Inferring gene networks with single-cell data: from mechanistic modelling to statistics

Ulysse Herbach

Biohasard

10 June 2021

Biological context

How do cells make decisions?

Example of decision making:

Differentiation: “stem” cell $\longrightarrow$ “mature” cell

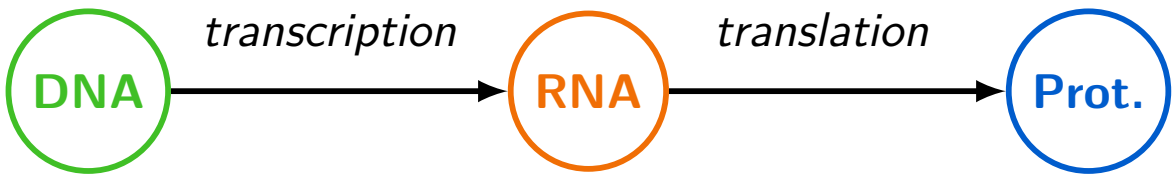
Biological context

How do cells make decisions?

Example of decision making:

Differentiation: "stem" cell $\longrightarrow$ "mature" cell

Fundamental diagram of molecular biology:



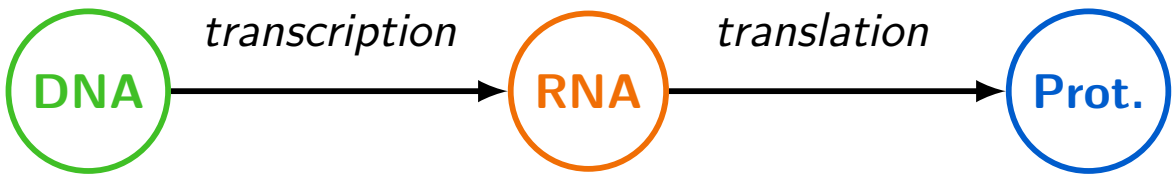
Biological context

How do cells make decisions?

Example of decision making:

Differentiation: "stem" cell $\longrightarrow$ "mature" cell

Fundamental diagram of molecular biology:



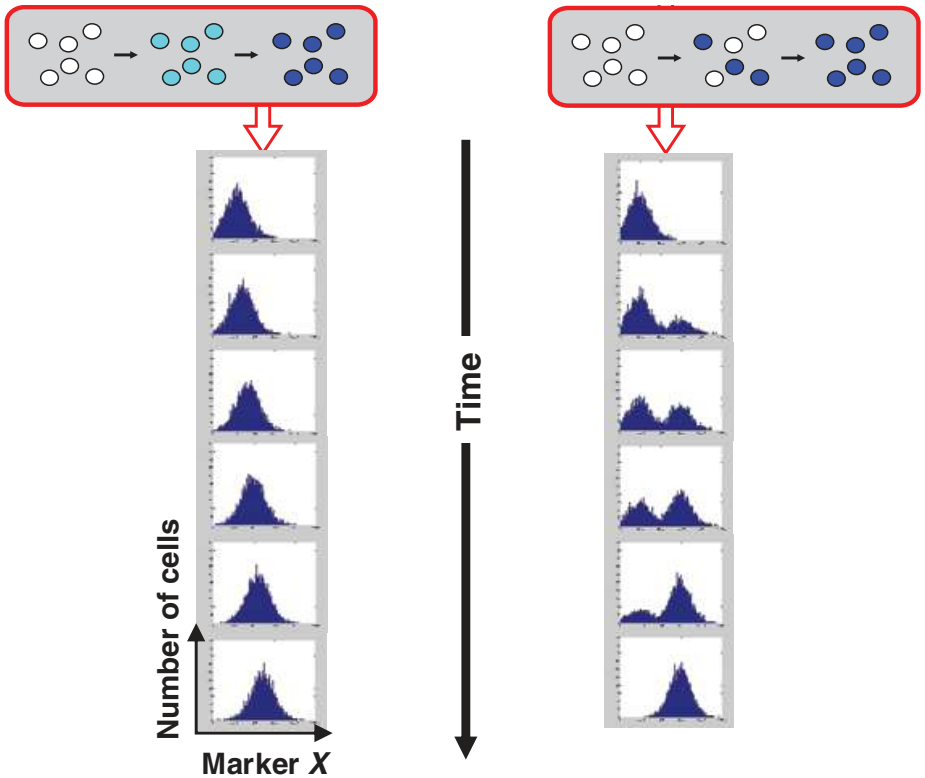
Basic idea of systems biology:

The behaviour of a cell emerges from interactions between genes

1. Why a stochastic model?

Differentiation: change of paradigm

« old school »



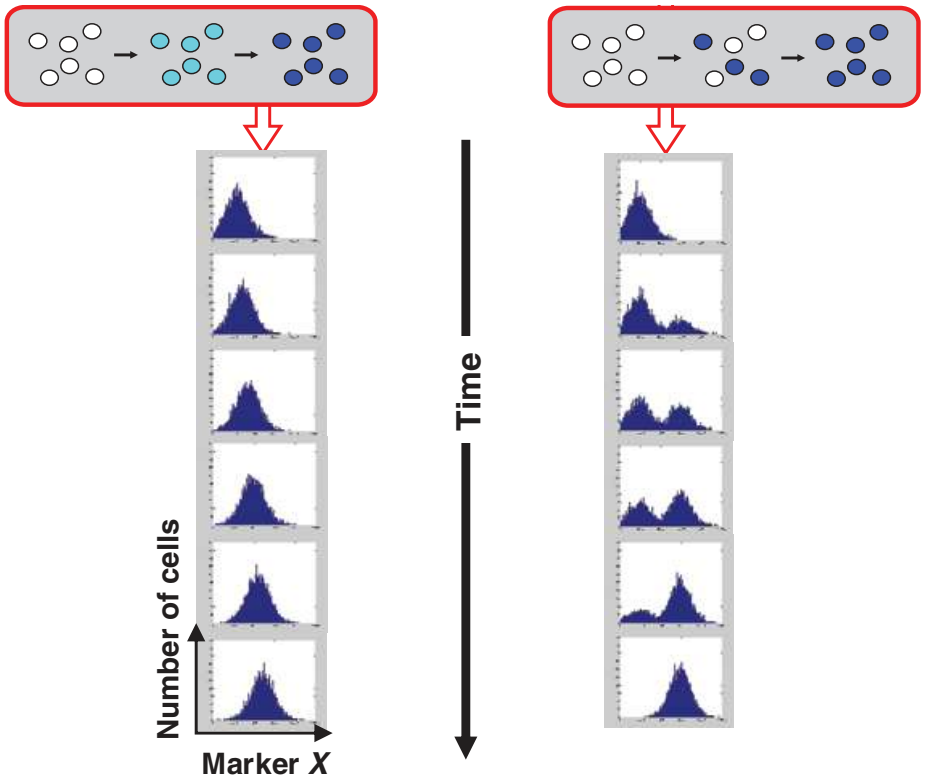
« new school »

S. Huang, Non-genetic heterogeneity of cells in development: more than just noise.

Development, 2009

Differentiation: change of paradigm

« *old school* »



« *new school* »

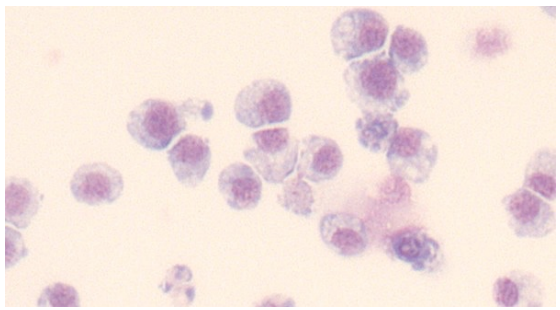
S. Huang, Non-genetic heterogeneity of cells in development: more than just noise. *Development*, 2009

Remark

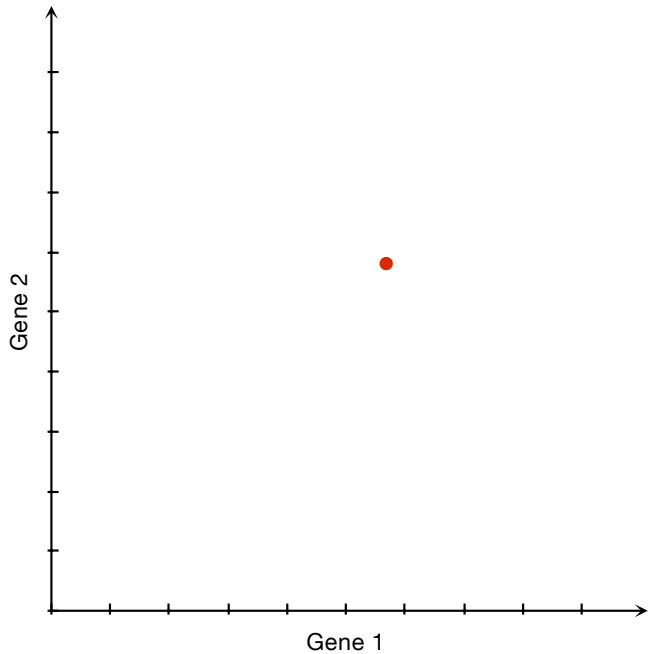
Seen on average (aka population or "bulk" data) as historically, these two paradigms are not distinguishable.

Population data

Cell population

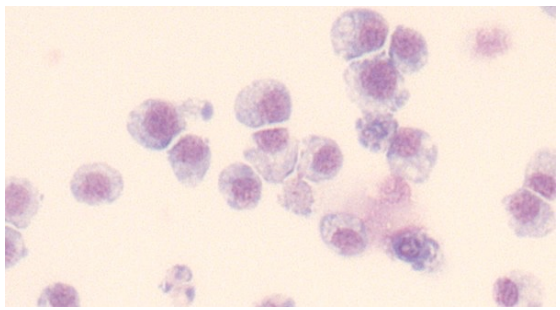


Population average

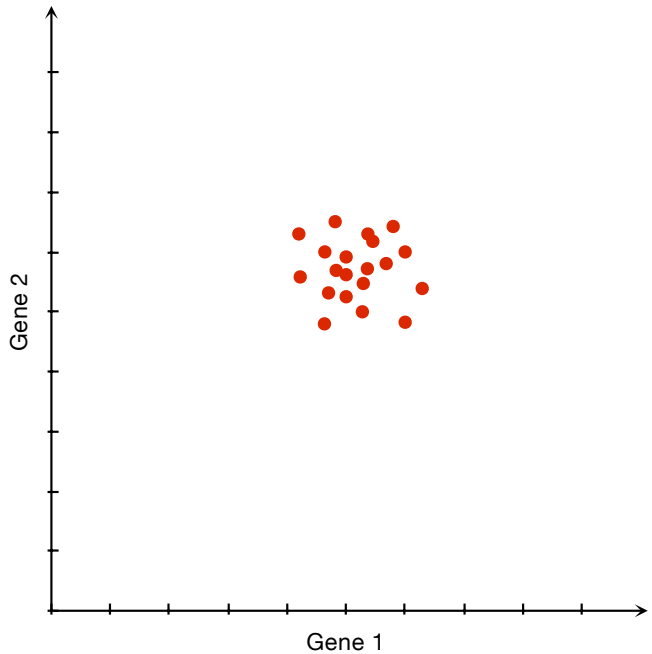


Population data

Cell population



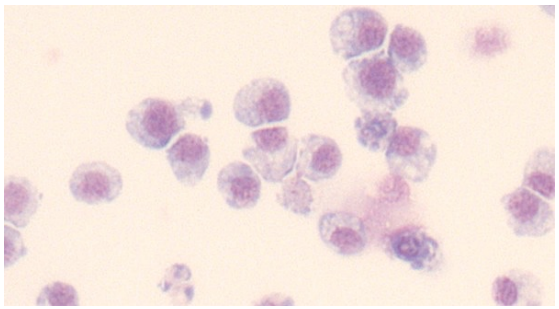
Repeats



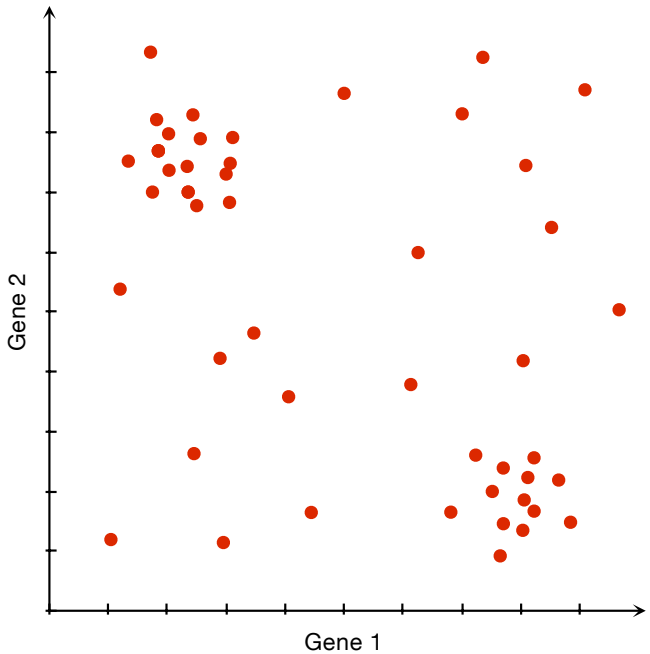
- Repeats only reveal technical noise
- To get biological variability, one has to *change conditions*

Single-cell data

Individual cells

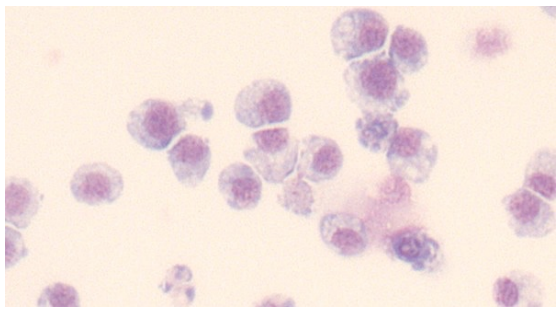


Single cells, 1 exp.

