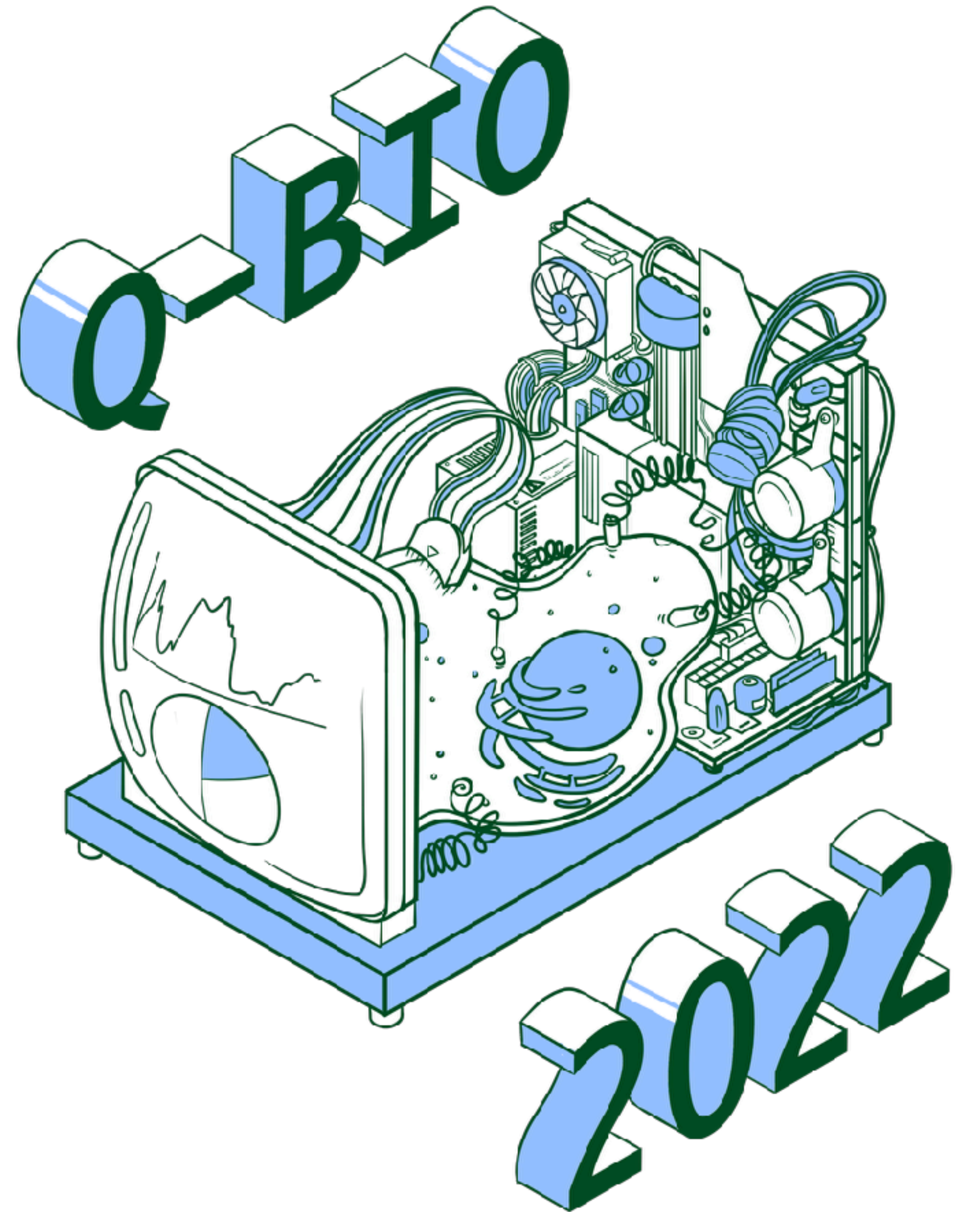


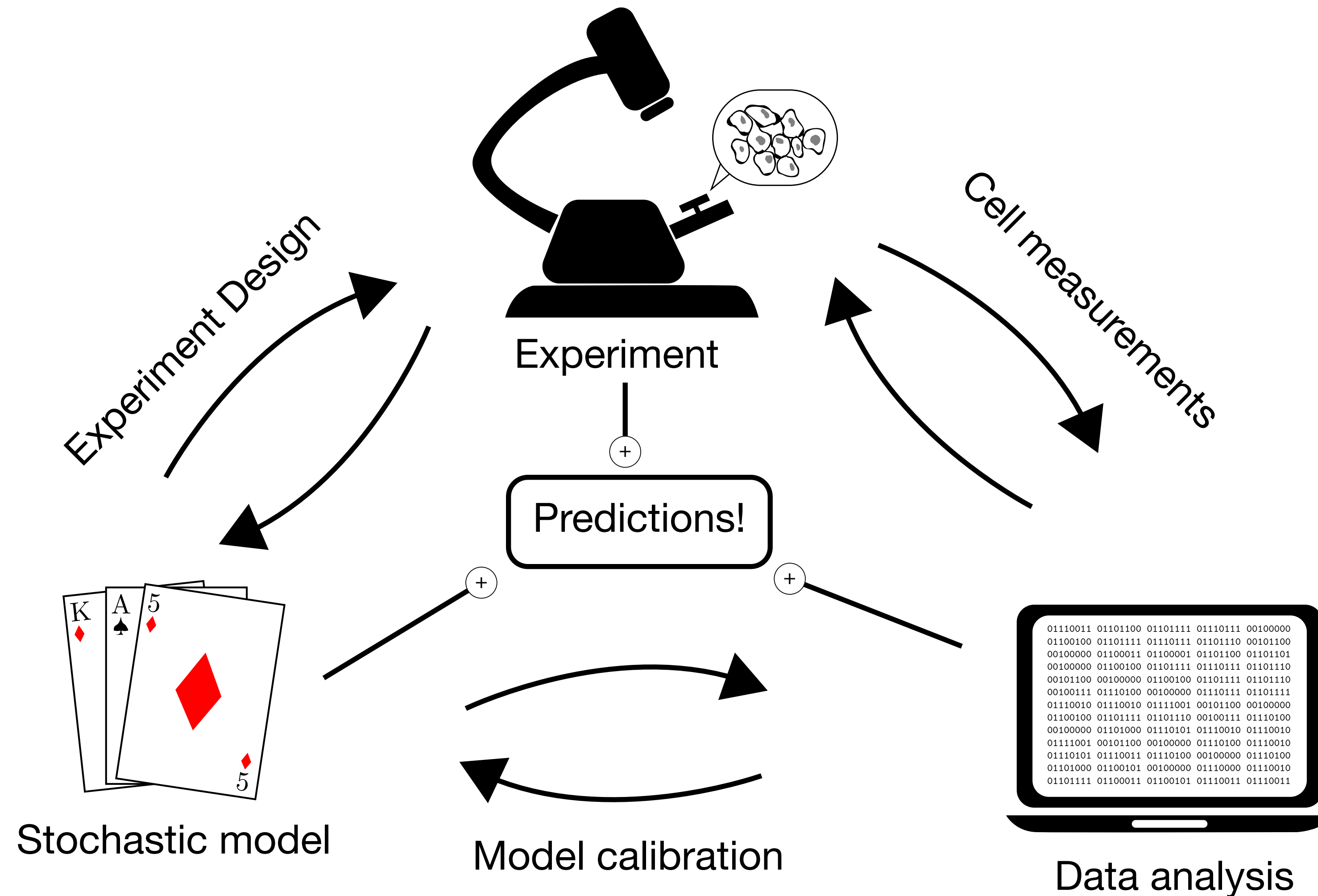
# Using Stochastic Models Single-Cell Data to Reduce Models and Design Experiments

*Zachary R Fox*

*Center for Nonlinear Studies & Information Sciences Group  
Los Alamos National Laboratory*

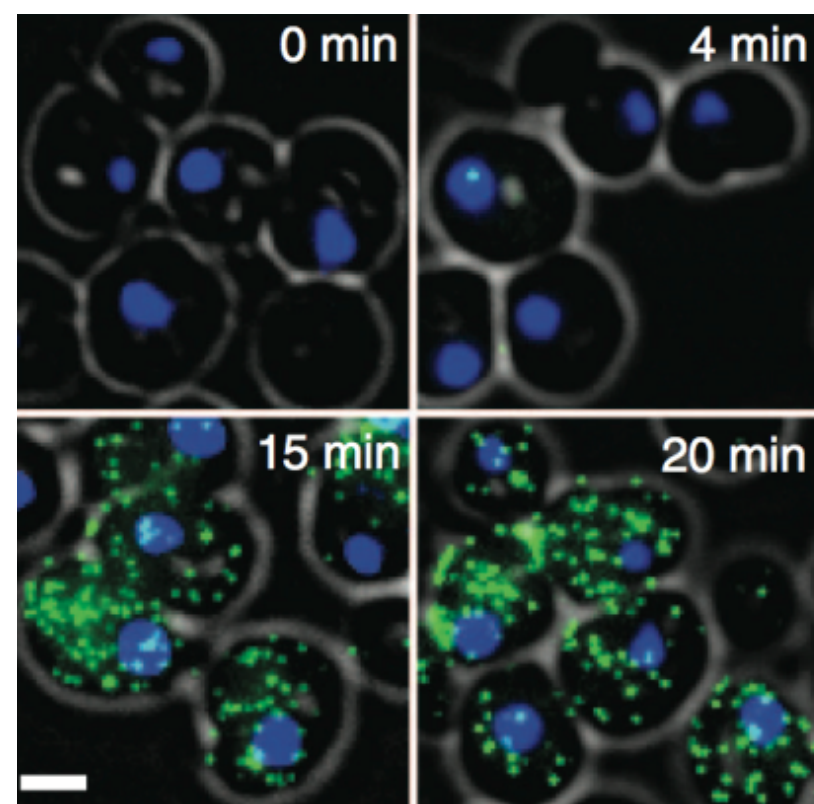


# Quantitative studies of meso-scale biological processes.

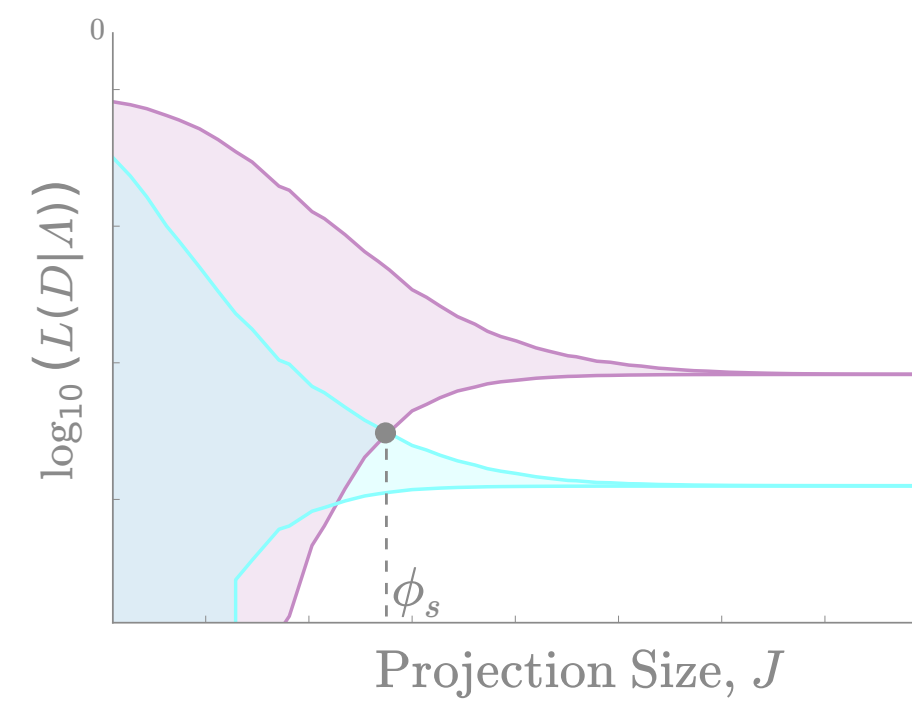


# Outline

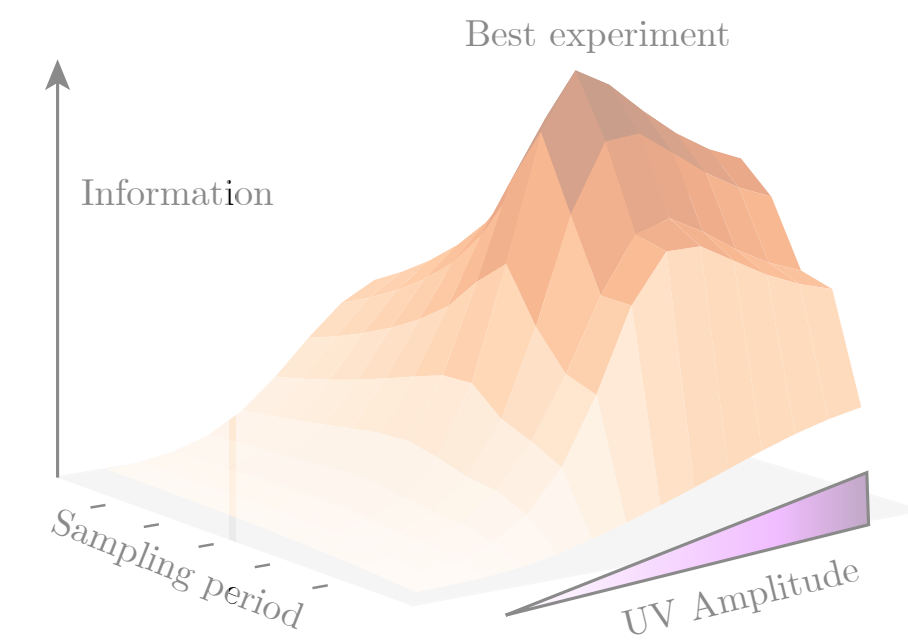
## Variability in biochemical reactions

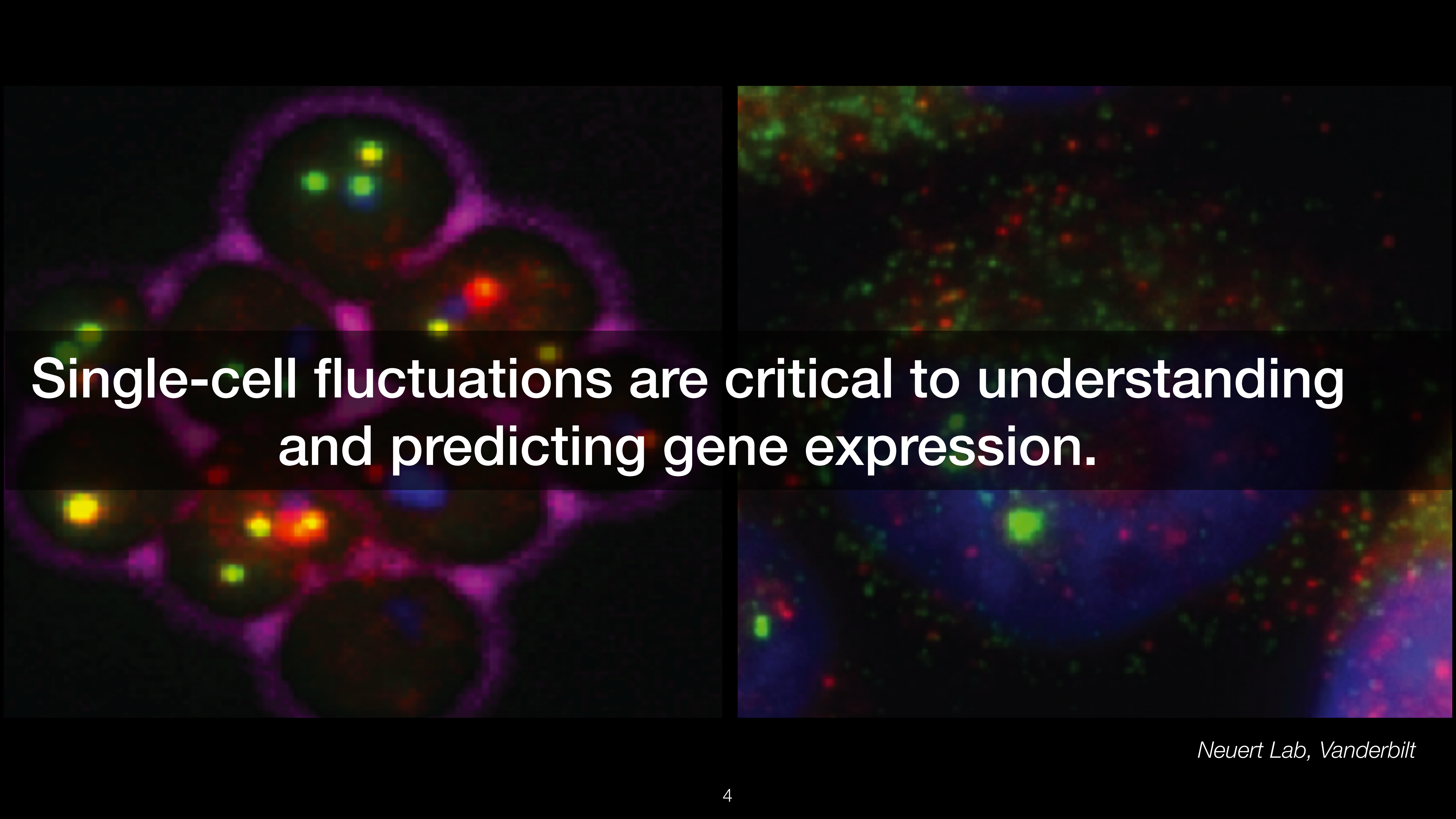


## Efficient model identification using error constraints

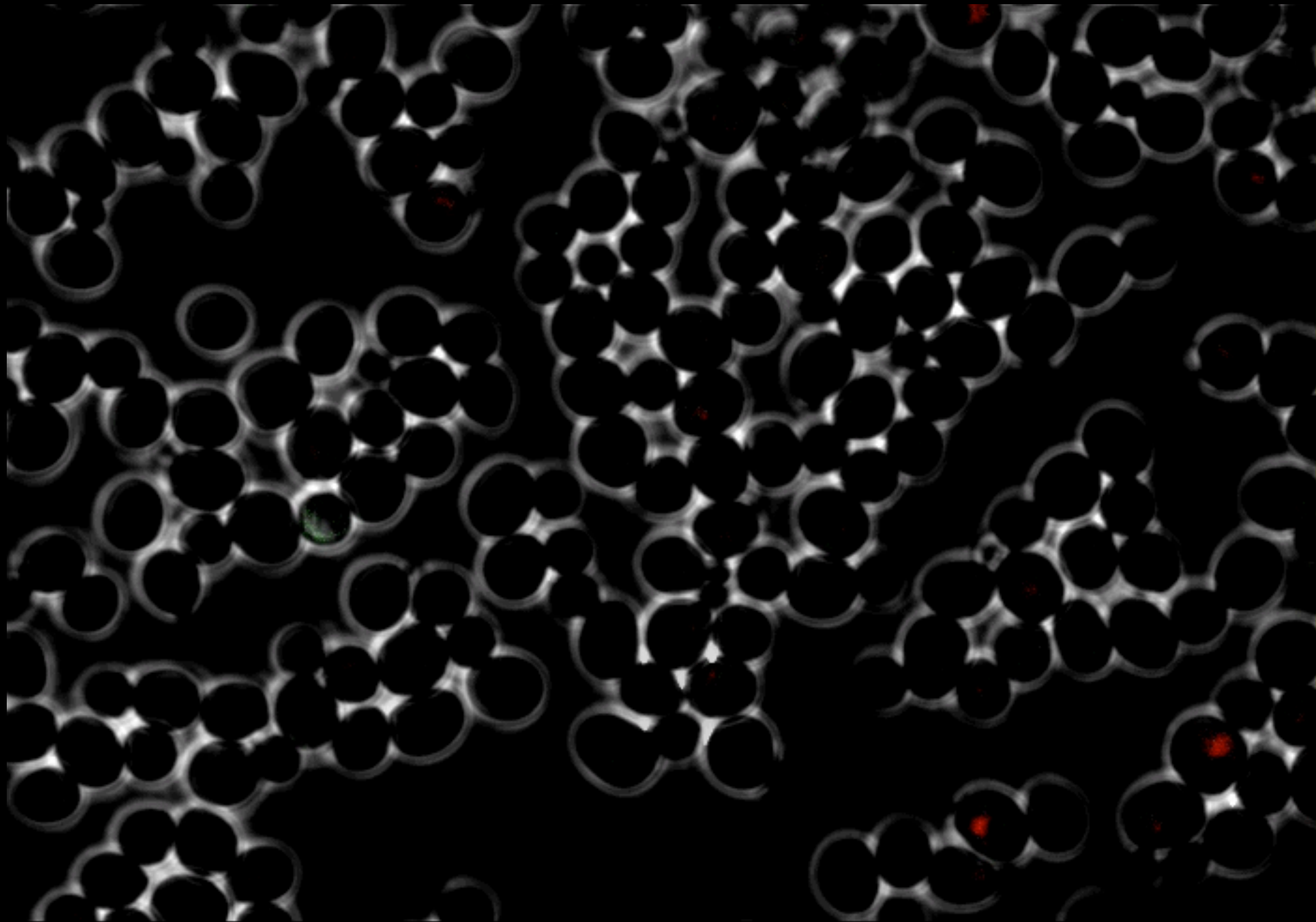


## Designing single-cell experiments with Fisher Information

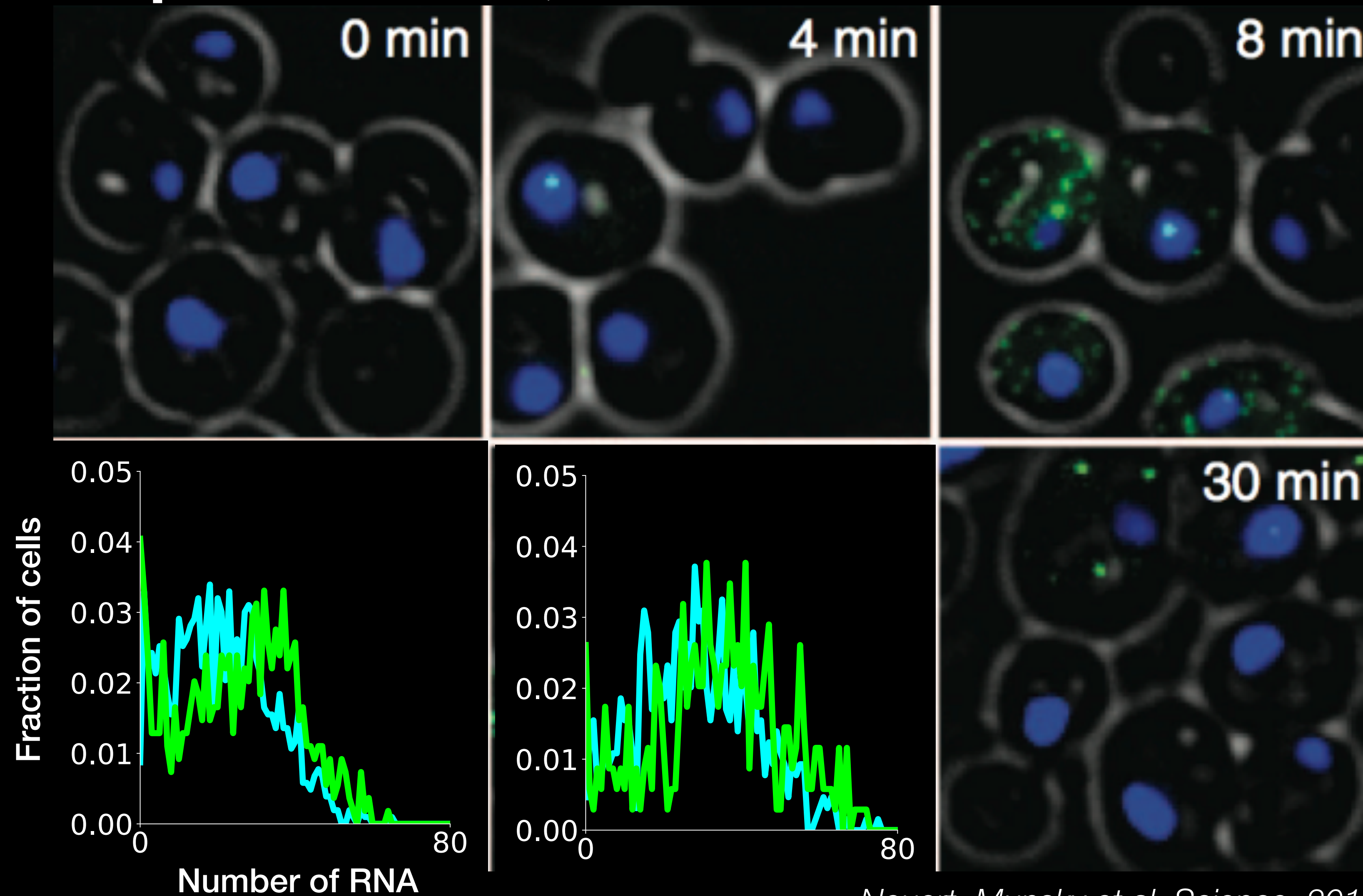




**Single-cell fluctuations are critical to understanding  
and predicting gene expression.**



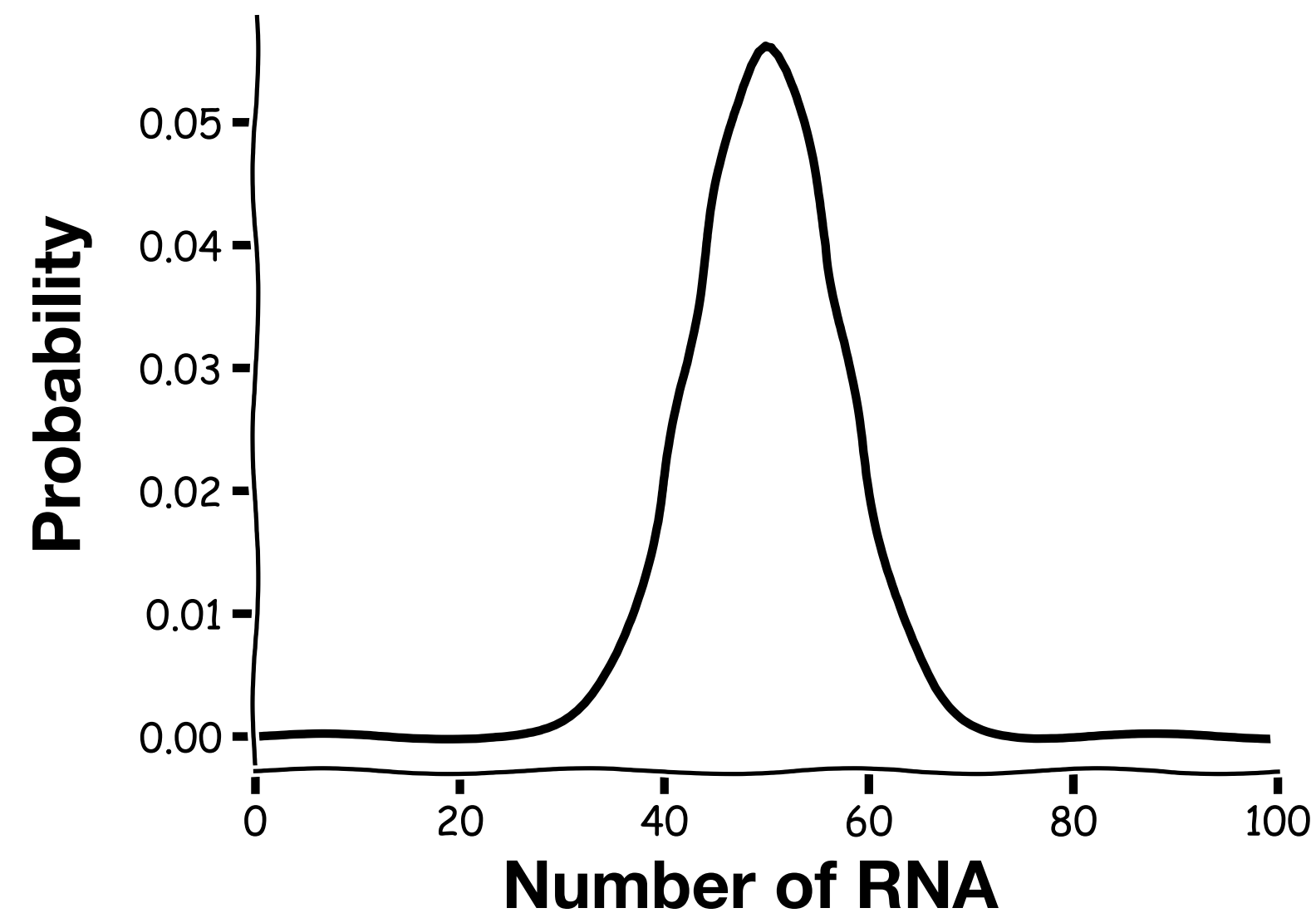
# Measured single-cell distributions are informative, reproducible, and non-Gaussian.



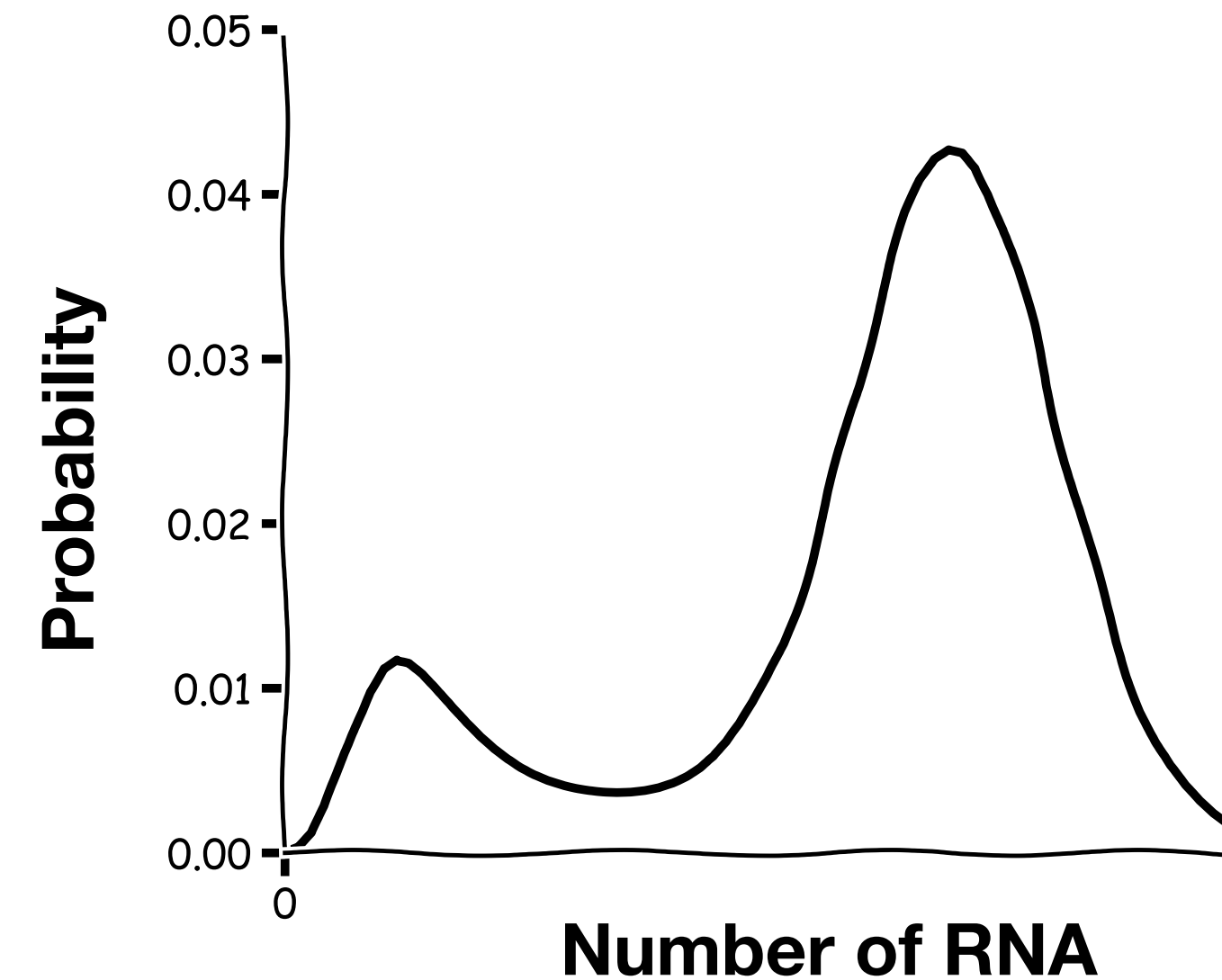
*Neuert, Munsky et al. Science, 2013.*

*Munsky, Li, **Fox**, Shepherd, Neuert. PNAS, 2018.*

# The shape of gene expression distributions is important.

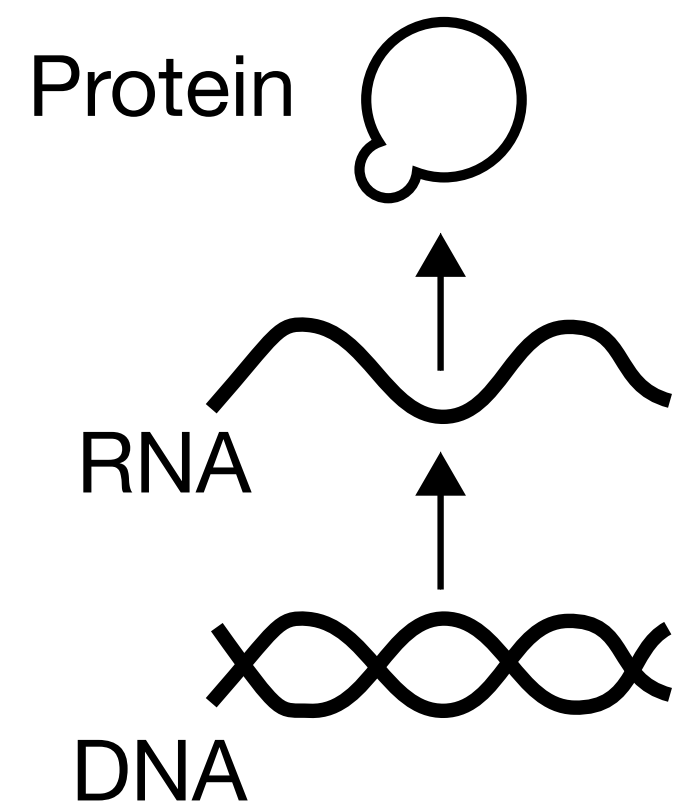


Gaussian distributions are completely characterized by their means and variances...



