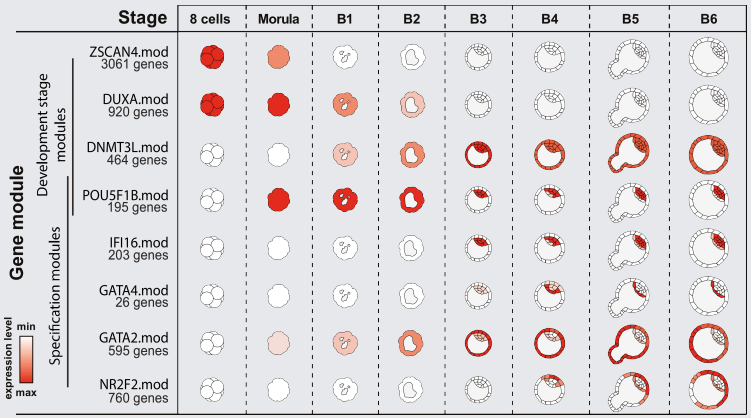
Expression of $\sim 20,000$ genes in $\sim 1,700$ cells from 128 multi-stage embryos

Principal results

- **Clustering** of cells
- Identification of **gene modules** $\rightarrow$ 438 transcription factors (TFs)
- **Pseudotime** evolution of cells at different developmental stages



Meistermann, *et al. Cell Stem Cell* (2021).



Meistermann, *et al. Cell Stem Cell* (2021).

State of the art[1] – modeling of single data

Data analysis

- Statistical, *e.g.* weighted correlation network analysis (WGCNA [Langfelder & Horvath, *BMC Bioinformatics*, 2008])
- Machine learning, *e.g.* reverse graph embedding (pseudotime [Qiu *et al.*, *Nature Methods*, 2017]), uniform manifold approximation and projection (UMAP [McInnes *et al.*, *arXiv preprint*, 2018])

Network inference

- Correlation, *e.g.* gene regulatory network (GRN) inference (SCENIC [Aibar *et al.*, *Nat Methods*, 2017])

Modeling

- Dynamic Boolean models via BoNesis requires average of gene expression and prior knowledge [Chevalier *et al.*, *ICTAI*, 2019]
- Mouse embryo development computational models requires genetic perturbations and knockdowns [Dunn *et al.*, *EMBO journal*, 2019]

[1] Not exhaustive

Goal: Boolean models of embryonic developmental stages

Challenges

- Single cell data specifity: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

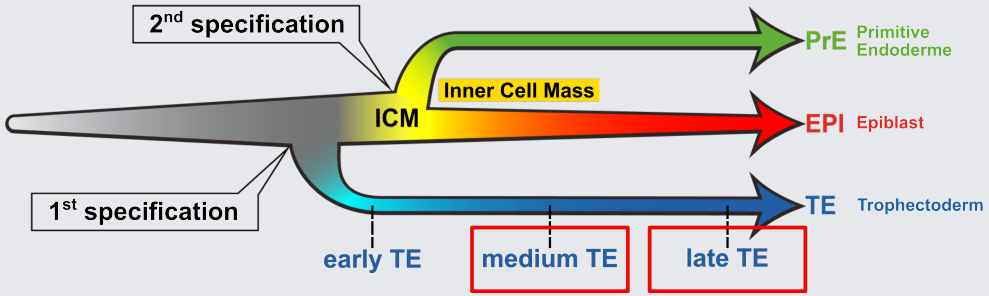
Goal: Boolean models of embryonic developmental stages

Challenges

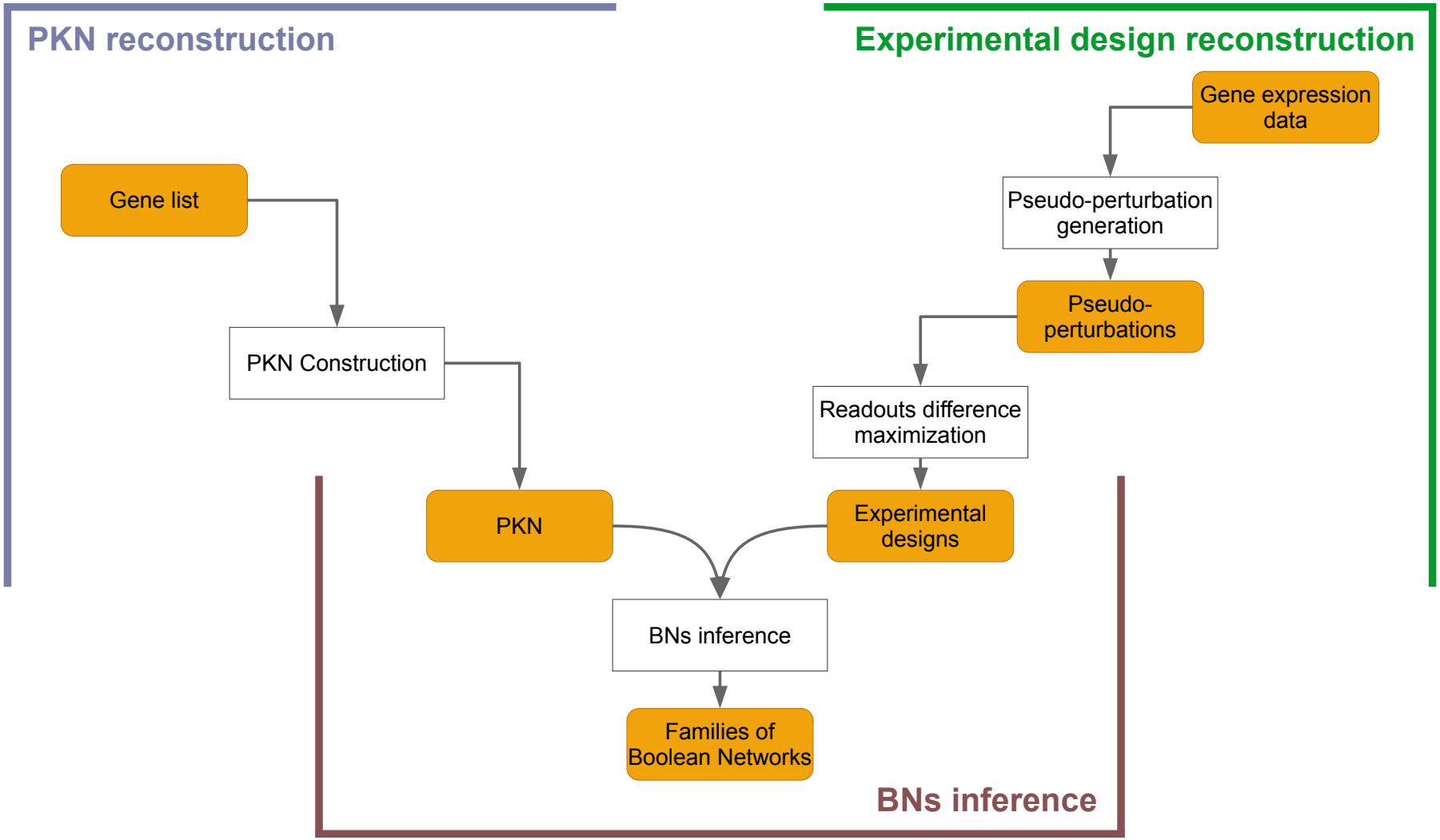
- Single cell data specifity: sparsity and redundancy
- High dimensional data: $\sim 20,000$ genes for $\sim 1,700$ cells
- Unavailable perturbations

