# Boolean Network Models of Human Preimplantation Development

Mathieu Bolteau

Computational Biology (COMBI) group

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

✉ mathieu.bolteau@ls2n.fr – 🌐 www.mathieubolteau.xyz

DEVSTEM Meeting

Tuesday, March 12th 2024




# Motivations

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

# Motivations

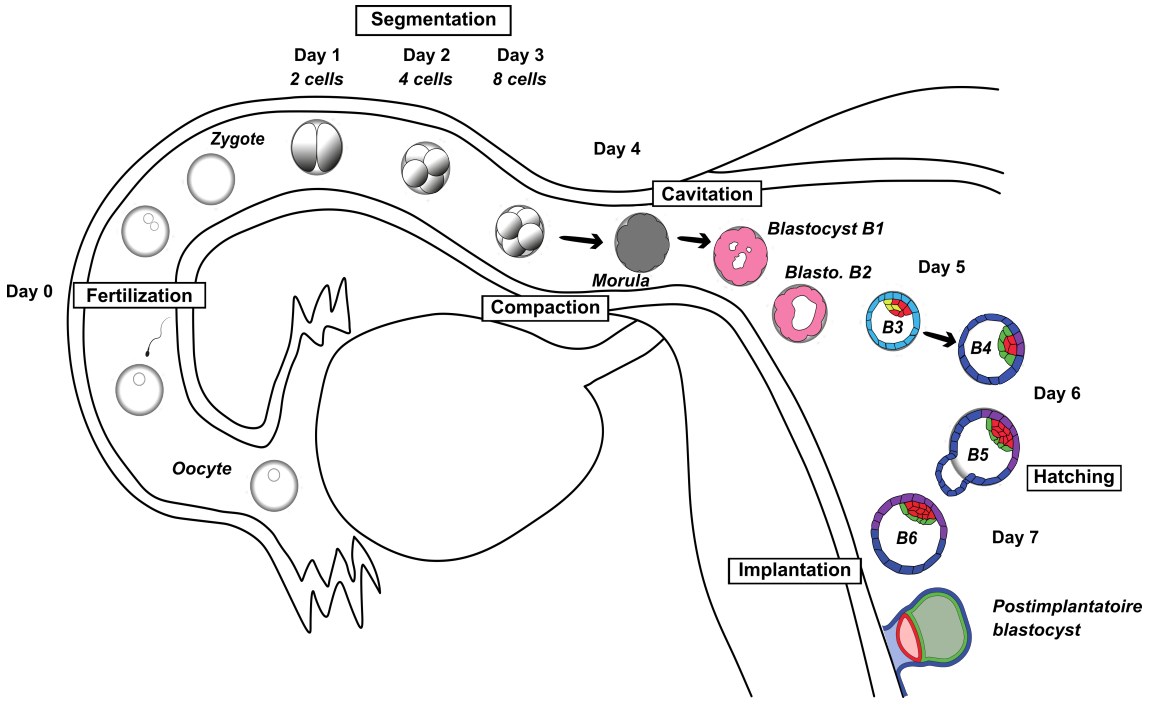
Need to **better understand preimplantation development**
(especially cell fate transition)

---

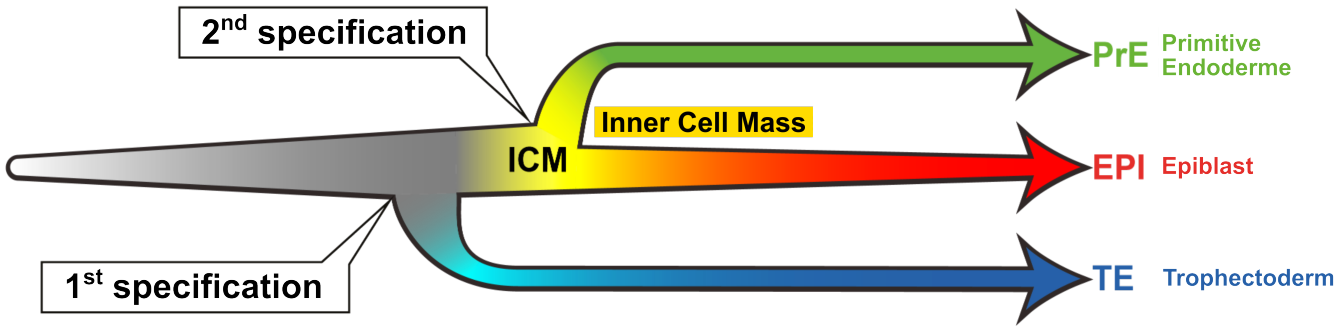
Research on human embryos is **limited** (experiments, law, ethics)

**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

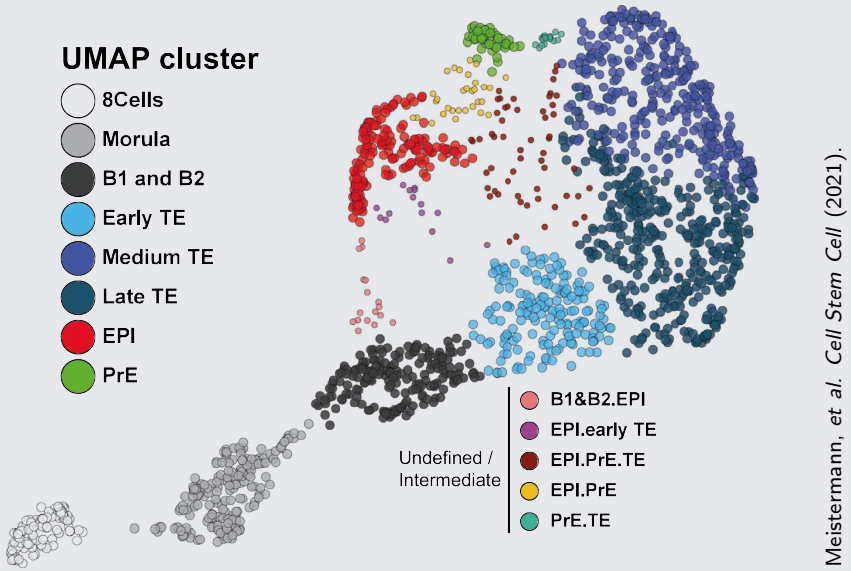
[Meistermann, *et al. Cell Stem Cell*, 2021]

## 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



# Concepts

## Prior-Knowledge Network (PKN)

Signed and oriented graph, where nodes correspond to biological entities (e.g., genes, proteins), and edges represent causal or functional relationships between these entities . [Radulescu *et al.*, *JRSI*, 2006]

# Concepts

## Prior-Knowledge Network (PKN)

Signed and oriented graph, where nodes correspond to biological entities (e.g., genes, proteins), and edges represent causal or functional relationships between these entities . [Radulescu *et al.*, *JRSI*, 2006]

## Boolean Network (BN)

A Boolean network $B$, of dimension $n$ is defined as $B = (N, F)$ where: $N = \{v_1, \ldots, v_n\}$ is a finite set of nodes (variables or genes) and $F = \{f_1, \ldots, f_n\}$ is a set of Boolean functions $f_i : \mathbb{B}^n \rightarrow \mathbb{B}$, with $\mathbb{B} = \{0, 1\}$, describing the evolution of variable $v_i$. [Kauffman, *JTB*, 1969]

# Concepts

## Prior-Knowledge Network (PKN)

Signed and oriented graph, where nodes correspond to biological entities (e.g., genes, proteins), and edges represent causal or functional relationships between these entities . [Radulescu *et al.*, *JRSI*, 2006]

## Boolean Network (BN)

A Boolean network $B$, of dimension $n$ is defined as $B = (N, F)$ where: $N = \{v_1, \ldots, v_n\}$ is a finite set of nodes (variables or genes) and $F = \{f_1, \ldots, f_n\}$ is a set of Boolean functions $f_i : \mathbb{B}^n \rightarrow \mathbb{B}$, with $\mathbb{B} = \{0, 1\}$, describing the evolution of variable $v_i$. [Kauffman, *JTB*, 1969]

## Pseudo-perturbation

Boolean vector that encodes the expression status of a set of $k$ genes within a particular cell. A pseudo-perturbation of a cell in a class *matches* with a second pseudo-perturbation from another class. [Bolteau *et al.*, *JCB*, submitted]

# Concepts

## Answer Set Programming (ASP)

A declarative programming paradigm used to solve difficult (primarily NP-hard) search problems. It is based on rules and constraints permitting to restrict the problem from an initial solution set. [Baral, *Cambridge University Press*, 2003]

# Concepts

## Answer Set Programming (ASP)

A declarative programming paradigm used to solve difficult (primarily NP-hard) search problems. It is based on rules and constraints permitting to restrict the problem from an initial solution set. [Baral, *Cambridge University Press*, 2003]

```
<head> :- <body>.
```

# Concepts

## Answer Set Programming (ASP)

A declarative programming paradigm used to solve difficult (primarily NP-hard) search problems. It is based on rules and constraints permitting to restrict the problem from an initial solution set. [Baral, *Cambridge University Press*, 2003]

<head> :- <body>.