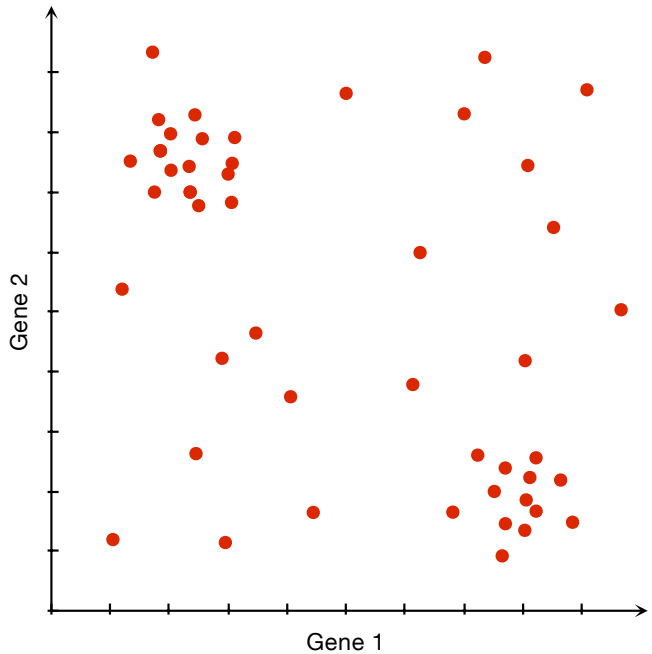


Single-cell data

Individual cells

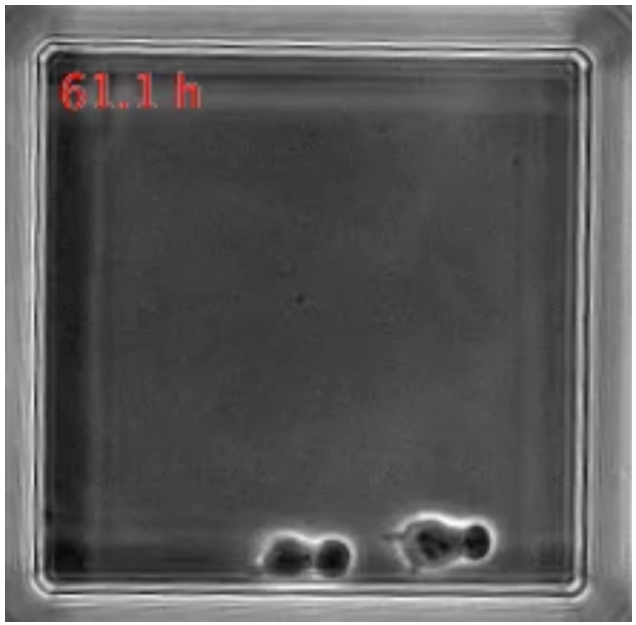


Single cells, 1 exp.



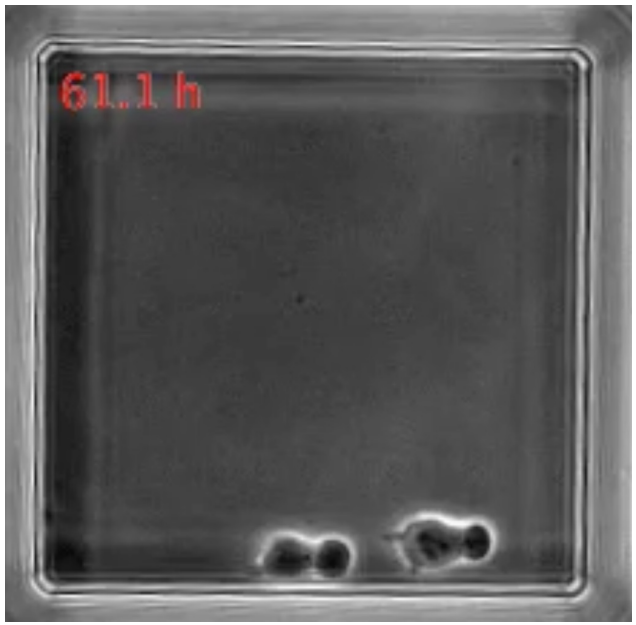
- Variability looks important...
- Sub-populations may appear: “molecular phenotypes”

A visual example



Moussy *et al*, Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. *PLOS Biology*, 15(7), 2017

A visual example

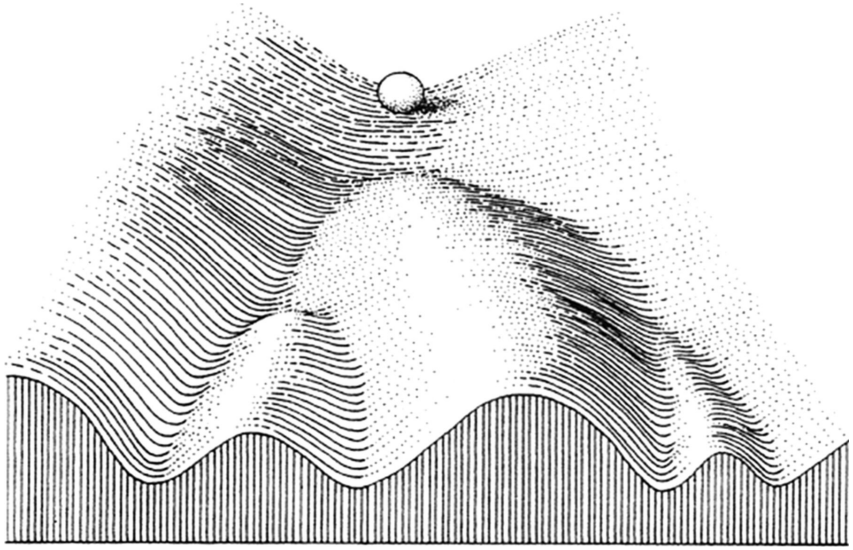


Moussy *et al*, Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. *PLOS Biology*, 15(7), 2017

New paradigm

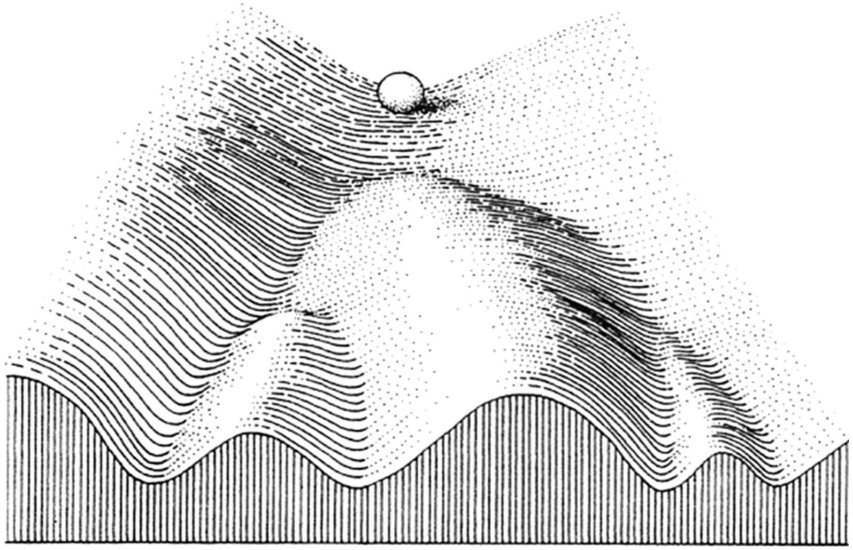
Gene expression is a stochastic phenomenon!

A modern view of Waddington landscapes

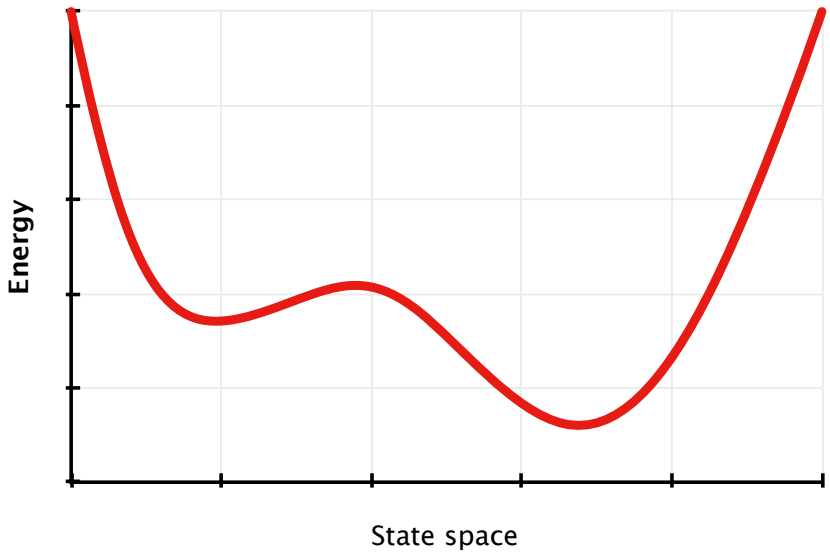


$\leftarrow$ 1942 !!!

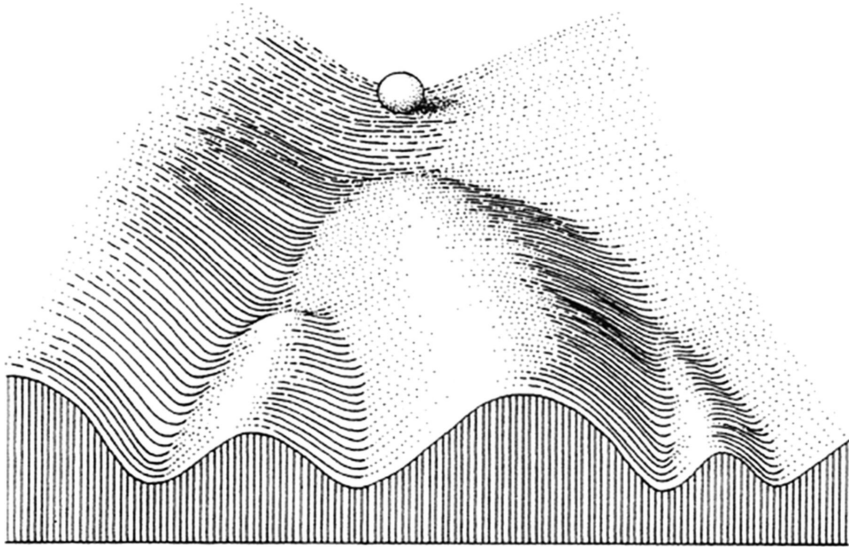
A modern view of Waddington landscapes



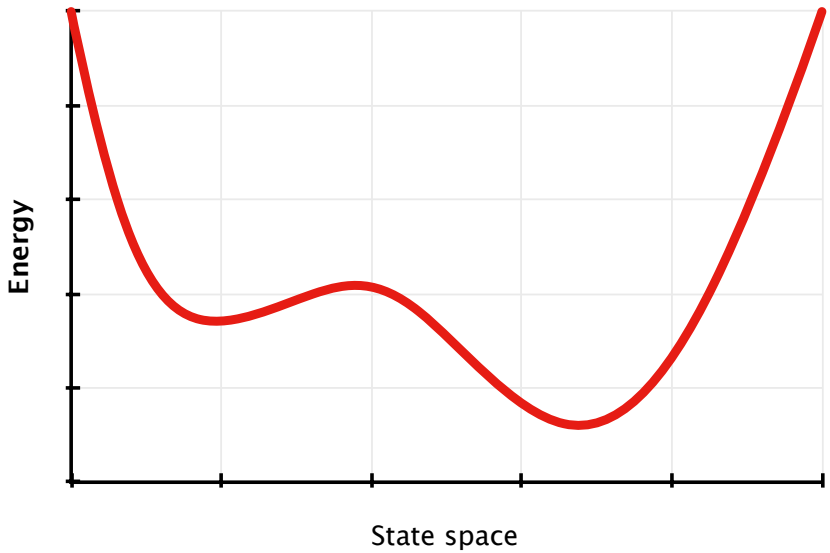
← 1942 !!!

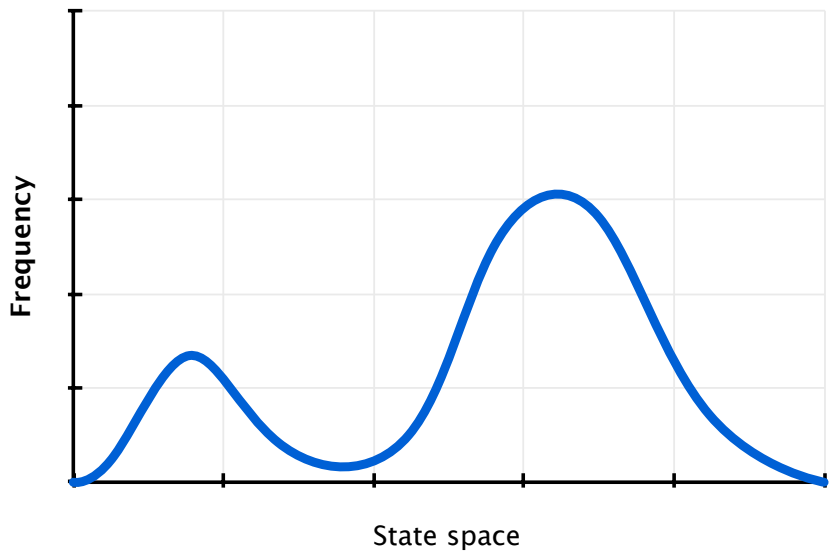


A modern view of Waddington landscapes



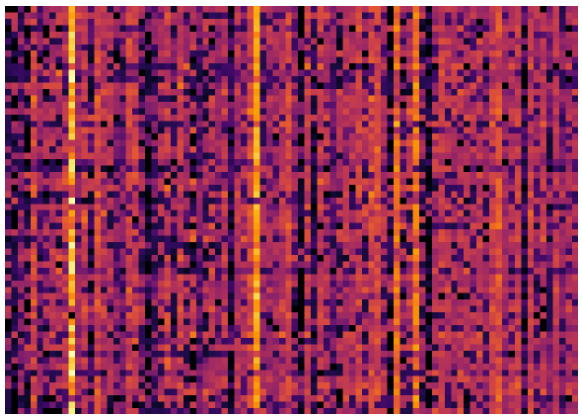
← 1942 !!!