Proposed solution

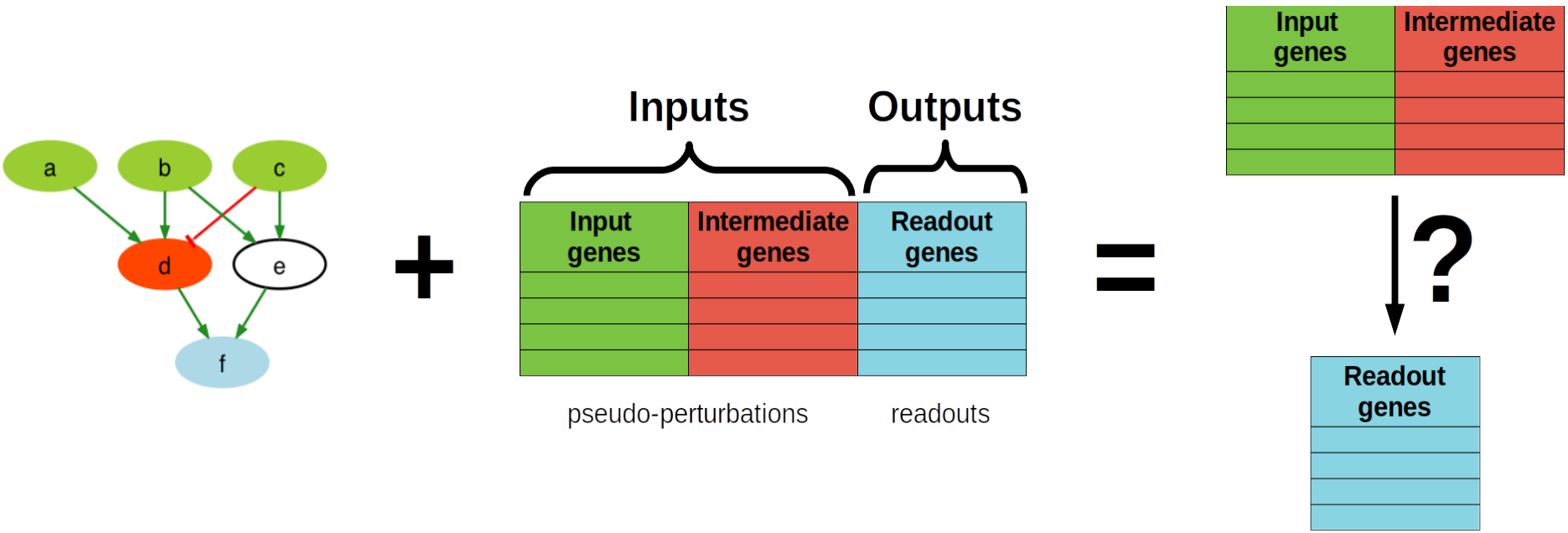
- Distinguish between **two developmental stages**
- Build **specific network models** for each stage
- Identify **regulatory mechanisms** that differentiate both models
- Application on TE maturation: medium (M^{TE}) and late (L^{TE}) TE



Pipeline


PKN reconstruction
Experimental design reconstruction
Gene expression data
Gene list
Pseudo-perturbation generation
Pseudo-perturbations
PKN Construction
Readouts difference maximization
PKN
Experimental designs
BNs inference
Families of Boolean Networks
BNs inference

Learning predictive models



Prior-Knowledge Network

Experimental design

Predictive model

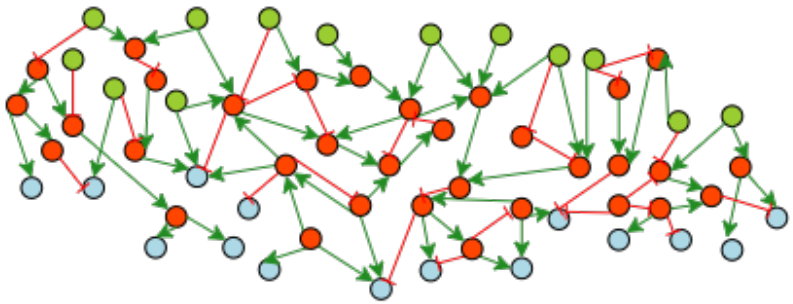
- Signed and directed causal interactions among genes
- Gene expression for a developmental stage