An example: select only "medium TE" and "late TE" cells

```
1  % Facts representing the knowledge
2  cell(c1). cell(c2). cell(c3). % 3 cells
3  class(early_TE). class(medium_TE). class(late_TE). % 3 classes
4  be_part(c1,early_TE). be_part(c2,medium_TE). be_part(c3,late_TE).
5  % Generate all possible set of selected cells
6  {sel_cell(C,CL) : cell(C), be_part(C,CL)}.
7  % Forbid answers with early TE class sel_cell()
8  :-sel_cell(_,early_TE).
9  % Print selected cells
10 #show sel_cell/4.
```

# State of the art – 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]

# 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

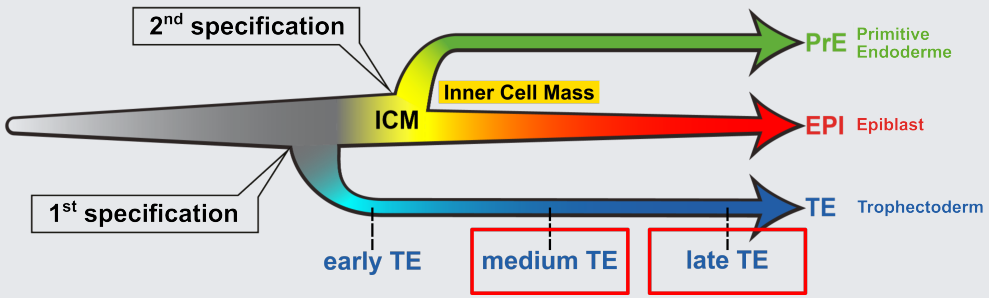
## Challenges

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

## Proposed solution

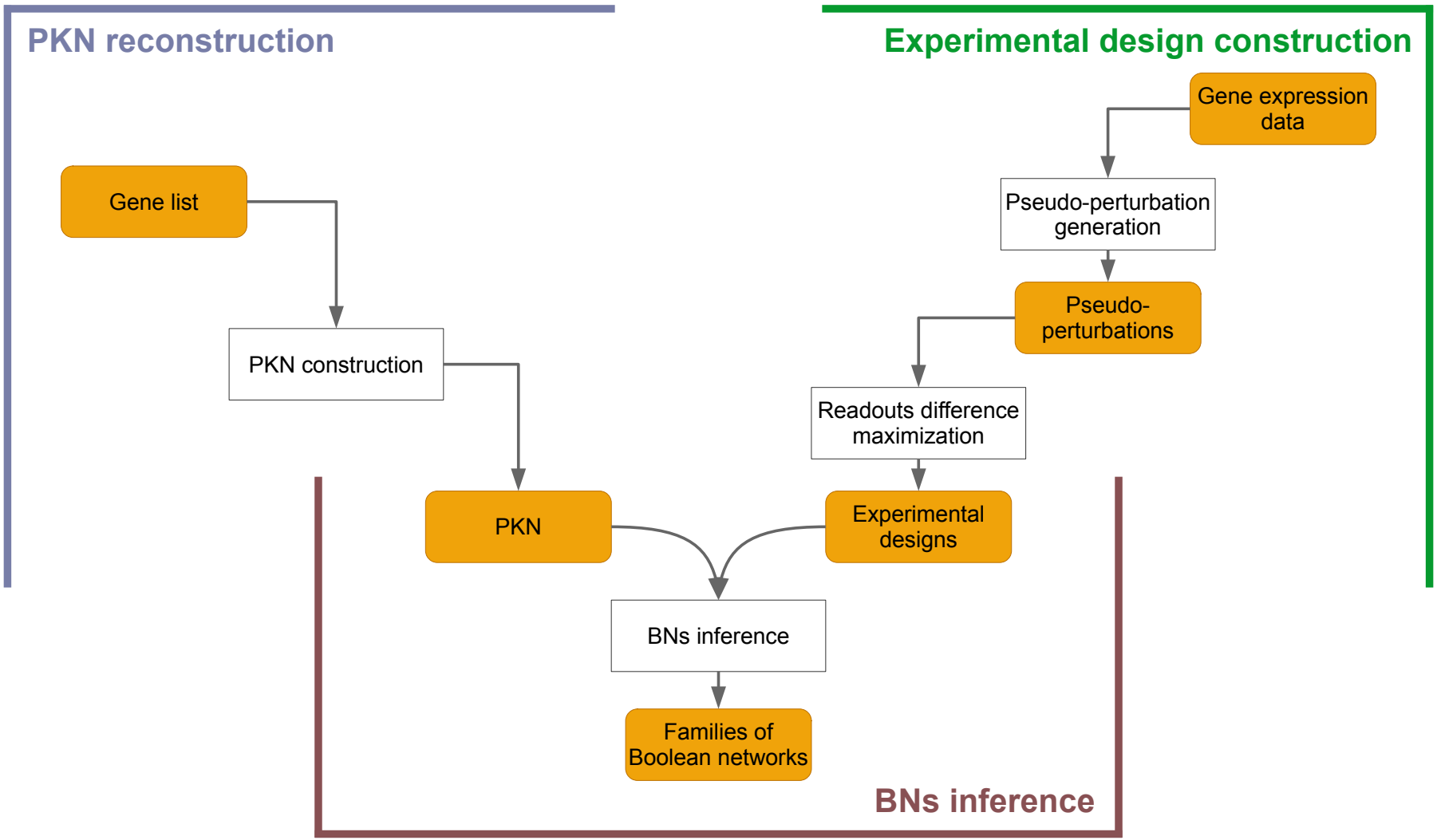
[Bolteau *et al.* *ISBRA'23*, 2023.]

- 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

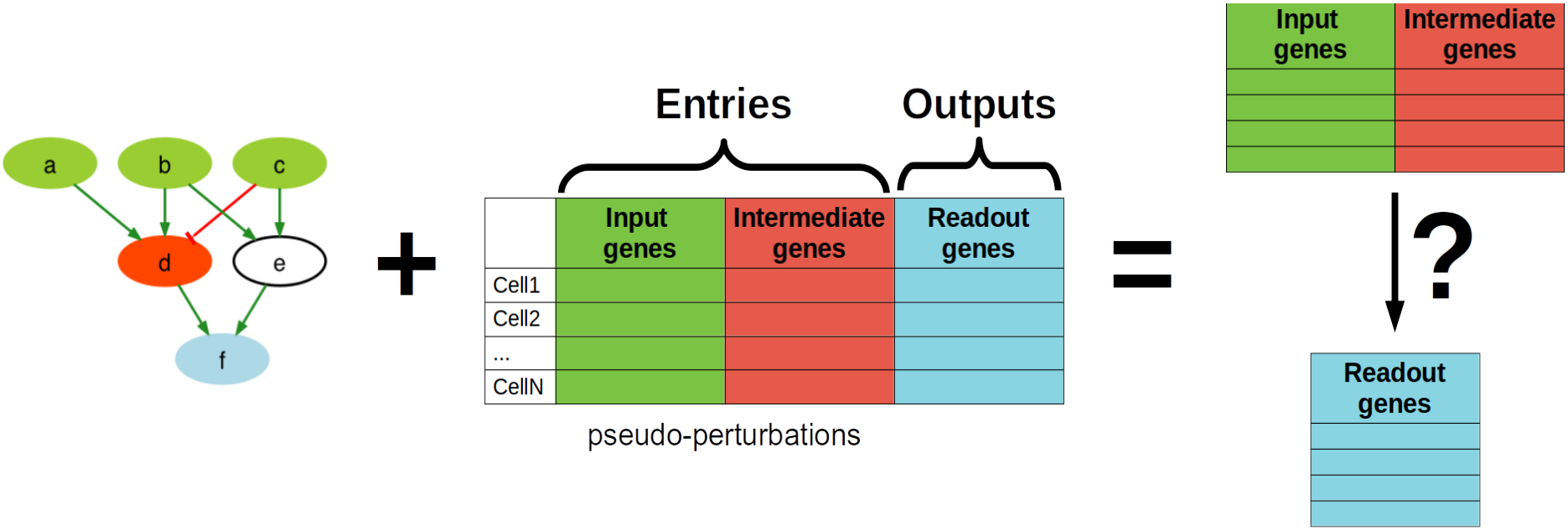


# Pipeline

[Bolteau et al. *ISBRA'23*, 2023.]