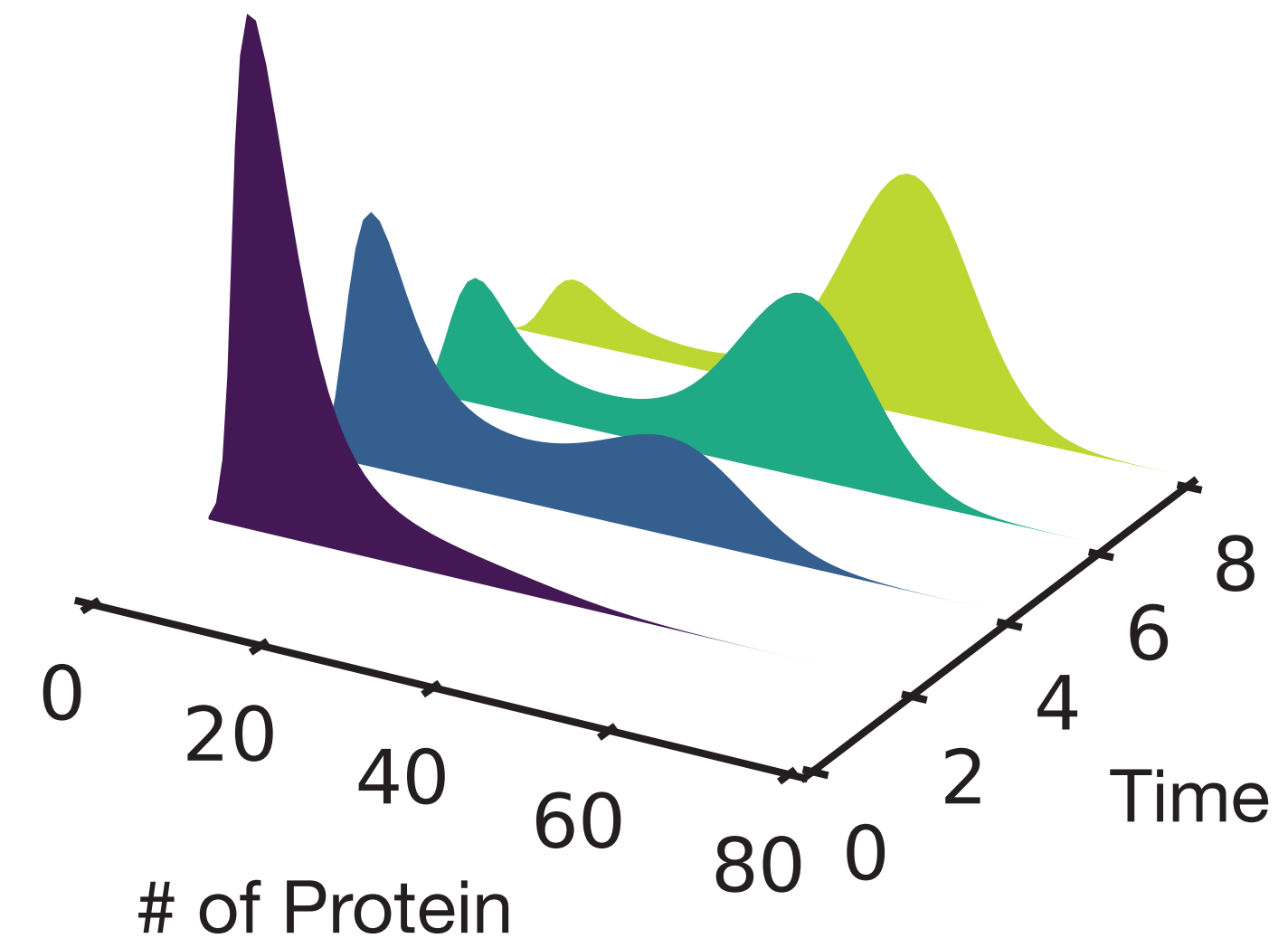
... but more complex distributions require analyses that go beyond the first moments of their distributions.

# Stochastic modeling allows us to fit and predict probability distributions of biomolecules.



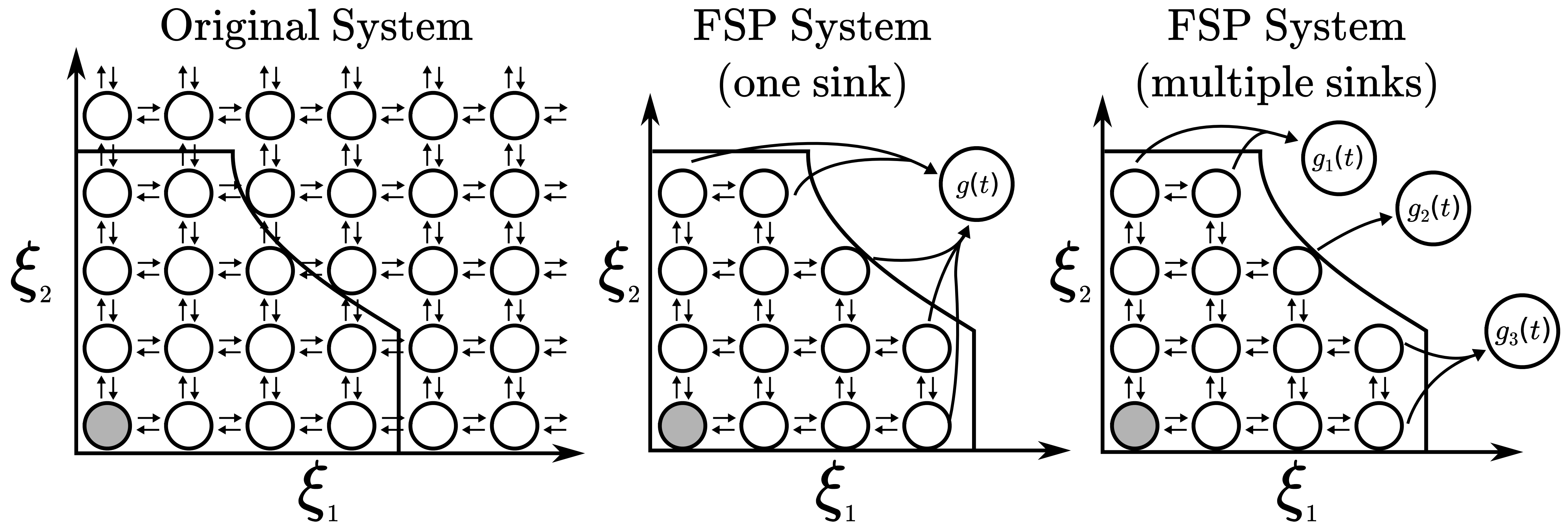
$$\frac{d\mathbf{p}}{dt} = \mathbf{A}(u(t), \boldsymbol{\theta})\mathbf{p}$$

The Chemical Master Equation (CME) describes the time varying statistics of such processes





# The finite state projection approach to solving the chemical master equation



# The FSP approach enables precise approximation of the CME, its sensitivity, and timing distributions.

## Approximate Master Equation

$$\frac{d}{dt} \begin{bmatrix} \mathbf{p}(t) \\ g(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{JJ} & \mathbf{0} \\ -\mathbf{1}^T \mathbf{A}_{JJ} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}(t) \\ g(t) \end{bmatrix}$$

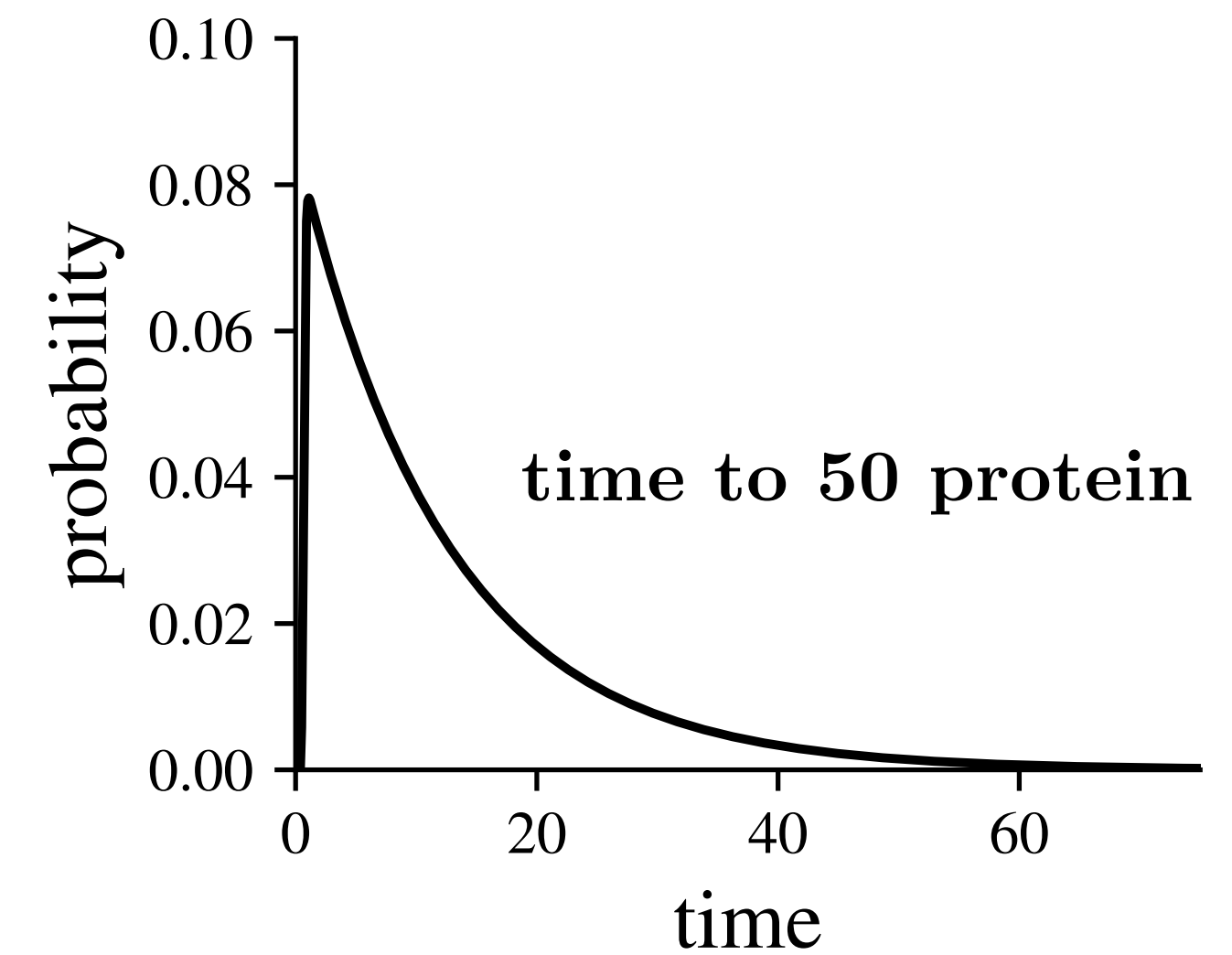
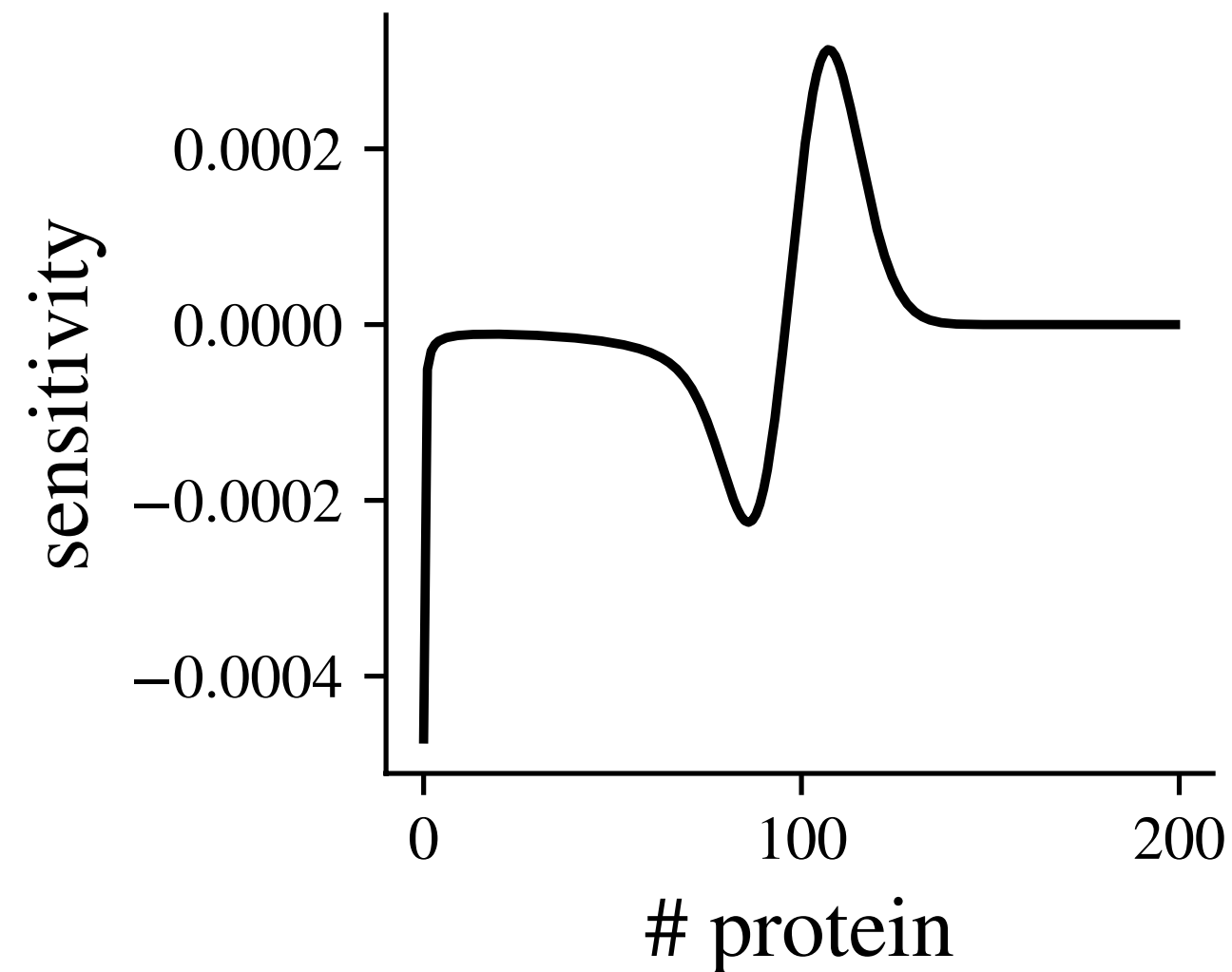
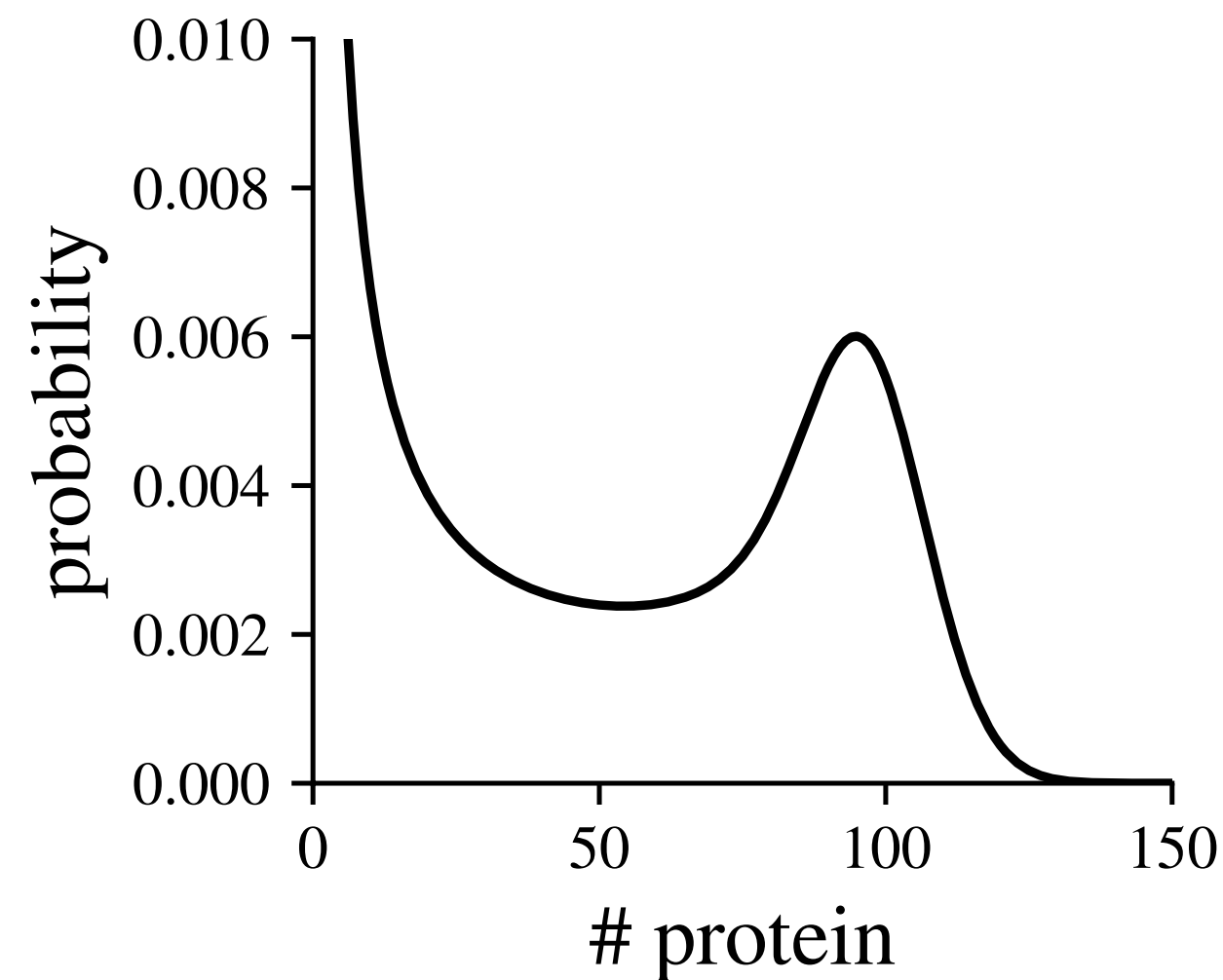
## Sensitivity Analysis

$$\frac{d}{dt} \begin{bmatrix} \mathbf{p}(t) \\ \mathbf{S}_i(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{A}_i & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{p}(t) \\ \mathbf{S}_i(t) \end{bmatrix}$$

where  $\mathbf{S}_i = \frac{\partial \mathbf{p}}{\partial \theta_i}$ ,  $\mathbf{A}_i = \frac{\partial \mathbf{A}}{\partial \theta_i}$

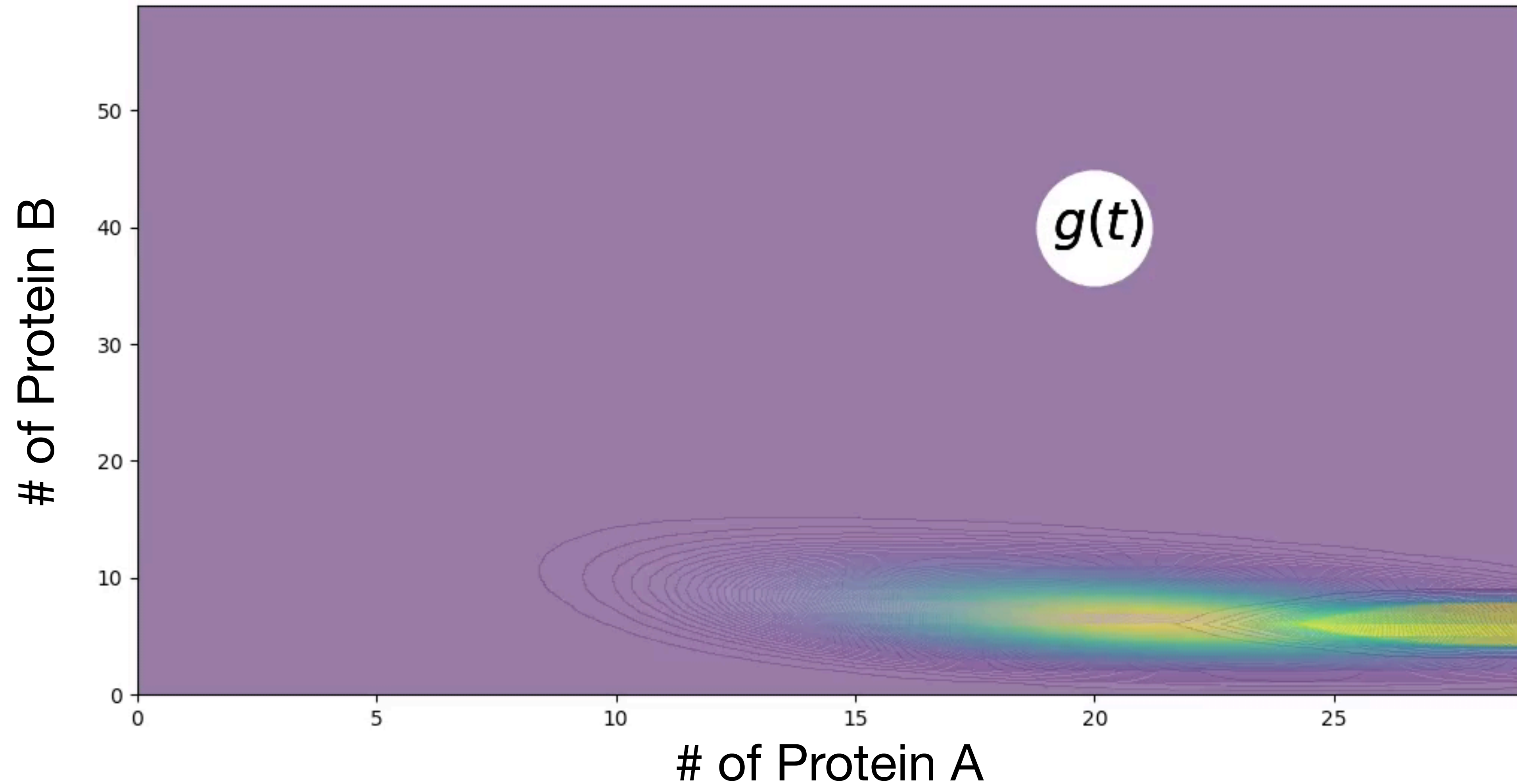
## Timing distribution

$$f(t) = -\mathbf{1}^T \mathbf{A}_{JJ} \exp(\mathbf{A}_{JJ}t) \mathbf{p}_0$$

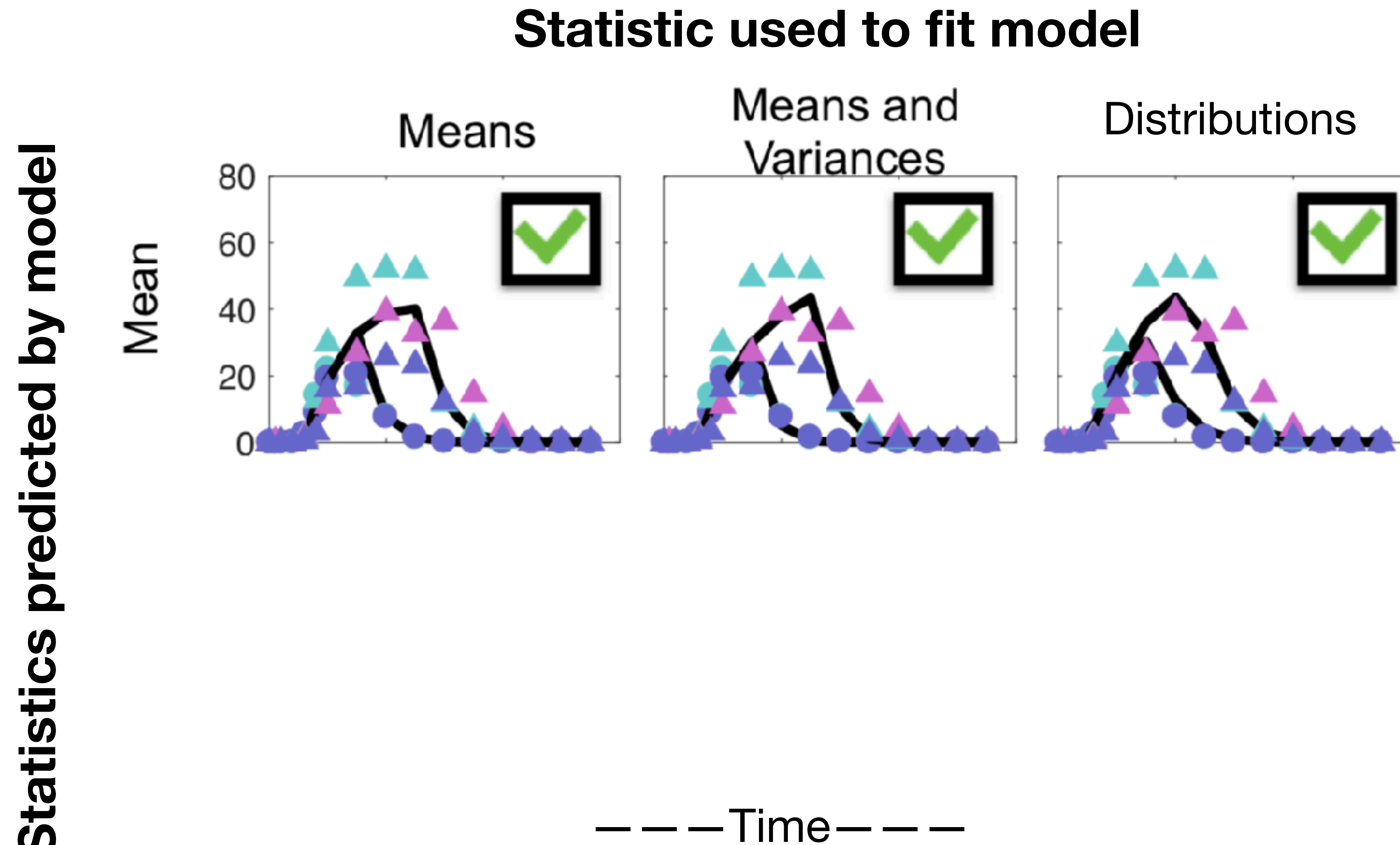


# Visualizing the FSP error

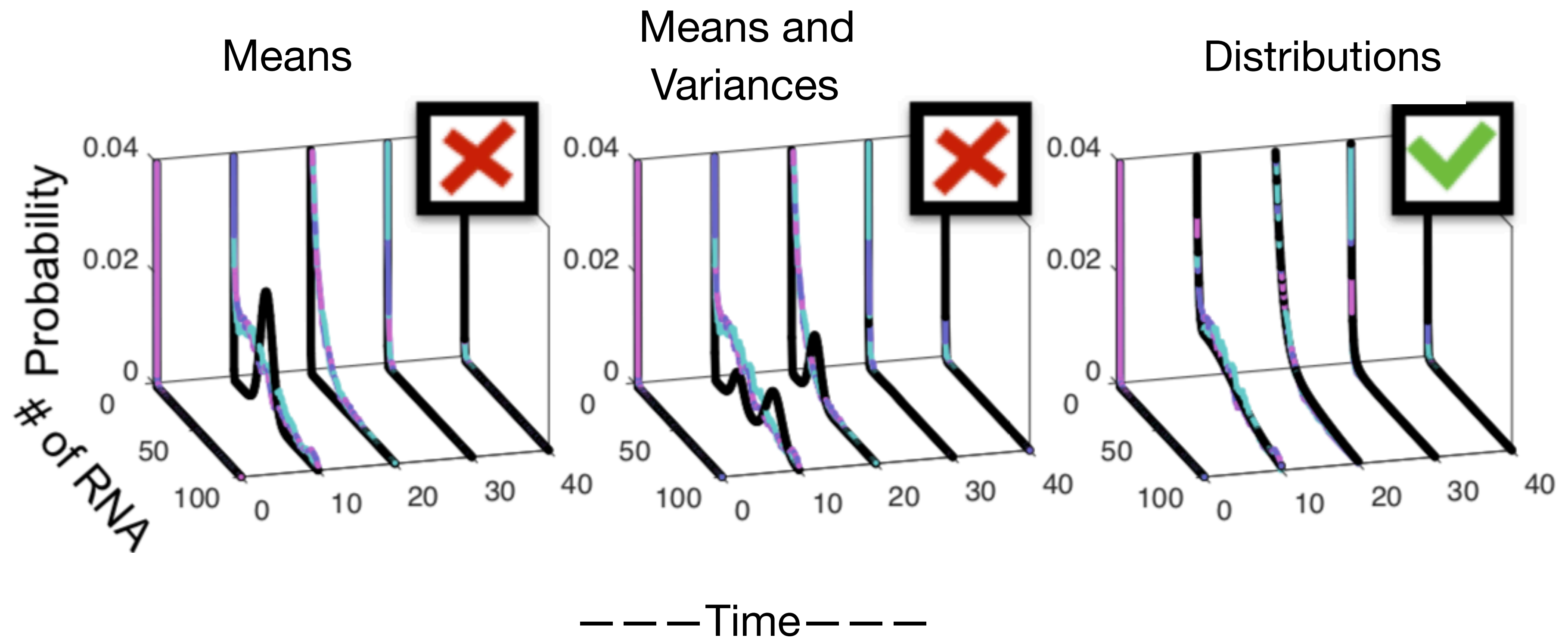
$t = 1$



# Summary statistics of distributions are not sufficient to characterize gene expression



Instead, all fluctuation information measured can be used to accurately identify models.