Step 1. PKN reconstruction

Gene list

PKN construction

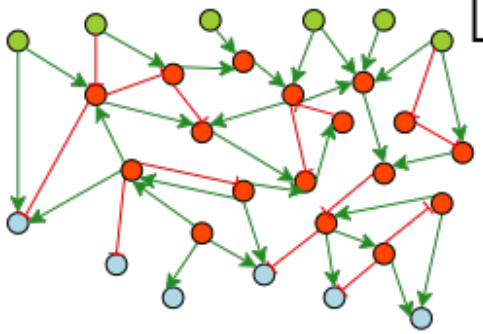
pyBRAvo tool



PKN

PKN reduction

Python



Reduced PKN

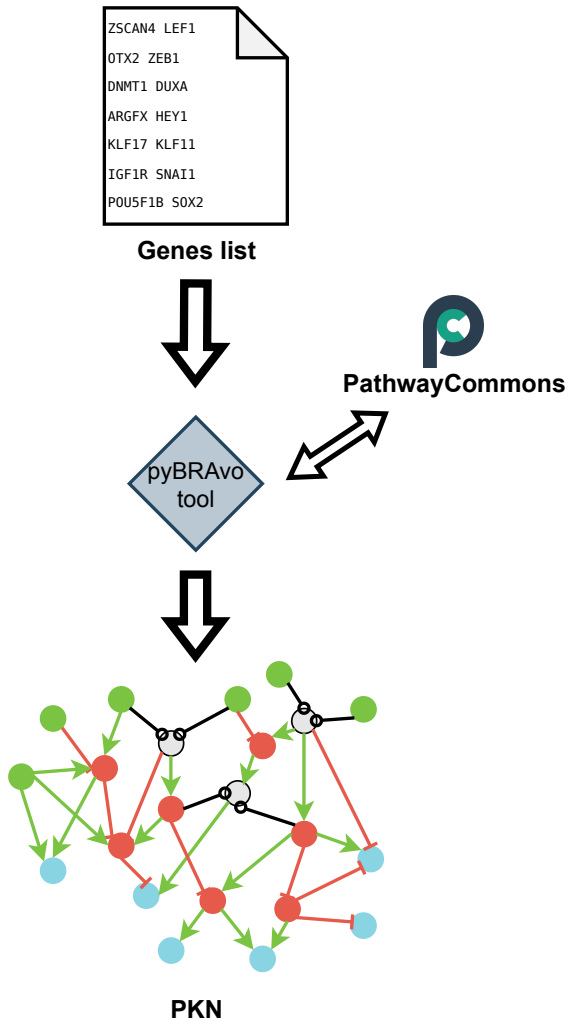
Step 1. PKN reconstruction

Query on PathwayCommons database, via pyBRAvo [Lefebvre et al. *Database*, 2021]

Output PKN

- Labeled (activation/inhibition) and oriented graph
- Nodes: genes (inputs + intermediates + readouts), protein-complexes
- Edges: Transcription regulation

→ inputs & intermediates genes: input for experimental design (Step 2)
→ readouts genes: output for experimental design (Step 2)



Step 2. Experimental design reconstruction

Idea

Extract pseudo-perturbation experiments from scRNAseq data given the PKN structure (Step 1)

Data preprocessing

- Binarization of input + intermediate genes (5 genes). Basic approach: gene is expressed (1) if at least 2 reads are present; else it is absent (0).
- Normalization of readout genes

Step 2.Pseudo-perturbation generation

Inputs + intermediates | Readouts

Cell	A	B	C	D	E	F	G	H	Class
1	1	1	0	1	0	0.8	0.4	0.6	A
2	1	0	1	0	0	0.2	0.5	0.3	A
3	1	1	0	1	1	0.8	0.3	0.9	A
4	1	1	1	0	1	0.6	0.1	0.2	B
5	1	0	0	1	1	0.7	0.8	0.5	B
6	0	0	1	0	1	0.7	0.2	0.3	B

Logic program (k=3)

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	A
2	1	1	0	0.2	0.5	0.3	A

Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B
4	1	1	0	0.6	0.1	0.2	B

- 3 selected genes: A, C, D ($k = 3$)
- Matching cells (1,5), (2,4) ← pseudo-perturbations
- Optimal number of matching cells: 2

Step 2. Pseudo-perturbations generation algorithm

Main rules (x4)

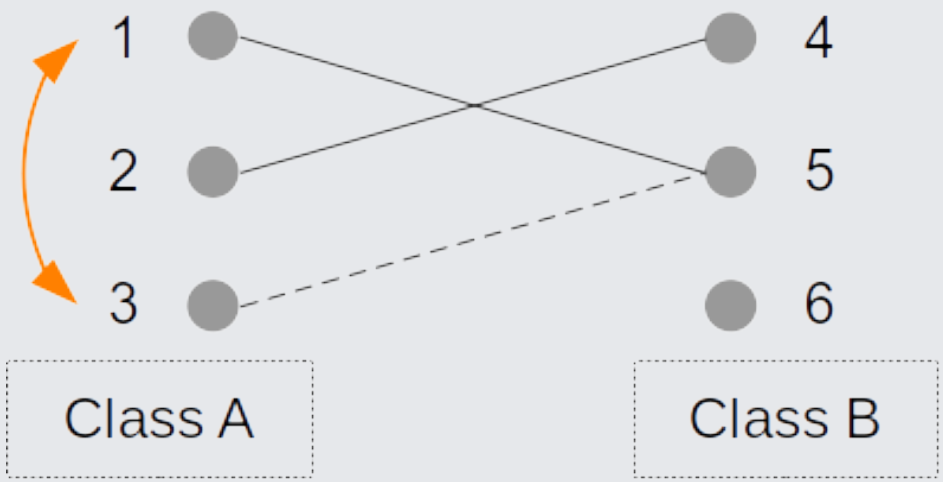
- k-genes: Select k genes among all possible combinations of input + intermediate genes.
- Reachability: input $\rightarrow$ intermediate.
- Matching cells: Select pairs of cells (c_1, c_2), $c_1 \in A$, $c_2 \in B$, for which the (binarized) expression matches for each of the *k-genes*.
- Filter redundancy: The set of k (binarized) expressions should differ for all *matching cells* of the same class.

Optimizations (x2)

- Maximize the number of *matching cells* of either A or B class.
- Maximize the difference of expression of the readout genes of the *matching cells* across all redundant pairs.

Step 2. Maximizing the readouts difference

Redundancy



N. of matching cells : 2 (max)

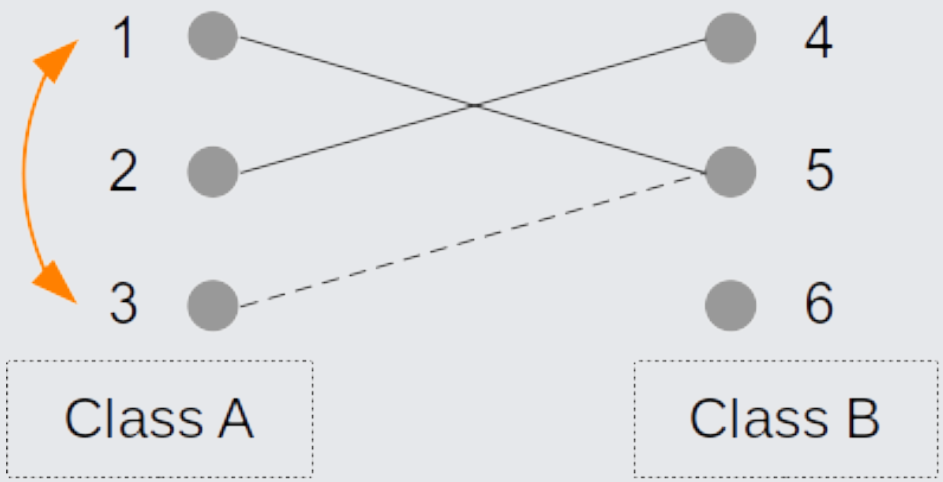
Solution 1 : (1,5), (2,4)