# Learning predictive models



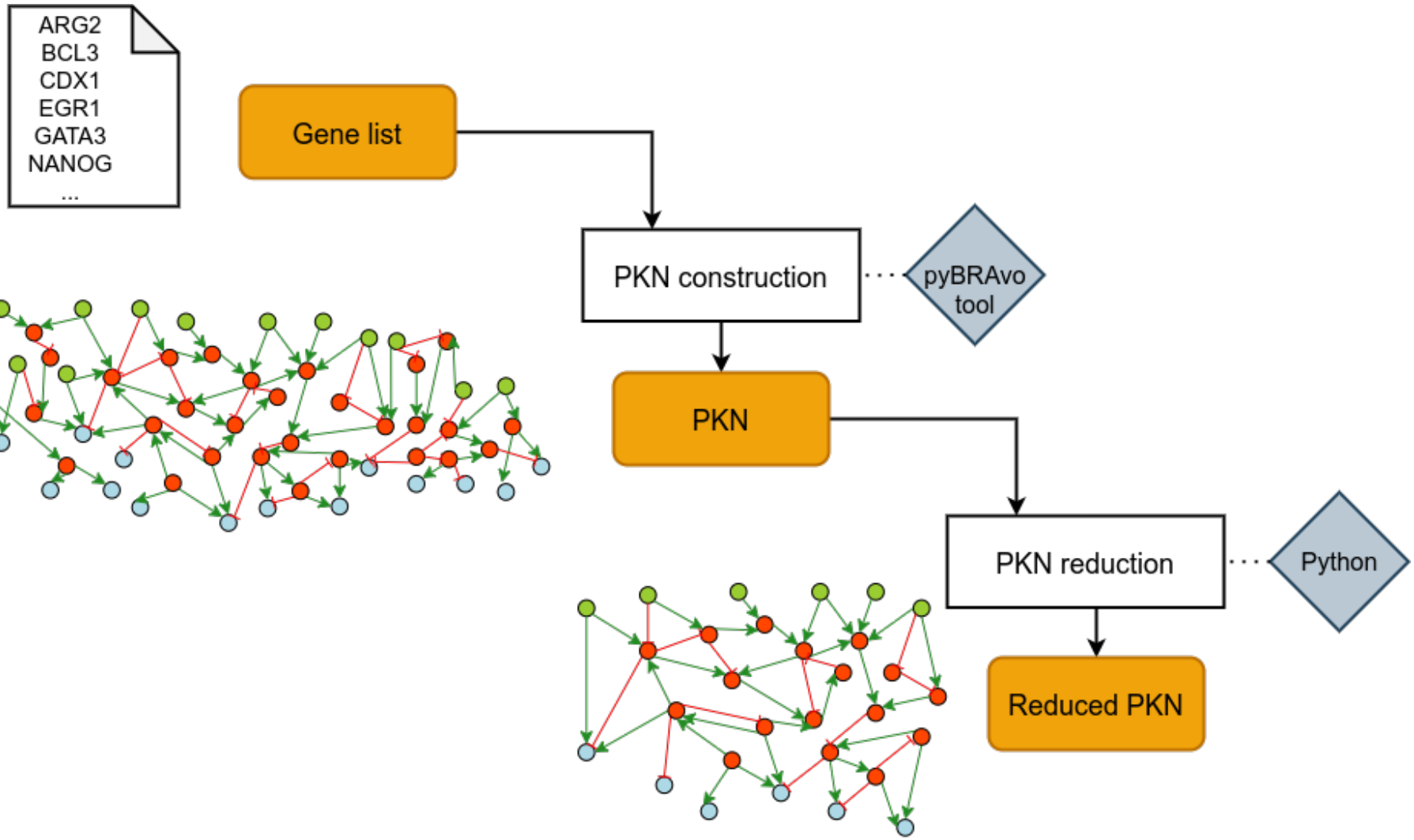

Prior-Knowledge Network

Experimental design

Predictive model

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

# Step 1. PKN reconstruction



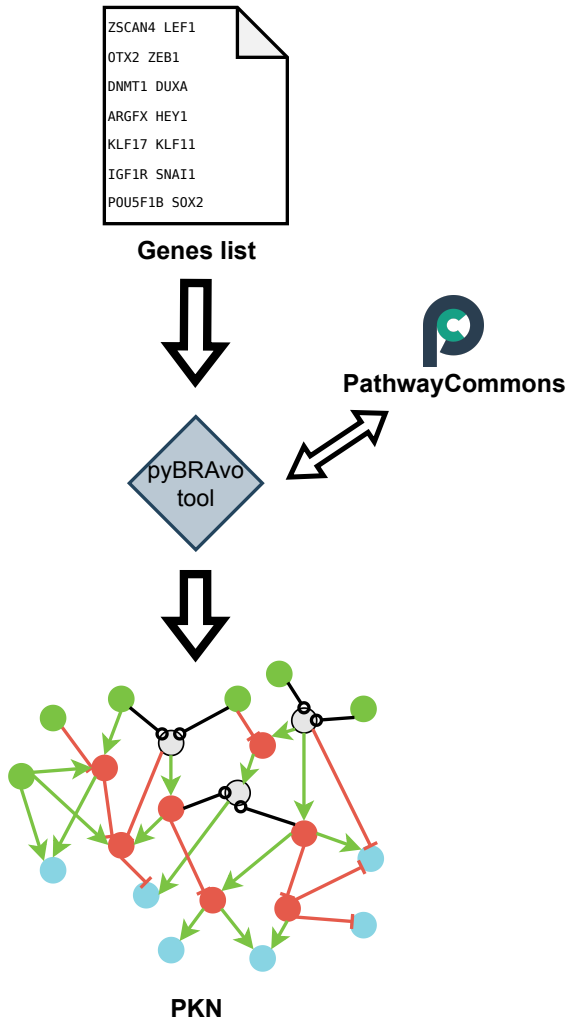
# Step 1. PKN reconstruction

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

Parameter: max depth

Output PKN

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




→ inputs & intermediates genes: entry 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)

# 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 ($k$ genes)

$$binarized = \begin{cases} 0, & \text{if } raw < 2, \\ 1, & \text{otherwise.} \end{cases}$$

# 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 ($k$ genes)

$$binarized = \begin{cases} 0, & \text{if } raw < 2, \\ 1, & \text{otherwise.} \end{cases}$$

- Normalization of readout genes (2 options)

"Min-Max" normalization

$$normalized = \frac{raw - min}{max - min}$$

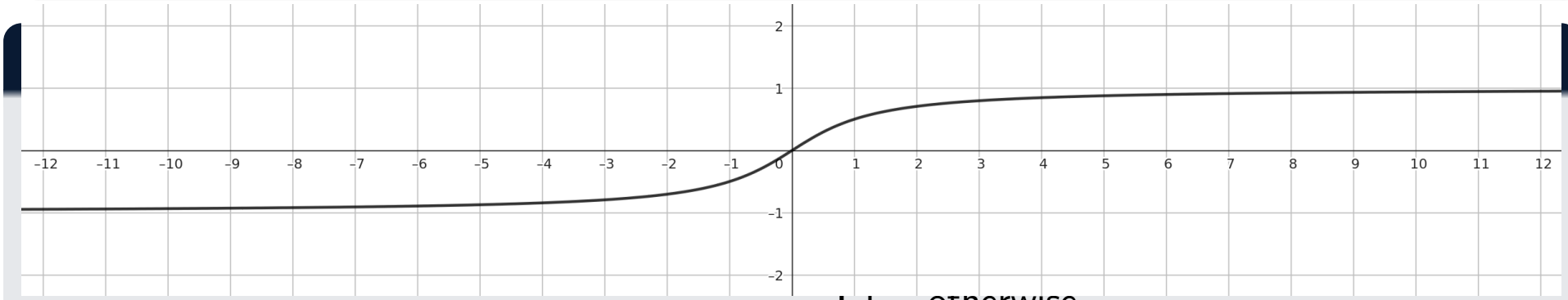
"Arctangeant" normalization

$$normalized = \frac{2}{\pi} \times \arctan(raw)$$

# Step 2. Experimental design reconstruction

## Idea

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




$\begin{cases} 1, & \text{otherwise.} \end{cases}$

- Normalization of readout genes (2 options)

"Min-Max" normalization

$$normalized = \frac{raw - min}{max - min}$$

"Arctangeant" normalization

$$normalized = \frac{2}{\pi} \times \arctan(raw)$$

# Step 2. Pseudo-perturbation identification

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 |

**ASP program**
***k*=3**
**Selected genes: A, C & D**

| 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 pseudo-perturbations: 2

# Step 2. Pseudo-perturbations identification algorithm

**Main rules (x4)**

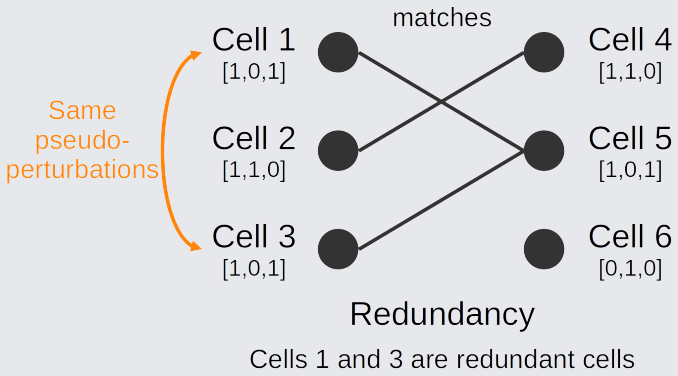
- k-genes: Select $k$ genes among all possible combinations of input + intermediate genes.
- Reachability: input $\rightarrow$ intermediate.
- Matching cells: Select pairs of cells $(c_1, c_2)$, $c_1 \in A$, $c_2 \in B$, for which the (binarized) expression matches for each of the *k-genes*.
- Filter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

**Optimization (x1)**

- Maximize the number of *pseudo-perturbations*.