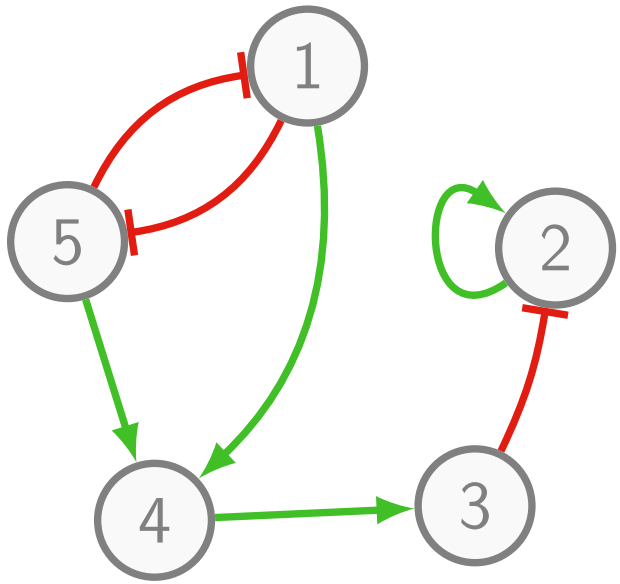
Statistical question

Gene expression levels



?

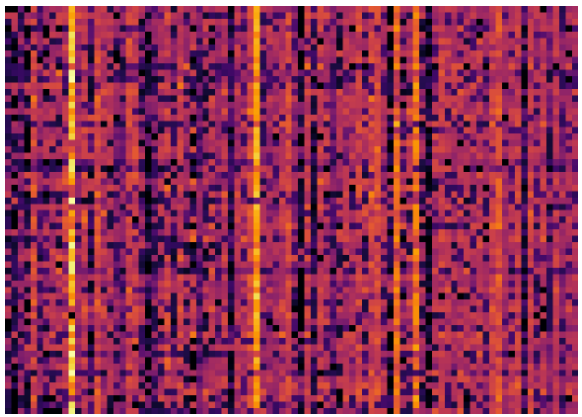
Gene regulatory network



"Gold standard" dilemma: use real data (*uncertain network*) or simulated data (*known network but unrealistic data*)?

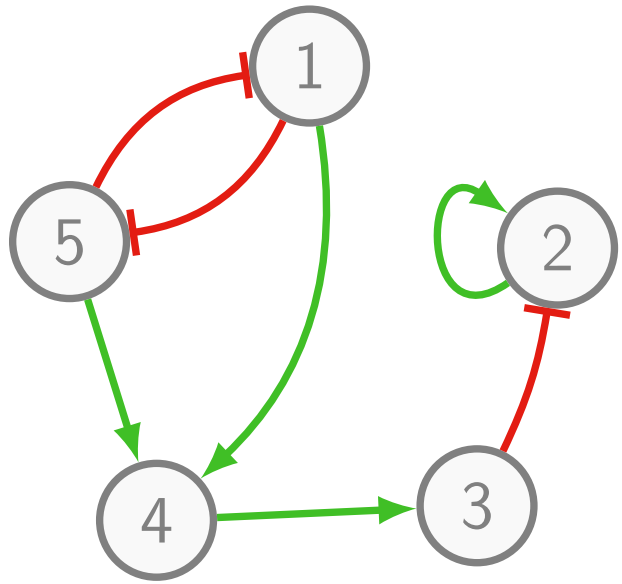
Statistical question

Gene expression levels



?

Gene regulatory network

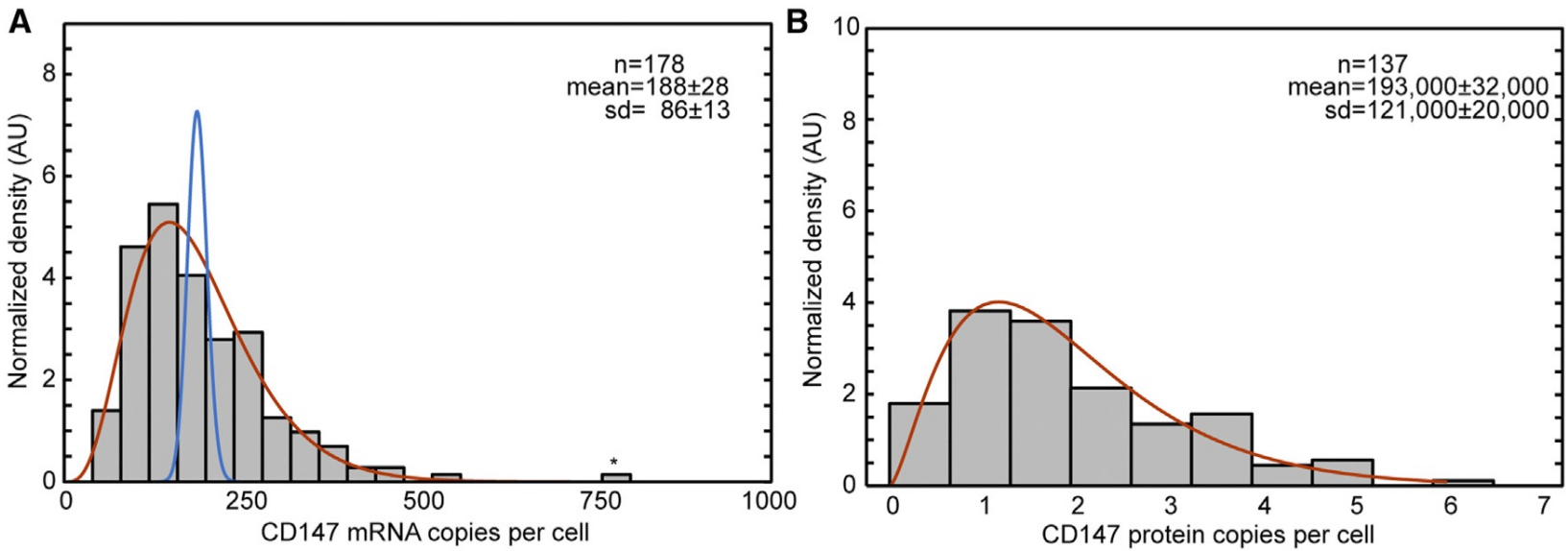


"Gold standard" dilemma: use real data (*uncertain network*) or simulated data (*known network but unrealistic data*)?

Our approach

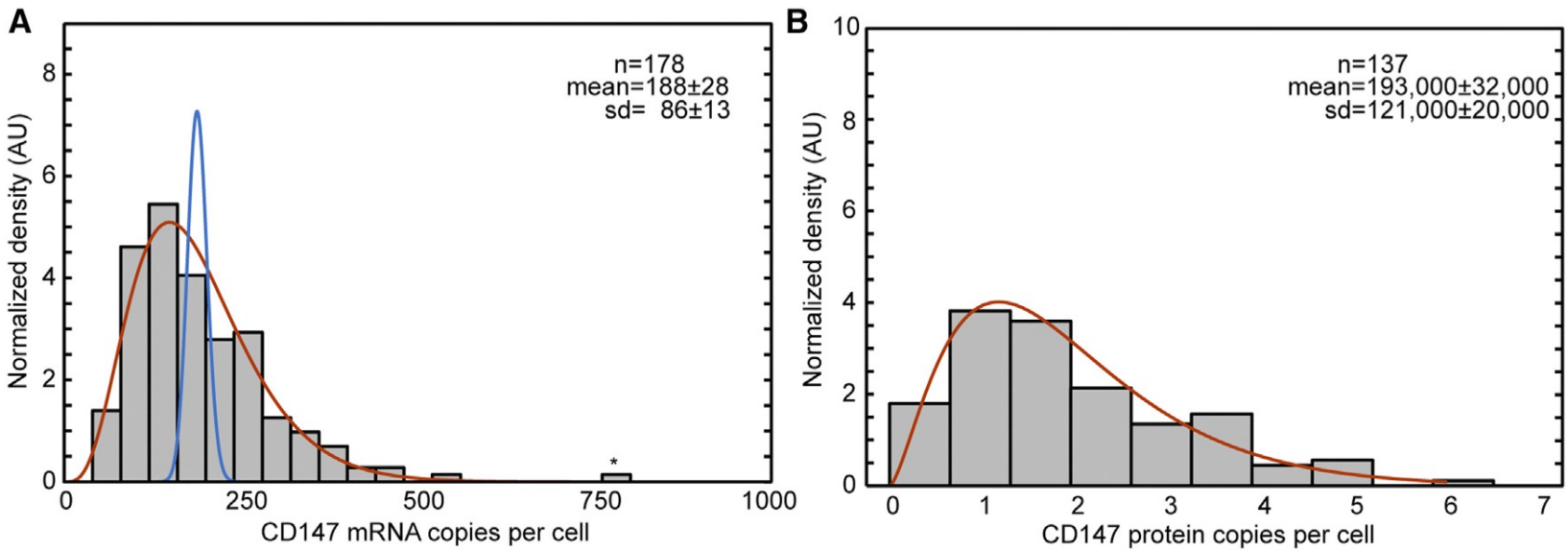
1. Build a mechanistic (*hence stochastic*) gene network model
2. **Calibrate** the model, which will correspond to **infer** a network

What kind of stochasticity?



Albayrak *et al*, Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell*, 61:914–924, 2016

What kind of stochasticity?



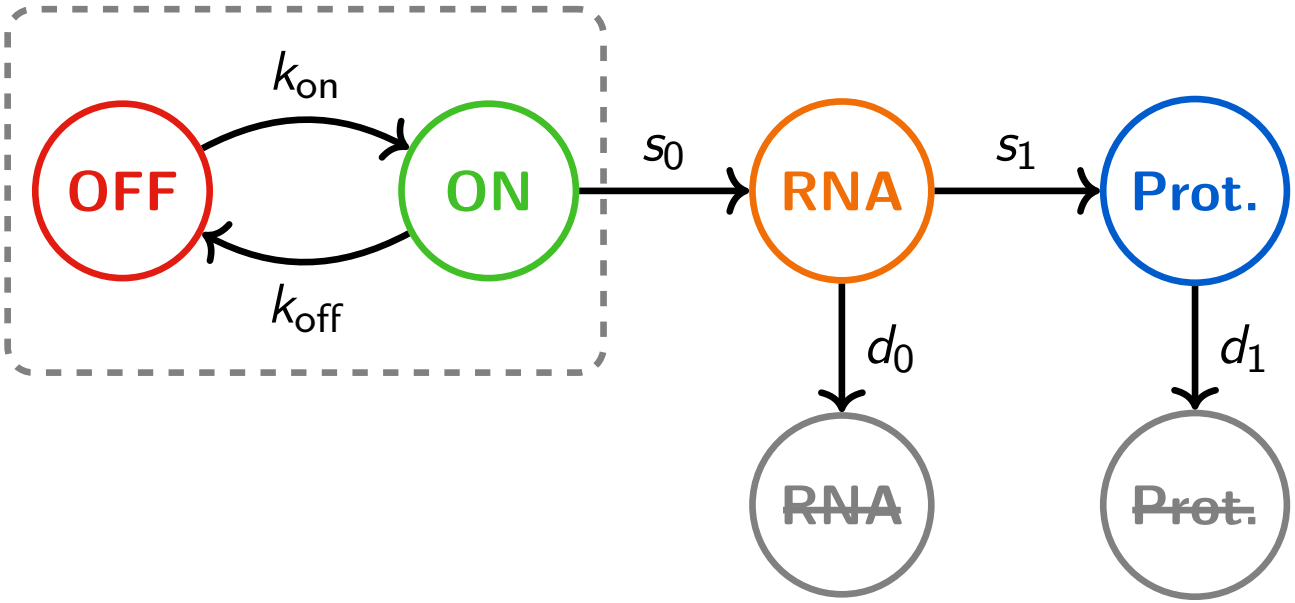
Albayrak *et al*, Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell*, 61:914–924, 2016

Remarks

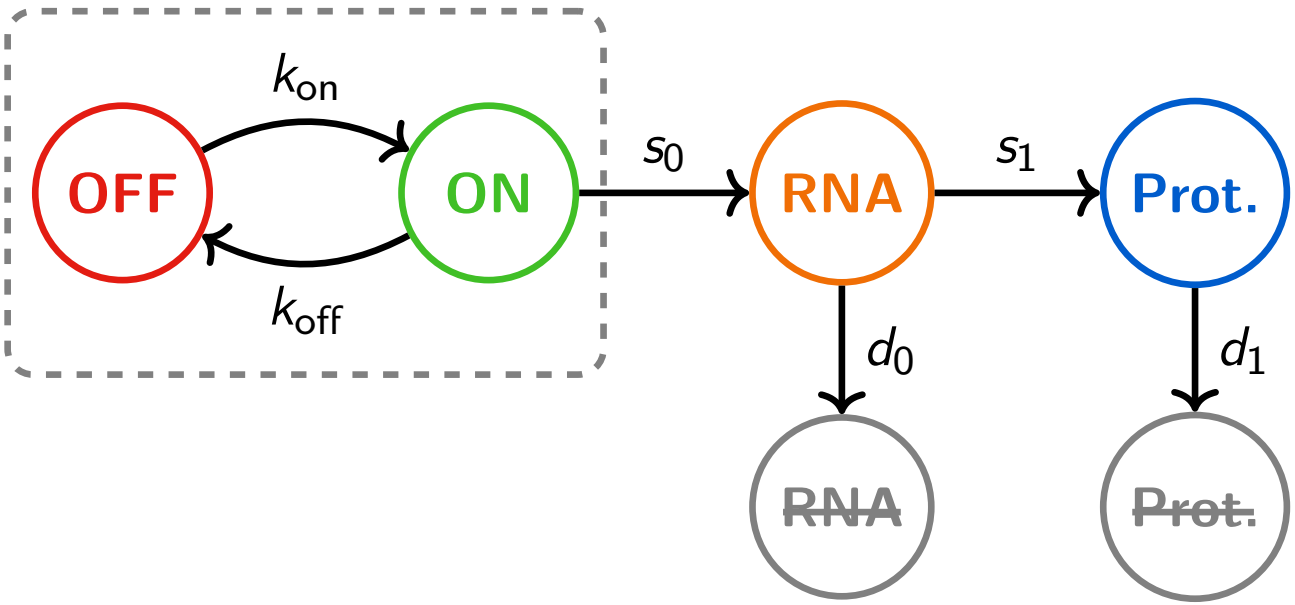
- Typical distributions are **not Poisson** but rather **Gamma**
- Sometimes they appear as **mixtures** of Gamma distributions
- For this gene: $\langle \mathrm{mRNA} \rangle \approx 10^2$ and $\langle \mathrm{Protein} \rangle \approx 10^5$ copies/cell