Solution 2 : (3,5), (2,4)

Redundancy measure :
Class A : 3 out of 3
Class B : 2 out of 3

Step 2. Maximizing the readouts difference

Redundancy



N. of matching cells : 2 (max)

Solution 1 : (1,5), (2,4)

Solution 2 : (3,5), (2,4)

Redundancy measure :
Class A : 3 out of 3
Class B : 2 out of 3

Readout difference maximization

diff(1,5) = |0.8-0.7| + |0.4-0.8| + |0.6-0.5| = 0.6

Cell	A	C	D	F	G	H	Class
1	1	0	1	0.8	0.4	0.6	A

Cell	A	C	D	F	G	H	Class
5	1	0	1	0.7	0.8	0.5	B

diff(3,5) = |0.8-0.7| + |0.3-0.8| + |0.9-0.5| = 0.9

Cell	A	C	D	F	G	H	Class
3	1	0	1	0.8	0.3	0.9	A

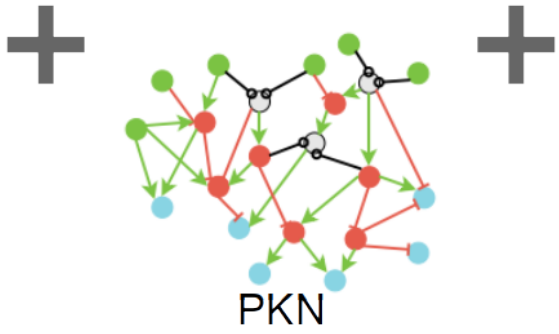
Step 3. BNs inference using Caspo [Guziolowski et al. *Bioinformatics*, 2013.]

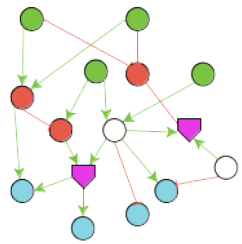
Cell	A	C	D	F	G	H	Class
3	1	0	1	0.8	0.3	0.9	A
2	1	1	0	0.2	0.5	0.3	A

Experimental design for Class A

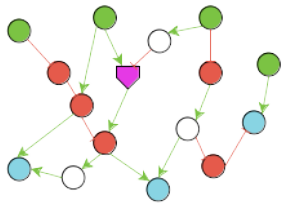
Cell	A	C	D	F	G	H	Class
4	1	1	0	0.6	0.1	0.2	B
5	1	0	1	0.7	0.8	0.5	B