# Step 2. Maximizing the readouts difference

## Redundancy




2 solutions:

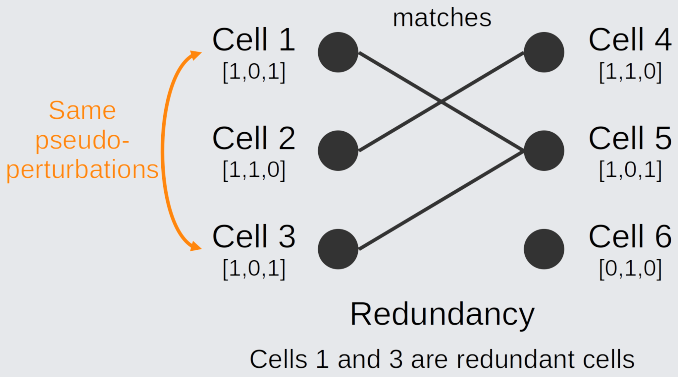
- (1,5), (2,4)
- (3,5), (2,4)

Pseudo-perturbations representativity:

- Class A: 100% (3/3)
- Class B: 66% (2/3)

# Step 2. Maximizing the readouts difference

## Redundancy




2 solutions:

- (1,5), (2,4)
- (3,5), (2,4)

Pseudo-perturbations representativity:

- Class A: 100% (3/3)
- Class B: 66% (2/3)

## Readout difference maximization

| 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 |

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

**vs.**

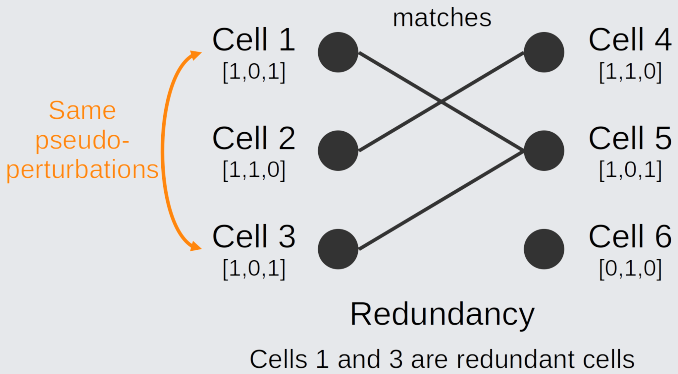
readout_diff(3,5) = |0.8-0.7| + |0.3-0.8| + |0.9-0.5| = **0.9**

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

# Step 2. Maximizing the readouts difference

## Redundancy




2 solutions:

- (1,5), (2,4)
- (3,5), (2,4)

Pseudo-perturbations representativity:

- Class A: 100% (3/3)
- Class B: 66% (2/3)

## Readout difference maximization

| 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 |

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

vs.

readout_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 |
| 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 |

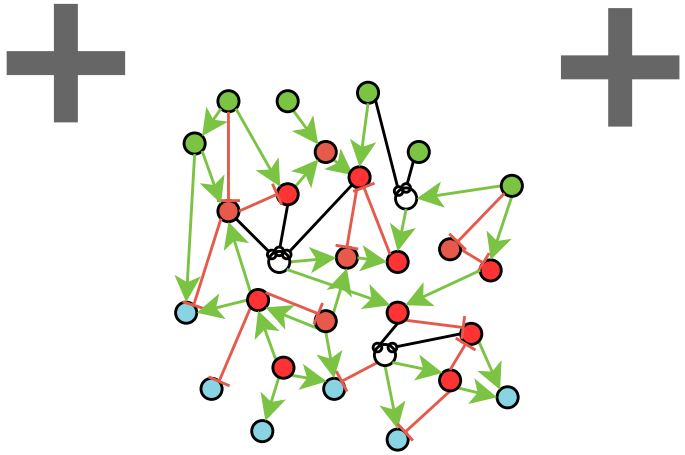
# Step 3. BNs inference using Caspo [Guziolowski *et al.* *Bioinformatics*, 2013.]

| 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

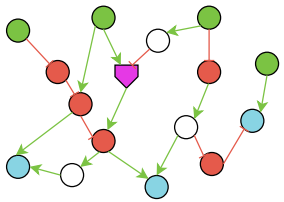
| Cell | A | C | D | F | G | H | Class |
|---|---|---|---|---|---|---|---|
| 5 | 1 | 0 | 1 | 0.7 | 0.8 | 0.5 | B |
| 4 | 1 | 1 | 0 | 0.6 | 0.1 | 0.2 | B |

Experimental design for Class B

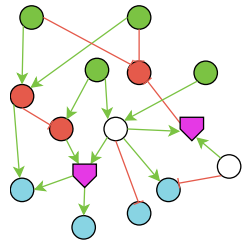


PKN

Learning Boolean Networks (Caspo)



Familiy of Boolean Networks for Class A



Familiy of Boolean Networks for Class B

# Reconstructed PKN – Max depth parameter [Bolteau *et al.*, 2024, *in prep.*]

Input: 438 transcription factors involved in human embryonic dev.

2 values tested:

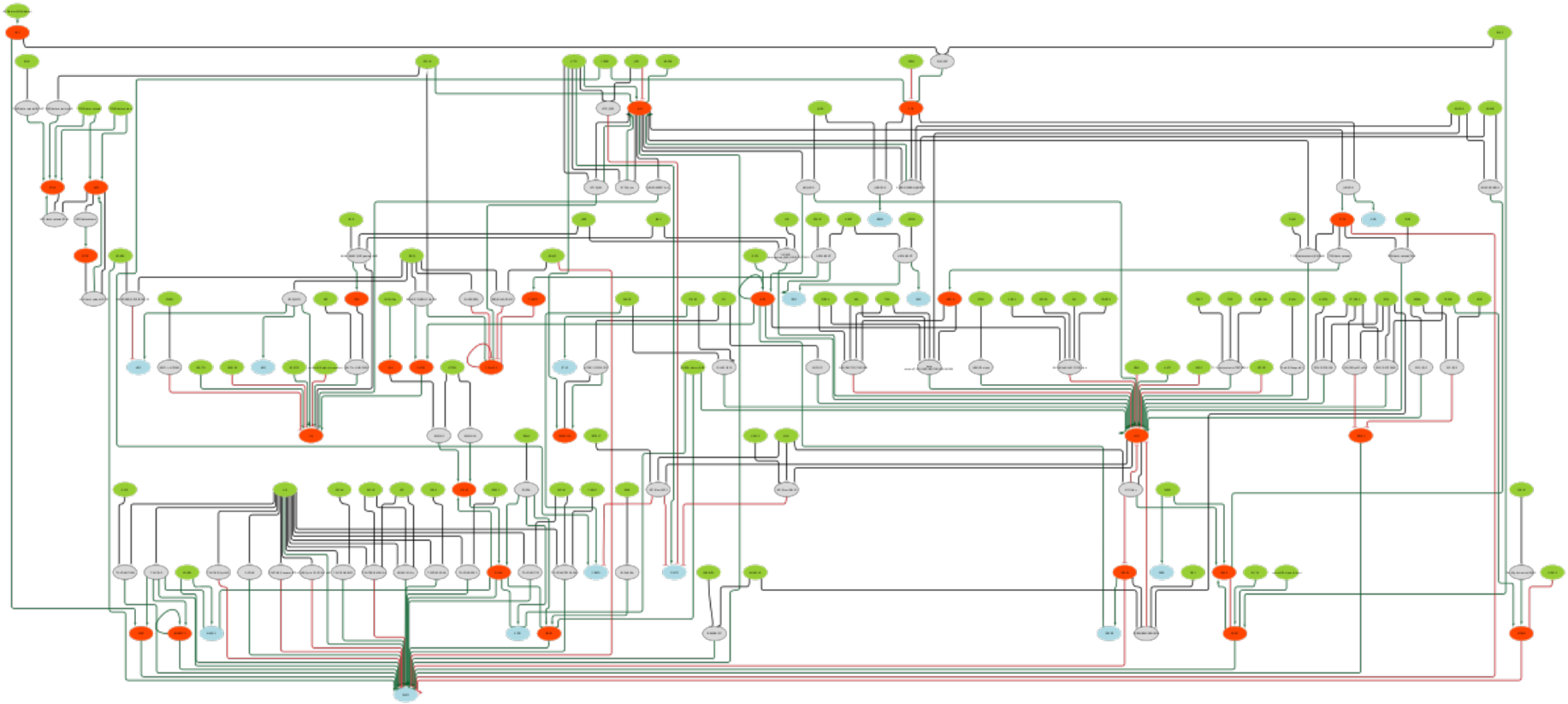
$max\ depth = 2$ (2 recursive queries) $\rightarrow PKN^2$
$max\ depth = 0$ (no depth, total reconstruction) $\rightarrow PKN^0$

| PKN | #nodes | #edges | #inputs | #intermediates | #readouts | #protein-complexes |
|---|---|---|---|---|---|---|
| $PKN^2$ | 191 | 285 | 84 | 27 | 14 | 66 |
| $PKN^0$ | 233 | 369 | 93 | 37 | 19 | 85 |

