DISSERTATION

INTEGRATING DISCRETE STOCHASTIC MODELS WITH SINGLE-CELL AND SINGLE-MOLECULE EXPERIMENTS

Submitted by

Zachary R. Fox

Graduate Degree Program in Bioengineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2019

Doctoral Committee:

Advisor: Brian Munsky

Laurie Stargell
Jesse Wilson
Ashok Prasad

ABSTRACT

INTEGRATING DISCRETE STOCHASTIC MODELS WITH SINGLE-CELL AND SINGLE-MOLECULE EXPERIMENTS

Modern biological experiments can capture the behaviors of single biomolecules within single cells. Much like Robert Brown looking at pollen grains in water, experimentalists have noticed that individual cells that are genetically identical behave seemingly randomly in the way they carry out their most basic functions. The field of stochastic single-cell biology has been focused developing mathematical and computational tools to understand how cells try to buffer or even make use of such fluctuations, and the technologies to measure such fluctuations has vastly improved in recent years. This dissertation is focused on developing new methods to analyze modern single-cell and single-molecule biological data with discrete stochastic models of the underlying processes, such as stochastic gene expression and single-mRNA translation. The methods developed here emphasize a strong link between model and experiment to help understand, design, and eventually control biological systems at the single-cell level.

ACKNOWLEDGEMENTS

This work relied very much on abundant collaborations with several students and mentors. I first would like to acknowledge Dr. Abhyudai Singh and his group at the University of Delaware for my first real academic experiences. I had never heard of this field of systems biology before then, and my time in Dr. Singh's group starting shaping me as an independent academic relatively early in my career. Cesar, Mohammad, and Khem also showed me how a even a group of theorists can still collaborate and enjoy many of the aspects of team science. From the start of my time at Colorado State, I have had the pleasure of working with Dr. Gregor Neuert at Vanderbilt University, who never ceased to astound me with his knowledge from both the experimental and theoretical side. Gregor has been a second mentor and has graciously hosted me in his lab several times which have been a highlight of my time at CSU. I would also like to thank Gregor's students over the years who have patiently explained the biological aspects of their research. I had the pleasure of attending three q-bio summer schools hosted at CSU, which were full of many fascinating conversations and new friends. At CSU, I would like to thank Dr. Patrick Shipman, Dr. Tom Chen, Dr. Carol Wilusz, Dr. Stu Tobet, along with Denise Morgan, Kate Sherill, and Sara Mattern for their help and support. I would also like to thank my current and former committee members, Dr. Ashok Prasad, Dr. Laurie Stargell, Dr. Jesse Wilson, and Dr. Diego Krapf, who have provided excellent advice and mentorship throughout my time at CSU. I have been lucky to have many friends from CSU that have been helpful in thinking about my work, especially Patrick Stockton, Jake Sebasta and Kristen Jackson, and also those inside the Munsky group, Lisa Weber, Will Raymond, Michael May, Mohammad Tanhaemami, Charis Ellis, Jaron Thompson, Elliot Djokic, Huy Vo, Luis Aguilera and Linda Forero-Quintero. Finally, I would like to thank Dr. Brian Munsky, who has been exactly the right blend of kind, helpful, thoughtful, caring and inspiring. I cannot imagine having a better mentor, and he has created an example of a group leader that I will aspire to. I could not have done any of this without funding sources, including the NIH, NSF,

and Keck Foundation. I'd like to thank the GAUSSI program as a whole for creating a fantastic research program and their generous support, as well as Dr. Rudolph and the OVPR's support.

I would also like to thank the many other friends and family members who have supported me through my degree, especially my brother, Chad, parents Jim and Diana, and to my other family, Chuck, Donna, Kyle and Katelyn. I do not think I would have been successful without all of you. Finally, I would like to thank my wife Korie for all of her love and support in all aspects of my life.

DEDICATION

I would like to dedicate this dissertation to Korie, Oswald, and Juniper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Many physical, chemical, and biological processes are characterized by discrete particles that randomly fluctuate in space, time, or number. These microscopic fluctuations often provide the key to understand and modify mechanisms that control macroscopic phenomena. By and large, stochastic fluctuations in discrete numbers of specific genes, RNA, or proteins across genetically identical populations of cells play an important role in the understanding of gene regulation [1–5]. As an example, consider the fate of the bacteria *E. coli*, in which a small, seemingly random subset of the population is in a persistent state that can evade antibiotic treatment [6]. This is often thought of as a bet-hedging strategy, as the persistent cells are able to survive attacks at the cost of slower growth and division, while the remainder of the population is able to grow and divide while remaining vulnerable to attack [6–8]. However, what molecular mechanisms determine which cells become persistent, and which grow and divide?

To solve this kind of question, it is important to understand the underlying stochastic processes that influence critical biological systems. Only recently have modern experimental techniques, such as flow cytometry, single-cell RNA sequencing, and single-molecule fluorescence in-situ hybridization (smFISH) [9–11] allowed for the precise quantification of the fluctuations of biomolecules like DNA, RNA and protein at the single-cell and single-molecule level. The “rules" governing these processes, such as mass action kinetics, transcription factor based gene expression regulation, and much more can be modeled and then compared to high resolution data, which may invalidate the different hypotheses (or models) about how a biological process works. However, the details of how to model such processes depend on the type of data that is being collected, the computational feasability of the model, and the underlying statistics of the process that is being measured. Several approaches have been developed to fit models to the statistical moments [12–14], stochastic trajectories [15], or full probability distributions [11, 16] of data collected with these experimental techniques. Despite the progress of these computational and modeling approaches,

our ability to quantify single biomolecules in single cells has created a need for more tools that include the biological details that can be measured experimentally.

In systems where fluctuations are not critical to our understanding their underlying mechanisms, mathematical modeling has been used to gain insight about the system and design experiments to collect better data [17], and even predict how the system responds to new inputs. Such models allows scientists and engineers to use component parts to compose novel systems that can perform pre-programmed tasks. In biology, this type of model-driven approach has had some success, from sustained oscillations of gene expression [18] to identifying certain types of cancer cells [19]. A major roadblock to applications of systems biology is a lack of our ability to develop predictive models. Methods that are able to predictively model biology can revolutionize personalized medicine, agriculture, and biofuel production by applying systems engineering principles. One challenge in biology is that fluctuations in biomolecule numbers, even across isogenic populations, are often non-Gaussian, which necessitates the use of modeling approaches that do not make assumptions about the shapes of the underlying distributions [20]. Furthermore, phenomena are often discrete, and not continuous, which leads to interesting behaviors when particle numbers are low. In fact, the choices that one makes in computational analyses can have a profound impact on our ability to infer model parameters and make useful predictions [20, 21].

In light of the current challenges created by modern data and computational resources, this dissertation develops new theoretical and computational tools to improve the current state-of-the-art for stochastic modeling of gene expression in biological systems. Because fluctuations in biological systems are critically informative for building predictive understanding in biology, each method developed here uses a fluctuation based analysis to model, interepret, and even design experiments for modern measurement approaches.

Within these goals, the methods developed fall in to two main categories, the first of which is to develop new theory that integrates discrete stochastic models and single-cell data. The goal with the projects associated with this aim is to *incorporate data into model reduction strategies*. The general idea of these tools is that the measurements that are collected can be used to help constrain

models in various ways. Such tools allow the modeller to identify models and their parameters and design experiments more efficiently. Chapter 4 discusses the *Finite state projection based bounds on the likelihood of observing single-cell data* [22], which develops new bounds on the likelihood of oberving a measured data set given a particular model of stochastic gene expression. These bounds utilize single-cell data, such as smFISH measurements, to constrain the acceptable modeling error needed to identify models. They can be used to rapidly eliminate much of parameter space that matches data poorly with minimal computational expense. In Chapter 5, we develop projection based reduction of chemical master equation models using single-cell data. In this work, we show how single-cell data can be used to construct a reduced basis that describes the important dynamics of the system. We then project the FSP onto this data-defined basis and use the reduced model to identify model parameters.

The second set of analyses uses stochastic models to design single-cell experiments. Chapters 6 7 develops a method to use discrete stochastic models to design optimal experiments with Fisher information. Fisher information is a common tool in statistics and engineering that uses a model of a system to determine the expected information that can be gained by performing a particular experiment. Often, the Fisher information is used to determine the precision to which model parameters can be estimated within a particular experimental setting. We derive the necessary equations to compute the Fisher information for stochastic models of gene expression and then demonstrate how it can be used to design experiments for several common models of gene expression. Finally, we apply the Fisher information to experimentally measured RNA distributions in the canonical HOG-MAPK stress response system in yeast. Our form of the FIM for stochastic gene expression is the only analysis that uses all of the fluctuation information contained in distributions, and leads to different experiment design decisions than one would find using methods that make assumptions about the shape of the distributions of biomolecules being measured.

Chapter 8 develops methods to integrate stochastic models of single-molecule translation with novel single-particle translation measurements. We develop a stochastic codon-dependent model of single ribosomes as they move along mRNA and elongate proteins. These models incorporate

synthetic sequences that encode epitope regions that bind antibody-like probes. Recent experimental capabilities use this principle to image single polysomes within single cells [23–26]. However, a major limitation of these experiments is the number of antibody-like probes that are available, which fundamentally leads to a small number of genes that can be measured in single-cells. The purpose of our research in this area is to use stochastic models to find predictable fluctuation fingerprints in the fluorescence intensity measurements that arise in different genes as they translate. These different fingerprints allow us to tell apart two different genes as they translate in single-cells, even if they have been labeled with the same antibody-like probes.

The next chapter introduces the chemical master equation (CME), which has been the workhorse of systems biology in recent years. Because the CME is difficult to solve directly (it often consistents of an infinite set of ordinary differential equations), we often use the finite state projection approach (FSP) [27]. The FSP truncates the CME into a finite number of equations. Chapter 2 also discusses other common analyses of the CME, including the stochastic simulation algorithm and approaches based on the dynamics of the moments of the CME. Chapter 3 introduces several likelihood functions which may be used to compare modern experimental data with stochastic models. These functions depend on the assumptions of the model and the resolution (i.e. bulk measurements, single-cell fluorescence measurements, or single molecule measurements) of the data, and are used throughout this dissertation.

Chapter 2

The Chemical Master Equation

2.1 Chemical Master Equation

Like many single-molecule kinetic events, gene expression is often modeled as a Markov process, where each discrete state $\mathbf{x}_i = \begin{bmatrix} \xi_1 & \xi_2 & \xi_3 & \dots & \xi_N \end{bmatrix}_i^T$ corresponds to the integer numbers of N chemical species (e.g., RNA or protein). Transition events between states are different reactions such as transcription, translation or degradation, and these reactions can be indexed by $\mu \in \{1, 2, ..., M\}$. These reactions occur with propensities $w_\mu(\mathbf{x}_i)dt$, which is the probability that the μ^{th} reaction occurs in the next infinitesimal time step $(t, t+dt)$ given the current state $\mathbf{x}_i$. State transitions are described as $\mathbf{x}_i \rightarrow \mathbf{x}_j = \mathbf{x}_i + \boldsymbol{\psi}_\mu$, where $\boldsymbol{\psi}_\mu$ is the stoichiometry vector that describes the change in population after the μ^{th} reaction. In such models, each node has a continuous valued probability $p(\mathbf{x}_i, t)$ that evolves in time according to the linear ODE known as the chemical master equation (CME), [28, 29]

$$\frac{dp(\mathbf{x}_i, t)}{dt} = \sum_{\mu=1}^{M} w_\mu(\mathbf{x}_i - \boldsymbol{\psi}_\mu)p(\mathbf{x}_i - \boldsymbol{\psi}_\mu, t) - \sum_{\mu=1}^{M} w_\mu(\mathbf{x}_i)p(\mathbf{x}_i, t). \tag{2.1}$$

By enumerating all possible $\{\mathbf{x}_1, \mathbf{x}_2, ...,\} \in \mathbf{X}$ and corresponding probabilities, $\mathbf{p} = \begin{bmatrix} p(\mathbf{x}_1, t) & p(\mathbf{x}_2, t) & ... \end{bmatrix}^T$, the CME can be posed in matrix form as $\frac{d}{dt}\mathbf{p}(t) = \mathbf{A}\mathbf{p}(t)$, where $\mathbf{A}$ is known as the *infinitesimal generator* (examples of $\mathbf{A}$ are provided in later sections).

The CME dimension is often infinite, making it impossible to solve directly for most systems. The finite state projection (FSP) approach was developed to allow one to approximate the CME solution within strict error bounds [11, 27, 30].

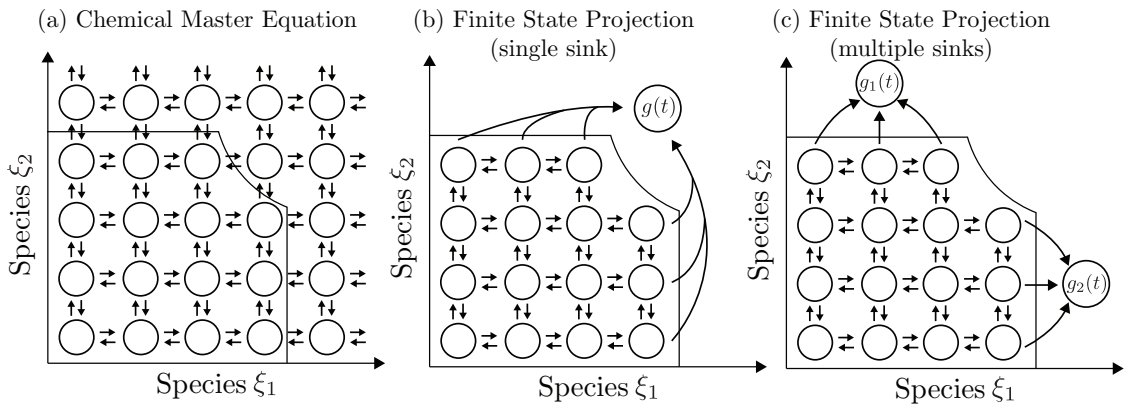

Figure 2.1: *Demonstration of the finite state projection approximation to the chemical master equation.* (a) Graph representation of a master equation with two species and infinite states. (b) Finite state projection with a subset of the full state state space, and any reaction that leaves the set of states indexes by J must go into the sink state $g(t)$. (c) Same as (b), except with multiple sinks.

2.2 Finite State Projection

In its formulation, the FSP approach selects a finite set of indices, $J = \{j_1, ..., j_L\}$ with which it separates the full state space $\mathbf{X}$ into two exhaustive and disjoint sets, $\mathbf{X}_J = \{\mathbf{x}_{j_1}, ..., \mathbf{x}_{j_L}\}$ and its complement $\mathbf{X}_{J'}$. Under this reorganization, the full master equation can be written

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}_J(t)\\ \mathbf{p}_{J'}(t)\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{JJ} & \mathbf{A}_{JJ'}\\ \mathbf{A}_{J'J} & \mathbf{A}_{J'J'}\end{bmatrix}\begin{bmatrix}\mathbf{p}_J(t)\\ \mathbf{p}_{J'}(t)\end{bmatrix}. \tag{2.2}$$

To approximate the CME, the FSP approach forms a finite state Markov process, where all nodes in $\mathbf{X}_{J'}$ are aggregated to one or more sink states g that record the probability mass that leaves $\mathbf{X}_J$. However, the FSP approach requires all probability mass within g to remain in g as time proceeds. The new, reduced FSP-CME becomes

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}_J^{\text{FSP}}\\ g(t)\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{JJ} & \mathbf{0}\\ -\mathbf{1}^T\mathbf{A}_{JJ} & 0\end{bmatrix}\begin{bmatrix}\mathbf{p}_J^{\text{FSP}}\\ g(t)\end{bmatrix}. \tag{2.3}$$

The resulting approximation in Eq. (2.3) provides three key insights into the exact CME solution. First, it provides a lower bound on the true solution,

$$\begin{bmatrix} \mathbf{p}_J^{\mathrm{FSP}}(t) \\ \mathbf{0} \end{bmatrix} \leq \begin{bmatrix} \mathbf{p}_J(t) \\ \mathbf{p}_{J'}(t) \end{bmatrix} \quad \text{for all} \;\; t > 0. \tag{2.4}$$

This can be easily interpreted by noting that probability can only leave $\mathbf{X}_J$ in the FSP-CME (Eq. 2.3), but can return from $\mathbf{X}_{J'}$ to $\mathbf{X}_J$ in the original CME (Eq. 2.2). Second, the FSP provides an exact measure of the approximation error,

$$\left| \begin{bmatrix} \mathbf{p}_J(t) \\ \mathbf{p}_{J'}(t) \end{bmatrix} - \begin{bmatrix} \mathbf{p}_J^{\mathrm{FSP}}(t) \\ \mathbf{0} \end{bmatrix} \right|_1 = \left| \mathbf{p}_J(t) - \mathbf{p}_J^{FSP}(t) \right|_1 + \left| \mathbf{p}_{J'}(t) \right|_1 \tag{2.5}$$

$$= \left| \mathbf{p}_J(t) \right|_1 + \left| \mathbf{p}_{J'}(t) \right|_1 - \left| \mathbf{p}_J^{FSP}(t) \right|_1 \tag{2.6}$$

$$= 1 - \left| \mathbf{p}_J^{FSP}(t) \right|_1 \tag{2.7}$$

$$= g(t), \tag{2.8}$$

where $|\cdot|_1$ denotes the absolute sum of a vector. Finally, as states are added to the set $\mathbf{X}_J$, the error $g(t)$ decreases monotonically. Proofs of these results are provided in [30, 31]. The FSP yields a finite set of linear ordinary differential equations. In the case of a non-time varying infinitesimal generator matrix, the solution to the FSP for an initial condition $\mathbf{p}_J^{\mathrm{FSP}}(0)$ at time t is simply the matrix exponential,

$$\mathbf{p}_J^{\mathrm{FSP}}(t) = \exp\left(\mathbf{A}_{JJ} t\right) \mathbf{p}_J^{\mathrm{FSP}}(0). \tag{2.9}$$

However, for many interesting systems shown in Chapters 4 and 6, the generator $\mathbf{A}$ is time-varying. In these situations, we are limited to numerically integrating the set of ODEs in Eq. 2.3.

The state space of the FSP, $\mathbf{X}_J$, is easily defined through use of polynomial projection shapes [30],

$$\mathbf{X}_J = \left\{ \mathbf{x}_i \text{ such that } f_k(\mathbf{x}_i) \leq c_k \text{ for all constraints } k = 1, 2 \ldots, K \right\}. \tag{2.10}$$

Here $\{f_k(\mathbf{x}_i)\}$ is a set of polynomials of the population counts, and the constraints $\{c_k\}$ are weights on these polynomials that may be increased (decreased) to include more (fewer) states. In practice, each k^{th} constraint can be associated with its own sink, g_k, which aggregates all states that satisfy the $\{1,\ldots,(k-1)\}^{\text{th}}$ constraints, but not the k^{th} constraint. The value of $g_k(t_f)$ then quantifies probability of violation for the k^{th} constraint, which in turn guides the systematic increase of c_k. With this expansion, the FSP algorithm presented in [27, 30] can be used to select an $\mathbf{X}_J$ to make $g(t) = \sum_k g_k(t)$ small for a specified finite time. However, lower error will require more states and greater computational expense, which is described in detail in Chapter 4.

2.3 Moments of the Chemical Master Equation

Often, the FSP is computationally intractable to solve. In such cases, one may turn to statistical moments of the time varying distribution $p(t)$ defined by the CME, which are often able to be efficiently computed. For systems with affine linear propensity functions (i.e. the propensity function $w(\mathbf{x},t) = w_0(t)\mathbf{x} + w_1(t)$, the moments of the CME can be computed to arbitrary order. The uncentered moments of the CME, $\mathbb{E}\{\mathbf{x}^{\mathbf{m}}\}$, where $\mathbf{m} = [m_1, m_2, ..., m_{N_s}]$ is a vector of integers corresponding to the m_i^{th} power of the i^{th} species in $\mathbf{x}$, and the entire moment $\mathbf{x}^{\mathbf{m}}$ is found according to the following formula [12]:

$$\frac{d\mathbb{E}\{\mathbf{x}^{\mathbf{m}}\}}{dt} = \mathbb{E}\left\{\sum_{j=1}^{N_r} w_j(\mathbf{x})\left[\prod_{i=1}^{N_s}(\xi_i + \mathbf{\Psi}_{ij})^{m_i} - \prod_{i=1}^{N_s}\xi_i^{m_i}\right]\right\}. \tag{2.11}$$

In the next chapter, I will show how these moments have been used to maximize the likelihood of a stochastic model given single-cell data. For example, consider a simple birth and death process, with two reactions

$$\mathcal{R}1 : \varnothing \xrightarrow{k_{\text{r}}} X;\ \ \mathcal{R}2 : X \xrightarrow{\gamma} \varnothing.$$

This process has a stoichiometry matrix and propensity matrix given by

$$\mathbf{\Psi} = \begin{bmatrix} 1 & -1 \end{bmatrix} \tag{2.12}$$

$$\mathbf{W} = \begin{bmatrix} 0 \\ \gamma \end{bmatrix} x + \begin{bmatrix} k_{\mathrm{r}} \\ 0 \end{bmatrix} \tag{2.13}$$

Applying these to Eq. 2.11 we find the following dynamics for the mean of the process, $\mathbf{m} = [1]$ as

$$\begin{aligned} \frac{d\mathbb{E}\{x\}}{dt} &= \mathbb{E}\left\{k_r\left((x+1)-x\right) + \gamma x\left((x-1)-x\right)\right\} \\ &= k_{\mathrm{r}} - \gamma\mathbb{E}\{x\}. \end{aligned} \tag{2.14}$$

Interestingly, this form of the equation exactly matches the macroscopic ODEs corresponding to the same system. This is true whenever propensity functions are linear with x [32]. Similarly, the second uncentered moment of the process can be found

$$\begin{aligned} \frac{d\mathbb{E}\{x^2\}}{dt} &= \mathbb{E}\left\{k_r\left((x+1)^2 - x^2\right) + \gamma x\left((x-1)^2 - x^2\right)\right\} \\ &= -2\gamma\mathbb{E}\left\{x^2\right\} + \left(2k_{\mathrm{r}} + \gamma\right)\mathbb{E}\left\{x\right\} + k_{\mathrm{r}}. \end{aligned} \tag{2.15}$$

While moment-based approaches can be useful in calculating solutions to the CME, when the propensities are nonlinear, one must turn to approximations such as moment closures [12] to find the moment dynamics. Even with exact moments, the number of moments that needs to be computed to accurately represent the underlying distribution can be large, and leasds to a large, dense set of equations to be integrated that computationally comparable to solving the full master equation [20]. In such situations, it may be useful to turn to stochastic simulations of the chemical master equation.

2.4 Simulating the Chemical Master Equation

Perhaps the most common approach to solving the chemical master equation is to find sample paths from the time varying probability distribution. This is achieved through the stochastic simulation algorithm (SSA), often called the Gillespie algorithm [33]. Each trajectory simulated using this algorithm is a sample path from the solution to the chemical master equation $p(x,t)$. Algorithm 1 outlines a simple SSA implementation, called the direct method. Essentially, this approach uses two randomly generated number to determine when does the next reaction happens and which reaction occurs. From these two pieces of information, the state is updated.

Algorithm 1 Stochastic Simulation Algorithm

Initialize $\mathbf{x} = \mathbf{x}_0$, $t = t_0$, $\mathbf{w} = \mathbf{w}(\mathbf{x}_0, t_0)$
while $t < t_f$ **do**
  $r_1 = \text{unif}(0,1)$
  $r_2 = \text{unif}(0,1)$
  $a_0 = |\mathbf{w}(\mathbf{x},t)|_1$
  $\tau = \min\left[\log(1/r_1)/a_0, t_f - t\right]$
  $t = t + \tau$
  $k = 1$
  while $r_2 < w_k/a_0$ **do**
    $w_k = \sum_{i=1}^{k} \mathbf{w}_i(\mathbf{x},t)$
    $k = k + 1$
  end while
  $\mathbf{x} = \mathbf{x} + \boldsymbol{\psi}_k$
end while

By running many simulations, one can approximate the solution to the CME with high fidelity, though many trajectories may be required to achieve low enough error, especially when the distribution being sampled has long tails. The error after N samples of the trajectory is $\mathcal{O}(N^{-1/2})$, as are all Monte Carlo algorithms. As an example, to estimate a probability of 10^{-4}, one needs

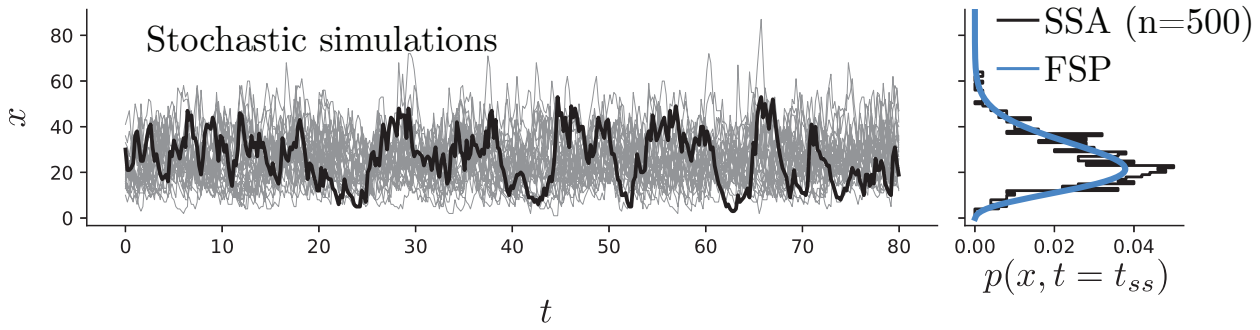

Figure 2.2: *Demonstration of the SSA and FSP approaches to solving the chemical master equation.* Sample trajectories simulated by the SSA are shown in grey, with one example in black. After some time, the system has equilibrated and the distribution is stationary. The right panel show a normalized histogram of 500 SSA trajectories compared to the FSP solution with error less than 10^{-6}.

to run 10^8 trajectories. Sample trajectories and the FSP solution are shown in Fig. 2.2. Because of the computational challenges presented by Monte Carlo approaches to simulating the chemical master equation, approximation schemes have been developed to more efficiently generate sample paths. Perhaps the most commonly used is τ-leaping, in which each reaction is taken to occur a Poisson-distributed number of times in the small time period τ [34–37]. However, care must be taken in choosing appropriate values of τ because the propensities are often functions of the value of the state, and therefore may change substantially if a large number of reactions occur in the time τ. A classic pathological example that many of the above articles deal with is if the propensities are such that more degradation reactions happen in time τ than the number of productions plus the number of proteins already in the system, in which case the total number of protein become negative.

Many different approaches to modeling using the CME have be used, and ultimately the correct choice depends on the computational resources available, the fidelity of the data that is being used (i.e. single-cell vs. bulk measurements, discrete molecule counting vs. intracellular fluorescence), and the importance of stochasticity in the problem. This dissertation is mainly concerned with systems in which using the FSP approach is the best option, though throughout compares results to those one would get using a moments based approach or a stochastic simulation based approach. The next chapter utilizes these different approaches to find the likelihood of different types of

data sets, for bulk measurements with Gaussian errors to discrete molecule counting of individual RNA.

Chapter 3

Likelihood-based identification of stochastic models of gene expression

To make a model of a biological process useful, one must match experimentally observed variables to those in the model, i.e. fit the model to the data. However, there are many approaches to model calibration or fitting that one may take, depending on the type of data that is being considered, the computational cost of solving the model, and whether the uncertainty in the kinetic parameters of the model is of interest or only a single point estimate of the parameters. To even further complicate the problem, one must decide which model or models to use in the first place, and how to rigorously discriminate between multiple models. The famous quote from Jon von Neumann “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk" aludes to the tradeoff between model complexity and overfitting a model. However, in many ways this quote neglects yet another challenge, which is that models with more parameters may fit more features of the data, but actually finding those regions of parameter space can be extremely difficult. The approach of our work is to start with the assumption that we do not have the true model, but instead a model which can be useful particular aspects of the system, and that can be invalidated. To find such kinetic parameters, and to perform model selection, one needs to determine the quality of the fit and the associated uncertainties for a particular model and a particular data set.

This chapter derives likelihoods of different types of data for models of stochastic gene expression under different assumptions about the characteristics of the data that is being fit. Once the likelihood function that is appropriate to use is established, it can be applied in maximum likelihood frameworks, Bayesian inferenence, or other optimization schemes of interest.

3.1 Derivation of Likelihoods

In Chapters 4-6 we are interested in analyzing snapshot measurements of independent cell populations, such as those collected using smFISH, over multiple time points. The smFISH technique uses small oligo-nucleotides with attached fluorophores that hybridize to an RNA of interest [9,10]. These fluorescent probes bind to the complementary sequence of the RNA of interest in the cells, producing diffraction limited spots that can be counted in each cell to quantify the discrete number of RNA in a given cell. In cells with large numbers of RNA, it may be difficult to discern the numbers of spots of RNA that appear in each cell. However, cells must be fixed for the oligo-nucelotide probes to enter them, and therefore each temporal measurement contains unique cells, and often in this case one assumes that the measurements are independent in time, as no single-cell temporal correlations are available. One other advantage of this approach is that it does not require genetic modifications to the genes that are being studied, as is common for time-lapse microscopy techniques that use GFP. We assume that measurements at each time point $\mathbf{t} \equiv [t_1, t_2, \ldots, t_{N_t}]$ are independent. Part or all of the species in $\mathbf{x}_i^L$ may be measured, where $L \subseteq (1, 2, \ldots, N_s)$ is set of N_o observable indices. Measurements of N_c cells can be concatenated into a matrix $\mathbf{D}_t \equiv [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{N_c}]_t$ of the observable dimensions at each measurement time t.

For FSP models, the likelihood of indpendent measurements and its logarithm for N_c measured cells can be written directly is simply the product of the probabilities, where state $\mathbf{x}_i^L$ was observed y_j times at time t:

$$\ell(\mathbf{D}|\boldsymbol{\theta}) = \prod_{t=t_1}^{t_{N_t}} \prod_{j \in \mathcal{J}_D} p(\mathbf{x}_j^L, t|\boldsymbol{\theta})^{y_j} \tag{3.1}$$

$$\log \ell(\mathbf{D}|\boldsymbol{\theta}) = \sum_{t=t_1}^{t_{N_t}} \sum_{j \in \mathcal{J}_D} y_j \log(p(\mathbf{x}_j^L, t|\boldsymbol{\theta})), \tag{3.2}$$

where $\mathcal{J}_D$ is the set of states observed in the data. The vector $p(\mathbf{x}^L)$ is the marginal distribution of the observable species from the joint probability vector $p(\mathbf{x})$. The summation in Eq. 3.1 can be rewritten as a product $\mathbf{y} \log p(\mathbf{x}^L)$, where $\mathbf{y} \equiv [y_0, y_1, \ldots, y_M]$ (i.e. binned data).

3.2 Moment-based approaches to likelihood

As previously discussed, in many situations computing the full solution to the FSP is computationally intractable, and one must attempt to instead identify model parameters $\boldsymbol{\theta}$ by matching the moments of the chemical master equation to the moments of single-cell data. In the limit of large numbers of molecules reacting in a well-mixed solution, the linear noise approximation (LNA) may be applied to CME [29]. In such cases, molecule numbers are considered to be Gaussian, and the well-known Gaussian form of the likelihood may be applied [38]. If the observed data is assumed to come from a multivariate Gaussian distribution with means $\boldsymbol{\mu}(t;\boldsymbol{\theta}) = [\mu_1(t;\boldsymbol{\theta}), \mu_2(t;\boldsymbol{\theta}), \dots \mu_{N_s}(t;\boldsymbol{\theta})]^T$ and covariance matrix $\boldsymbol{\Sigma}(t;\boldsymbol{\theta})$, such as those in Eqs. 2.11, the likelihood is given by:

$$\ell(\mathbf{D};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{t=t_1}^{t_{N_t}} \prod_{i=1}^{N_c} (2\pi^{N_o}|\boldsymbol{\Sigma}(t)|)^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(\mathbf{d}_i(t) - \boldsymbol{\mu}(t))^T \boldsymbol{\Sigma}^{-1}(t)(\mathbf{d}_i(t) - \boldsymbol{\mu}(t))\right) \quad (3.3)$$

In [13, 14, 39], the authors suggest approximating the likelihood where the sample mean and variance are taken to be jointly Gaussian, i.e. the random vector $\mathbf{z} = [\boldsymbol{\mu}_\mathbf{s}, \boldsymbol{\Sigma}_s]^T$, $Z \sim \mathcal{N}(\mathbf{z}, \mathbf{C})$, and $\mathbf{C}$ is the covariance matrix:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\boldsymbol{\mu}_s\boldsymbol{\mu}_s} & \mathbf{C}_{\boldsymbol{\mu}_s\boldsymbol{\Sigma}_s} \\ (\mathbf{C}_{\boldsymbol{\mu}_s\boldsymbol{\Sigma}_s})^T & \mathbf{C}_{\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_s} \end{pmatrix}. \quad (3.4)$$

The submatrices on the diagonal correspond to the variance of $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$, and the off diagonal terms correspond to correlations between the sample means and variances.

In [14], they derive the elements of each of these matrices in terms of the moments of the underlying model distribution $p(\mathbf{x}|\boldsymbol{\theta})$ for models with one or two species.

For example, consider the variance/covariance of the sample mean is $\mathbf{C}_{\boldsymbol{\mu}_s\boldsymbol{\mu}_s}$, where we have the data matrix with N measurements $\mathbf{X} = [\mathbf{x}_1^T\ \mathbf{x}_2^T \dots \mathbf{x}_N^T]^T$, where each row in the matrix $\mathbf{X}$ corresponds to a different measurement. The sample mean $\bar{\mathbf{x}}$ can be written $\mathbf{1}^T\mathbf{X}/N$, where $\mathbf{1}$ is a column vector of ones of size N. Without loss of generality, let $\mathbb{E}\{\mathbf{X}\} = 0$, and

$$\mathbf{C}_{\boldsymbol{\mu}_s\boldsymbol{\mu}_s} = \frac{1}{N^2}\left(\mathbb{E}\left\{\mathbf{1}^T\mathbf{X}\mathbf{X}^T\mathbf{1}\right\} - \mathbb{E}\left\{\mathbf{1}^T\mathbf{X}\right\}\mathbb{E}\left\{\mathbf{1}^T\mathbf{X}\right\}^T\right) \tag{3.5}$$

$$= \frac{1}{N^2}\mathbb{E}\left\{\mathbf{1}^T\mathbf{X}\mathbf{X}^T\mathbf{1}\right\} = \frac{N}{N^2}\mathbb{E}\left\{\mathbf{X}\mathbf{X}^T\right\} \tag{3.6}$$

$$= \frac{1}{N}\boldsymbol{\Sigma}. \tag{3.7}$$

Similar procedures can be used find the rest of the $\mathbf{C}$. One challenge with this approach is highlighted with the practicality of using measured sample variance when the population variance is large. This is demonstrated in Fig. 3.1, which shows the distribution of sample variances for the induced RNA Hog-MAPK data from our work in [20]. Essentially, a broad distribution of sample variances can lead to a high probability of sampling a sample variance that is lower than the true variance, which can bias the maximization of the estimation of parameters.

3.3 Inference of time-series data

All of the likelihood functions discussed up to this point discuss data that are independent in time. Most often, this means that one cannot track single cells over multiple time points, but rather to take a measurement one must fix the cells (as is the case with smFISH data or single-cell RNA sequencing data) or that the cells have no identity (as with basic flow cytometry data). However, time correlated data are very common in fluorescence time-lapse microscopy data, and can provide a large wealth of data. While this area is still being actively researched [38, 40, 41], I will briefly outline an FSP-based approach to inferring likelihood from time-series data.

Consider a single time-series measurement of a single fluorescent protein, $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_n)]$. Given a stochastic model of the abundance of the particular protein y, the data can be considered a single-sample path of the full time-varying probability distribution $p(y, t|\boldsymbol{\theta})$. Note that although this section is written from the perspective of a single fluorescent protein in a single cell, it can easlily be extended to measurements of multiple proteins or proteins and RNA in single cells. Because temporal correlations may last the entire trajectory of gene expression that has been measured, the likelihood at a given time point depends on the entire path until the final time,

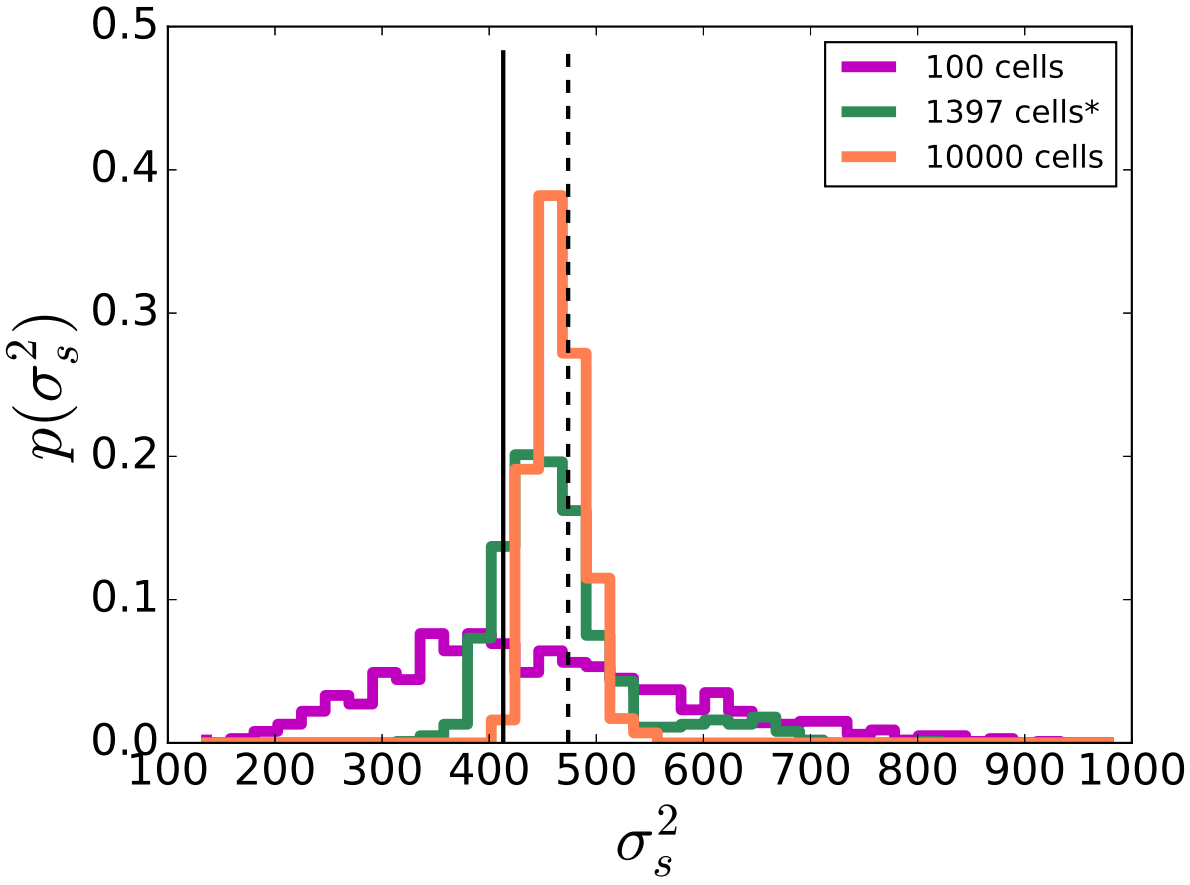

Figure 3.1: *The effect of finite data on estimates of the variance,* σ_s^2. Distributions of sample variances 1,000 measurements of 100 cells (purple), 1,397 cells (green) and 10,000 cells (orange) were computed. 1,397 cells were measured experimentally for the 0.2M condition at t =15 min. When many cells are measured, the distribution of σ_s^2 is approximately Gaussian (orange). However, with less measurements (green and magenta) these distributions are not only more broad (as expected by the central limit theorem), but also skewed. This skewness arises because of the long tails often observed in the data. This means that a relatively small random sampling of such distributions will underestimate the variance of the distribution. However, infrequently the tail of this distribution will be measured, and the sample variance will be much larger than the true variance. On average, the estimator is unbiased; the mean of all three distributions is the same.

$$\ell(\mathbf{y}|\boldsymbol{\theta}) = p(y_1, y_2, ...y_N|\boldsymbol{\theta}) \tag{3.8}$$

Under the Markov assumption, the probability of moving from y_1 to y_2 depends only on $p(y_1)$, and therefore the probability of observing y_2 is the probability that the system was in y_1 and moved to y_2, or the transition probability $p(y_2|y_1)$. Therefore, the likelihood can be written as a series of transition probabilities,

$$\ell(\mathbf{y}) = p(y_1)p(y_2|y_1), ...p(y_N|y_{N-1}) \tag{3.9}$$

$$= p(y_1)\prod_{i=2}^{N} p(y_i|y_{i-1}) \tag{3.10}$$

and the log-likelihood is (3.11)

$$\log \ell(\mathbf{y}) = \log p(y_1) + \sum_{i=2}^{N} \log p(y_i|y_{i-1}). \tag{3.12}$$

(3.13)

Noting again that the solution of the FSP (with non-time-varying propensity functions) is given by the $\mathbf{p}(t_f) = \exp(\mathbf{A}(\boldsymbol{\theta})t_f)\mathbf{p}(0)$, the matrix $\mathbf{Q}(\boldsymbol{\theta}) = \exp(\mathbf{A}(\boldsymbol{\theta})\Delta t)$ maps the solution of the master equation at time t to the solution at time time $t + \Delta t$, i.e. $\mathbf{p}(t + \Delta t) = \mathbf{Q}(\boldsymbol{\theta})\mathbf{p}(t)$. Therefore $\mathbf{Q}(\boldsymbol{\theta})$ is a matrix of the transition probabilities in the time Δt. In the case of fluorescent time-lapse experiments, Δt corresponds to the measurement sampling period. Therefore, the log-likelihood of time-lapse data found using the FSP can be found from

$$\log \ell(\mathbf{y}) = \log p(y_1) + \sum_{i=2}^{N} \log \mathbf{Q}_{y_i, y_{i-1}}. \tag{3.14}$$

This approach is exact in the sense that it does not make any assumptions beyond those of standard chemical reaction kinetics. It remains to be seen how approximations of this approach via linear noise approximations changes the inference of model parameters. For the other likelihood-based inference methods discussed above, there are a plethora of examples throughout this dissertation.