Experimental design for Class B





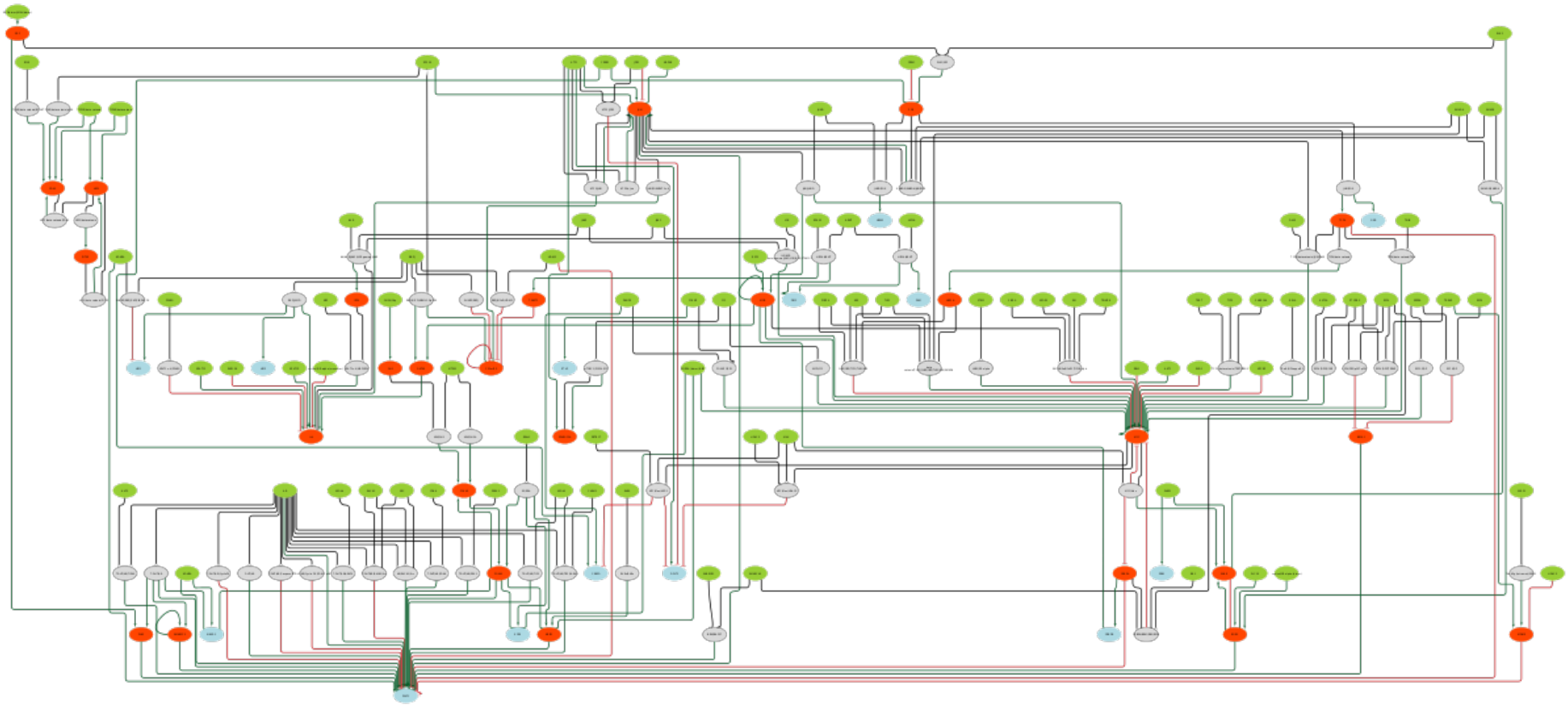
Boolean Networks for Class A



Boolean Networks for Class B

Reconstructed PKN

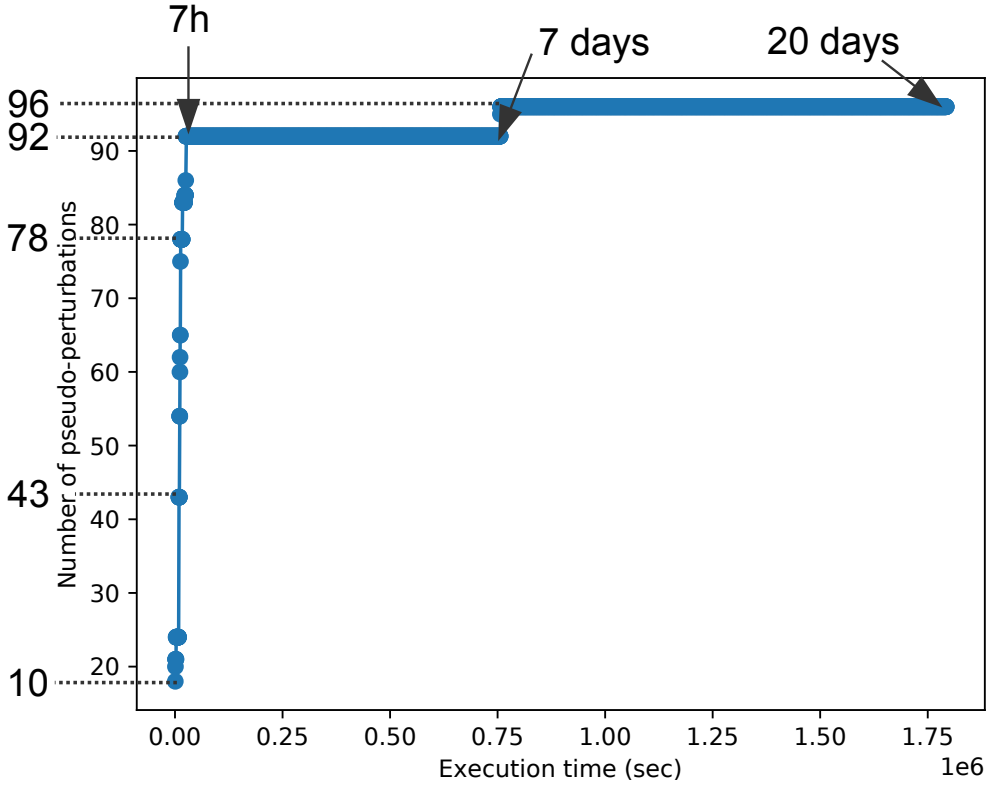
- 233 nodes : inputs (85), intermediates (36), readouts (19)
- 369 edges



Pseudo-perturbations search

Inputs

- $\#M^{TE}$ *cells* $= 348$
- $\#L^{TE}$ *cells* $= 332$
- $k = 10$: 10 genes selected from 121 input and intermediate genes
- Search space: $\binom{121}{10} = 1.27x10^{14}$ possible choices

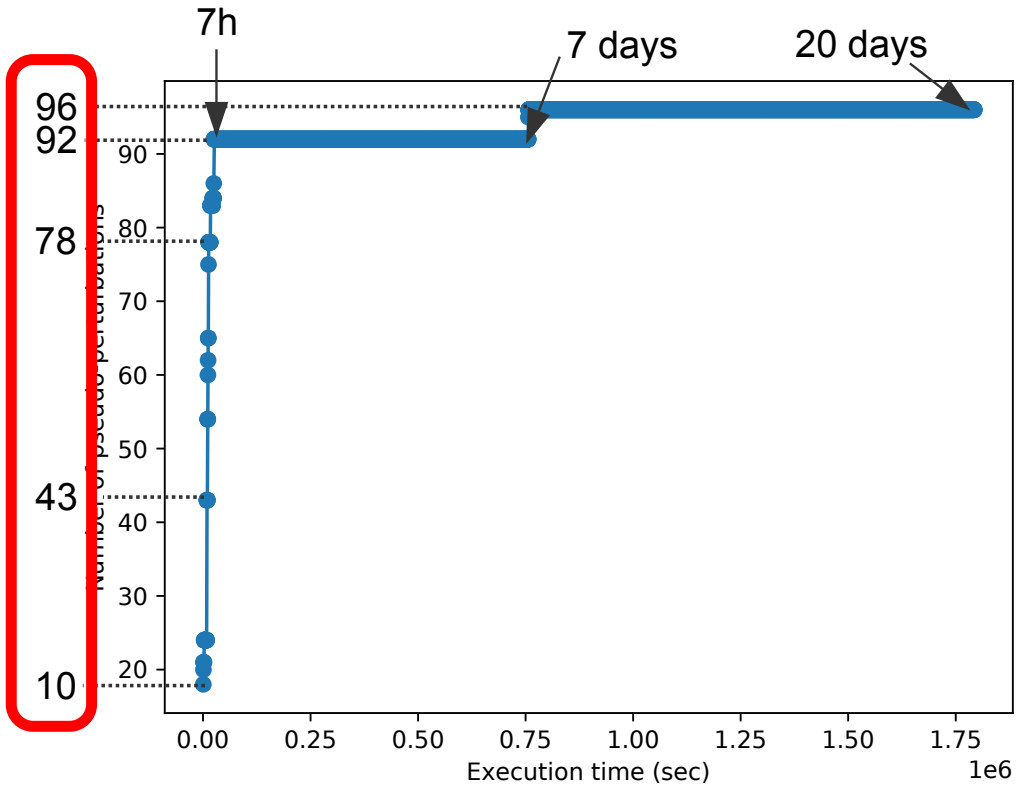


Convergence of the number of pseudo-pertubations over time.