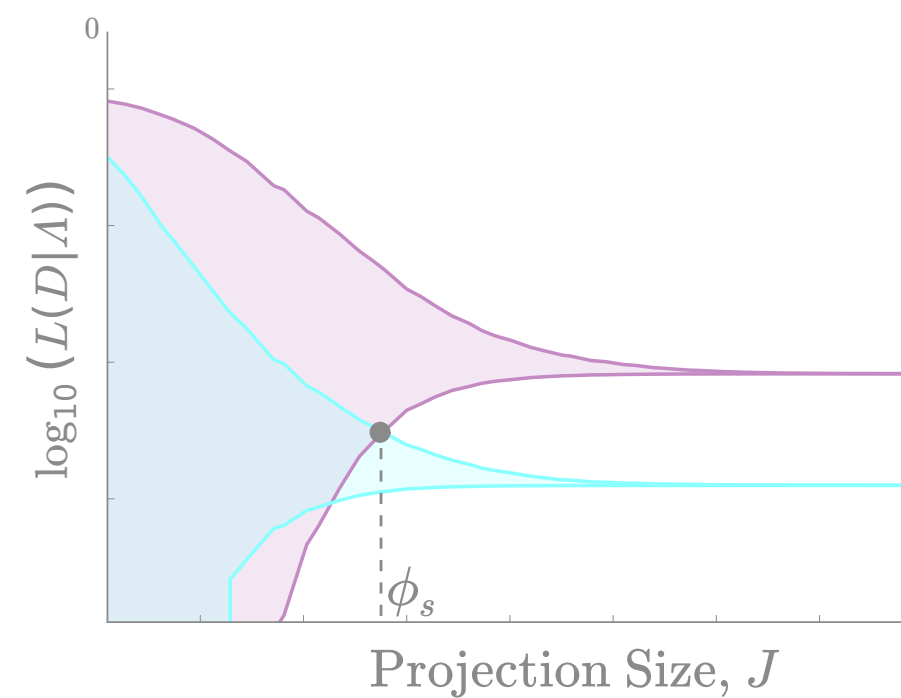


# Outline

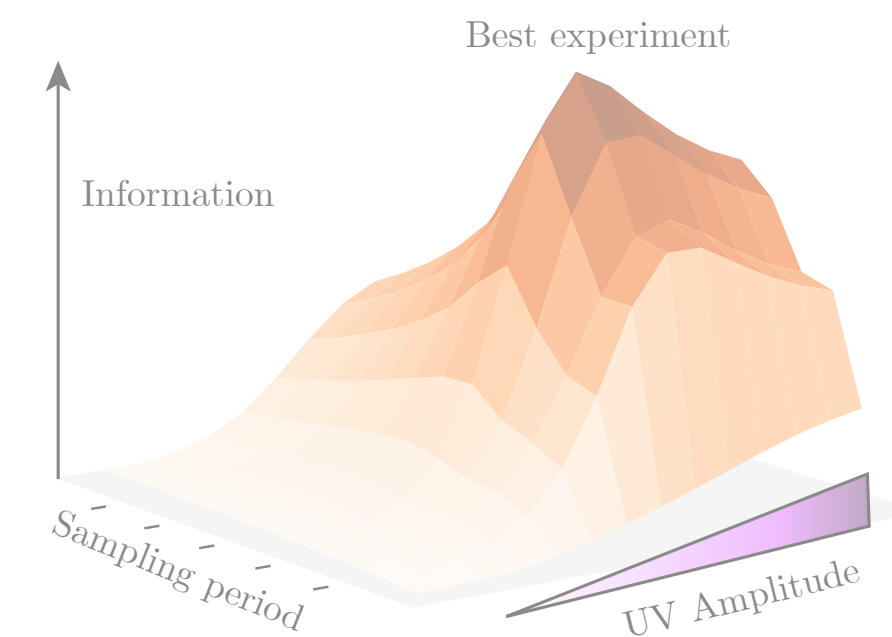
Variability in biochemical reactions

- **This variability is paramount to understanding and identifying models of gene expression.**
- **The FSP allows for computation of full probability distributions.**

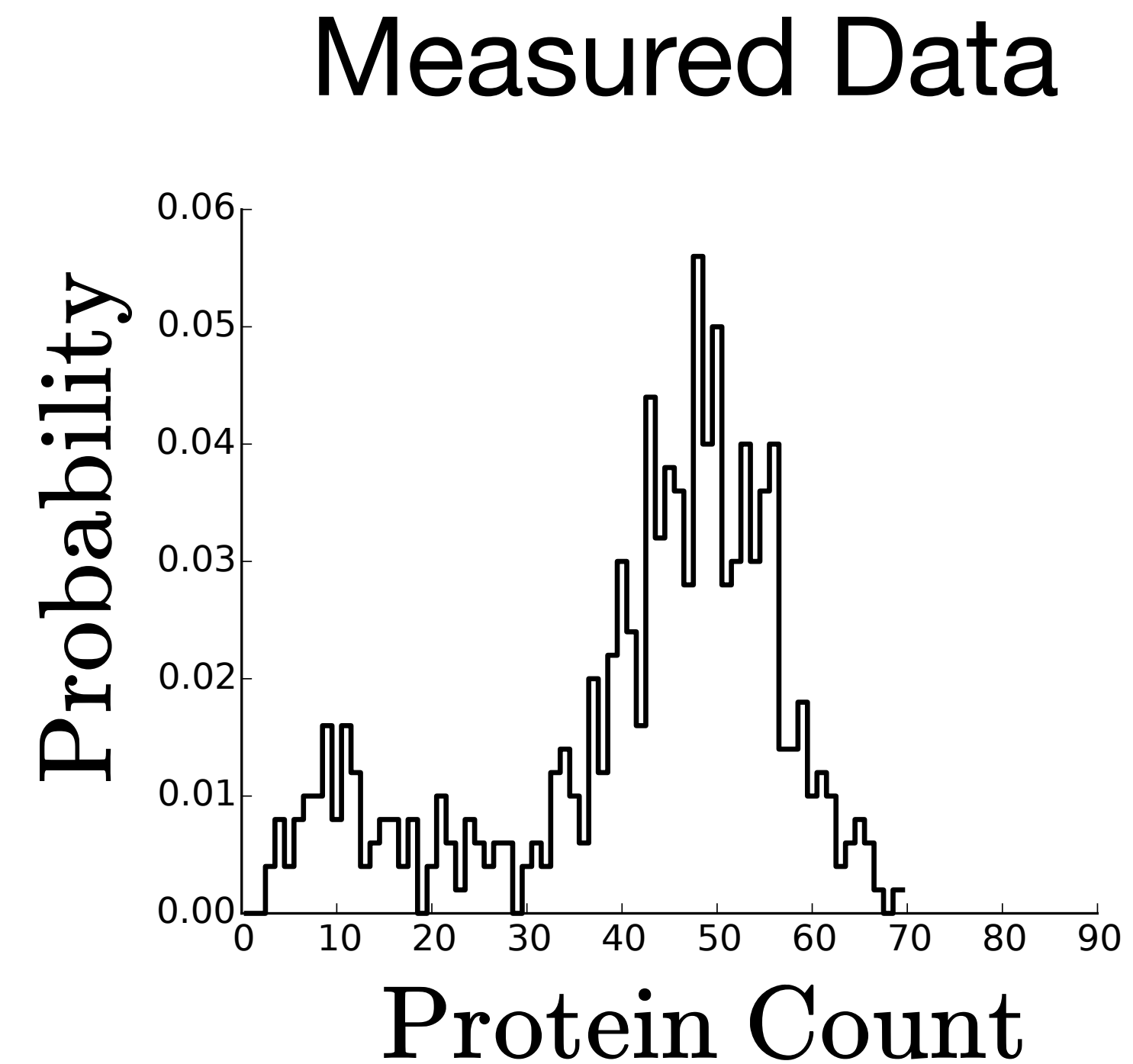
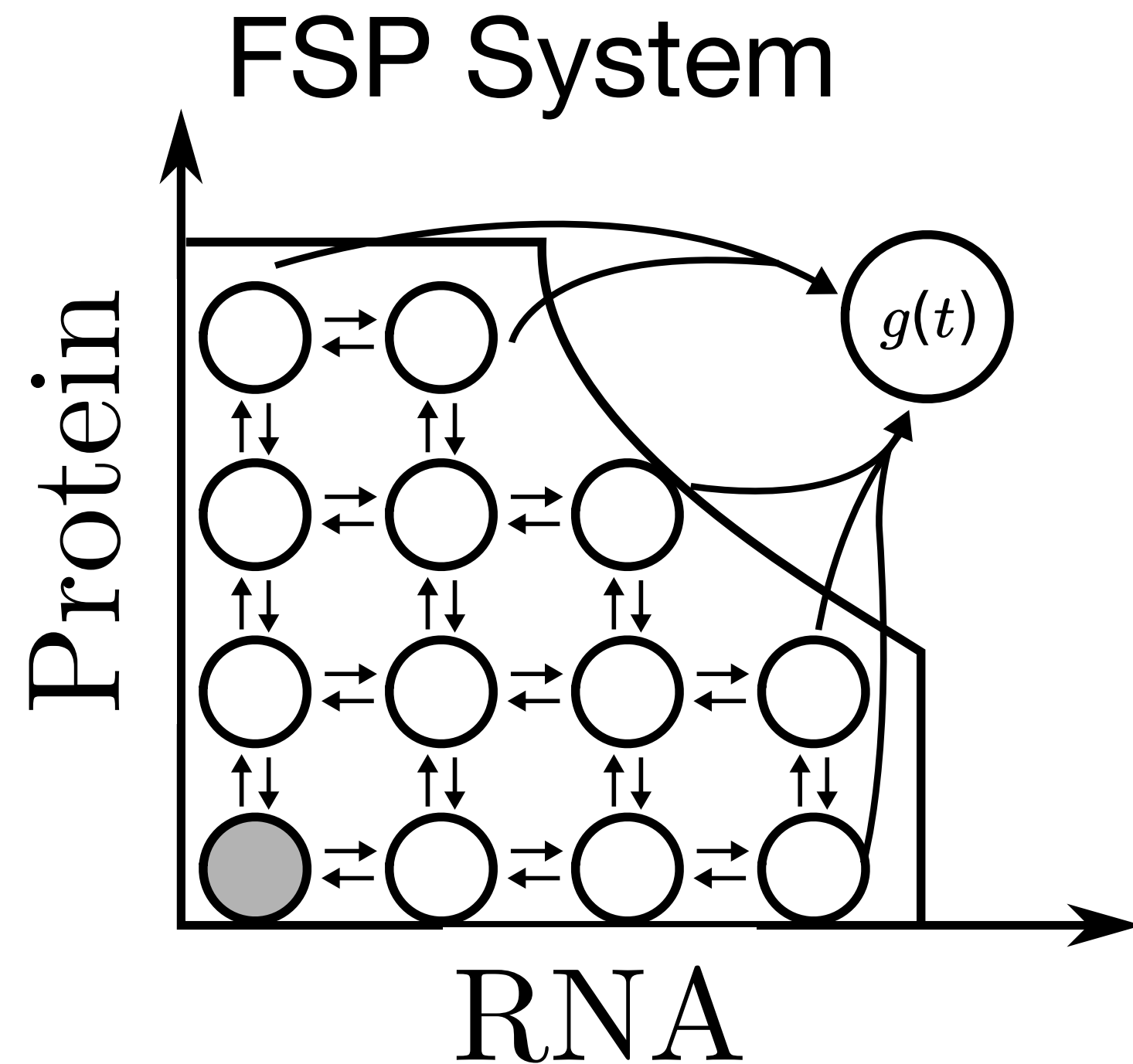
Efficient model identification using error constraints



Designing single-cell experiments with Fisher Information



# How can the known FSP error be used with single-cell data to help constrain models?



$$\sum_{t \in T} \sum_{i \in \mathcal{I}_D} d_{t,i} \log p_J^{FSP}(\mathbf{x}_i, t | \theta)$$

The lower bound on the likelihood follows from the definition of the FSP as a lower bound on the true probabilities.

$$\begin{bmatrix} p_J^{FSP}(\mathbf{x}, t) \\ \mathbf{0} \end{bmatrix} \leq \begin{bmatrix} p_J(\mathbf{x}, t) \\ p_{J'}(\mathbf{x}, t) \end{bmatrix} \text{ for all } t > 0$$

$$\sum_{t \in T} \sum_{i \in \mathcal{I}_D} d_{t,i} \log p_J^{FSP}(\mathbf{x}_i, t | \theta) \leq \sum_{t \in T} \sum_{i \in \mathcal{I}_D} d_{t,i} \log p(\mathbf{x}_i, t | \theta)$$



**An upper bound can be found by solving the following optimization problem:**

$$\left| \begin{bmatrix} p_J(\mathbf{x}, t) \\ p_{J'}(\mathbf{x}, t) \end{bmatrix} - \begin{bmatrix} p_J^{FSP}(\mathbf{x}, t) \\ \mathbf{0} \end{bmatrix} \right| = g(t)$$

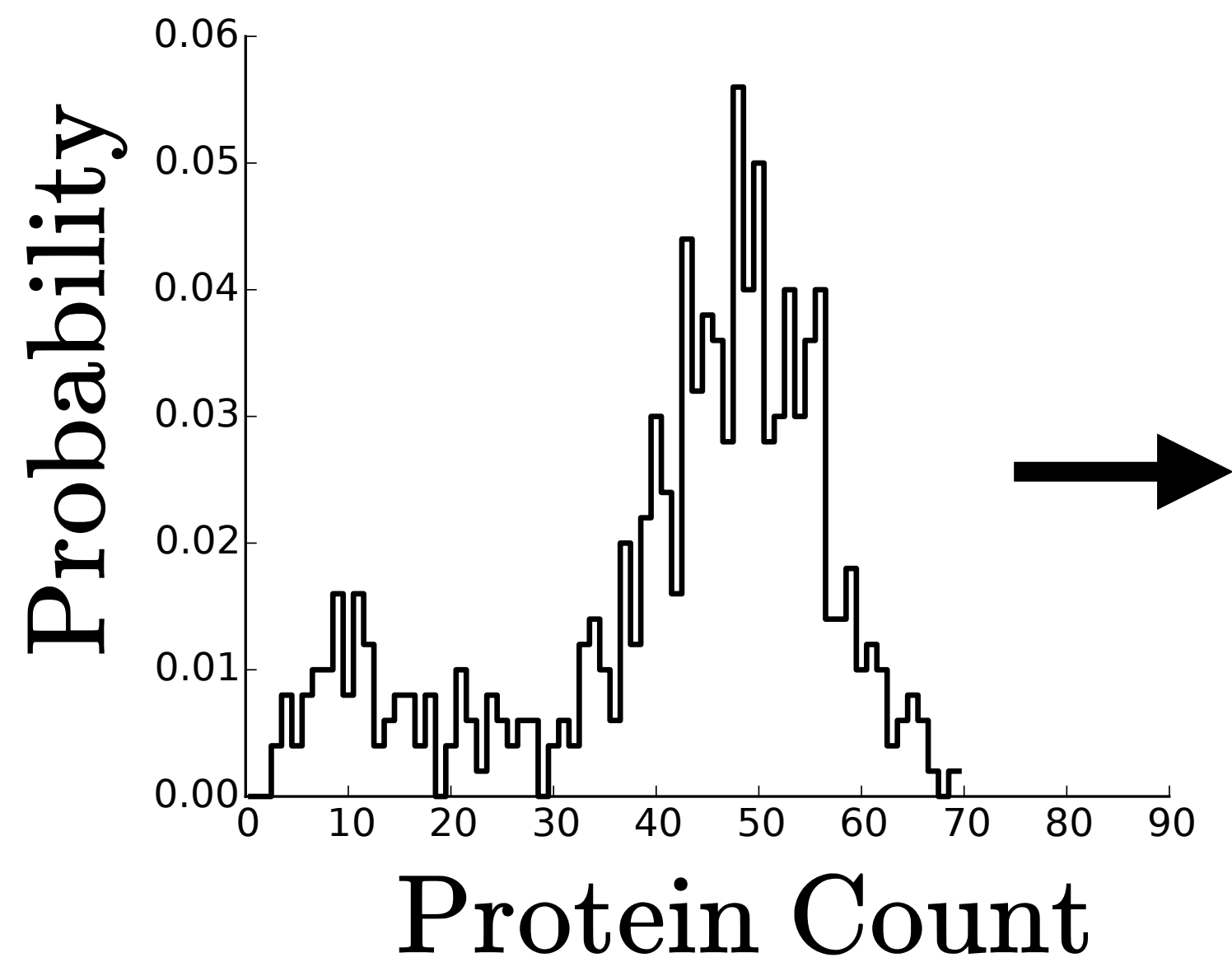
$$\max_{\boldsymbol{\varepsilon}} \sum_{t \in T} \sum_{j \in \mathcal{J}_D} d_{t,j} \log \left( p_J^{FSP}(\mathbf{x}_j, t | \theta) + \varepsilon_{t,j} \right)$$

$$\text{Such that } \sum_k \varepsilon_{t,k} = g(t) \text{ and } \varepsilon_{t,k} \geq 0.$$

**This is solved using an iterative water-filling algorithm.**

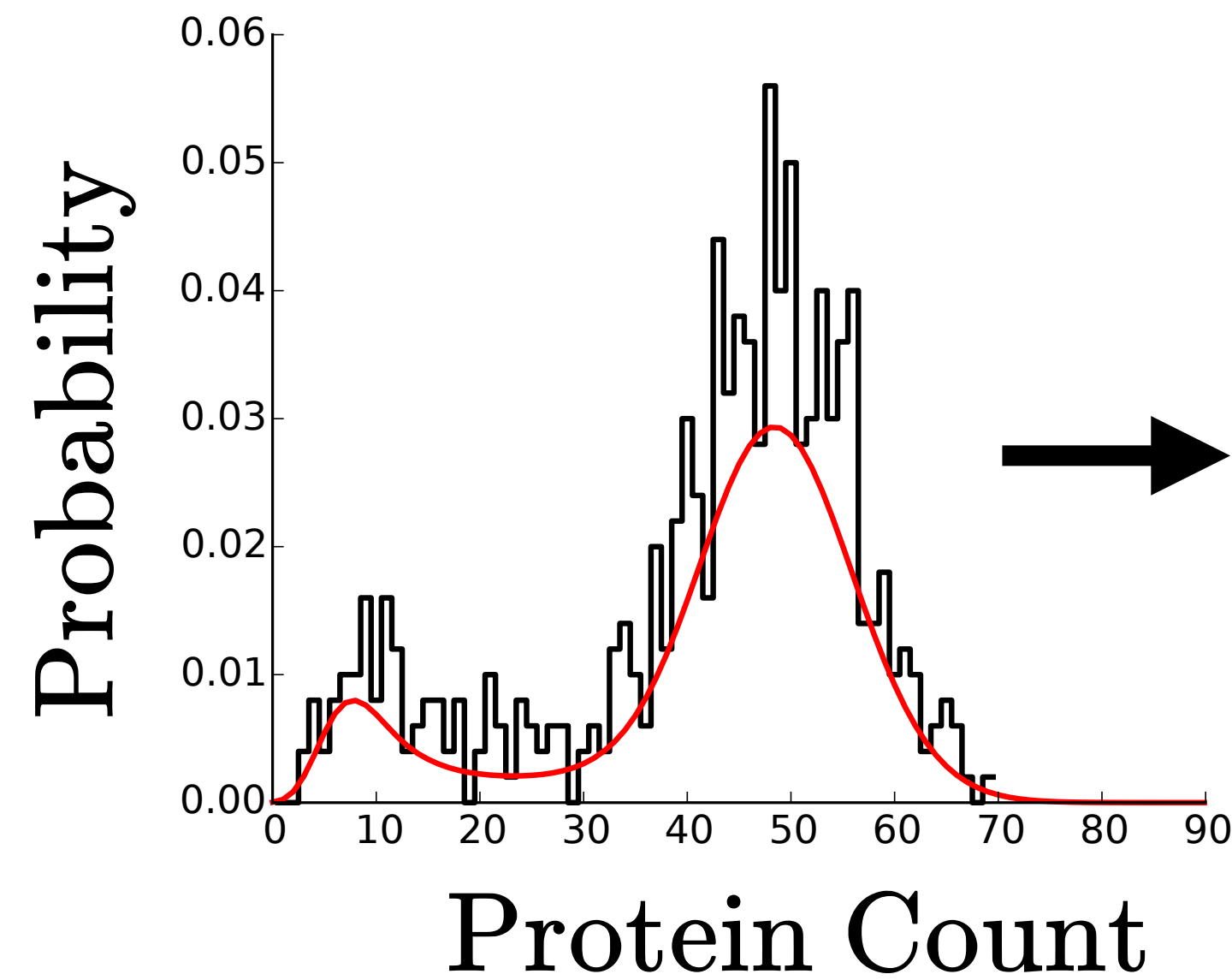
# The optimal error redistribution adds probability where it will most affect the likelihood function.

Simulated Data



Simulated Data

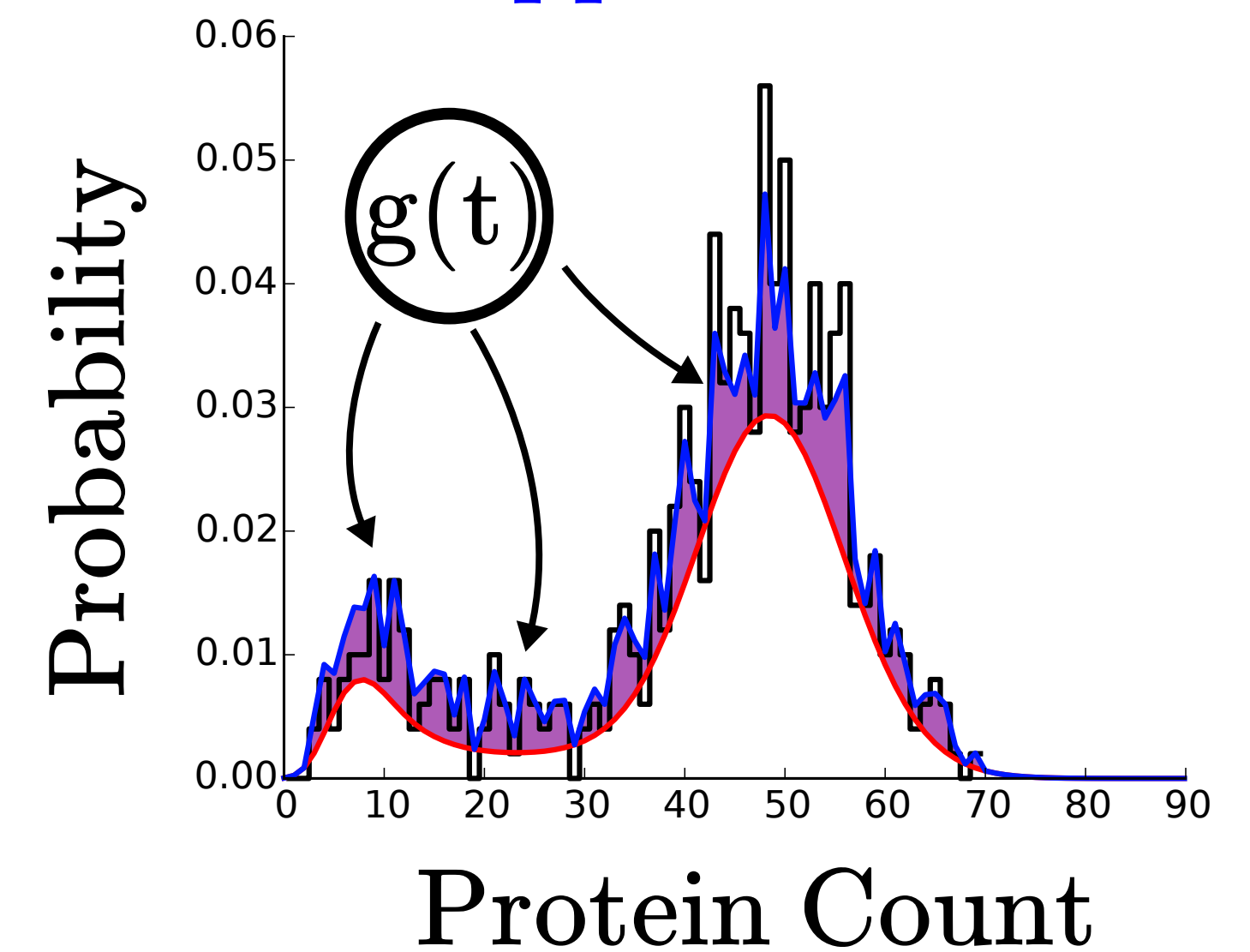
Lower Bound



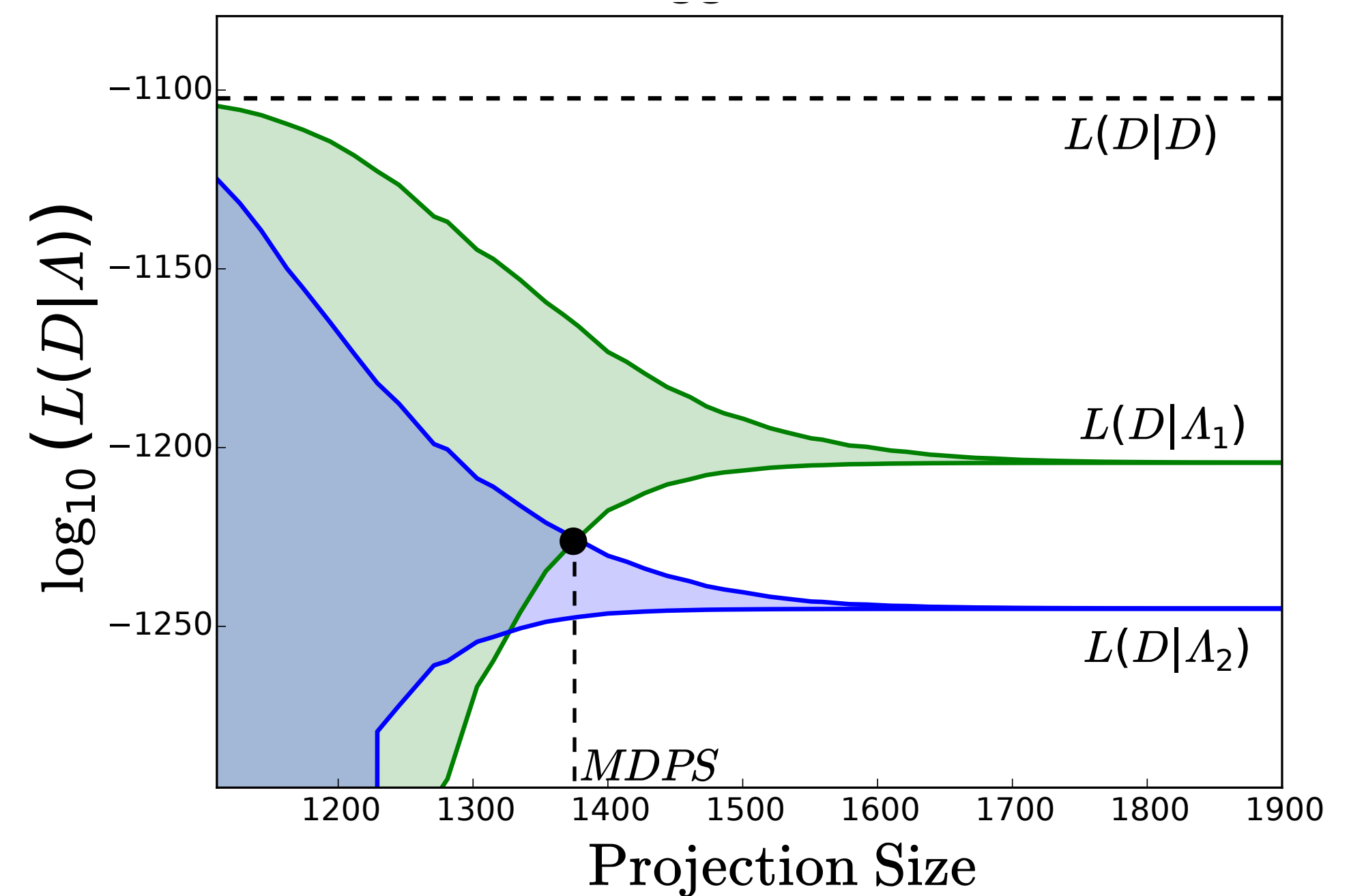
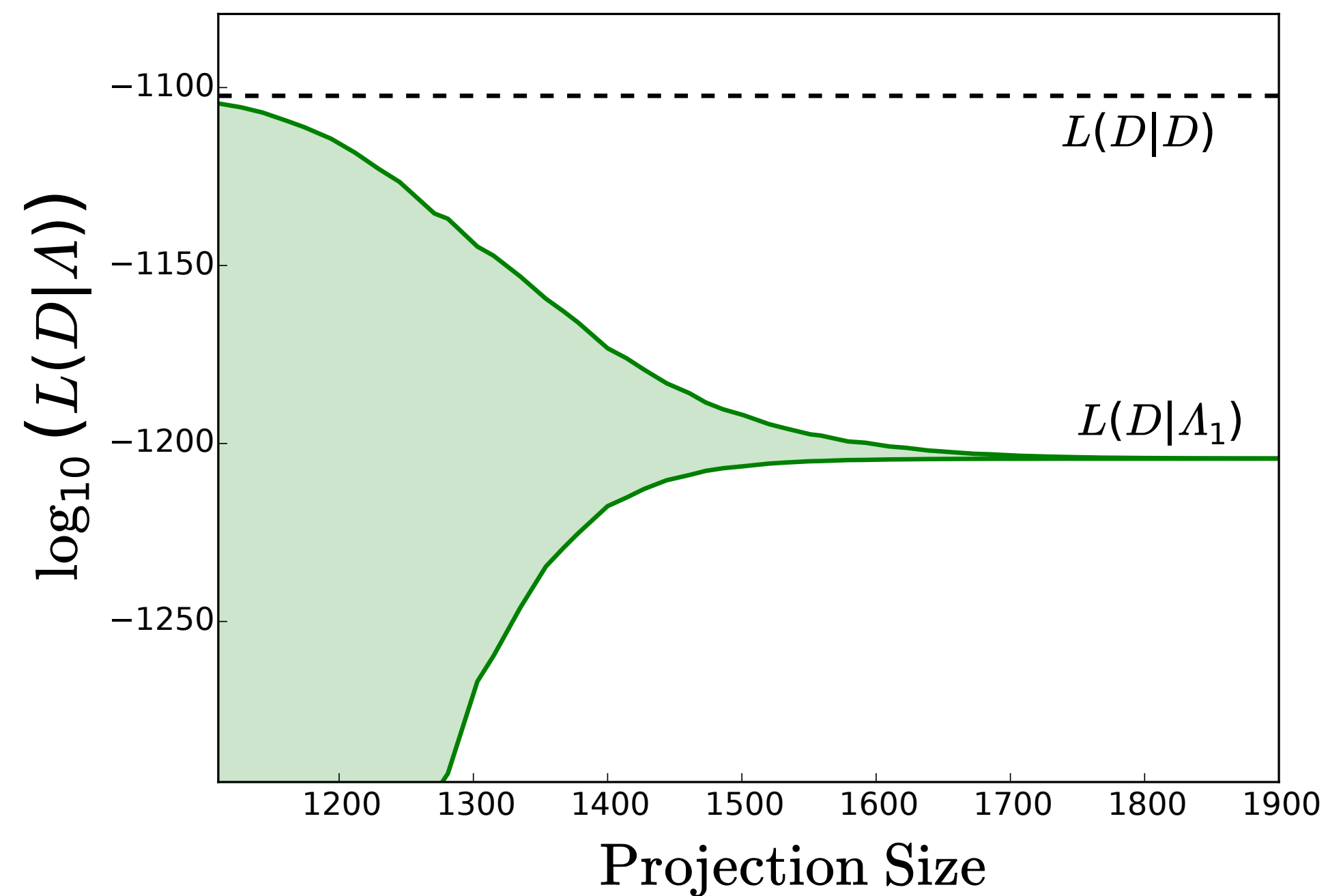
Simulated Data

Lower Bound

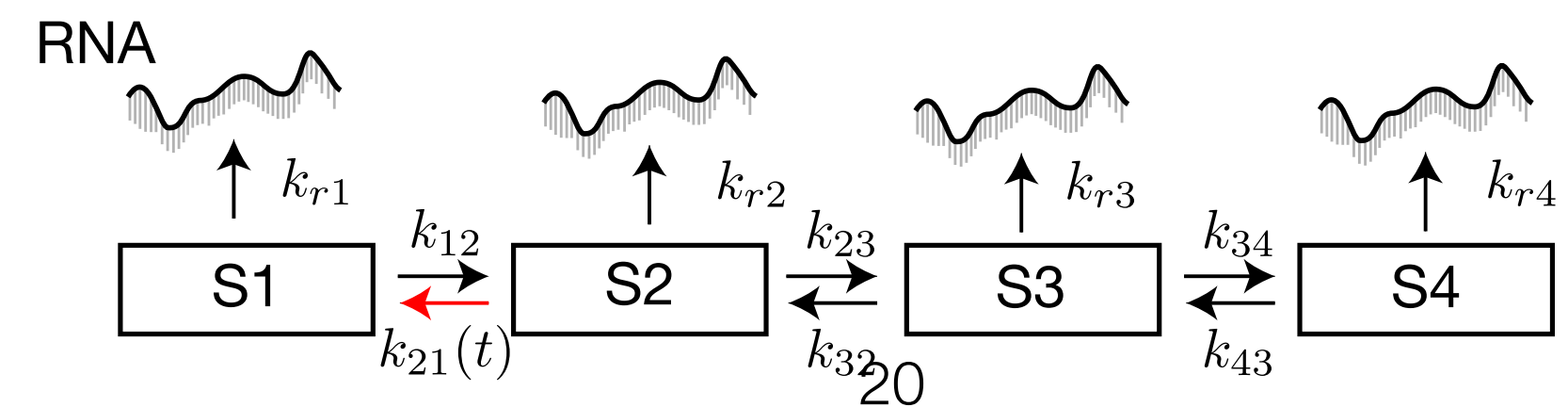
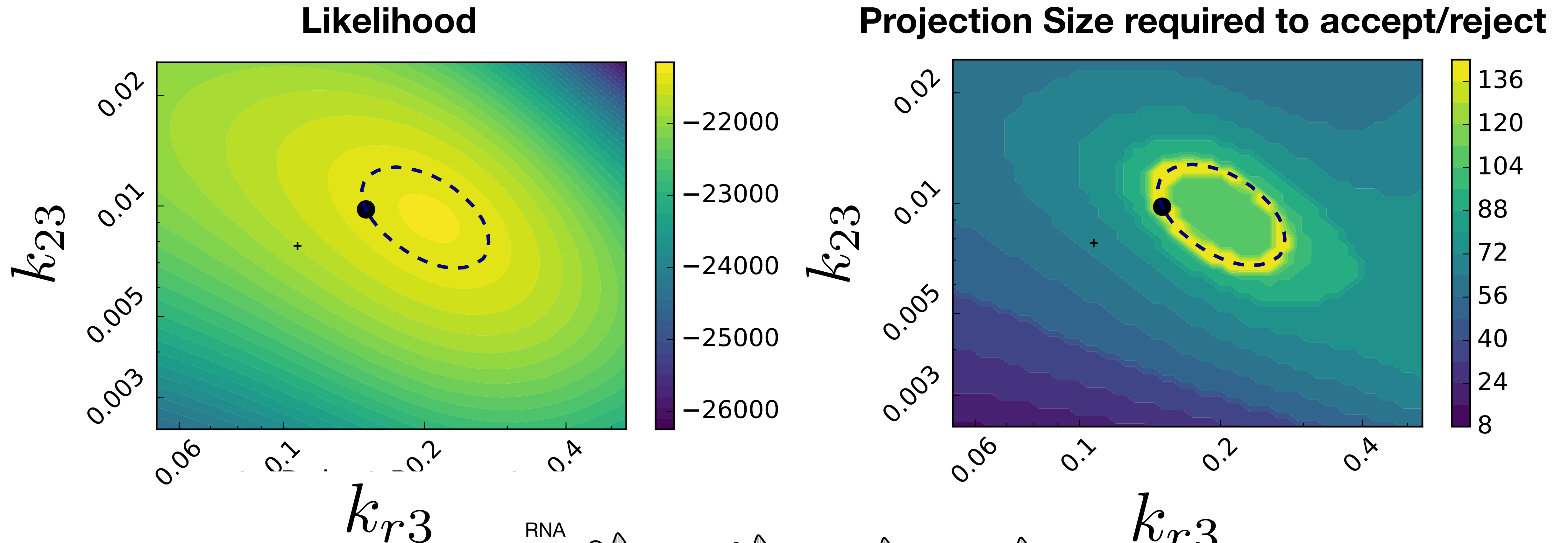
Upper Bound



# This bound informs the accuracy required to tell apart two models.



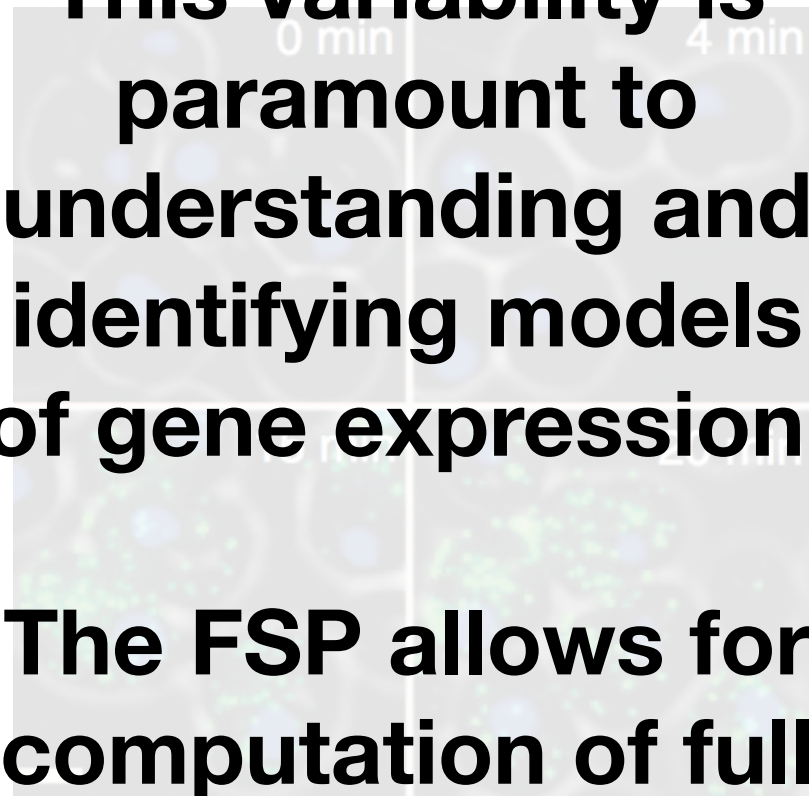
# FSP-bounds reduce the error needed for model discrimination for the Hog1-p model.