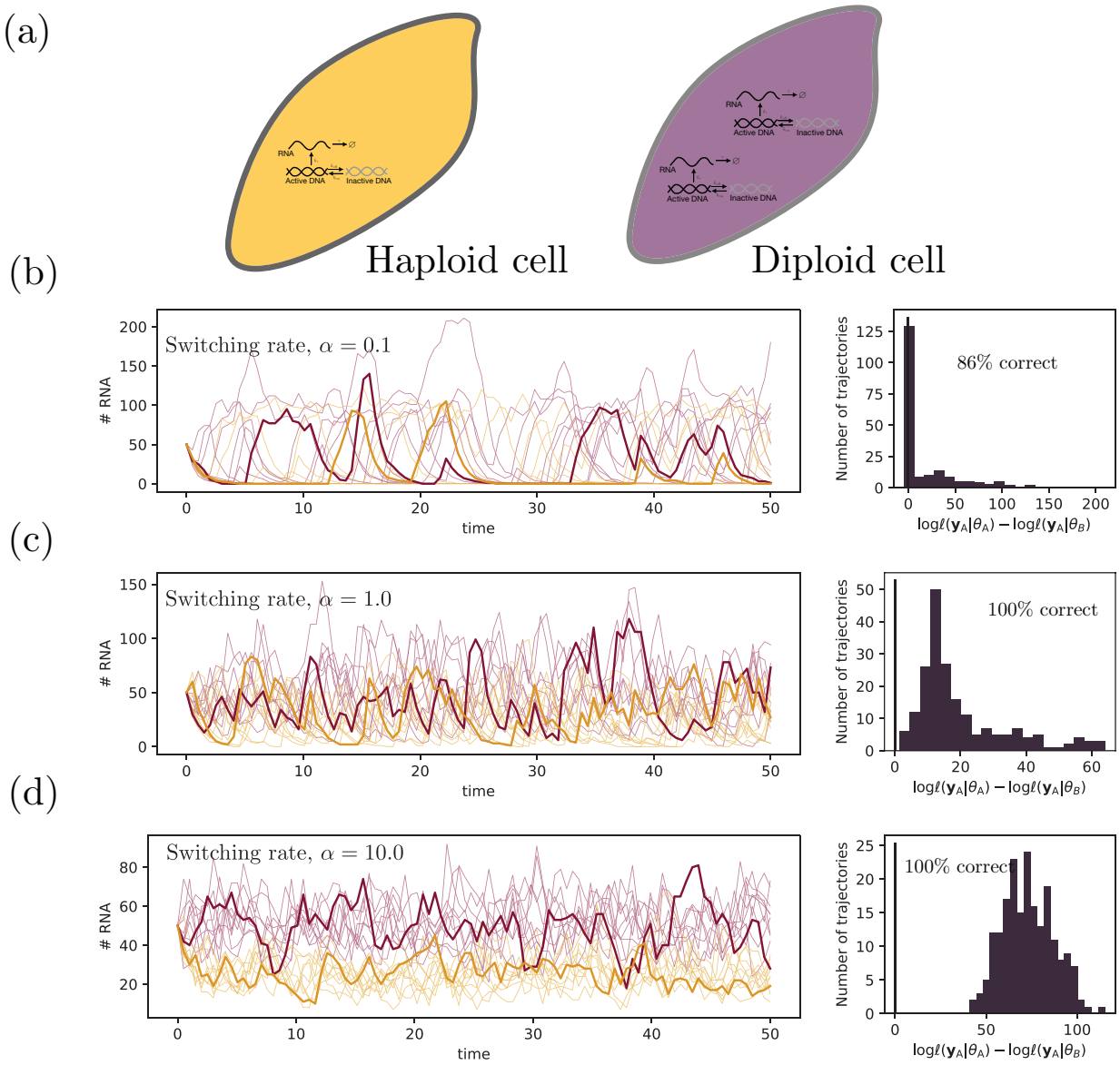

Figure 3.2: *Time series inference of single-cell data.* (a) Haploid cells with a single copy of a gene and a diploid cell with two copies of the same gene. (b-d) Example trajectories of the different cell types haploid (yellow) and diploid (purple), with $k_{\text{on}} = \alpha$, $k_{\text{off}} = 3\alpha$, $k_r = 100$, and $\gamma = 1$. In (b), $\alpha = 0.1$, (c) $\alpha = 1$, (d) $\alpha = 10$. For each value of α, the likelihood that a given trajectory came from the correct model $\log \ell(\mathbf{y}_A|\theta_A)$ or for the incorrect model $\log \ell(\mathbf{y}_A|\theta_B)$ was computed for both haploid and diploid cells. The difference was binned and plotted on the right.

However, as motivation for the time-series inference, consider the following experiment, where there is a population of cells that either have one copy of a given gene (haploid cells) or two copies of that gene (diploid cells). If we assume that the production rate and switching dynamics of a two-state promoter are equal for each gene, we are left with the standard 'bursting gene expression' model of transcription, shown in Fig. 3.2(a) [1], and diploid cells undergo a slightly different stochastic process with three effective gene states, corresponding to a single copy actively transcribing RNA, both copies actively transcribing RNA, and neither copy actively transcribing RNA, which makes up the following set of biochemical reactions:

$$\begin{aligned}
\mathcal{R}1: \quad & G_{\text{off,off}} \xrightarrow{2k_{\text{on}}} G_{\text{off,on}} \\
\mathcal{R}2: \quad & G_{\text{off,on}} \xrightarrow{k_{\text{off}}} G_{\text{off,off}} \\
\mathcal{R}3: \quad & G_{\text{off,on}} \xrightarrow{k_{\text{on}}} G_{\text{on,on}} \\
\mathcal{R}4: \quad & G_{\text{on,on}} \xrightarrow{2k_{\text{off}}} G_{\text{off,on}} \\
\mathcal{R}5: \quad & G_{\text{on,off}} \xrightarrow{k_{\text{r}}} G_{\text{on,off}} + RNA \\
\mathcal{R}6: \quad & G_{\text{on,on}} \xrightarrow{2k_{\text{r}}} G_{\text{on,off}} + RNA \\
\mathcal{R}7: \quad & RNA \xrightarrow{\gamma} \varnothing.
\end{aligned} \tag{3.15}$$

Thus, we have two different models of the biochemical processes, θ_{hap} for the haploid cells and θ_{dip} for the diploid cells. For each of these two models, we simulated time series trajectories using the SSA [33],1, shown in Fig. 3.2, where each trajectory corresponds to RNA abundance in either a haploid or diploid single cell. The likelihood of each trajectory was computed under both model assumptions, $\log \ell(\mathbf{y}|\theta_{\text{dip}})$ and $\log \ell(\mathbf{y}|\theta_{\text{hap}})$. We then subtract the likelihood of the correct model, labeled θ_A whether the trajectory was simulated with θ_{dip} or θ_{hap} from the incorrect model, labeled θ_B. If the likelihood of the correct model θ_A is higher than the likelihood of the incorrect model θ_B, their difference will be positive and the trajectory was correctly classified, shown in Fig. 3.2(b-d), right panels. To test the effect of different promoter switching rates, each panel (b-d) has a different promoter switching rate, α, which affects the model parameters as $k_{\text{on}} = \alpha$, $k_{\text{off}} = 3\alpha$.

As promotor switching increases, the trajectories are much easier to identify as belonging to a diploid or haploid cell, as the processes are essentially Poisson with two different mean expression levels. However, at slow switching rates, the dynamics between the two time series are much more similar, though still identifiable using this approach.

While this work demonstrates one potential use of the FSP to infer time-series data, the likelihood function in Eq. 3.14 could be used to infer model parameters from single-cell trajectories using maximum likelihood approaches, or to find posteriors of model parameters in a Bayesian setting. Furthermore, most current experiments measure a single fluorescence signal that changes over time in a single cell, as opposed to discrete RNA or protein numbers as this approach assumes. However, the deconvolution of total fluorescence into protein numbers has been used to fit flow cytometry measurements with the FSP in the past [42], and the same idea could be applied to single-cell time series.

The next chapter introduces a new upper bound on the likelihood in Eq. 3.1, by recognizing that the FSP solution only provides a lower bound in the likelihood of single-cell data. We derive this upper bound and find a novel algorithm to rapidly compute it. This upper bound depends on the model error, $g(t)$, which in turn depends on the amount of states in the FSP and ultimately the computational expense to solve the model.

Chapter 4

Finite state projection based bounds to compare chemical master equation models using single-cell data [1]

4.1 Introduction

A little over ten years ago, the finite state projection (FSP [27]) approach was introduced to approximate the solution of the Chemical Master Equation (CME [28, 29]) and to capture the dynamics of discrete molecular events that control single-cell gene regulation. Since that time, the FSP has received substantial attention; has seen numerous computational improvements; and has become a benchmark tool in the analysis of stochastic gene regulation. Most recently, the FSP has been used to fit and predict experimental data in yeast, bacteria, and human cells [43]. The main utility of the FSP is to provide precise bounds on the accuracy of its approximation as well as a systematic approach to improve that accuracy. However, improved accuracy comes with increased computational cost, and no attention has been given to how one could optimize this tradeoff. Careful evaluation of this tradeoff is needed to improve the rigor and efficiency with which FSP models can be matched to experimentally measured data. In this work, we develop new FSP-based bounds on the likelihood of single-cell data given a stochastic model; we show how these bounds can be used to reduce computational costs; and we demonstrate how the co-design of FSP tools and experimental data can lead to efficient inference of discrete stochastic models from experimental single-cell data.

[1] This work was published in PLoS Computational Biology in 2019.

4.2 FSP-Derived Bounds on the Likelihood

To quantify how experimental data affects the accuracy requirement for the FSP, we consider single-molecule, single-cell data, such as that obtained using the technique of smFISH. This technique allows experimentalists to count the number of specific RNA molecules in individual cells, as described in Chapter 3. The solution for the FSP is $\mathbf{p}_J^{\text{FSP}}$ and is guaranteed to be a lower bound on the model's true solution $\mathbf{p} = [p(\mathbf{x}_1), p(\mathbf{x}_2), \ldots]$ by Eq. (2.4). The log-likelihood in Eq. 3.1 is monotonic in each $p(\mathbf{x}_i)$; therefore $\mathbf{p}_J^{\text{FSP}}$ provides a lower bound on the log-likelihood of $\mathbf{D}$ given the model,

$$\text{LB}_J(\mathbf{D}) \equiv \sum_{i \in \mathcal{I}_\mathbf{D}} d_i \log p_J^{\text{FSP}}(\mathbf{x}_i) \leq \sum_{i \in \mathcal{I}_\mathbf{D}} d_i \log p(\mathbf{x}_i). \tag{4.1}$$

However, in Eq. (2.3) the FSP also provides the exact error in the solution of a particular model described by the CME. By redistributing the known FSP error back onto the CME solution in an optimal manner, an upper bound on $\log L(\mathbf{D})$ can be derived,

$$\begin{aligned} \text{UB}_J(\mathbf{D}) \equiv & \max_{\{\varepsilon_i\}} \sum_{i \in \mathcal{I}_\mathbf{D}} d_i \log \left(p_J^{\text{FSP}}(\mathbf{x}_i) + \varepsilon_i \right) && \geq \sum_{i \in \mathcal{I}_\mathbf{D}} d_i \log p(\mathbf{x}_i) \\ & \text{such that: } \sum_i \varepsilon_i = g \text{ and } \varepsilon_i \geq 0, \end{aligned} \tag{4.2}$$

where ε_i is the probability error redistributed to state $\mathbf{x}_i$. To optimize the redistribution of g and determine $\text{UB}_J(\mathbf{D})$, we use a modified water-filling algorithm similar to those used to determine the amount of power to send to different channels in communications systems [44,45]. To simplify notation, we define the FSP probability for each state as $p_i \equiv p_J^{\text{FSP}}(\mathbf{x}_i)$ and the corresponding partial objective as $f_i \equiv d_i \log(p_i + \varepsilon_i)$. To determine which states have the highest impact on the likelihood, the derivative of f_i with respect to ε_i is computed from Eq. (4.2) to get

$$\left. \frac{\partial f_i}{\partial \varepsilon_i} \right|_{\varepsilon_i = 0} = \begin{cases} \frac{d_i}{p_i + \varepsilon_i} = \frac{d_i}{p_i} & \text{for } i \in \mathcal{I}_\mathbf{D} \\ 0 & \text{for } i \notin \mathcal{I}_\mathbf{D} \end{cases}, \tag{4.3}$$

and we define N_D as the number of distinct observations (i.e., the size of $\mathcal{I}_{\mathbf{D}}$). These values are then ranked in decreasing magnitude according to

$$\frac{\tilde{d}_1}{\tilde{p}_1} \geq \frac{\tilde{d}_2}{\tilde{p}_2} \geq ... \geq \frac{\tilde{d}_r}{\tilde{p}_r} \geq ... \geq \frac{\tilde{d}_{N_D}}{\tilde{p}_{N_D}}, \tag{4.4}$$

where the notation $\tilde{\cdot}$ refers to the data-ordered state space $\tilde{\mathbf{X}}_{\mathbf{D}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_r, ..., \tilde{\mathbf{x}}_{N_D}\}$. An optimal redistribution of g will equalize the first n terms of Eq. (4.4) and satisfy the linear constraints

$$\left.\begin{aligned} &\tilde{d}_{r+1}\epsilon_r - \tilde{d}_r\epsilon_{r+1} = \tilde{d}_r\tilde{p}_{r+1} - \tilde{d}_{r+1}\tilde{p}_r \\ &\epsilon_r \geq 0 \end{aligned}\right\} \text{for } r \in \{1, ..., n-1\}. \tag{4.5}$$
$$\sum_{j=1}^{n} \varepsilon_j = g$$

For example, when $n = 4$, ε_i can be directly solved from the following linear equation:

$$\begin{bmatrix} \tilde{d}_2 & -\tilde{d}_1 & 0 & 0 \\ 0 & \tilde{d}_3 & -\tilde{d}_2 & 0 \\ 0 & 0 & \tilde{d}_4 & -\tilde{d}_3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} = \begin{bmatrix} \tilde{d}_1\tilde{p}_2 - \tilde{d}_2\tilde{p}_1 \\ \tilde{d}_2\tilde{p}_3 - \tilde{d}_3\tilde{p}_2 \\ \tilde{d}_3\tilde{p}_4 - \tilde{d}_4\tilde{p}_3 \\ g \end{bmatrix}. \tag{4.6}$$

In this formulation, the number of states to which probability is redistributed, n, is the largest dimension for which the solution of Eq. (4.6) is strictly positive for all ε_i. If the states $\mathbf{X}_J$ used by the approximation do not span the support of the distribution of data, there will be s states for which $p_i = 0$ and $\frac{\partial f_i}{\partial \epsilon_i}\Big|_{\varepsilon=0}$ is infinite, and $\{\varepsilon_i\}$ will always include some mass for those states. Algorithm 1 provides pseudocode for the proposed error redistribution approach. At most, the FSP error redistribution algorithm requires $N_D - s$ iterations, and in practice computation of the upper bound on the likelihood takes only a fraction of the time needed for the FSP solution itself, especially for cases where the data corresponds to partial state observations.

Algorithm 2 FSP Error Redistribution Algorithm

Rank $\left.\frac{\partial f_i}{\partial \varepsilon_i}\right|_{\varepsilon_i=0\ \forall\ i}$

$n = 1$, $\varepsilon_1 = g$

while $\frac{\partial f}{\partial \varepsilon_{n+1}} > \frac{\partial f}{\partial \varepsilon_n}$ **and** $n < N_D$ **do**

$n \rightarrow n+1$

Solve for ε_1,ε_2,...,ε_n using Eq. 4.5

end while

$\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]$

The FSP-derived bounds on likelihoods have several important implications for the comparison of stochastic models to single-cell data. Let $\boldsymbol{\Lambda}$ denote a particular combination of a model and its parameters, and let $L(\mathbf{D}|\boldsymbol{\Lambda})$ denote the likelihood of $\mathbf{D}$ given $\boldsymbol{\Lambda}$. In the case when $\mathbf{X}_J = \varnothing$ (*i.e.*, the FSP set is empty), all of the probability mass must be redistributed, and the FSP-derived upper bound is given by:

$$\begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \ldots & \varepsilon_{N_D} \end{bmatrix} = \frac{1}{\sum_{i=0}^{N_D} \tilde{d}_i} \begin{bmatrix} \tilde{d}_1 & \tilde{d}_2 & \ldots & \tilde{d}_{N_D} \end{bmatrix}. \tag{4.7}$$

This result is easily understood – the maximum of the log-likelihood function occurs when the distribution of the data exactly matches the model, and in this case the FSP upper bound describes the best any potential model can ever do. To interpret bounds for non-trivial FSP projections, we make use of the facts that (i) the FSP approximation lower bound p_i increases monotonically as $\mathbf{X}_J$ is expanded [27] and (ii) the likelihood increases monotonically as each p_i increases. As a result, $\text{LB}_J(\mathbf{D})$ and $\text{UB}_J(\mathbf{D})$) are guaranteed to be monotonically increasing and decreasing functions of the projection size. Fig. 4.1 illustrates the converging upper- and lower-bounds for the likelihoods of two FSP models as the size of the index set J, or equivalently the size of $\mathbf{X}_J$, is increased.

4.2.1 Using FSP-Derived Bounds for Model Discrimination

For any two models and their parameter sets, $\mathbf{\Lambda}_i$ and $\mathbf{\Lambda}_j$, we define the set of sufficient discriminating projections, $\mathbf{\Phi}(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$, as any pair of projection index sets, J_i and J_j, that guarantees the correct ranking of likelihoods for the two models,

$$\begin{aligned}\mathbf{\Phi}(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j) \equiv \{J_i, J_j\} \text{ such that } &\text{UB}_{J_i}(\mathbf{\Lambda}_i) < \text{LB}_{J_j}(\mathbf{\Lambda}_j) \\ \text{or } &\text{UB}_{J_j}(\mathbf{\Lambda}_j) < \text{LB}_{J_i}(\mathbf{\Lambda}_i)\end{aligned} \tag{4.8}$$

Intuitively, these are any two projections such that the worst possible likelihood for one model is greater than best possible likelihood of the other. In Fig. 4.1, the red and green circles denote pairs of projections sufficient to guarantee that parameter set $\mathbf{\Lambda}_2$ is more likely than $\mathbf{\Lambda}_1$. Because $\mathbf{\Phi}(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ can contain an infinite set of such pairs with varying projection sizes, we define a minimal symmetric discriminatory projection, $\phi_s(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$, as

$$\phi_s(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j) \equiv \text{smallest set } J \text{ such that } \{J, J\} \in \mathbf{\Phi}(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j). \tag{4.9}$$

In Fig. 4.1, the blue circle denotes $\phi_s(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$. Finally, in many sequential parameter searches, previous FSP models may already be computed to high accuracy, and it may not be necessary to demand the same accuracy for subsequent models. For this case, we define a minimum non-symmetric projection size such that:

$$\phi_i(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j) \equiv \text{smallest set } J_i \text{ such that } \{J_i, J_j\} \in \mathbf{\Phi}(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j). \tag{4.10}$$

The utility of this particular discriminatory projection size definition becomes important in parameter search problems, where it enables sequential likelihood evaluations to be conducted at minimal projection sizes. In Fig. 4.1, the green circles denotes the $\phi_2(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ a particular combination where the size of $\mathbf{X}_{J_2}$ is minimized given a previous computation for $\mathbf{X}_{J_1}$. As the examples below will demonstrate, utilization of these minimal discriminating projections can substantially reduce

the computational effort in parameter inference by eliminating much of the potential parameter space with smaller projection sizes.

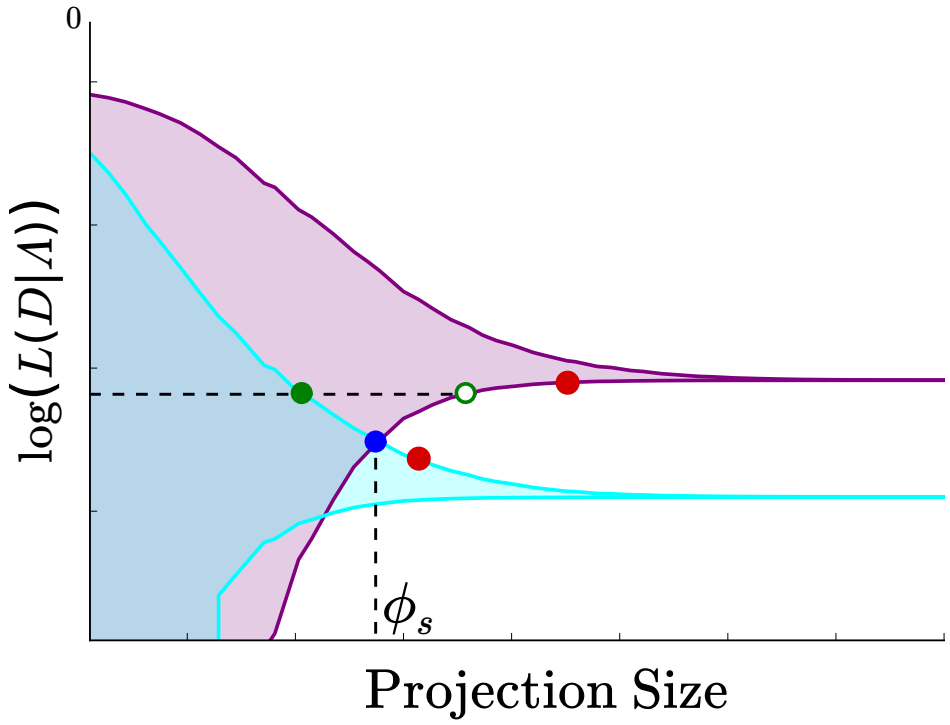

Figure 4.1: *Schematic of discriminatory projection sizes.* Monotonically upper and lower FSP bounds on the likelihood are shown for two parameter sets $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. The red and green circles illustrate two pairs of projections in $\mathbf{\Phi}(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ that enable exact ranking of the two parameter sets. The blue circle illustrates the minimal symmetric discriminatory projection, $\phi_s(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$. The filled green circle illustrates the minimal nonsymmetric discriminatory projection, $\phi_2(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$, needed to discriminate the system given the previous analysis of $\mathbf{\Lambda}_1$ (open green circle).

4.2.2 Relationship of FSP bounds to other CME truncations

The FSP upper and lower bounds present an opportunity to better understand relationships between the FSP and other CME approximations in the context of single-cell data. For example, several groups have imposed limits of species [46] or total molecule populations [47], which result in truncated master equations with reflecting boundary conditions (in contrast to the FSP's absorbing boundary condition). In a similar vein, one could renormalize the FSP solution to make use of the original FSP computation. For this renormalization, the CME can be written in terms similar

to the FSP solution as:

$$\frac{d}{dt}\mathbf{p}_J^{\text{renorm}} = \mathbf{A}_{JJ}\mathbf{p}_J^{\text{renorm}} + \alpha\mathbf{p}_J^{\text{renorm}} \tag{4.11}$$

where $\alpha = |\mathbf{A}_{JJ}\mathbf{p}_J^{\text{renorm}}|_1$ is the rate of flow of probability out of $\mathbf{X}_J$, which is now redistributed back into $\mathbf{X}_J$ according to the probability $\mathbf{p}_J^{\text{renorm}}$. However, the solution of this non-Markovian system is identical to simply renormalizing $\mathbf{p}_J^{\text{FSP}}$ at all times:

$$\mathbf{p}_J^{\text{renorm}} = \frac{\mathbf{p}_J^{\text{FSP}}}{|\mathbf{p}_J^{\text{FSP}}|_1} \tag{4.12}$$

This can easily be shown by substituting Eq. (4.12) into Eq. (4.11).

Because the FSP likelihood bounds provide the best- and worst-case redistribution of exiting probability, likelihoods computed by reflection, renormalization or any other arbitrary strategy are guaranteed to lay between the computed FSP bounds. Therefore, it is possible that reflecting boundaries may provide an improved approximation of the true likelihood for a particular combination of data and model. Unfortunately, the likelihoods of reflected or renormalized solutions are not necessarily monotonic and the likelihoods of renormalized solutions for two different parameter sets or models may change rank depending on the projection size. These issues will be addressed further in Section (4c).

4.3 Application of FSP-Derived Bounds

To demonstrate the application of the FSP bounds, this section uses simulated data and three examples of stochastic gene regulation: an unregulated birth-death model, a genetic toggle switch, and a non-linear self-activating gene.

4.3.1 Birth-Death Model

The variability in mRNA copy numbers for housekeeping genes is well captured by the standard single-species birth and death model. This model consists of two reactions that describe

transcription and degradation as shown in Fig. 4.2(a),

$$\mathcal{R}1: \ \varnothing \xrightarrow{w_1} x \quad \mathcal{R}2: \ x \xrightarrow{w_2} \varnothing$$

where the propensity functions $\mathbf{w} = \{w_1, \ w_2\}$ are

$$w_1 = k_r; \quad w_2 = \gamma x.$$

Table 4.1 shows three different parameter sets for this example. A well-known analytical solution for this model, assuming $x(0) = 0$, is the time-varying Poisson distribution

$$p(x_i|k_r, \gamma) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \tag{4.13}$$
$$\text{where } \lambda = \frac{k_r}{\gamma}(1 - e^{-\gamma t_f}).$$

The infinitesimal generator for this model can be written as

$$\mathbf{A}_{ji} = \begin{cases} -w_1(x_i) - w_2(x_i) & \text{for } i = j \\ w_1(x_i) & \text{for } (i,j) \text{ such that } x_j = x_i + 1 \\ w_2(x_i) & \text{for } (i,j) \text{ such that } x_j = x_i - 1 \\ 0 & \text{otherwise} \end{cases}. \tag{4.14}$$

The FSP formulation for this model can be written from Eq. (4.14) as

$$\frac{d}{dt}\begin{bmatrix} \mathbf{p}_0 \\ \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \\ g(t) \end{bmatrix} = \begin{bmatrix} -k_r & \gamma & 0 & \dots & 0 \\ k_r & -k_r - \gamma & 2\gamma & \ddots & 0 \\ 0 & k_r & -k_r - 2\gamma & \ddots & 0 \\ \vdots & \ddots & \ddots & N_m\gamma & \vdots \\ 0 & 0 & k_r & -k_r - N_m\gamma & 0 \\ 0 & 0 & 0 & k_r & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}_0 \\ \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_N \\ g(t) \end{bmatrix}, \tag{4.15}$$

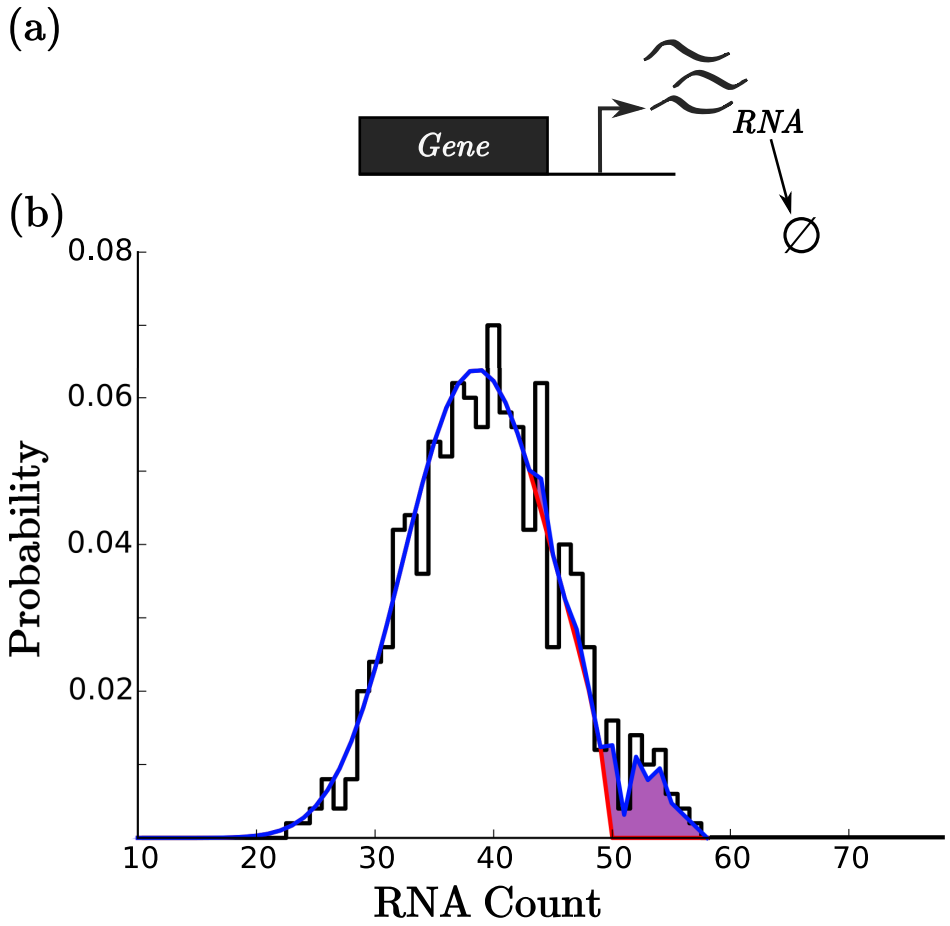

Figure 4.2: *FSP bounds for the birth/death model.* (a) Schematic of the RNA birth and death process. (b) Probability distributions for the RNA birth and death process at $t = 1$. Simulated data is in black. The lower bound Eq. (2.4) is shown in red, and the upper bound Eq. (4.2) is shown in blue. The shaded region denotes the redistribution of FSP error to maximize the likelihood of data. See Table 4.1 for parameters.

Where N_m is the size of the FSP truncation (*i.e.* $\mathbf{X}_J = \{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_{N_m}\}$). In this case, we used a single constraint function in Eq. (4.5) $f_1(x_i) = x_i \leq c_1$, where $c_1 = N_m$. To expand the state space for this model, c_1 is simply increased by one in each iteration.

For this model and the parameters provided in the bottom row of Table 4.1, we simulated 500 trajectories of the stochastic simulation algorithm [33], and plot the resulting “data" in Fig. 4.2(b) (black). For a projection sized defined by $N_m = 50$, Fig. 4.2(b) also shows the FSP lower computed using Eq. (2.4) in red and the FSP upper bounds from Eq. (4.2) in blue. Figure 4.3 demonstrates the convergence of the upper and lower FSP bounds as c_1 is increased for two different parameter sets. Increasing c_1 adds more states to $\mathbf{X}$ and monotonically decreases the error g. In turn, less error is

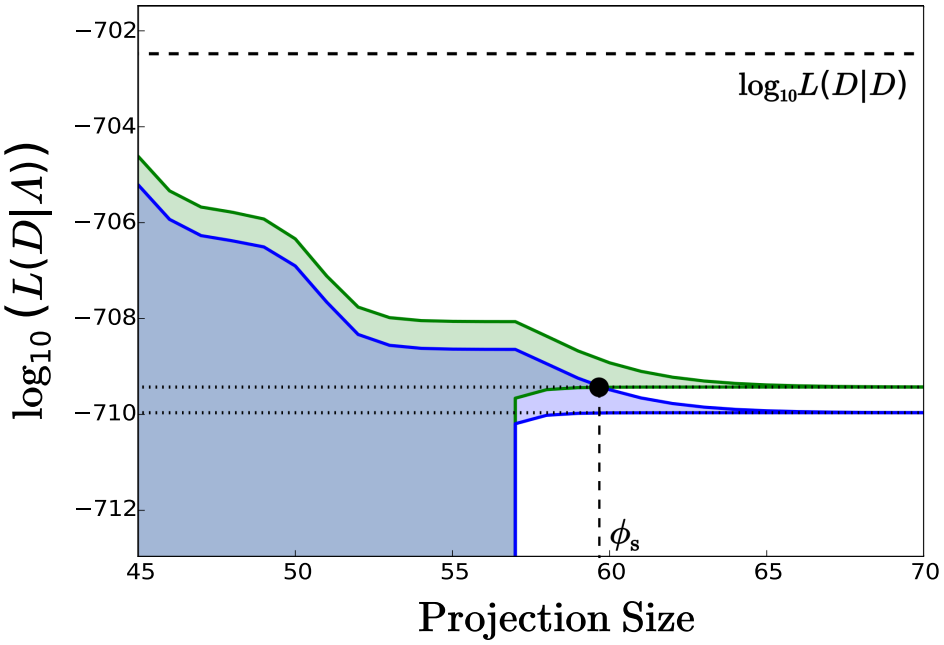

Figure 4.3: *Demonstration of the converging bounds for the birth and death model.* Upper and lower bounds on the likelihood of a simulated data set given two different parameter sets, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, as a function of the number of states included in the birth-death model. As the number of states increases, the upper and lower bounds monotonically converge to the true likelihood of each parameter set. The horizontal dashed lines are the likelihood values found using the analytical solutions in Eq. (4.13). ϕ_s indicates the minimum symmetric projection size that *guarantees* correct discrimination between the two parameter sets. See Table 4.1 for parameters.

Table 4.1: Birth and death model parameters

$\mathbf{\Lambda}$	$k_r(Nt^{-1})$	$\gamma(t^{-1})$
$\mathbf{\Lambda}_1$	50.0	0.5
$\mathbf{\Lambda}_2$	49.65	0.5
$\hat{\mathbf{\Lambda}}$	45.0	0.5

available to be distributed to the FSP solution, and $\text{UB}_J(\mathbf{\Lambda})$ and $\text{LB}_J(\mathbf{\Lambda})$ converge monotonically to the analytical value of $\log L(\mathbf{D}|\mathbf{\Lambda})$ as shown by the horizontal dashed lines.

4.3.2 Toggle Model

We next explore the application of the FSP bounds on the classic toggle model for two mutually inhibiting genes, *λcI* and *lacI*, as illustrated in Fig. 4.4(a). The first synthetic toggle switch was experimentally constructed by Gardner et al [48], but here we consider a simple model similar to that presented by Tian and Burrage [49]. For this model, each state is defined by the the discrete number of each protein, $\mathbf{x} = [\lambda\text{cI}\ \ \text{LacI}]$. The four reactions are:

$$\mathcal{R}1: \ \varnothing \xrightarrow{w_1} \lambda\text{cI}; \quad \mathcal{R}2: \ \lambda\text{cI} \xrightarrow{w_2} \varnothing;$$
$$\mathcal{R}3: \ \varnothing \xrightarrow{w_3} \text{LacI}; \quad \mathcal{R}4: \ \text{LacI} \xrightarrow{w_4} \varnothing;$$

where the propensity functions $\mathbf{w} = \{w_1, w_2, w_4, w_4\}$, are given by

$$w_1 = b_{\lambda\text{cI}} + \frac{k_{\lambda\text{cI}}}{1 + \alpha_{\text{LacI}}\text{LacI}^{\eta_{\text{LacI}}}}; \quad w_2 = \gamma_{\lambda\text{cI}} \cdot \lambda\text{cI};$$
$$w_3 = b_{\text{LacI}} + \frac{k_{\text{LacI}}}{1 + \alpha_{\lambda\text{cI}}\lambda\text{cI}^{\eta_{\lambda\text{cI}}}}; \quad w_4 = \gamma_{\text{LacI}} \cdot \text{LacI}.$$

The toggle model parameters are shown in Table 4.4, which have been used to generate simulated data shown in black in Fig. 4.4(b).

The infinitesimal generator for this toggle model can be written:

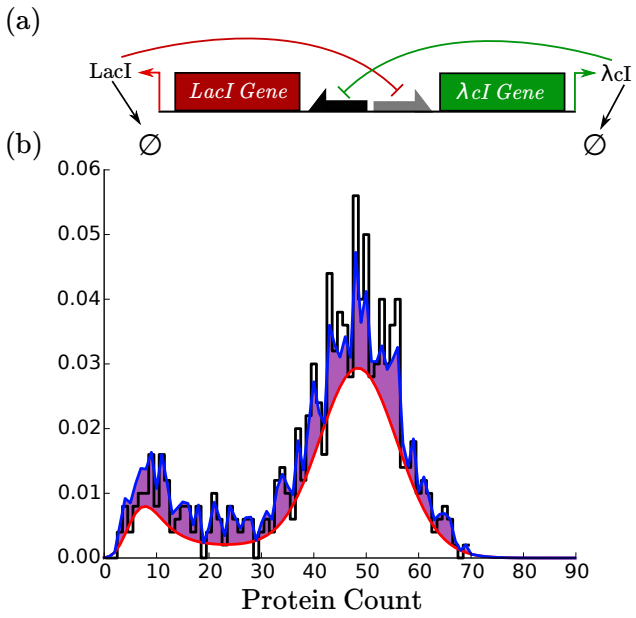

Figure 4.4: *FSP bounds for the toggle model.* (a) Schematic of the toggle model with two mutually repressing proteins, LacI and λcI. (b) Marginal probability distributions of LacI at $t = 8$ hrs. Simulated data is in black. The lower bound Eq. (2.4) is shown in red and the upper bound Eq. (4.2) is shown in blue. The shaded region denotes the redistribution of FSP error to maximize the likelihood of data. See Table 4.4 for parameters.

Table 4.2: Toggle model parameters

$\mathbf{\Lambda}$	$b_{\lambda\text{cI}}$ (s^{-1})	$k_{\lambda\text{cI}}$ (s^{-1})	α_{LacI} $(N^{-\eta_{\text{LacI}}})$	η_{LacI} ()	$\gamma_{\lambda\text{cI}}$ $(N^{-1}s^{-1})$	b_{LacI} (s^{-1})	k_{LacI} (s^{-1})	$\alpha_{\lambda\text{cI}}$ $(N^{-\eta_{\lambda\text{cI}}})$	$\eta_{\lambda\text{cI}}$ ()	γ_{LacI} $(N^{-1}s^{-1})$
$\mathbf{\Lambda}_1$	6.8e-5	1.6e-2	6.1e-3	2.1	6.7e-4	2.2e-3	1.7e-2	2.6e-3	3.0	3.8e-4
$\mathbf{\Lambda}_2$	6.8e-5	1.6e-2	6.1e-3	2.1	8.0e-4	2.2e-3	1.7e-2	2.6e-3	3.0	3.8e-4
$\hat{\mathbf{\Lambda}}$	6.8e-5	1.4e-2	6.1e-3	2.1	6.7e-4	2.2e-3	1.6e-2	2.6e-3	3.0	3.8e-4

$$\mathbf{A}_{ji} = \begin{cases} -\sum_{\mu=1}^{4} w_\mu(\mathbf{x}_i) & \text{for } i = j \\ w_1(\mathbf{x}_i) & \text{for } (i,j) \text{ such that } \mathbf{x}_j = \mathbf{x}_i + [1,\ 0] \\ w_2(\mathbf{x}_i) & \text{for } (i,j) \text{ such that } \mathbf{x}_j = \mathbf{x}_i + [-1,\ 0] \\ w_3(\mathbf{x}_i) & \text{for } (i,j) \text{ such that } \mathbf{x}_j = \mathbf{x}_i + [0,\ 1] \\ w_4(\mathbf{x}_i) & \text{for } (i,j) \text{ such that } \mathbf{x}_j = \mathbf{x}_i + [0,\ -1] \\ 0 & \text{elsewhere} \end{cases} \tag{4.16}$$

To apply the FSP to truncate the toggle model CME, we consider three constraint functions from Eq. (2.10), where c_1, c_2 and c_3 define the projection as

$$\mathbf{X}_J = \{\mathbf{x}_i\} \text{ such that } \begin{cases} f_1(\mathbf{x}_i) = (\text{LacI} - 4)(\lambda\text{cI} - 4) \leq c_1 \\ f_2(\mathbf{x}_i) = \text{LacI} \leq c_2 \\ f_3(\mathbf{x}_i) = \lambda\text{cI} \leq c_3 \end{cases} \tag{4.17}$$

These constraints are illustrated in Fig. 4.5. Fig. 4.4(b) shows the marginal probability distribution for LacI with $c_1 = 150$, $c_2 = 95$, and $c_3 = 55$. This plot shows the FSP lower bound in red and FSP upper bound in blue. Although Algorithm 1 distributes the error onto the joint probability distribution of both species, results are plotted only for the marginal distribution. By monotonically increasing c_k, more states are included, and the error $g(t)$ decreases. Fig. 4.6 shows the converging bounds for two parameter sets, where the total numbers of states satisfying the constraints in Eq. (4.17) is represented on the x-axis. For simplicity in presentation, c_2 and c_3 were initially set at high values of 95 and 55, respectively, and the expansion only modifies the criteria c_1. Similar results can be obtained with more general expansion routines, provided that c_k are constant or monotonically increasing for each k.