# Reconstructed $PKN^0$

[Bolteau *et al.*, 2024, *in prep.*]

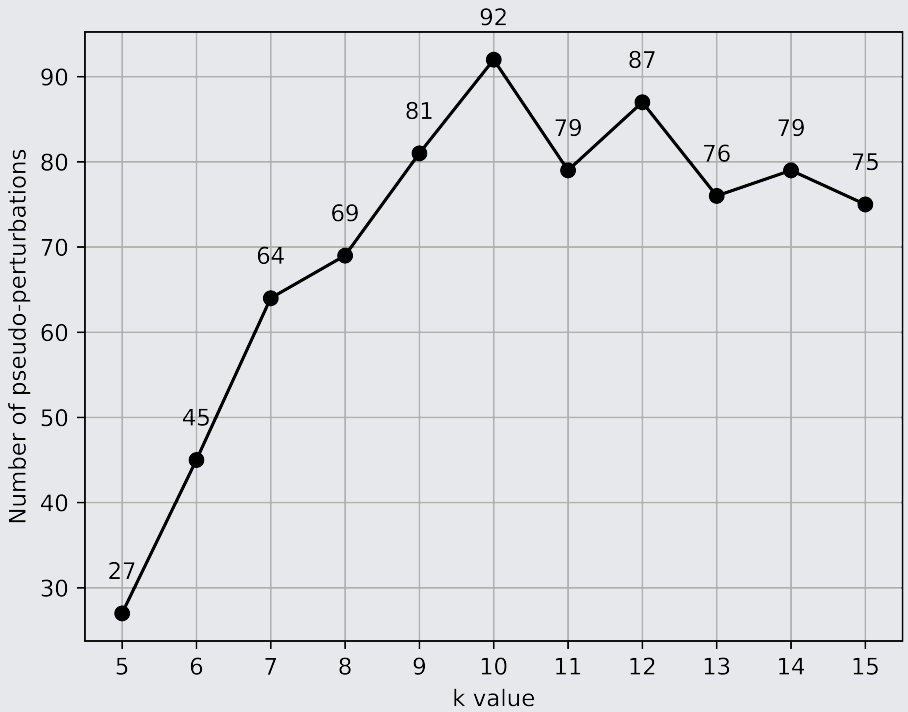
- 233 nodes : inputs (85), intermediates (36), readouts (19)
- 369 edges



# Pseudo-perturbations identification

[Bolteau *et al.*, 2024, *in prep.*]

## Exploration of the $k$-genes parameter



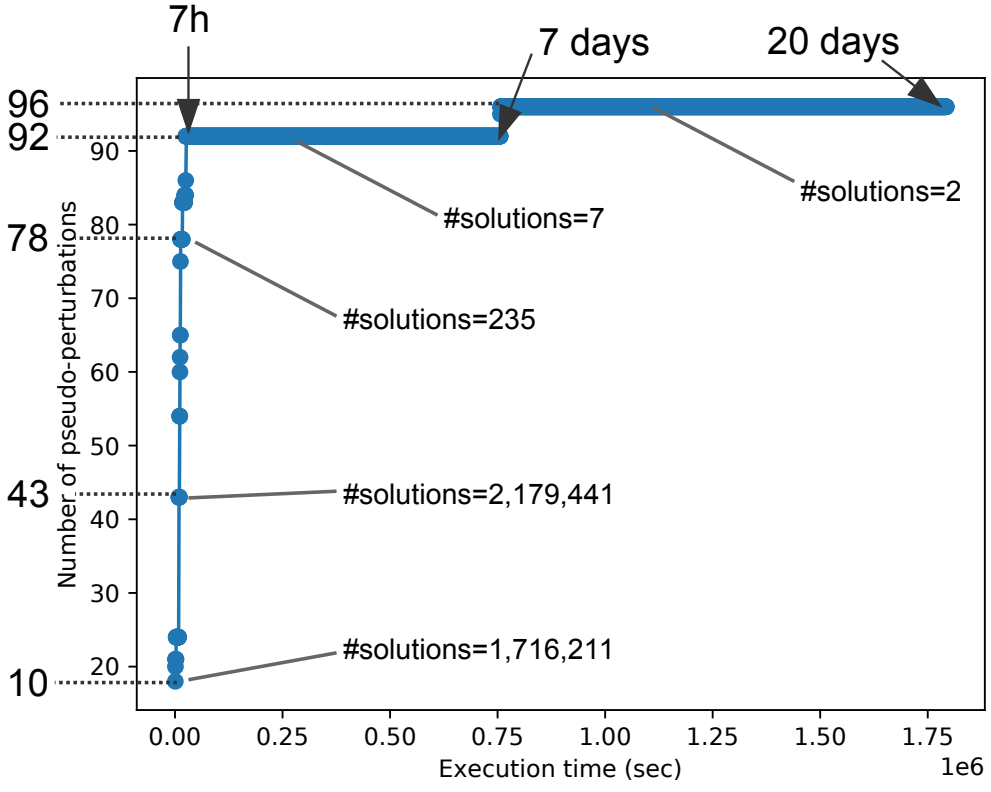

$k = 10$ is the best value to maximize the number of pseudo-perturbations

# Pseudo-perturbations identification

[Bolteau *et al.*, 2024, *in prep.*]

## 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



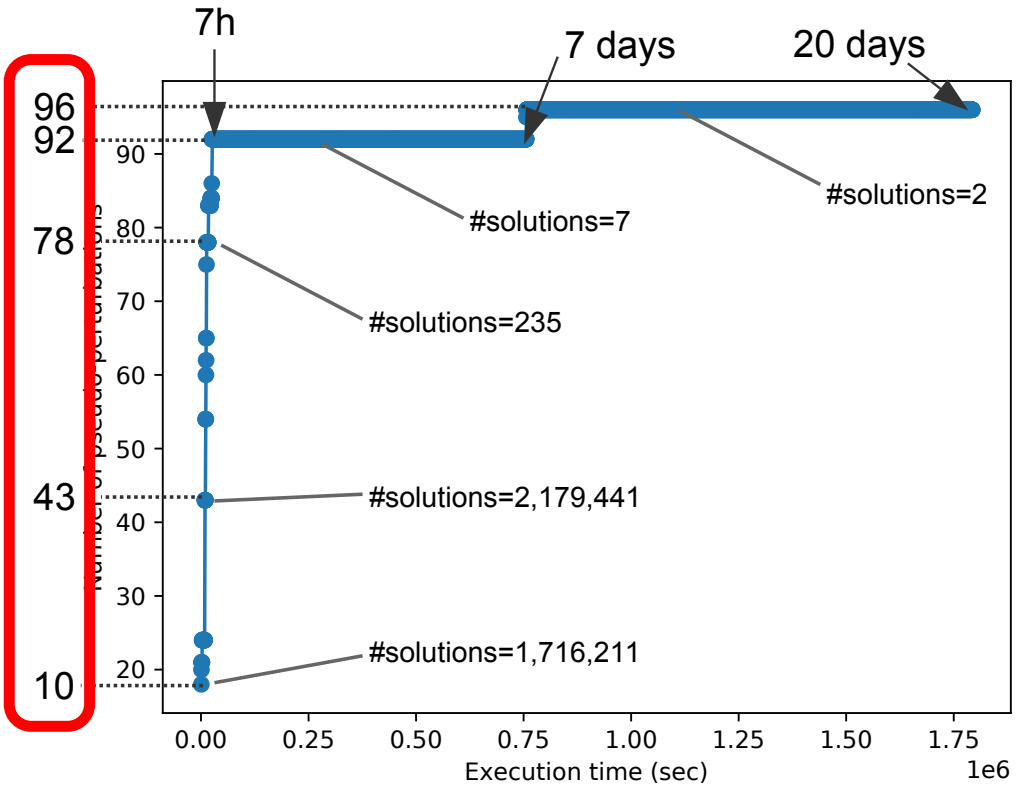

Convergence of the number of pseudo-pertubations over time.