# Outline

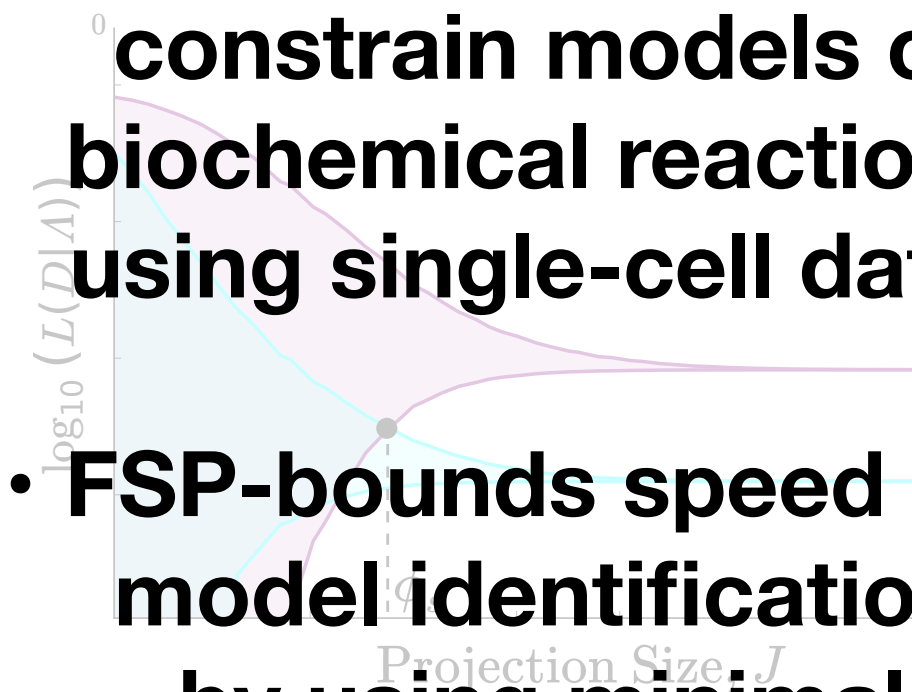
Variability in biochemical reactions

- **This variability is paramount to understanding and identifying models of gene expression.**
- **The FSP allows for computation of full probability distributions.**

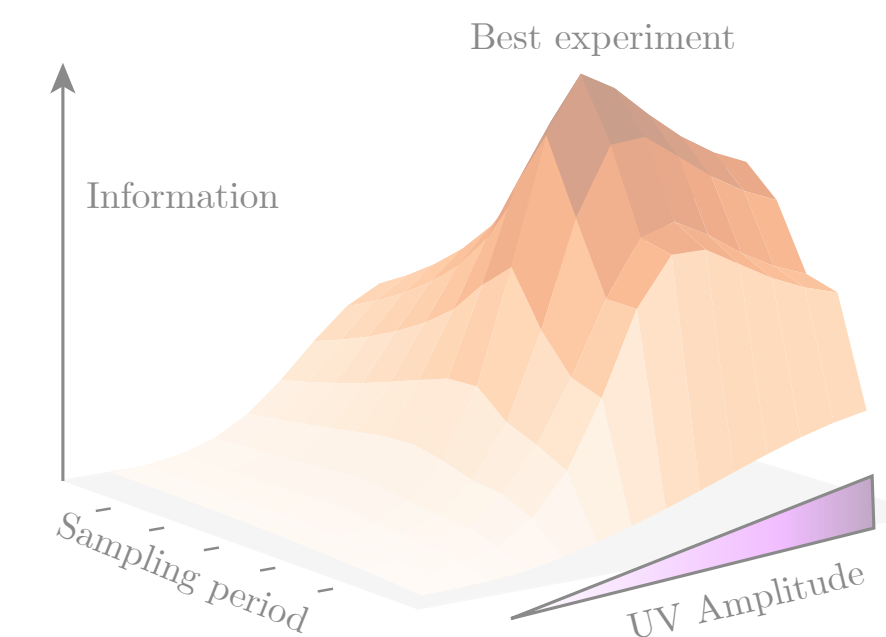


Efficient model identification using error constraints

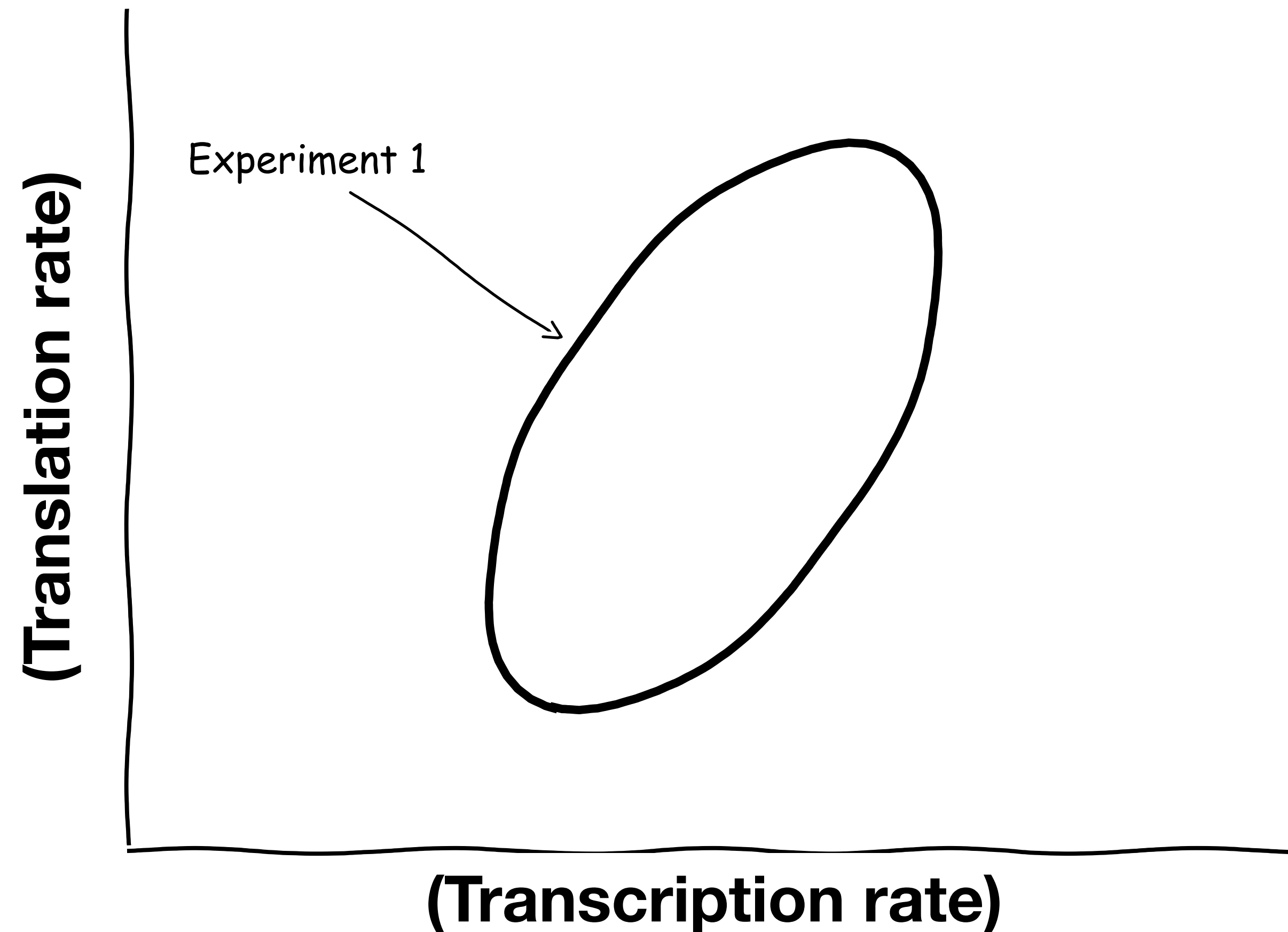
- **FSP errors allow us to constrain models of biochemical reactions using single-cell data.**
- **FSP-bounds speed up model identification by using minimal computational effort to compare models.**



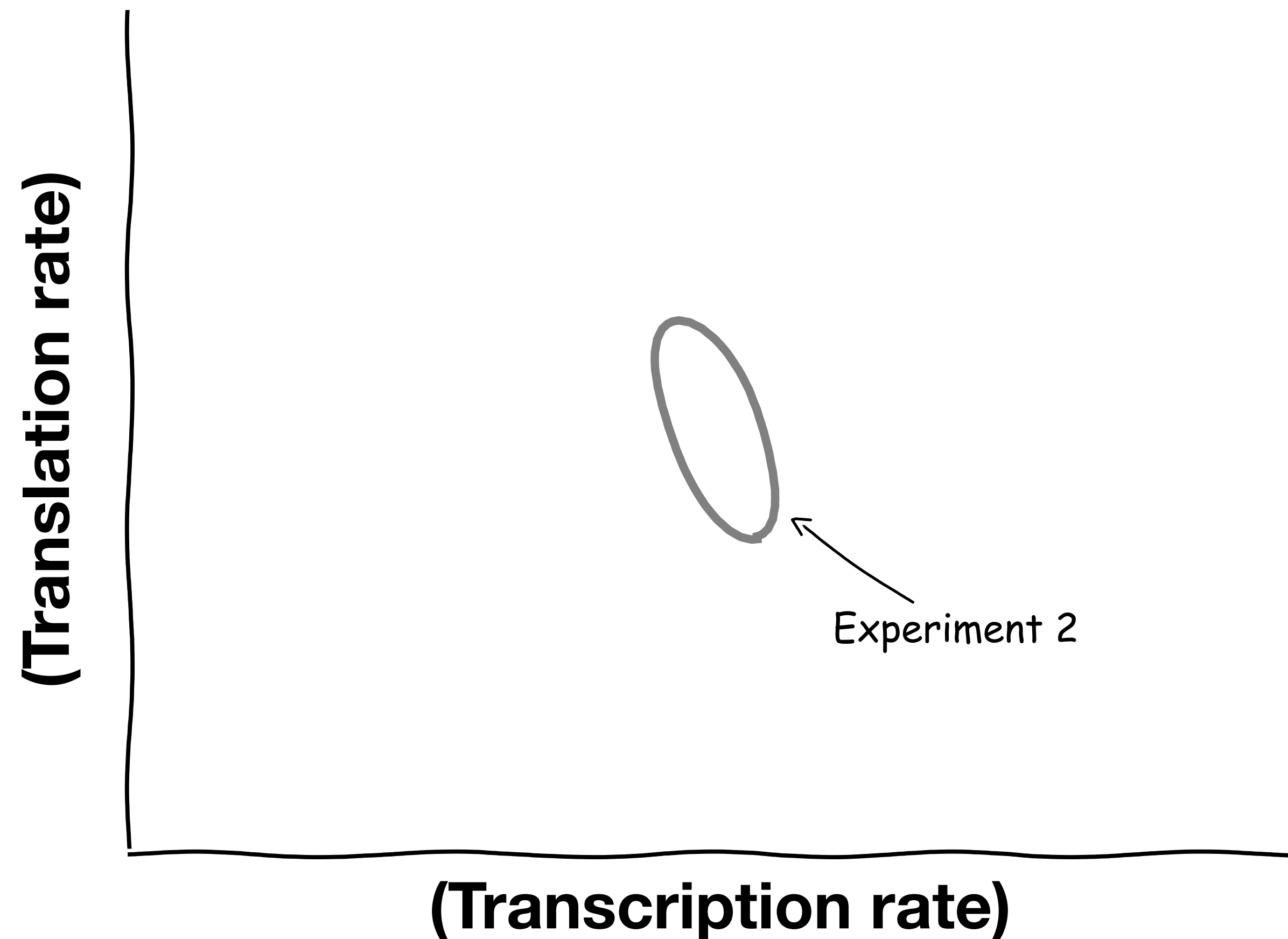
Designing single-cell experiments with Fisher Information



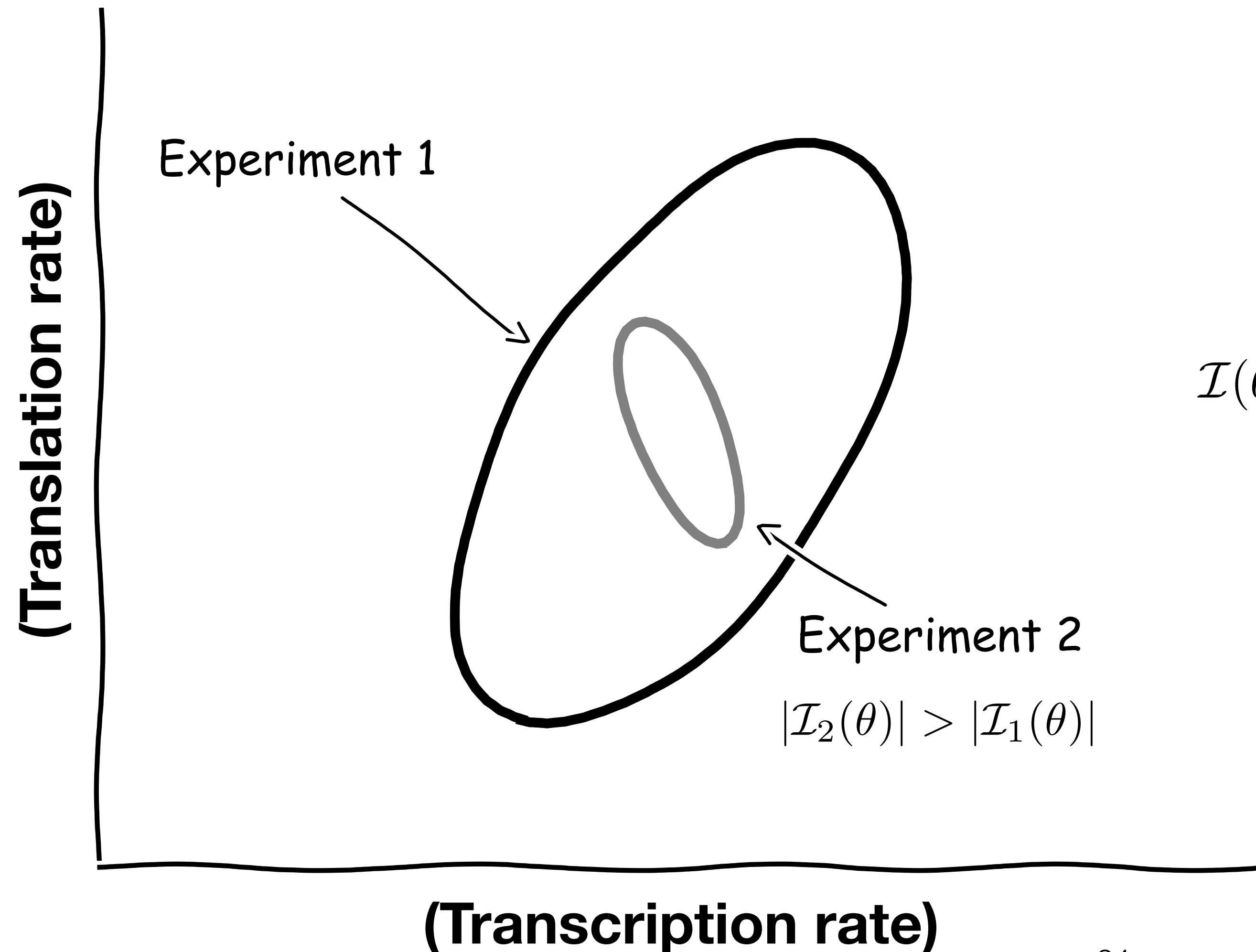
**Different single-cell experiments reveal different amounts of information about model parameters.**



Different single-cell experiments reveal different amounts of information about model parameters.



# Different single-cell experiments reveal different amounts of information about model parameters.



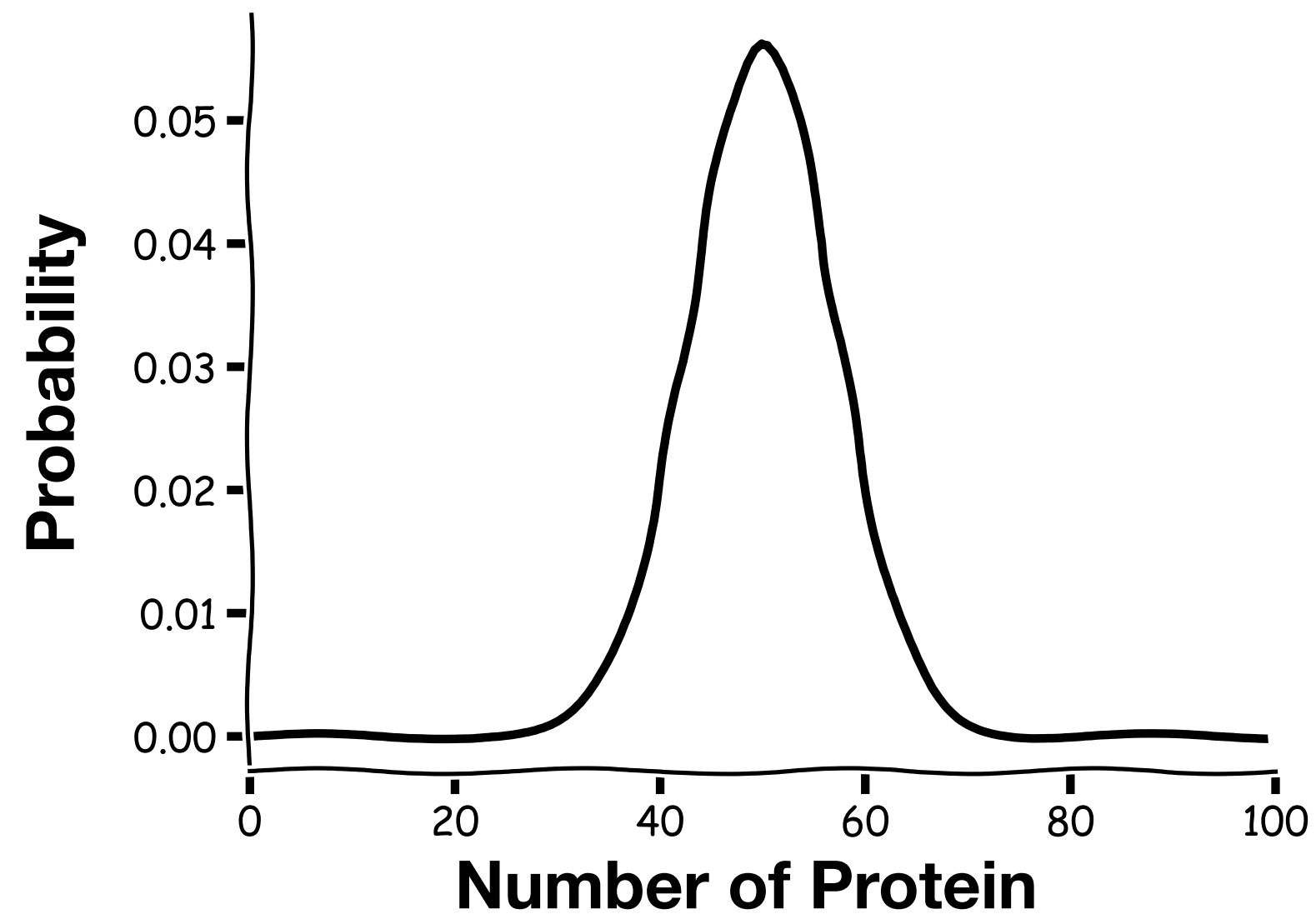
The Fisher information matrix (FIM) quantifies the expected uncertainty of potential experiments

$$\mathcal{I}(\theta) = \mathbf{E} \left[ \left( \nabla_{\theta} \log L(\mathbf{D}; \theta) \right)^T \left( \nabla_{\theta} \log L(\mathbf{D}; \theta) \right) \right]$$

Different FIM metrics estimate which experiments are better to answer specific questions.

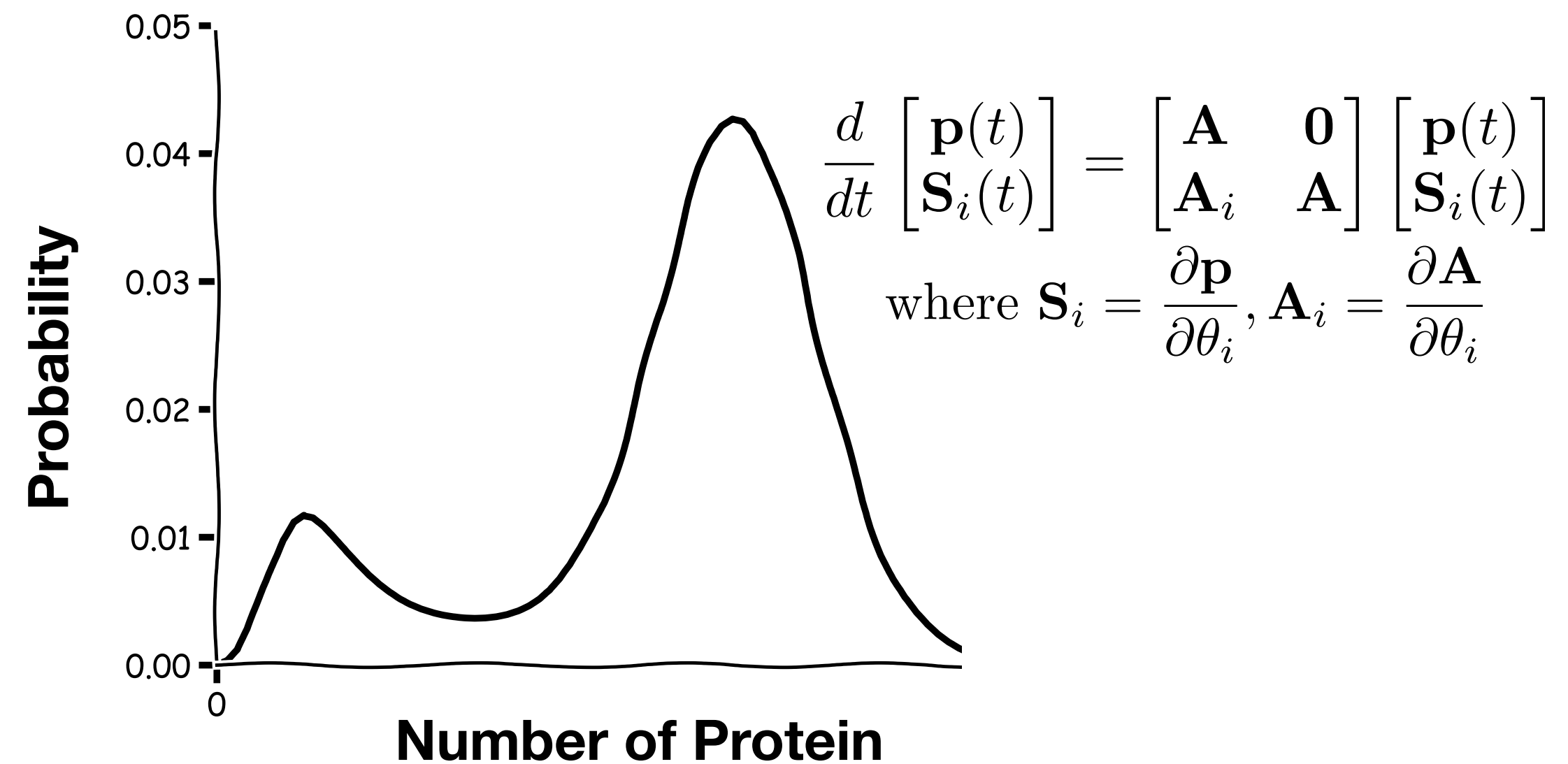


# The FSP-FIM extends FIM analyses beyond the limits of well-known distributions and simple stochastic processes.



The FIM for Gaussian distributions is well-known:

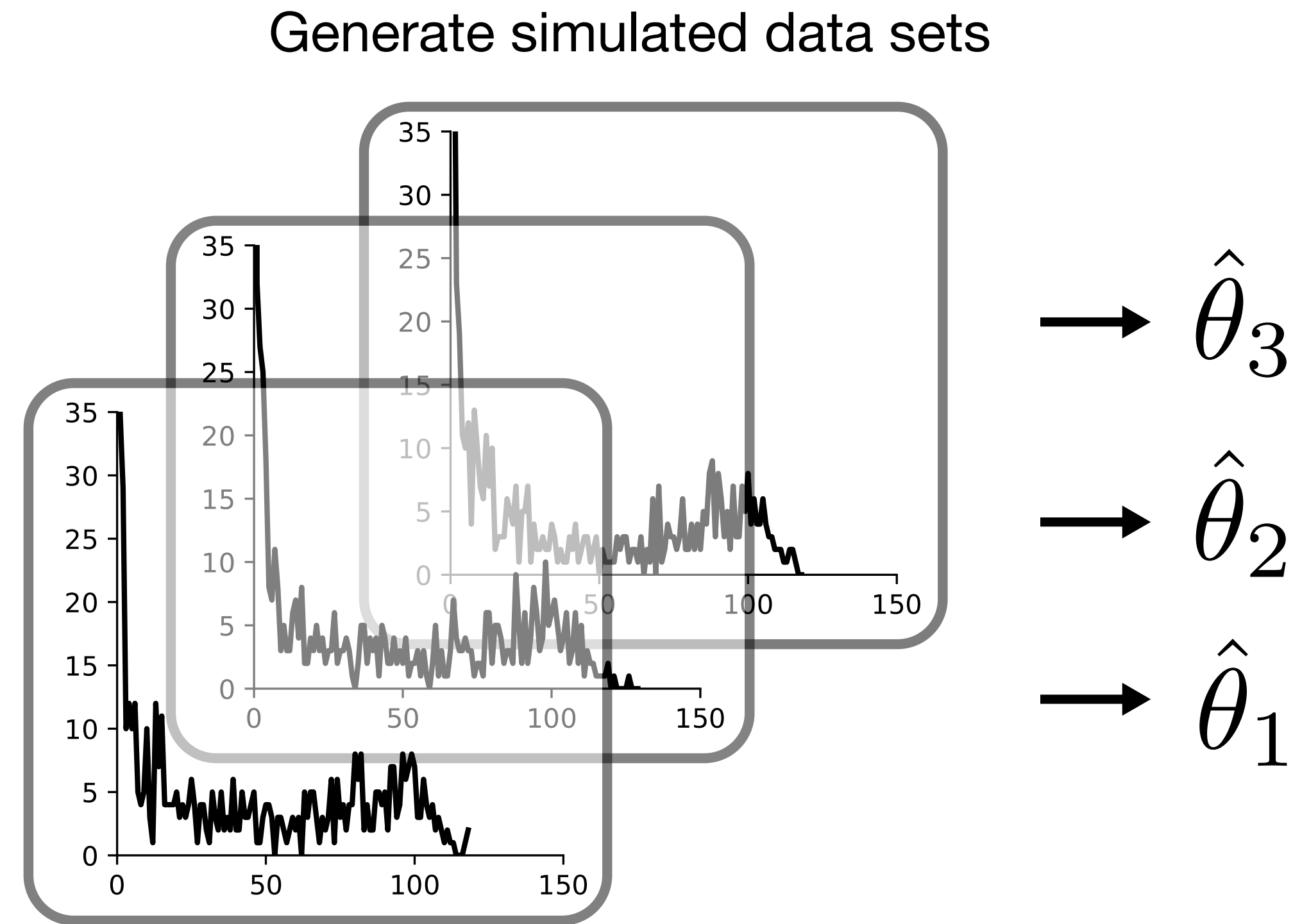
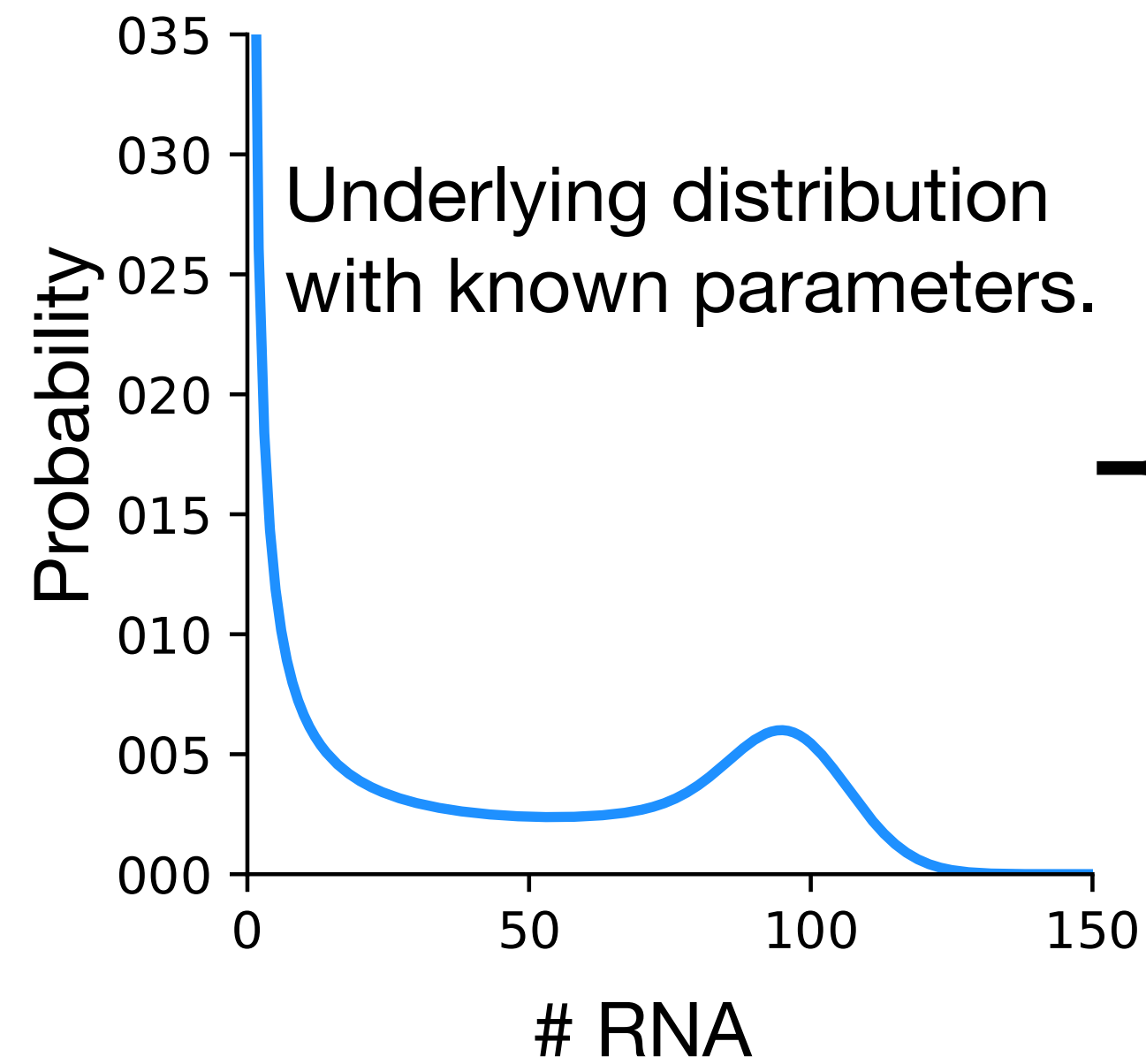
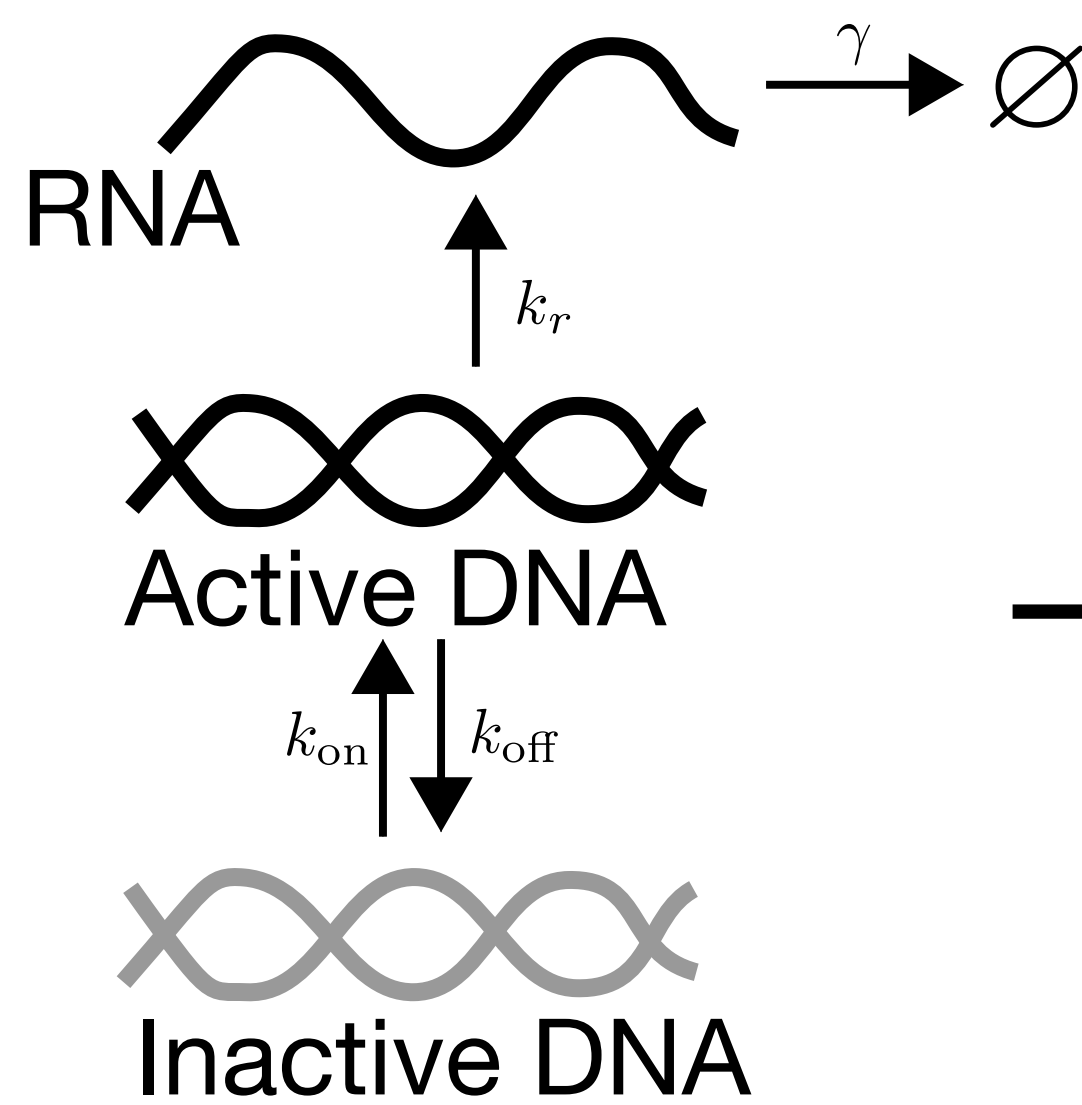
$$FIM_{i,j} = \frac{\partial \boldsymbol{\mu}^T}{\partial \theta_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j} + \frac{1}{2} \text{trace} \left( \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j} \right).$$