4.3.3 Comparing FSP Bounds to Other CME Truncation Approaches

To illustrate potential issues that arise through application of other CME truncation approaches, we consider a simple gene regulation model with nonlinear self-activation. In this model, the

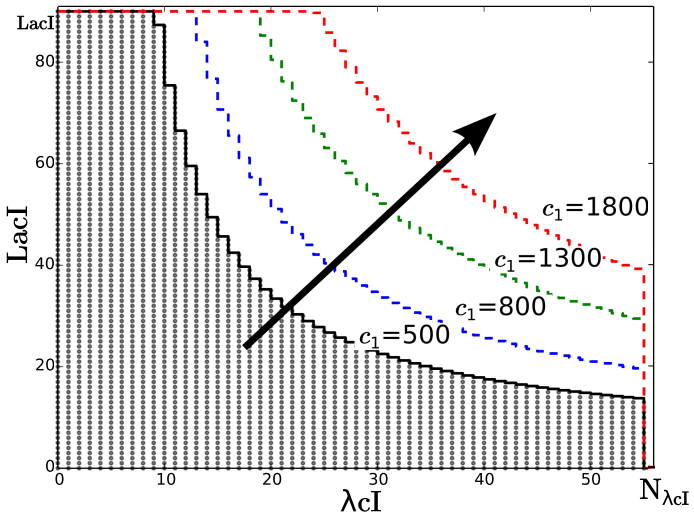

Figure 4.5: *FSP state space expansion.* Maximum species counts, as in Eqs. (4.17b) and (4.17c) are $c_2 = N_{\text{LacI}} = 95$ and $N_{\lambda_{\text{cI}}} = 55$. States included within a truncation of $c_1 = 500$ are in gray. As c_1 is increased, the boundary (dashed lines) increases to include more states, the FSP error decreases, and the FSP bounds converge to one another (see also Fig. 4.6).

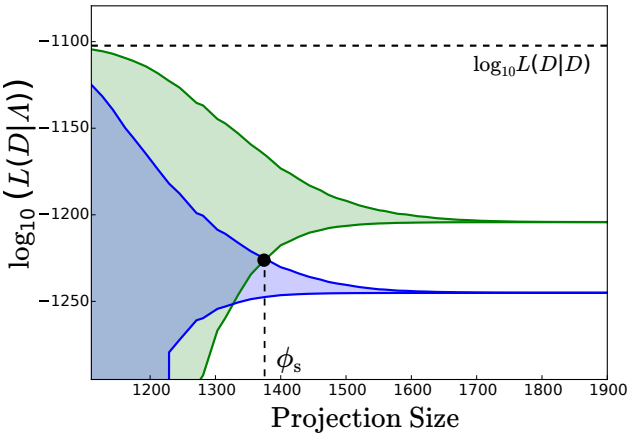

Figure 4.6: *Demonstration of the converging bounds for a two dimensional system.* Upper and lower bounds on the likelihood of a simulated data set given two different parameter sets, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, as a function of the number of states included in the toggle model. Parameters are given in Table III.

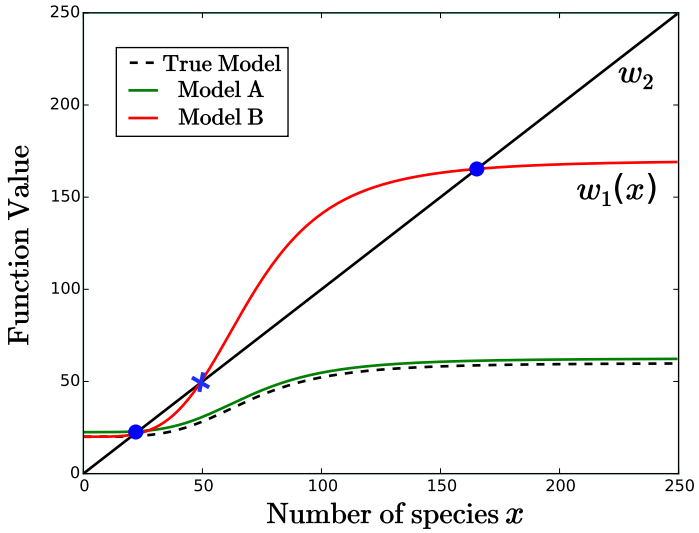

Figure 4.7: *Positive self-regulated gene expression.* First order degradation (black line) and positive feedback in production (red, green, dashed) can give rise to bistable (red) or monostable dynamics (green, dashed). Blue dots represent stable equilibria and the cross represents an unstable equilibrium. See Table 4.3 for parameters and Fig. 4.8(c) for examples of the corresponding distributions.

propensity of birth is given by the positive feedback function

$$w_1(x) = k_1 + k_2\Big(\frac{x^n}{m^n + x^n}\Big), \tag{4.18}$$

and the propensity of production is a first order process given by $w_2 = \gamma x$. In this formulation, k_1 is the rate of production for small values of x; $k_1 + k_2$ is the rate of production for large values of x; m is the value of x at which the rate of production is halfway between k_1 and $k_1 + k_2$; and the cooperativity factor n determines the steepness of the function as it moves from k_1 to $k_1 + k_2$. In the deterministic regime, this model of self-regulated gene expression can lead to one or two stable equilibria as illustrated in Fig. 4.7. In the stochastic regime, this corresponds to bimodal distributions in some parameter regimes, and unimodal distributions in other regimes. Table 4.3 provides three possible parameter sets for this model. We simulated data from the first of these models, $\mathbf{\Lambda}_{\text{True}}$, which admits a single stable point (see the dashed line in Fig. 4.7) and yields a unimodal distribution of data as shown in black in Fig. 4.8.

Next, we consider two perturbations of this true parameter set,

Table 4.3: Effective parameters for counterexample

$\mathbf{\Lambda}$	k_1	k_2	m	n	γ
$\mathbf{\Lambda}_{\text{True}}$	20	40	70	4	1
$\mathbf{\Lambda}_A$	22.5	40	70	4	1
$\mathbf{\Lambda}_B$	20	125	70	4	1

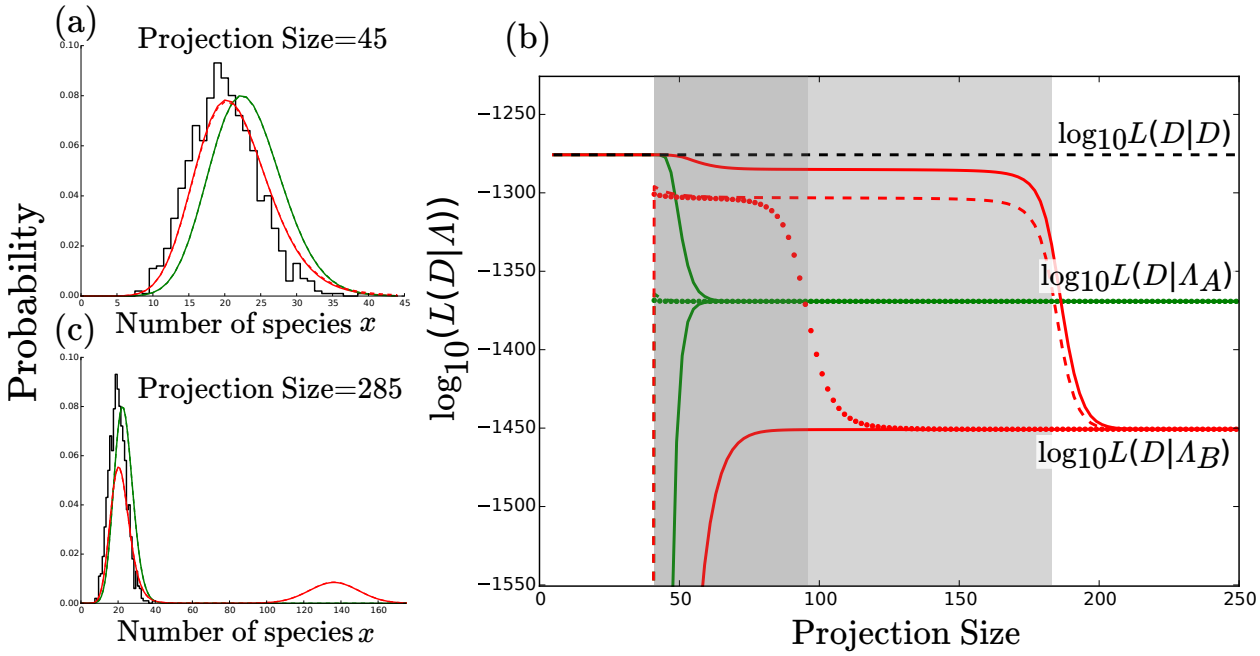

Figure 4.8: *Comparing FSP bounds with other CME truncation approaches.* (a) Distributions of response of the sel-f-regulated gene. Simulated data is in black. The renormalized FSP solutions for $\mathbf{\Lambda}_B$ is shown in red for $\mathbf{\Lambda}_B$ and in green for $\mathbf{\Lambda}_A$. Both methods use a projection size of 45. (b) Likelihood versus projection size for different CME truncations. The renormalized scheme is shown with dashed lines, and the reflecting scheme is shown with dotted lines. All schemes lie within the FSP bounds (solid lines) and eventually approach the correct likelihood values. Results for parameter set $\mathbf{\Lambda}_A$ are in green and for $\mathbf{\Lambda}_B$ are in red. At moderate projections sizes, the renormalized and reflecting boundary scheme appear to converge to a higher likelihood for $\mathbf{\Lambda}_B$ than for $\mathbf{\Lambda}_A$. At higher projection sizes, the trend is switched. (c) The same as (a), but now the projection size has been increased to XXX.

$$\begin{aligned} \mathbf{\Lambda}_A &= [\quad k_1(1-\varepsilon) \quad k_2 \quad n \quad m \quad \gamma \quad] \\ \mathbf{\Lambda}_B &= [\quad\quad k_1 \quad\quad \alpha k_2 \quad n \quad m \quad \gamma \quad] \end{aligned} \tag{4.19}$$

where $\mathbf{\Lambda}_A$ and $\mathbf{\Lambda}_B$ correspond to the red and green lines, respectively, in Figs. 4.7 and 4.8, and their parameters are given in Table 4.3. For $\mathbf{\Lambda}_B$ the system has a bimodal response and for $\mathbf{\Lambda}_A$, the response is unimodal. It is interesting to explore how the application of the renormalization scheme in Eq. (4.12) would affect comparison of these two models to the simulated data. For a projection size of 45, Fig. 4.8(a) suggests that $\mathbf{\Lambda}_B$ provides a better match to the data than $\mathbf{\Lambda}_A$. Moreover, the likelihood of the data given $\mathbf{\Lambda}_A$ and $\mathbf{\Lambda}_B$ appear to be nearly constant over a substantial portion of the projection space as shown by the dashed lines and the shaded region of Fig. 4.8(b). Based on this information, it would be easy to conclude that $\mathbf{\Lambda}_B$ is the more appropriate parameter set. However, only at large projections which include the second peak in the distribution for $\mathbf{\Lambda}_B$, does it becomes apparent that the $\mathbf{\Lambda}_A$ is the better choice. This scenario illustrates how the renormalization scheme can complicate parameter discrimination for certain combinations of models and data. Similar cautions also apply for reflecting boundary approximations of the CME (see dotted lines in Fig. 4.8). The strict upper and lower bounds provided by the FSP eliminates this ambiguity as a function of projection size.

It should also be noted that likelihood computations using reflection or renormalization based truncations require the support of the CME to include that of the experimental data. Otherwise, these approaches will match the FSP lower bound that suggests that the data is infinitely unlikely. Such a lower bound may appear uselessness at first, but as we will see in the next section, the FSP upper bound may still be sufficient for rigorous and efficient model selection.

4.4 FSP Likelihood Bounds in Parameter Searches

The FSP's constricting upper and lower bounds on the likelihood enable rigorous discrimination between two parameter sets, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$, without using unnecessarily strict error tolerances.

The following examples will demonstrate the utility of the sufficient discriminatory projections, $\mathbf{\Phi}(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$, $\phi_s(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$, and $\phi_i(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$.

4.4.1 Parameter Search for the Birth-Death and Toggle Models

We return to the simulated data presented in Figs. 4.2 and 4.4, but this time we apply the FSP for many different parameter combinations and for many different projections. Fig. 4.9 illustrates the practical strength of the minimal symmetric discriminating projection, $\phi_s(\mathbf{\Lambda}_i, \mathbf{\Lambda}_j)$. For the birth/death model in Figs. 4.9(a,b), parameter $\mathbf{\Lambda} = k_r$ is allowed to vary. Fig. 4.9(a) shows the likelihood of the data as a function of parameter $\mathbf{\Lambda}$, and Fig. 4.9(b) shows the size of the sufficient symmetric projection, $\phi_s(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}})$, needed to discriminate between $\mathbf{\Lambda}$ and a fixed parameter $\hat{\mathbf{\Lambda}}$. Similarly, for the toggle model, both the maximum rates of production for LacI and λcI were varied, $\mathbf{\Lambda} = \begin{bmatrix} k_{\lambda cI} & k_{LacI} \end{bmatrix}$. In this case, Fig. 4.9(c) shows contour plots of the likelihoods and Fig. 4.9(d) shows the corresponding contours of the size of $\phi_s(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}})$. For models whose likelihood is better or worse than $\hat{\mathbf{\Lambda}}$, Figs. 4.9(b,d) shows that the comparison can be made with smaller projection sizes. Considering that the solution of Eq. 2.3 has a complexity that is typically $\mathcal{O}(n^2)$ or worse depending upon the solution scheme [50], such reductions can lead to substantial computational savings. In past studies, the FSP has been solved to uniformly strict error tolerances such as 10^{-6} in Neuert et al. [11], yet consideration of the FSP bounds allows for error tolerances that are relaxed by several orders of magnitude.

4.4.2 FSP bounds on STL1 regulation in yeast

To demonstrate the application of the FSP bounds on real data, we examine a recent model and single-cell experimental data for Mitogen Activated Protein Kinase (MAPK) control of *STL1* gene regulation in *S. cerevisiae* (budding yeast). Yeast activate a variety of regulatory pathways to mitigate the osmotic pressure difference that arises from solute imbalance across the cell membrane. One such mechanism is the high-osmolarity glycerol (HOG) pathway in yeast [51]. Upon osmotic shock, the Hog1 kinase phosphorylates and localizes to the nucleus of the cell, where it initiates transcription of several genes, including *STL1*. After cells adapt to the new condition, the

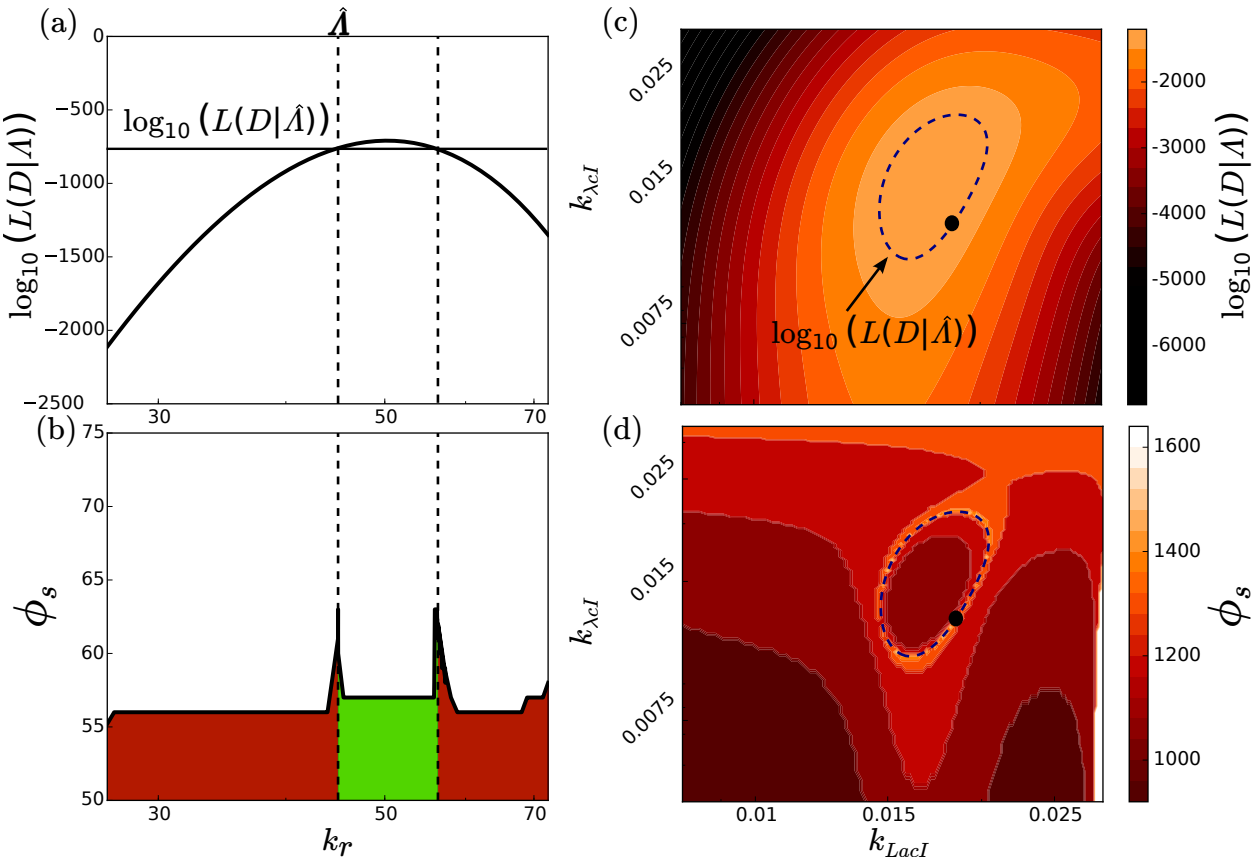

Figure 4.9: *Likelihoods and sufficient projection sizes.* (a) Log-likelihood versus RNA production rate k_r. The horizontal line denotes the likelihood at the fixed comparison parameter $\hat{\mathbf{\Lambda}}$. (b) The symmetric projection size ϕ_s required to compare parameter set $\mathbf{\Lambda}$ to $\hat{\mathbf{\Lambda}}$. For $\mathbf{\Lambda}$ such that $\log L(\mathbf{D}|\mathbf{\Lambda}) \cong \log L(\mathbf{D}|\hat{\mathbf{\Lambda}})$, larger projections are needed. The green (red) region represents parameter sets that are better (worse) than $\hat{\mathbf{\Lambda}}$. (c,d) Same as (a,b), except for the toggle model and two variable parameters, k_{LacI} and $k_{\lambda cI}$. Parameters inside (outside) the dashed contour represent parameter sets are better (worse) than the comparison set denoted with the black dots. In all plots, dashed lines denote parameter combinations with likelihoods that are equivalent to the reference set.

kinase leaves the nucleus, and the transcription pathways turn back off. Interestingly, while the nuclear localization and transcription initialization is a largely deterministic temporal signal, transcript abundance varies considerably between isogenic cells. This variability has been quantified in detail, using the smFISH technique to quantify transcript abundances in single cells at sixteen different time points at various times from zero to 60 minutes after osmotic shock [11].

The current time-varying CME model of the *STL1* regulation process allows the gene to switch between four possible states with different transcription rates as shown in Fig. 4.10(a). Reactions that change the gene from state i to j occur with propensities $\{k_{ij}\}$, and the transcription rates are given by k_{r_i}, for each of the $i = \{1, 2, 3, 4\}^{\text{th}}$ gene states. In this model, one particular transition rate k_{21} varies as a function of the Hog1p kinase in the nucleus as:

$$k_{21}(t) = \max\{0, \alpha + \beta \text{Hog1p}(t)\}, \tag{4.20}$$

where the temporal signal profile for Hog1p was measured experimentally [11] and is reproduced in Fig. 4.10(b). As a result of this dependence on a time-varying parameter, the infinitesimal generator for the CME is a function of time. The FSP truncation of the CME can be written as:

$$\frac{d}{dt}\begin{bmatrix} \mathbf{p}_0 \\ \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{P}_{N_m} \\ g(t) \end{bmatrix} = \begin{bmatrix} \mathbf{S}-\mathbf{T} & \mathbf{\Gamma} & \mathbf{0} & \dots & 0 \\ \mathbf{T} & \mathbf{S}-\mathbf{T}-\mathbf{\Gamma} & 2\mathbf{\Gamma} & \ddots & 0 \\ \mathbf{0} & \mathbf{T} & \mathbf{S}-\mathbf{T}-2\mathbf{\Gamma} & \ddots & 0 \\ \vdots & \ddots & \ddots & \mathrm{N_m}\mathbf{\Gamma} & 0 \\ \mathbf{0} & \dots & \mathbf{T} & \mathbf{S}-\mathbf{T}-\mathrm{N_m}\mathbf{\Gamma} & 0 \\ \mathbf{0} & \dots & \dots & \mathbf{1}^T\mathbf{T} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}_0 \\ \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{P}_{N_m} \\ g(t) \end{bmatrix}, \tag{4.21}$$

where the matrices $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{\Gamma}$ are given by:

$$\mathbf{S}(t) = \begin{bmatrix} -k_{12} & k_{21}(t) & 0 & 0 \\ k_{12} & -k_{21}(t) - k_{23} & k_{32} & 0 \\ 0 & k_{23} & -k_{32} - k_{34} & k_{43} \\ 0 & 0 & k_{34} & -k_{43} \end{bmatrix};$$

$$\mathbf{T} = \begin{bmatrix} k_{r_1} & 0 & 0 & 0 \\ 0 & k_{r_2} & 0 & 0 \\ 0 & 0 & k_{r_3} & 0 \\ 0 & 0 & 0 & k_{r_4} \end{bmatrix};$$

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma & 0 & 0 & 0 \\ 0 & \gamma & 0 & 0 \\ 0 & 0 & \gamma & 0 \\ 0 & 0 & 0 & \gamma \end{bmatrix}. \tag{4.22}$$

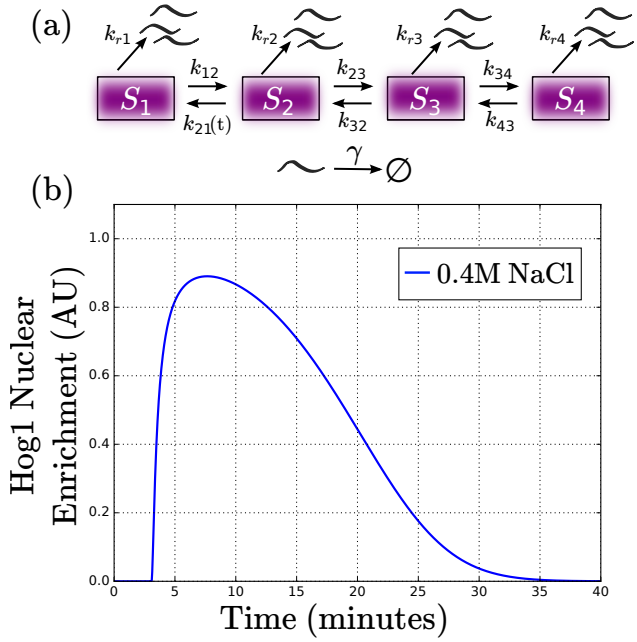

Figure 4.10: *Gene regulation in the HOG-STL1 system.* (a) The four-state model of Hog1p-induced *STL1* gene regulation, in which each gene state ($S_1 \ldots S_4$) has a distinct transcription rate. (b) The parameterized nuclear enrichment signal, Hog1p(t), that controls the rate of transition from S_2 to S_1. This signal was parameterized from experimental measurements at 0.4M NaCl by Neuert *et al* [11].

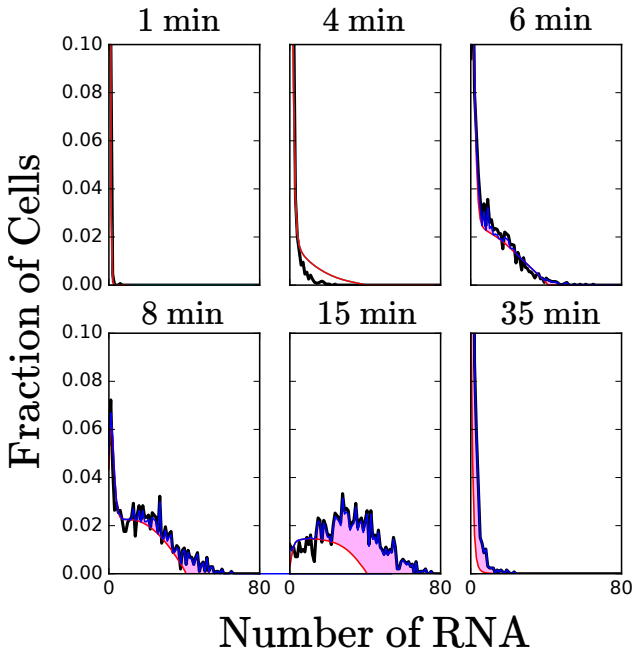

Figure 4.11: *FSP bounds on* STL1 *mRNA Distributions.* Experimentally measured distributions of *STL1* transcripts are in black for each time point [11]. The FSP lower bounds are shown in red and upper bounds are shown in blue. The Hog1-STL1 pathway is activated at $t = 0$ with a 0.4M treatment of NaCl (see also Fig. 4.10).

Fig. 4.11 shows examples of distributions for six time points during the osmotic shock response. Experimental data [11] (black) were collected using smFISH, and the FSP lower bound for a moderate projection size is shown in red. At each experimentally measured time point during the dynamic process, the FSP error $g(t)$ is computed based upon the FSP truncation, and the FSP upper bound on $\log L(\mathbf{D}|\mathbf{\Lambda})$ (shown in blue) is computed using Algorithm 1. In this illustrative example, we note that the projection size is substantially smaller than the support of the experimental data, yet the reduced FSP adequately captures the distribution, especially for the earlier time points. Because a good model must capture both early and late time points, this observation suggests that smaller projections may be quite informative for model discrimination.

To explore the impact that experimental data has on the required projection for the FSP, we use the non-symmetric minimal projection to determine the necessary projection size needed for parameter discrimination. In this case, each sequential comparison of two models begins with a previous FSP model that is already solved to a known precision. For example, in Fig. 4.12, we

plot the FSP error bounds versus the projection size for two Hog-STL1 model parameter sets, $\mathbf{\Lambda}_{\mathrm{N}}$ and $\mathbf{\Lambda}_{\mathrm{N+1}}$. If the likelihood is already known for the N^{th} parameter set (horizontal dashed line in Figs. 4.12(a,b)), then the FSP model for the $(\mathrm{N}+1)^{\mathrm{th}}$ parameter set need only be solved with a projection size corresponding to $\phi_{\mathrm{N+1}}$. Figure 4.12 represents this minimal nonsymmetric projection size with black circles. In many cases, such as that shown in Fig. 4.12(b), the new parameter is worse than the previous case, and the necessary discriminating projection size can be much smaller than the support of the experimental data. Such situations where the next parameter set is worse that the current set are the norm in a typical parameter search.

Fig. 4.13(a) shows the likelihood of the experimental data versus two parameters in the Hog-STL1 model, and Fig. 4.13(b) shows the size of the necessary discriminating projection for the new $\mathbf{\Lambda}_{\mathrm{N+1}}$ given that the old $\mathbf{\Lambda}_{\mathrm{N}}$ is at the black circle. For new parameters that give smaller likelihoods than $\mathbf{\Lambda}_{\mathrm{N}}$ (*i.e.* those outside of the halo), parameter discrimination can be achieved with projection sizes that are a fraction of the support of the data. In fact the median projection needed to compare the old and new models is 70, compared to a data support size of 107.

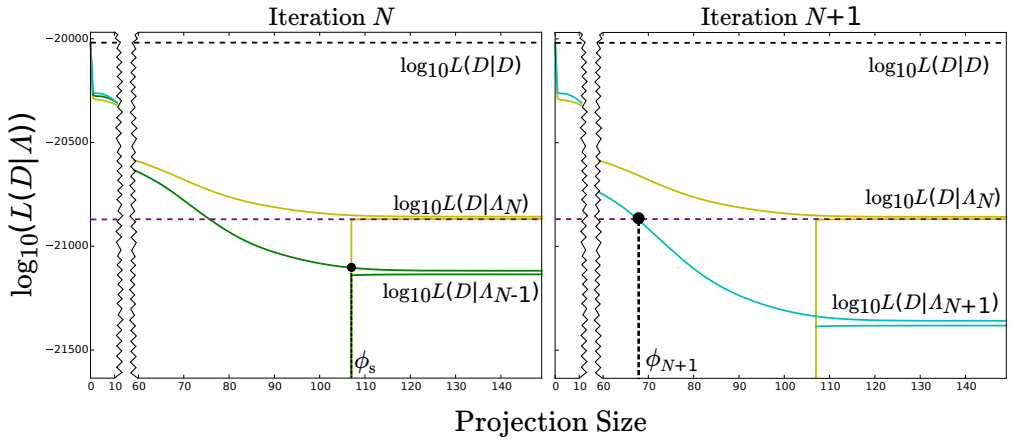

Figure 4.12: *Using FSP-bounds to search Hog1-STL1 models.* (a) FSP upper and lower bounds versus projection size for old parameters $\mathbf{\Lambda}_{N-1}$ and new parameters $\mathbf{\Lambda}_{N}$. In this case, $\mathbf{\Lambda}_{N}$ is better, and sufficient discrimination is made at ϕ_s, which corresponds to the support of the experimental data. (b) Comparison of the bounds for old parameters $\mathbf{\Lambda}_{N}$ and new parameters $\mathbf{\Lambda}_{N+1}$. In this case, reusing the FSP bounds for $\mathbf{\Lambda}_{N}$ makes it possible to reject $\mathbf{\Lambda}_{N+1}$ at a projection size that is less than the support of the experimental data.

Table 4.4: Hog model parameters

$\mathbf{\Lambda}$	k_{12}	α	β	k_{23}	k_{32}	k_{34}	k_{43}	k_{r1}	k_{r2}	k_{r3}	k_{r4}	γ
$\mathbf{\Lambda}_{N-1}$	2.096	5406.3	25116.6	0.00979	0.00868	0.0448	0.465	9.16e-4	0.01232	0.1372	1.953	5.53e-3
$\mathbf{\Lambda}_{N}$	2.096	5406.3	18116.6	0.00779	0.00668	0.0448	0.465	9.16e-4	0.01232	0.1072	1.953	5.53e-3
$\mathbf{\Lambda}_{N+1}$	2.096	5406.3	30116.6	0.01379	0.01068	0.0448	0.465	9.16e-4	0.01232	0.1372	1.953	5.53e-3
$\mathbf{\Lambda}_{fixed}$	2.096	5406.3	18116.6	0.02	0.00668	0.0448	0.465	9.16e-4	0.01232	0.1072	1.953	5.53e-3

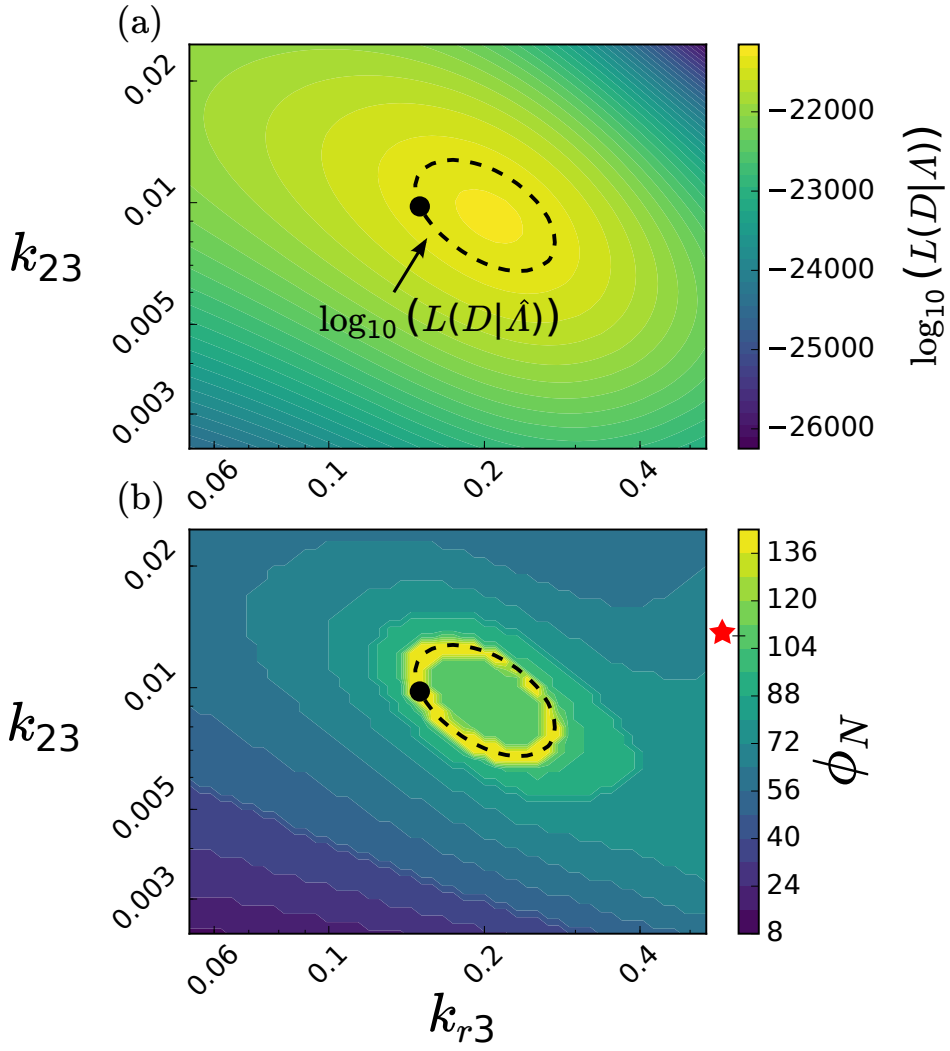

Figure 4.13: *Likelihoods and sufficient projection sizes for Hog1-STL1 data.* (a) The likelihood of smFISH data for 2500 different parameter combinations of k_{r_3} and k_{23}. (b) The size of $\phi_{\mathrm{N}+1}$ versus $\mathbf{\Lambda}_{\mathrm{N}+1} = [k_{r_3}, k_{23}]$, where $\mathbf{\Lambda}_{\mathrm{N}} = \hat{\mathbf{\Lambda}}$. Over most of the parameter space, sufficient discrimination does not require the full support of the experimental data ($N_m = 107$).

4.5 Summary and Conclusions

In recent years, substantial interest has arisen to integrate discrete stochastic models with single-cell experimental data. This has motivated many approaches to solve the chemical master

equation, including stochastic simulations, moment closure analyses, and the finite state projection approach. Progress in this arena will continue in the future to open new biochemical processes for discrete stochastic, computational analyses. However, until now there has been little discussion of how accurate models need to be in order to adequately interpret experimental data. In this article, we have explored the benefit by which careful consideration of experimental data can help to reduce computational complexity and enable more efficient and rigorous comparison of multiple models in the context of experimental data. We have shown that this advance can substantially reduce the complexity of model identification for single-cell gene regulation models using real data, and we believe this approach opens new doors for gene regulation models in many pathways and organisms.

In light of the above results, it would interesting to reexamine other approaches to fit stochastic models to single-cell data. For example, a common and highly flexible tool for this task is the stochastic simulation algorithm (SSA [33]). As one runs more and more SSA trajectories, the collected statistics converge to the solution of the CME, and the computed likelihood of the data given the model will also converges to the correct value. Unlike the FSP approach derived here, convergence of the SSA or other kinetic Monte Carlo approaches will not be monotonic, and long distribution tails can be very difficult to estimate. However, although the SSA does not provide a direct computation or bounds on its computational error, one can estimate the rate of convergence with increasing numbers of trajectories. With these one could imagine that the insight gained from the optimal redistribution of the FSP error could be adapted to explore similar *ad hoc* redistribution methods for the SSA. Such analyses provide intriguing paths for future theoretical and computational investigations.

Finally, it is now well established that stochastic models can help to better understand single-cell gene regulatory responses. Here, we have complemented this fact by showing how single-cell data may inform the design of rigorous and yet more efficient computational analyses. Together, these insights offer further motivation for tighter integration and co-design of computational and experimental investigations of biological phenomena.

Chapter 5

Fast Parameter Identification of Models of Stochastic Gene Regulatory Networks Using Data-Driven Radial Basis Function Model Reduction [2]

5.1 Introduction

The ability to model gene regulatory networks has significant ramifications in scientific fields such as molecular biology and medicine. When species exist in large numbers, as often encountered in biochemical engineering, they can be treated as continuous quantities modeled by deterministic ordinary differential equations [52]. However, certain biochemical species of interest such as RNAs exist only in low copy numbers and the effect of intrinsic noise is significant, thus requiring a probabilistic modeling approach [53].

Temporally-varying populations for many single-cell biological processes can be modeled with continuous-time, discrete-state Markov processes [54–56]. Each state is the integer vector whose entries are the number of molecules of all species. Finding the probability distribution over these states amounts to solving the forward Kolmogorov equation, known in biochemistry as the chemical master equation (CME [57, 58]). The CME is a first-order, linear, infinite-dimensional system of ordinary differential equations that describes the time evolution of the probability distribution of the corresponding Markov process. Analytical solutions to the CME are known only for the simplest models [59]. For more elaborate systems, the total number of states grows exponentially with the number of species and becomes intractable, a situation known as the curse of dimensionality.

[2] The ideas presented in this chapter formed the foundations of a later publication titled "Bayesian estimation for stochastic gene expression using multifidelity models." in the Journal of Physical Chemistry B, with Huy Vo as the lead author. Like the method described in this chapter, the published work uses projection based model reduction, but focuses on Krylov subspace-based projections that were built from full, expensive FSP evalutions. The work presented in this chapter uses single-cell measurements to define the basis onto which the FSP dynamics are projected, surpassing any need to fully evaluate the large expensive evalutions.

For most biological networks, the CME is solved indirectly via sampling trajectories of the Markov process using the stochastic simulation algorithm [60], variants such as τ-leaping [61], or continuous approximations such as the chemical Langevin equation [62]. However, these kinetic Monte Carlo methods have slow convergence and lack strict error control when approximating entire probability distributions.

Alternatively, one can seek to compute directly the solution of the CME using a model order reduction method known as the finite state projection (FSP) [63, 64]. The principle of the FSP is to keep only states with significant probabilities and discard the rest of the state space, thus effectively truncating the CME into a finite problem. There are multiple methods for the CME that build on this principle [65–67]. However, even following truncation, the number of states required by the FSP may still be huge.

A promising approach to further reduce the FSP is to interpolate on a sparse set of nodes with interpolants generated by an appropriate family of bases. Multiple interpolation methods for the CME already exist [68–70], but here we explore the projection of the CME onto a linear space of radial basis functions (RBF) [71]. RBF interpolation of high-dimensional data is standard in the field of machine learning [72–74], and in computational fluid dynamics [75] due to their accurate representation of high-dimensional features and their efficiency and ease of implementation. Although RBF projection has been demonstrated to reduce the CME [76, 77], we introduce an approach with which define improved RBF centers and shape parameters. Our approach employs a modified version of adaptive residual subsampling [78] to determine RBF bases to capture empirical histogram of single-cell data. We then use these RBF bases to reduce the CME into a smaller, more solvable system of equations.

For real biological systems, the likelihood of data given a model is computed by comparing CME predictions to measured histograms. Model parameters and their uncertainties are then inferred by maximizing this likelihood [79] using optimization routines such as Matlab's *fminsearch* [80] or the Metropolis-Hastings algorithm [81]. Here, we propose a new implementation of the Metropolis-Hastings algorithm, in which the CME is first projected onto data-driven RBF

bases. We show that this *RBF-Metropolis-Hastings* approach significantly reduces the runtime in comparison to standard FSP-Metropolis-Hastings analyses.

A critical challenge in identifying biological models of gene regulation is the enormous parameter space that arises from a large number of continuously valued parameters. Moreover, certain parameter combinations may be well-constrained by experimental data, while other combinations are far less certain. One common approach to parameter estimation and uncertainty quantification is the Metropolis-Hastings MCMC algorithm [81]. Here, we use the Metropolis-Hastings algorithm to generate parameter distributions for the full FSP model and the RBF-reduced FSP model.

When applied to discrete stochastic models, the most expensive component of the MCMC approach is the evaluation of the likelihood of each parameter set, which requires a new CME solution corresponding to each parameter combination. Since a typical MCMC requires a large number of samples, any speedup in the CME solver would have tremendous impact on the performance of MCMC parameter identification. Our work seeks to implement this speedup by replacing the existing FSP solver with the RBF-based method, with the goal of using this method to quickly approach the correct parameter set.

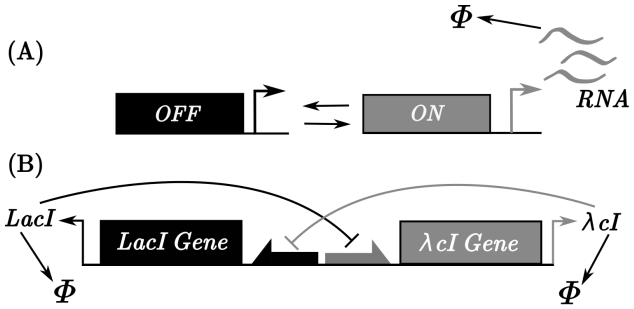

Figure 5.1: (A) The generalized two-state bursting gene expression model captures RNA transcription and degradation for a single gene that can switch between active (ON) and inactive (OFF) states. (B) The genetic toggle switch model by Gardner *et al*. consists of two mutually repressing promoters. Cooperative repression of the promoters is modeled using repressive Hill functions as shown in the text.