2. Mechanistic network model

Building block for gene networks



Building block for gene networks



Remarks

- This model is **simple enough** to tackle it mathematically
- Rates k_{on}, k_{off} represent a set of many **underlying reactions**
- Can **reproduce data** when set in "bursty" regime ($k_{off} \gg k_{on}$)

Keeping only the most important noise

The only rare species is the **promoter state** $E(t) = 0$ or 1.

$$E(t) : 0 \xrightarrow{k_{\text{on}}} 1, \quad 1 \xrightarrow{k_{\text{off}}} 0$$

$$M'(t) = s_0 E(t) - d_0 M(t)$$

$$P'(t) = s_1 M(t) - d_1 P(t)$$

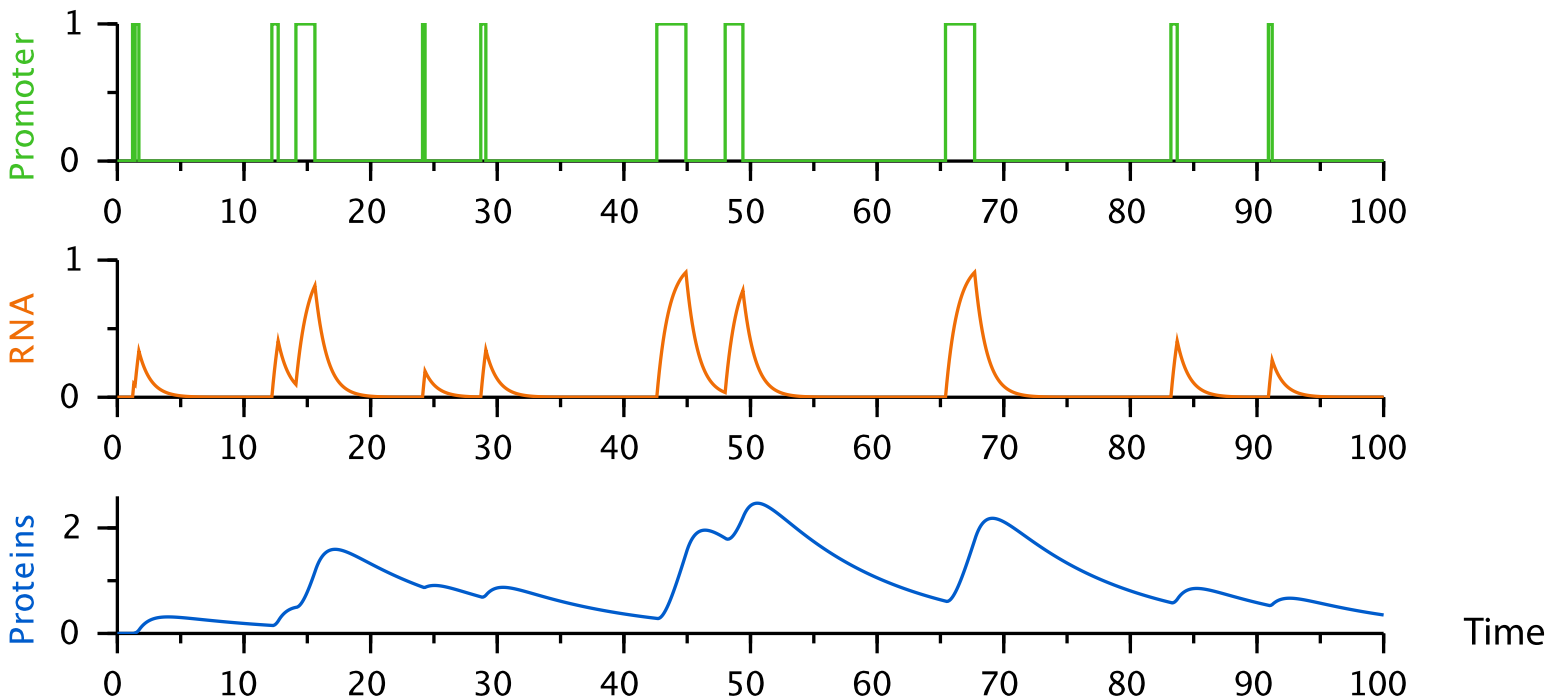
Keeping only the most important noise

The only rare species is the **promoter state** $E(t) = 0$ or 1.

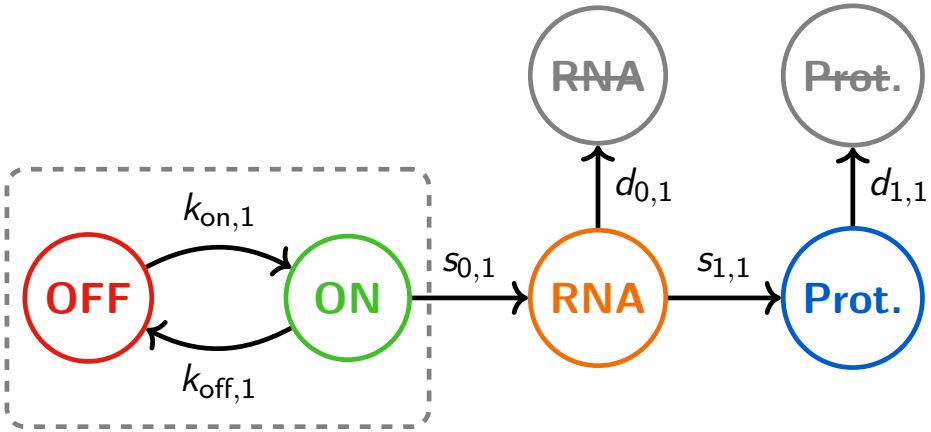
$$E(t) : 0 \xrightarrow{k_{\text{on}}} 1, \quad 1 \xrightarrow{k_{\text{off}}} 0$$

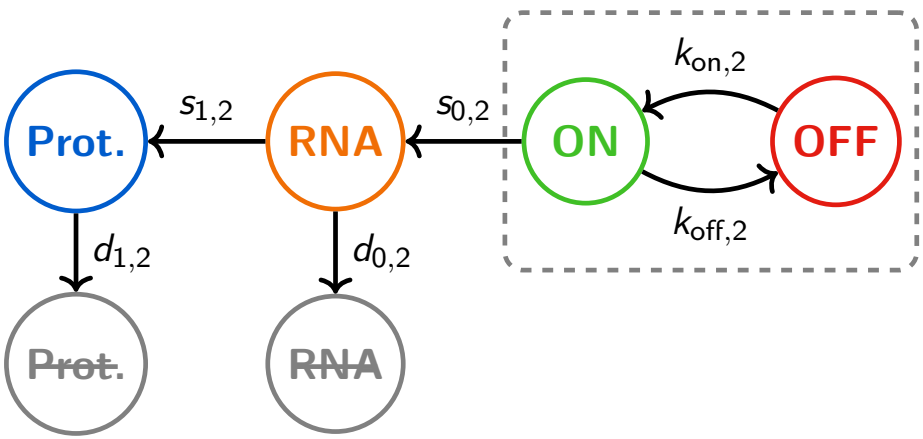
$$M'(t) = s_0 E(t) - d_0 M(t)$$

$$P'(t) = s_1 M(t) - d_1 P(t)$$

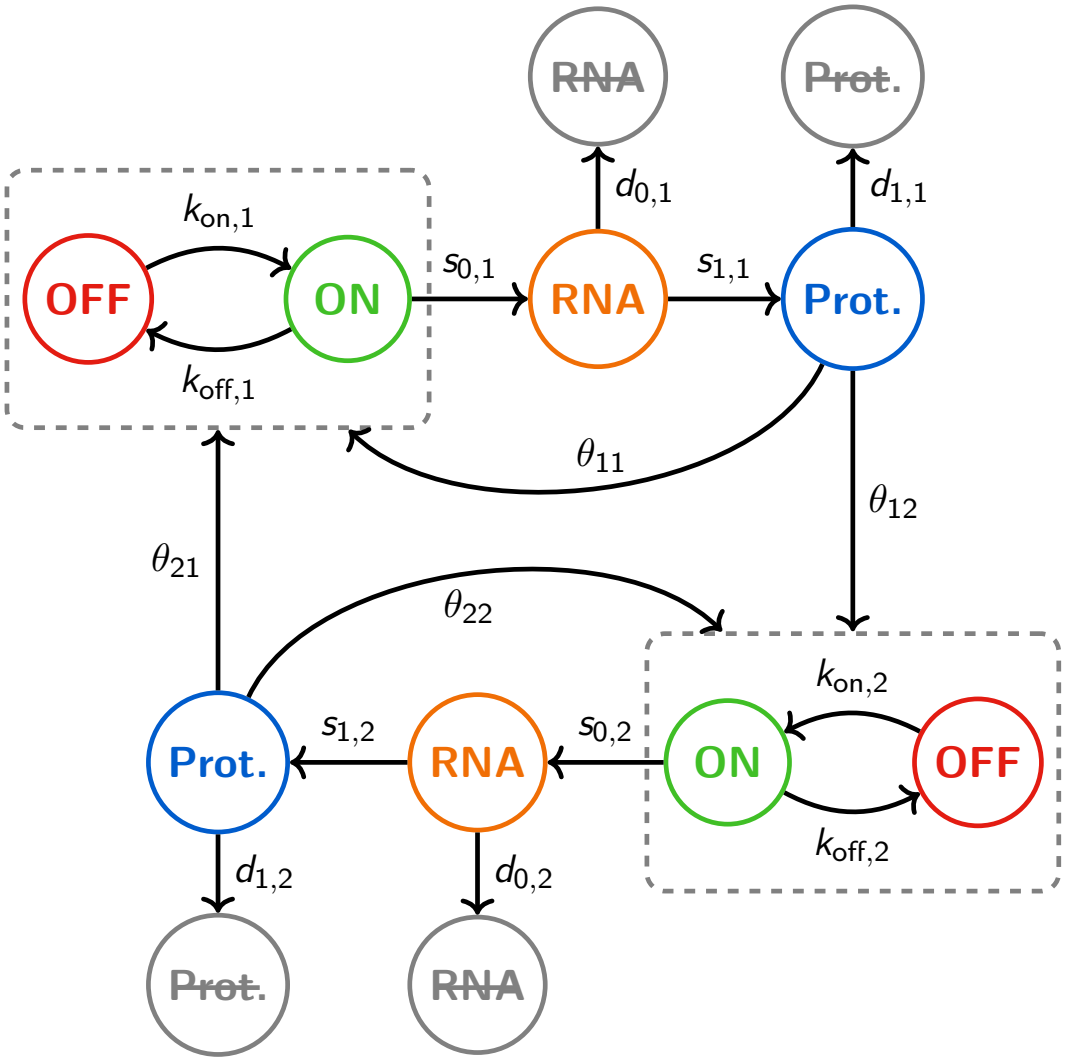


Interacting genes



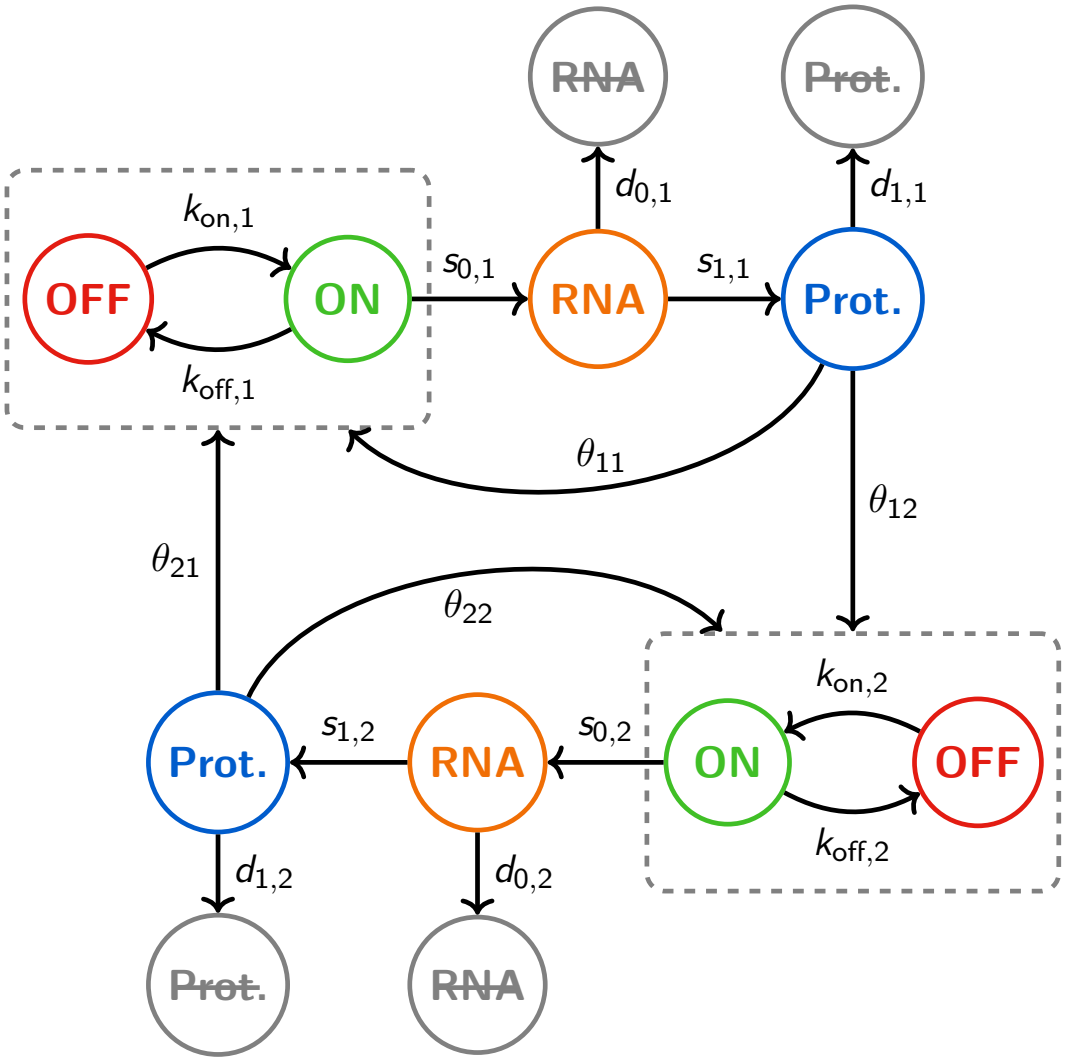


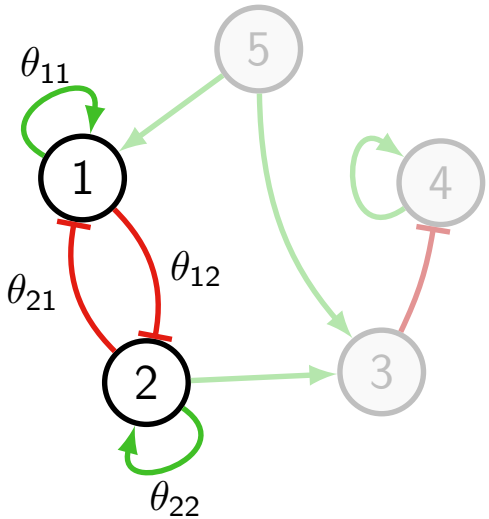
Interacting genes



Principle: $k_{\text{on},i}$, $k_{\text{off},i}$ functions of proteins parameterized by $\theta = (\theta_{ij})_{1 \leqslant i,j \leqslant n}$