The FSP-FIM analyzes information for discrete distributions of any shape:

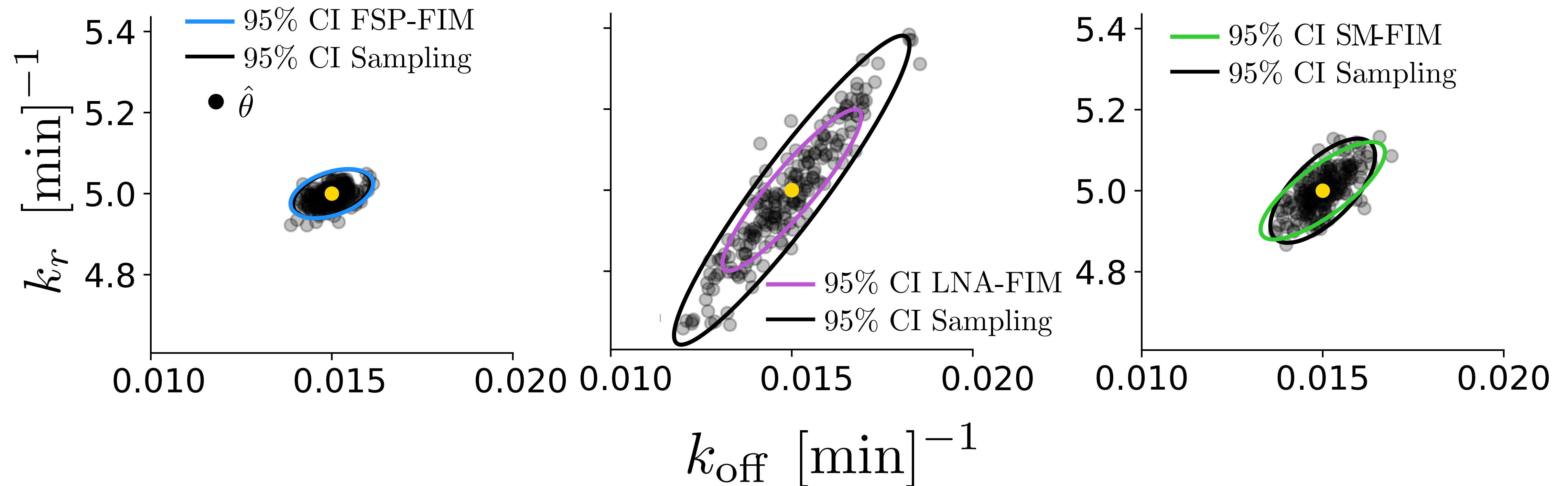
$$FIM_{i,j} = n \sum_{k=1}^N \frac{1}{p(\mathbf{x}_k; \boldsymbol{\theta})} \mathbf{S}_i^k \mathbf{S}_j^k$$

# The asymptotic normality of the maximum likelihood estimator can be used to confirm the FSP-FIM for non-Gaussian data.

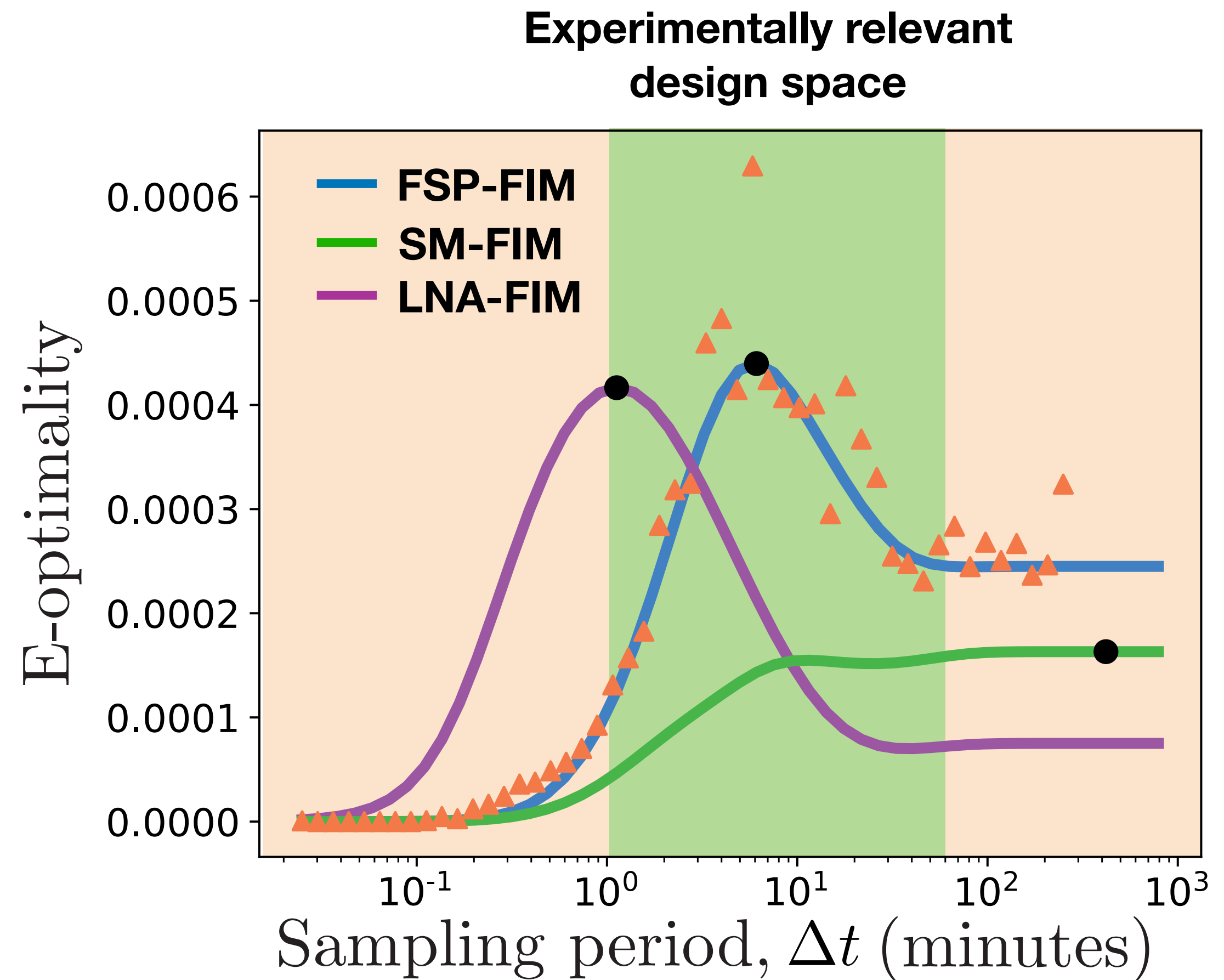
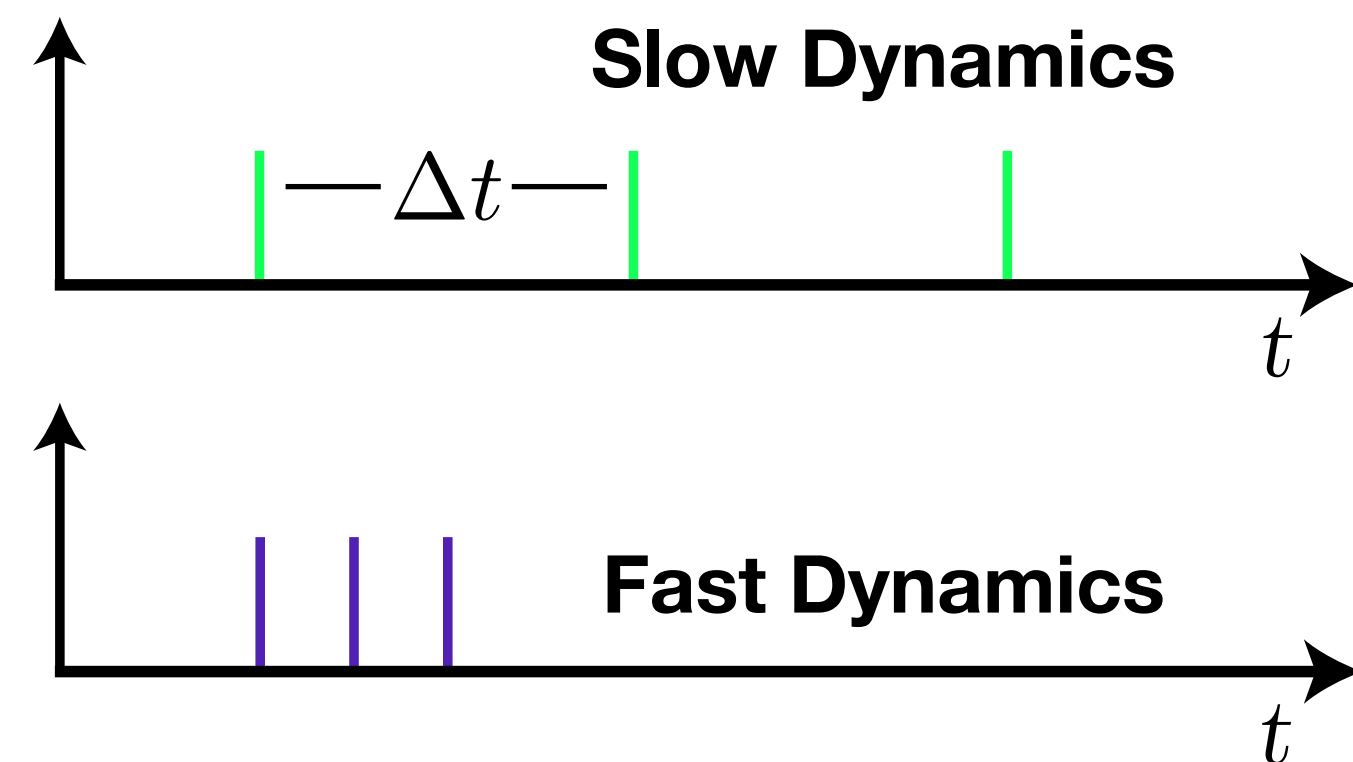


$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{dist} \mathcal{N}(0, I(\theta^*)^{-1})$$

The FSP-FIM provides a more accurate FIM approximation, even when models are linear and moments are computable.

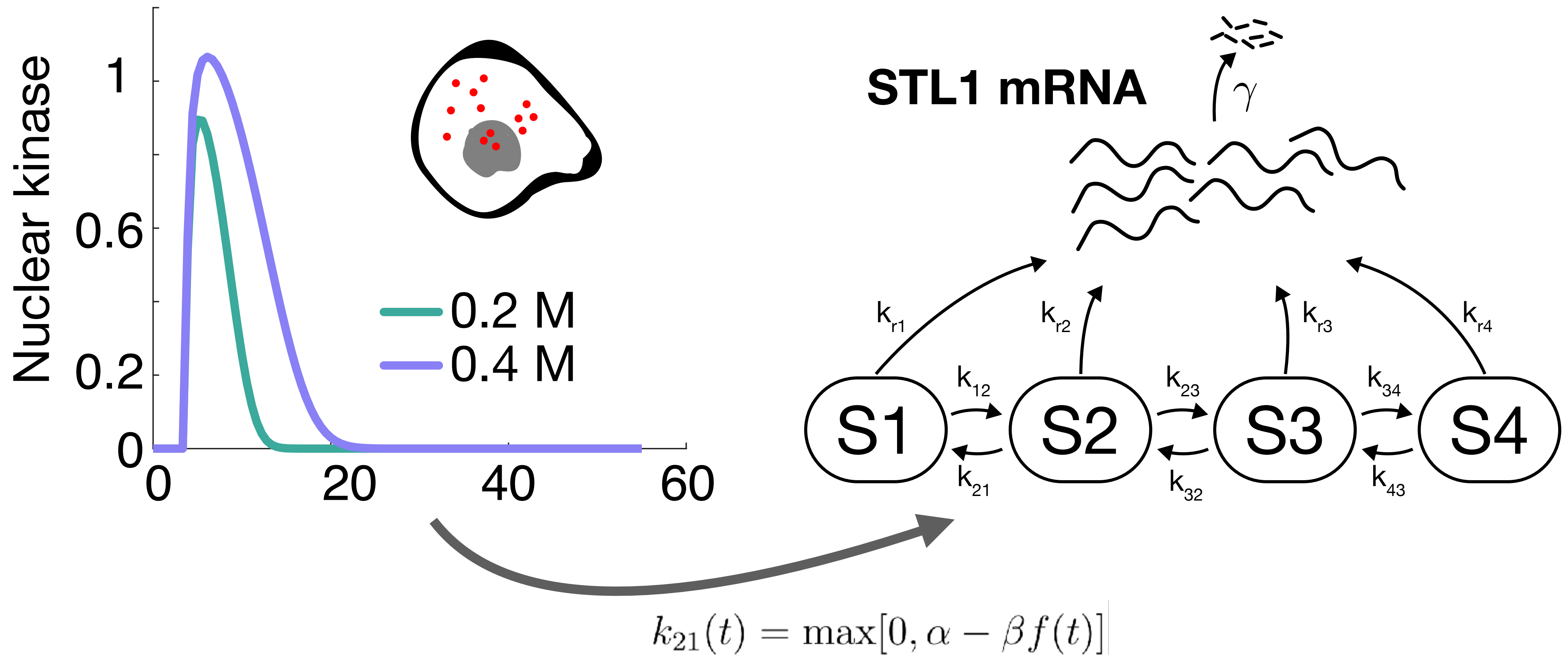


# The FSP-FIM analysis identifies more informative experiment designs, which were missed by alternate approaches.



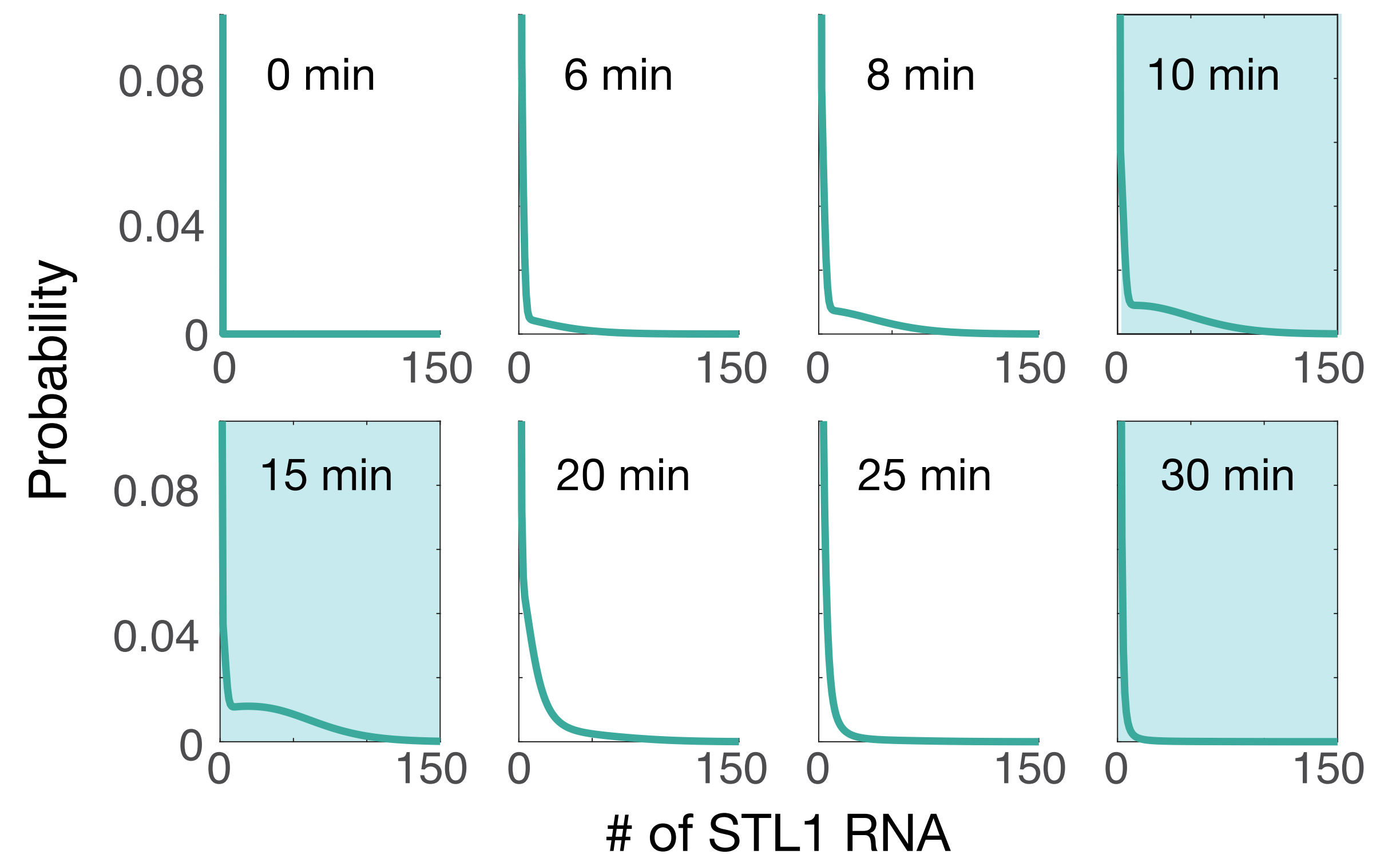
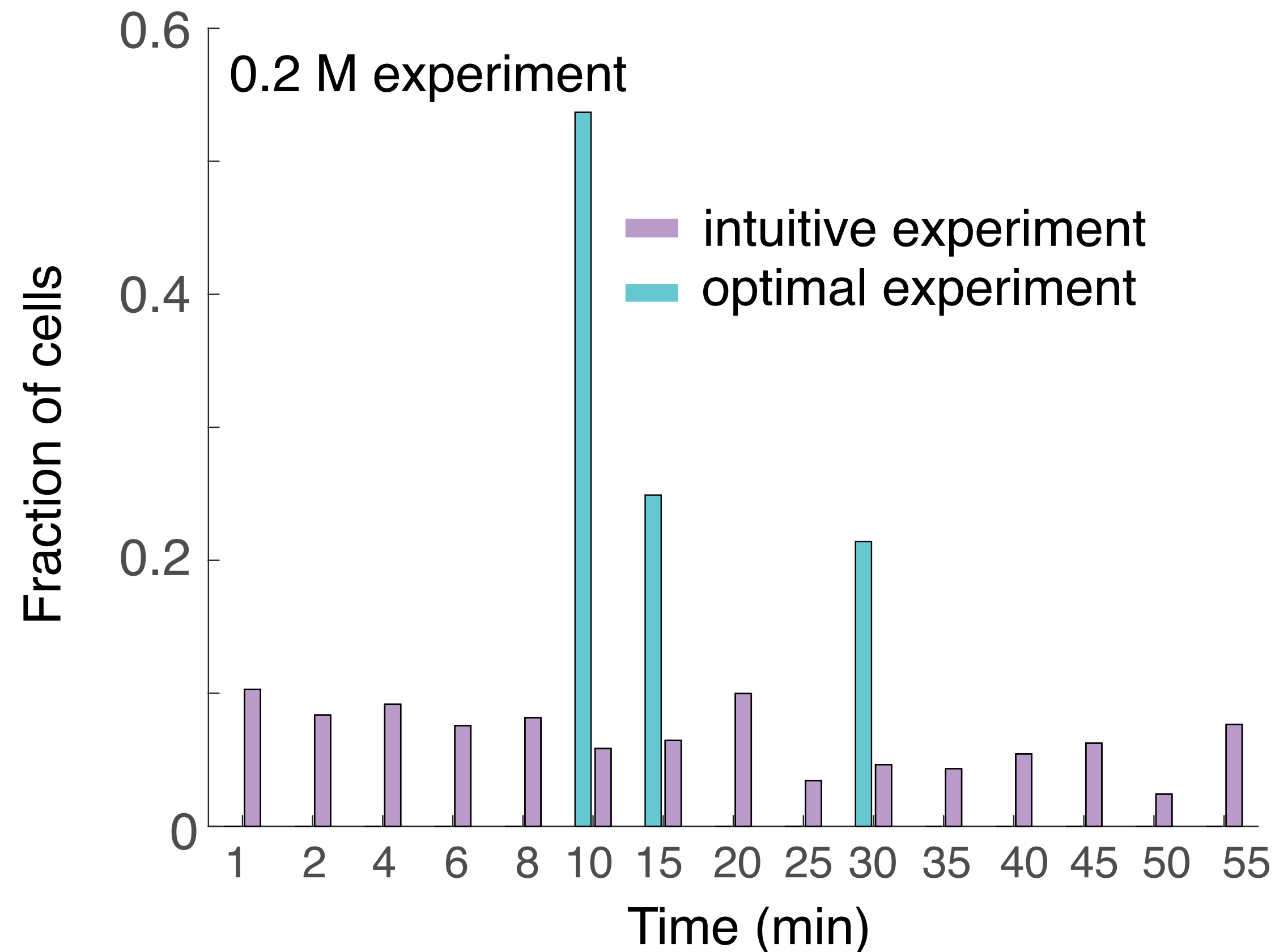
**Can we use the FSP-FIM to  
understand time-varying *controlled*  
systems?**

# Deterministic, time-varying input signal affects the stochastic model of mRNA expression.

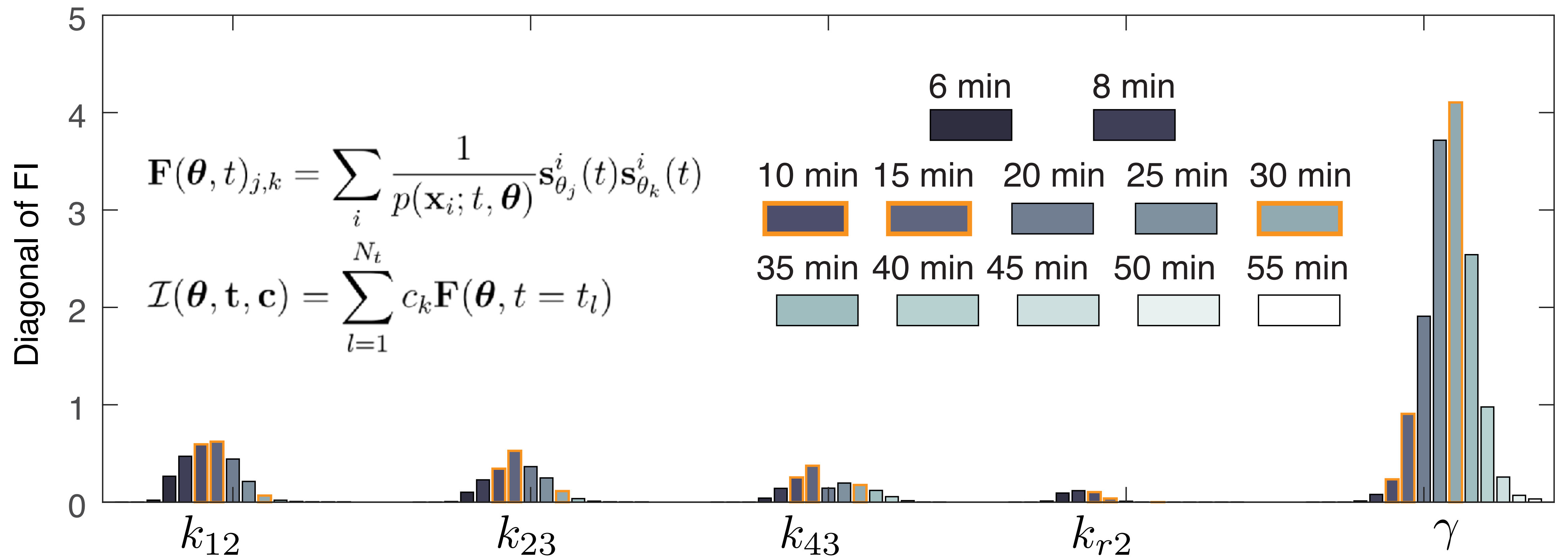


# The FSP-FIM can be used to design experiments to learn about model parameters.

$$\mathbf{F}(\boldsymbol{\theta}, t)_{j,k} = \sum_i \frac{1}{p(\mathbf{x}_i; t, \boldsymbol{\theta})} \mathbf{s}_{\theta_j}^i(t) \mathbf{s}_{\theta_k}^i(t) \quad \mathcal{I}(\boldsymbol{\theta}, \mathbf{t}, \mathbf{c}) = \sum_{l=1}^{N_t} c_k \mathbf{F}(\boldsymbol{\theta}, t = t_l)$$

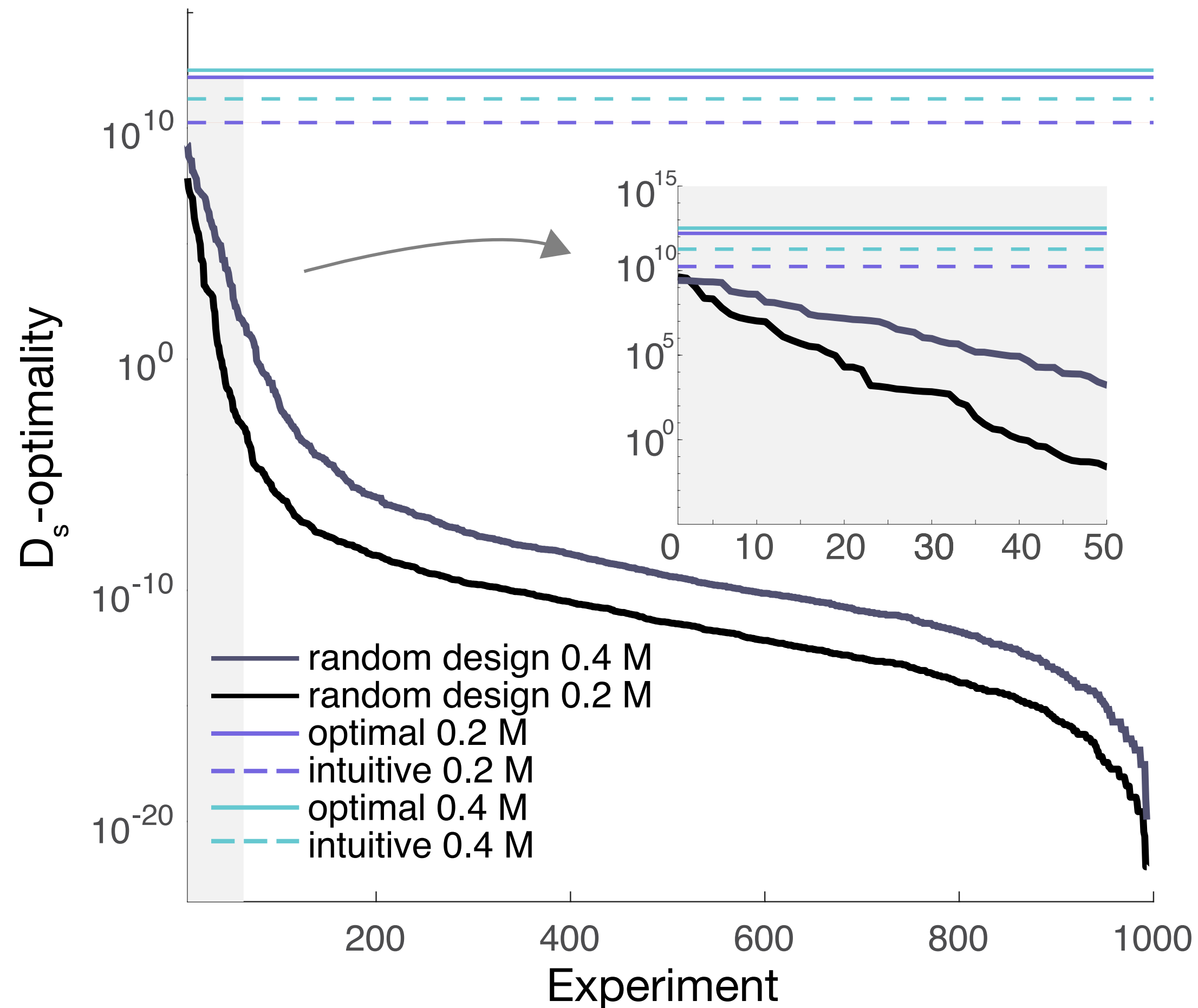


# Fisher information analysis reveals which dynamics are informative.





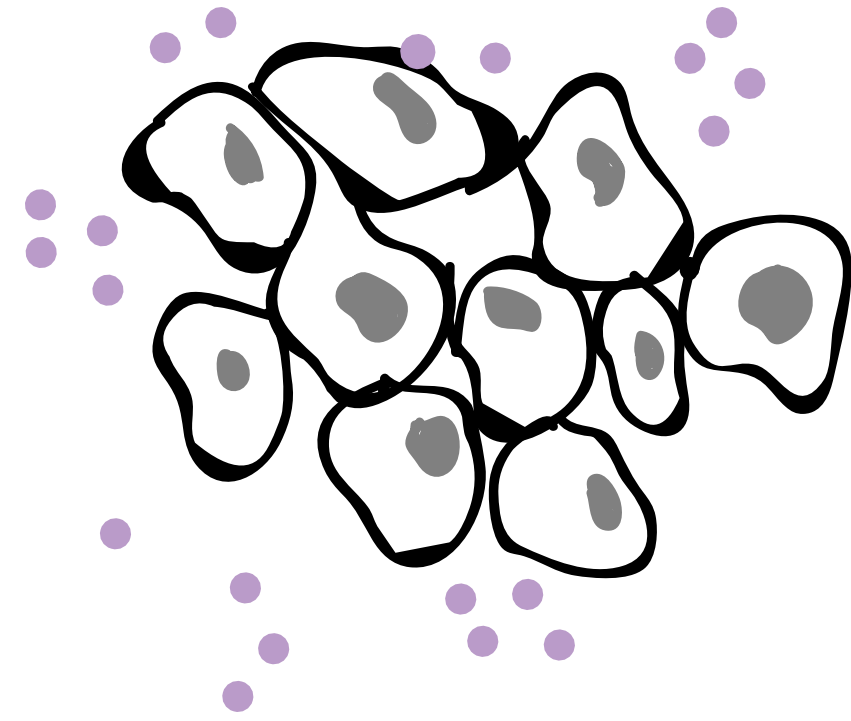
# Optimal experiment designs are better than random experiment designs.