5.2 Interpolation Using Radial Basis Functions

The FSP truncates the state space of the CME to be finite. However, even the finite state space used in the FSP can still require an enormous number of states to obtain reasonable model accuracy. In order to alleviate the state space explosion, we employ projection-based model reduction of the FSP.

There is a large body of literature that uses projection-based model reduction of the FSP to improve computational performance. Recently proposed methods are Krylov subspaces, wavelets, polynomial spaces [65, 69, 70] and many others. A useful projection-based model reduction should allow going back and forth between the original model and the reduced model with ease, and it should not lose track of important features of the full model in the reduced model through excessive deformations of the state space. The present work is original in that we allow for the first time the use of single-cell data to guide the selection of the basis functions. Our choice of projection uses a meshless reduction method with radial basis functions. The advantages of the RBF-based model reduction are the ease of going back and forth between the full and the reduced model using the RBF projection operator, as well as its accurate representation of the important features of the state space, even in the reduced model. In this section, we give and overview of RBF interpolation and discuss a scheme for selecting centers, and describe how the RBF-FSP follows naturally from this interpolation.

5.2.1 Overview of RBF Interpolation

The curse of dimensionality causes the FSP state space to become extremely large, even for a small number of species. As a result, computing the probability distributions for such models is extremely computationally expensive. However, often times the underlying dynamics of the CME are much less complex than dictated by the full FSP when looking from the perspective of a suitable basis [82, 83]. It is reasonable to expect that reductions, such as interpolation via change of basis can retain accuracy, while reducing the computational burden of solving the full master equation.

One natural choice of such a basis family is the radial basis function. Radial basis functions are easily implemented, and scale well with the dimensionality of the function they interpolate [72,76,77,84]. This makes them very attractive for interpolating the multi-dimensional probability distributions that result from the FSP truncation to the CME. Although this basis family has been very recently used in the context of a CME solver [76,77], little has been said about how to choose the basis parameters to give reasonable accuracy and efficiency.

RBF interpolation is mesh free, requiring only the choice of RBF centers and of the tuning of the scaling parameters that indicate the width of the function supports, which are often referred to as the shape parameters [78]. Though not mathematically fully understood, the practice of choosing variable shape parameters often results in well-conditioned basis. Our approach is to choose the RBF centers based upon multi-dimensional probability distributions of discrete, single-cell data. The algorithm is adaptive, requiring successive steps of refinement and coarsening of RBF centers and corresponding tuning of the scaling parameter. We then implement adaptive RBF interpolation for single and multi-dimensional probability distributions that result from the CME.

5.2.2 RBF-Based Reduction of the FSP

The RBF-based reduction of the FSP (RBF-FSP) can be developed as follows. We enumerate the states J from the full FSP as $\{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n\}$. To reduce the FSP further, we consider a subset of the FSP state-space K, enumerated by $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_r\}$, to be the r radial basis centers. Each basis center $\mathbf{x}_k$ can be associated with a vector $\mathbf{v}_k$ of length n whose entries are given by the Gaussian function centered at $\mathbf{x}_k$ as

$$\mathbf{v}_k(i) = \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/\epsilon_k). \tag{5.1}$$

We collect these vectors into the radial basis interpolation matrix $\mathbf{\Phi} = [\mathbf{v}_1 \ldots \mathbf{v}_n]$, which has dimension $n \times r$. The interpolation matrix $\mathbf{\Phi}$ is positive definite and invertible. Therefore, $\mathbf{v}_1, \ldots \mathbf{v}_n$ form a basis for $\mathbb{R}^n$, which we will refer to as the *radial basis*. This interpolation matrix maps the FSP probability vector $\mathbf{p}_J^{FSP}$ to a reduced representation $\mathbf{q}$,

$$\mathbf{q} = \mathbf{\Phi}^{-1}\mathbf{p}. \tag{5.2}$$

Thus, we can define a the state matrix for the reduced system,

$$\mathbf{B} = \mathbf{\Phi}^{-1}\mathbf{A}\mathbf{\Phi}. \tag{5.3}$$

The dynamics of the FSP in the radial basis is given by

$$\frac{d}{dt}\mathbf{q}(t) = \mathbf{B}\mathbf{q}(t), \quad \mathbf{q}(0) = \Phi^{-\mathrm{L}}\mathbf{p}(0). \tag{5.4}$$

This reduces the FSP to a $r \times r$ dimensional dynamical system, which we call the RBF-FSP.

5.2.3 Choosing RBF Centers and Scaling Parameters

The choice of RBF centers and of scaling parameters is paramount for the interpolation, yet to our knowledge no systematic method exists to determine them. In particular, fitting a high dimensional probability distribution with different peak heights and widths will require a choice of RBF centers that is refined enough to capture all the peaks and adaptive in scale to capture the widths of the peaks. Driscoll *et al*. suggest a hierarchical multilevel algorithm with local refinement and coarsening to choose RBF for interpolation in problems with multiple localized features [78]. Following inspiration from the Driscoll algorithm, we implement an adaptive mesh algorithm for the interpolation of the CME, with two key differences. First, our approach utilizes discrete, single-cell data to choose the RBF centers, as computing the exact solution to use for interpolation may be computationally intractable in higher dimensional systems due to the curse of dimensionality. Second, we are confined to discretely-valued centers, both by discrete data and discrete-state models. At each iteration, the algorithm samples the error at new points between the current RBF centers and accepts or rejects them based upon a predefined error threshold. Choosing the right error thresholds is crucial in smoothing-out data using RBF interpolation. If the threshold is too strict, the interpolant will try to reproduce the noisy characteristics of the data; if the threshold is

too relaxed, the interpolant will smooth away relevant features of the data such as multi-modality. Figure 5.2 shows the results of applying the RBF interpolation refinement algorithm to simulated data for two different models. In Figure 5.2A, the algorithm is applied to data simulated with the from the bursting gene expression model (see Figure 5.1A). This interpolation results in a smooth representation of the data in green with five RBF centers. In combination with the five RBFs for the observable space of mRNA populations, there are two gene states ('on' and 'off'). When applied to the CME, this RBF with therefore yield a reduced dimension for $\mathbf{B}$ that is 10×10. We also implemented the adapted RBF refinement algorithm to interpolate data simulated with the SSA for a genetic toggle switch model comprised of two mutual repressors (Figure 5.1B). Figure 5.2B shows the original data, and Figure 5.2C shows the interpolation of that data on the RBF basis set. In this case, through systematic refinement and coarsening, 137 RBF centers are sufficient to capture the two-dimensional joint probability distribution of the toggle model.

The main advantage of the proposed interpolation technique is that it easily generalizable to more than two dimension. In fact, the number of RBF centers required for interpolation scales approximately linearly with the number of species to which it is applied. In turn, this reduced dimension can then alleviate the state space explosion often encountered with complex biochemical reaction networks. Once the RBF has been identified to produce an effective interpolation of the single-cell data, the same RBF can be applied to reduce the CME model, as discussed in the examples below.

5.3 Numerical Examples

For each numerical example below, we simulated data using the SSA [60], and we then apply a modified version of Driscoll *et al*. algorithm to the simulated data at a single time point to choose a basis for the reduced model.

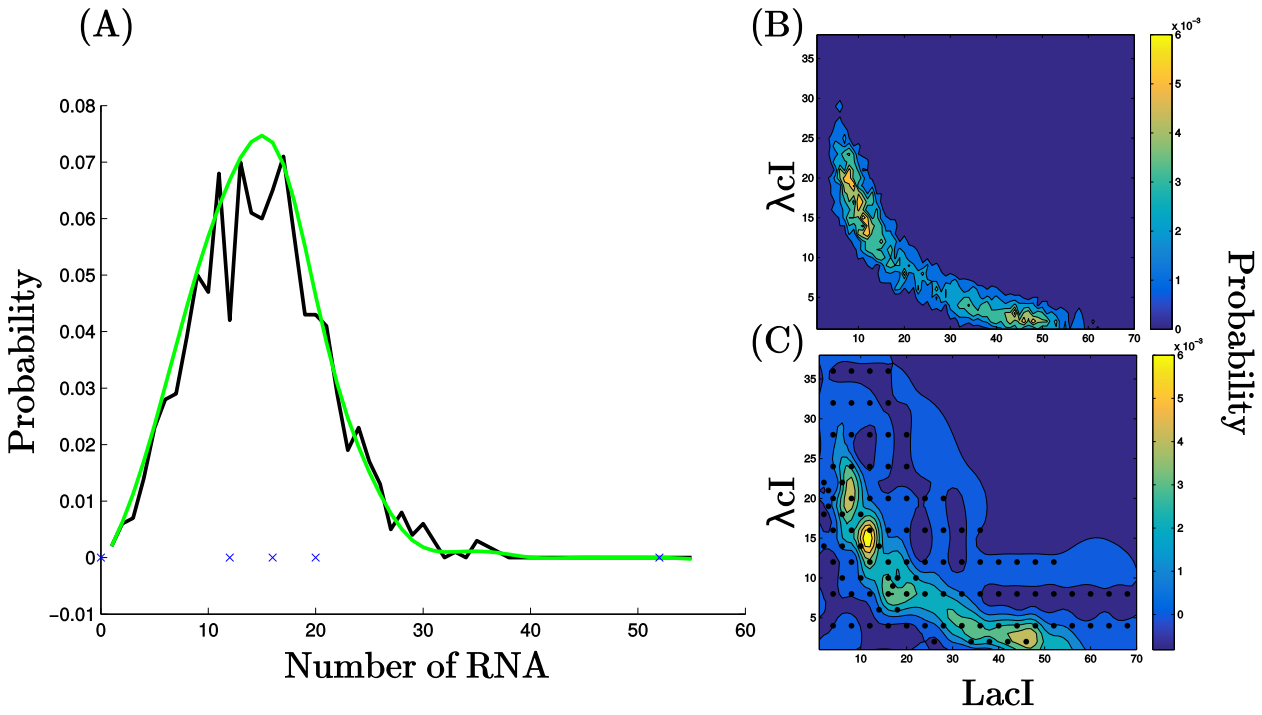

Figure 5.2: RBF-based interpolation of simulated single-cell data. (A) For the bursting gene expression model, we compare the data generated with the stochastic simulation algorithm (SSA) (black) with the approximation using radial basis function interpolation (green) at $t = 10$s. The five RBF centers are positioned at the blue crosses. (B) Data for the genetic toggle switch is generated for 1000 trajectories of the SSA at $t = 4$hrs. We plot the joint probability mass of the two repressors LacI and λcI. (C) The RBF interpolation for the genetic toggle switch. We use 137 radial basis functions centered at the black dots.

5.3.1 Bursting Gene Expression

The bursting gene expression model arises from changes in the state of a gene's promoter, such as the binding/unbinding of transcription factors. When the gene is 'on', RNA is actively translated at rate k_r. For an RNA molecule x, this simple two-state view of a gene can result in a variety of RNA dynamics depending on the system parameters [85]. This gene regulatory network can be written by the following set of biochemical reactions describing the state of the gene $\mathcal{G}$ and RNA abundance x.

$$\mathcal{R}1: \ \mathcal{G}_{\text{off}} \xrightarrow{w_1} \mathcal{G}_{\text{on}} \quad \mathcal{R}2: \ \mathcal{G}_{\text{on}} \xrightarrow{w_2} \mathcal{G}_{\text{off}}$$

$$\mathcal{R}3: \ \mathcal{G}_{\text{on}} \xrightarrow{w_3} \mathcal{G}_{\text{on}} + x \quad \mathcal{R}4: \ x \xrightarrow{w_4} \varnothing$$

where the propensities $\mathbf{w} = \{w_1, \ w_2, \ w_3, \ w_4\}$ are $w_1 = k_{\text{on}}$, $w_2 = k_{\text{off}}$, $w_3 = k_r$ and $w_4 = \gamma x$.

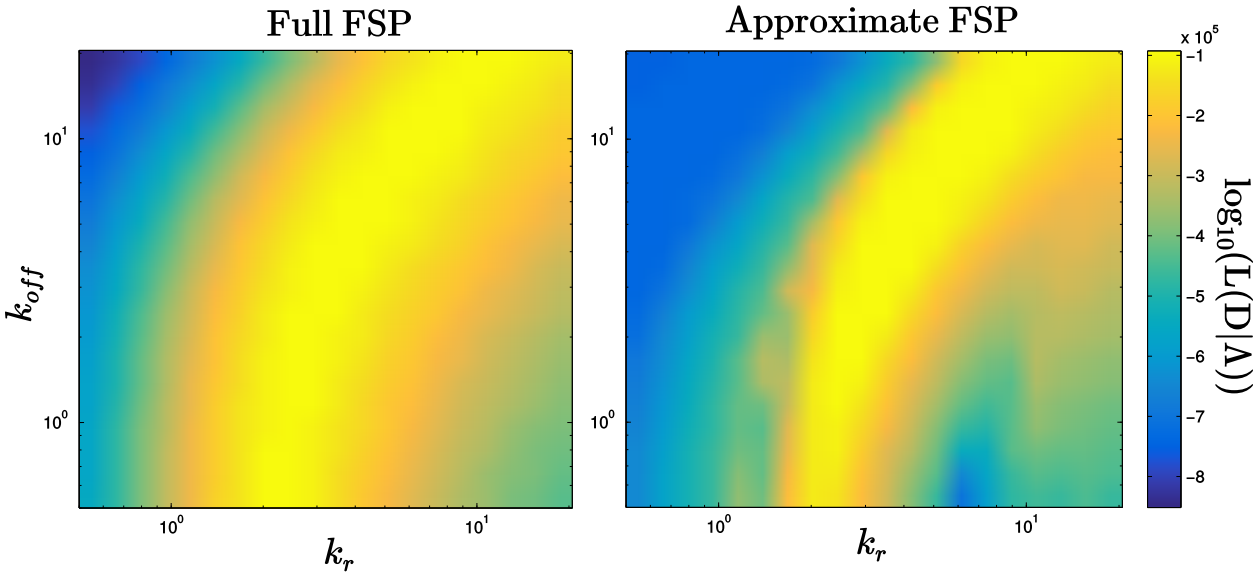

Figure 5.3: Parameter sweep with the full FSP (left) and the RBF-FSP (right) for the bursting gene expression model for 30 time-points from $t = 0$s to $t = 10$s. Colors correspond to the likelihoods for parameters k_r and k_{off}. Each parameter is varied one order of magnitude above and below the true parameters $\mathbf{\Lambda}_{\text{true}}$ (center of plot).

Data was simulated for 30 linearly-spaced time points between $t = 0$ and $t = 10$ with the SSA. Figure 5.2A shows the RBF-based representation (green line) of the simulated data (black line) with the five centers selected using the adaptive residual subsampling algorithm. We then tested 2500 different combinations of the transcription rate k_r and gene deactivation rate k_{off} spanning one order of magnitude above and below the 'true' parameter values from which the data was generated. Figures 5.3A and 5.3B show the resulting log-likelihood values for the data given the full and reduced models, respectively. In Figure 5.5A, the parameters that maximize the log-likelihood for the full and reduced models are shown in blue dots and green lines, respectively. The best parameters identified and their associated likelihoods and computational times, t_{ID}, are given in Table 5.1. As only five RBF centers were required to represent the data at all times (compared to the full FSP state space of 80), the time required to identify the parameters t_{ID} with a parameter sweep was much lower than with the full model.

5.3.2 Mutually-Repressing Toggle Switch

For a second test model, we examine the well-known genetic toggle switch circuit of Gardner *et al*. [86]. There are two mutually repressive promoter species *lacI* and λ*cI* (Figure 5.1B) with

	$k_r(s^{-1})$	$k_{\text{off}}(s^{-1})$	$L(\mathbf{D}\|\mathbf{\Lambda})$	$t_{\text{ID}}(s)$
Full	9.54	1.30	-9.27e4	172
Reduced	9.54	1.43	-9.30e4	16.0
“True"	10.0	1.50	-9.26e4	-

Table 5.1: Parameters identified and their associated likelihoods for a parameter sweep over 2500 parameter combinations for k_r and k_{off} with the full FSP and the RBF-FSP.

stochastic interactions. The reactions of this biochemical network are then given by:

$$\mathcal{R}1: \ \varnothing \xrightarrow{w_1} \lambda\text{cI} \quad \mathcal{R}2: \ \lambda\text{cI} \xrightarrow{w_2} \varnothing$$

$$\mathcal{R}3: \ \varnothing \xrightarrow{w_3} \text{LacI} \quad \mathcal{R}4: \ \text{LacI} \xrightarrow{w_4} \varnothing,$$

where the propensities $\mathbf{w} = \{w_1, w_2, w_4, w_4\}$ are given above the arrows, and

$$w_1 = b_x + \frac{k_x}{1 + \alpha_{yx}\text{LacI}^{n_{yx}}}$$

$$w_2 = \gamma_x \cdot \lambda\text{cI}$$

$$w_3 = b_y + \frac{k_y}{1 + \alpha_{xy}\lambda\text{cI}^{n_{xy}}}$$

$$w_4 = \gamma_y \cdot \text{LacI}.$$

With these reactions and the parameters listed in Table 5.2, we simulated data using the SSA (shown in Figure 5.2B).

We then ran the Metropolis-Hastings algorithm combined with the RBF-FSP solution at each step to find the parameter values that maximize the likelihood function. Our results for identifying the probability distributions for the best parameters using Metropolis-Hastings are in Figure 5.5B and 5.5C for the full FSP and the reduced, RBF-FSP solutions, respectively. The values of the identified parameters of the genetic toggle switch are in Table 5.2.

We then compare against the parameters identified by the usual procedure, where the FSP stands in place of the RBF approximation. The RBF procedure is almost twice as fast as the existing technique, while the parameters it identifies are near to the “true" parameter values (Figure 5.4).

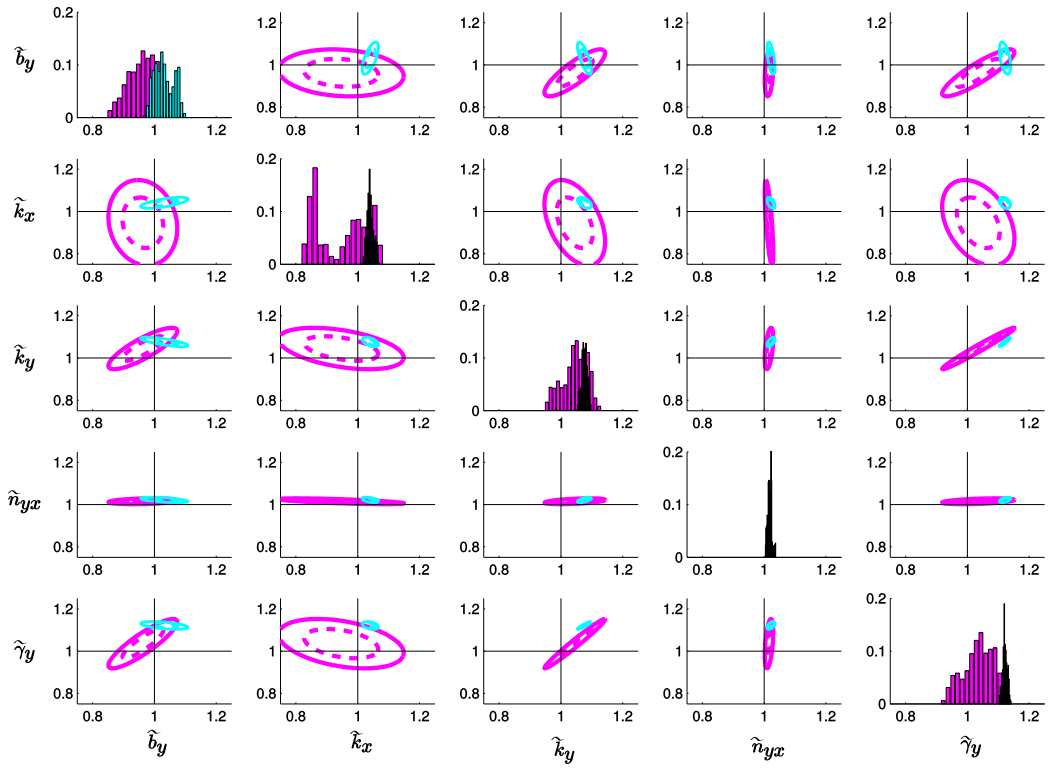

Figure 5.4: Parameter distributions from the Metropolis Hastings search show that the RBF-FSP approaches $\mathbf{\Lambda}_{\text{true}}$. The 95% (solid lines) and 65% (dashed lines) for the parameter distributions sampled with MCMC for the full FSP (magenta) and the RBF-FSP (cyan). Single parameter histograms are shown on the diagonal. Parameter distributions are scaled relative to $\mathbf{\Lambda}_{\text{true}}$ such that each exact parameter has a value of unity.

	b_y	k_x	k_y	η_{yx}	γ_y	$L(\mathbf{D}\|\mathbf{\Lambda})$	t_{ID}(min)
True Model	2.20e-3	1.60-2	1.70e-2	2.1	3.8e-4	-3.022e4	-
Reduced Model	2.22e-3	1.65e-2	1.85e-2	2.17	4.31e-4	-3.114e4	287
Full Model	1.89e-3	1.46e-2	1.60e-2	2.10	3.46e-4	-3.0871e4	444

Table 5.2: Parameters identified using the Metropolis-Hastings algorithm for the toggle model. Parameters were selected as the best choices in the latter half of the MCMC chain.

Figure 5.4 demonstrates the ability of the RBF-FSP to identify parameters in the toggle model. The 95% and 65% confidence intervals were computed for the second half of a 100000 iteration-long MCMC chain which has been thinned to 10000 samples. The parameters are normalized by the "true" parameter values in Figure 5.4, i.e. $\tilde{\mathbf{\Lambda}} = \frac{\mathbf{\Lambda}}{\mathbf{\Lambda}_{\text{true}}}$

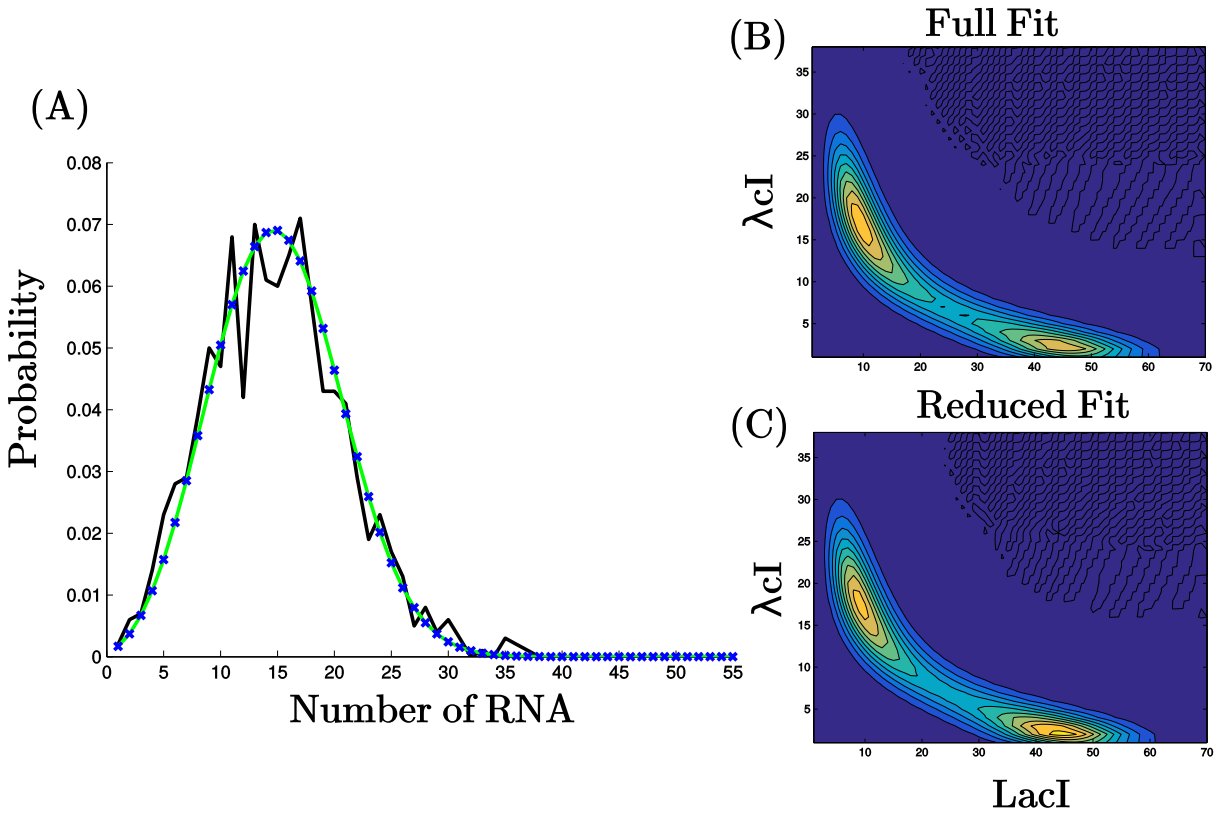

Figure 5.5: Best fits for the bursting gene expression and toggle models. (A) Simulated data is given by the black histogram. Modeled probability distributions correspond maximized likelihood of observing the data $L(D|\mathbf{\Lambda})$. Parameters were identified from parameter sweeps using the RBF-FSP (green) or the full FSP (blue). (B,C) Same as A, except for the toggle model, and using parameters identified during 100,000 runs of the Metropolis-Hastings algorithm.

5.3.3 Toggle Model with Time-Varying Inputs

Next, we consider a special case of the toggle model in which the basal rate of production of $LacI$ varies in time:

$$b_x = b_{x0}(1 - \sin(2\pi\omega t), \tag{5.5}$$

where ω is the frequency of the time-varying input signal. This example demonstrates the use of the RBF-FSP for time-varying infinitesimal generators, $\mathbf{A}(t)$ and $\mathbf{B}(t)$, in which matrix exponentiation and Krylov methods are inapplicable, and for which the CME must be numerically integrated, such as with MATLAB's `ode23s`. For this example, the RBF centers were selected using a single data snapshot in time, $t = 4$hr. We then tested 1000 values of k_x in a parameter sweep using the full FSP and RBF-FSP separately, and the results of this parameter sweep are shown in Figure 5.6. We

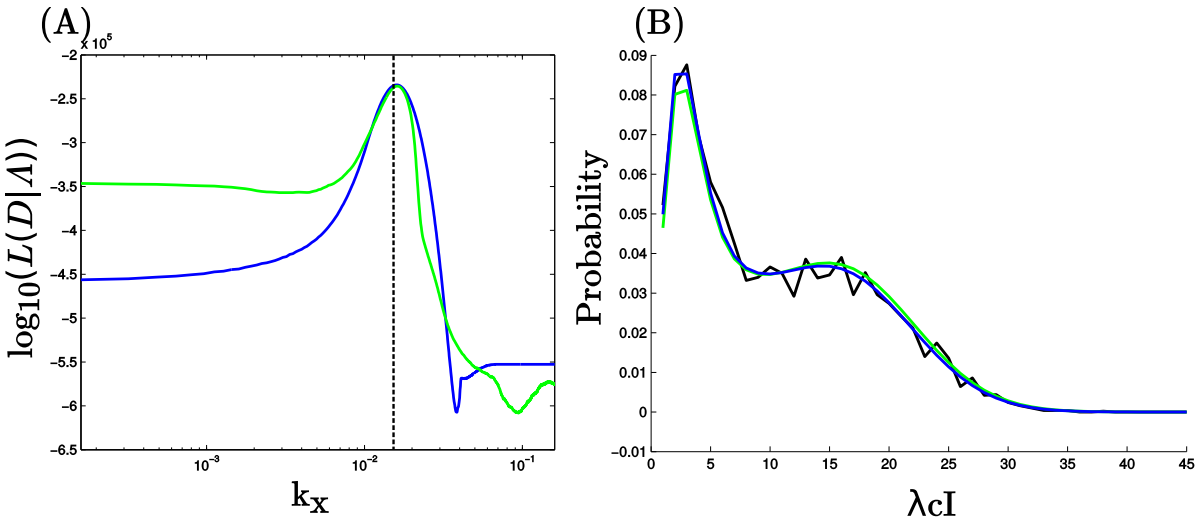

Figure 5.6: Parameter identification with a time-varying toggle model and 5000 modified SSA simulations. (A) The log-likelihood as a function of k_x using the full FSP (blue) and the RBF-FSP (green). The two approaches agree approximately near the maximum, which coincides with the 'true' value of k_x. (B) Marginal distributions for simulated data (black) as well as for the model identified with the full FSP (blue) and RBF-FSP (green).

found that when the parameters are such that they closely match the simulated data, the full FSP (blue line) and reduced RBF-FSP (green line) solutions are in close agreement for their computed likelihoods (see Figure 5.6A). However, when the parameters are far removed from their correct values the RBF-FSP computation is much less accurate. The marginal probability distributions obtained with the best parameters identified with both methods are presented in Figure 5.6B. For this example, we observed computational speedups of more than twenty-fold using the RBF-FSP: the parameter sweep with the full FSP took 8.33 hr while the search with the RBF-FSP took just 0.37 hr.

5.4 Discussion

When examined at the single-cell level, biochemical processes are subject to single-molecule events and discrete stochastic phenomena. These stochastic dynamics can be measured using modern single-cell and single-molecule experiments and they can be described by the chemical master equation (CME). However, inferring gene regulation parameters from single-cell data requires thousands of CME solutions and enormous computational effort. In this article, we pro-

posed a means to use single-cell data to define a small set radial basis functions onto which the CME can be projected prior to numerical analysis or parameter inference. We applied this RBF-reduced CME approach to three example models, including bursting gene expression, a genetic toggle switch, and a toggle switch with time varying rates. For each, we showed that we could use simulated data to define RBF basis sets that capture the most important dynamics of the CME, and that using these RBFs as part of the parameter inference scheme could lead to substantial reductions in computational effort. We expect that this approach will be highly valuable to quickly evaluate stochastic models to compare to single-cell data. Moreover, because the number of RBFs needed to interpolate higher dimensional data scales linearly with the number of dimensions, it is envisioned that this data-drive reduction of the CME could provide a key step toward overcoming the curse of dimensionality in the analysis and identification of stochastic gene regulation models.

Chapter 6

The finite state projection based Fisher information matrix approach to estimate and maximize the information in single-cell experiments [3]

6.1 Introduction

Recent labeling and imaging technologies have greatly increased capabilities to measure biological phenomena at the single-cell and single-molecule levels. When conducted under different conditions, single-cell experiments can probe processes for different spatial or temporal resolutions, for different population sizes, under different stimuli, at different times during a response, and for myriad other controllable or observable factors [11, 13, 16, 87–90]. As these experiments have become more capable to precisely perturb or measure different biological species, they have also become more expensive, which imposes a limit on the number and type of experiments that can be conducted in any given study. Clearly, not all experiment designs provide the same information, and different experiments may be "optimal" to answer different questions about the system. However, the inherent diversity of modern experiments makes it difficult to intuit which experiments will be most informative and in which circumstances. Computational tools for model-driven experiment design could help to select more informative experiments, provided that existing tools can be adapted to overcome the unique challenges presented by single-cell data.

One model-driven approach to optimal experiment design is to use the *Fisher information matrix* (FIM), which describes the precision to which a model's parameters can be estimated for any particular experiment [14, 38, 91–94]. To improve estimates of model parameters, the FIM can be used iteratively in a Bayesian framework by specifying maximally informative experimental

[3] This work was published in PLoS Computational Biology in 2019.

conditions, collecting data under these conditions, using new data to constrain parameters, and using the newly constrained parameters to design the next round of experiments [14,39,41,93,94]. The formalism of the FIM for experiment design has been used to great effect in engineering disciplines, such as radar, astrophysics, and optics [95–97]. In principle, similar analyses could introduce a natural feedback in the co-design of single-cell experiments and discrete stochastic models, but for this to work, accurate analyses are needed to extract more meaning from the data and to provide better predictions about how biological systems will behave under new conditions.

Experimentally observed cell-to-cell variability has been well demonstrated to provide substantial quantitative insight to constrain and identify the mechanisms and parameters of gene regulation models [11,13,16,22,42,87–89,98]. Therefore, the FIM analysis for the optimal design of single-cell experiments should explicitly consider such single-cell variability. Standard FIM analyses assume continuous-valued observables with Gaussian-distributed *measurement* noise. However, in contrast to most classical engineering applications, the distributions of integer-valued RNA or protein levels across an isogenic cell population can be highly complex and subject to intrinsic and extrinsic variations, with nonlinear interactions that lead to multiple peaks and long tails [9–11,43]. Because the FIM is not computable for general discrete stochastic processes with non-Gaussian distributions, computational biologists have applied various approximations to estimate the FIM. A few recent biological studies use the Linear Noise Approximation [29] to treat single-cell distributions as Gaussian, which allows for the use of standard Fisher information analyses [38]. This approach, which we refer to as the LNA-FIM, should be valid for large numbers of molecules, but it is unlikely to be accurate for systems with high intrinsic noise corresponding to low gene, RNA, or protein counts. A different approach to estimate the FIM uses the central limit theorem (CLT) to approximate the sample mean and covariance to be jointly Gaussian and uses higher-order moments of the chemical master equation to estimate the likelihood of these moments [14]. This approach, which we refer to as the sample moments approach (SM-FIM), should be valid for large numbers of cells as can be collected in high-throughput experimental approaches, such as flow cytometry. However, when distributions have long asymmetric tails and sample sizes are

limited, higher moments become very difficult to estimate and can lead to surprising model estimation errors [20]. Beyond these few Gaussian assumptions, there has been little work devoted to improve the design of time-varying single-cell experiments for systems with arbitrary probability distributions.

In this study, we introduce a formulation of the Fisher information for use with discrete stochastic models and data sets containing intrinsic variability that is measurable with single-biomolecule resolution. Our approach utilizes the finite state projection (FSP) approach [27] to solve the chemical master equation (CME) [28,29], and compute the likelihood of single-cell data given a discrete stochastic model [11,22,43]. The FSP solves for the probability distribution over discrete numbers of biomolecules to any arbitrary error tolerance. By utilizing the full probability distributions, as opposed to finite order or approximate moments of these distributions, our approach makes no assumptions and works well for distributions with multiple peaks or long tails.

In the next section, we introduce the FSP and derive the sensitivities of the FSP solution to small perturbations in parameters. Next, we derive the likelihood function and its local sensitivity for discrete stochastic models and discrete data. These allow us to formulate and compute the FSP-FIM. Next, we use a combination of analytical results and numerical simulations to verify the FSP-FIM for two common models of gene expression. Finally, we demonstrate how the FSP-FIM can be applied to design nontrivial experiments for a simulated system with nonlinear reaction rates.

6.2 Derivation of the Fisher Information for FSP Models

The FIM, which describes the amount of information that can be expected by performing a particular experiment with N_c cells, is defined as

$$\mathcal{I}(\boldsymbol{\theta}) = N_c \mathbb{E}\left\{ \left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta})\right)^T \left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta})\right) \right\}, \tag{6.1}$$

where the expectation is taken over $p(\mathbf{X};\boldsymbol{\theta})$, corresponding to the density from which future (or hypothetical) data could be sampled. For FSP models, this density is the discrete distribution found by solving Eq. 2.3. Equation 6.1 is positive semi-definite and is additive for collections of independent observations [91]. The inverse of the FIM is known as the Cramèr-Rao bound (CRB), which provides a useful lower bound on the variance for any unbiased estimator of model parameters [92]. The notion of information stems from the fact that new experiments should increase the FIM, corresponding to additional knowledge about $\boldsymbol{\theta}$ and a tighter CRB. More specifically, the well-known asymptotic normality of the maximum likelihood estimator (MLE) states that as the number of measurements N_c increases, the MLE estimates will converge in distribution to a multivariate normal probability density with a variance given by the CRB,

$$\sqrt{N_c}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{dist} \mathcal{N}(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}), \tag{6.2}$$

where $\hat{\boldsymbol{\theta}}$ is the $\boldsymbol{\theta}$ that maximizes Eq. 3.1 and $\boldsymbol{\theta}^*$ are the "true" model parameters that produced the observed data [91, 92]. Designing experiments to maximize a given metric of the FIM can be expected to provide a more accurate estimate of $\boldsymbol{\theta}$, where different definitions of 'accuracy' (i.e., different vector norms for parameter errors) can be implemented through the choice of different FIM metrics.

To derive the FIM requires one must take the partial derivative of the log-likelihood (Eq. 3.1) with respect to the parameters $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{p_0}\frac{\partial p_0}{\partial \theta_1} & \frac{1}{p_0}\frac{\partial p_0}{\partial \theta_2} & \cdots & \frac{1}{p_0}\frac{\partial p_0}{\partial \theta_{N_p}} \\ \frac{1}{p_1}\frac{\partial p_1}{\partial \theta_1} & \frac{1}{p_1}\frac{\partial p_1}{\partial \theta_2} & \cdots & \frac{1}{p_1}\frac{\partial p_1}{\partial \theta_{N_p}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{p_N}\frac{\partial p_N}{\partial \theta_1} & \frac{1}{p_N}\frac{\partial p_N}{\partial \theta_2} & \cdots & \frac{1}{p_N}\frac{\partial p_N}{\partial \theta_{N_p}} \end{pmatrix}. \tag{6.3}$$

The expression $\nabla_{\boldsymbol{\theta}} p(\mathbf{X};\boldsymbol{\theta})$ is the *sensitivity matrix*, $\mathbf{S}$, which has dimensions $N \times N_\theta$, where N is the dimension of the CME or its FSP projection. We derive an equation similar to that presented

in [99] to define the time evolution of the sensitivity for each state's probability density, $p(\mathbf{x}_l;\boldsymbol{\theta})$, to each parameter θ_j. However, unlike previous analyses that rely on stochastic simulations and finite difference approaches, the FSP enables direct approximation of the sensitivities. Using the sensitivity matrix, the entries of the FIM can be computed as:

$$\mathcal{I}(\boldsymbol{\theta})_{ij} = N_c \mathbb{E}\left\{ \left(\frac{1}{p(\mathbf{x}_l;\boldsymbol{\theta})} \right)^2 \mathbf{S}_{li}\mathbf{S}_{lj} \right\}. \tag{6.4}$$

Taking the expectation over all l on $(1, N)$ yields the elements of the FIM:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta})_{ij} &= N_c \sum_{l=1}^{N} \left(\frac{1}{p(\mathbf{x}_l;\boldsymbol{\theta})} \right)^2 \mathbf{S}_{li}\mathbf{S}_{lj} p(\mathbf{x}_l;\boldsymbol{\theta}), \\ &= N_c \sum_{l=1}^{N} \frac{1}{p(\mathbf{x}_l;\boldsymbol{\theta})} \mathbf{S}_{li}\mathbf{S}_{lj}, \end{aligned} \tag{6.5}$$

which quantifies Fisher information for the model evaluated at a single time point. For smFISH data, each time point is independent. If $N_c(t_k)$ cells are measured at each k^{th} time point, the FIM is summed, and the total information is computed as:

$$\mathcal{I}(\boldsymbol{\theta})_{ij} = \sum_{k=1}^{N_t} N_c(t_k) \sum_{l=1}^{N} \frac{1}{p(\mathbf{x}_l;t_k,\boldsymbol{\theta})} \mathbf{S}_{li}(t_k)\mathbf{S}_{lj}(t_k). \tag{6.6}$$

The Fisher information can be found using Eq. 6.6 for any model for which the FSP (Eq. 2.3) can be solved. This formulation explicitly quantifies how the number of cells and number of time points impact the information, and is easily extended to include other experiment design aspects such as the interval of successive measurements or changes in applied inputs, as we will demonstrate in the following sections. Because one is often interested in the relative sensitivity of parameters rather than the absolute sensitivity, a logarithmic parameterization of the FIM can easily be obtained from Eq. 6.6.

$$\mathcal{I}(\log\boldsymbol{\theta}) = \mathbf{E}\Big[\Big(\nabla_{\log\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta}) \Big)^T \Big(\nabla_{\log\boldsymbol{\theta}} \log p(\mathbf{X};\boldsymbol{\theta}) \Big) \Big] \tag{6.7}$$

The logarithmic parameterization carries through to the computation of the sensitivity matrix,

$$\nabla_{\boldsymbol{\theta}} \log p(x;\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{p_0}\frac{\partial p_0}{\partial \log \theta_1} & \frac{1}{p_0}\frac{\partial p_0}{\partial \log \theta_2} & \cdots & \frac{1}{p_0}\frac{\partial p_0}{\partial \log \theta_{N_p}} \\ \frac{1}{p_1}\frac{\partial p_1}{\partial \log \theta_1} & \frac{1}{p_1}\frac{\partial p_1}{\partial \log \theta_2} & \cdots & \frac{1}{p_1}\frac{\partial p_1}{\partial \log \theta_{N_p}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{p_N}\frac{\partial p_N}{\partial \log \theta_1} & \frac{1}{p_N}\frac{\partial p_N}{\partial \log \theta_2} & \cdots & \frac{1}{p_N}\frac{\partial p_N}{\partial \log \theta_{N_p}} \end{pmatrix}. \tag{6.8}$$

Using the relationship $\frac{\partial f(x)}{\partial \log x} = x\frac{\partial f(x)}{\partial x}$, we can rewrite Eq. 6.8 as

$$\nabla_{\log \boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) = \mathbf{Q}\mathbf{S}\boldsymbol{\Theta}, \tag{6.9}$$

where

$$\boldsymbol{\Theta} \equiv \begin{pmatrix} \theta_1 & 0 & \dots & 0 \\ 0 & \theta_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \theta_{N_p} \end{pmatrix}$$

and $\mathbf{Q} = diag\{\frac{1}{\mathbf{p}}\}$. Therefore, the logarithmic parameterization is easily found by multiplying the i^{th} column in $\mathbf{S}$ by the corresponding parameter θ_i. The log-FSP-FIM can then be computed:

$$\mathcal{I}(\log \boldsymbol{\theta})_{i,j} = N_c \sum_{k=1}^{N} \frac{\theta_i \theta_j}{p(\mathbf{x}_k;\boldsymbol{\theta})} \mathbf{s}_i^k \mathbf{s}_j^k = \theta_i \theta_j \mathcal{I}(\boldsymbol{\theta})_{ij}. \tag{6.10}$$

In the following sections, we will verify the FIM using several common models of gene expression, and demonstrate experiment designs using these approaches.