# Pseudo-perturbations identification

[Bolteau *et al.*, 2024, *in prep.*]

## Inputs

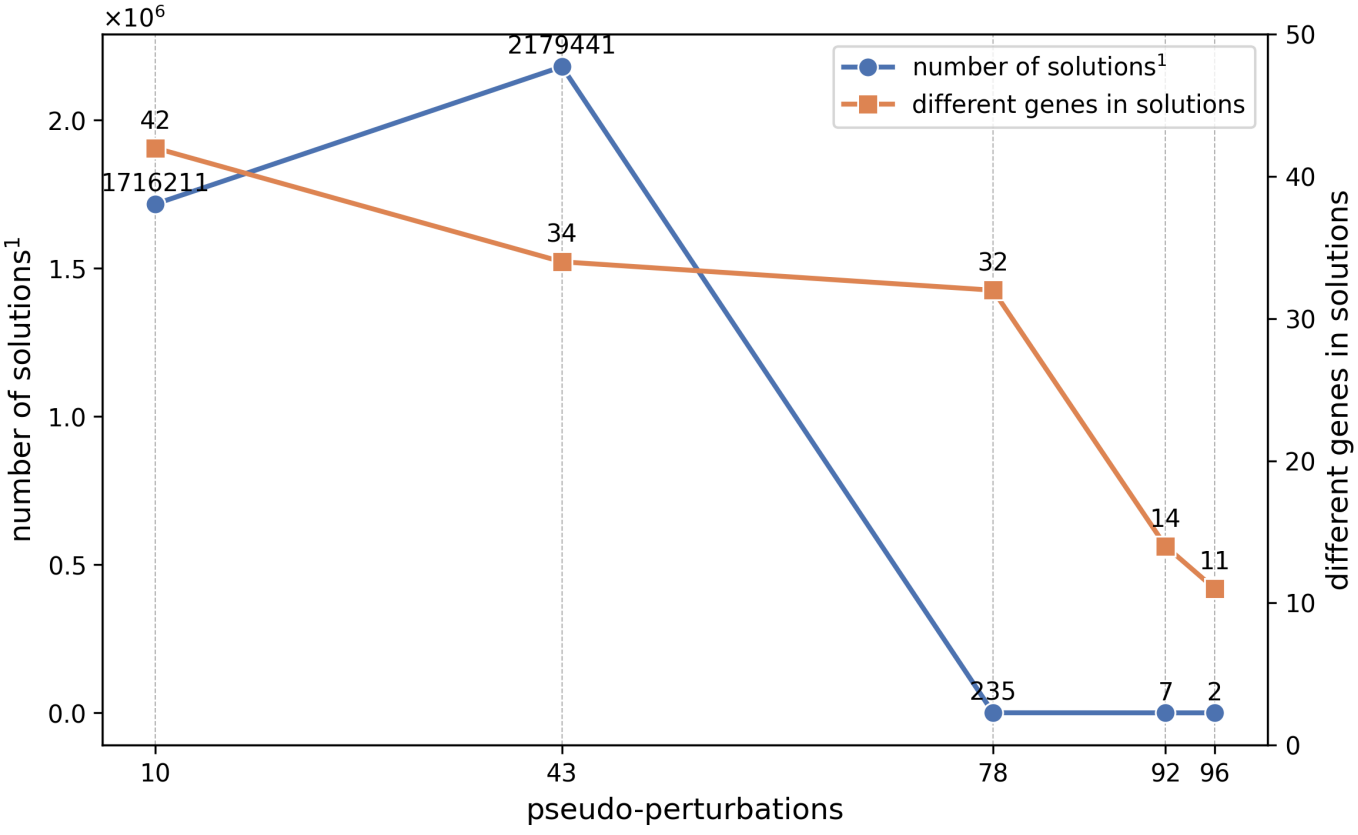
- $\#M^{TE}$ *cells* $= 348$
- $\#L^{TE}$ *cells* $= 332$
- $k = 10$: 10 genes selected from 121 input and intermediate genes
- Search space: $\binom{121}{10} = 1.27x10^{14}$ possible choices




Convergence of the number of pseudo-pertubations over time.

# Robustness of solutions

[Bolteau *et al.*, 2024, *in prep.*]

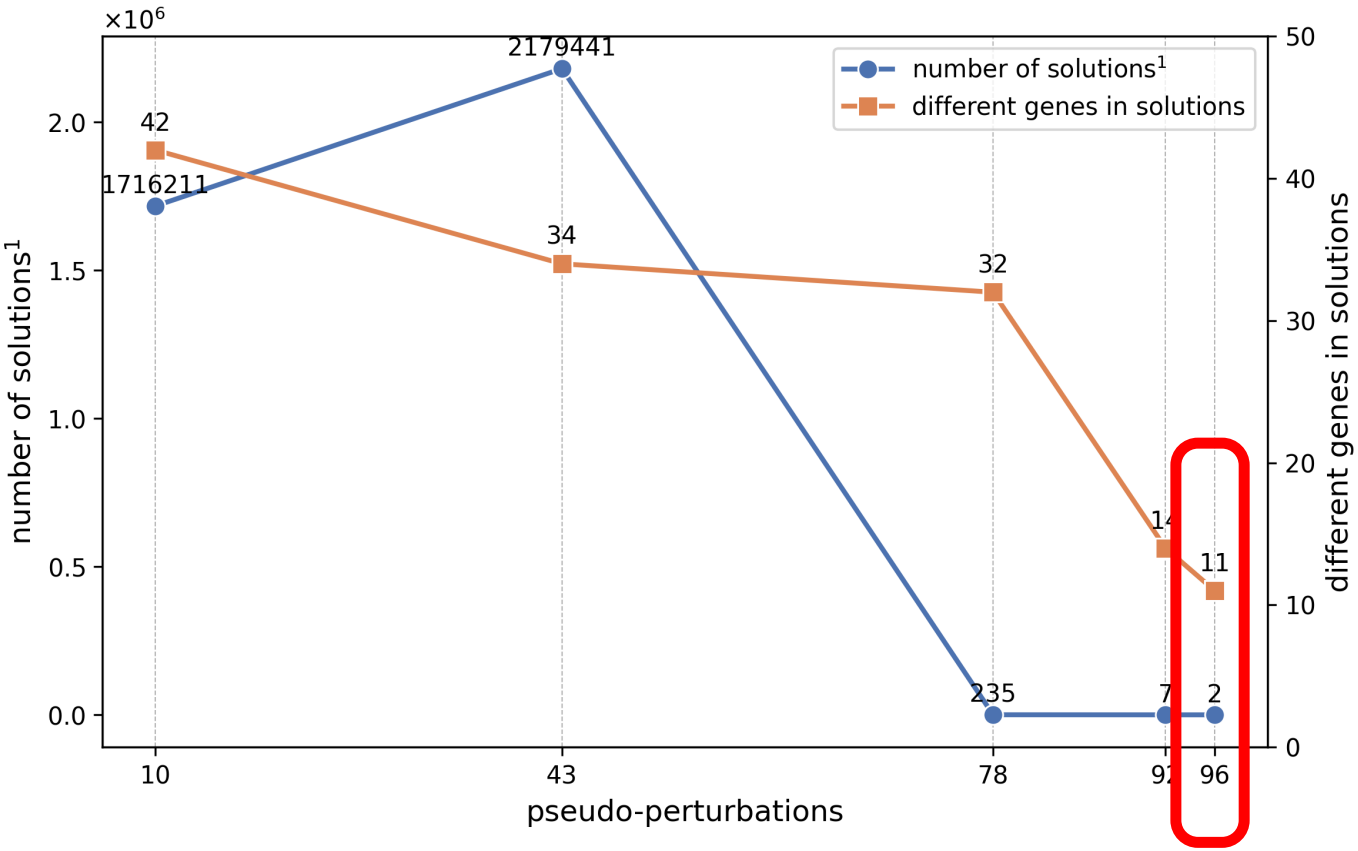



- The more pseudo-perturbations we have, the fewer different genes we have in the solutions
- Gene number explosion when few pseudo-perturbations

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

# Robustness of solutions

[Bolteau *et al.*, 2024, *in prep.*]

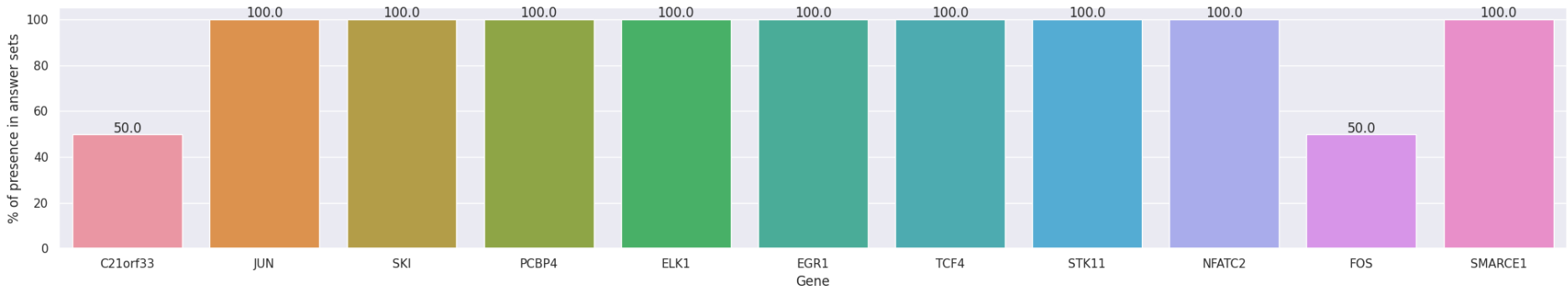



- The more pseudo-perturbations we have, the fewer different genes we have in the solutions
- Gene number explosion when few pseudo-perturbations

