DISSERTATION

EXPLOITING NOISE, NON-LINEARITY, AND FEEDBACK TO DIFFERENTIALLY CONTROL MULTIPLE DIFFERENT CELLS USING A SINGLE OPTOGENETIC INPUT

Submitted by

Michael P May

School of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor: Brian Munsky
Co-Advisor: Tim Stasevich

Diego Krapf
Patrick Shipman

ABSTRACT

EXPLOITING NOISE, NON-LINEARITY, AND FEEDBACK TO DIFFERENTIALLY CONTROL MULTIPLE DIFFERENT CELLS USING A SINGLE OPTOGENETIC INPUT

Motivated by Maxwells-Demon, we propose and solve a cellular control problem in which the exploitation of stochastic noise can break symmetry between two cells and allow for specific control of multiple cells using a single input signal. We find that a new type of noise-exploiting controllers are effective and can remain effective despite coarse approximations to the model's scale or extrinsic noise in key model parameters, and that these controllers can retain performance under substantial observer-actuator time delays. We also demonstrate how SIMO controllers could drive two-cell systems to follow different trajectories with different phases and frequencies by using a noise-exploiting controller. Together, these findings suggest that noise-exploiting control should be possible even in the case where models are approximate, and where parameters are uncertain. Having demonstrated the potential of noise-enhanced feedback control through computational modeling, we have also begun the next steps toward automating microscopy to implement this potential in experimental practice. Specifically, we demonstrate a new integrated pipeline to automate the image collection including: (i) quickly search in two-dimensions to find fields of view with cells of desired phenotypes, (ii) targeted collection of three-dimensional image data for these chosen fields of view, and (iii) streamlined processing of the collected images for rapid segmentation, spot detection and tracking, and cell/spot phenotype quantification.

DEDICATION

I would like to thank my wife, my father and my excellent advisors Brian Munsky and Tim Stasevich for all their support throughout my PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.0.1 Synthetic Biology

Synthetic biology is an interdisciplinary field covering biology, engineering, and control theory that has enhanced the ability to engineer and manipulate living organisms. By designing and constructing biological components and modules, changes to cellular behavior can be in incorporated in a logical manner similar to electronic circuits [1, 2]. Synthetic biology has been used in a wide array of applications, from developing biofuels [3], sustainable agriculture [2], and calculators with a display [4].

Orthogonal control in synthetic biology seeks to ensure independence for different components or modules within a biological system [5]. Orthogonality enables precise manipulation and programming of these components to perform specific functions without unintended cross-talk or disruption, enabling the construction of intricate and reliable biological circuits and systems that can be programmed using a literal programming language [6].

Modularity in synthetic biology breaks down complex biological processes into discrete, interchangeable modules or genetic components, each with a specific function. These modules can be standardized and easily combined to create novel biological systems or circuits, much like assembling building blocks. The modular approach offers several advantages, including flexibility in design, scalability, and the ability to rapidly prototype and modify biological functions [7].

1.0.2 Control Theory

Control theory is a multidisciplinary field that focuses on the design and analysis of systems to regulate and manipulate their behavior. At its core, control theory aims to develop algorithms, controllers, and strategies that ensure a system's output closely follows desired references, despite disturbances and uncertainties. It encompasses a wide range of applications [8], from engineering [9] to robotics [10] and biology [11, 12]. It relies on mathematical models, feedback loops, and

principles of optimization to achieve stability, robustness, and precision in controlling dynamic systems, making it a useful tool in scientific domains.

Enhancing control theory to account for noise and uncertainty is crucial in these systems. Noise and parameter uncertainties, inherent in real-world systems, can significantly affect the performance and reliability of control systems. To address this challenge, control strategies have been developed that explicitly model and mitigate the impact of measurement noise and parametric uncertainties. Techniques such as Kalman filtering [13], stochastic optimal control [14], and robust control [15] can be effective tools in such systems.

The management of noise in control theory initially found its solution in the application of robust control methods [15–17]. As control systems became more prevalent in various engineering and scientific domains, it became apparent that external disturbances, uncertainties in system parameters, and sensor measurement noise could severely impact system performance. In response to these challenges, robust control methods were developed to enhance the resilience and stability of control systems in the presence of such uncertainties. These methods encompassed a range of techniques, including H-infinity control, or μ-synthesis, which aimed to design controllers capable of maintaining desired system behavior even under the most adverse conditions. The advent of robust control enabled the development of systems capable of withstanding the inherent imperfections and fluctuations encountered in real-world applications.