6.2.1 Derivation of information for Gaussian fluctuations

The Gaussian distribution with mean and variance λ is defined

$$f(x,\lambda) = \frac{1}{\sqrt{2\pi\lambda}} e^{\frac{(x-\lambda)^2}{2\lambda}}. \tag{6.11}$$

Computing the FIM for this Guassian requires finding the derivative of the log-density

$$\log f(x,\lambda) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\lambda - \frac{1}{2}\left(\frac{x^2 - 2x\lambda + \lambda^2}{\lambda}\right) \tag{6.12}$$

with respect to λ,

$$\begin{aligned}\frac{\partial \log f(x,\lambda)}{\partial\lambda} &= -\frac{1}{2\lambda} - \frac{1}{2}(1 - \frac{x^2}{\lambda^2}) \\ &= -\frac{1}{2}(-\frac{x^2}{\lambda^2} + \frac{1}{\lambda} + 1)\end{aligned}$$

and squaring it:

$$\begin{aligned}\left(\frac{\partial \log f(x,\lambda)}{\partial\lambda}\right)^2 &= \frac{1}{4}\left(-\frac{x^2}{\lambda^2} + \frac{1}{\lambda} + 1\right)\left(-\frac{x^2}{\lambda^2} + \frac{1}{\lambda} + 1\right) \\ &= \frac{1}{4}\left(\frac{x^4}{\lambda^4} - \frac{2x^2}{\lambda^3} - \frac{2x^2}{\lambda^2} + \frac{1}{\lambda^2} + \frac{2}{\lambda} + 1\right).\end{aligned} \tag{6.13}$$

To take the expected value, we need the second and fourth moments of the normal distribution, which are $\lambda^2 + \lambda$ for the second uncentered moment and $\lambda^4 + 6\lambda^3 + 3\lambda^2$ for the fourth uncentered moment. Thus, we have:

$$\begin{aligned}\mathbb{E}\Big[\Big(\frac{\partial \log f(x,\lambda)}{\partial\lambda}\Big)^2\Big] &= \frac{1}{4}\Big(\frac{\lambda^4 + 6\lambda^3 + 3\lambda^2}{\lambda^4} - \frac{2(\lambda^2+\lambda)}{\lambda^3} - \frac{2(\lambda^2+\lambda)}{\lambda^2} + \frac{1}{\lambda^2} + \frac{2}{\lambda} + 1\Big) \\ &= \frac{1}{4}\Big(\frac{4}{\lambda} + \frac{2}{\lambda^2}\Big) = \frac{1}{\lambda} + \frac{1}{2\lambda^2}.\end{aligned}$$

6.2.2 Derivation of information for a Poisson distribution

The Poisson distribution is defined:

$$f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{6.14}$$

Again, by taking the log

$$\log f(x,\lambda) = x\log\lambda - \lambda - \log x! \tag{6.15}$$

Now, take the derivative with respect to λ

$$\frac{\partial \log f(x,\lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1, \tag{6.16}$$

and squaring this term yields:

$$\Big(\frac{\partial \log f(x,\lambda)}{\partial \lambda}\Big)^2 = \frac{x^2}{\lambda^2} - \frac{2x}{\lambda} + 1. \tag{6.17}$$

As the FIM is the expected value of this quantity, and the mean and variance of the Poisson distribution are given by λ,

$$\begin{aligned}\mathbb{E}\Big[\Big(\frac{\partial \log f(x,\lambda)}{\partial \lambda}\Big)^2\Big] &= \mathbb{E}\Big[\frac{x^2}{\lambda^2}\Big] - \mathbb{E}\Big[\frac{2x}{\lambda}\Big] + 1 \\ &= \frac{\lambda^2+\lambda}{\lambda^2} - 2 + 1 \\ &= \frac{1}{\lambda}.\end{aligned} \tag{6.18}$$

6.3 Derivation of sensitivities for FSP models

The change of probability $p(\mathbf{x}_l)$ with respect to small changes in parameter θ_j describes the sensitivity of the l^{th} state in the Markov process to the j^{th} parameter [99, 100]. These local sensitivities can be calculated by transforming the linear ODEs describing the time evolution of the probabilities of the state space $\frac{d}{dt}\mathbf{p}(t) = f\big(\mathbf{p}(t),\boldsymbol{\theta},t\big)$ into a set of ODEs describing the time evolution of the sensitivities. Given an initial condition, the solution to the CME is:

$$\mathbf{p}(t;\boldsymbol{\theta}) = \mathbf{p}(t_0) + \int_{t_0}^{t} f(\mathbf{p}(s;\boldsymbol{\theta}),\boldsymbol{\theta},s)ds \tag{6.19}$$

Taking partial derivatives with respect to $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}}\mathbf{p}(t;\boldsymbol{\theta}) = \int_{t_0}^{t} \Big[\nabla_{\boldsymbol{\theta}} f(\mathbf{p}(s;\boldsymbol{\theta}),\boldsymbol{\theta},s) + \nabla_{\mathbf{p}} f(\mathbf{p}(s;\boldsymbol{\theta}),\boldsymbol{\theta},s)\nabla_{\boldsymbol{\theta}}\mathbf{p}(s;\boldsymbol{\theta})\Big] ds. \tag{6.20}$$

We can now describe the sensitivities $\mathbf{S} \equiv \nabla_{\boldsymbol{\theta}}\mathbf{p}$ as they evolve with time, by taking the time derivative of the equation above. For the FSP, the right-hand side $f(\mathbf{p}(t;\boldsymbol{\theta}),\boldsymbol{\theta},t) = \mathbf{A}(\boldsymbol{\theta},t)\mathbf{p}(t)$, and

$$\nabla_{\boldsymbol{\theta}} f(t,\mathbf{p}(t;\boldsymbol{\theta}),\boldsymbol{\theta}) = \left(\nabla_{\boldsymbol{\theta}}\mathbf{A}(\boldsymbol{\theta})\right)\mathbf{p}(t) \tag{6.21}$$

$$\nabla_{\mathbf{p}} f(t,\mathbf{p}(t;\boldsymbol{\theta}),\boldsymbol{\theta}) = \mathbf{A}(\boldsymbol{\theta}) \tag{6.22}$$

In many cases, including all models formulated using mass-action kinetics, the generator $\mathbf{A}$ can be written as a linear combination of the model parameters, i.e. $\mathbf{A} = \sum \theta_i \mathbf{B}_i$, and the derivative with respect to the i^{th} parameter can be found,

$$\frac{\partial}{\partial\theta_i}\mathbf{A} = \frac{\partial}{\partial\theta_i}(\theta_i\mathbf{B}_i) = \mathbf{B}_i. \tag{6.23}$$

Using this notation, Eq. 6.20 is reduced to the set of linear ODEs for each parameter θ_i,

$$\frac{d}{dt}\begin{pmatrix}\mathbf{p}(t)\\ \mathbf{S}_i(t)\end{pmatrix} = \begin{pmatrix}\mathbf{A} & 0\\ \mathbf{B}_i & \mathbf{A}\end{pmatrix}\begin{pmatrix}\mathbf{p}(t)\\ \mathbf{S}_i(t)\end{pmatrix}. \tag{6.24}$$

In practice, Eq. 6.24 can be computed in parallel for each parameter, and the computation of sensitivities for K parameters is equivalent to solving K sparse systems of ODEs, each twice the size of the FSP computation.

6.3.1 Moment-based FIM Approximations

Current state-of-the-art approaches for single-cell, single-molecule experiment design rely on computing moments of the CME. Approaches that use ODE reaction kinetics (in a deterministic

model setting) [101–103], linear noise approximations [38, 41], or higher order moments [14] all make use of the well-known Gaussian form of the FIM

$$FIM_{i,j} = \frac{\partial \boldsymbol{\mu}}{\partial \theta_i}^T \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j} + \frac{1}{2} trace\Big(\boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_j}\Big). \tag{6.25}$$

In the higher order approach, developed by Ruess et al [14] takes the sample mean and sample variance to be jointly Gaussian, and thus requires the computation of up to the 4th moments in Eq. 2.11.

6.4 Verifications and applicationss of the FSP-FIM

6.4.1 The FSP-FIM captures the exact information for constitutive gene expression

To demonstrate and validate the FSP-FIM method, we begin with a simple birth and death model for constitutive gene expression as shown in Figure 6.1. This model, which has been fit to capture the variability for many housekeeping genes [87, 98], consists of two reactions, corresponding to the constant transcription and first order decay of RNA,

$$\begin{aligned} \mathcal{R}1: &\quad gene \xrightarrow{k_r} gene + RNA \\ \mathcal{R}2: &\quad RNA \xrightarrow{\gamma} \emptyset. \end{aligned}$$

The production and degradation parameters are defined as $\boldsymbol{\theta} = [k_r, \gamma]$.

Given an initial condition of zero RNA for this process, the population of RNA at any later time is a random integer sampled from a Poisson distribution,

$$p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{6.26}$$

where λ is the time varying average population size,

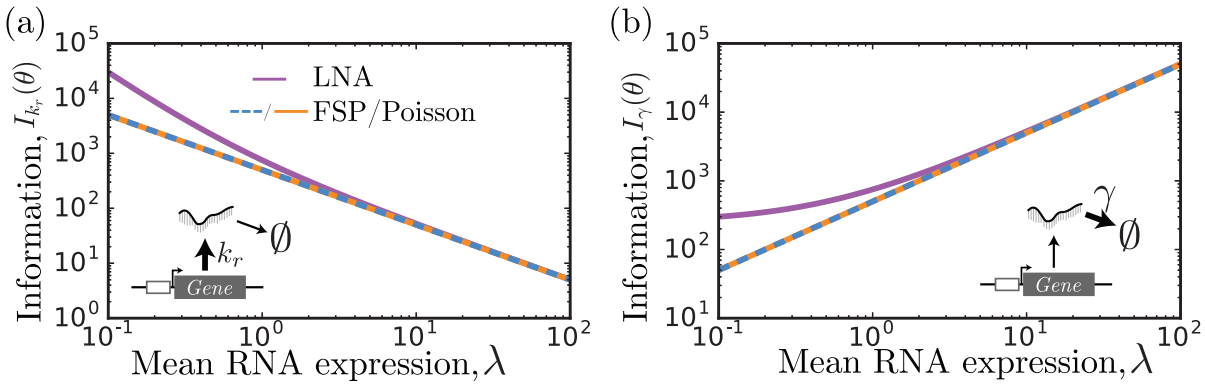

Figure 6.1: *Fisher information for a model of birth and death.* The Fisher information for the two model parameters k_r (a) and γ (b) for various values of the mean expression level, λ. The analytical form of the FIM for a Gaussian approximation and that computed using Eq. 6.25 (purple line) match to one another. The value computed using the FSP-FIM (blue) matches to the exact form of the analytical Poisson distribution (orange dashed). As λ becomes large, all four approaches are consistent.

$$\lambda(t, k_r, \gamma) = \frac{k_r}{\gamma}[1 - \exp(-\gamma t)]. \tag{6.27}$$

We have chosen the constitutive gene expression model to verify the FSP-FIM because the exact solution for the Fisher information for Poisson fluctuations can be derived in terms of λ as [91]:

$$\mathcal{I}_{\text{Poisson}(\lambda)} = \frac{1}{\lambda}. \tag{6.28}$$

Figure 6.1 shows the exact value of Fisher information (orange) versus the mean expression level for the two parameters k_r and γ. Figure 6.1 also shows that the FSP-FIM (blue) matches the exact solution for the information on both parameters at all expression levels, which verifies the FSP-FIM for this known analytical form.

Having demonstrated that the FSP-FIM matches to the exact solution, it is instructive to compare how well the previous LNA-FIM and SM-FIM estimates match to the exact FIM computation. For the Poisson distribution, the mean and variance are both equal to λ. Using this fact, the FIM can be approximated using the LNA-FIM for normal distributions (see Eq. 6.25). This expression reduces to

$$\mathcal{I}_{\mathcal{N}(\lambda,\lambda)} = \frac{1}{\lambda} + \frac{1}{2\lambda^2}, \tag{6.29}$$

when both the mean and variance are λ. As λ becomes large, the Poisson distribution becomes well approximated by a normal distribution [92]. Equations 6.28-6.29 show that for this limit of large λ, the first term in Eq. 6.29 dominates, and $\mathcal{I}_{\mathcal{N}}$ reduces to $\mathcal{I}_{\mathrm{Poisson}}$, yielding nearly equivalent values for the expected information. However at low mean expression $\lambda \leq 1$, the strictly positive Poisson and the symmetric Gaussian distributions are less similar, and the Gaussian approximation predicts more information than is actually possible given the exact Poisson distribution. These trends are shown in Fig. 6.1, where the LNA-FIM approach only matches to the exact solution at high expression levels (compare orange and purple lines). For this example, the sample-moments based FIM (SM-FIM) is exact and matches to the analytical and FSP-FIM solutions at all expression levels [14].

6.4.2 The FSP-FIM matches the simulated information for bursting gene expression

Next, we consider a slightly more complicated model of bursting gene expression, in which a single gene undergoes stochastic transitions between active and inactive states with rates k_{on} and k_{off}. This switching model, which is depicted in Fig. 6.2(a), has been studied in detail [98, 104–110], and it has been used to capture single-cell smFISH measurements in mammalian cells [107, 111], yeast cells [11, 106], and bacterial cells [112]. When active, the gene transcribes RNA with constant rate k_r and these RNA degrade in a first order reaction with rate γ. The four reactions of the system are:

$$\mathcal{R}1: \quad g_{\text{off}} \xrightarrow{k_{\text{on}}} g_{\text{on}} \tag{6.30}$$

$$\mathcal{R}2: \quad g_{\text{on}} \xrightarrow{k_{\text{off}}} g_{\text{off}} \tag{6.31}$$

$$\mathcal{R}3: \quad g_{\text{on}} \xrightarrow{k_r} g_{\text{on}} + RNA \tag{6.32}$$

$$\mathcal{R}4: \quad RNA \xrightarrow{\gamma} \emptyset. \tag{6.33}$$

For the examples below, we use the baseline parameters given by: $k_{\text{on}} = 0.05\alpha$ min^{-1}, $k_{\text{off}} = 0.15\alpha$ min^{-1}, $k_r = 5.0$ min^{-1}, and $\gamma = 0.05$ min^{-1}. In particular, the mRNA degradation rate, which sets the overall time-scale, was chosen to be representative of the average decay times (approximately 20 minutes) for mRNA in yeast [113].

For the bursting gene expression model, rescaling the transition rates k_{on} and k_{off} by a common factor does not affect the mean expression level, because the fraction of time spent in the active state remains unchanged. This fraction can be written

$$f_{\text{on}} \equiv \frac{\alpha k_{\text{on}}}{\alpha k_{\text{on}} + \alpha k_{\text{off}}} = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}}, \tag{6.34}$$

and is the same for any $\alpha > 0$. For the parameters given above, the average expression at steady state is given by $k_r f_{\text{on}}/\gamma = 25$. However, rescaling the transition rates does change the shape of the distribution as shown in Fig. 6.2(b-d) [98]. When switching is slow, the gene stays in the “on" and “off" states long enough to observe individual high and low peaks corresponding to the “on" and “off" states, as in shown in Fig. 6.2(b). However, for intermediate switching rates, the gene does not spend enough time in the “off" state for bursts to decay or enough time in the “on" state for large populations to accumulate (see Fig. 6.2(c)). At fast switching rates the “on" and “off" states come to a fast quasi-equilibrium, and the time-averaged system approaches a Poisson process, where the effective production rate is $k_r f_{\text{on}}$. For the bursting gene expression model, all moments of the distributions can be computed exactly from Eq. 2.11, even when the RNA distributions are highly non-Gaussian [12].

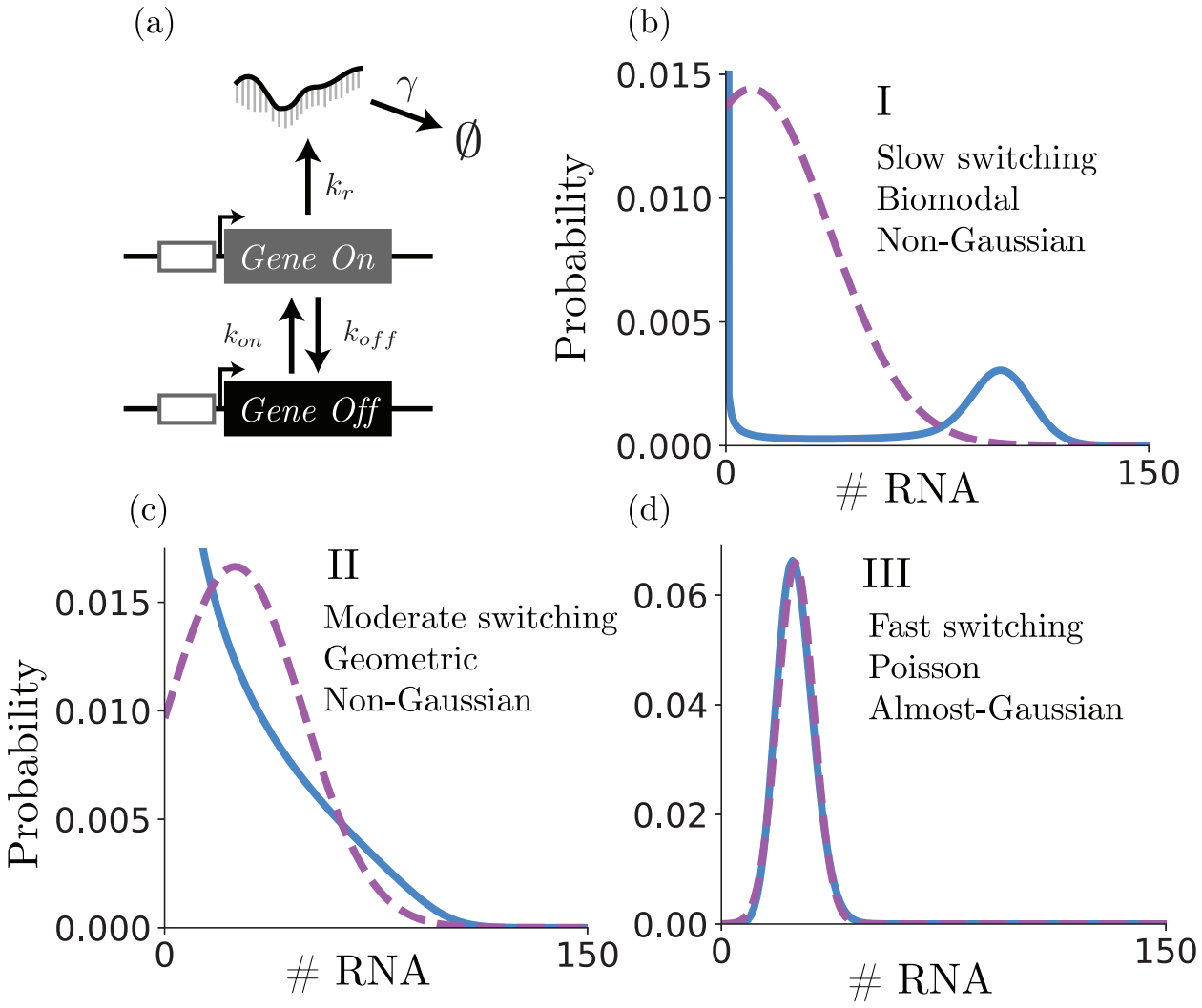

Figure 6.2: *Bursting gene expression.* (a) Schematic of the standard bursting gene expression model. Parameters are defined as given in the text to yield an "on" fraction of 0.25 and a mean expression of 25 mRNA per cell. (b) At slow switching rates, unique "on" and "off" modes are apparent, and distributions of molecule numbers are bimodal. (c) For intermediate switching rates, the distributions are geometric. (d) At high switching rates, the distributions are nearly Poisson (d). For each switch rate scale (labeled I, II, or III), the distribution of RNA computed with the FSP (blue) is compared to a Gaussian with the same mean and variance (purple).

Since the previous example has already verified the accuracy of the FSP-FIM when the expression has a Poisson distribution, we now verify the FSP-FIM for the slow switching case in which the distribution is bimodal ($\alpha = 0.1$). To our knowledge an exact FIM solution is not known for the bursting gene expression model, so we evaluate the different FIM approximations by finding the sampling distribution of the MLE, and we compare the covariance of this distribution to the

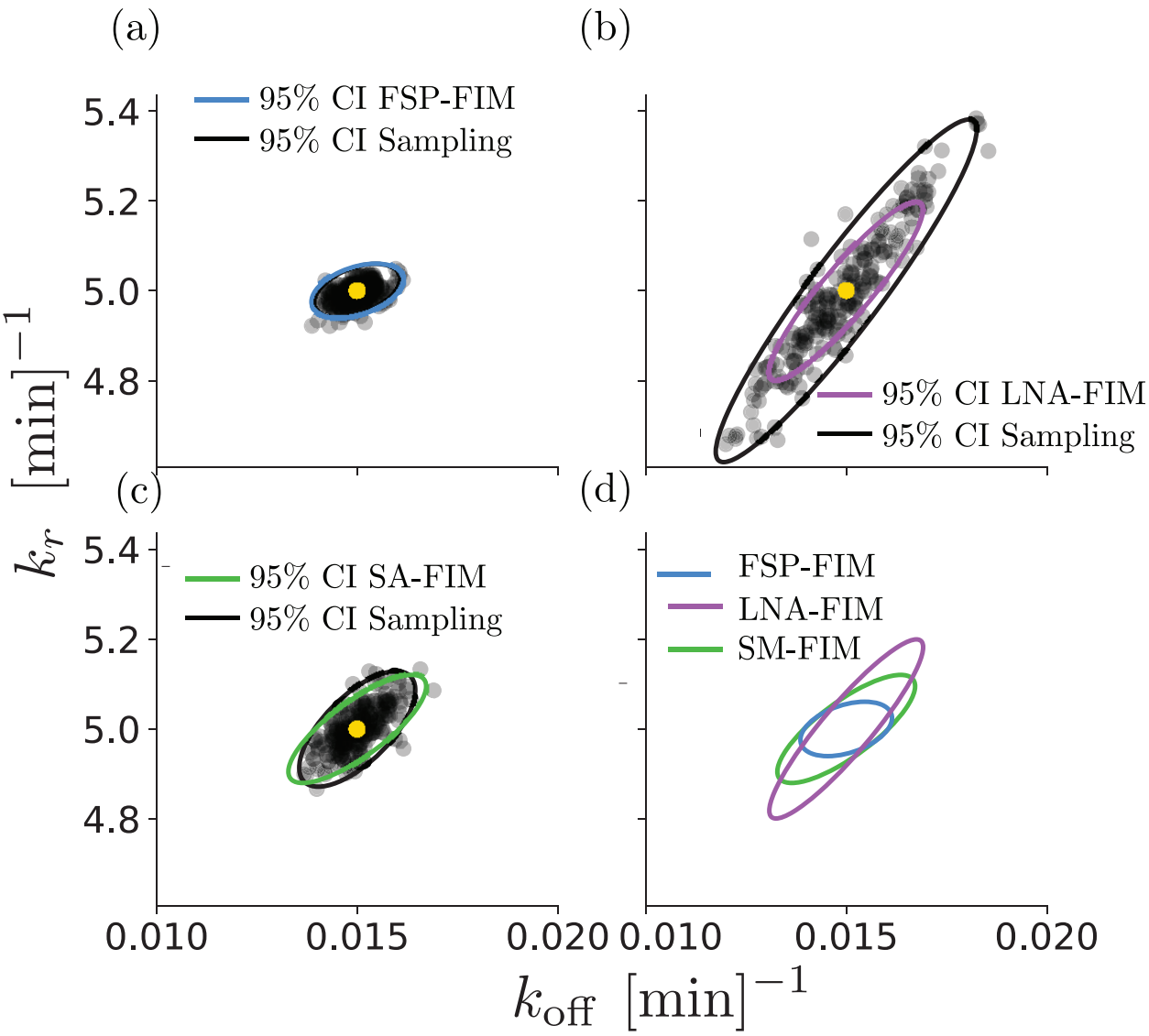

Figure 6.3: *Verification of the FSP-FIM for models with non-Gaussian distributions.* The inverse of the FIM is a lower bound on the variance of the MLE estimator. Here, we simulate 200 data sets with 1,000 cells in each data set. We then find the MLE $\hat{\boldsymbol{\theta}}$ (scatter plots) for each, and compare the covariance of these samples to the inverse of the FIM for the (a) FSP-, (b) LNA-, and (c) SM-FIM approaches. Panel (d) shows the FIM matrices for all approximations on the same axes. Simulated data were generated using the parameters given in the main text and at 10 time points evenly distributed between 0 and 200 minutes.

inverse of the FIM [92]. To do this, we sample from $p(\mathbf{X}; t, \boldsymbol{\theta}^*)$ under reference parameter set $\boldsymbol{\theta}^*$ to generate 200 simulated data sets, each with independent RNA measurements for 1,000 cells. We then allow k_{off} and k_r to be free parameters, and we find $\hat{\boldsymbol{\theta}}$ for each of the 200 data sets. Figure 6.3 compares the 95% confidence intervals found by taking the inverse of the FIM and through MLE estimation using simulated data for the FSP likelihood (Eq. 3.1) shown in Fig. 6.3(a), the

LNA-based likelihood (Eq. 3.3 in the Methods section) shown in Fig. 6.3(b), and the SM-based likelihood (Eq. 3.3 in the Methods section, Supplementary Eq. 10) shown in Fig. 6.3(c). Figure 6.3(a) shows that the CRB predicted by the FSP-FIM matches almost perfectly to the confidence intervals determined by MLE analysis of independent data sets. Figure S3 (left column) shows that this estimate is consistently accurate over multiple different experiment designs. In contrast, the LNA-FIM dramatically overestimates the information and predicts confidence intervals that are much smaller than are actually possible (Figs. 3(b) and S3, center column). The SM-FIM does a better job than the LNA in that it matches the MLE analysis for some experimental conditions (Fig. 6.3(c)) but much less well for other conditions (Fig. S3, right column). We note that the three different FIM estimates yield different principle directions and different magnitudes for parameter uncertainty (Fig. 6.3(d)), but in all cases the FSP-MLE matches to the FSP-FIM and results in the tightest MLE estimation.

Having verified the FSP-FIM for the bursting gene expression model with multiple parameter sets, we next explore how the information changes as a function of the system parameters. Figure 6.4 shows the determinant of the FIM (also known as the D-optimality or information density) for the bursting gene expression model as a function of the switch rate scaling factor, α, using the LNA-FIM (purple), SM-FIM (green) and FSP-FIM (blue) approximations. In the limit of fast switching (i.e. $\alpha \rightarrow \infty$), the expected information converges to nearly the same value for all approaches, as expected for a Poisson distribution with high effective population size of $\lambda = 25$ RNA. However, in the non-Gaussian regimes with slow switch rates, the LNA-FIM over-estimates and SM-FIM under-estimates the information compared to the verified FSP-FIM approach. We note that these differences arise despite the fact that the bursting gene expression model has linear propensity functions, which allows for closed and exact computation of the statistical moments.

6.4.3 The FSP-FIM Can Design More Informative Single-Cell Experiments

Next, having verified the FSP-FIM for its ability to accurately estimate the FIM for different parameter sets, we explore the use of the FSP-FIM to design experiments that maximize infor-

mation. Specifically, we will use classical FIM-based experiment design approaches to choose single-cell experiments first for the bursting gene expression model above, and then for a nonlinear toggle model for which moments can no longer be computed exactly. We consider two different metrics of the FIM, which are frequently used in model-driven experiment design [14, 93]. The first of these is E-optimality, which corresponds to the smallest eigenvalue of the FIM. By finding the experiment which maximizes this eigenvalue, the information is increased in the principle direction of parameter space in which the least information is known (i.e. the parameter uncertainty is highest). The second FIM criteria is D-optimality, which corresponds to the determinant of the FIM. By maximizing the determinant of the FIM over the experiment design space, one finds an experiment which minimizes the volume of the uncertainty in parameter space. We note that many other experimental design criteria are possible, and the choice of criteria depends on what one desires to learn about the system.

Optimizing the sampling rate for bursting gene expression. Our first demonstration of FSP-FIM based experiment design is to select the optimal single-cell sampling period with which to identify the parameters of the bursting gene expression model. For this, we have chosen to analyze E-optimality criteria, which seeks to maximize the smallest eigenvalue of the FIM. We consider a potential experiment design space consisting of 60 logarithmically distributed sampling periods Δt from 2×10^{-2} minutes and 7×10^{2} minutes. For each sampling period, a total of five evenly spaced temporal measurements would be taken. Figure 6.5(a) compares the information expected versus the sampling period using the different FIM approximations: LNA-FIM (purple), SM-FIM (green) and FSP-FIM (blue). For each potential experiment, we then simulate 200 data sets for 1,000 cells each by sampling $p(\mathbf{X};t,\boldsymbol{\theta}^*)$, use Eq. 3.1 to find the MLE parameter estimate for each data set, and then compute the covariance matrix from the MLE parameter sets. This covariance matrix is inverted, and its minimum eigenvalues are depicted as orange triangles in Fig. 6.5(a). Figure 6.5(b) also shows a scatterplot to compare the relationship between the MLE-observed information and the predicted information for all FIM approaches. Once again, the FSP-FIM consistently matches the observed E-optimality at all experimental conditions. However, the LNA approach is much less

consistent, sometimes over-estimating and sometimes under-estimating the real information for the different experimental conditions. The SM-FIM consistently underestimates the true information for this example, although it is not clear if this trend would hold for all sets of parameters and experimental conditions.

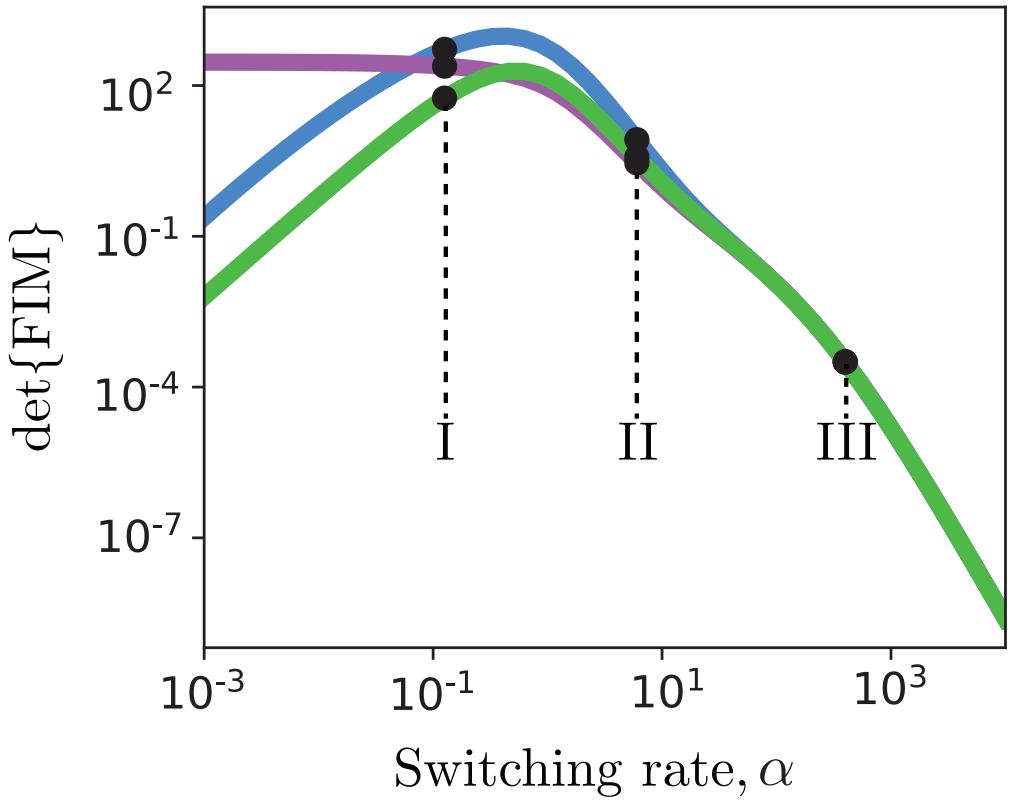

Figure 6.4: *FIM analysis of the bursting gene model.* The determinant FIM for the LNA-FIM (purple), FSP-FIM (blue), and SM-FIM (green) as a function of the gene switching rate scale, α. Labels I, II, III correspond to the switch rates for which distributions are plotted in Figs. 6.2(a-c). Parameters are given in the main text and data are assumed to be collected at 10 equally separated time points between 0 and 200 minutes.

From Fig. 6.5(a), it is clear that the amount of expected information depends strongly on the sampling period. When the sampling period is much longer than the characteristic time to reach the steady state distribution ($\Delta t \gg 1/\gamma$), the information does not change because all snapshots are already close to steady state. When the sampling period is too short ($\Delta t \ll 1/\gamma$), there is insufficient time for the distributions to change and the information tends to zero. Despite conserving these trends, the three different FIM analyses result in substantially different optimal experiments for

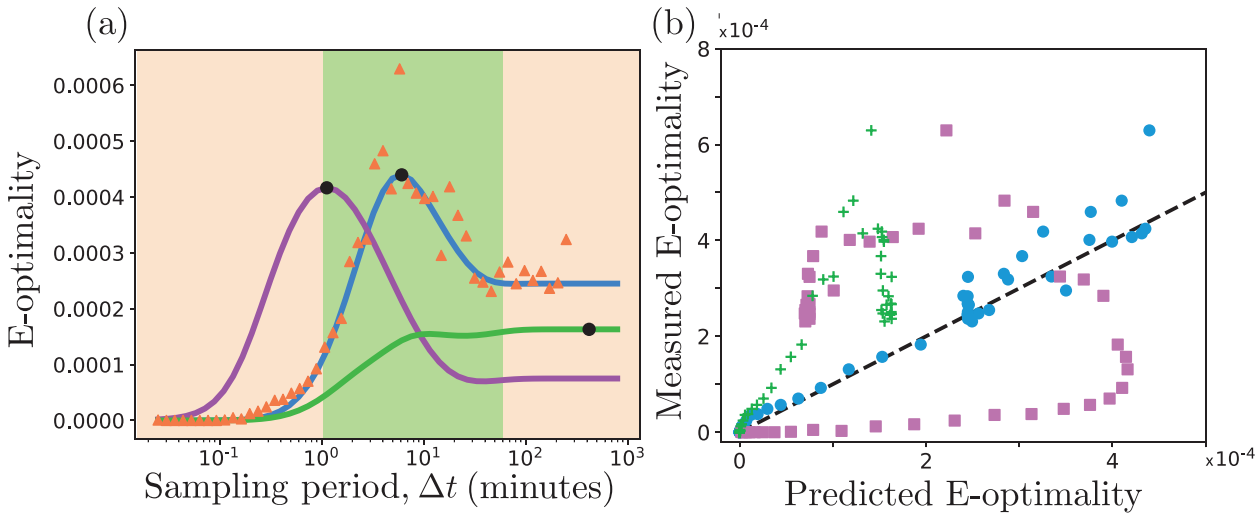

Figure 6.5: *Designing experiments with the FSP-FIM.* (a) E-optimality (i.e., smallest eigenvalue of the FIM) for the standard bursting gene expression model versus sampling period, Δt, using FSP-FIM (blue), LNA-FIM (purple), and SM-FIM. Maximizing E-optimality corresponds to minimizing variance in the in the most variable direction of parameter space. The orange triangles show MLE-based confirmation of the E-optimality, using 200 simulated data sets for each sampling period. The green shaded region represents experiments that are feasible using smFISH, from minute resolution [11] to hour resolution [112] (b) Comparison of the FSP-FIM (x-axis) versus the observed information (y-axis) for various sampling periods using the FSP-FIM (blue circles), LNA-FIM (purple squares), and SM-FIM (green crosses). Kinetic parameters are given in the main text.

the E-optimality design criteria. Using the FSP-FIM, the optimal experiment is $\Delta t = 6.1$ minutes, which we verified using the MLE sampling approach (compare orange triangles and blue line in Fig. 6.5(a)). This optimal design is well-aligned with smFISH experimental technique, which can capture cell populations with one minute resolution [11] to one hour resolution [112]. However, the LNA-FIM selects a much faster sampling period of $\Delta t = 1.1$ minutes, and the SM-FIM selects a much slower sampling period of $\Delta t = 420$ minutes. Thus, the FSP-FIM not only provides more information compared to moments-based approaches, but it also provides a better estimate of the expected information. In turn, these improved estimates can help to avoid potentially misleading experiments and select optimal designs.

The FSP-FIM accurately estimates information for systems with nonlinearities and bimodal responses. To demonstrate the utility of the FSP-FIM approach for models with nonlinear reaction propensities and multiple species, we turn to the toggle model first introduced by Gardner et al [48], with a stochastic formulation by Tian and Burrage [49]. Figure 6.7(a) shows a schematic

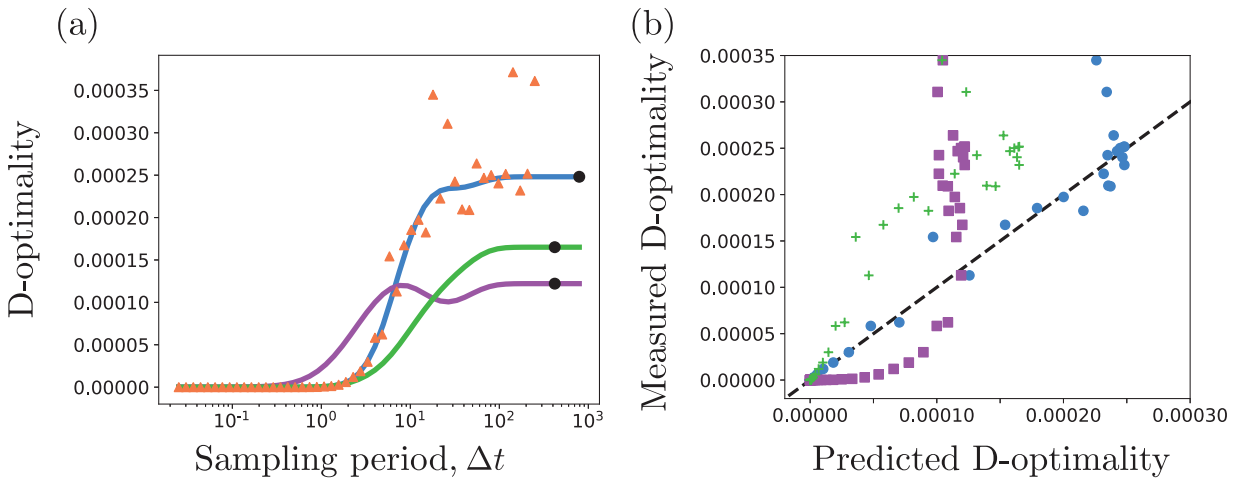

Figure 6.6: Optimal experiment design for the bursting gene expression model using the determinant of the FIM, D-optimality. (a) The D-optimality criteria for the FSP-FIM (blue), LNA-FIM (purple) and SM-FIM (green) for different sampling periods Δt. Orange triangles represent the D-optimality confirmed using 200 simulated data sets for each potential sampling period. Optimal sampling periods are given by black circles. (b) Comparison of the FSP-FIM at the the reference parameter set (x-axis) and the observed information (y-axis) for various sampling periods using the FSP-FIM (blue circles), LNA-FIM (purple squares), and SM-FIM (green crosses). Kinetic parameters are $k_{on} = 0.05\ \text{min}^{-1}$, $k_{off} = 0.15\ \text{min}^{-1}$, $k_r = 5$ molecules/min, and $\gamma = 0.05\ \text{min}^{-1}$.

of the toggle model, which consists of two mutually repressing proteins, $x \equiv$ LacI and $y \equiv \lambda$cI, where the production of each species depends non-linearly on the concentration of its competitor. The reactions in the toggle model can be written

$$\mathcal{R}1 : \emptyset \xrightarrow{w_1} x; \qquad \mathcal{R}2 : x \xrightarrow{w_2} \emptyset; \tag{6.35}$$

$$\mathcal{R}3 : \emptyset \xrightarrow{w_3} y; \qquad \mathcal{R}4 : y \xrightarrow{w_4}; \emptyset \tag{6.36}$$

where

$$w_1 = b_x + \frac{k_x}{1 + \alpha_{yx} y^{\eta_{yx}}}; \qquad w_2 = \gamma_x x; \tag{6.37}$$

$$w_3 = b_y + \frac{k_y}{1 + \alpha_{xy} x^{\eta_{xy}}}; \qquad w_4 = \gamma_y(\text{UV}) y. \tag{6.38}$$

In this formulation, we have assumed that the degradation of λcI is controlled by an ultraviolet (UV) radiation through the light-induced circuit described by Kobayashi et al [114]. Similar to

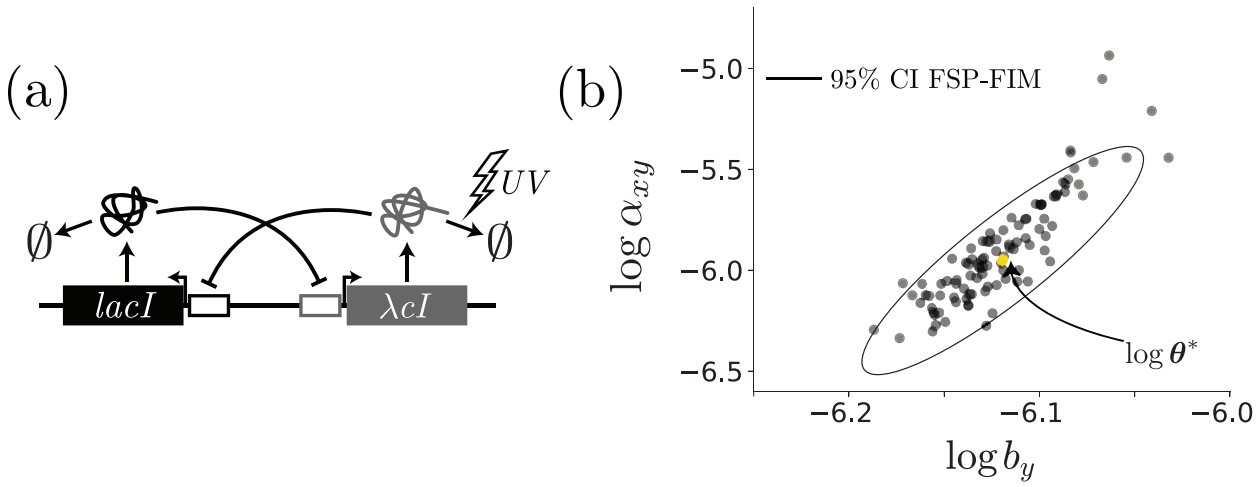

Figure 6.7: *Validation of a toggle model.* (a) Model schematic of the two genes, *lacI* and λcI, which are mutually repressing [48]. Degradation of λcI is controlled by UV radiation. (b) Verification of the FSP-FIM (black ellipse) for 200 MLE estimates of 1,000 cells each (black dots) for two free model parameters, α_{xy} and b_y.