[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 $L^{TE}$

# Pseudo-perturbation representativity (redundancies)

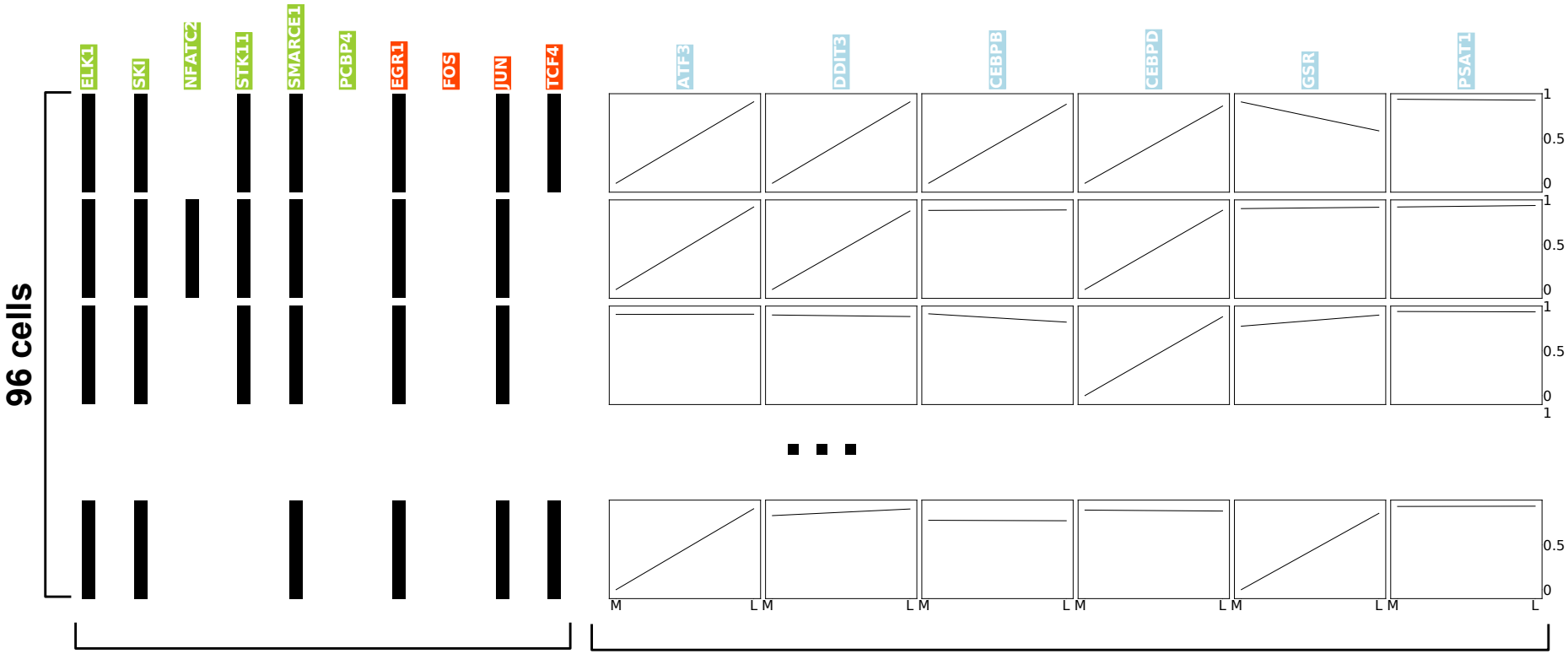
| Solution | $M^{TE}$ (%) | $L^{TE}$ (%) | Total (%) |
|---|---|---|---|
| **1** | **266 (76%)** | **246 (74%)** | **512 (75%)** |
| 2 | 235 (68%) | 248 (75%) | 483 (71%) |

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

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

# Experimental designs

[Bolteau *et al.*, 2024, *in prep.*]

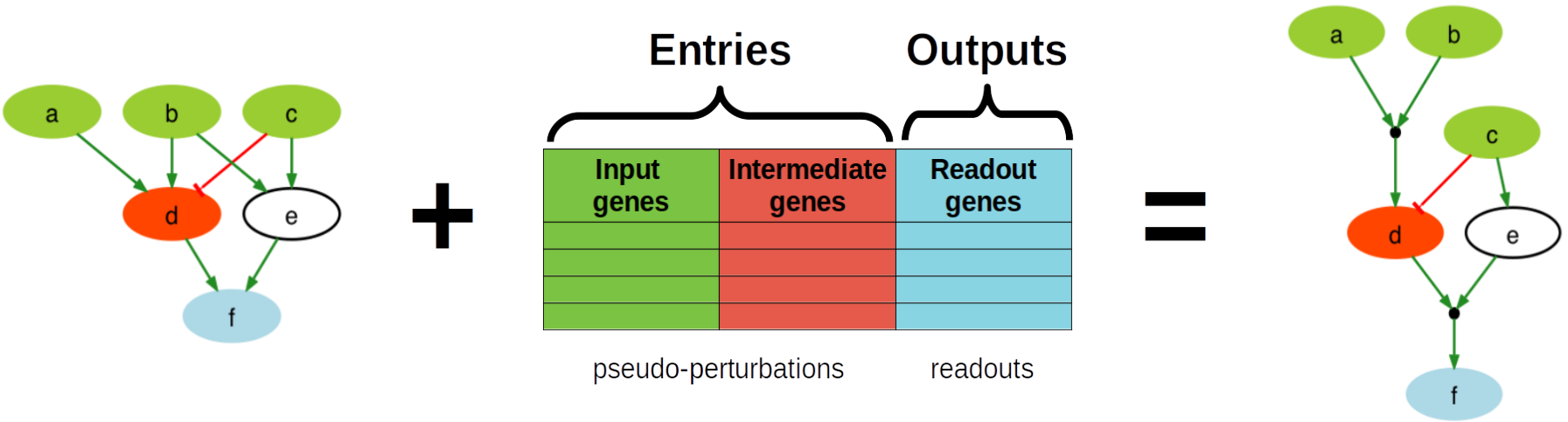



**10 inputs and intermediates genes with identical expression for $M^{TE}$ and $L^{TE}$ ($k$=10)**

**Expression evolution between $M^{TE}$ (left) and $L^{TE}$ (right) for 6 readout genes (susbset)**

# Learning Boolean logic models

[Bolteau *et al.*, 2024, *in prep.*]




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 – "Min-Max" normalization

[Bolteau *et al.*, 2024, *in prep.*]

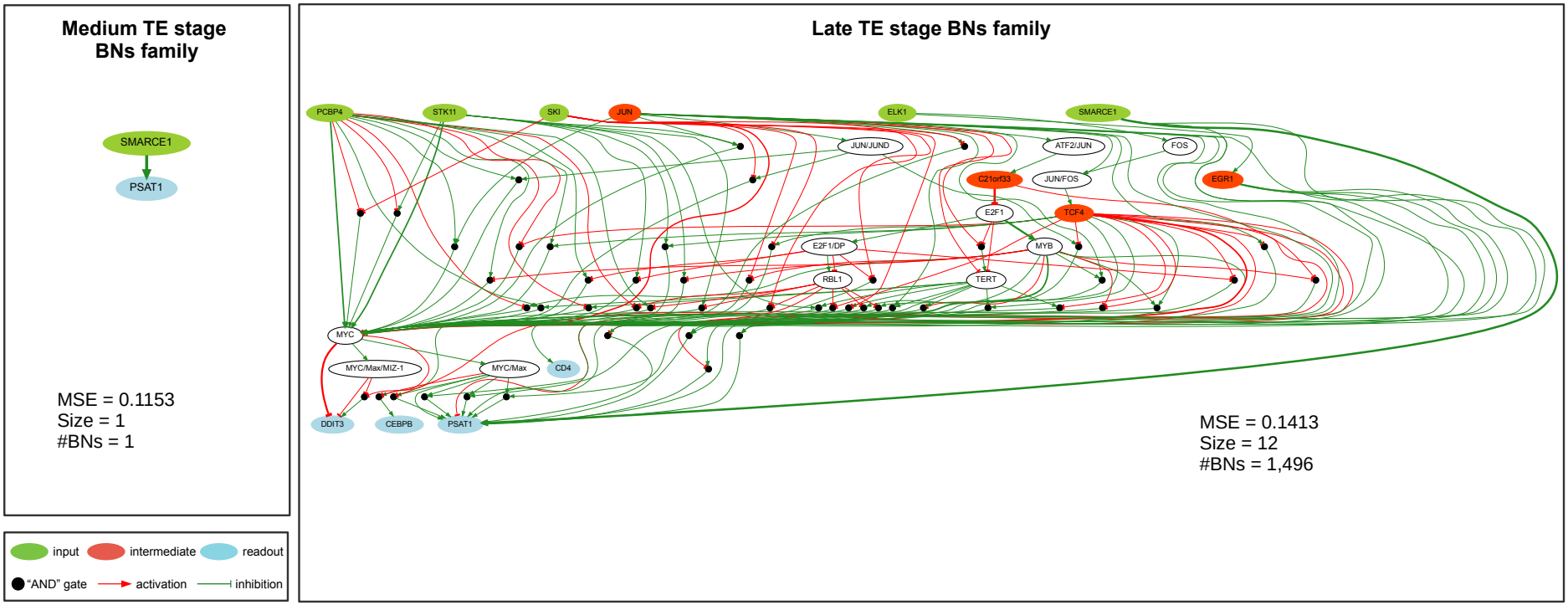
| Solution | MSE | | Size | | #Networks | |
|---|---|---|---|---|---|---|
| | $M^{TE}$ | $L^{TE}$ | $M^{TE}$ | $L^{TE}$ | $M^{TE}$ | $L^{TE}$ |
| 1 | 0.1153 | 0.1413 | 1 | 12 | 1 | 1,496 |
| 2 | 0.1180 | 0.1400 | 1 | 5 | 1 | 199 |

MSE: distance between readout predictions through the BNs and the "real" readout values.
Size: number of logic clauses.