Pseudo-perturbations search

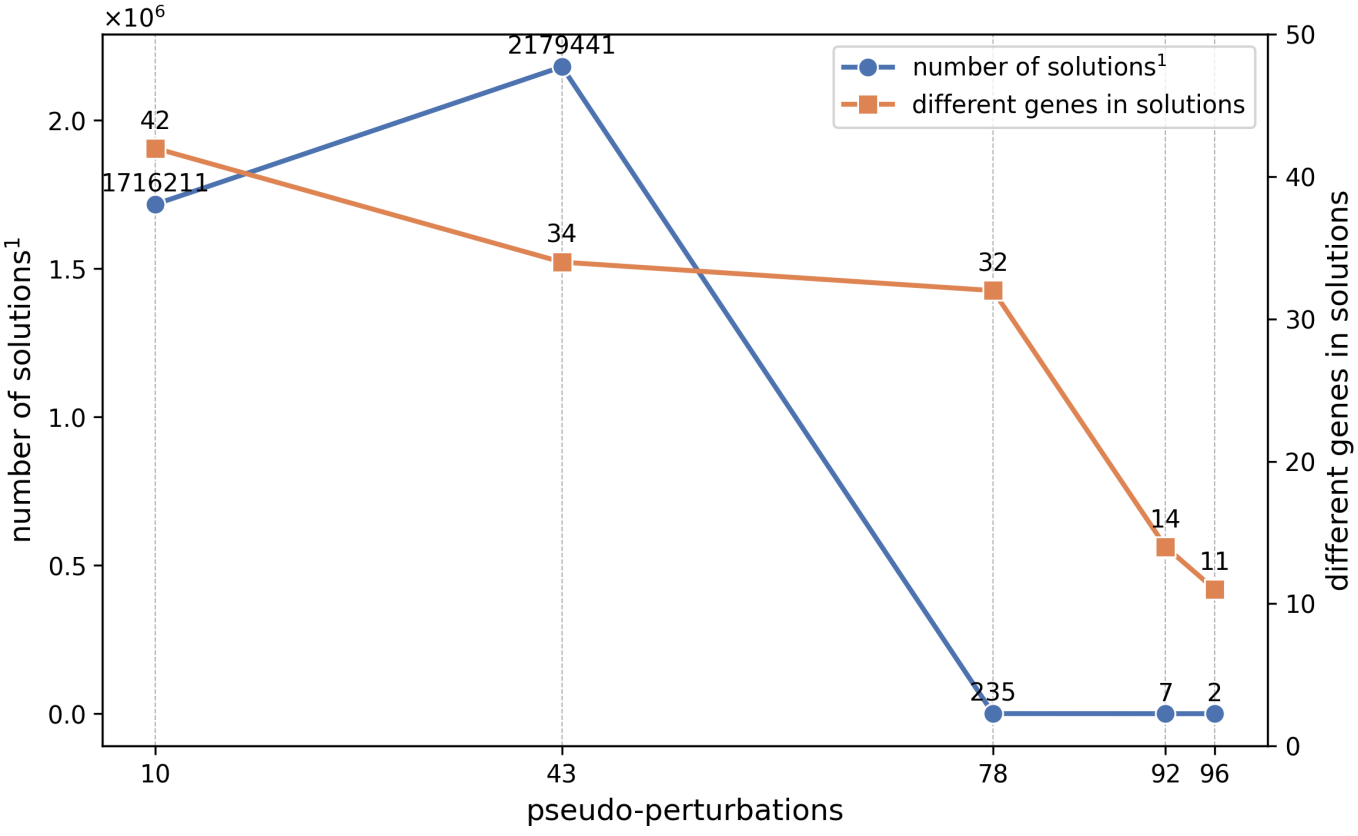
Inputs

- $\#M^{TE}$ *cells* $= 348$
- $\#L^{TE}$ *cells* $= 332$
- $k = 10$: 10 genes selected from 121 input and intermediate genes
- Search space: $\binom{121}{10} = 1.27x10^{14}$ possible choices



Convergence of the number of pseudo-pertubations over time.