Interacting genes





Principle: $k_{on,i}$, $k_{off,i}$ functions of proteins parameterized by $\theta = (\theta_{ij})_{1 \leqslant i,j \leqslant n}$

Network model (dimensionless version)

We note n the number of genes in the network and write:

- $\mathbf{E} = (E_1, \ldots, E_n) \in \{0,1\}^n$ (promoters)
- $\mathbf{M} = (M_1, \ldots, M_n) \in [0,1]^n$ (mRNA)
- $\mathbf{P} = (P_1, \ldots, P_n) \in [0,1]^n$ (proteins)

We then consider the process $(\mathbf{E}(t), \mathbf{M}(t), \mathbf{P}(t))_{t \geqslant 0}$ defined by:

$$\forall i \in [\![1, n]\!], \quad \begin{cases} E_i(t) : 0 \xrightarrow{k_{\text{on},i}(\mathbf{P}(t))} 1, \quad 1 \xrightarrow{k_{\text{off},i}(\mathbf{P}(t))} 0 \\ M_i{}'(t) = d_{0,i}(E_i(t) - M_i(t)) \\ P_i{}'(t) = d_{1,i}(M_i(t) - P_i(t)) \end{cases}$$

Some known results

Theorem (Benaïm, Le Borgne, Malrieu and Zitt, 2015)

Suppose that the functions $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ are continuous and > 0 on $[0,1]^n$. Then $(\mathbf{E}(t), \mathbf{M}(t), \mathbf{P}(t))_{t \geqslant 0}$ is an **ergodic PDMP**.

Some known results

Theorem (Benaïm, Le Borgne, Malrieu and Zitt, 2015)

Suppose that the functions $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ are continuous and > 0 on $[0,1]^n$. Then $(\mathbf{E}(t), \mathbf{M}(t), \mathbf{P}(t))_{t \geqslant 0}$ is an **ergodic PDMP**.

When promoters and mRNA are faster than proteins:

Deterministic limit (Faggionato, Gabrielli and Crivellari, 2010)

At the limit $[d_{1,i} / \min(d_{0,i}, k_{\mathrm{on},i}, k_{\mathrm{off},i})] \to 0$, proteins follow

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \Phi(\mathbf{P}) \quad \text{where} \quad \Phi_i(\mathbf{P}) = d_{1,i} \left(\frac{k_{\mathrm{on},i}(\mathbf{P})}{k_{\mathrm{on},i}(\mathbf{P}) + k_{\mathrm{off},i}(\mathbf{P})} - P_i \right).$$

There is also a diffusion limit (Pakdaman, Thieullen and Wainrib, 2012)

Some known results

Theorem (Benaïm, Le Borgne, Malrieu and Zitt, 2015)

Suppose that the functions $k_{\mathrm{on},i}$ and $k_{\mathrm{off},i}$ are continuous and > 0 on $[0,1]^n$. Then $(\mathbf{E}(t), \mathbf{M}(t), \mathbf{P}(t))_{t\geqslant 0}$ is an **ergodic PDMP**.

When promoters and mRNA are faster than proteins:

Deterministic limit (Faggionato, Gabrielli and Crivellari, 2010)

At the limit $[d_{1,i}/\min(d_{0,i}, k_{\mathrm{on},i}, k_{\mathrm{off},i})] \to 0$, proteins follow

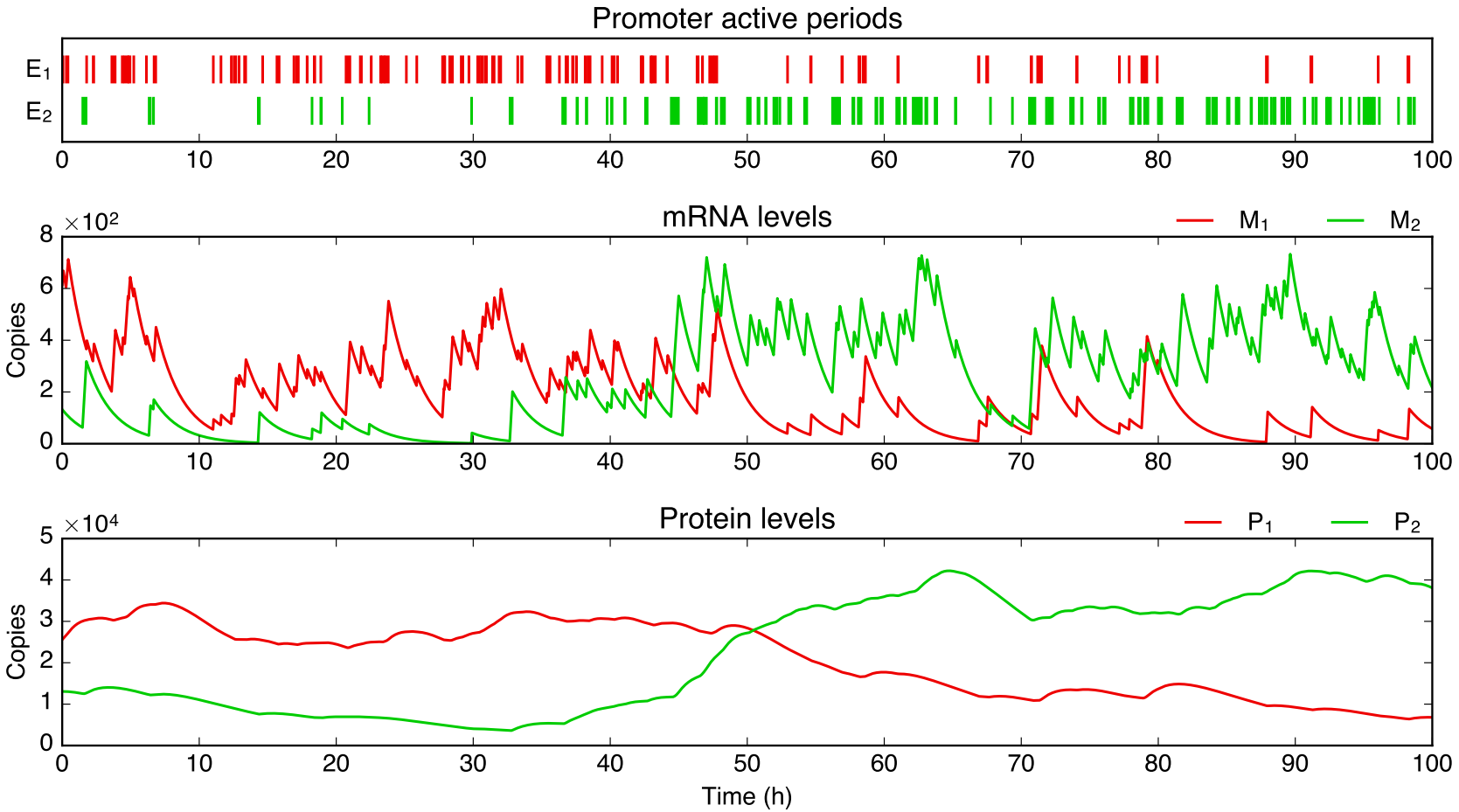
$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \Phi(\mathbf{P}) \quad \text{where} \quad \Phi_i(\mathbf{P}) = d_{1,i}\left(\frac{k_{\mathrm{on},i}(\mathbf{P})}{k_{\mathrm{on},i}(\mathbf{P}) + k_{\mathrm{off},i}(\mathbf{P})} - P_i\right).$$

There is also a diffusion limit (Pakdaman, Thieullen and Wainrib, 2012)

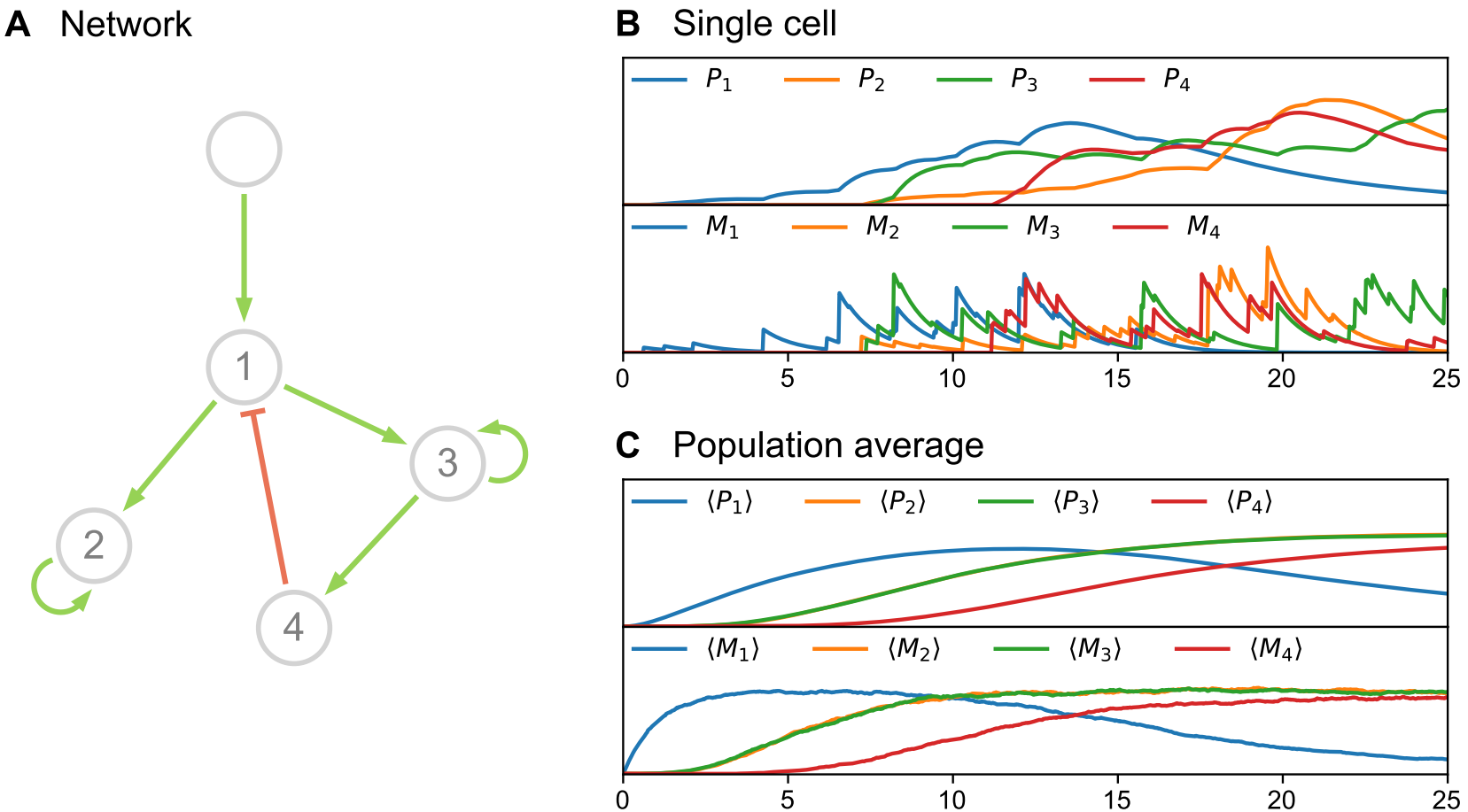
Idea: *we place ourselves in the case* $d_{1,i} \ll \min(d_{0,i}, k_{\mathrm{on},i}, k_{\mathrm{off},i})$, *but without passing directly to the limit* (in practice $d_{1,i}/d_{0,i} \approx 0.2$).