[30], we assume that the UV level affects the degradation of λcI according to the function:

$$\gamma_y(\mathrm{UV}) = 3.8 \times 10^{-4} + \frac{0.002\mathrm{UV}^2}{1250 + \mathrm{UV}^3}, \tag{6.39}$$

where the minimum degradation rate has been chosen to match dilution due to the *E. coli* half life of 30 min [30]. The remaining parameters used for this example are given by $\boldsymbol{\theta}^*$ in Table 6.1. The system's initial condition at $t = 0$ is assumed to be the equilibrium distribution when no UV is applied. For this biological system and these parameters, different levels of UV radiation will give rise to different dynamics. At low levels of radiation, switching to the high LacI state is rare, and the distribution tends to have a single peak. At intermediate levels of radiation, switching between low and high levels of LacI expression is possible, and LacI distributions may be bimodal. Finally, at high levels of radiation, the system very quickly switches into the high LacI state.

Because this model has complex nonlinear propensity functions, the statistical moments cannot be calculated in closed form, and the LNA-FIM and SM-FIM estimates are no longer expected to provide accurate estimates for information or optimal experiment designs. In contrast, the FSP analysis remains unchanged, and the FSP-FIM can be computed exactly as above. As before, we

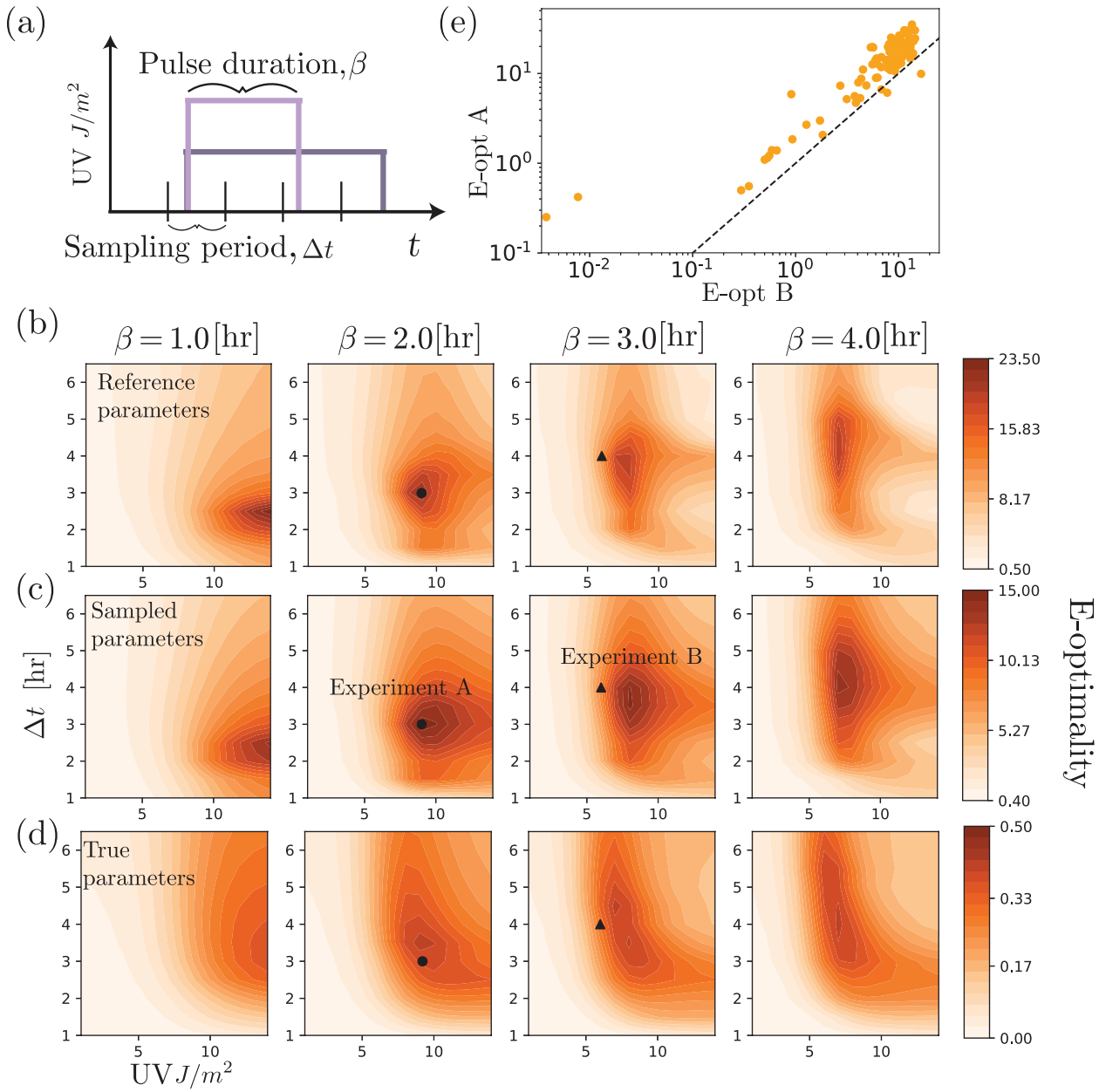

Figure 6.8: *Experiment design for the nonlinear genetic toggle model.* (a) Degradation rate of λcI is controlled by UV as shown in Fig. 6.7(a). The magnitude and duration (β) of UV exposure are free experiment design parameters, along with the time between measurements Δt. (b) E-optimality (the smallest eigenvalue of the FIM) versus the 3-dimensional experiment design space, where the FIM is computed using (b) the reference parameter set, (c) by averaging the E-optimality over 100 unique parameter sets and (d) using the "true" parameter values. The black circle is the optimal design chosen according to (c). The black triangle denotes a nearby, but less informative, experiment. (e) For the experiments corresponding to the black circle and triangle in (b-d), E-optimality values are shown for each sampled parameter set.

verify the FSP-FIM for this nonlinear case using a set of 200 simulated data sets measured at 1 hr, 4 hr, and 8 hr, each with 1,000 cells, and we found MLE parameter estimates $\hat{\boldsymbol{\theta}}$ for each simulated data set. Figure 6.8(a) shows this verification in a simple case with two free parameters, b_y and α_{xy}, and Fig. 6.9 shows the verification where all parameters free except for Hill coefficients η_{xy} and η_{yx}. In this and all subsequent analysis of the toggle model, we have used the logarithmic parameterization of the FIM (Eq. 6.10).

Next, we aim to design more complex experiments for the toggle model described above. We consider an experiment design space where the measurement sampling period (Δt), pulse duration (β), and pulse magnitude (UV) can all be changed, as illustrated in Fig. 6.8(a). Each pulse of UV starts at $t = 1$ hr. We then compute the FSP-FIM for each experiment $\{\text{UV}, \beta, \Delta t\}$.

To capture the more realistic situation where parameters are unknown prior to experimentation, we next explore how parameter uncertainty affects the estimation of the FIM and the design of optimal experiments. To begin, we assume that parameters have been partially estimated from a simple initial experiment corresponding to measurements of the unperturbed steady state at zero UV input to the system. In practice, similar preliminary parameter estimates could be acquired from literature, from previous less-optimized experiments, or by comparison to related pathways or organisms. For our analysis, the prior estimate for parameters is described by a multivariate lognormal distribution with a geometric mean of $\hat{\boldsymbol{\theta}}_0$ given in Table 6.1. Parameters sampled from this distribution are substantially different from the "true" parameter, $\boldsymbol{\theta}^*$, which is also shown in Table 6.1. Figure 6.8(b) shows the E-optimality criteria for parameter set $\hat{\boldsymbol{\theta}}_0$ as a function of the experiment design parameters $\{\text{UV}, \beta, \Delta t\}$. Next, we sampled 100 random sets of parameters from the prior distribution (Fig. 6.10), and we computed the E-optimality for each set. Figure 6.8(c) presents expected information versus experiment design averaged over these 100 parameter sets. For comparison, Fig. 6.8(d) shows the information versus experiment designs if one had exact knowledge of the true parameters.

From Figs. 6.8(b-d), we observe that relative estimates of the FIM remain consistent despite substantial changes to the parameters from which the FIM is computed. To explore this observa-

	$\boldsymbol{\theta}^*$	$\hat{\boldsymbol{\theta}}_0$	units
b_y	6.80×10^{-5}	9.86×10^{-4}	s^{-1}
b_x	2.20×10^{-3}	3.19×10^{-3}	s^{-1}
k_y	1.60×10^{-2}	1.60×10^{-2}	s^{-1}
k_x	1.70×10^{-2}	2.50×10^{-2}	s^{-1}
α_{xy}	6.10×10^{-3}	8.28×10^{-3}	$N^{-\eta_{xy}}$
α_{yx}	2.60×10^{-3}	2.46×10^{-3}	$N^{-\eta_{xy}}$
η_{xy}	2.10	2.10	-
η_{yx}	3.00	3.00	-
γ_x	3.80×10^{-4}	5.57×10^{-4}	$N^{-1}\mathrm{s}^{-1}$

Table 6.1: Parameters for the toggle model. $\boldsymbol{\theta}^*$ is the "true" parameter set from which data is generated, and $\hat{\boldsymbol{\theta}}_0$ is the MLE parameter set fit to a baseline data set generated assuming 0 UV (see Fig. 6.10 for further discussion). Here, N is used to denote the units of single-molecules.

	Single experiment	Dual greedy	Dual simultaneous
$\left\{\begin{array}{c}\beta\\ \Delta\\ \mathrm{UV}\end{array}\right\}$	$\left\{\begin{array}{c}2\text{ hr}\\ 3\text{ hr}\\ 9\ J/m^2\end{array}\right\}$	$\left\{\begin{array}{c}4\text{ hr}\\ 5.5\text{ hr}\\ 14\ J/m^2\end{array}\right\}$	$\left\{\begin{array}{c}1\text{ hr}\\ 2.5\text{ hr}\\ 9\ J/m^2\end{array}\right\}, \left\{\begin{array}{c}4\text{ hr}\\ 2.5\text{ hr}\\ 13\ J/m^2\end{array}\right\}$
E-opt	14.9	32.0	36.8

Table 6.2: Comparing sequential experiment design approaches.

tion more closely, we selected the experiment that maximizes the averaged E-optimality in Fig. 6.8(c). This experiment is denoted by a black circle in Figs. 6.8(b-d), and we compare it to another similar experiment design, shown by the black triangle in Fig. 6.8(b-d). Figure 6.4.3 shows the expected parameter uncertainty for these two designs and shows that the optimal experiment reduces variance in some parameter directions by more than an order of magnitude compared to the sub-optimal experiment. To explore how different parameters change the ranking of these two experiments, we analyze the ranking of Experiment A and Experiment B not only based on their average E-optimality value as in Fig. 6.8(c), but at each of 100 random parameter combinations. Figure 6.8(e) shows that for 97 of the 100 parameter samples, the relative ranking of the experiments is consistent, even though the absolute value of the E-optimality criteria varies over several orders of magnitude.

The analysis shown in Fig. 6.8 assumes a fixed initial distribution at $t = 0$, which was specified by the steady state distribution under the true parameters in the absence of UV radiation. Under this

$$UV = 0 \ J/m^2$$

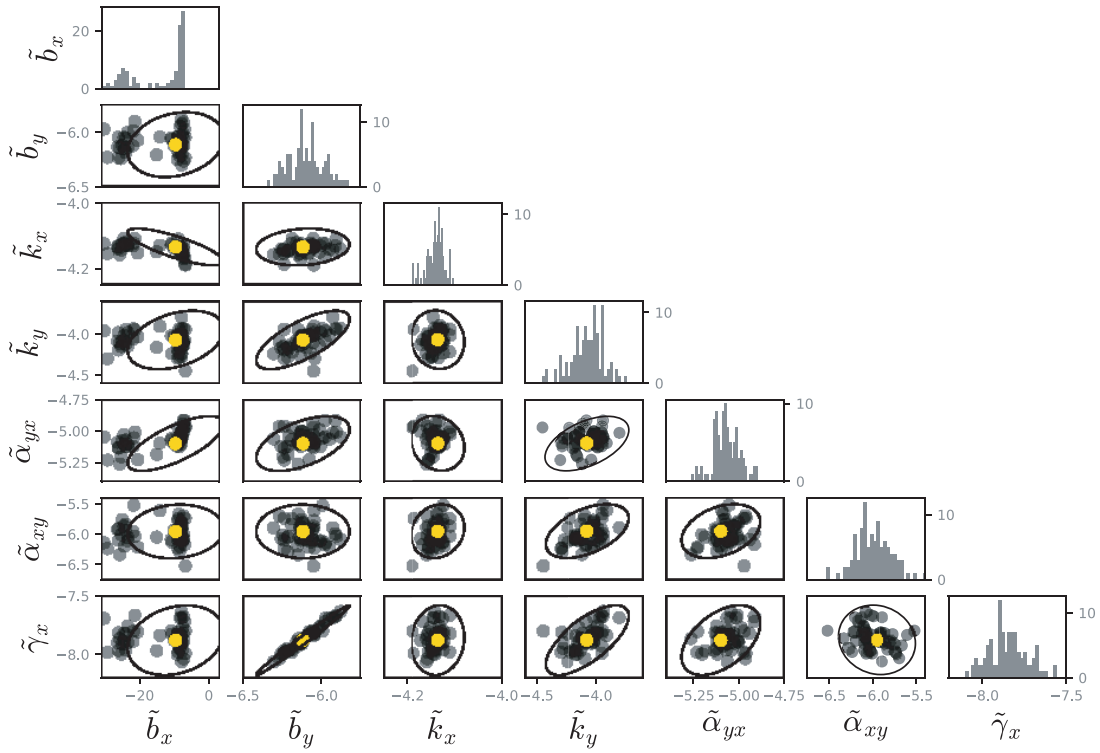

Figure 6.9: Verification of the FSP-FIM for the seven free parameters for the toggle model. Each black circle corresponds to the logarithm of an MLE estimate, $\log \hat{\boldsymbol{\theta}}$ for 100 different simulated data sets. The gold circle corresponds to the reference parameter set, $\log \boldsymbol{\theta}^*$. The 95% ellipse corresponding to the log-FSP-FIM is shown in black. The tilde corresponds to the log of each parameter, i.e. $\tilde{b}_x = \log b_x$.

assumption, the initial sensitivity matrix $S(0)$ in Eq. 6.24 was set to zero. Figure 6.12 extends the analysis to compute the initial sensitivity $\mathbf{S}_{\theta_i}(0) = \partial \mathbf{p} / \partial \theta_i$ at steady state, which slightly increases the estimate of information for the early time points, but has relatively little effect on the choice of optimal experiment design.

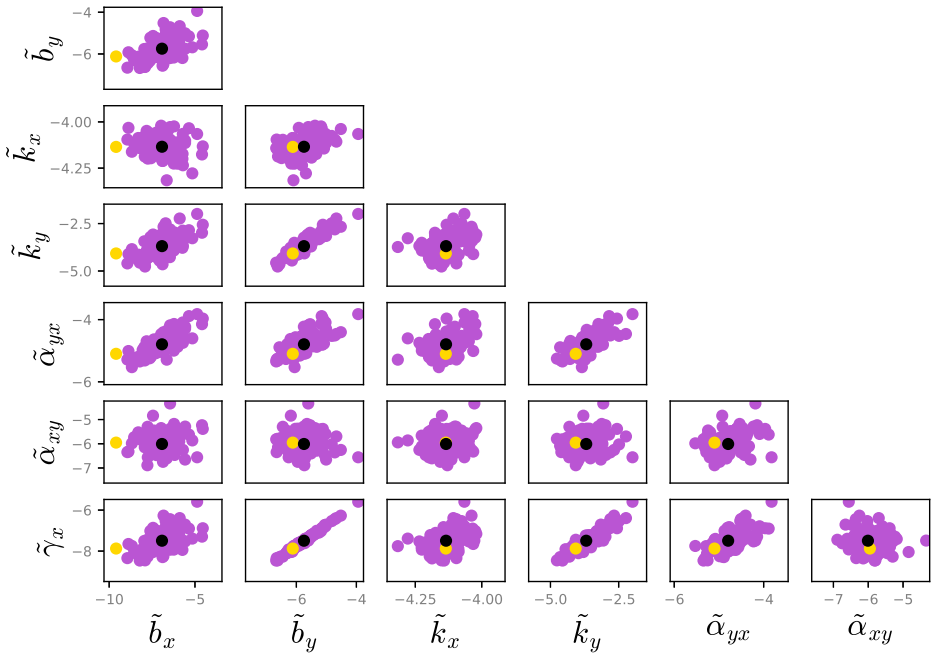

Figure 6.10: Parameters were sampled 100 times from a log-normal distribution evaluated about the reference parameter set $\hat{\boldsymbol{\theta}}_0$, shown in black. The covariance of this distribution was chosen according to the inverse of the FIM evaluated for an experiment with 0 UV, $t = [1, 4, 8]$ hr, and 100 measurements at each time point. For reference, the gold parameters are the 'true' parameters for the model. The tilde corresponds to the log of each parameter, i.e. $\tilde{b}_x = \log b_x$.

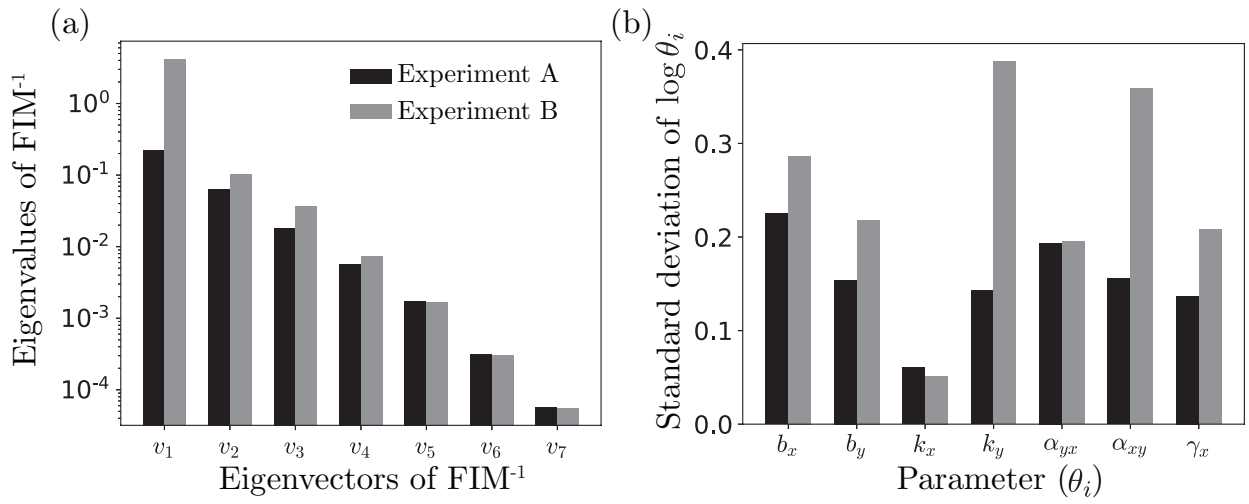

Figure 6.11: (a) The eigenvalues of the inverse of the Fisher information (i.e. the CRB) for the two experiments in Fig. 7(c) in the main text. Note that the y-axis is on a logarithmic scale. Lower values correspond to lower parameter uncertainty. (b) The effect of Experiment A and B on standard deviations of $\log \theta_i$.

We next seek to understand how optimal experiments depend on one's plans to perform multiple experiments. The "single experiment" in Table 6.2 refers to designing a single experiment, $\mathcal{E}_1$, to maximize the expected FIM design criteria, such as finding the maximal combination in Fig. 6.8(c). The "dual greedy" approach also chooses the same $\mathcal{E}_1$ and then seeks to find the most complementary additional experiment, $\mathcal{E}_2$, to maximize the overall FIM design criteria. Finally, the "dual simultaneous" search finds the optimal combination of any two possible experiments, $\hat{\mathcal{E}}_1$ and $\hat{\mathcal{E}}_2$ to maximize the design criteria. Since the optimal choice for $\hat{\mathcal{E}}_1$ and $\hat{\mathcal{E}}_2$ can admit the other choices, it must yield at least as high a design criteria as $\mathcal{E}_1$ and $\mathcal{E}_2$. By comparing the three design strategies for the current toggle model, we find indeed that the simultaneous approach discovers a substantially more informative experiment than does the greedy approach. In other words, the overall expected value of an experiment, can depend not only on the current parameter values, but also upon which other experiments one intends to conduct. If one has plans to do multiple experiments, it may be better to consider the potential information from all experiments as a whole rather than to design each experiment one at a time.

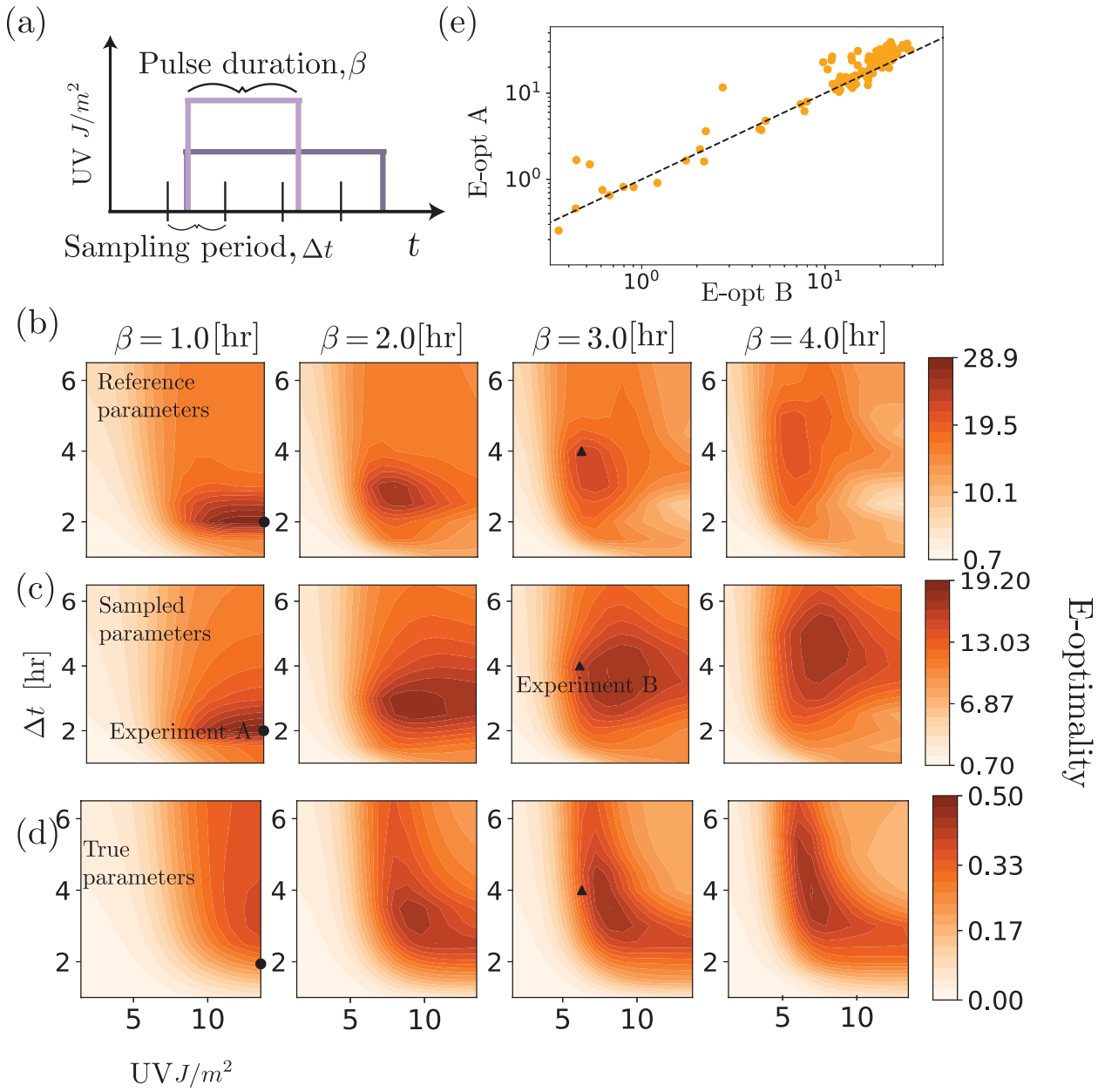

Figure 6.12: Toggle model experiment design with non-zero initial sensitivities (a) Degradation rate of λcI is controlled by UV as shown in Fig. 7(a). The magnitude and duration (β) of UV exposure are free experiment design parameters, along with the time between measurements Δt. (b) E-optimality (the smallest eigenvalue of the FIM) versus the 3-dimensional experiment design space, where the FIM is computed using (b) the reference parameter set, (c) by averaging the E-optimality over 100 unique parameter sets and (d) using the "true" parameter values. The black circle is the optimal design chosen according to (c). The black triangle denotes a nearby, but less informative, experiment. (e) For the experiments corresponding to the black circle and triangle in (b-d), E-optimality values are shown for each sampled parameter set.

6.5 Discussion

Fluctuations in biological systems complicate modeling by introducing substantial variability in gene expression among individual cells within a homogeneous population. This variability contains valuable and quantifiable insights [98], but data with multiple peaks and long tails, such as those collected using smFISH, are often poorly described by modeling approaches that only make use of low-order moments of such distributions [20]. The FSP approach [27] has previously been used to identify and predict gene expression dynamics for complex and rich single-molecule, single-cell data [11, 111, 112]. In this work, we have developed the FSP-based Fisher information matrix, which extends the FSP analysis to allow rigorous design of experiments that are optimally informative about the model's parameters.

The FSP-FIM uses a novel sensitivity analysis, which requires solving a system of ODEs that is twice the size of the FSP dimension for each parameter, and therefore should be computationally tractable for any problem to which the FSP can be applied. The local sensitivity of each parameter is independent of the other parameters, so the computation is easily parallelized among multiple processors. We verified that the FSP-FIM approach matches the information for the constitutive gene expression model, whose response follows a Poisson distribution (Fig. 6.1), and for which the FIM can be computed exactly. The FSP-FIM also matches to classical FIM approaches that assume normally distributed data (LNA-FIM) or very large data sets (SM-FIM) in the limiting case when the data distributions are close to being Gaussian (Figs. 6.1-6.4). For systems where data is not Gaussian and for which there is no exact FIM formula, we showed that the FSP-FIM is more accurate than traditional approaches (Figs. 6.4, 6.5), which we validated by generating many independent data sets and comparing the inverse of the FSP-FIM to the variance in the MLE estimates (Figs. 6.3 and 6.7).

We showed that the choice of FIM analysis can lead to different optimal experiment designs (Fig. 6.5). For example, Fig. 6.5 shows that the LNA-FIM can substantially overestimate the information of certain experiments compared to the full, correct information obtain using the FSP-FIM, which could mislead researchers to choose experiment designs that are much worse than they ex-

pect. In practice, overestimation of the Fisher information can have the further deleterious effect of overconfidence in poor parameter estimates, which can result in model bias and poor predictions as we observed recently in [20]. Furthermore, the LNA-FIM is not self-consistent, and often overestimates the information even compared to the ellipse found from sampling the MLE with the Gaussian likelihood function. On the other hand, we found that the SM-FIM under-estimated the information for the bursting gene model, but the amount of underestimation varied substantially with experimental conditions, which could cause researchers to reject otherwise informative experiments. In contrast to these moment-based approaches, the MLE sampling using the FSP approach always provided the best parameter estimates (Figs. 3 and S3), and the FSP-FIM was always consistent with the confidence intervals verified by sampling (Figs. 1 6.1, 6.3, 6.5), even for the case of nonlinear reaction propensities for which exact moments cannot be found (Figs. 6.7(a), and 6.9).

In our analysis of the non-linear toggle model, we allowed for the independent control of three experimental variables (Fig. 6.8a), and found experiments that optimize particular criteria of the FIM. Furthermore, we showed that other experiments very near to the optimal experiment in the design space can be significantly less informative than the optimal experiment (Figs. 6.8(b-e) and 6.4.3. Choosing between such similar experiment designs is non-trivial and would be difficult or impossible using intuition alone. On the other hand, we explored the effects of parameter uncertainty on FSP-FIM-based experiment design, and we found that parameter rankings are relatively robust to parameter uncertainty, even when the absolute value of the FSP-FIM is sensitive (Fig. 6.8).

We found that that the choice of optimal experiments depends on the number of experiments to be completed (Table 6.2). For example, the optimal set of two experiments may not contain the optimal single experiment. Sometimes, the FIM for a given experiment may be singular or nearly singular, indicating that the model under investigation is unidentifiable for the current parameterization and experiment design. In such a case, the FIM-eigenvectors corresponding to the near-zero eigenvalues indicate specific linear combinations of parameters that are poorly constrained (e.g.,

'sloppy' directions [21]). If a second complementary experiment can shift the orientation of these sloppy vectors, then those parameters may yet be uncovered through combinations of multiple experiments. Alternatively, if a given combination of parameters remains unidentifiable for all admissible experiments, then these irrevocably sloppy directions may be used to reformulate the model into one that has a reduced set of fully identifiable parameters. We note that as one conducts new experiments and collects new data, parameter posteriors will need to be updated. As this occurs, optimal experiments may also need to be adjusted (e.g., through application of a Bayesian experiment design framework [115]), and future developments are needed to incorporate FSP-FIM computations within such iterative frameworks.

Our results show that the FSP-FIM performs better than previous approaches for gene regulation models with low molecule counts or nonlinear reaction rates. Previous studies have demonstrated many realistic systems for which such FSP can be used to identify and predict stochastic dynamics in numerous biological systems [11, 16, 20, 42, 111, 112, 116–118]. Each of these studies has used different experimental input signals, such as temporal salinity profiles [11, 20], temperature [112], or chemical induction [42, 111]. Modern optogenetic experiments promise to allow for even more robust and flexible control of input signals to control cellular behavior [90, 119, 120]. For such studies, the FSP-FIM could now be used to optimize these signals to achieve maximally informative experiments.

Like any other tool, the FSP-FIM also has its associated challenges. Our initial investigations focused on intrinsic stochastic fluctuations of small biochemical processes, and we used simulated data to verify our new computational tools. For models with large molecular counts of four or more species or with the addition of mechanisms to account for extrinsic variability, existing methods to solve the FSP-FIM will remain intractable until more efficient probability density based CME analyses can be developed to address such problems [82, 121–124]. Until higher dimension CME approaches are developed, approximate moment-based experiment design methods, such as the SM-FIM and LNA-FIM, may remain the only available options to design experiments for large biochemical pathways. We also note that real experiments come with additional sources of noise,

such as the errors or uncertainties associated with experimental measurements. For example, in smFISH data analysis, image processing settings give rise to variability in final RNA counts due to density dependent co-localization of RNA molecules. This measurement uncertainty may have a non-negligible effect on parameter inference, and future controlled experiments are needed to elucidate the degree to which such effects depend on optical imaging settings. Fortunately, such variabilities are easily incorporated within the framework of the FSP analysis. For example, previous work has used a simple linear transformation to adapt FSP analyses to include the effects of noisy GFP fluorescence emission and background autofluorescence when comparing integer-valued biochemical models to flow cytometry data in arbitrary continuous units of fluorescence [42]. Once adapted to take these transformations into account, the FSP-FIM could be used to design experiments to minimize the effects of measurement noise.

New experimental capabilities are creating an enormous potential to probe single-cell biological responses. These capabilities are making it ever more difficult to choose what species in the system to measure, whether to measure joint distributions (i.e. measure the RNA counts from multiple genes in the same cells) or marginal distributions (only measure RNA counts from a single gene at a time), or in what condition. Furthermore, different experiments have different costs, and the experimentalists must not only optimize their information about model parameters, but also consider the trade-off between collecting more data and the cost of a given experiment. By providing a new computational tool to iteratively improve models and design experiments for an important class of biological problems, the FSP-FIM will help to improve quantitative predictive modeling of gene expression.

Chapter 7

Optimal Allocation of Single-Cell Measurements for the HOG-MAPK Pathway in S. Cerevisae

7.1 Introduction

The HOG pathway is a pathway commonly studied to understand nuclear localization in response to phosphorylation. We have previously used stochastic models of kinase activated transcription to predict adaptive responses across entire cell populations in yeast [11,20,43]. However, this effort took hours of significant computational effort on both the image processing of smFISH measurements and fitting of experimental data. In this work, we use the newly developed FSP-FIM (Chapter 6 to show how we could have used this model to optimize the design of experiments to use less measurements to infer the model.

Downstream activation of stress response genes such as STL1 and CTT1 depend on nuclear localization dynamics of MAPK, a kinase which is phosporylated in response to salt stress and localizes to the nucleus. Depending on the concentration and rate of application of this stress, the nuclear localization dynamics of MAPK can be different (shown in Fig. 7.1b). While the first study optimizes the experiments to minimize the uncertainty in model parameters, in this study we find the optimal smFISH measurement times and cell numbers to minimize uncertainty about the amount of salt in the environment, and verify this method.

7.2 Background

7.2.1 Finite State Projection models of osmotic stress response in yeast.

Gene expression and regulation is a complicated process in which transcription factors, chromatin modifiers, and more interact with DNA, RNA, protein the cellular environment to carry out different functions. For example, stress response genes are activated in bacteria in response to

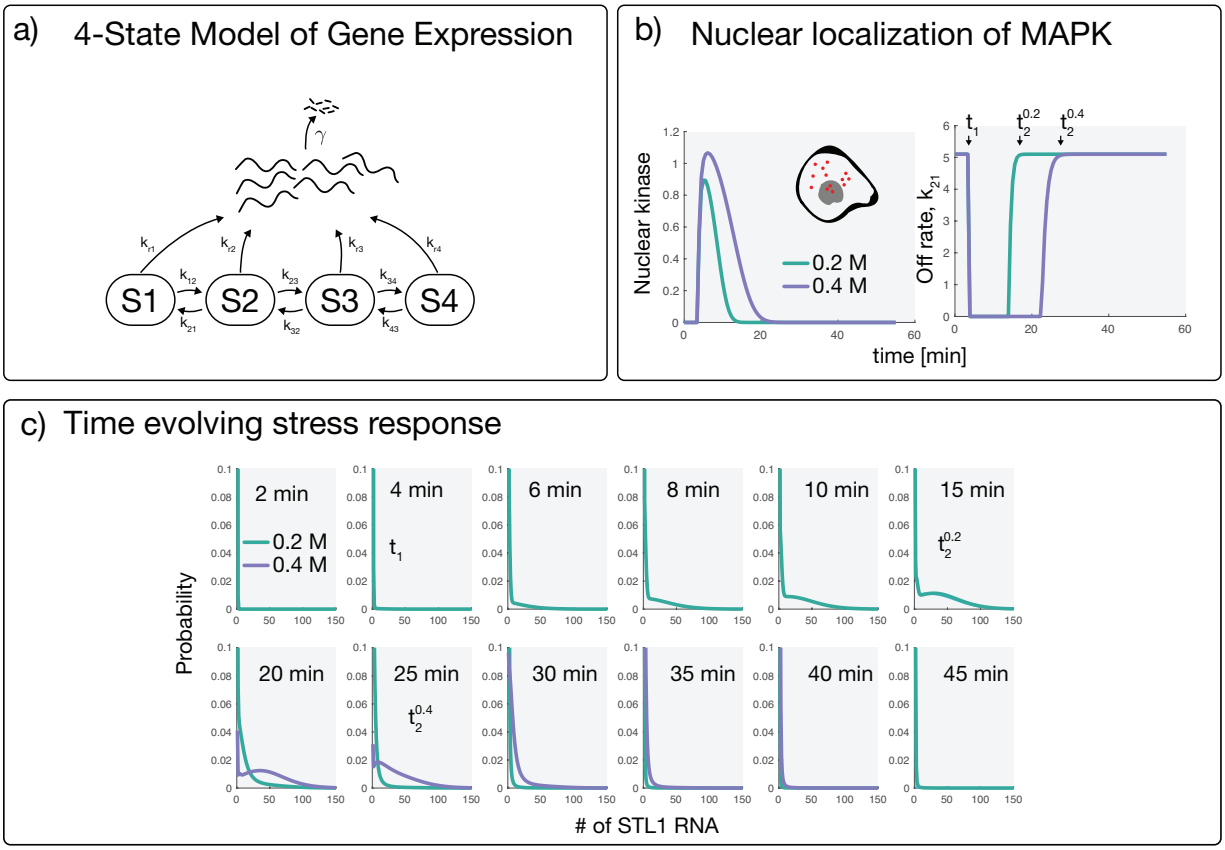

Figure 7.1: *Stochastic modeling of osmotic stress response genes in yeast.* (a) Four state model of gene expression, where each state creates mRNA with a different transcription rate. These mRNA degrade with rate γ. (b) The effect of measured MAPK nuclear localization (left) on the the rate of switching from S2 to S1 (right) under both 0.2M and 0.4M osmotic stress. (c) Time evolution of the STL1 RNA in response to the 0.2M and 0.4M salt stress.

heat shock [112]. Furthermore, stress response genes often behave stochastically across isogenic cell populations, and therefore models of such cellular responses must capture stochastic behavior. The chemical master equation framework of stochastic chemical kinetics has been the workhorse of stochastic modeling of gene expression, whether through simulated sample paths of its solution via the stochastic simulation algorithm [33], moment approximations [12], or finite state projections [27]. Recently, it has come to light that for some systems it is extremely important to consider the full distribution of biomolecules when fitting CME-based models [20,125]. For signal activated transcription in the HOG-MAPK stress response pathway in yeast, an FSP model has been used to fit and predict mRNA distributions at a variety of salt concentrations [11,20].

This model of osmotic stress response consists of transitions between four different gene states, shown in Fig. 7.1(a). The probability of a transition from the i^{th} to the j^{th} gene state occuring in the infinitesimal time dt is given by $k_{ij}dt$. Each state also has a corresponding transcription rate of mRNA, k_{ri}, and the mRNA degrade with rate γ. Further descriptions and validations of this model are given in the supporting information and in [11,20,43]. To accurately fit and predict RNA levels across cell populations, the authors in [11] cross-validated across a number of different potential models with different numbers of gene states and time varying parameters. The most predictive of these was the model shown in Fig. 7.1(a),in which the transition rate from the second gene activation state to the first gene activation state is a function of nuclear MAPK levels, $f(t)$. The nuclear localization of MAPK affects this transition with a threshold function,

$$k_{21}(t) = \max[0, \alpha - \beta f(t)]. \tag{7.1}$$

Figure 7.1(b) shows the nuclear localization dynamics of MAPK (i.e. $f(t)$) at 0.2M and 0.4M, with simulated nuclear localization dynamics fit to a model (from [20]), and Fig. 7.1(c) shows the value of $k_{21}(t)$ for each salinity level. This rate results in a time-vary generator for the master equation dynamics.

The generator matrix for the FSP system can be written as a sum of multiplied by corresponding parameters,

$$\mathbf{A}(t) = \sum_{i=1}^{K} \theta_i(t)\mathbf{A}_i. \tag{7.2}$$

The different $\mathbf{A}_i$ are matrices of 1's and -1's. The dynamics of the CME are then given by $\frac{d\mathbf{p}}{dt} = \mathbf{A}(t)\mathbf{p}$.