	FIM 0.2 M	FIM 0.4 M
optimal 0.2 M	12.2	12.3
optimal 0.4 M	12.0	12.5
intuitive 0.2 M	10.2	11.1
intuitive 0.4 M	10.6	11.3

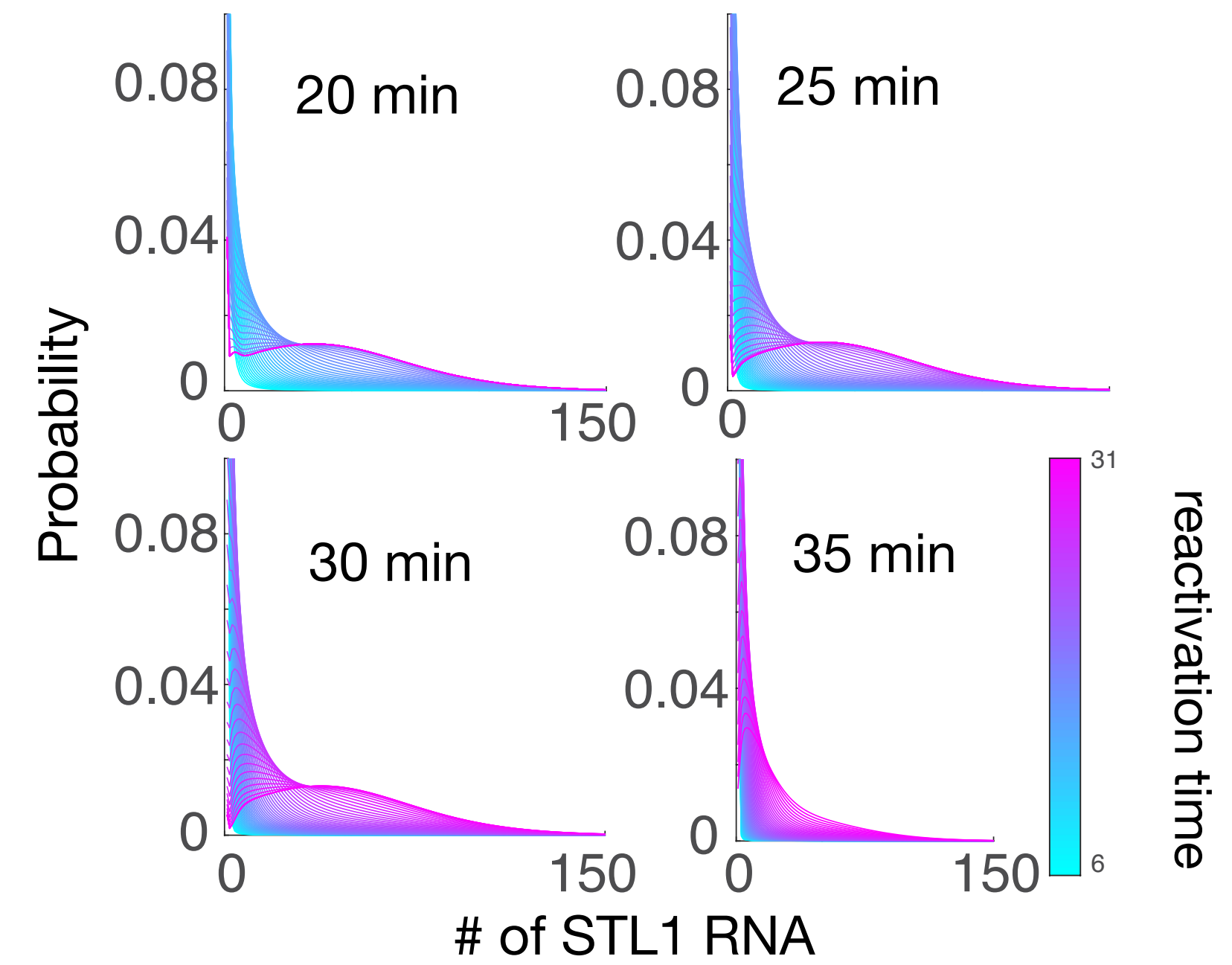
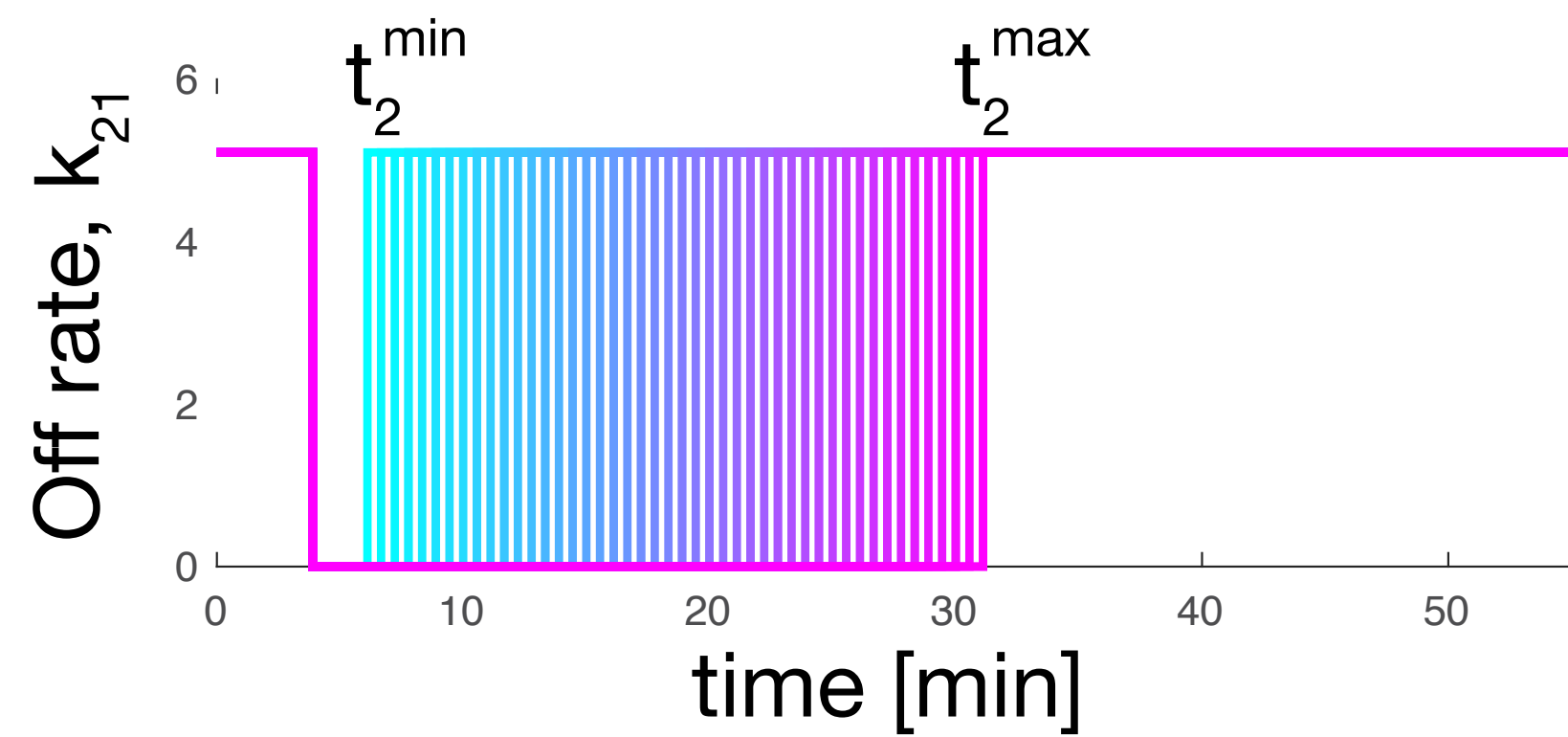
Color scale for  $\log_{10}(D_s\text{-optimality})$ : 10.5 to 12.5

# The FSP-FIM can also be used to optimize measurements to learn about the environment.

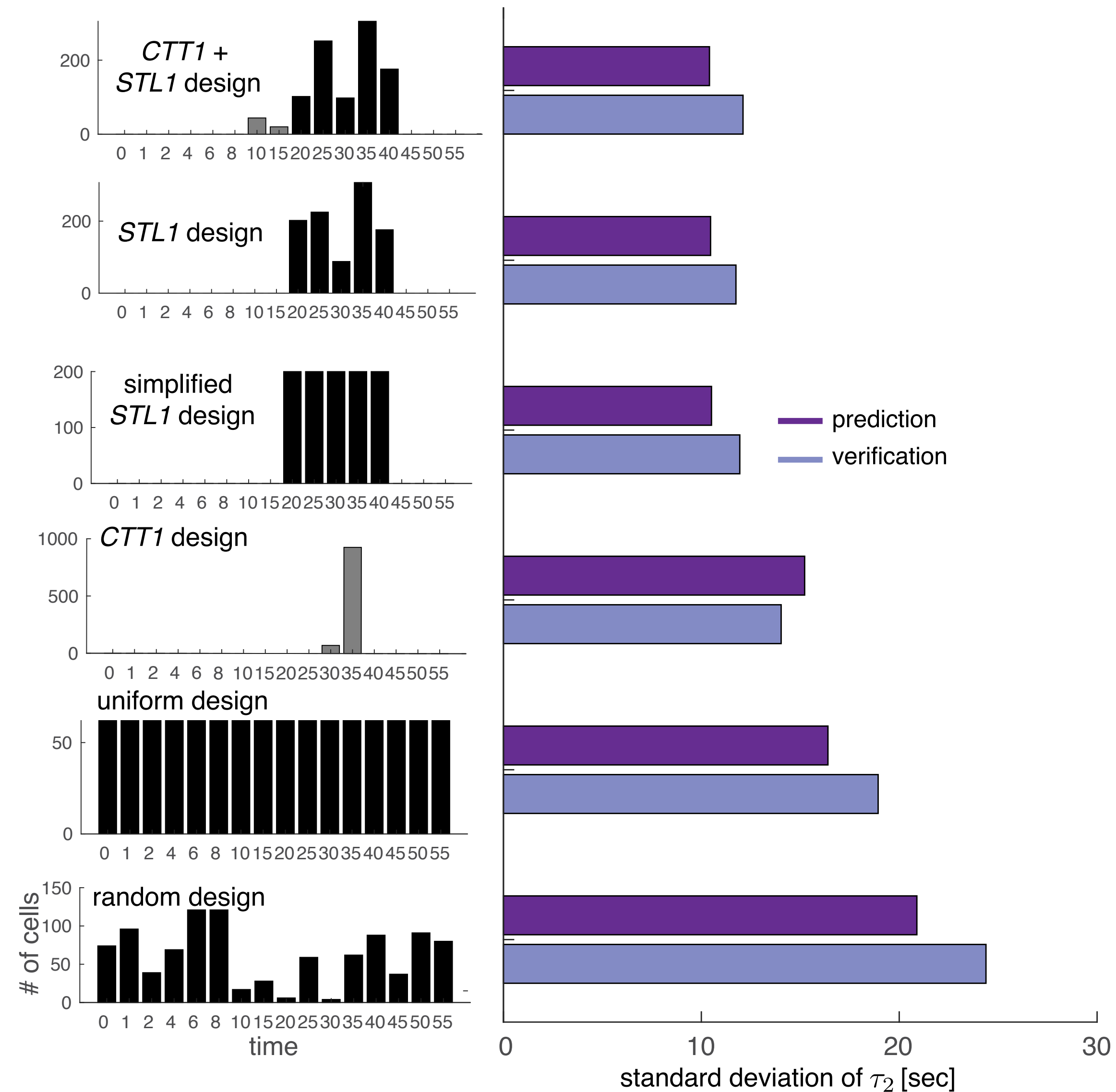


$$k_{21}(t) = \max[0, \alpha - \beta f(t)]$$

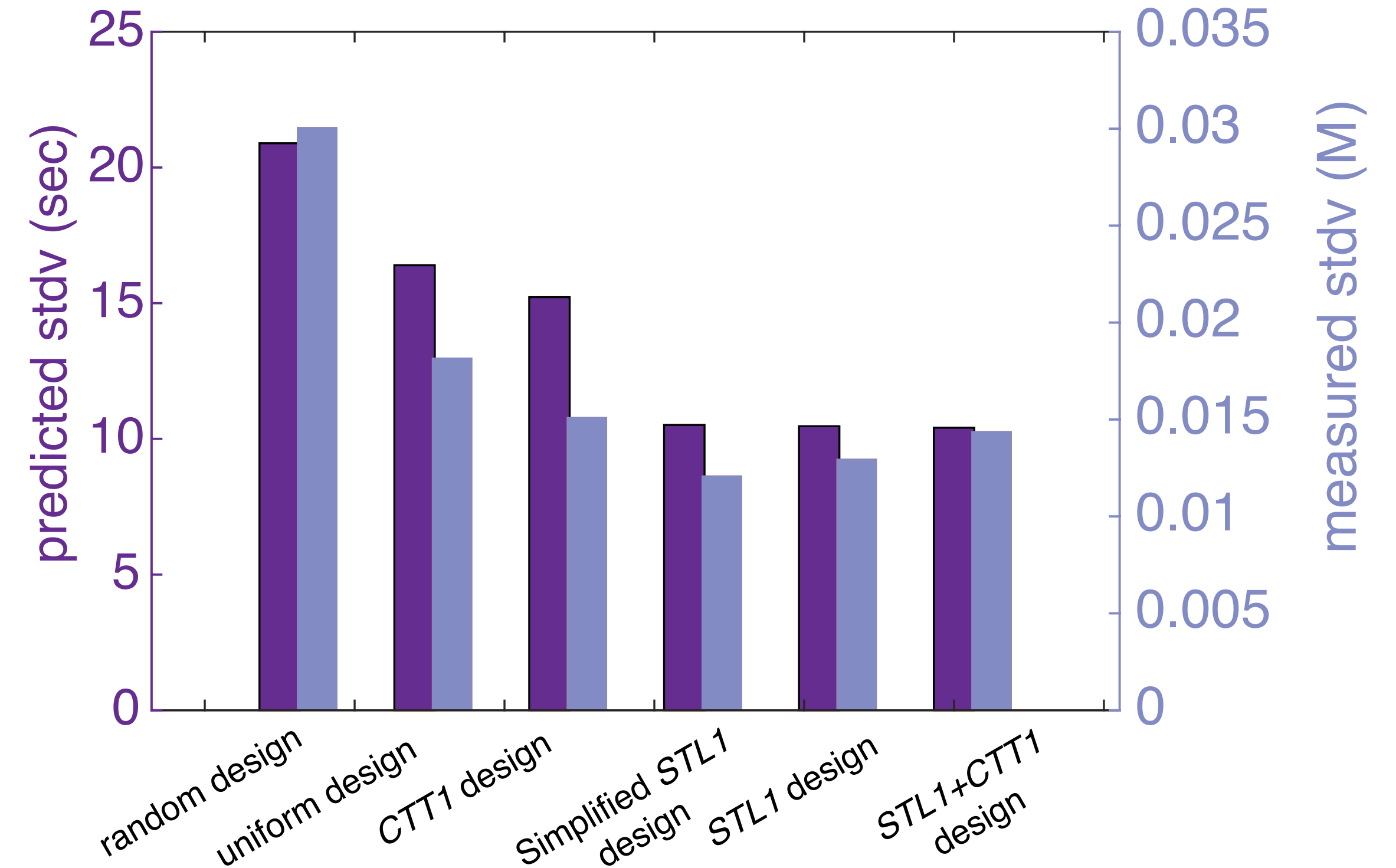
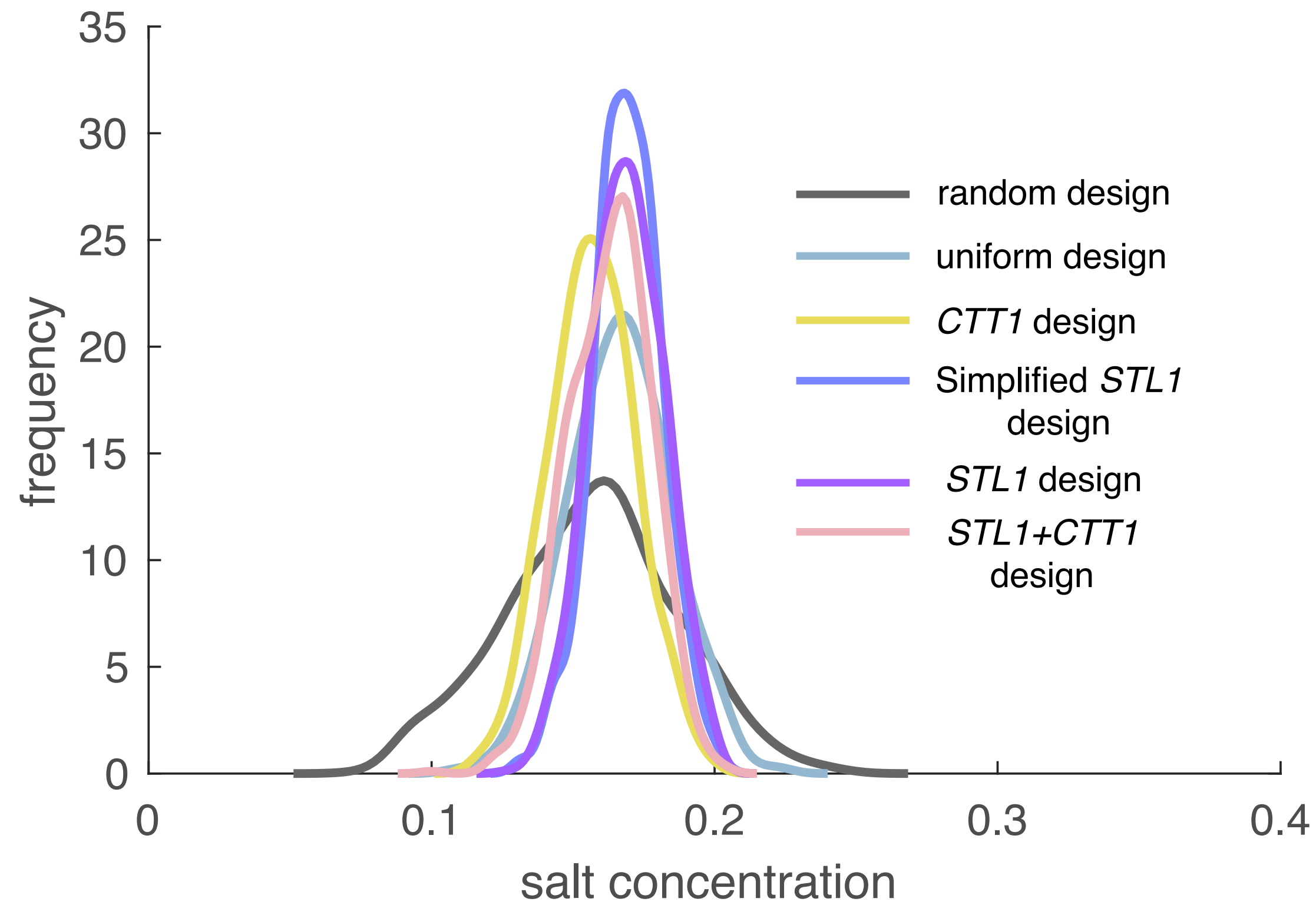
$$\mathbf{c}_{\text{opt}} = \arg \min_{\mathbf{c}, \sum c_i = 1} \int_{t_2^{\min}}^{t_2^{\max}} \frac{1}{t_2^{\max} - t_2^{\min}} \mathcal{I}^{-1}(\mathbf{c}; t_2 = t, \boldsymbol{\theta}) dt$$



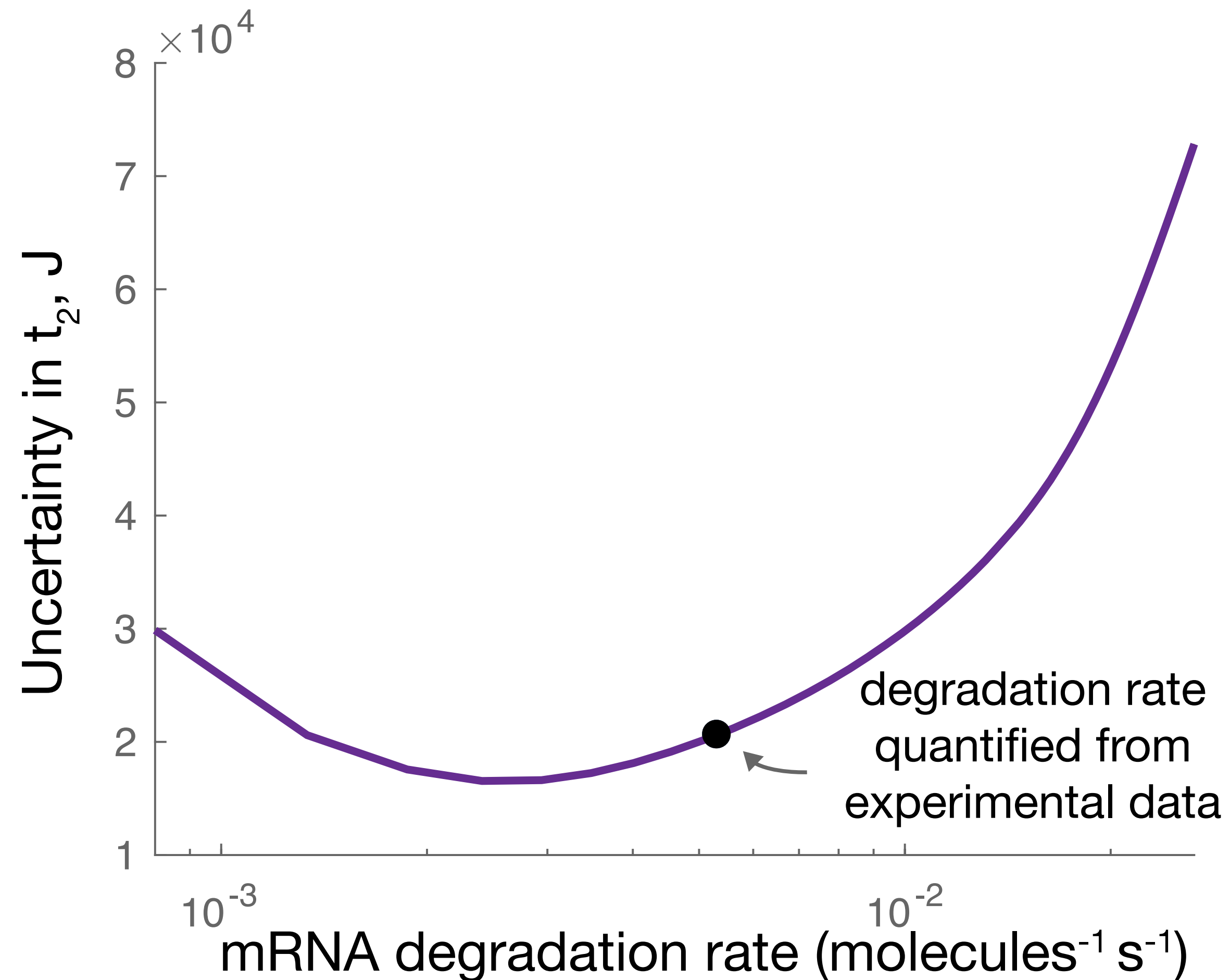
# Using different experiment designs to learn deactivation times.



# Experimental verification for biosensor experiments.

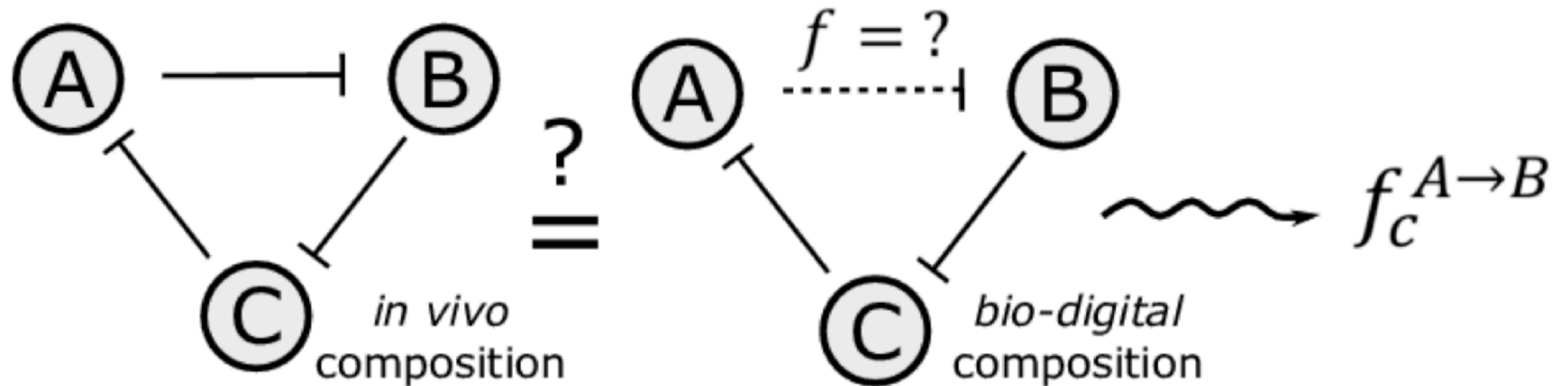


This method can also be used to find the degradation rate of mRNA which reduces uncertainty in deactivation time.

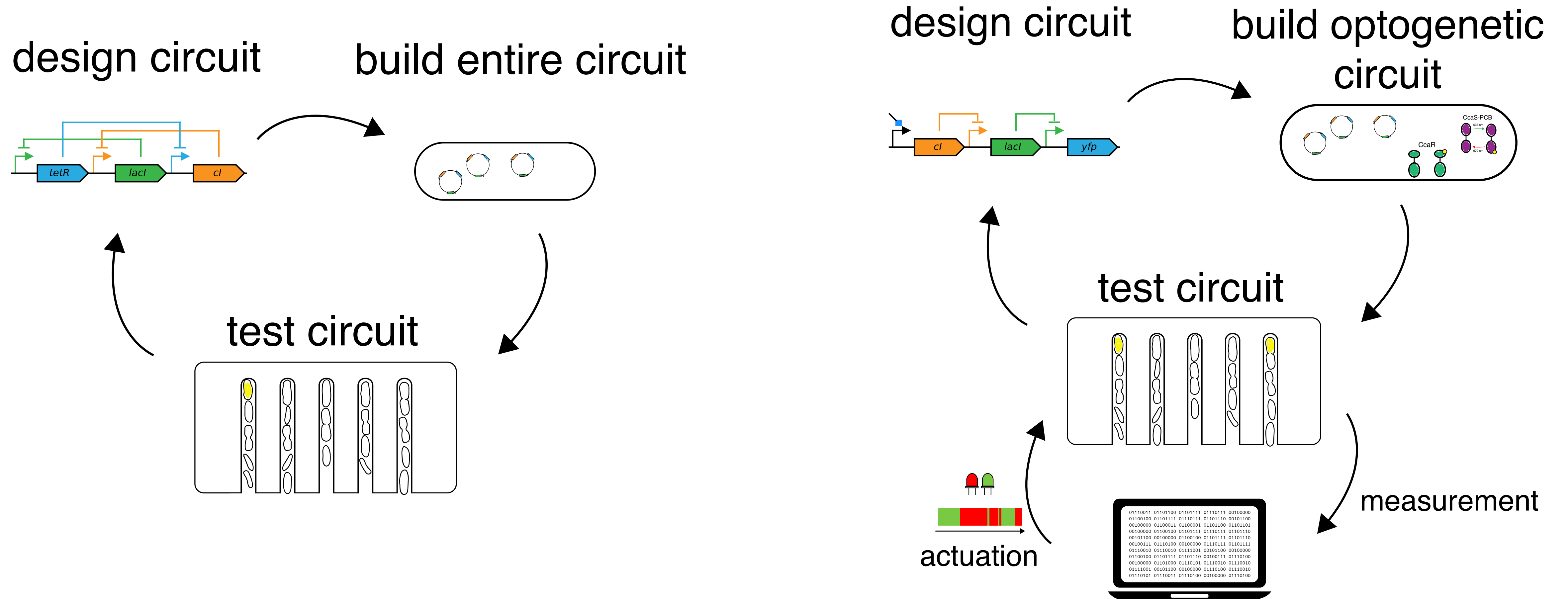


**What can we learn about model  
parameters from timing  
distributions?**

# Composability is a primary challenge in constructing synthetic biological circuits

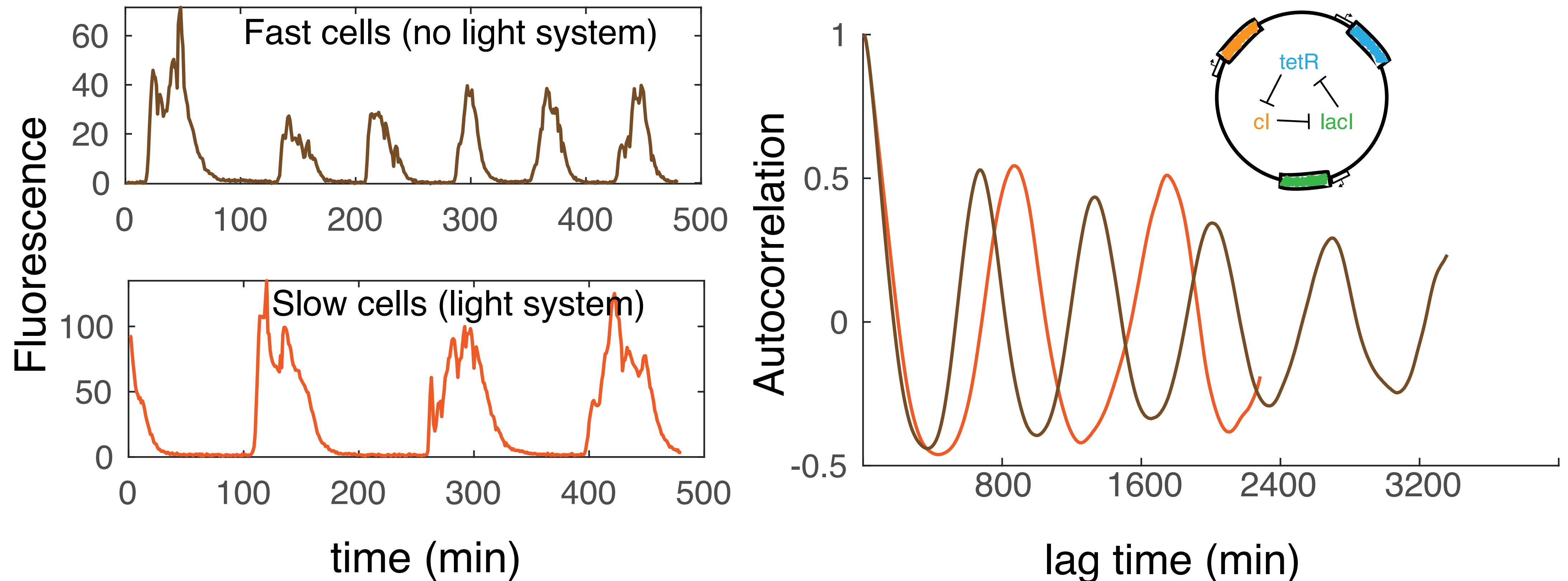


# Optogenetics could be used to rapidly prototype circuit designs

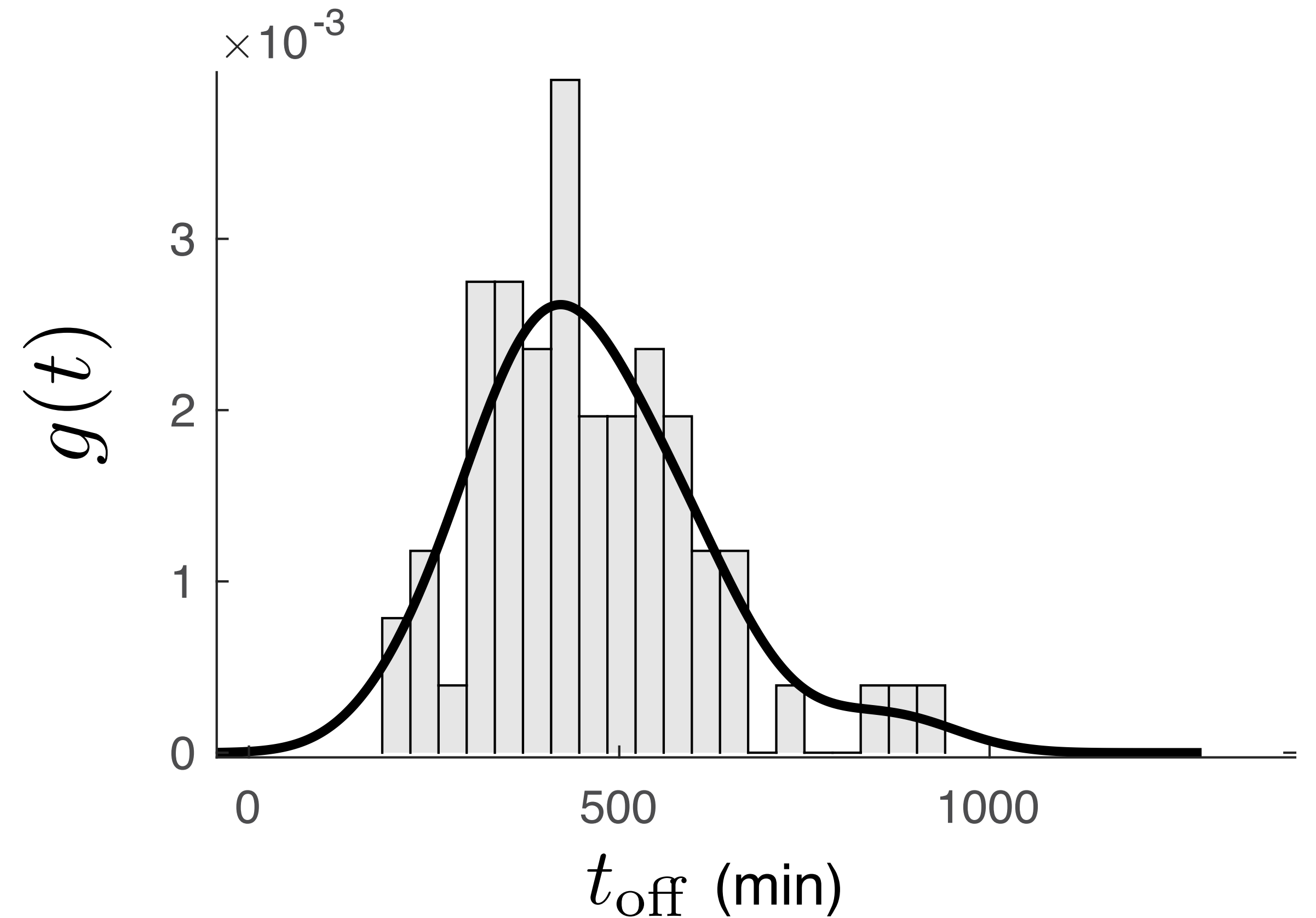
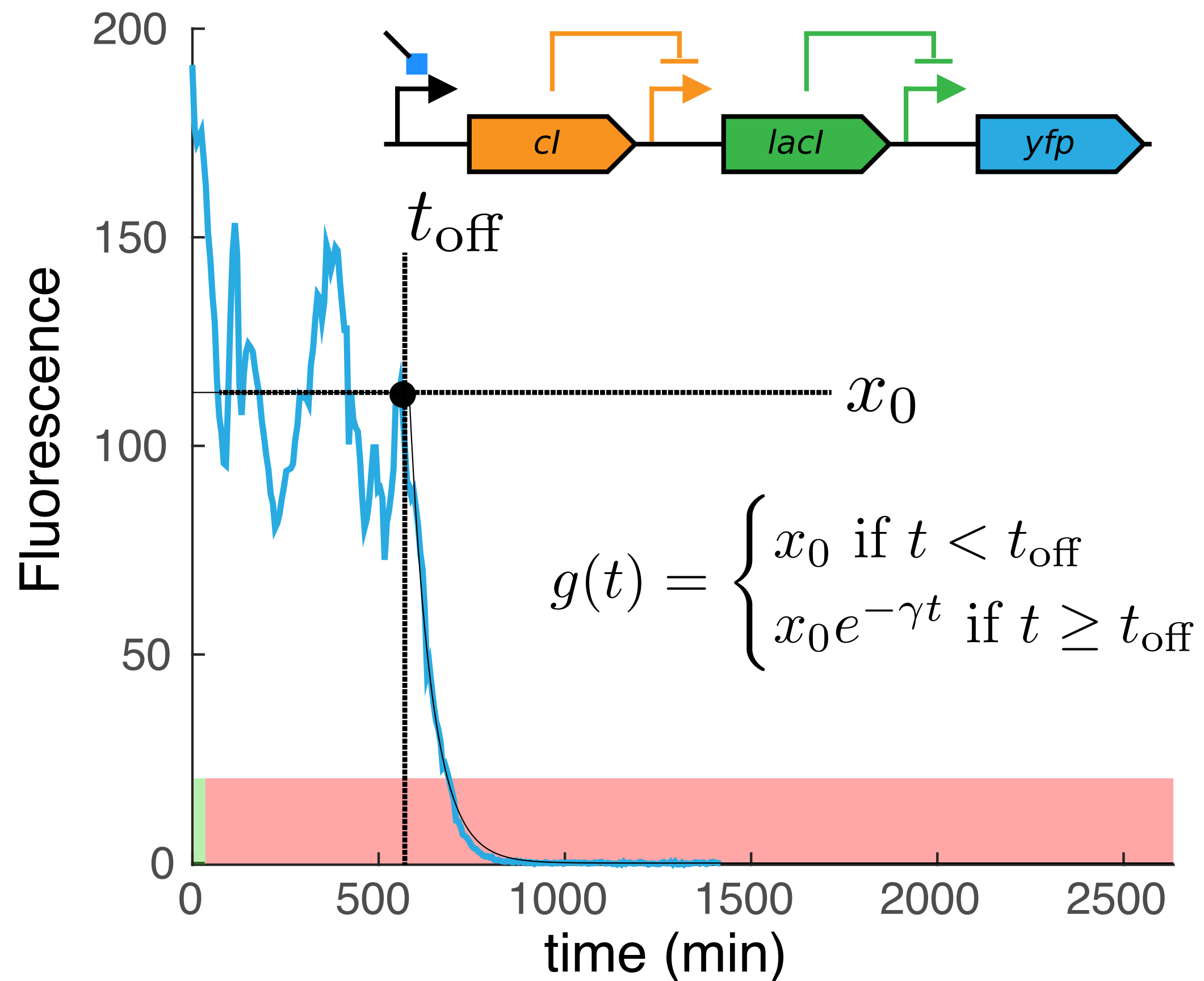