Example 1: two-gene "toggle switch"

$$\theta = \begin{pmatrix} + & - \\ - & + \end{pmatrix}$$



Example 2: four genes with stimulus



Two crucial simplifications

1. When $d_{1,i} \ll d_{0,i}$, we have an "intermediate" simplification:

$$\forall i \in [\![1, n]\!], \quad \begin{cases} E_i(t) : 0 \xrightarrow{k_{\text{on},i}(\mathbf{P}(t))} 1, \quad 1 \xrightarrow{k_{\text{off},i}(\mathbf{P}(t))} 0 \\ {P_i}'(t) = d_{1,i}(E_i(t) - P_i(t)) \end{cases}$$

2. We can then define (when $k_{\text{on},i} \ll k_{\text{off},i}$) :

$$\mathcal{L}(\mathbf{M}|\mathbf{P}) = \bigotimes_{i=1}^{n} \text{Gamma}\left(\frac{k_{\text{on},i}(\mathbf{P})}{d_{0,i}}, \frac{k_{\text{off},i}(\mathbf{P})}{d_{0,i}}\right)$$

Two crucial simplifications

1. When $d_{1,i} \ll d_{0,i}$, we have an "intermediate" simplification:

$$\forall i \in [\![1, n]\!], \quad \begin{cases} E_i(t) : 0 \xrightarrow{k_{\text{on},i}(\mathbf{P}(t))} 1, \quad 1 \xrightarrow{k_{\text{off},i}(\mathbf{P}(t))} 0 \\ P_i{}'(t) = d_{1,i}(E_i(t) - P_i(t)) \end{cases}$$

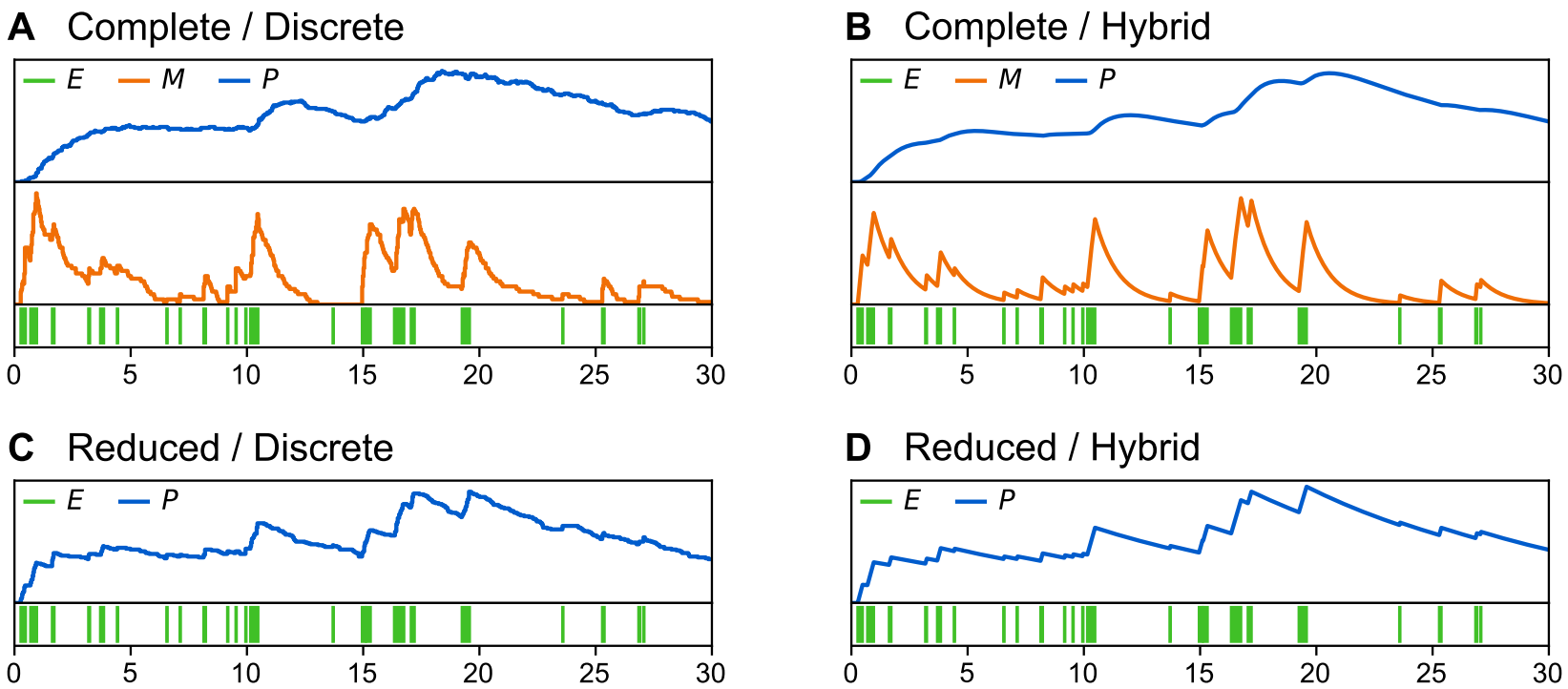
2. We can then define (when $k_{\text{on},i} \ll k_{\text{off},i}$) :

$$\mathcal{L}(\mathbf{M}|\mathbf{P}) = \bigotimes_{i=1}^{n} \text{Gamma}\left(\frac{k_{\text{on},i}(\mathbf{P})}{d_{0,i}}, \frac{k_{\text{off},i}(\mathbf{P})}{d_{0,i}}\right)$$

Remark

The reduced model $(\mathbf{E}(t), \mathbf{P}(t))_{t \geqslant 0}$ is still a PDMP with the same properties (ergodicity, same deterministic limit).

Comparing models



Remark

We shall use:

- Model A or B for **data simulation** ("gold standard")
- Model C or D for **building inference algorithms**

3. Deriving a statistical model

Inference strategy

1. Obtain a simple **analytical approximation** of the stationary distribution $p(\mathbf{x}, \mathbf{y}|\theta)$ of mRNA $\mathbf{x} = (x_i)$ and proteins $\mathbf{y} = (y_i)$
2. Replace $\theta = (\theta_{ij})$ by a **variational parameter** $\alpha(t) = (\alpha_{ij}(t))$
3. Use $p(\mathbf{x}, \mathbf{y}|\alpha(t))$ as a **statistical likelihood** to be maximized

Inference strategy

1. Obtain a simple **analytical approximation** of the stationary distribution $p(\mathbf{x}, \mathbf{y}|\theta)$ of mRNA $\mathbf{x} = (x_i)$ and proteins $\mathbf{y} = (y_i)$
2. Replace $\theta = (\theta_{ij})$ by a **variational parameter** $\alpha(t) = (\alpha_{ij}(t))$
3. Use $p(\mathbf{x}, \mathbf{y}|\alpha(t))$ as a **statistical likelihood** to be maximized

Statistical model for cell k observed at time t_k

$$p(\mathbf{y}_k) = \prod_{i=1}^{n} {y_{ki}}^{c_i\sigma_{ki}-1} e^{-c_i y_{ki}} \frac{{c_i}^{c_i\sigma_{ki}}}{\Gamma(c_i\sigma_{ki})}$$

$$p(\mathbf{x}_k|\mathbf{y}_k) = \prod_{i=1}^{n} {x_{ki}}^{a_i\sigma_{ki}-1} e^{-b_i x_{ki}} \frac{{b_i}^{a_i\sigma_{ki}}}{\Gamma(a_i\sigma_{ki})}$$

Inference strategy

1. Obtain a simple **analytical approximation** of the stationary distribution $p(\mathbf{x}, \mathbf{y}|\theta)$ of mRNA $\mathbf{x} = (x_i)$ and proteins $\mathbf{y} = (y_i)$
2. Replace $\theta = (\theta_{ij})$ by a **variational parameter** $\alpha(t) = (\alpha_{ij}(t))$
3. Use $p(\mathbf{x}, \mathbf{y}|\alpha(t))$ as a **statistical likelihood** to be maximized

Statistical model for cell k observed at time t_k

$$p(\mathbf{y}_k) = \prod_{i=1}^{n} {y_{ki}}^{c_i\sigma_{ki}-1} e^{-c_i y_{ki}} \frac{{c_i}^{c_i\sigma_{ki}}}{\Gamma(c_i\sigma_{ki})}$$

$$p(\mathbf{x}_k|\mathbf{y}_k) = \prod_{i=1}^{n} {x_{ki}}^{a_i\sigma_{ki}-1} e^{-b_i x_{ki}} \frac{{b_i}^{a_i\sigma_{ki}}}{\Gamma(a_i\sigma_{ki})}$$

Interaction function (choice of $k_{\mathrm{on},i}$)

$$\sigma_{ki}(\mathbf{y}_k) = \frac{\exp(\beta_i + \sum_j \alpha_{ji}(t_k) y_{kj})}{1 + \exp(\beta_i + \sum_j \alpha_{ji}(t_k) y_{kj})}$$

Self-consistent field approximation

Aim: approximate the stationary distribution $p(\mathbf{y})$ of $(\mathbf{P}(t))_{t \geqslant 0}$

Hartree approximation (Walczak, Sasai and Wolynes, 2005)

Locally independent promoters but which are subject to a common "proteomic field": in other words $p(\mathbf{y}) \approx h(\mathbf{y})$ with

$$h(\mathbf{y}) = \prod_{i=1}^{n} \frac{{y_i}^{a_i(\mathbf{y})-1}(1-y_i)^{b_i(\mathbf{y})-1}}{\mathrm{B}(a_i(\mathbf{y}), b_i(\mathbf{y}))}$$

where $a_i(\mathbf{y}) = k_{\mathrm{on},i}(\mathbf{y})/d_{1,i}$ and $b_i(\mathbf{y}) = k_{\mathrm{off},i}(\mathbf{y})/d_{1,i}$.

Self-consistent field approximation

Aim: approximate the stationary distribution $p(\mathbf{y})$ of $(\mathbf{P}(t))_{t\geqslant 0}$

Hartree approximation (Walczak, Sasai and Wolynes, 2005)

Locally independent promoters but which are subject to a common "proteomic field": in other words $p(\mathbf{y}) \approx h(\mathbf{y})$ with

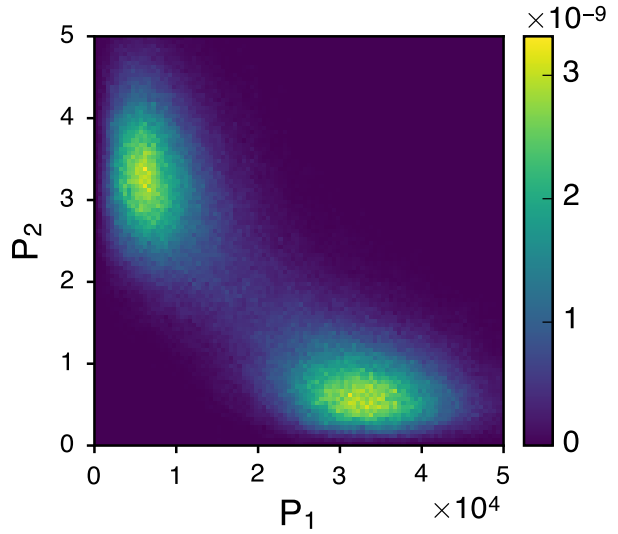
$$h(\mathbf{y}) = \prod_{i=1}^{n} \frac{{y_i}^{a_i(\mathbf{y})-1}(1-y_i)^{b_i(\mathbf{y})-1}}{\mathrm{B}(a_i(\mathbf{y}), b_i(\mathbf{y}))}$$

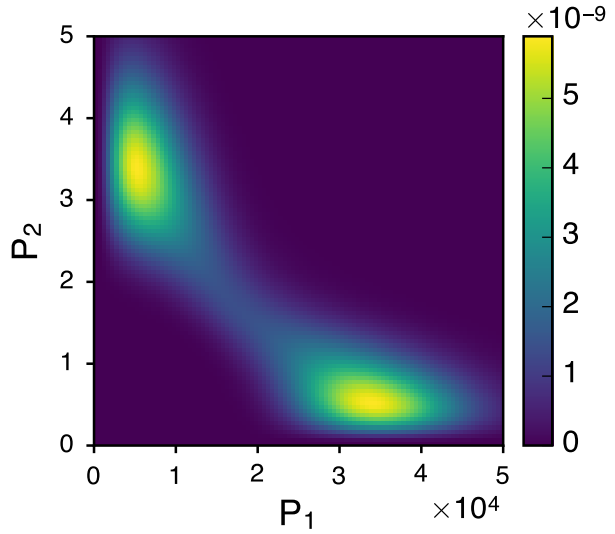
where $a_i(\mathbf{y}) = k_{\mathrm{on},i}(\mathbf{y})/d_{1,i}$ and $b_i(\mathbf{y}) = k_{\mathrm{off},i}(\mathbf{y})/d_{1,i}$.

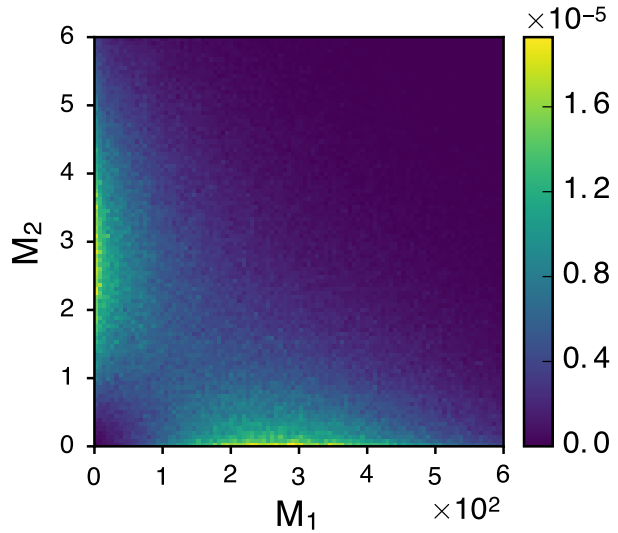
Why it should work: a concentration result

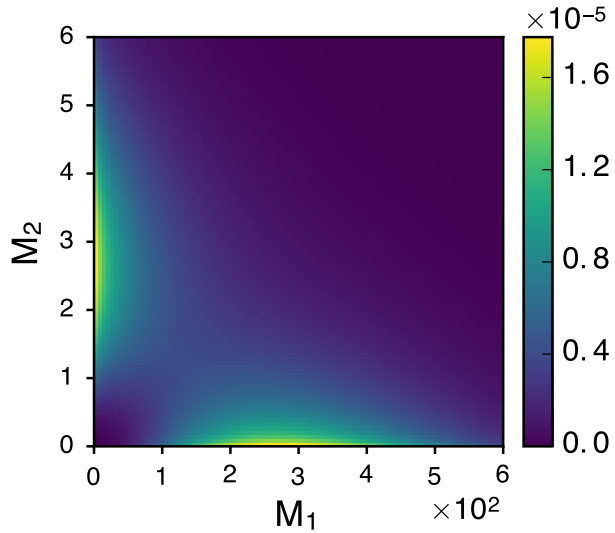
At the limit $d_{1,i} \ll \min(d_{0,i}, k_{\mathrm{on},i}, k_{\mathrm{off},i})$, the function h converges to a sum of Dirac measures $\delta_{\bar{\mathbf{y}}_k}$ where the $\bar{\mathbf{y}}_k$ are exactly the **fixed points of the previous deterministic system** (i.e. $\Phi(\bar{\mathbf{y}}_k) = 0$).