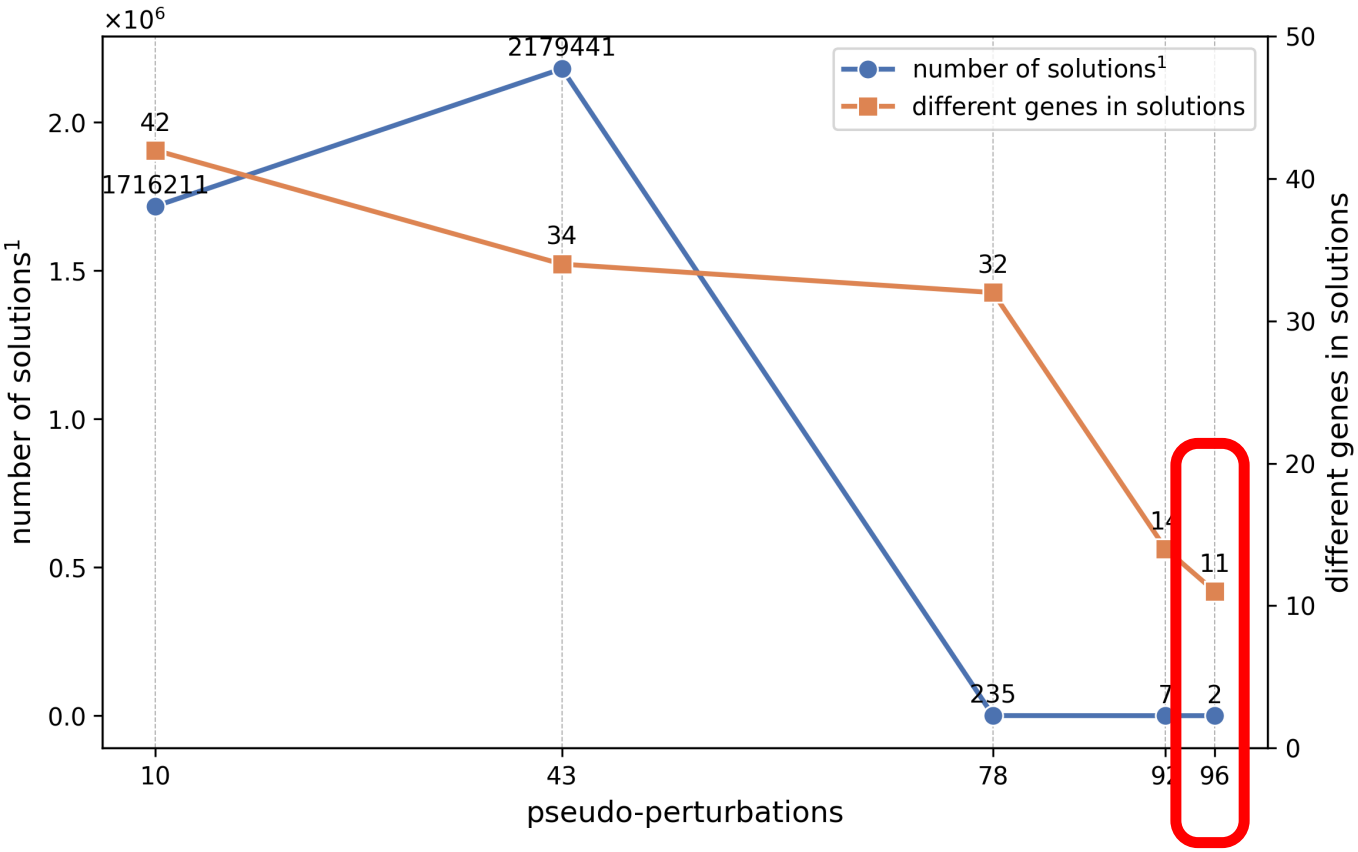
Robustness of solutions



- The more pseudo-perturbations we have, the fewer different genes we have in the solutions
- Gene number explosion when few pseudo-perturbations

[1] 7 days of run on a computer cluster comprising 160 CPUs and 1.5 To of RAM

Robustness of solutions

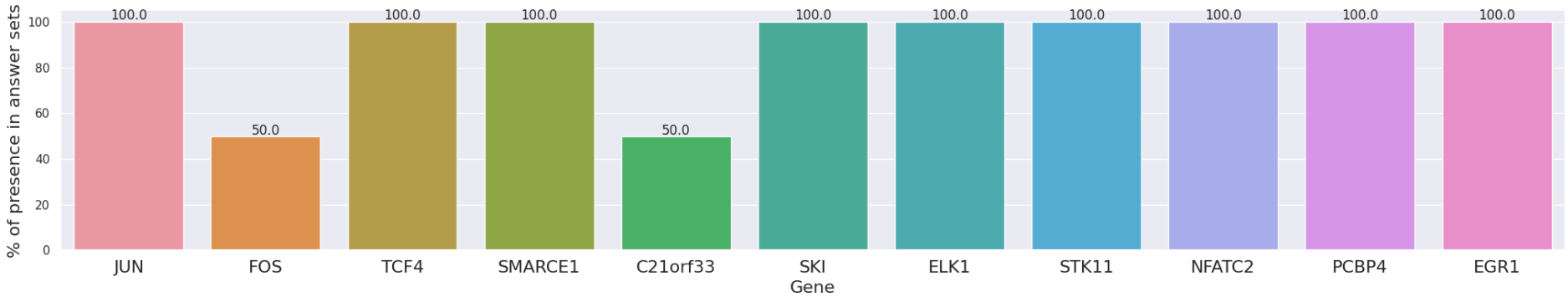


- The more pseudo-perturbations we have, the fewer different genes we have in the solutions
- Gene number explosion when few pseudo-perturbations

[1] 7 days of run on a computer cluster comprising 160 CPUs and 1.5 To of RAM

96 pseudo-perturbations sub-optimal solution

- Number of solution $= 2$
- Different genes in solutions $= 11$



11 characteristic genes
to have the same Boolean behavior in M^{TE} and L^{TE}

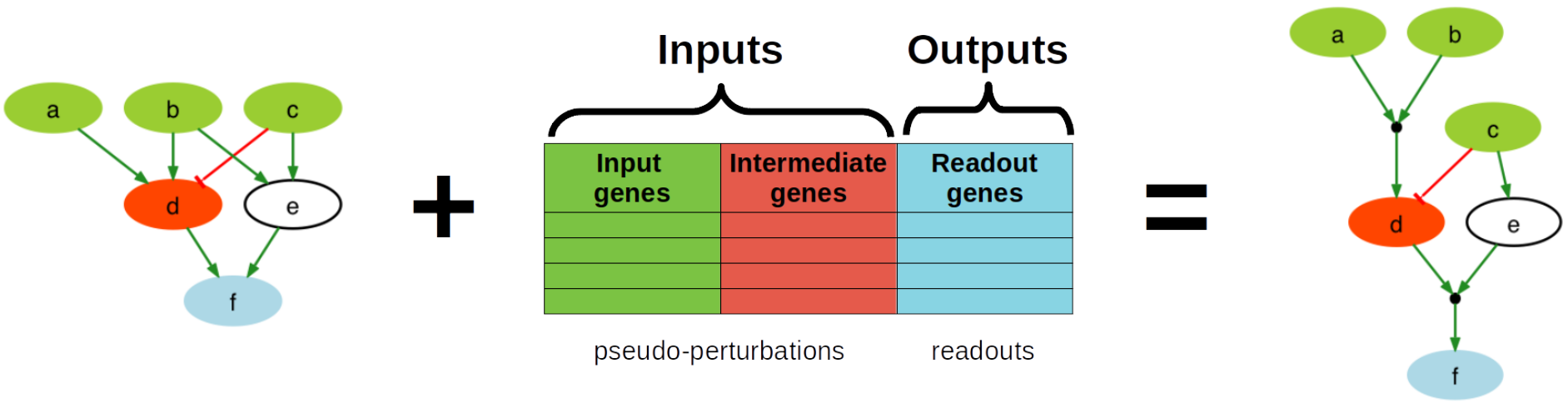
Cell representativity (redundancies)

Solution	M^{TE} (%)	L^{TE} (%)	Total (%)
1	**266 (76%)**	**246 (74%)**	**512 (75%)**
2	235 (68%)	248 (75%)	483 (71%)

- $\#M^{TE} cells = 348$
- $\#L^{TE} cells = 332$
- $\#Total\ cells = 680$

On average, 73% of representativity for the total number of cell at each stage.