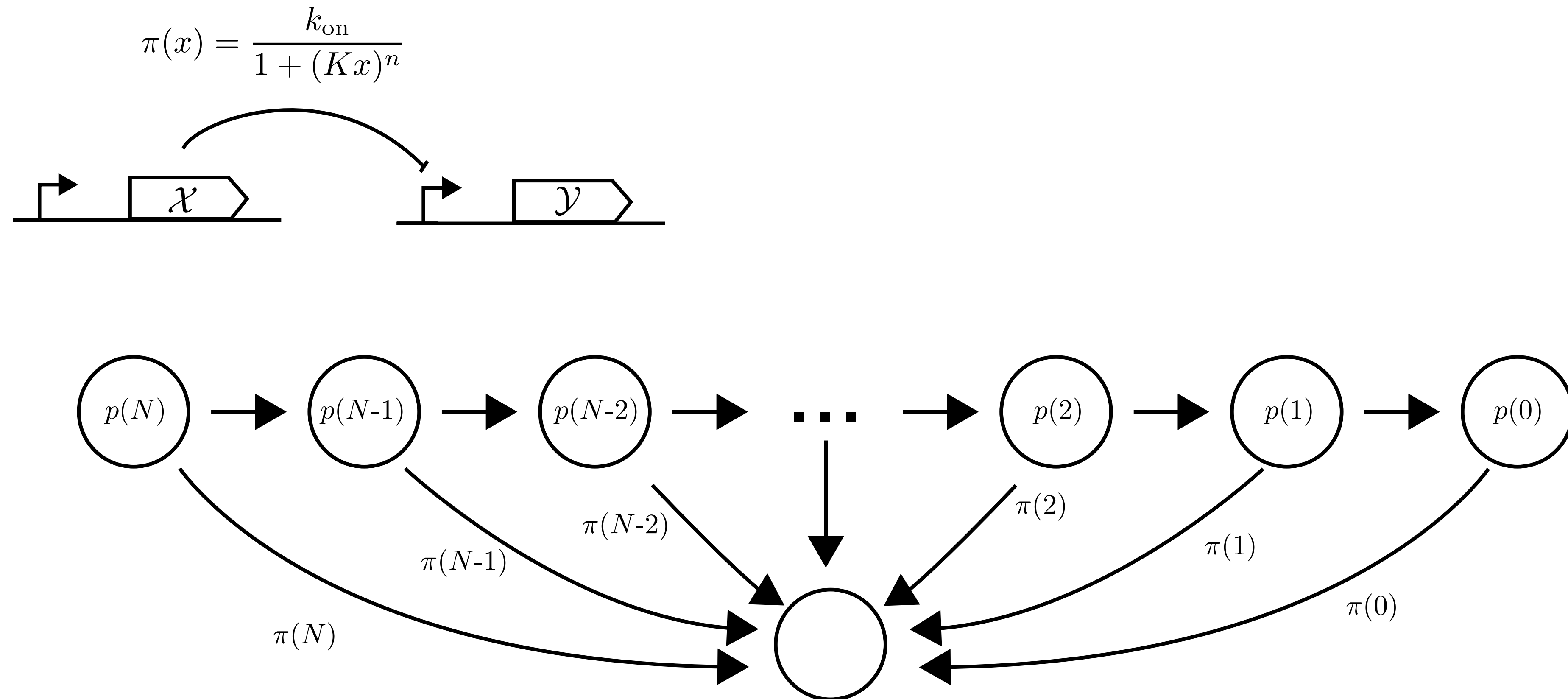
# A repressilator with light controlled system in the background oscillates regularly



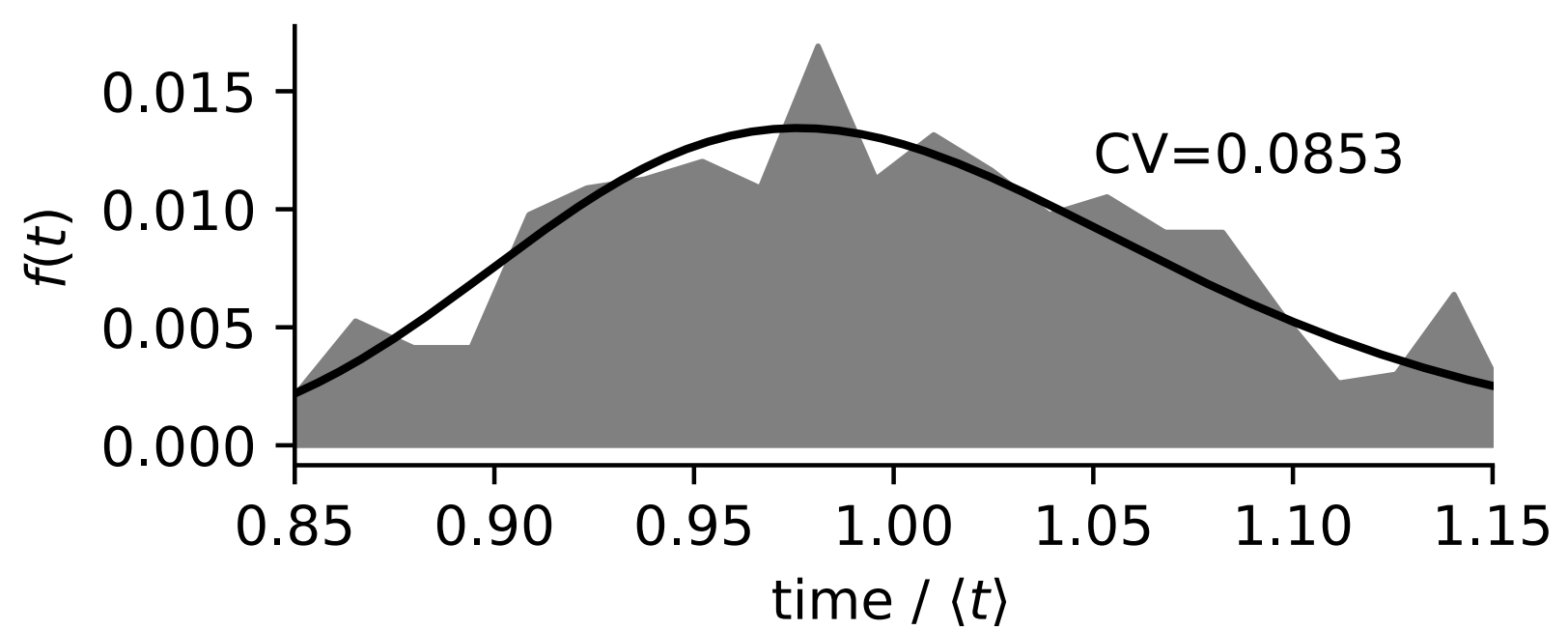
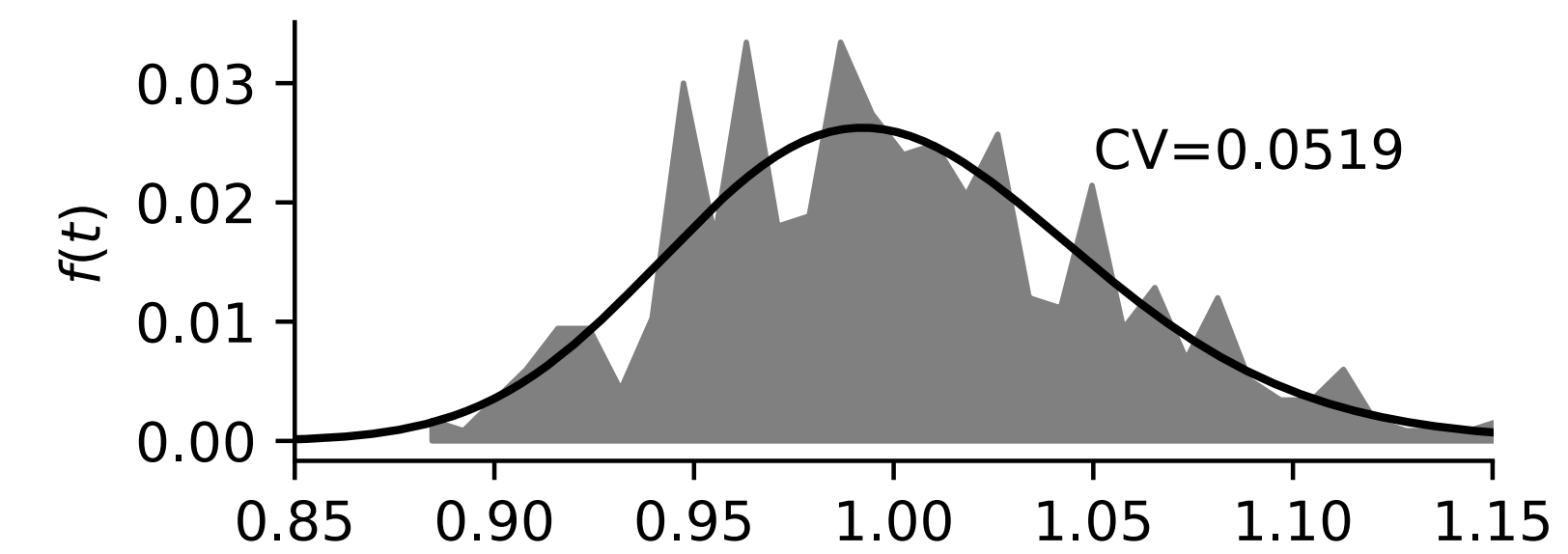
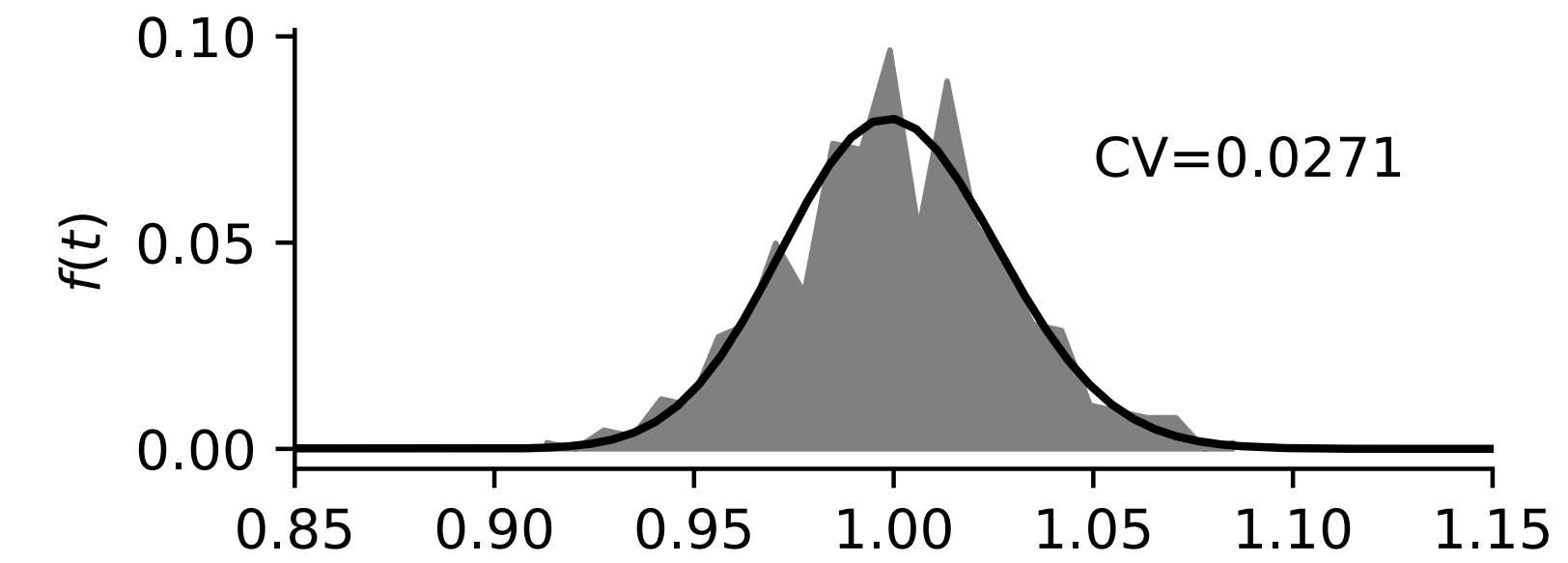
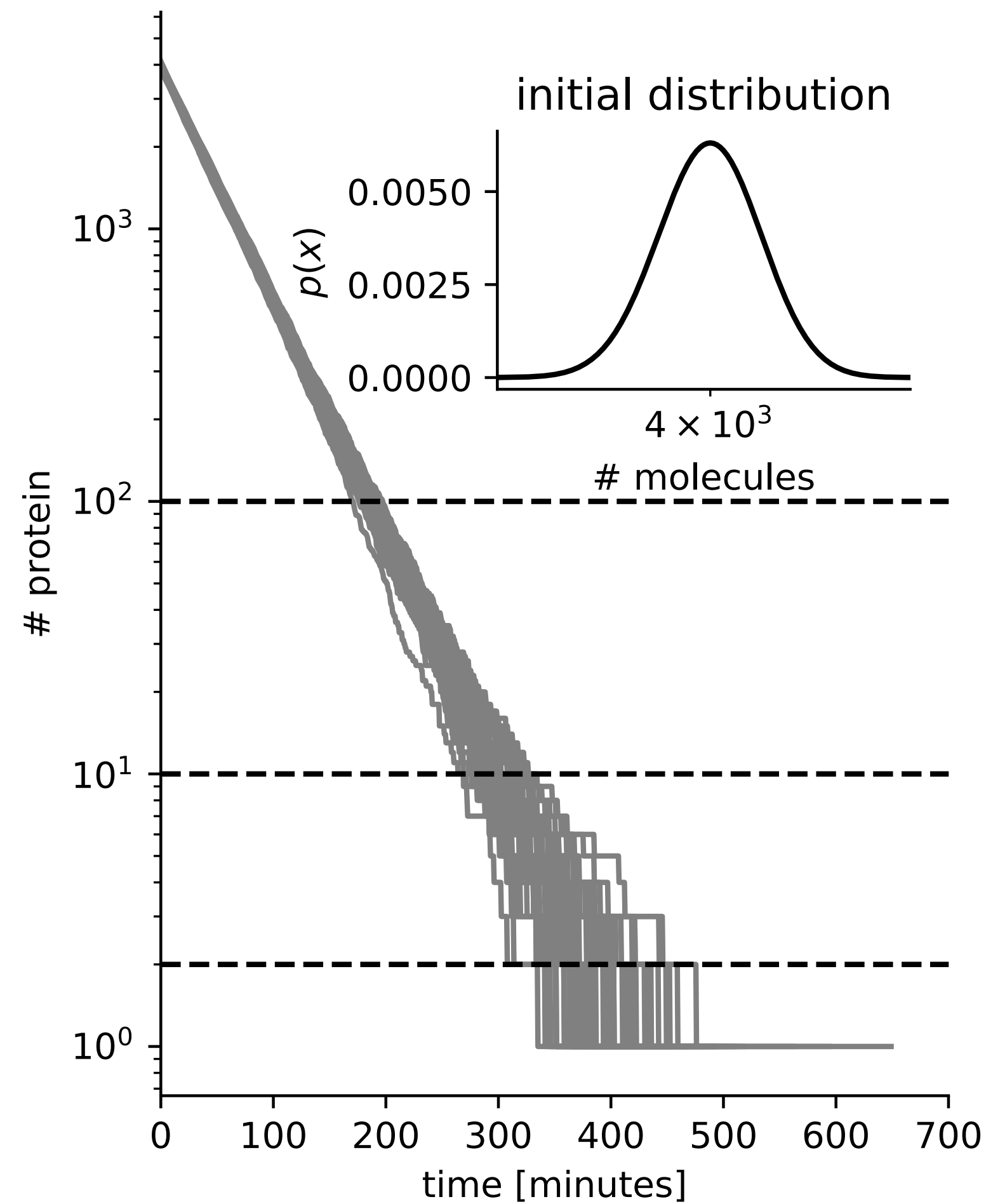
# Cutting the circuit can help quantify individual elements



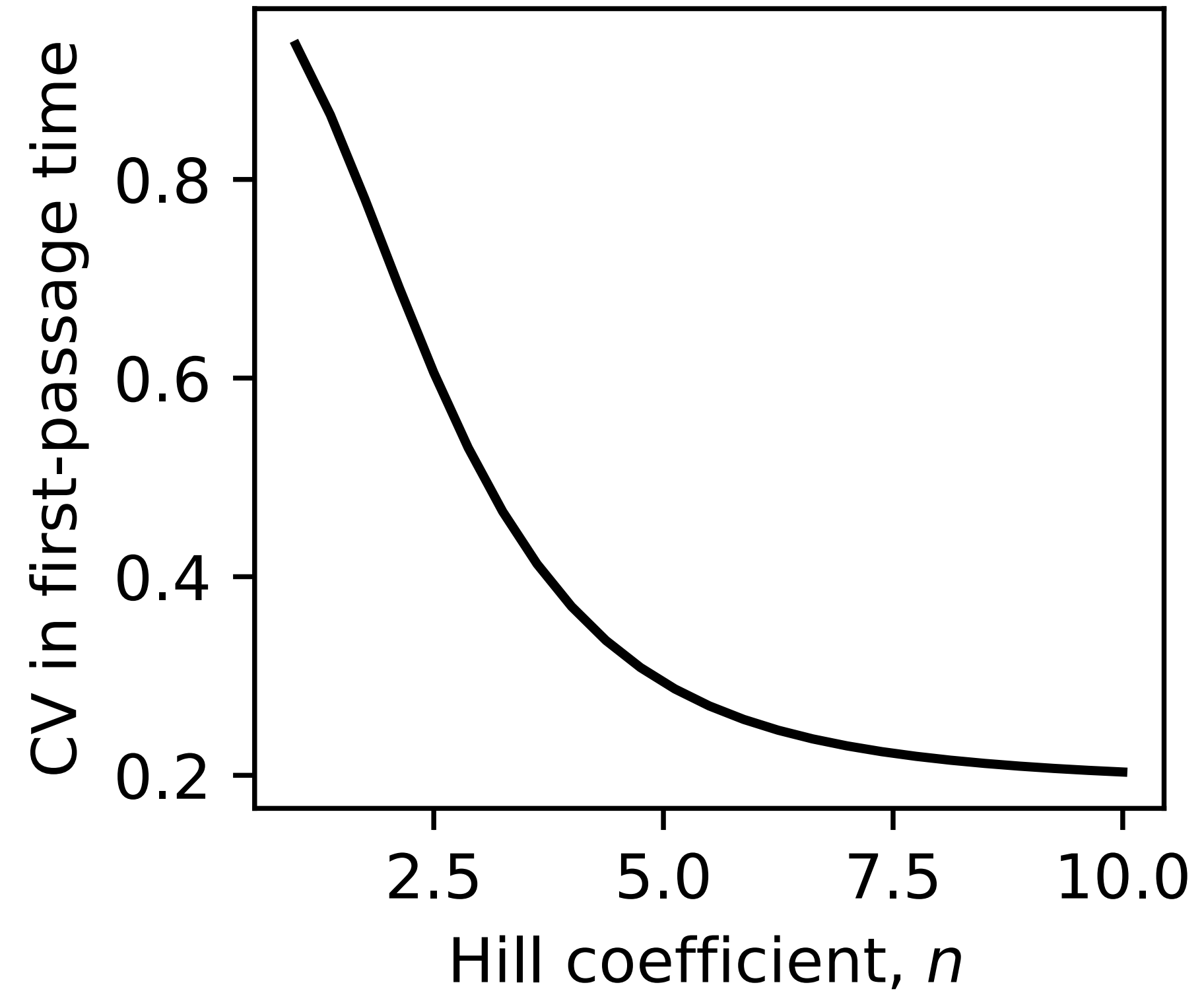
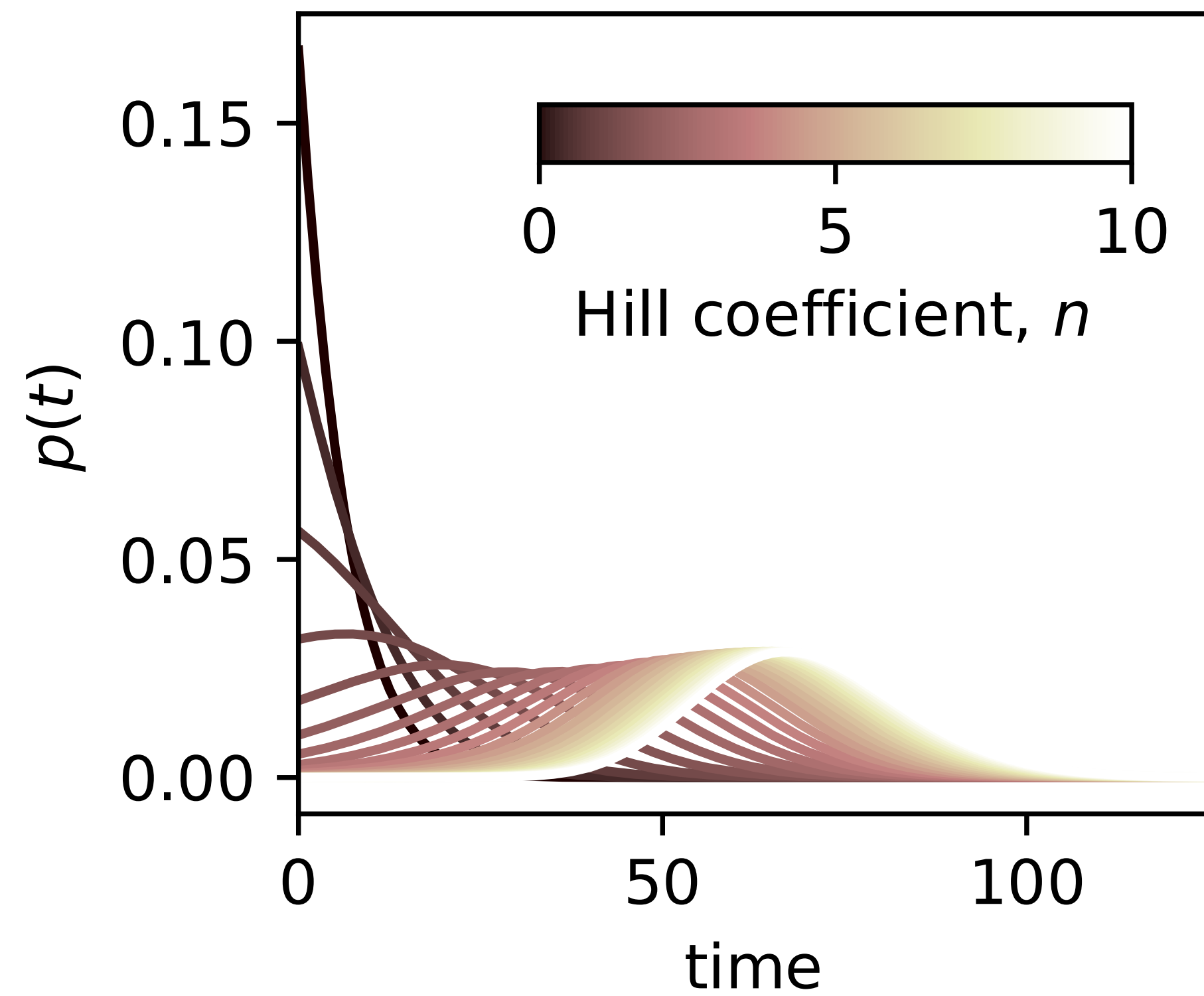
# Simple stochastic model of derepression of a promoter.



# Different repression thresholds have different statistics.



# First-passage times are sensitive to Hill function parameters



# Fisher information for first passage time distributions

$$f(t) = -\mathbf{1}^T \mathbf{A} \exp(\mathbf{A}t) \mathbf{p}_0,$$

$$\mathcal{I} = \mathbb{E} \left[ \nabla_{\boldsymbol{\theta}} \log f(t; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log f(t; \boldsymbol{\theta})^T \right]$$

$$\mathcal{I}_{i,j} = \int_T \frac{1}{f(t; \boldsymbol{\theta})} f_{\theta_i}(t; \boldsymbol{\theta}) f_{\theta_j}(t; \boldsymbol{\theta}) dt$$

$$f_{\theta_i} = -\mathbf{1}^T \left[ \mathbf{A}_{JJ}^{\theta_i} \mathbf{p}(t) + \mathbf{A} \mathbf{s}_{\theta_i}(t) \right]$$

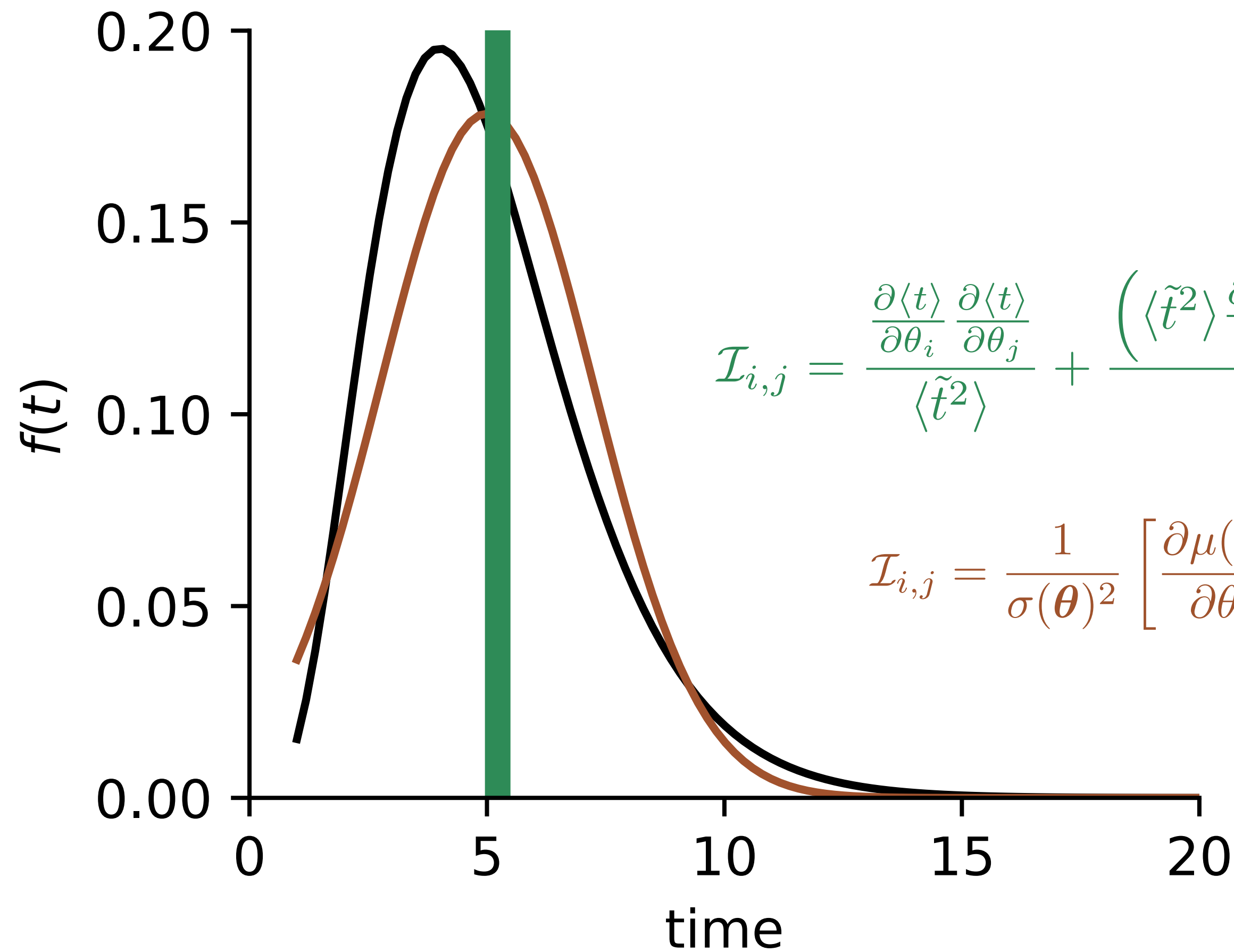
For the Markov chain model, the first passage time distribution is a phase-type distribution.

The Fisher information matrix is defined as the expected (w.r.t. **data**) first derivative of the log-likelihood function.

For phase type distributions, the FIM can be found as an integral over the sensitivities of  $f(t)$  normalized by their probabilities.

The sensitivities can be found from the sensitivities of the original Markov chain.