Distributions: exact vs. approximate

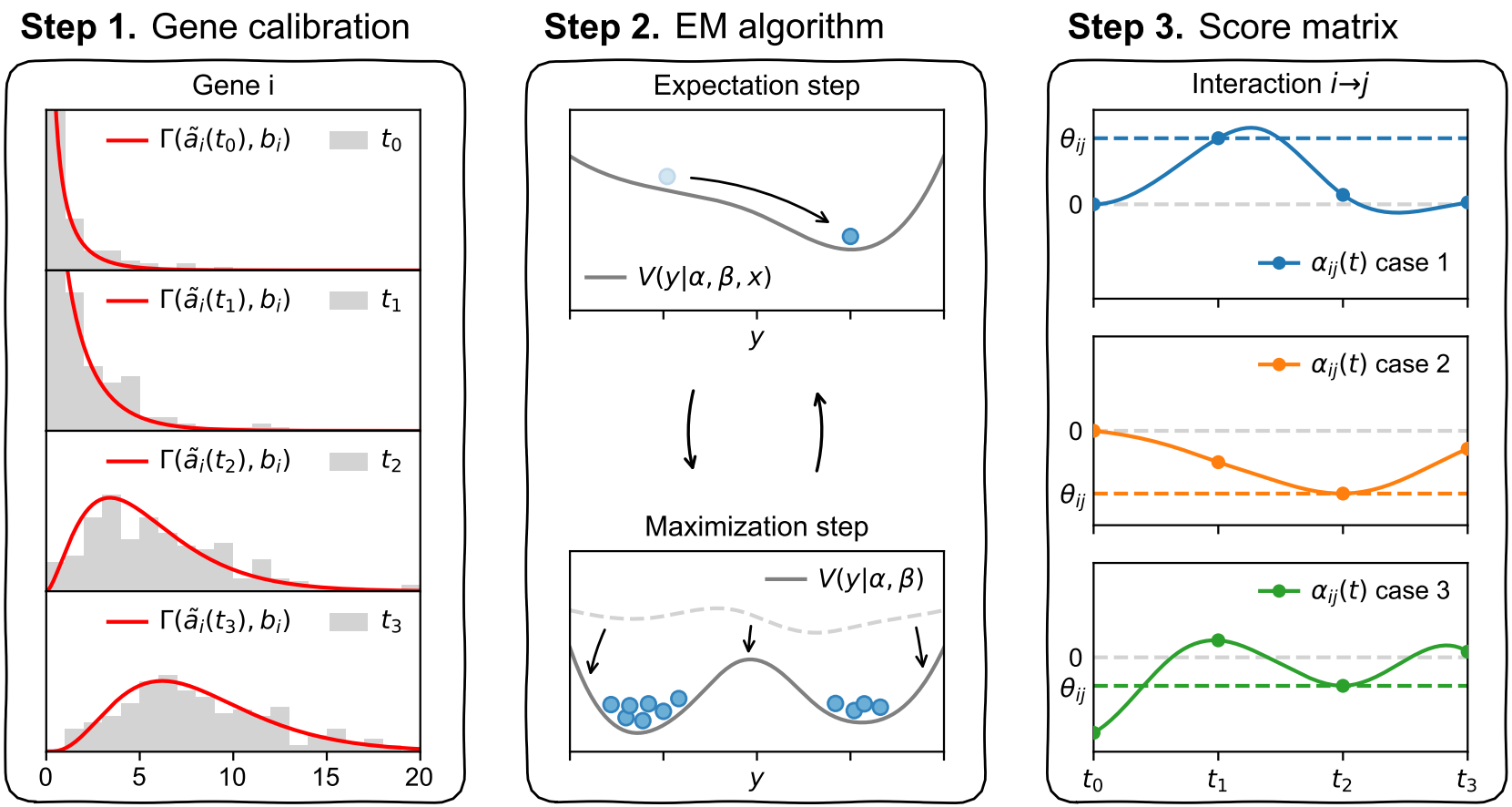




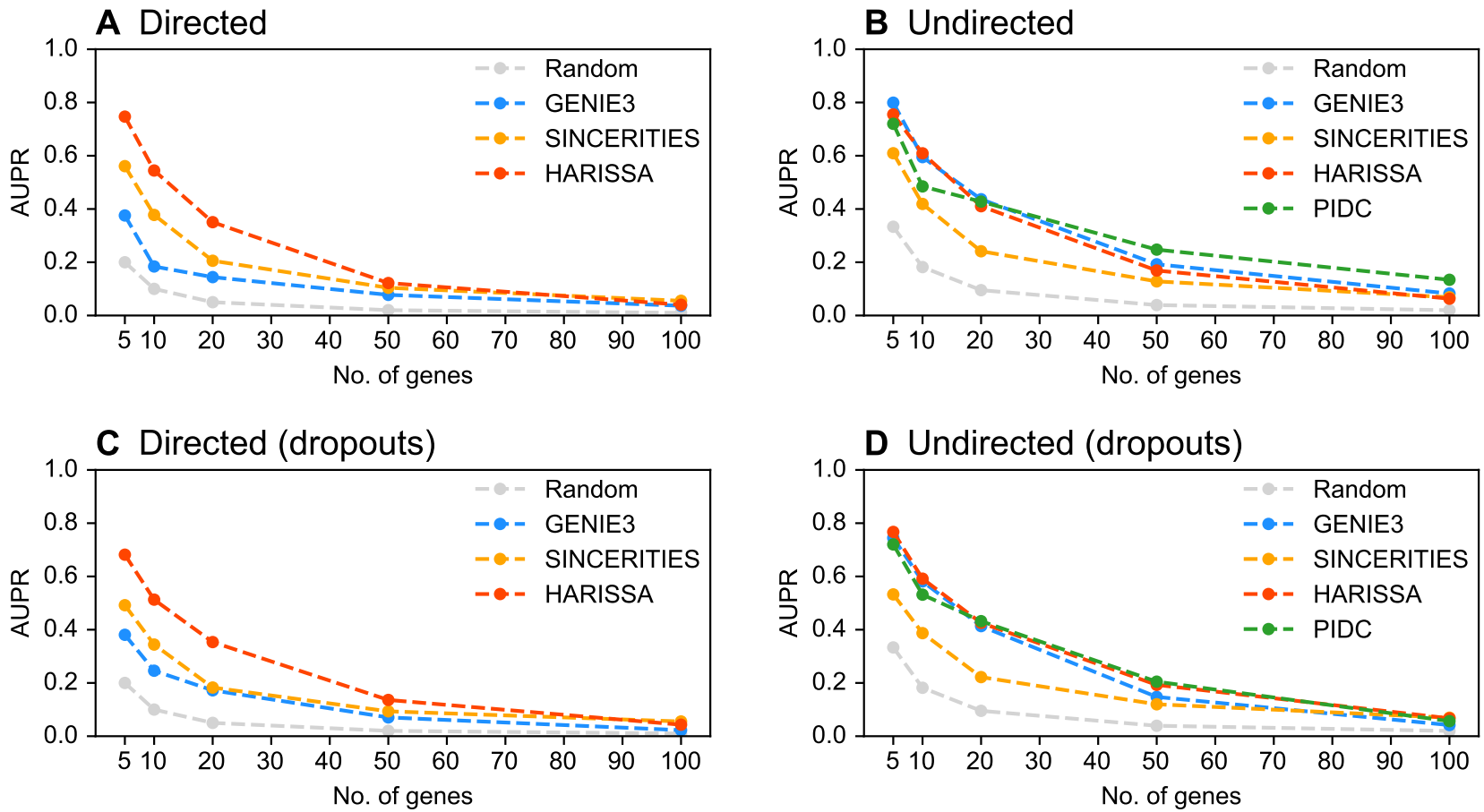




Inference in practice

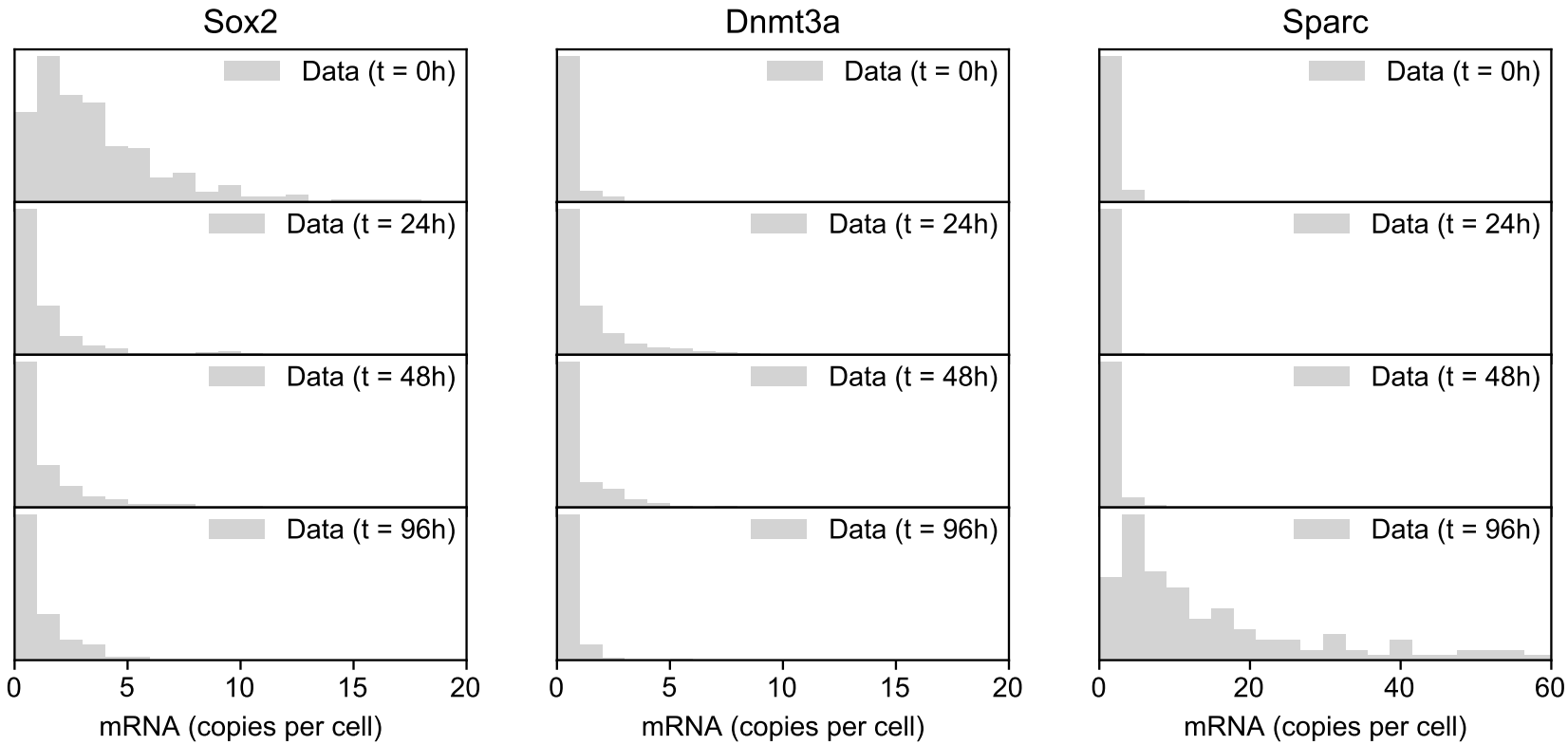


Small benchmark (PDMP network model)



Data: 10 time points with 100 cells per time point (1000 cells sampled per data set)
Networks: random directed trees (uniform distribution) with stimulus and activations

Real data

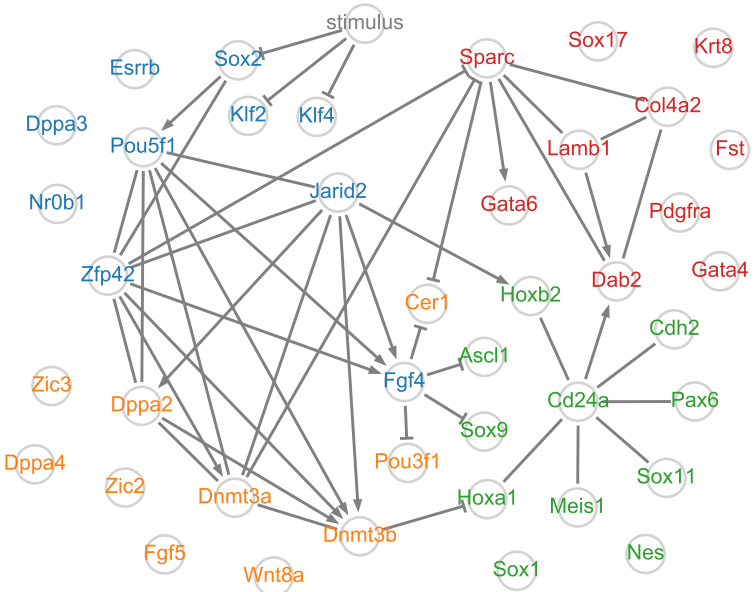


Semrau *et al.*, Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature Communications*, 8(1)2017

Data: 9 time points with 272 cells on average per time point (between 137 and 335)
Inference: particular subset of 41 genes considered in [Semrau *et al.*, 2017]