7.2.2 Finite State Projection based Fisher Information for signal-activated stochastic gene expression.

The Fisher information matrix (FIM), is a common tool in engineering and statistics to find estimates of parameter uncertainties prior to collecting data, which allows one to find the experimental settings which make these predicted uncertainties as small as possible. Recently, it has been applied to biological systems to estimate kinetic rate parameters in stochastic gene expression systems [14, 38, 39, 41, 125]. In general, the FIM for a single measurement is defined

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left\{ \left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})\right)^T \left(\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})\right) \right\}. \tag{7.3}$$

As the number of measurements N_c is increased such that maximum likelihood estimates (MLE) of parameters are unbiased, the distribution of MLE estimates is known to be normally distributed with a variance given by the inverse of the Fisher information matrix, i.e.

$$\sqrt{N_c}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{dist} \mathcal{N}(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}). \tag{7.4}$$

In chapter 6, we developed the FSP-based Fisher information matrix (FSP-FIM), which allows one to use the FSP solution $p(\mathbf{x}, t)$ and the piecewise-sensitivity matrix $\mathbf{S}$ to find the Fisher information matrix for stochastic gene expression systems. The dynamics of the sensitivity of each state in the process to the i^{th} kinetic parameter $\frac{d\mathbf{p}}{d\theta_i}$ is given by

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}^i\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{A}_i & \mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}^i\end{bmatrix}, \tag{7.5}$$

where $\mathbf{A}_i = \frac{\partial \mathbf{A}}{\partial \theta_i}$, and for linear models is the same as the $\mathbf{A}_i$ given in Eq. 7.2 [125]. The FSP-FIM at a single time point is given by

$$\mathbf{F}(\boldsymbol{\theta}, t)_{i,j} = \sum_{l=1}^{N} \frac{1}{p(\mathbf{x}_l; t, \boldsymbol{\theta})} \mathbf{s}_l^i(t)\mathbf{s}_l^j(t). \tag{7.6}$$

The FIM for a sequence of measurements taken independently (i.e. for smFISH data) at times $\mathbf{t} = [t_1, t_2, \ldots, t_{N_t}]$ is then

$$\mathcal{I}(\boldsymbol{\theta}, \mathbf{t}, \mathbf{c}) = \sum_{k=1}^{N_t} c_k \mathbf{F}(\boldsymbol{\theta}, t = t_k). \tag{7.7}$$

where $\mathbf{c} = [c_1, c_2, \ldots, c_{N_t}]$ is the number of cells measured at each of the measurement times. For smFISH experiments, the vector $\mathbf{c}$ plays an important role in the design of the study, as it corresponds to the time points to be measured (i.e. which times are optimal to fix the cells), and how many cells to count the RNA in at those selected times. The next section aims to find the optimal $\mathbf{c}$ for STL1 mRNA in yeast cells.

7.3 Results

7.3.1 Verification of the FSP-FIM for time-varying stochastic gene expression

Our work in Chapter 6 was limited to models of stochastic gene expression that had constant $\mathbf{A}$. Here, we extend this to time-varying $\mathbf{A}$ to the adaptive stress response gene STL1 in yeast with a time varying signal given in Eq. 7.1. For this analysis, we considered a subset of all model parameters, shown in Fig. 7.1(a). Model parameters simultaneously fit to experimentally measured 0.2M and 0.4M STL1 mRNA from [20] were used as a reference set of parameters (yellow dots in Fig. 7.2), $\boldsymbol{\theta}^*$. These reference parameters were used to generate 50 unique simulated data sets. The parameters that maximized the likelihood for each simulated data set were then found as a set of $\hat{\boldsymbol{\theta}}$, shown as gray dots in Fig. 7.2. Using the asymptotic normality of the maximum likelihood estimator and its relationship to the FIM (Eq. 7.4), we then compared the 95% CIs of inverse of the Fisher information to those of the MLE estimates, shown by the blue and green ellipses in Fig. 7.2(a). We also compared the eigenvalues of the inverse of the Fisher information to the eigenvalues of the covariance matrix of MLE estimates in Fig. 7.2(b). With the FSP-FIM verified

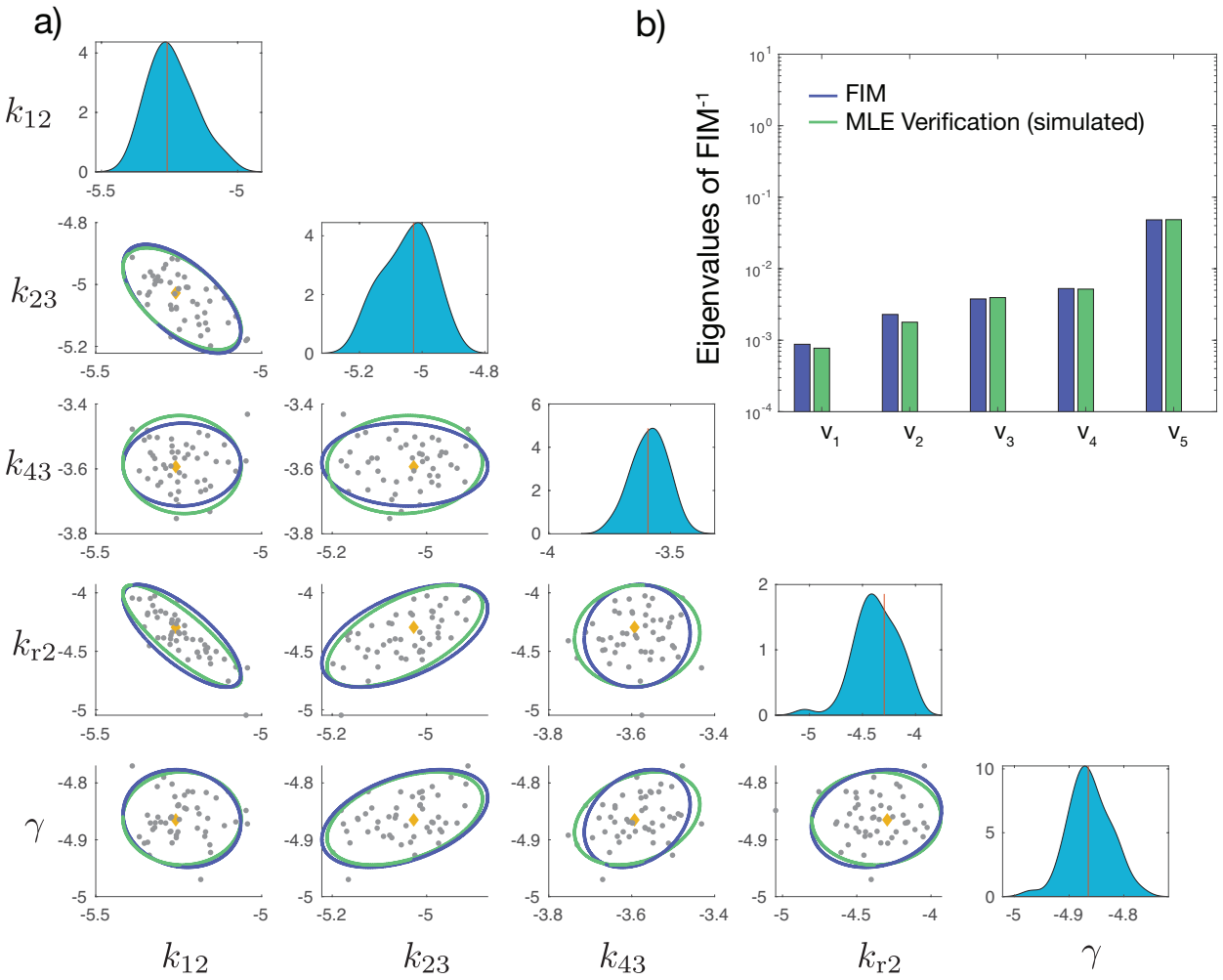

Figure 7.2: *Verification of the FSP-FIM for the time-varying HOG-MAPK model.*(a) Scatter plots and density plots of the spread of MLE estimates for 50 simulated data sets for a subset of model parameters. The ellipses show the 95% CI for the inverse of teh FIM (blue) and covariance of scatter plot (green). The yellow dot indicates the parameters at which the FIM and simulated data sets were generated. (b) Ranked eigenvlaues for the covariance of MLE estimates (green) and inverse of the FIM (blue).

for the Hog-MAPK model, we next explore how the FIM can be used to optimally allocate the number of cells to measure at different times after osmotic shock.

7.3.2 Designing optimal measurements for the HOG-MAPK pathway to design smFISH experiments in *S. cerevisae*

To test the FSP-FIM in the realistic context of time-varying gene expression, we consider a simulated course of smFISH data for osmotic stress response of yeast in yeast. We start with a single experimental replicate of smFISH data at 0.2 M NaCl concentration, with a known set of underlying model parameters, which were taken from simultaneous fits to 0.2M and 0.4M data in [20]. Parameters for the single simulated data set were found by maximizing the likelihood 3.1 using iterative genetic algorithms and simplex-based searches [20]. These baseline parameters were then used to optimize the allocation of measurements at different time points $t = [1, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$ minutes after NaCl induction. To show the practical application of these approaches, we first designed experiments to maximize the information about a subset of the model parameters, sometimes referred to as D_s-optimality. This metric corresponds to maximizing the product of the eigenvalues of the FIM.

The number of cells to be measured at a discrete set of time points for the system can be optimized using a greedy approach, in which measurements are added one at a time according to the time point that increases the metric of interest the most. Mathematically, our goal is to find

$$\max_{\mathbf{c}} |\mathcal{I}(\mathbf{c};\boldsymbol{\theta})|_{D_s} \text{ such that } \sum_{i=1}^{N_t} \mathbf{c}_i = 1 \tag{7.8}$$

where $\mathbf{c}$ is a vector of length N_t where each entry corresponds to the fraction of total measurements to be allocated at $t = \mathbf{t}_i$, and $|\mathcal{I}(\mathbf{c};\boldsymbol{\theta})|_{D_s}$ refers to the product of the eigenvalues FIM. To illustrate this approach, we first allocated cell measurements according to D_s-optimality. The fraction of cells that is optimal for a 0.2M NaCl input compared to the experimentally measured number of cells is shown in Fig. 7.3. While each available time point was allocated a non-zero fraction of

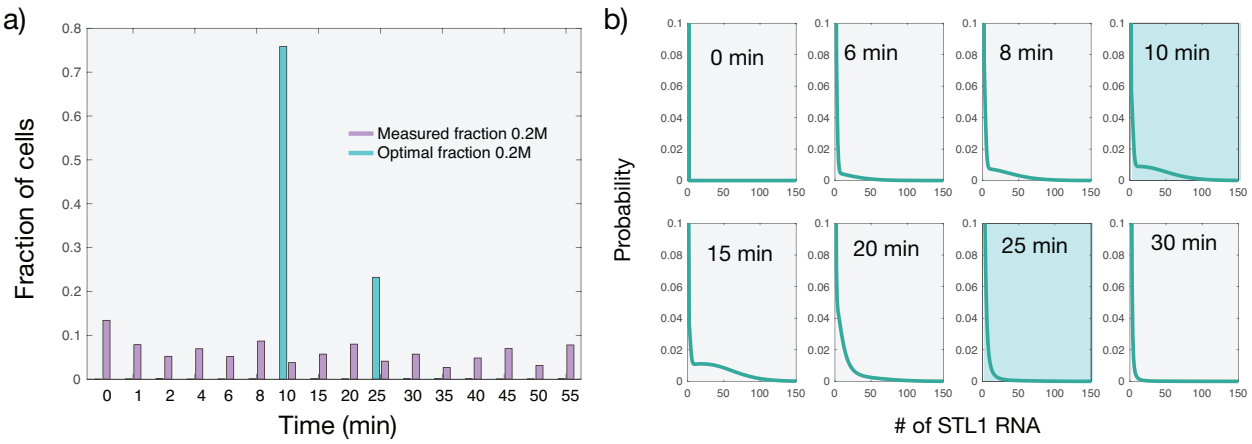

Figure 7.3: *Optimizing the allocation of cell measurements at different time points.* (a) Comparison of optimal fractions cells to measure (blue) at different time points compared to experimentally measured numbers of cells (red). (b) Model fits (blue) to experimental data (black) at a subset of time points. The blue boxes denote the time points of optimal measurements.

measurements, two time points at $t = 10$ min and $t = 25$ min were vastly more informative than the other available time points. To verify this result, we simulated 50 data sets of 1,000 cells each and found the MLE estimates for each sub-sampled data set. We compared the spread of these MLE estimates to the inverse of the optimized FIM, shown in Fig. 7.2. The increase in information of the optimal 0.2M experiment is compared to the baseline, 'intuitive' experiment is shown in Fig. 7.4(a). The optimal experiment only requires measurement of 2 time points compared to the full experiment, in which 16 time points were measured. We next compare the intuitive experiment design to a random experiment design, in which measurements are randomly distributed among different time points, and compare the D_s-optimality. Fig. 7.4(b) shows that the intuitive experiment is more informative than a random experiment, but is still significantly less informative than the optimal experiment.

7.3.3 Designing optimal biosensor experiments

Thus far, we have considered a set of experiments to find the optimal experiment regarding the information about model parameters. We next use the model to design an optimal series of smFISH measurements to sense the cellular environment. In the HOG-MAPK transcription model, we model the way that extracellular osmolarity ultimately affects stress response gene transcription

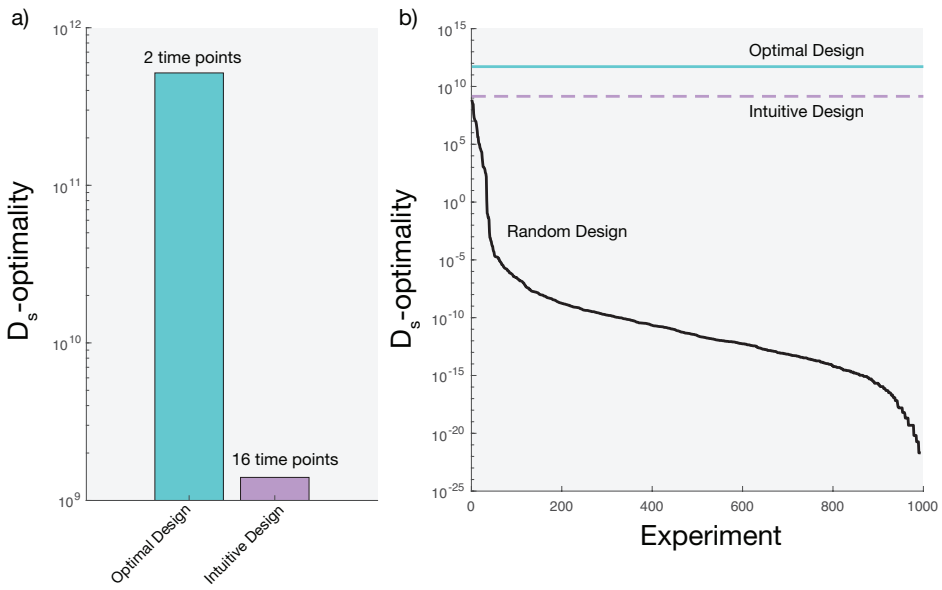

Figure 7.4: *Information gained by performing optimal experiments compared to actual experiments* (a) D_s-optimality for the optimal design using only two time points compared to the measured number of cells using all 16 time points. (b) Comparing the information of the optimal experiment design (blue), intuitive experiment design (purple), and random experiment designs (black).

levels through the time-varying parameter $k_{21}(t)$ in Eq. 7.1. Figure 7.1b shows the effect 0.2M and 0.4M salt concentrations on k_{21} activation. Higher salt concentrations delay the time at which $k_{21}(t)$ becomes nonzero. Using this fact, we approximate the function $k_{21}(t)$ as the sum of three Heaviside step functions,

$$k_{21}(t) = u(t) - u(t - t_1) + u(t - t_2), \tag{7.9}$$

where t_1 is a fixed delay of the time it takes for nuclear kinase levels to reach a particular threshold, and t_2 is the time they drop below that threshold. Our goal in this section is to find an experiment which reduces the uncertainty in t_2 for a range of values of t_2, shown in Fig. 7.5(a). We are assuming that t_2 is related to the salt concentration experienced by the cell, as shown in Fig. 7.1b and 7.5(b) in which 0.2M salt inputs have a lower t_2 than 0.4M salt inputs. To estimate the uncertainty in t_2 given our model, we find the sensitivity of the distribution of mRNA abundance to the variable t_2. This FSP-sensitivity requires finding $\mathbf{s}_{t_2} = \frac{\partial \mathbf{A}}{\partial t_2}$. As $k_{21}(t)$ is the only part of $\mathbf{A}$

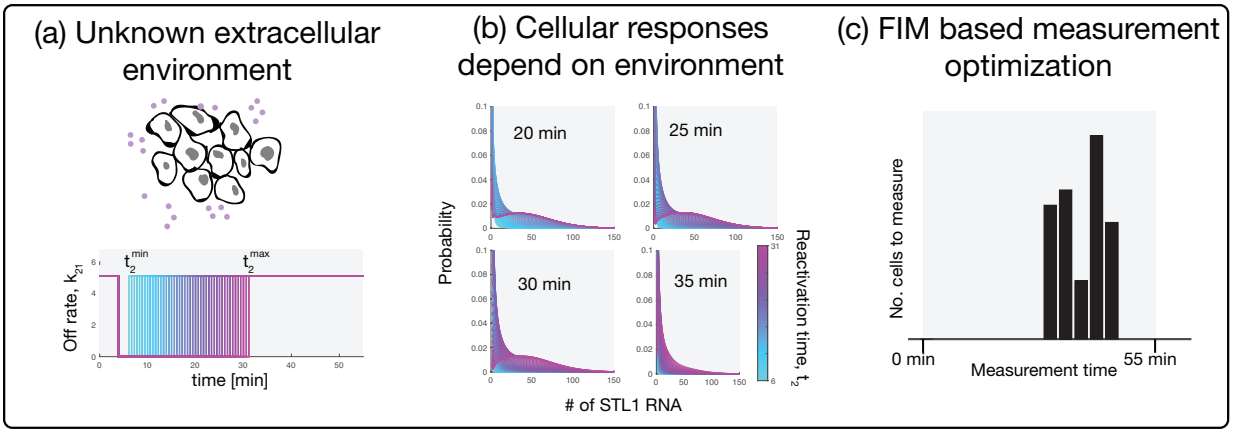

Figure 7.5: *Overview of optimal design for biosensing experiments in the osmotic stress response in yeast.* (a) Unknown salt concentrations in the environment give rise to different reactivation times, t_2. These different reactivation times cause downstream STL1 expression dynamics to behave differently as shown in panel (b). These different responses can be used to resolve experiments that reduce the uncertainty in t_2.

that depends explicitly on t_2, we only need

$$\frac{\partial k_{21}(t)}{\partial t_2} = \delta(t_2), \tag{7.10}$$

and therefore $\mathbf{s}_{t_2} = \frac{\partial \mathbf{A}}{\partial t_2}$ is only non-zero at $t = t_2$. The time evolution sensitivity of the system to t_2 is then

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{t_2}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{t_2}\end{bmatrix} \quad \text{with} \quad \mathbf{s}(0) = \mathbf{s}_{t_2}\mathbf{p}(t_2), \tag{7.11}$$

and time is integrated from t_2 to the final time of interest. This suggests that the time evolution of the sensitivities only really depends on probability vector $p(\mathbf{x}, t = t_2)$ and the generator matrix $\mathbf{A}$, i.e. $\frac{d\mathbf{s}_{t_2}}{dt} = \mathbf{A}\mathbf{s}_{t_2}$. If the Fisher information at each measurement time is written into a vector $\mathbf{f} = [f_1, f_2, \ldots, f_{N_t}]$ and the number of measurements is the vector of equal measurements, $\mathbf{c} = [c_1, c_2, \ldots, c_{N_t}]$, the information for a given t_2 value is the sum-product of these two vectors,

$$\mathcal{I}(t_2) = \sum_{k=1}^{N_t} \mathbf{c}_k \mathbf{f}_k = \mathbf{c}^T \mathbf{f}. \tag{7.12}$$

Because our goal is to find an experiment that is optimal given a range of possible t_2's (which we have assumed is linearly related to the salt concentration in the environment), the time until reactivation is treated as a uniform random variable over the range of reasonable activation times, $\mathcal{T} \sim \text{unif}(t_2^{\min}, t_2^{\max})$, where $t_2^{\min}$ and $t_2^{\max}$ correspond to the minimum and maximum t_2 values we consider. To find the experiment that reduces our uncertainty in t_2, we integrate the FIM in Eq. 7.12 over the uncertainty in $\mathcal{T}$,

$$\mathbf{c}_{\text{opt}} = \min_{\mathbf{c}} \int_{t_2^{\min}}^{t_2^{\max}} p(t) \mathcal{I}^{-1}(\mathbf{c}; t_2 = t, \boldsymbol{\theta}) dt \tag{7.13}$$

$$= \min_{\mathbf{c}} \int_{t_2^{\min}}^{t_2^{\max}} \mathcal{I}^{-1}(\mathbf{c}; t_2 = t, \boldsymbol{\theta}) dt, \tag{7.14}$$

because we have assumed that $p(t)$ is uniform. The objective function of our minimization is the integral

$$J = \int_{t_2^{\min}}^{t_2^{\max}} \mathcal{I}^{-1}(\mathbf{c}; t_2 = t, \boldsymbol{\theta}) dt, \tag{7.15}$$

which corresponds to the uncertainty about the value of t_2 for a given $\mathbf{c}$. We then used the same greedy approach described in Section 7.3.2 to find the optimal $\mathbf{c}$. To verify this approach, we sampled a random value of t_2, which we call t_2^{true}. For this value of t_2, we then simulate 100 random data sets of according to each of the five experiment designs in Fig. 7.6. For each of the random data sets, we asked which FSP solution for the range of t_2 values was most likely using Eq. 3.1, which we call $\hat{t}_2^*$, and made a table $[t_2^{true} \;\; \hat{t}_2^*]$. The error in the estimation is then the MSE of the columns of this table, as shown in purple in Fig. 7.6. The optimal design and the simplified design perform much better than a uniform design or random experiment designs. The simplified design refers to a design which uses the same time points as the optimal design, but with equal numbers of measurements at each time.

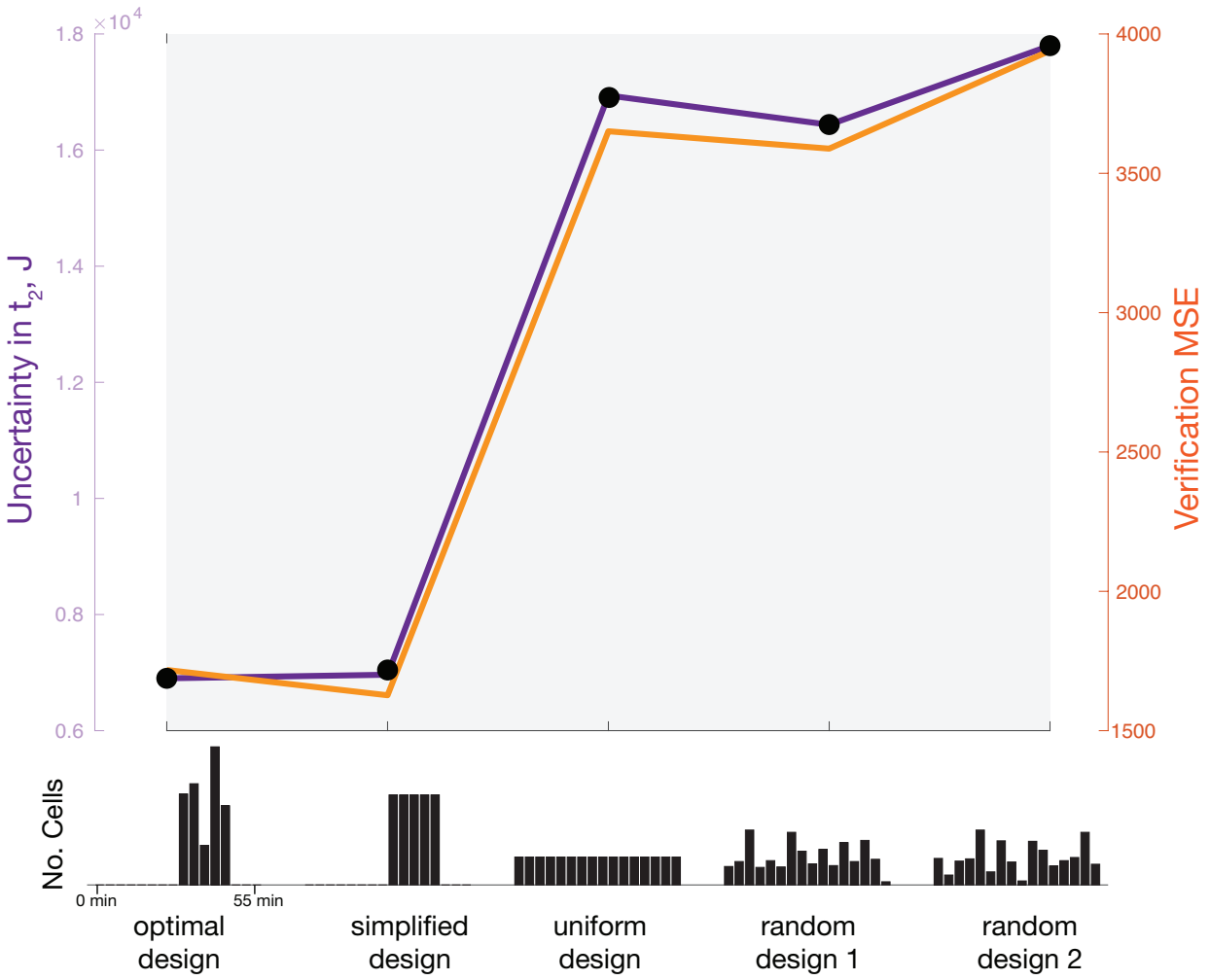

Figure 7.6: *Verification of the uncertainty in t_2 for different experiment designs.* The top panel shows the value of J in Eq. 7.15 for different experiment designs (bottom panel) in purple, and the MSE values for verification are shown in orange.

7.4 Discussion

The methods developed in this work presents a principled, model-driven approach to allocating single-cell measurements in a time-varying stochastic system. We demonstrate these theories on a well-established model the osmotic stress response in yeast cells, and particularly for the STL1 gene, which is activated upon the nuclear localization of phosporylated MAPK [11, 20]. For this system, we showed how to optimally allocate the number of cells measured at different times to maximize the information about a subset of model parameters. We then compared these optimal designs to the actual experiments performed and randomly generated experiments shown in Fig. 7.4b. We found that while the experiments performed were much better than you would expect by random chance, they still had lower Fisher information than the optimal experiment. Similarly, the optimal experiment design consisted of measuring only two time points in the process many times, compared to a more intuitive design of relatively uniform measurements, as shown in Fig. 7.4a. This suggests that the optimal design not only found an experiment that increases the Fisher information, but also is experimentally 'cheaper' than an intuitive design.

We then used Fisher information to design experiments to learn about the cellular environments. Using the same osmotic shock response model, we found the optimal experiment to reduce the uncertainty about the adaptation time, given a range of possible adaptation times. We then compared this experiment design to other experiment designs, including an intuitive design, where all time points are treated equally, and two other random experiments in Fig. 7.6. This method of using design experiments to use cells as biosensors in stress environments could be extremely useful in a biomedical contexts, where the time and amount of sample can be hugely important. This work also provides another example of model-driven experiment design.

Chapter 8

Using Fluctuations to Expand the Color Palette of Single-Molecule Microscopy [4]

8.1 Introduction

Recent technology developments allow the quantification of single-RNA and proteins in live cells. The MS2 system [126–129] encodes stem loop structures into a gene of interest, which are subsequently bound by fluorescently tagged MS2 coat proteins upon transcription of RNA. This approach allows one to see single RNAs as they are transcribed and transported within the cell [130].

In the same spirit, recent works have developed the technology to visualize single polysomes [23–25]. For this method, the gene is modified to produce proteins that bind antibody-like probes. This technology has been combined with the MS2 system to visualize the entire central dogma of molecular biology at single-molecule resolution. However, these new technologies have created a need for more advanced computational approaches to help design future studies. For example, where in the gene of interest should one add MS2 sequence or FLAG sequence to answer a particular biological question?

One limitation of antibody-like based live-cell protein measurements is the relatively small numbers of colors available to make the measurements [23–25], which fundamentally limits the number of genes that can be measured in a single cell, and therefore limits the scope of questions that can be addressed with this technique. However, different genes have different sequences, codon dependencies and lengths, all of which give rise to different fluctuations in the single-polysome intensity traces measured with the nascent polypeptide chain tracking described above.

[4] This chapter first summarizes the modeling approach developed by a collaboration of the Munsky Group (computational) and Stasevich Group (experimental), primarily led by Luis Aguilera on the computational side. This chapter extends that analysis to multiplex the number of genes that can be measured in single cells.

The characteristics of these fluctuations, such as their autocorrelation times, mean intensisty levels, and variance intensity levels can be used to discriminate between proteins without using different fluorophore. These statistics can be used for multiplexing, i.e. discriminating between multiple genes without using different colors of molecules. Such multiplexing could expand the number of proteins that can be imaged in different cells. In this work, we develop a stochastic model of translation, describe the statistics of this process, and develop a novel computational pipeline that accurately classifies experimental trajectories that have been trained on simulated trajectories.

8.1.1 Stochastic Model of Translation Dynamics

Single-molecule translation is stochastic process in which ribosomes bind with messenger mRNA, and polymerize polypeptides one amino acid at a time. Each amino acid addition, or elongation event, of the protein can be modeled as a stochastic event which occurs with probability $w_i(\mathbf{x}_i)dt$ in the infinitesimal time interval dt. Ribosomes bind the mRNA with probability $k_{init}dt$, and dissociate with probabilty $w_t x_n dt$. Elongation of multiple nascent proteins along a single mRNA molecule may be modeled as a discrete, stochastic process in which ribosomes bind, elongate polypeptide chains, and terminate. By enumerating the position of each codon along the mRNA x_i, the process can be written as a series of chemical reactions

$$\varnothing \xrightarrow{k_{init}(1-x_1)} x_1 \xrightarrow{w_i(x_2,\ldots,x_{nf+1})} x_2 \xrightarrow{w_{i+1}(x_{i+1},\ldots,x_{nf+1})} x_i \ldots \xrightarrow{w_t} \varnothing. \tag{8.1}$$

This general formulation of the model allows for several important biological features of elongation, such as the effects of “ribosome exclusion", in which a ribosome may not advance to the $i + 1^{th}$ position if there is another ribosome in the $i + n_f$ codons in front of it. This causes the propensities of each step to be non-linear functions,

$$w_0 = k_i \prod_{j=1}^{nf}(1 - x_j), \tag{8.2}$$

$$w_i = k_e(i) \cdot x_i \prod_{j=1}^{nf}(1 - x_{i+j}), \text{ for } i = 1, ..., N-1;. \tag{8.3}$$

Furthermore, one can incorporate codon-specific elongation rates for each step of the process, where the propensity coefficient $k_e(i)$ depends on the relative abundance of that particular codon in the human genome,

$$k_e(i) = \bar{k_e} \cdot (u(i)/\bar{u}). \tag{8.4}$$

Finally, the termination of elongation can be found using a single rate,

$$w_t = k_t \cdot x_N. \tag{8.5}$$

This set of biochemical reactions can be used to track the positions of individual ribosomes as they move along the RNA, creating the polypeptide. At a given time t, this approach describes the binary occupancy of each codon position along the mRNA. The vector $\mathbf{x}(t)$ is a vector of 1's and 0's of length N. However, single polysome measurements do not resolve single ribosomes, so to compare the vector $\mathbf{x}(t)$ is not an observable quantity. Instead, diffraction limited fluorescent spots are quantified. To compare simulations to such single molecule data, the vector of ribosome positions $\mathbf{x}(t)$ needs to be mapped to fluoresence intensities $I(t)$. Every time a ribosome passes the epitope region of a gene, FLAG-tags bind to amino acids, and increases the fluorescence intensity of the polysome. The locations of epitope regions can then be used to map ribosome positions to fluoresence intensities using a probe design vector $\mathbf{c}$ of length N:

$$I(t) = \sum_{i}^{N} c_i x_i = \mathbf{c}^T \mathbf{x}. \tag{8.6}$$

The vector $\mathbf{c}$ is the cumulative sum of a vector of length N with 1's in the epitope regions and 0's elsewhere. This formulation of a mechanistic stochastic model of translation can be simulated using the stochastic simulation algorithm using the software rSNAPSIM [131], to obtain the intensities $I(t)$ over time. Intensities for two genes, KDM5B and H2B are shown in Fig. 8.2, along with experimentally measured intensity trajectories, distributions,and autocorrelations measured for the two genes.

8.2 Autocorrelation of translation dynamics

When ribosome loading is sparse, higher-order interaction of ribosomes is rare, and the nonlinearities in Eq. 8.2 have a lesser effect on the dynamics. Under such circumstances, it is possible to derive a simplified linear system model for the elongation dynamics, which is nonphysical in the sense that a ribosome could pass another ribosome while elongating. In the linear model, the propensity of an elongation step is $w_i(x_i) = k_i x_i$, and the ability of a ribosome to add another amino acid only depends on the current position of the ribosome, and not on the footprint of other ribosomes.

For this simplified process, we can define a stoichiometry matrix that describes the change in $\mathbf{x}$ for every reaction,

$$\mathbf{S}_{i,j} = \begin{cases} 1 \text{ for all } i = j \\ -1 \text{ for all } \ i = j - 1, \end{cases} \tag{8.7}$$

where i corresponds to each codon in the protein of interest. Each row of the stoichiometry matrix corresponds to an elongation event of an individual ribosome from the i^{th} to the $i + 1^{\text{th}}$ codon. The propensities of each reaction can be written in the affine linear form

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{W}_1 \mathbf{x}, \tag{8.8}$$

where $\mathbf{w}_0$ is a column vector of zeros with the first entry k_i and $\mathbf{W}_1$ is a matrix

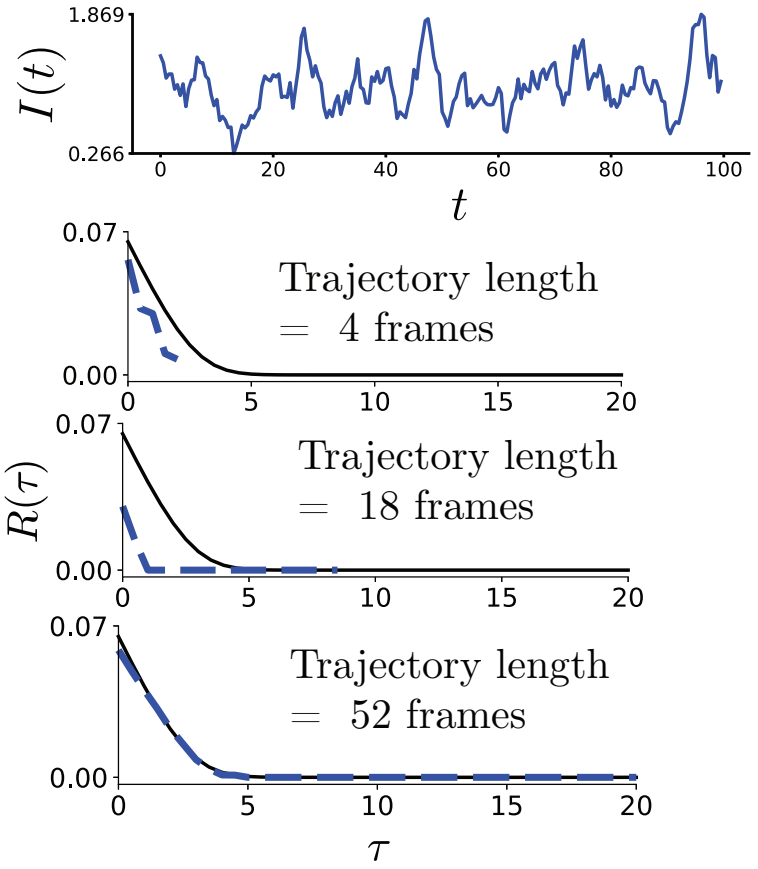

Figure 8.1: *The effect of different trajectory lengths on autocorrelation.* Autocorrelations are shown a single stochastic trajectory (top) with varying measurement lengths, shown in blue. The black line shows the autocorrelation derived from the model above.

$$\mathbf{W}_{1i,j} = \begin{cases} -k_e(i) \text{ for all } i = j \\ \\ k_e(i) \text{ for all } \ i = j - 1. \end{cases} \tag{8.9}$$

Under these assumptions, the first two moments of the intensity $I(t)$ can be found:

$$\mathbb{E}\{I(t)\} = \mathbb{E}\{\mathbf{cx}\} = \mathbf{c}\mathbb{E}\{\mathbf{x}\} \tag{8.10}$$

$$\mathbb{E}\{I(t)^2\} = \mathbb{E}\{\mathbf{cxx}^{\mathrm{T}}\mathbf{c}^{\mathrm{T}}\} = \mathbf{c}\mathbb{E}\{\mathbf{xx}^{\mathrm{T}}\}\mathbf{c}^{\mathrm{T}}. \tag{8.11}$$

The autocorrelation dynamics of the process are defined in terms of the intensity $I(t)$ as in Eq. 8.6 and can be decomposed in terms of the ribosome position vector $\mathbf{x}$ as

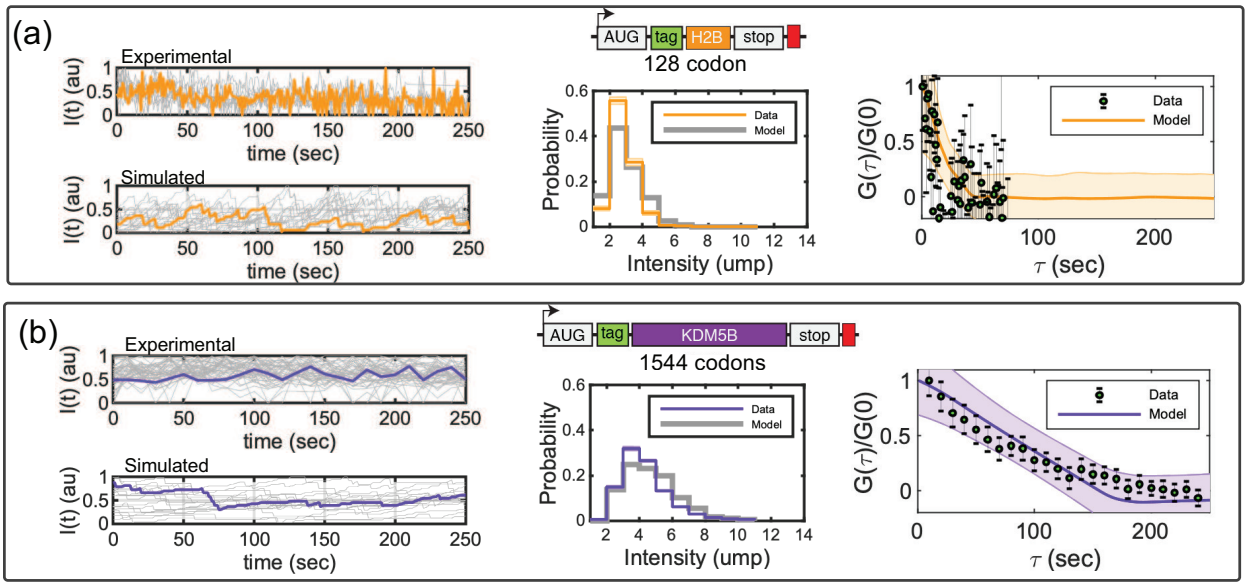

Figure 8.2: *Comparing experimental and simulated statistics of single-model translation.* Experimental and simulated trajectories (left) intensity distributions (center) and autocorrelations (right) for two genes H2B and KDM5B. Adapted from our work in [131].

$$G(\tau) = \mathbb{E}\{\mathbf{c}^{\mathrm{T}}\mathbf{x}(t)\mathbf{x}(t+\tau)^{\mathrm{T}}\mathbf{c}\} \tag{8.12}$$

$$= \mathbf{c}\mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t+\tau)^{\mathrm{T}}\}\mathbf{c}^{\mathrm{T}}. \tag{8.13}$$

Noting that $\mathbf{c}$ is a constant with respect to τ, it is only necessary to find the auto- and cross-correlations of the ribosome positions. Following the regression theorem [32], these correlations are given by the solution to the set of ODEs,

$$\frac{d\Sigma(\tau)}{d\tau} = \phi\Sigma(\tau) \tag{8.14}$$

given the initial condition is the steady-state covariance of the process, i.e.

$$\Sigma(0) = \lim_{t\to\infty} \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^{\mathrm{T}}\} \tag{8.15}$$

and the autonomous matrix of the process $\phi = \mathbf{SW}$. Because the system is linear, the steady-state covariance Σ_0 is given by the solution to the Lyapunov equation

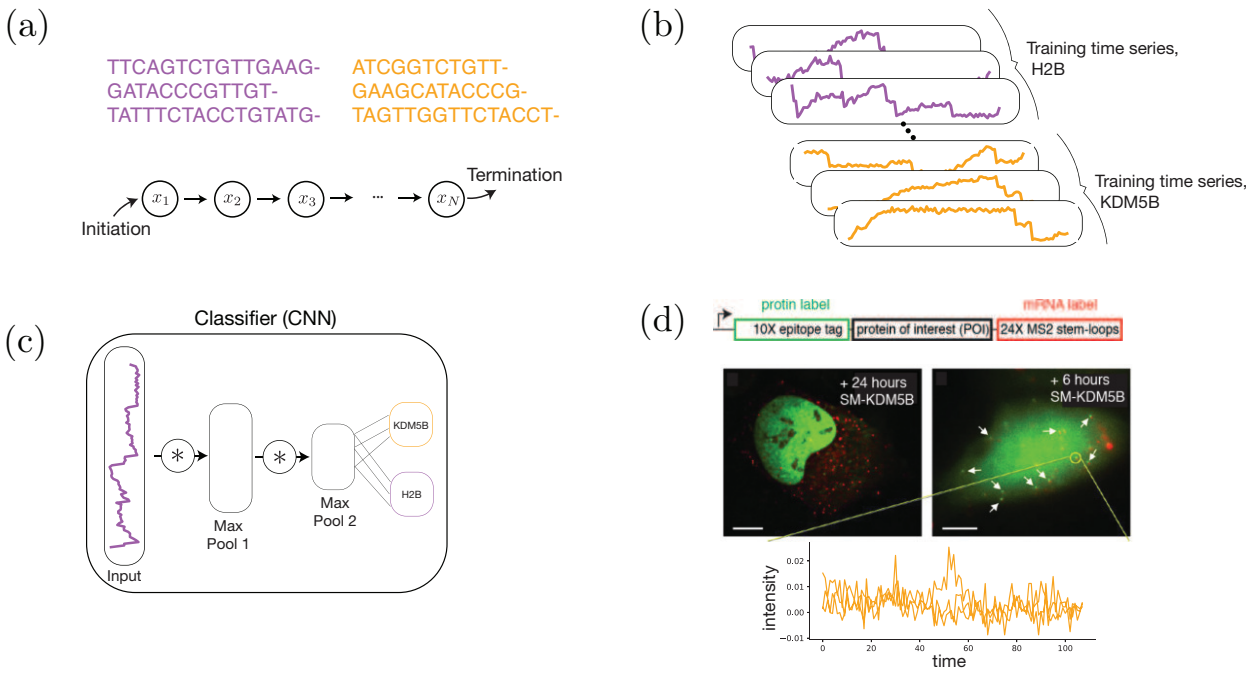

Figure 8.3: *Outline of CNN based approach to classify polysomes.* (a) Given two different gene sequences, a stochastic model of protein elongation can simulate intensity trajectories. These intensity trajectories can be used to generate training data (b) that can then classified using a convolutional neural network (CNN), shown in (c). After training on purely simulated data, the CNN can be used to classify intensity trajectories measured in murine cells.

$$\mathbf{S}\mathbf{W}_1\Sigma(0) + \Sigma(0)\mathbf{W}_1^T\mathbf{S}^T + \mathbf{S}\mathrm{diag}(\mathbf{W}_1\mathbb{E}\{\mathbf{x}\} + \mathbf{w}_0)\mathbf{S}^T = 0. \tag{8.16}$$

Integrating Eq. 8.14, the autocorrelation of the intensity $R(\tau)$ can be found using Eq. 8.12. In practice, autocorrelations are difficult to measure from a single fluorescent signal [23, 127, 131], due to finite signal length, photobleaching effects, and measurement noise. The effect of finite measurement time on autocorrelation is shown for an arbitrary gene in Fig. 8.1. While autocorrelation dynamics are not always accurate for a single trajectory, they can be averaged across ensembles of traejctories to give insight about the time scales of the fluctuations in intensity signal [23]. Because the aim of this study is identify single-trajectories withing cells, we now turn to other computational methods to discriminate between multiple genes, which use the biophysical model described above to generate data for a statistical model that can accurately classify single trajectories.