# Three Fisher informations for three approximations of the likelihood.

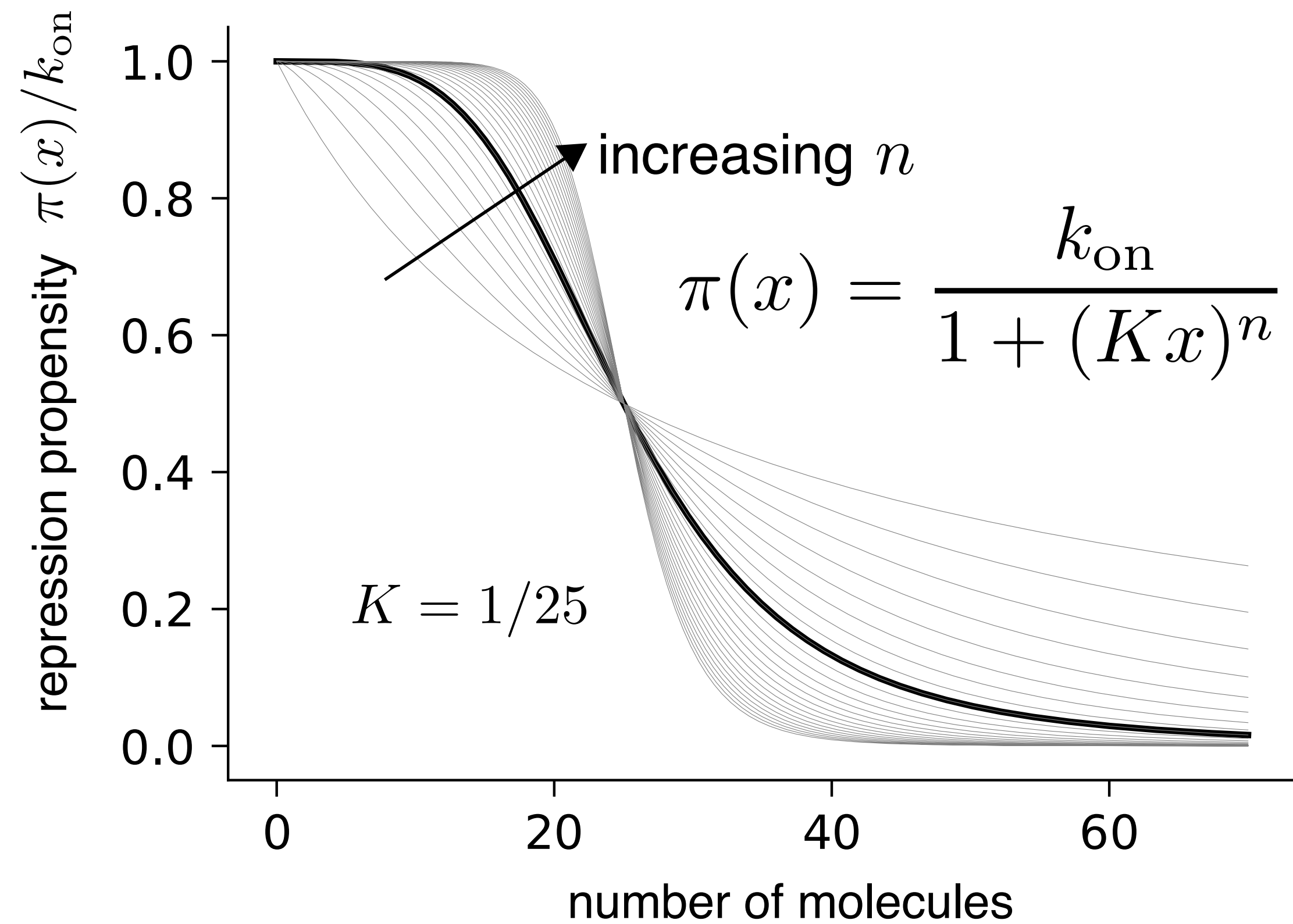
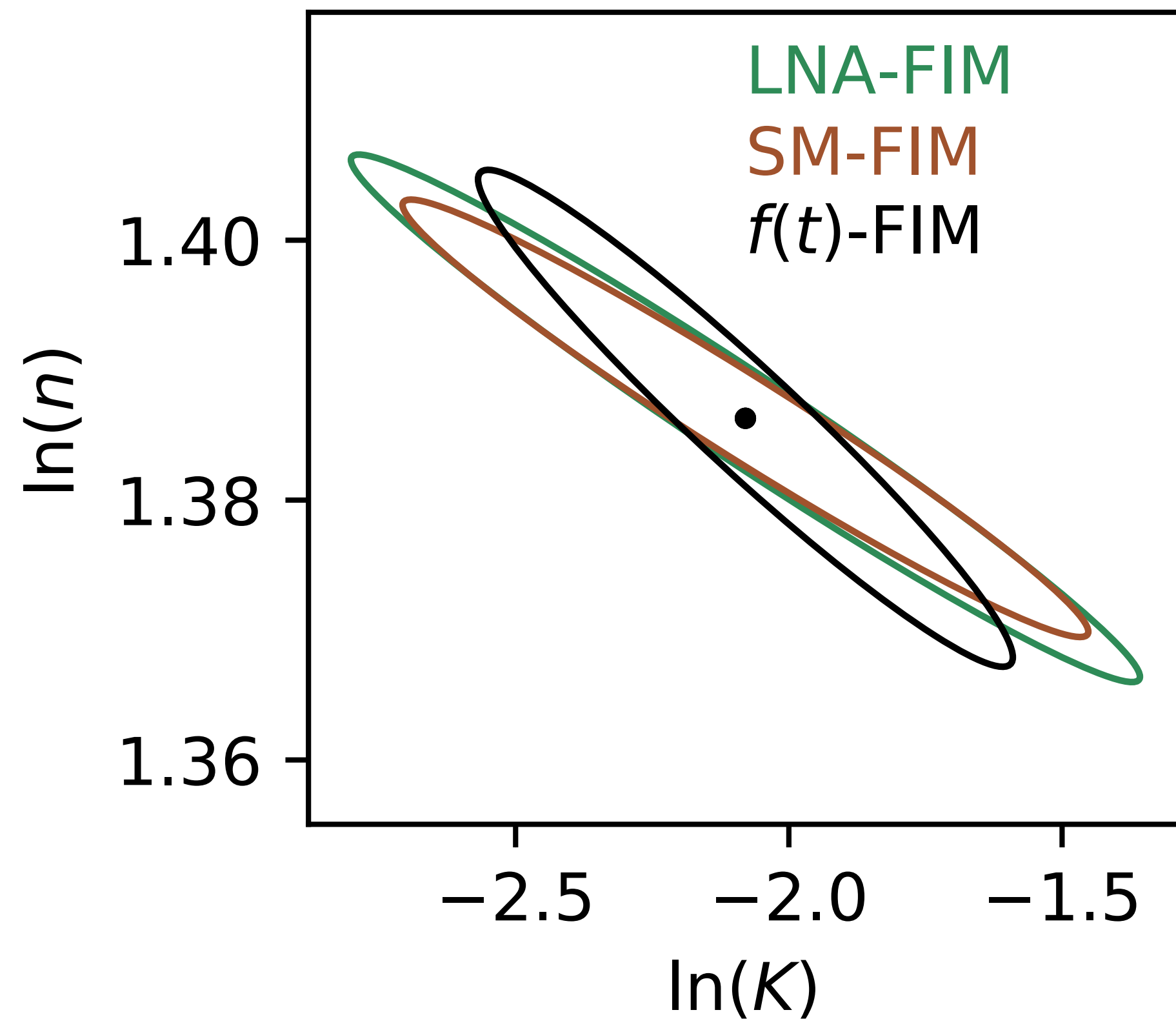


$$\mathcal{I}_{i,j} = \int_T \frac{1}{f(t; \boldsymbol{\theta})} f_{\theta_i}(t; \boldsymbol{\theta}) f_{\theta_j}(t; \boldsymbol{\theta}) dt$$

$$\mathcal{I}_{i,j} = \frac{\frac{\partial \langle t \rangle}{\partial \theta_i} \frac{\partial \langle t \rangle}{\partial \theta_j}}{\langle \tilde{t}^2 \rangle} + \frac{\left( \langle \tilde{t}^2 \rangle \frac{\partial \langle \tilde{t}^2 \rangle}{\partial \theta_i} - \frac{\partial \langle t \rangle}{\partial \theta_i} \langle \tilde{t}^3 \rangle \right) \left( \langle \tilde{t}^2 \rangle \frac{\partial \langle \tilde{t}^2 \rangle}{\partial \theta_j} - \frac{\partial \langle t \rangle}{\partial \theta_j} \langle \tilde{t}^3 \rangle \right)}{\langle \tilde{t}^2 \rangle^2 (\langle \tilde{t}^4 \rangle - \langle \tilde{t}^2 \rangle^2) - \langle \tilde{t}^2 \rangle \langle \tilde{t}^3 \rangle^2} + \mathcal{O}(1)$$

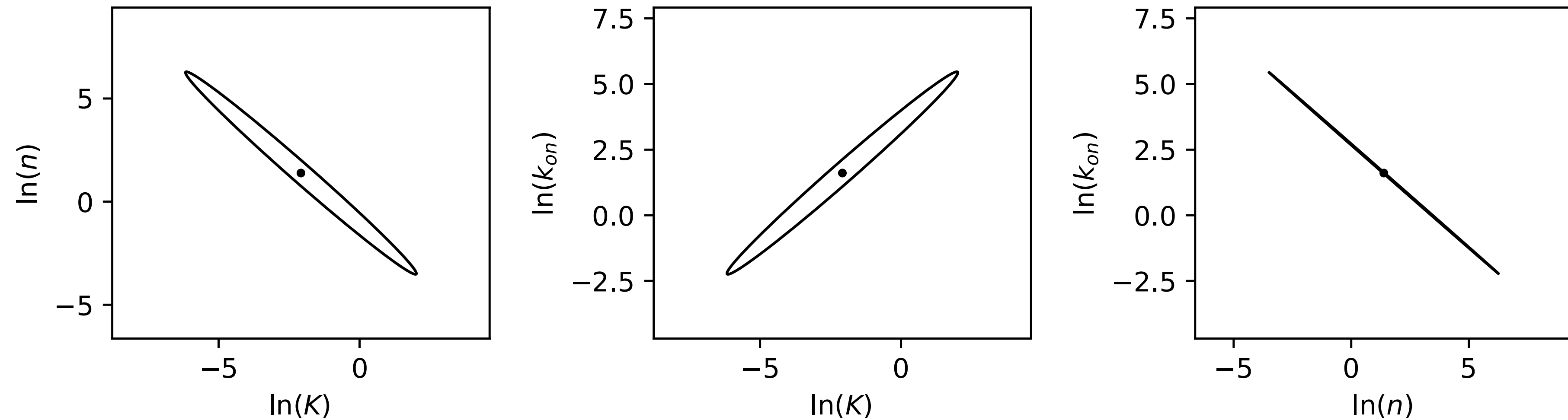
$$\mathcal{I}_{i,j} = \frac{1}{\sigma(\boldsymbol{\theta})^2} \left[ \frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_j} + \frac{1}{2} \left( \frac{1}{(\sigma^2(\boldsymbol{\theta}))^2} \frac{\partial \sigma^2(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \sigma^2(\boldsymbol{\theta})}{\partial \theta_j} \right) \right]$$

# Some Hill function parameters can be learned from single cell data.





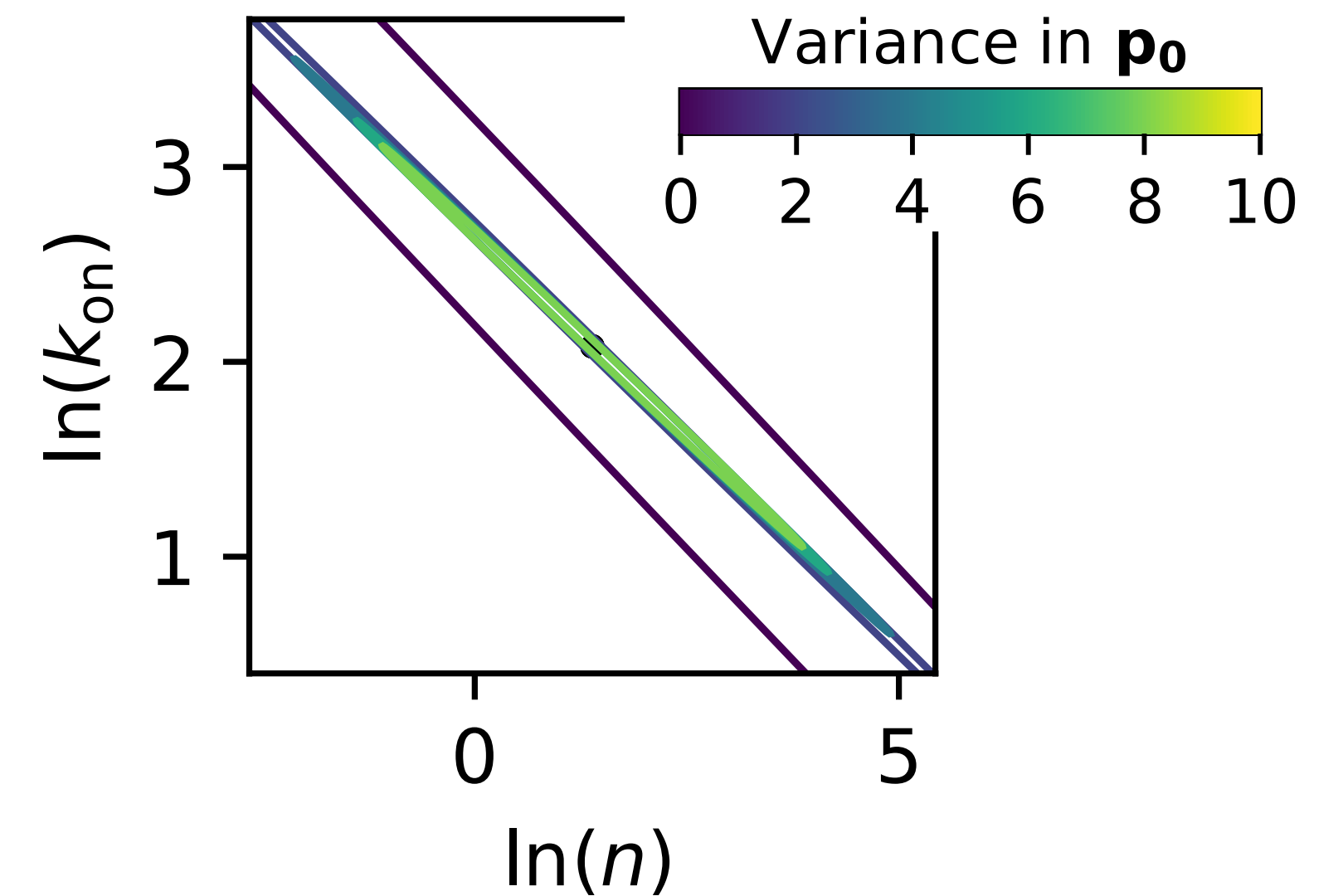
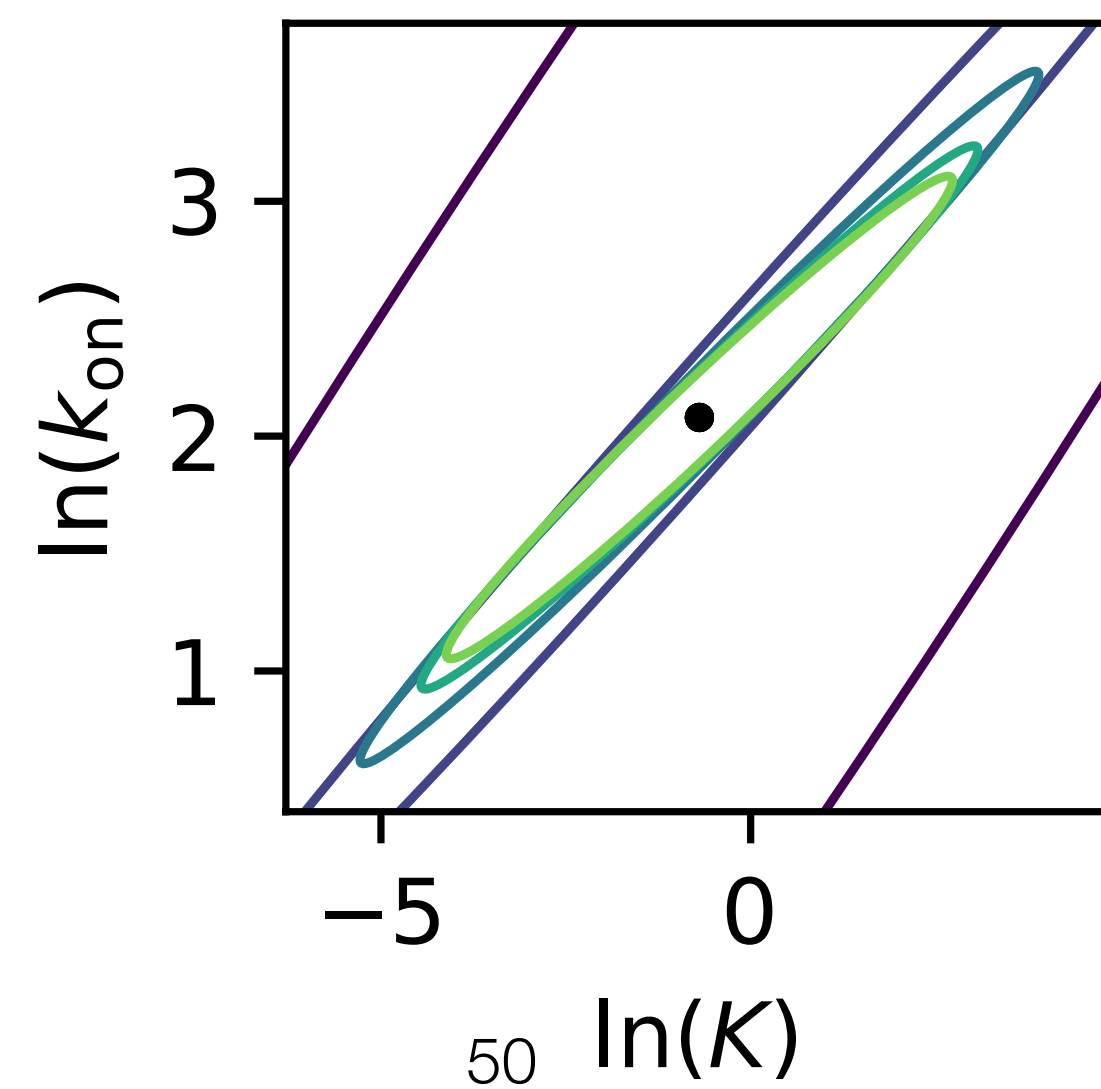
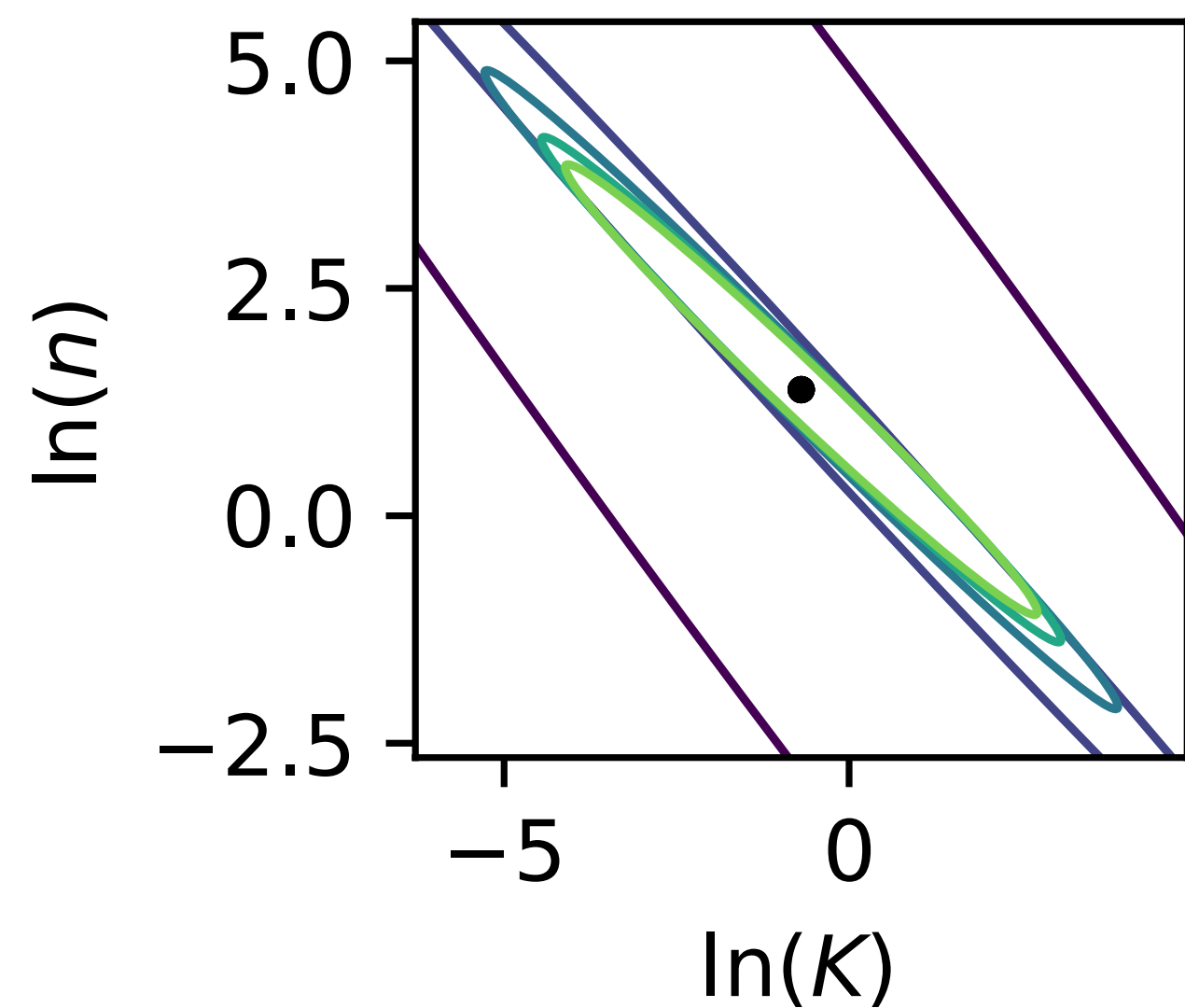
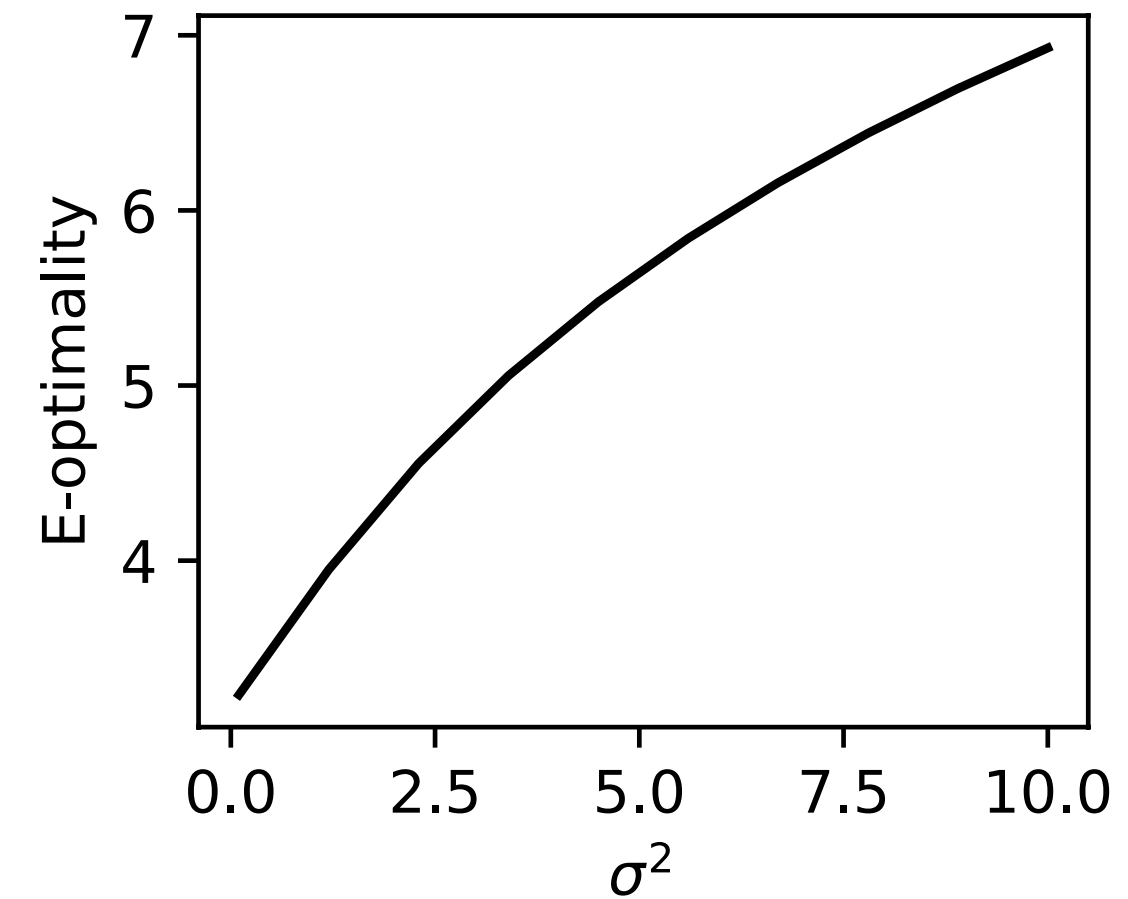
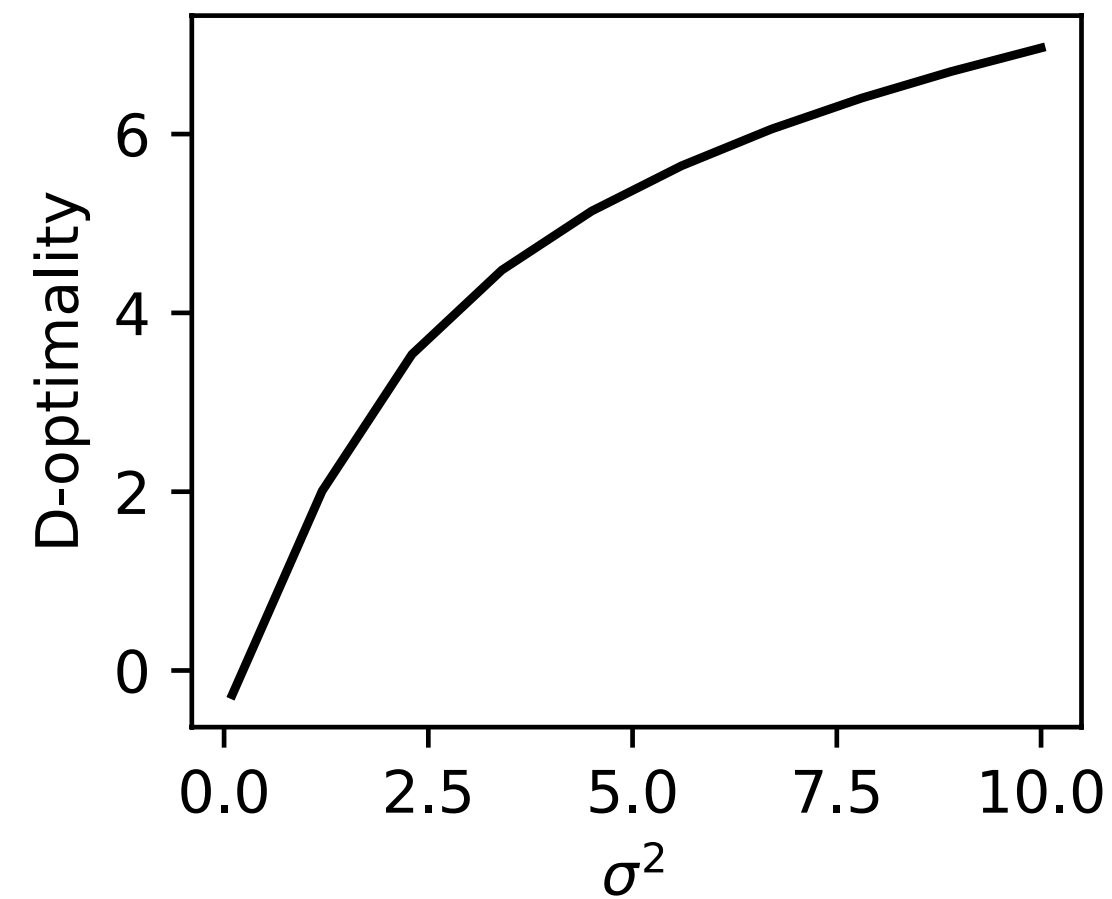
**However, when all three Hill function parameters are free, they can be difficult to identify.**



**Is there a better experiment to learn these parameters?**

# Parameter estimation improves with increased variance in the initial distribution of molecule numbers.

$$\pi(x) = \frac{k_{\text{on}}}{1 + (Kx)^n}$$



# What information about the initial distributions of molecule numbers are contained in the first passage times?

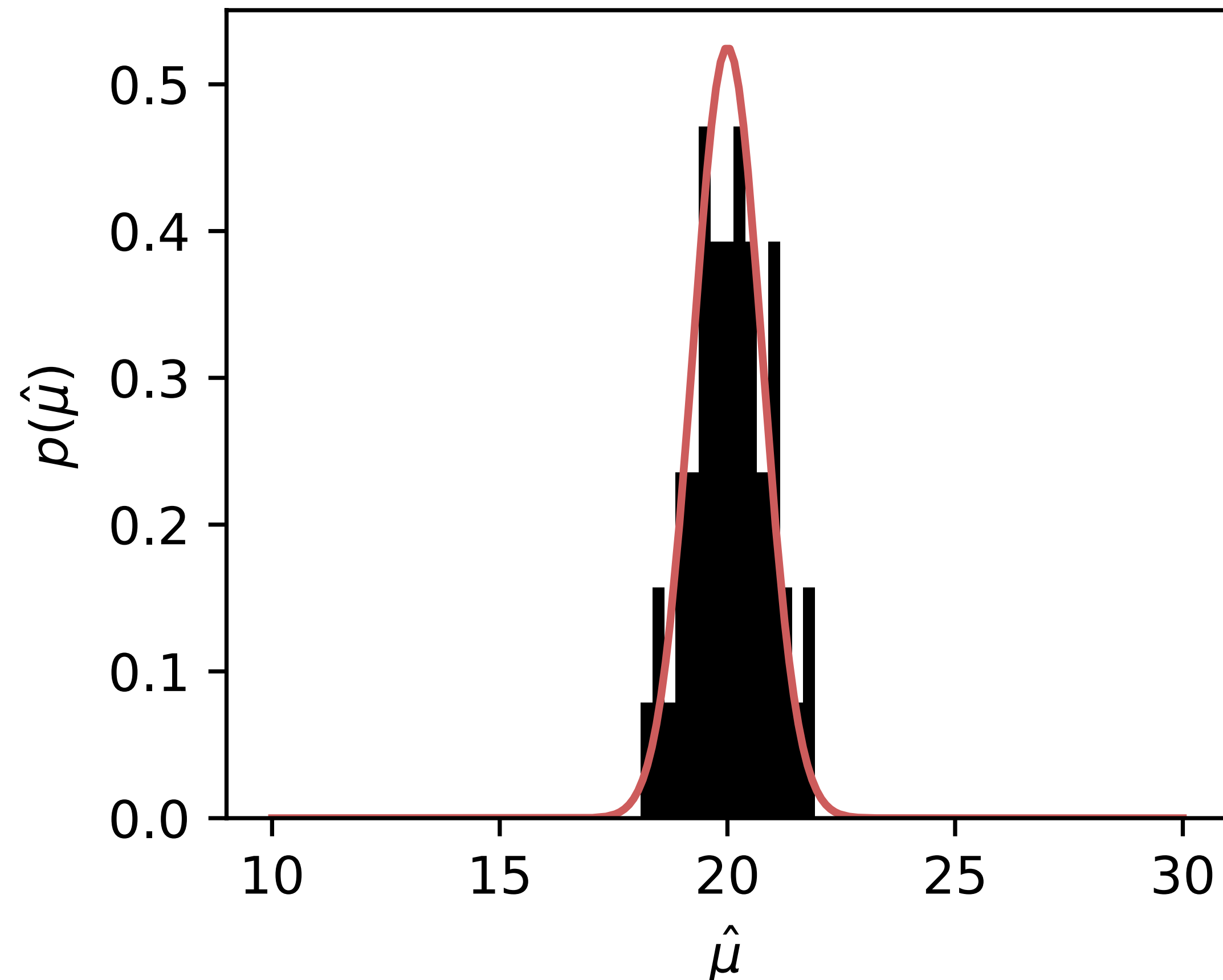
Estimating the mean of the number of initial molecules from first passage time distributions.

FPT's were constructed from 50 simulated data sets of 100 trajectories each.

The maximum likelihood estimate of the mean FPT was found for each data point.

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{dist} \mathcal{N}(0, I(\theta^*)^{-1})$$

**Asymptotic normality of the maximum likelihood estimator**



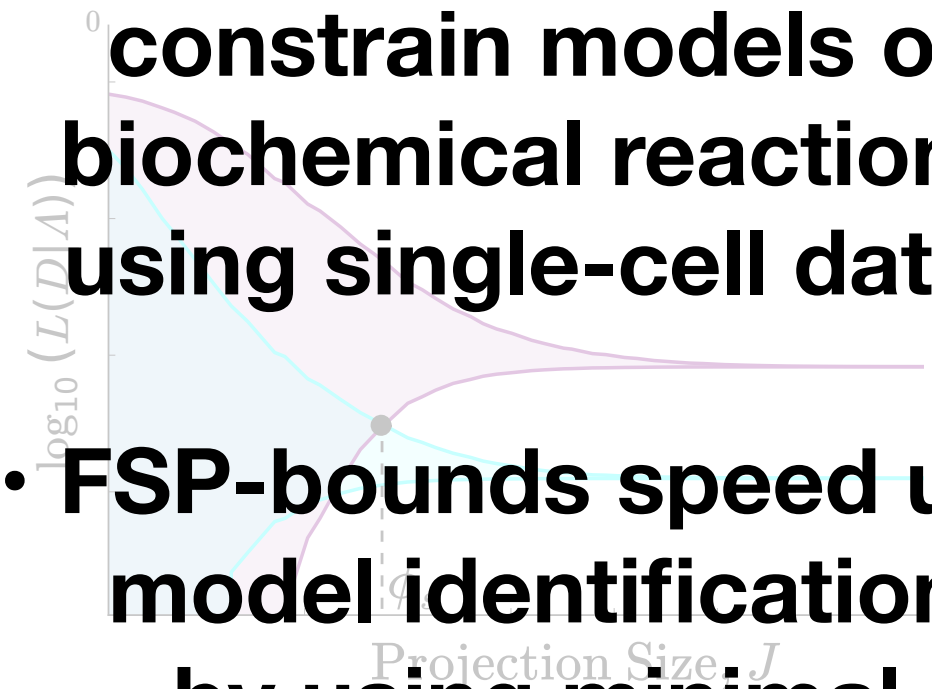
# Outline

Variability in biochemical reactions

- **This variability is paramount to understanding and identifying models of gene expression.**
- **The FSP allows for computation of full probability distributions.**

Efficient model identification using error constraints

- **FSP errors allow us to constrain models of biochemical reactions using single-cell data.**
- **FSP-bounds speed up model identification by using minimal computational effort to compare models.**



Designing single-cell experiments with Fisher Information

- **The FSP-based FIM makes no assumptions about the shape of gene expression distributions.**
- **The FSP-FIM can be used to collect optimally-informative single-cell data.**

# Acknowledgements

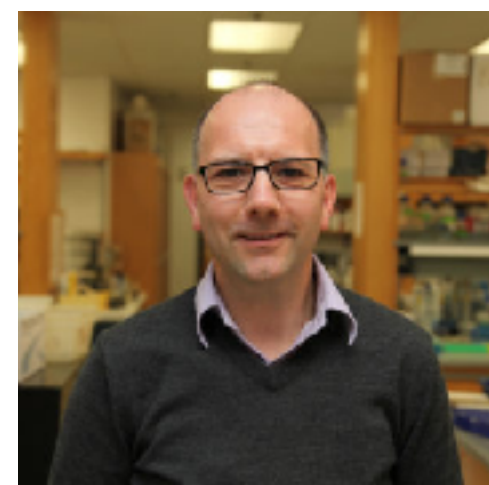
Munsky and Stasevich groups at Colorado State University, Fort Collins, CO



InBio Team



Gregor Neuert, Vanderbilt



Contact & Feedback: [zachfox@gmail.com](mailto:zachfox@gmail.com)

[zachfox.github.io](https://zachfox.github.io)