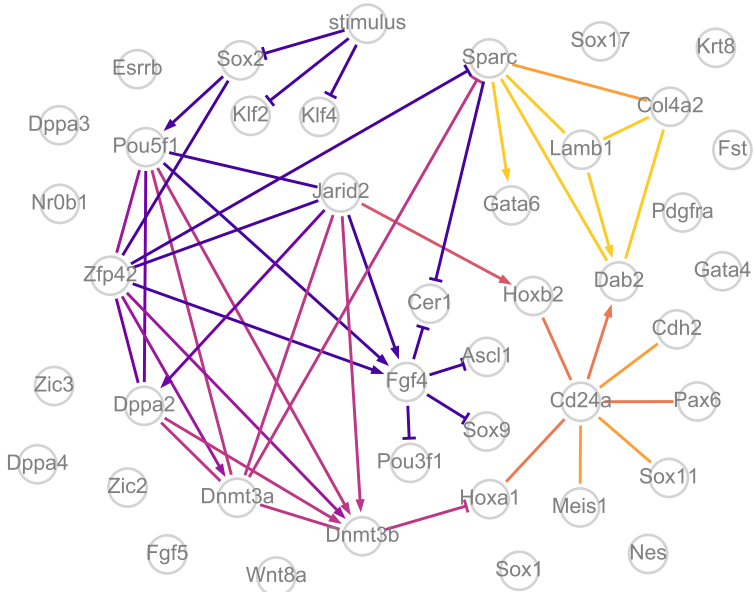
First result: two viewpoints

A Inferred network



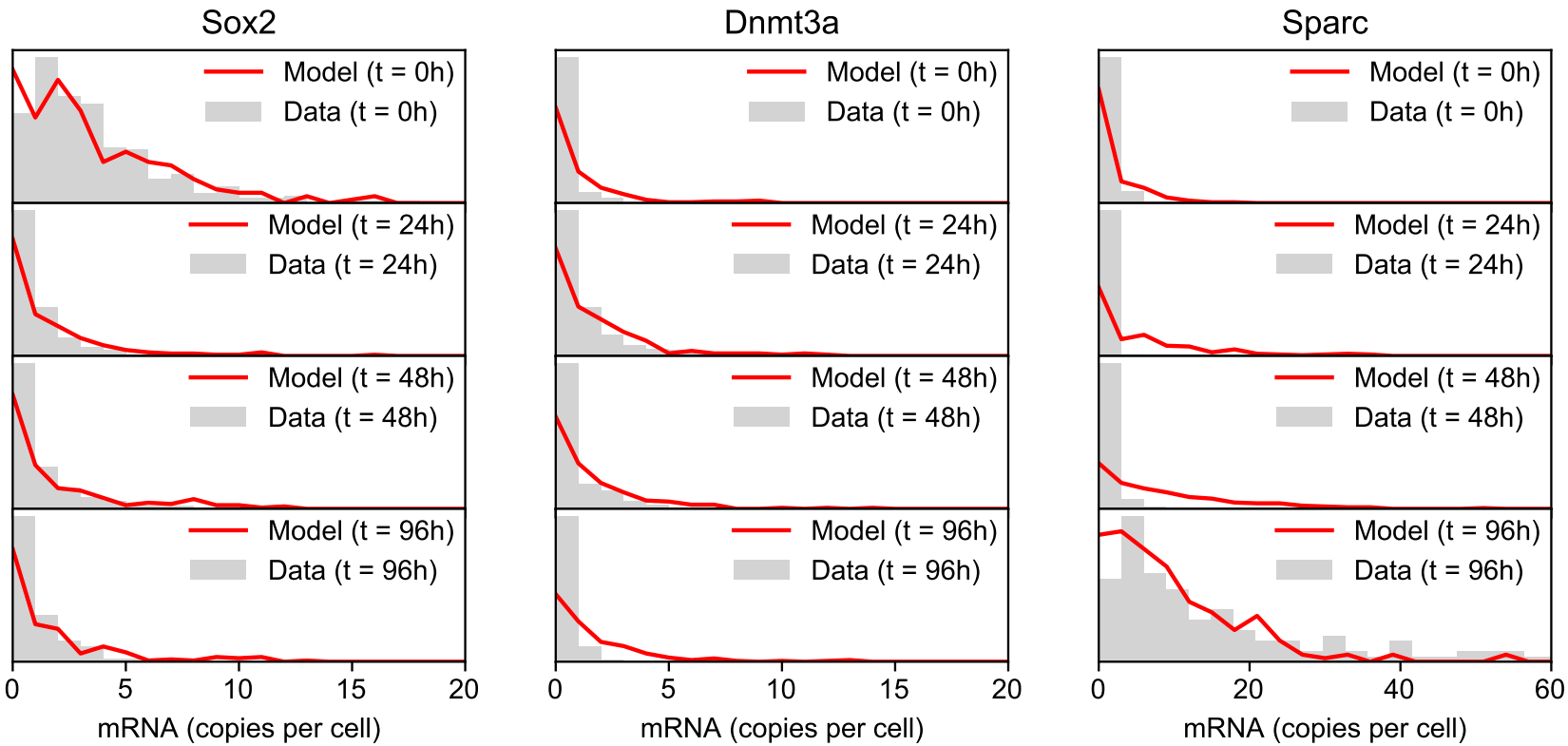
Pluripotency
Extraembryonic endoderm
Post-implantation epiblast
Neuroectoderm

B Time decomposition



6h 12h 24h 36h 48h 60h 72h 96h

Back to the mechanistic model

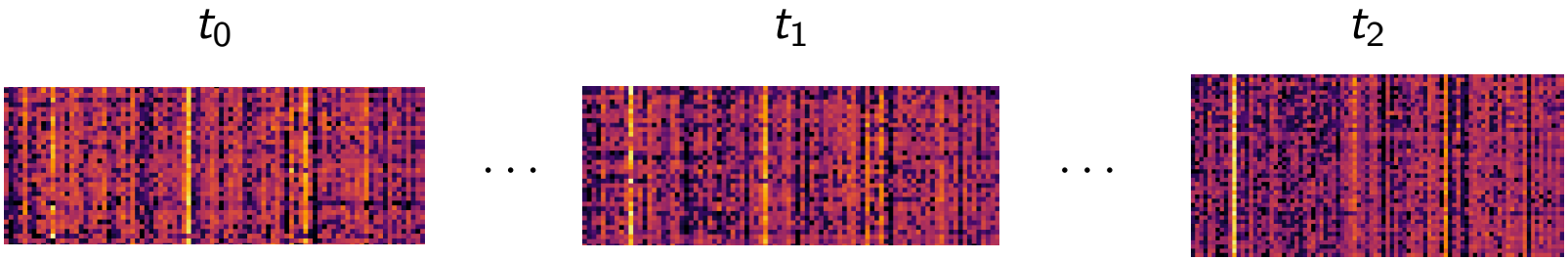


Prospects

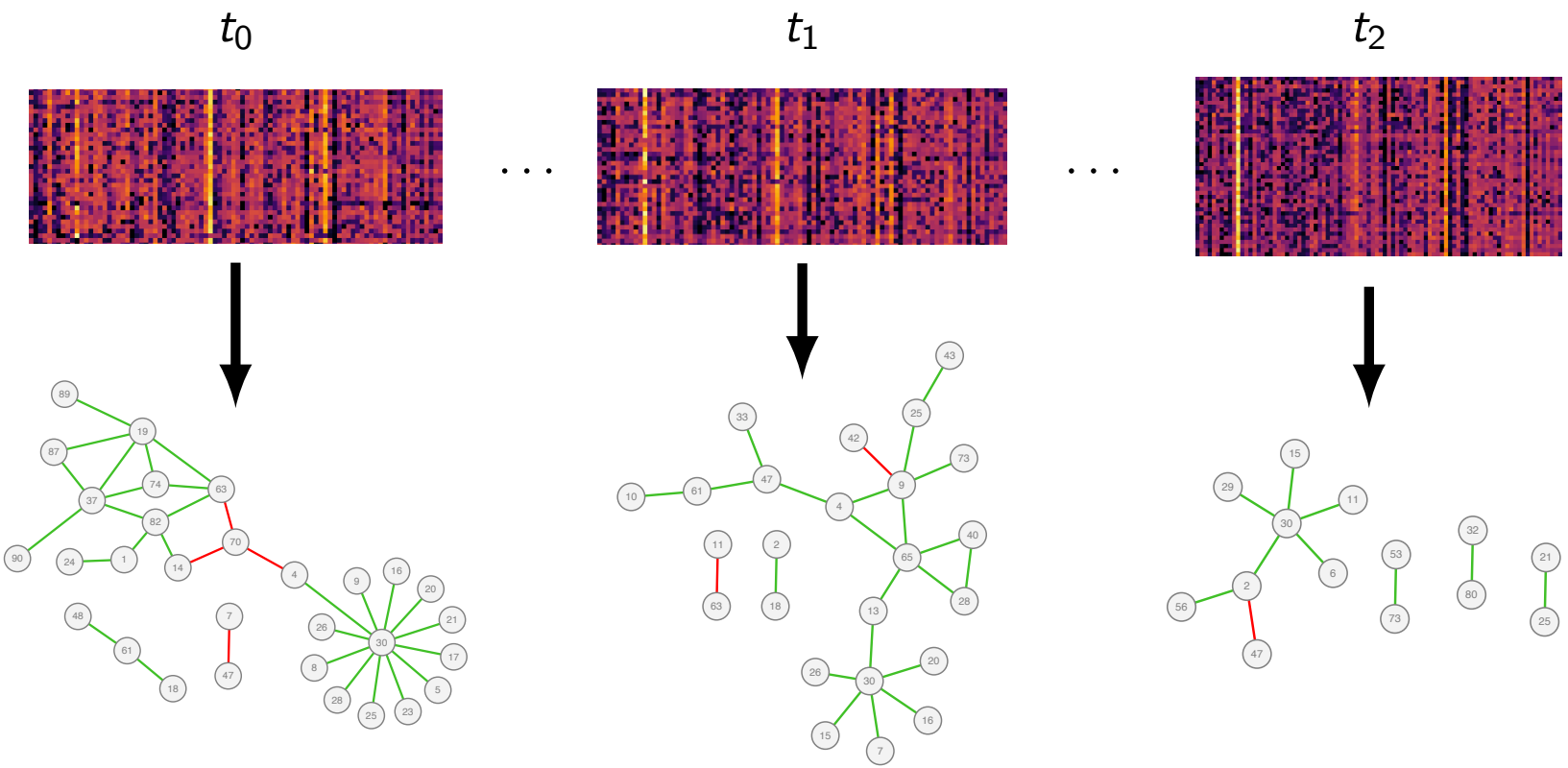
Calibration seems not too bad. Can now make **predictions**...

An open question...

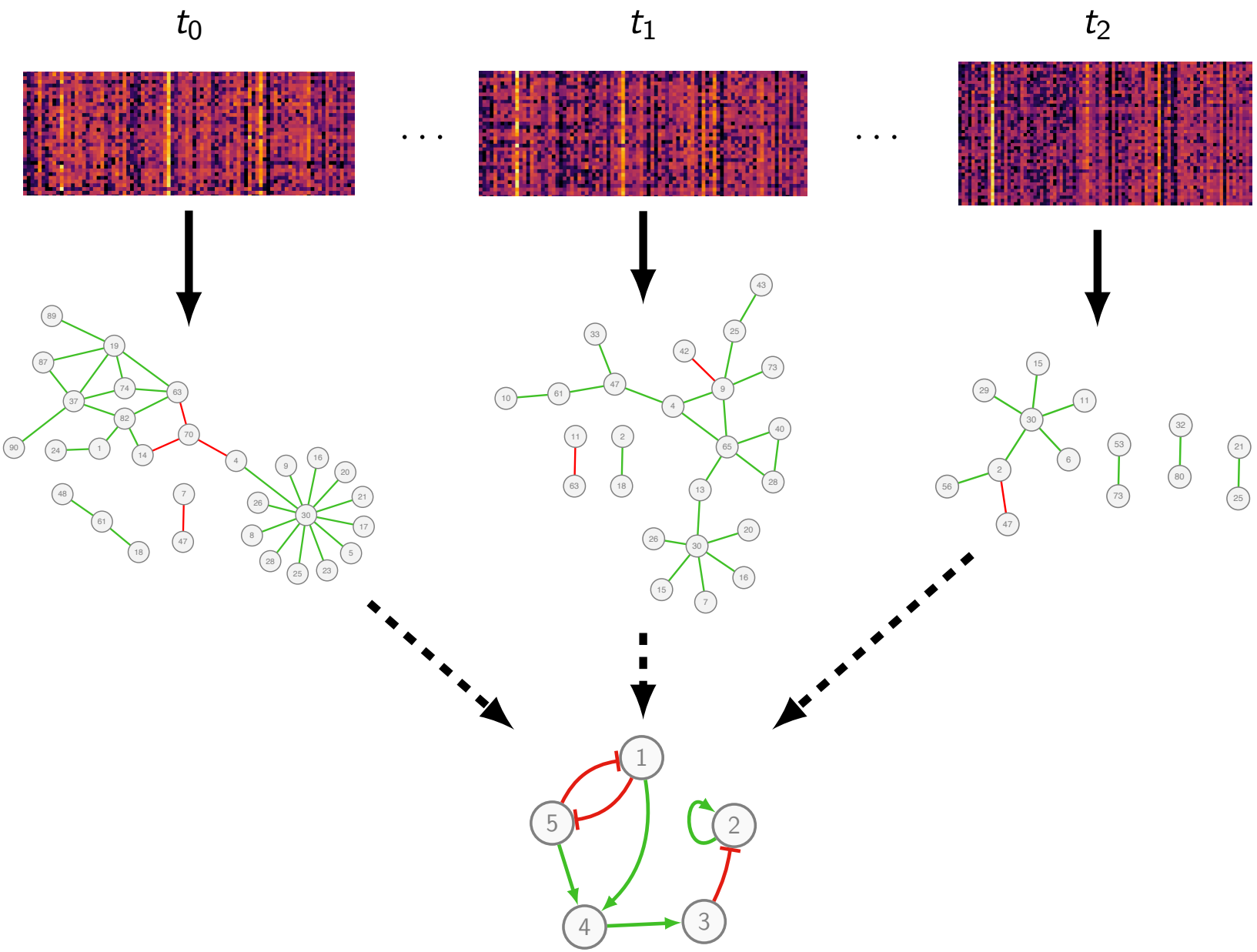
How to optimally exploit the time information?



How to optimally exploit the time information?



How to optimally exploit the time information?



Thank you!

Herbach, U., Bonnaffoux, A., Espinasse, T., and Gandrillon, O. (2017).

Inferring gene regulatory networks from single-cell data: a mechanistic approach.

BMC Systems Biology, 11(1):105.

Bonnaffoux, A., Herbach, U., Richard, A., Guillemin, A., Gonin-Giraud, S., Gros, P.-A., and Gandrillon, O. (2019).

WASABI: a dynamic iterative framework for gene regulatory network inference.

BMC Bioinformatics, 20(1):220.

Ventre, E. (2021).

Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics.

bioRxiv preprint, https://doi.org/10.1101/2021.06.01.446414.