#Networks: number of inferred (sub-)optimal BNs.

- Larger MSE for $L^{TE} \rightarrow L^{TE}$ more difficult to fit
- More redundancies for $L^{TE}$ (number of BNs) $\rightarrow$ different ways to explain the "entry-output" relation with Boolean gates
- $M^{TE}$ size seem irrelevant

# Inferred BNs for solution 1 – "Min-Max" normalization




- More readouts implicated in $L^{TE}$ stage
- Greater BNs variability for $L^{TE}$ $\rightarrow$ Gain of function
- $M^{TE}$ BNs family biologically irrelevant

# Caspo metrics – "Artangeant" normalization [Bolteau *et al.*, 2024, *in prep.*]

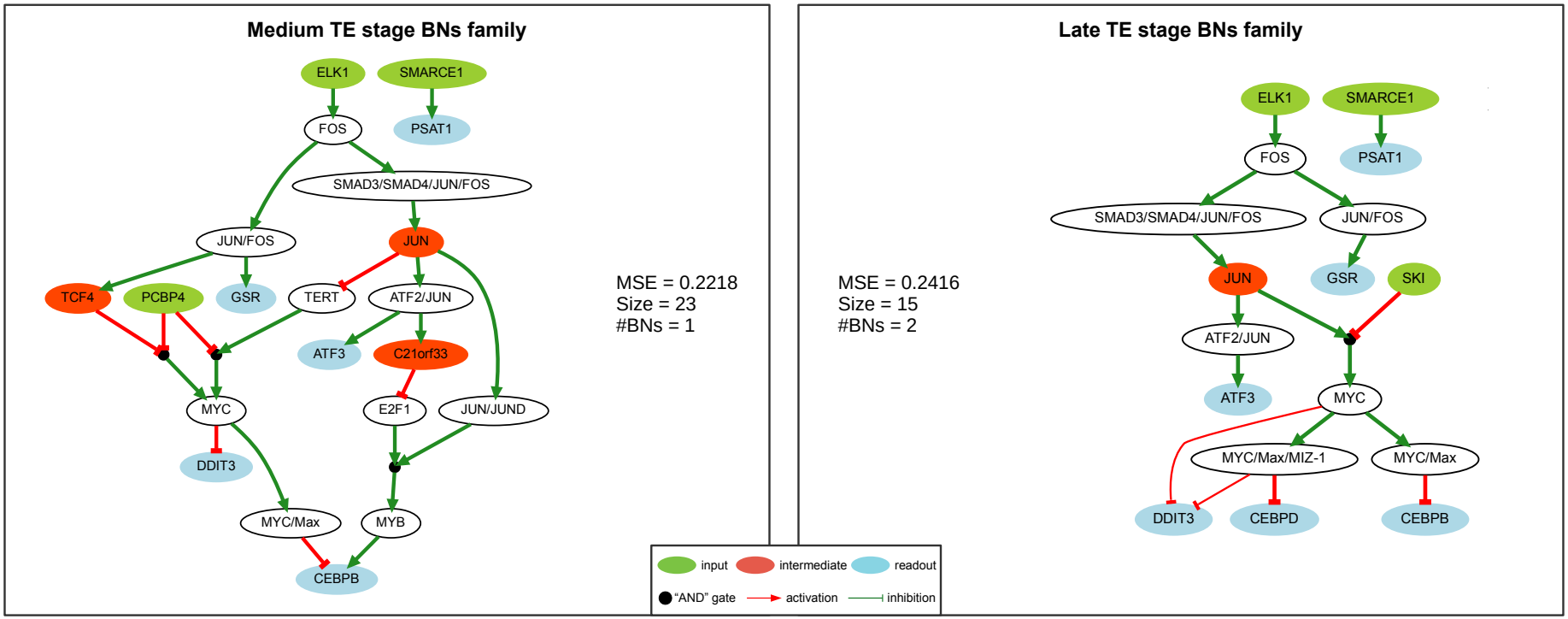
| Solution | MSE | | Size | | #Networks | |
|---|---|---|---|---|---|---|
| | $M^{TE}$ | $L^{TE}$ | $M^{TE}$ | $L^{TE}$ | $M^{TE}$ | $L^{TE}$ |
| 1 | 0.2218 | 0.2416 | 23 | 15 | 1 | 2 |
| 2 | 0.2242 | 0.2516 | 21 | 16 | 214 | 70 |

MSE: distance between readout predictions through the BNs and the "real" readout values.
Size: number of logic clauses.

#Networks: number of inferred (sub-)optimal BNs.

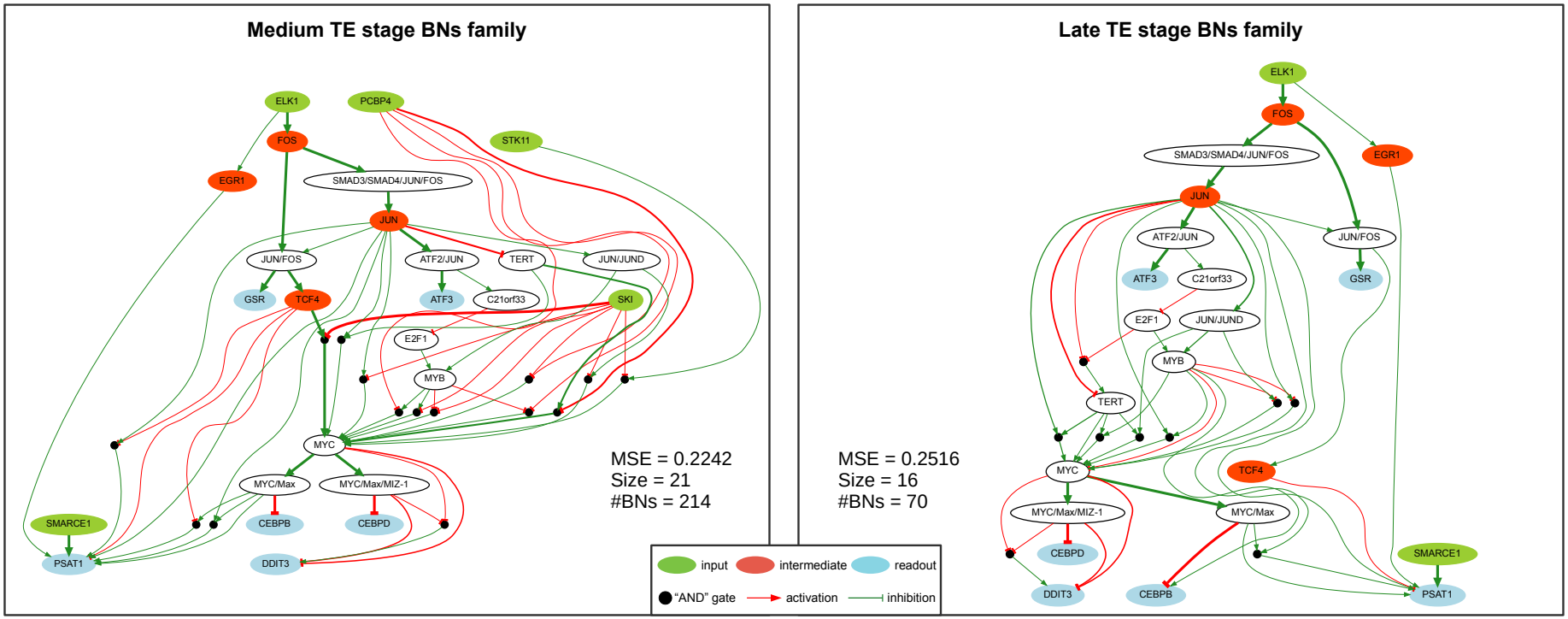
- Larger MSE for $L^{TE} \rightarrow L^{TE}$ more difficult to fit
- Size and #BNs greater for $M^{TE}$

# Inferred BNs for solution 1 – "Arctangeant" normalization




- Disconnected BNs
- Need for in-depth analysis

# Inferred BNs for solution 2 – "Arctangeant" normalization




- More inputs implicated in $M^{TE}$ stage, same number of intermediates and readouts
- Greater BNs variability for $M^{TE}$

# Conclusion

## Pseudo-perturbation generation

- Efficient algorithm to select cells and genes to generate pseudo-perturbations $\rightarrow$ 92 pseudo-perturbations in 7 hours
- Robustness of the generated solutions $\rightarrow$ from $+2$ millions of solutions to only 2
- 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
- Case-study adaptable method: optional PKN reconstruction step
- 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

- Deepen the results obtained for the normalization "Arctangeant"
- A more accurate PKN
- Study the impact of different discretization methods
- 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