8.3 Convolutional neural networks to multiplex single-molecule translation

The role of neural networks in modern machine learning approaches has vastly increased in recent years. In particular, convolutional neural networks (CNN) [132, 133] have become extremely popular for image recognition, but also image generation [134]. CNN's have been applied to classify histological samples [135–137], segment cells for microscopy data [138]. The disadvantage of neural network based algorithms is the ambiguity that one faces in trying to interpret the resulting weights and biases from the network, though recent work has started to develop some insight about what different layers of the network are doing [139]. In general, CNN's can be thought of as optimal filters for classifying the image of interest. This property makes them ideal for time-series classification of a stochastic process, as they are able to find the relevant frequency relationships within the data, compared to a recurrent neural network or LSTM, which essentially is a set number of neural networks that "unfold" in time [140]. Because of the explosion of methods in the field of machine learning, there are many types of algorithms and sub-algorithms that can be used for any problem, and there seems to be some art in deciding which method is best for which problem. In our case, CNN's are likely to be successful for classifying trajectories, though other approaches, such as hidden Markov Models, standard neural networks applied to the frequency decompositions of the data, or other approaches may also be successful.

There are some aspects of our problem which make it interesting from a computational/theoretical standpoint beyond applying a black-box method to some data and obtaining classifications of different trajectories. Our goal is to train and validate the model on simulated data exclusively, and then see how well the same neural network is able to classify experimentally measured trajectories. This approach requires the model to be extremely representative of the data, which we show in [131], and importantly that there is way to find all parameters of the model prior to collecting data. Ultimately, this will require estimating the average elongation rate, initiation rate, and termination rate for a gene based on it's sequence alone. The advantages of such a model are vast, as

one can then use it to design experiments to optimally distinguish between multiple genes in single cells.

We start by training a CNN on simulated trajectories for the KDM5B and H2B genes, shown in Fig. 8.3(b). The neural network architecture was extremely simple, consisting of two convolutional layers, two max pool (averaging) layers, and a single dense layer for classification. The network was implemented with the Keras [141] package for TensorFlow [142]. Validation for these two genes is shown in Fig. 8.4(a) and (b), and 100% accuracy is achieved. As a proof of concept, we then take this network and ask how well it is able to classify experimentally measured trajectories of H2B (N=10) and KDM5B (N=18), shown in Fig. 8.4(b) and (c). While all trajectories are correctly classified for KDM5B, only 60% were correctly classified for H2B. While the goal of this approach is to tell apart trajectories in the same cell and the same color, we do not currently have experimental data with two genes in the same cell with different color tags, and therefore cannot validate the classifications from the same cell. Despite the experimental challenges associated with measuring gene, the results in Fig. 8.4 are sufficent to motivate further exploration of this approach as data quality improves. Having validated the network, we next ask how different experimental parameters can affect classification results.

8.3.1 Experiment design using convolutional neural networks

One major question for fluorescence microscopy is how to choose the frame rate of the camera to measure trajectories. For many live cell experiments that rely on fluorophores, the number of photons released by a given particle decreases as they are exposed to light, and therefore the signal decreases. This decrease in signal creates noisier measurement. Therefore, there is an inherent tradeoff between measurement time and experimental noise. Often times, it is of interest to numerically remove this effect by fitting and normalizing fluoresence intensity measurements to an exponential curve [23, 143].

To simulate the effect of variability in the number of photons that are detected by the camera given a single (or multiple) fluorophores, we add white noise (poisson limit of large numbers of

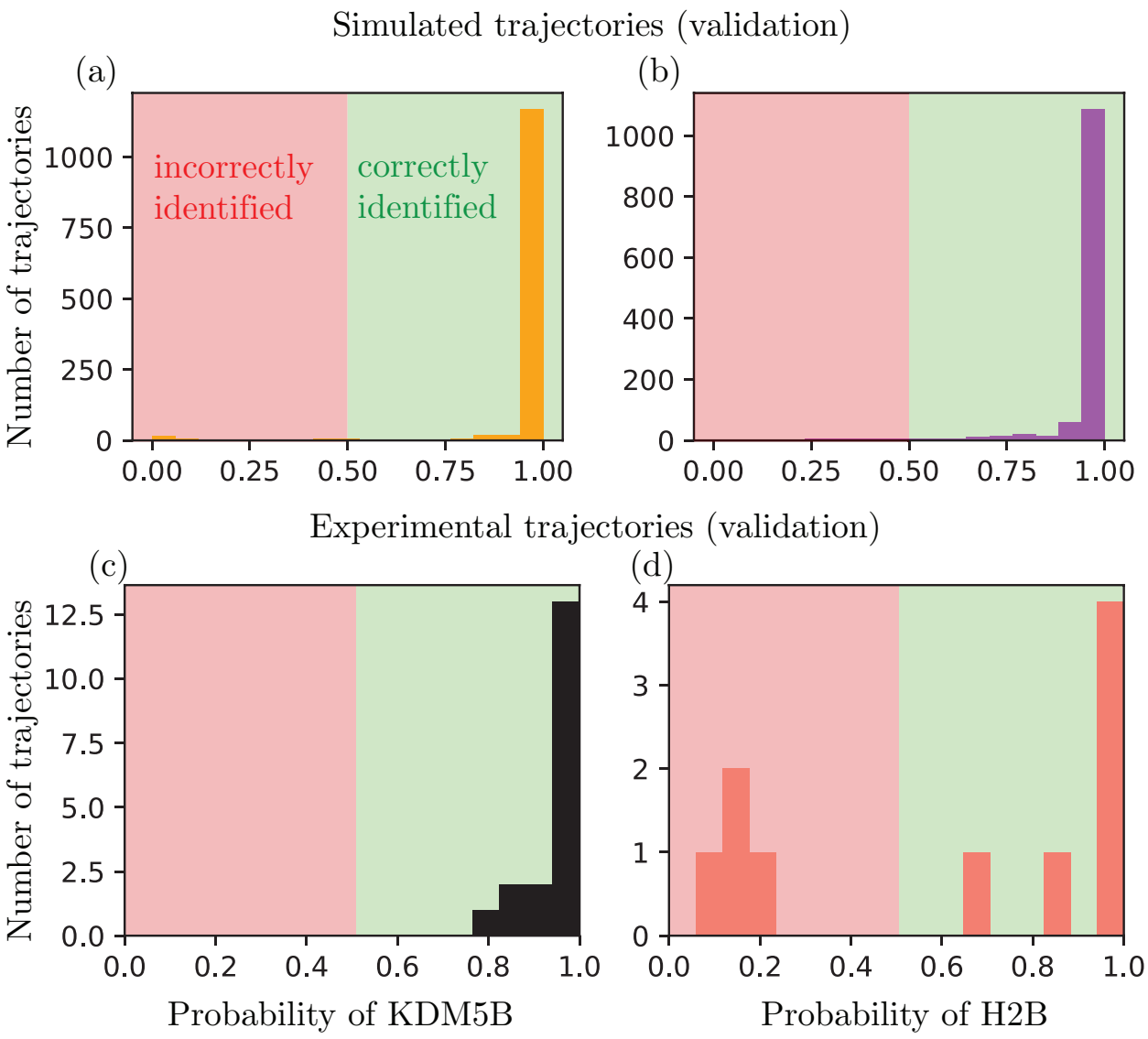

Figure 8.4: (a-b) Classification results for simulated trajectories held out for validation, where the model was trained to discriminate between KDM5B (left) and H2B (right). Panels (c-d) show the results for experimentally measured trajectories of KDM5B and H2B.

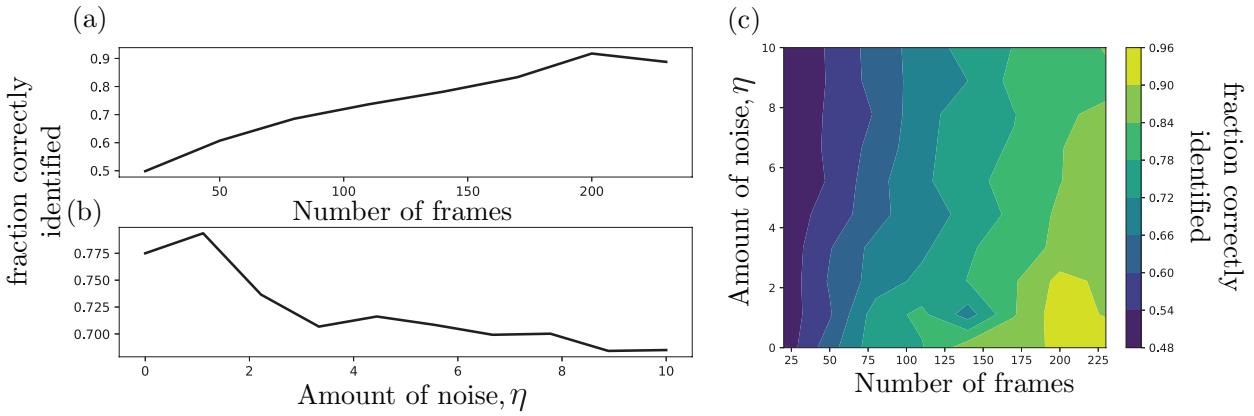

Figure 8.5: *Experimental considerations for multiplexing single-molecule translation.* Correct classification percentage as a function of trajectory length (a) for H2B and KDM5B genes with 2.2X noise, noise (b) with a fixed trajectory length 110 frames, and with both noise and number of frames being varied (c).

photons) to the simulated trajectories,

$$\tilde{I}(t) = I(t) + \eta, \tag{8.17}$$

where η is a normally distributed random variable that does not have any temporal correlation. As the noise level increases, the ability to correctly identify the trajectories decreases, shown in Fig. 8.5(b). Figure 8.5(a) shows the improvement of identification of single trajectories as a function of the trajectory length. In a more realistic scenario, as trajectory length increases, the noise also increases, as one must use lower laser power to avoid photobleaching effects. However, as this tradeoff is yet to be rigorously experimentally quantified, we compute identification accuracy accross a range of measurement times and experimental errors, shown in Fig. 8.5. The tradeoff between trajectory length and measurement noise is a (unknown) curve on that contour plot.

8.4 Discussion

Single-molecule imaging of the entire central dogma of molecular biology has only recently been possible with the advent of single-polysome translation imaging methods [23–25]. Here, we demonstrate a stochastic model of the translation process, and drawing on our work [131], show that it can match the statistics of measured polysome intensity trajectories. We showed that this

model can be used to generate autocorrelations under a linear model assumption, which in principle is different for genes of different lengths. We then ask how this model can be used to multiplex single-molecule translation experiments, in which there are a finite number of colors that can be used to image within single cells, which limits the number of genes that can be measured. While autocorrelations can be rapidly generated from the approximate model (Eq. 8.14), effects of noise and finite measurement times make them difficult to compare to experimental data, shown in Fig. 8.1. Because of the challenges associated with autocorrelation based identification of single-cell trajectories, we turn to a hybrid method that uses the stochastic model we define in Eq. 8.1 to generate training data that accurately represents what one can expect to see experimentally [131], but can be generated in essentially unlimited quantity, as compared to the relatively difficult to obtain experimental trajectories. The other advantage of an approach based on simulated training data is that one can change the experimental settings, such as noise and measurement time to find conditions which are optimal for discriminating between multiple genes in the same cell Fig. 8.5. However, the classification of simulated trajectories is only meaningful if the same neural network model can be used to classify experimental trajectories. In Fig. 8.4, we showed that the trained network can distinguish between experimentally measured KDM5B and H2B trajectories that came from different cells with moderate accuracy.

As our ability to include more accurate measurement and biological details into the stochastic model improves, the model that is trained on the simulated data should more accurately reflect classifications that we can expect from experimental trajectories. In addition to improving the biophysical stochastic model, in the future we will include more experiment design features, using multicolor probe designs, such as those used by Lyon et al [26]. Multiple probe colors and their positions within the protein of interest will affect the fluctuation characteristics and could lead to coordinated designs to measure 10's of genes in the same cells. Furthermore, the neural networks to classify the genes can be added to image acquisition software, leading to real-time classification of genes and experiment designs. Paired with optogenetic technology [90, 144], these methods

could be used analyze and control gene expression in live cells with feedback control to drive cells towards particular cell fates and coordinate with other cells.

This work demonstrates a combination of mechanistic modeling and machine learning. The mechanistic model provides ample stochastic trajectories to train the convolutional network, which would otherwise require a large number of measured trajectories to train. This idea may be useful for other data types, where data is limited and a mechanistic model can be readily defined, but there is no clear way to apply the mechanstic model to the data.

Chapter 9

Conclusions and Future Work

This dissertation has developed a new set of computational tools to better understand and make use of computational models of gene expression. The advances presented here are meant to make biological modeling, even for systems with noisy processes like bursting transcription/translation, single-molecule translation, more integrated with experimental data. In general, a main goal of computational and systems biology is to develop useful models that accurately descibe experimental observations by logically gathering known information about the system that is being studied, and using it to predict how this system will behave in different environments. However, predictive modeling has thus far had mild success, which is often attributed the extreme complexity of biological systems. To address the complexity in biology, the modeller is ofter tempted to add enough details to capture all the known biological information, which requires a huge number of kinetic parameters. These parameters are almost impossible to infer from data because they are poorly constrained by the relatively low dimensional quantitative data that is available, and the models are often analyzed at the level of average expression of the relevant biomolecules in the system. Another reason that less detailed models of gene expression have had limited successes may be that the analysis approach is very important, and that using average behavior is not a good proxy for the underlying behavior of the system. This dissertation develops new methods for analyzing relatively small stochastic models of gene expression with a high level of precision, to rigorously co-design modern single-cell, single-molecule experiments and biological models of their underlying processes.

The FSP bounds on the likelihood of single cell data in Chapter 4 can be used to speed up the identification of stochastic models of gene expression by using the data to inform the accuracy of the model itself. While the bounds themselves are novel, the idea of using data not only as a quantity to fit the model, but also to constrain its computational cost is powerful idea that can be applied in other ways. For example, Chapter 5 uses the data to define a lower dimensional

basis on which we projected the FSP dynamics. This project uses a small number of radial basis functions to interpolate the entire, high-dimensional state space. However, a rigorous error model using radial basis functions was not able to be developed. Furthermore, it would be interesting to investage, for both the FSP bounds and the project-based model reduction, to investigate how constraints that come from multiple types of data can be imposed on FSP solutions. For example, one could imagine using bulk assays to learn the mean of the process, and use that information to help define basis centers or constrain the FSP bounds. Furthermore, the FSP bounds have yet to be applied to a Bayesian inference scheme, in which one could rapidly find posterior parameter distributions.

This work also takes some steps forward in using identified predictive, stochastic models and using them to design experiments that are as informative as possible. The FSP-based Fisher information approach presented in Chapter 6 has been used to optimize measurement times in a simulated model of bursting gene expression and a simulated toggle model. We also use the FIM to design optimal optogenetic-controlled degradation in the simulated toggle system. Finally, we studied the allocation of measurements at different times in an experimental yeast system, and validating the FSP-FIM for time-varying inuts. In the future, the FSP-FIM could be used in the setting of fluorescence-activated cell sorting to define a population of cells with stochastic gene expression dynamics that are optimal to learn about a particular feature. In a similar vein, the FSP-FIM can be used to develop optimal image analysis by only spending computational power to count individual RNA in cells that are likely to be informative. Finally, often times we are not concerned about the uncertainty in the parameters of the model, but rather in the uncertainty in the predictions the model will make. In that vein, we are developing a prediction-uncertainty reduction method that makes use of the FIM.

For models of single molecule translation in Chapter 8, we developed novel methods to multiplex the measurements of multiple genes in the same cell. By using a mechanistic model to simulate the gene expression of the two genes, H2B and KDM5B, we generated intensity trajectories that have realistic fluctuation characteristics compared to the single-polysome translation

measurements [23–25]. We then showed how a simplified model that does not allow for ribosomal exclusion can be used to find the autocorrelation function for different genes. Finally, we showed that the stochastic mechanistic model can be used to generate training data for a machine learning algorithm, which was then able to classify experimentally measured trajectories. This novel approach could be used to expand the number of genes that are able to be measured in a single cell from one or two genes to ten to twenty genes. By measuring more genes in single cells, it is possible to understand more about the dynamic regulation of genes than is otherwise possible.

This dissertation is primarily concerned with developing novel computational and theoretical methods to model and analyze modern single cell data. Often, computational methods and models are built in a vacuum, isolated from data and the messy world of biological measurement. Our approach is to build these tools so that they can be applied with experimentalists in mind. A major part of bridging computational methods to use for actual experiments requires easy-to-use softwares to develop models and design experiments. Along these lines, a major step forward (and one that is currently in progress), is the creation of software that can define a stochastic model, input single-cell measurements, identify model parameters, and design future experiments. Such software should have an attractive graphical user interface. For the single molecule translation problem, our group has started developing the rSNAPSIM package, which allows one to simulate and analyze intensity trajectories for any gene of interest. We have also started to develop a software called the Stochastic System Identification Toolkit, which provides graphical model construction and basic model fitting approaches. Eventually, the FSP-bounds, projection based model reduction, and FSP-based FIM will all be used to enhance this software.

In conclusion, the works here present first steps in filling out the toolkit for analysis of full probability distributions of biomolecules across populations of cells. As quantitive methods to measure single molecules in single cells has improved, there is a need to develop better methods to analyze and interpret data. Quantitative modeling and prediction of how biological systems behave will revolutionize medicine and agriculture, especially as our ability to manipulate and design DNA improves.

Bibliography

[1] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[2] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, July 2011.

[3] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, October 2008.

[4] Gábor Balázsi, Alexander van Oudenaarden, and James J Collins. Cellular Decision Making and Biological Noise: From Microbes to Mammals. *Cell*, 144(6):910–925, March 2011.

[5] Avigdor Eldar and Michael B Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, September 2010.

[6] Alexander Harms, Etienne Maisonneuve, and Kenn Gerdes. Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science*, 354(6318), 2016.

[7] Richard Losick and Claude Desplan. Stochasticity and cell fate. *Science (New York, N.Y.)*, 320(5872):65–68, April 2008.

[8] Edo Kussell, Roy Kishony, Nathalie Q Balaban, and Stanislas Leibler. Bacterial persistence: a model of survival in changing environments. *Genetics*, 169(4):1807–1814, April 2005.

[9] Arjun Raj, Patrick van den Bogaard, Scott A Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, October 2008.

[10] A M Femino, F S Fay, K Fogarty, and R H Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, April 1998.

[11] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, February 2013.

[12] Abhyudai Singh and João P Hespanha. Approximate moment dynamics for chemically reacting systems. *Automatic Control, IEEE Transactions on*, 56(2):414–418, 2011.

[13] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–8345, 2012.

[14] Jakob Ruess, Andreas Milias-Argeitis, and John Lygeros. Designing experiments to understand the variability in biochemical reaction networks. *Journal of The Royal Society Interface*, 10(88):20130588, 2013.

[15] Gabriele Lillacci and Mustafa Khammash. Parameter estimation and model selection in computational biology. *PLoS Comput Biol*, 6(3):e1000696, 2010.

[16] Mariana Gomez-Schiavon, Liang-Fu Chen, Anne E West, and Nicolas E Buchler. BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. *Genome biology*, 18(1):164, September 2017.

[17] Valerii Fedorov, Yuehui Wu, and Rongmei Zhang. Optimal dose-finding designs with correlated continuous and discrete responses. *Statistics in medicine*, 31(3):217–234, February 2012.

[18] M B Elowitz and S Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, January 2000.

[19] Tasuku Kitada, Breanna DiAndreth, Brian Teague, and Ron Weiss. Programming gene and engineered-cell therapies with synthetic biology. *Science (New York, N.Y.)*, 359(6376), February 2018.

[20] Brian Munsky, Guoliang Li, Zachary R Fox, Douglas P Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, June 2018.

[21] Gutenkunst, R, Waterfall J, Casey, F, Brown, K, Myers, C, and Sethna, J. Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology*, 3(10):1871–1878, October 2007.

[22] Zachary Fox, Gregor Neuert, and Brian Munsky. Finite state projection based bounds to compare chemical master equation models using single-cell data. *Journal of Chemical Physics*, 145, 2016.

[23] Morisaki, Tatsuya, Lyon, Kenneth, DeLuca, Keith F, DeLuca, Jennifer G, English, Brian P, Zhang, Zhengjian, Lavis, Luke D, Grimm, Jonathan B, Viswanathan, Sarada, Looger, Loren L, Lionnet, Timothee, and Stasevich, Timothy J. Real-time quantification of single RNA translation dynamics in living cells. *Science (New York, N.Y.)*, 352(6292):1425–1429, June 2016.

[24] Xiaowei Yan, Tim A Hoek, Ronald D Vale, and Marvin E Tanenbaum. Dynamics of Translation of Single mRNA Molecules In Vivo. *Cell*, 165(4):976–989, May 2016.

[25] Wu, B, Eliscovich, C, Yoon, Y J, and Singer, R H. Translation dynamics of single mRNAs in live cells and neurons. *Science (New York, N.Y.)*, 352(6292):1430–1435, June 2016.

[26] Kenneth Lyon, Luis U. Aguilera, Tatsuya Morisaki, Brian Munsky, and Timothy J. Stasevich. Live-cell single rna imaging reveals bursts of translational frameshifting. *bioRxiv*, 2018.

[27] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, January 2006.

[28] Donald A McQuarrie. Stochastic Approach to Chemical Kinetics. *Journal of Applied Probability*, 4(3):413, December 1967.

[29] N G Van Kampen and Nicolaas Godfried. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992.

[30] Brian Munsky. In Michael E Wall, editor, *Quantitative biology: from molecular to cellular systems*, chapter 11. CRC Press, 2012.

[31] B Munsky and M Khammash. Transient analysis of stochastic switches and trajectories with applications to gene regulatory networks. *IET Systems Biology*, 2(5):323–333, September 2008.

[32] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004.

[33] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977.

[34] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, July 2001.

[35] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Avoiding negative populations in explicit Poisson tau-leaping. *The Journal of chemical physics*, 123(5):054104, August 2005.

[36] Christian A Yates and Kevin Burrage. Look before you leap: a confidence-based method for selecting species criticality while avoiding negative populations in τ-leaping. *The Journal of chemical physics*, 134(8):084109, February 2011.

[37] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, 124(4):044109, January 2006.

[38] Michał Komorowski, Maria J Costa, David A Rand, and Michael P H Stumpf. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8645–8650, May 2011.

[39] Jakob Ruess, Francesca Parise, Andreas Milias-Argeitis, Mustafa Khammash, and John Lygeros. Iterative experiment design guides the characterization of a light-inducible gene expression circuit. *Proceedings of the National Academy of Sciences of the United States of America*, 112(26):8148–8153, June 2015.

[40] Zechner, Christoph, Unger, Michael, Pelet, Serge, Peter, Matthias, and Koeppl, Heinz. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nature methods*, 11(2):197–202, February 2014.

[41] Christoph Zimmer. Experimental design for stochastic models of nonlinear signaling pathways using an interval-wise linear noise approximation and state estimation. *PloS one*, 11(9):e0159902, September 2016.

[42] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology*, 5(318):318, 2009.

[43] Brian Munsky, Zachary Fox, and Gregor Neuert. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods*, 85:12–21, 2015.

[44] Robert G Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[45] Wei Yu, Wonjong Rhee, Stephen Boyd, and John M Cioffi. Iterative water-filling for gaussian vector multiple-access channels. *IEEE Transactions on Information Theory*, 50(1):145–152, 2004.

[46] Verena Wolf, Rushil Goel, Maria Mateescu, and Thomas A Henzinger. Solving the chemical master equation using sliding windows. *BMC systems biology*, 4(1):42, 2010.

[47] Youfang Cao and Jie Liang. Optimal enumeration of state space of finitely buffered stochastic molecular networks and exact computation of steady state landscape probability. *BMC Systems Biology*, 2(1):1, March 2008.

[48] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a Genetic Toggle Switch in Escherichia coli. *Nature*, 403(6767):339–342, January 2000.

[49] Tianhai Tian and Kevin Burrage. Stochastic models for regulatory networks of the genetic toggle switch. *Proceedings of the National Academy of Sciences*, 103(22):8372–8377, 2006.

[50] C Moler and C Van Loan. Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, 45(1):3–49, February 2003.

[51] Serge Pelet, Fabian Rudolf, Mariona Nadal-Ribelles, Eulàlia de Nadal, Francesc Posas, and Matthias Peter. Transient activation of the HOG MAPK pathway regulates bimodal gene expression. *Science (New York, N.Y.)*, 332(6030):732–735, May 2011.

[52] Desmond J. Higham. Modeling and Simulating Chemical Reactions. *SIAM Rev.*, 50(2):347–368, Jan 2008.

[53] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[54] Brian Munsky, Brooke Trinh, and Mustafa Khammash. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, 5(318):318, 2009.

[55] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander Van Oudenaarden. Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science*, 339(6119):584–587, 2013.

[56] Douglas P. Shepherd, Nan Li, Sofiya N. Micheva-Viteva, Brian Munsky, Elizabeth Hong-Geller, and James H. Werner. Counting small RNA in pathogenic bacteria. *Anal. Chem.*, 85:4938–4943, 2013.

[57] Donald A Mcquarrie and Daniel T Gillespie. Stochastic Theory and Simulations of Chemical Kinetics. *J. Appl. Probab.*, 478(1967):413–478, 1967.

[58] D.T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1-3):404–425, September 1992.

[59] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.*, 54(1):1–26, Jan 2007.

[60] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

[61] Y. Cao, D.T. Gillespie, and L.R. Petzold. Adaptive explicit-implicit tau-leaping method with automatic tau selection. *J. Chem. Phys*, 126(22):224101, 2007.

[62] Daniel T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297, 2000.

[63] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, Jan 2006.

[64] Brian Munsky and Mustafa Khammash. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *J. Comput. Phys.*, 226(1):818–835, Sep 2007.

[65] K. Burrage, M. Hegland, S. MacNamara, and R.B. Sidje. A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems. In A.N. Langville and W.J. Stewart, editors, 150^{th} *Markov Anniversary Meeting, Charleston, SC, USA*, pages 21–38. Boson Books, 2006.

[66] V. Kazeev, M. Khammash, M. Nip, and C. Schwab. Direct solution of the chemical master equation using quantized tensor trains. *PLoS Comp. Bio.*, 10(3), 2014.

[67] R.B. Sidje and H.D. Vo. Solving the chemical master equation by a fast adaptive finite state projection based on the stochastic simulation algorithm. *Math. Biosci.*, 269:10–16, 2015.

[68] B. Munsky and M. Khammash. The finite state projection approach for the analysis of stochastic noise in gene networks. *IEEE Transactions on Automatic Control*, 53(Special Issue):201–214, Jan 2008.

[69] T. Jahnke and T. Udrescu. Solving chemical master equations by adaptive wavelet compression. *J. Comp. Phys.*, 229(16):5724–5741, 2010.

[70] Jose Juan Tapia, James R. Faeder, and Brian Munsky. Adaptive coarse-graining for transient and quasi-equilibrium analyses of stochastic gene regulation. *2012 IEEE 51st IEEE Conf. Decis. Control*, 836:5361–5366, 2012.

[71] Gregory E. Fasshauer. *Meshfree Approximation Methods with MATLAB*. World Scientific, 2007.

[72] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[73] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, Mar 1991.

[74] Youngmin Cho and Lawrence K Saul. Kernel Methods for Deep Learning. *NIPS Conf.*, 9:342–350, 2009.

[75] E. J. Kansa. Multiquadrics-A scattered data approximation scheme with applications to computational fluid-dynamics-II solutions to parabolic, hyperbolic and elliptic partial differential equations. *Comput. Math. with Appl.*, 19(8-9):147–161, 1990.

[76] Jingwei Zhang, Layne T. Watson, Christopher A. Beattie, and Yang Cao. Radial basis function collocation for the chemical master equation. *Int. J. Comput. Met.*, 07(03):477–498, 2010.

[77] Ivan Kryven, Susanna Röblitz, and Christof Schütte. Solution of the chemical master equation by radial basis functions approximation with interface tracking. *BMC Syst. Biol.*, 9(1):67, 2015.

[78] Tobin A. Driscoll and Alfa R.H. Heryudono. Adaptive residual subsampling methods for radial basis function interpolation and collocation problems. *Comput. Math. with Appl.*, 53(6):927–939, 2007.

[79] B. Munsky, Z. Fox, and G. Neuert. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods*, 85:12–21, 2015.

[80] MATLAB. *version 8.1.0.604 (R2013a)*. The MathWorks Inc., Natick, Massachusetts, 2013.

[81] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[82] Munsky, B and Khammash, M. Transient analysis of stochastic switches and trajectories with applications to gene regulatory networks. *IET systems biology*, 2(5):323–333, September 2008.

[83] Golan Bel, Brian Munsky, and Ilya Nemenman. The simplicity of completion time distributions for common complex biochemical processes. *Physical biology*, 7(1):016003, March 2010.

[84] Martin Dietrich Buhmann. Radial basis functions. *Acta Numerica 2000*, 9:1–38, 2000.

[85] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[86] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[87] Zenklusen, D, Larson, D R, and Singer, R H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12):1263–1271, December 2008.

[88] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, December 2005.

[89] Leah M Octavio, Kamil Gedeon, and Narendra Maheshri. Epigenetic and conventional regulation is distributed among activators of FLO11 allowing tuning of population-level heterogeneity in its expression. *PLoS genetics*, 5(10):e1000673, October 2009.

[90] Armin Baumschlager, Stephanie K Aoki, and Mustafa Khammash. Dynamic Blue Light-Inducible T7 RNA Polymerases (Opto-T7RNAPs) for Precise Spatiotemporal Gene Expression Control. *ACS synthetic biology*, 6(11):2157–2167, November 2017.

[91] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[92] G. Casella and R. L. Berger. *Statistical inference*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1990.

[93] Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *The FEBS Journal*, 276(4):923–942, February 2009.

[94] Bernhard Steiert, Andreas Raue, Jens Timmer, and Clemens Kreutz. Experimental Design for Parameter Estimation of Gene Regulatory Networks. *PloS one*, 7(7):e40052, July 2012.

[95] Michele Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Physical Review D*, 77(4), 2008.

[96] Rod Frehlich. Cramer-Rao bound for Gaussian random processes and applications to radar processing of atmospheric signals. *IEEE Transactions on Geosciences and Remote Sensing*, 31(6):1123–1131, Nov 1993.

[97] Yoav Shechtman, Steffen J Sahl, Adam S Backer, and W E Moerner. Optimal point spread function design for 3D imaging. *Physical review letters*, 113(13):133902, September 2014.

[98] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science (New York, N.Y.)*, 336(6078):183–187, April 2012.

[99] Rudiyanto Gunawan, Yang Cao, Linda Petzold, and Francis J Doyle. Sensitivity analysis of discrete stochastic systems. *Biophysical journal*, 88(4):2530–2540, April 2005.

[100] Vicente Costanza and John H Seinfeld. Stochastic sensitivity analysis in chemical kinetics. *The Journal of chemical physics*, 74(7):3852–3858, April 1981.

[101] Joshua F Apgar, David K Witmer, Forest M White, and Bruce Tidor. Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular bioSystems*, 6(10):1890–1900, October 2010.

[102] Bandara, Samuel, Schlöder, Johannes P, Eils, Roland, Bock, Hans Georg, and Meyer, Tobias. Optimal Experimental Design for Parameter Estimation of a Cell Signaling Model. *PLoS computational biology*, 5(11):e1000558, November 2009.

[103] F P Casey, D Baird, Q Feng, R N Gutenkunst, J J Waterfall, C R Myers, K S Brown, R A Cerione, and J P Sethna. Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET systems biology*, 1(3):190–202, May 2007.

[104] J Peccoud and B Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[105] T B Kepler and T C Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–3136, December 2001.

[106] J M Raser. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004.

[107] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS biology*, 4(10):e309, October 2006.

[108] V Shahrezaei and P S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[109] Srividya Iyer-Biswas, F Hayot, and C Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79(3):2323, March 2009.

[110] Ido Golding. Deciphering the stochastic kinetics of gene regulation. *Biophysical journal*, 112(3):342a, February 2017.

[111] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-Fos transcriptional bursts. *Cell reports*, 8(1):75–83, July 2014.

[112] Douglas P Shepherd, Nan Li, Sofiya N Micheva-Viteva, Brian Munsky, Elizabeth Hong-Geller, and James H Werner. Counting small RNA in pathogenic bacteria. *Analytical chemistry*, 85(10):4938–4943, May 2013.

[113] Yulei Wang, Chih Long Liu, John D. Storey, Robert J. Tibshirani, Daniel Herschlag, and Patrick O. Brown. Precision and functional specificity in mrna decay. *Proceedings of the National Academy of Sciences*, 99(9):5860–5865, 2002.

[114] Hideki Kobayashi, Mads Kærn, Michihiro Araki, Kristy Chung, Timothy S Gardner, Charles R Cantor, and James J Collins. Programmable cells: interfacing natural and en-

gineered gene networks. *Proceedings of the National Academy of Sciences*, 101(22):8414–8419, June 2004.

[115] J Vanlier, C A Tiemann, P A J Hilbers, and N A W van Riel. A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, April 2012.

[116] Chunbo Lou, Brynne Stanton, Ying-Ja Chen, Brian Munsky, and Christopher A Voigt. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology*, 30(11):1137–1142, November 2012.

[117] Heng Xu, Samuel O Skinner, Anna Marie Sokac, and Ido Golding. Stochastic kinetics of nascent RNA. *Physical review letters*, 117(12), September 2016.

[118] Leonardo A Sepúlveda, Heng Xu, Jing Zhang, Mengyu Wang, and Ido Golding. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*, 351(6278):1218–1222, March 2016.

[119] Marc Rullan, Dirk Benzinger, Gregor W Schmidt, Andreas Milias-Argeitis, and Mustafa Khammash. An optogenetic platform for real-time, single-cell interrogation of stochastic transcriptional regulation. *Molecular cell*, 70(4):745–756, May 2018.

[120] J Stewart-Ornstein, S Chen, J Bhatnagar, JS Weissman, and H El-Samad. Model-guided optogenetic study of PKA signaling in budding yeast. *Molecular Biology of the cell*, 28(1), 2017.

[121] Peles, S, Munsky, B, and Khammash, M. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *The Journal of chemical physics*, 125(20):204104, November 2006.

[122] Brian Munsky and Mustafa Khammash. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *Journal of Computational Physics*, 226(1):818–835, September 2007.

[123] B Munsky, J J Tapia, and J Faeder. Adaptive coarse-graining for transient and quasi-equilibrium analyses of stochastic gene regulation. *51st IEEE Conference on Decision and Control (CDC)*, 2012.

[124] H D Vo, Z R Fox, A Baetica, B Munsky bioRxiv, and 2018. Bayesian estimation for stochastic gene expression using multifidelity models. *biorxiv*, 2018.

[125] Zachary R Fox and Brian Munsky. The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments. *PLoS computational biology*, 15(1):e1006365, January 2019.

[126] E Bertrand, P Chartrand, M Schaefer, S M Shenoy, R H Singer, and R M Long. Localization of ASH1 mRNA particles in living yeast. *Molecular cell*, 2(4):437–445, October 1998.

[127] Daniel R Larson, Daniel Zenklusen, Bin Wu, Jeffrey A Chao, and Robert H Singer. Real-Time Observation of Transcription Initiation and Elongation on an Endogenous Yeast Gene. *Science (New York, N.Y.)*, 332(6028):475–478, April 2011.

[128] Hocine, Sami, Raymond, Pascal, Zenklusen, Daniel, Chao, Jeffrey A, and Singer, Robert H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature methods*, 10(2):119–121, February 2013.

[129] Antoine Coulon, Matthew L Ferguson, Valeria de Turris, Murali Palangat, Carson C Chow, and Daniel R Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife*, 3:e1002215, October 2014.

[130] Amir Mor, Shimrit Suliman, Rakefet Ben-Yishay, Sharon Yunger, Yehuda Brody, and Yaron Shav-Tal. Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nature cell biology*, 12(6):543–552, June 2010.

[131] Luis Aguilera, Will Raymond, Zachary Fox, Michael May, Elliot Djokic, Tatsuya Morisaki, Timothy J. Stasevich, and Brian. Munsky. Computational design and interpretation of single-RNA translation experiments. *In prep*, 2019.

[132] A Sharif Razavian, H Azizpour, and J Sullivan. CNN features off-the-shelf: an astounding baseline for recognition. *Proceeding of the Computer Vision Foundation*.

[133] A Krizhevsky and Hinton Sutskever, I and. Imagenet classification with deep convolutional neural networks. *Advances in neural networks (NIPS)*, 2012.

[134] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[135] Arkadiusz Gertych, Zaneta Swiderska-Chadaj, Zhaoxuan Ma, Nathan Ing, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E Walts, and Beatrice S Knudsen. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific reports*, 9(1):1483, February 2019.

[136] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 16(Pt 2):411–418, 2013.

[137] Noorul Wahab, Asifullah Khan, and Yeon Soo Lee. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. *Computers in biology and medicine*, 85:86–97, June 2017.

[138] Christopher D Malon and Eric Cosatto. Classification of mitotic figures with convolutional neural networks and seeded blob features. *Journal of pathology informatics*, 4(1):9, 2013.

[139] G Hinton, O Vinyals, and J Dean. 2015.

[140] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, November 1997.

[141] François Chollet et al. Keras. https://keras.io, 2015.

[142] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[143] Nathalie B Vicente, Javier E Diaz Zamboni, Javier F Adur, Enrique V Paravani, and Víctor H Casco. Photobleaching correction in fluorescence microscopy images. *Journal of Physics: Conference Series*, 90:012068, nov 2007.

[144] Remy Chait, Jakob Ruess, Tobias Bergmiller, Gašper Tkačik, and Călin C Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature communications*, 8(1):2557, November 2017.