Learning Boolean logic models



Prior-Knowledge Network

Experimental design

Optimal Boolean logic model

Meaning of "Optimal"

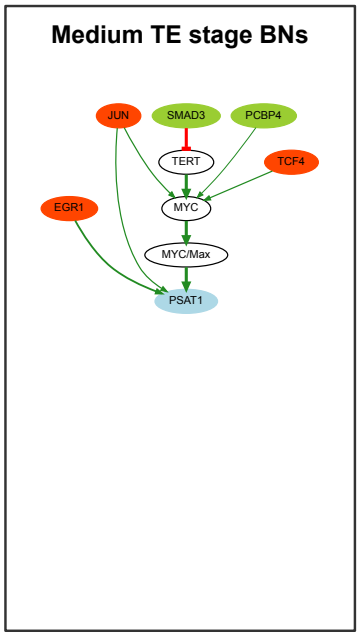
- **Biological Property**: consistency with experimental data
- **Parsimony Principle**: the minimal/simplest explanation

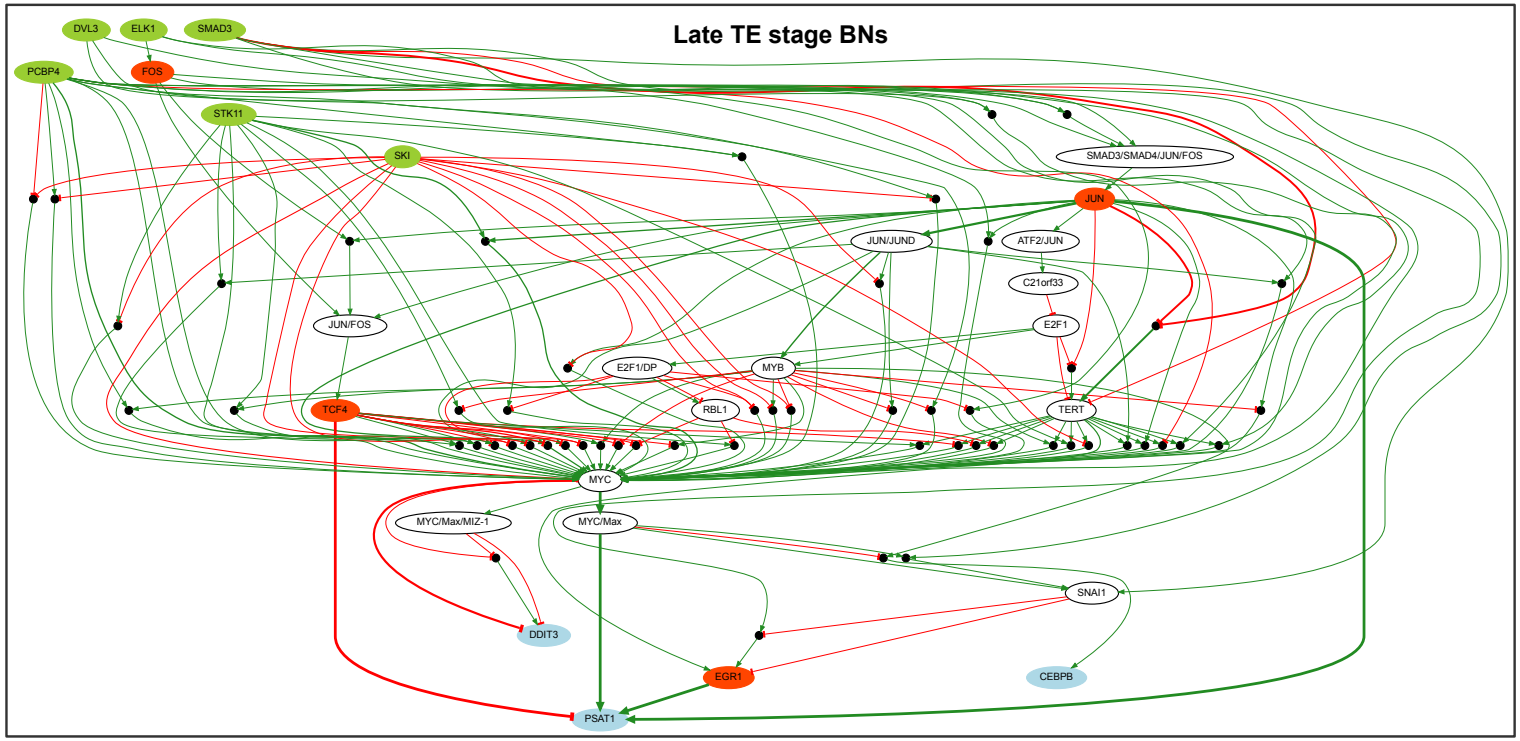
Caspo metrics for inferred BNs

Solution	MSE		Size		#Networks	
	M^{TE}	L^{TE}	M^{TE}	L^{TE}	M^{TE}	L^{TE}
1	0.1159	0.1410	6	20	5	17797
2	0.1180	0.1400	1	5	1	1

- Lower MSE for $M^{TE} \rightarrow M^{TE}$ model is simpler
- More redundancies for L^{TE} (number of BNs) $\rightarrow$ different ways to explain the "input-output" relation with Boolean gates

Inferred BNs for solution 1





- More readouts implicated in L^{TE} stage
- Greater BNs variability for $L^{TE} \rightarrow$ Gain of function
- L^{TE} seems more unstable $\rightarrow$ transition from L^{TE} to another stage

Conclusion

Pseudo-perturbation generation

- Efficient algorithm to select cells and genes to generate pseudo-perturbations $\rightarrow$ 96 pseudo-perturbations in 7 days
- Robustness of the generated solutions $\rightarrow$ from $+2$ millions of solutions to only 2
- Discovery of 11 genes whose on/off values remain identical for 96 cells across 2 classes
- Expression of 11 genes across 96 cells are representative of the cell populations (*e.g.* 72% in M^{TE} and 73% in L^{TE})
- Our method deals with single cell data and its specificities (redundancy and sparsity)

General method

- Proposed a method that learns Boolean networks of 2 stages using scRNAseq data and Prior Knowledge
- Mechanisms of TF-gene regulations distinguishing 2 developmental stages
- Complementarity with the state of the art
  - Boolean models without using perturbations
  - Method taking into account the diverse states of cell population

Perspectives

- Comparing different Prior knowledge (in progress)
- Study the impact of different discretization methods
- Evaluate the difference between the found BNs
- Apply the method on other developmental stages (different cell fate)

Aknowledgements

- Jérémie Bourdon @LS2N, Nantes University
- Carito Guziolowski @LS2N, Centrale Nantes
- Laurent David @CR2TI, CHU de Nantes, Nantes University
- ANR AIBY4 & ANR BOOSTIVF