Control systems adjust the control inputs depending on the state of the system (i.e., feedback), to manipulate the system's output and achieve the desired control performance. For single-input-single-output (SISO) systems a control input may have a simple effect, but for a multi-input-multi-ouput (MIMO) system, the control inputs may have more broad effects on the system. For a Linear-Time-Invariant system (LTI) these control actions acting on a system can be represented as a two linear matrices that can be optimized to produce the desired control performance. Non-Linear systems can be controlled by linearizing them down into Piecewise linear systems which can change over time. Stochastic non-linear systems can be linearized using the Finite State projection (FSP) [18, 19].

Many synthetic biological systems are controlled using recurring network topologies called motifs [20]. These motifs describe the elements within synthetic biology building blocks, consisting of specific arrangements of genes, regulatory elements, and feedback loops that precisely control cellular functions. High-level descriptions of these motifs, such as feedback and feedforward (coherent, incoherent, and adaptive) provide a system for categorizing and understanding their functional roles. New motifs are being discovered and proposed which can provide new and interesting behaviors [21, 22].

1.0.3 Optogenetics

Optogenetics is a field that combines genetics and optics to precisely control cellular activity in living organisms. It employs genetically engineered light-sensitive proteins, like T7 polymerase [23], which allow researchers to manipulate their activity by illuminating them with specific light wavelengths [24]. This has enabled precise and dynamic control over gene expression, enabling researchers to engineer and manipulate cellular processes at higher precision without the need for chemical diffusion. Their applications are far reaching, from inducing behavioral changes in mice [25], to controlling gene regulation [23, 26].

One way to use optogenetics to control cell signaling is to cluster proteins together using light-responsive protein tags. This effectively drops protein concentrations by confining them to small regions [27]. This can be used to inhibit or activate signaling pathways, depending on the specific proteins being targeted.

PID controllers have been used to regulate the intensity and duration of light exposure in optogenetics, providing a precise means to control the activation or deactivation of light-sensitive proteins. This enables researchers to finely tune the manipulation of cellular processes in optogenetic experiments, facilitating the study of dynamic biological control of gene regulation. So called, "cyborg" cells have been developed using computer controlled optognetics enhanced with PID controllers.

1.0.4 Noise and stochasicity

Noise and stochasticity are intrinsic to gene regulation [28], creating an unpredictable aspect within many cellular behaviors. In this realm, fluctuations in molecular processes and interactions can lead to varying gene expression levels, even in genetically identical cells. This inherent noise arises from factors such as random molecular collisions, environmental variability, and the discrete nature of chemical reactions. Understanding and quantifying this noise are critical for deciphering how cells make decisions [29], respond to stimuli, and maintain robustness in the face of uncertainties. This noise is often a detrimental effect, making biological experiments challenging to interpret and disrupting the functionality of synthetic biological systems.

While noise is traditionally viewed as a nuisance, recent research highlights its functional significance. Stochasticity can drive phenotypic diversity, facilitate rapid adaptation to changing environments, and even synchronize cellular responses in populations. Understanding and harnessing the impact of noise in gene regulation is essential for comprehending the robustness of biological systems, and it holds promise for designing more effective synthetic biological circuits in the future. Researchers employ mathematical techniques like Finite State Projection (FSP) [18, 19] Gillespie Stochastic Simulation Algorithm (SSA, [30, 31]), and statistical moments analyses, to model and analyze these stochastic systems [32, 33].

The Finite State Projection is a computational and mathematical method used in the analysis and modeling of systems with noise. The finite state projection begins with the Chemical Master Equation defined over infinite space of possible chemical states [34] and truncates these into a finite space of states plus an error term. Probability can flow out of any state except the error state. This analysis allows for the observation of how probability distributions evolve over time, while also providing insights into the degree of simulation accuracy. The finite state project has been used to simulate the evolution of protein and mRNA statistics over populations of cells [19], and to create new control problems for gene regulation.

The Gillespie Stochastic Simulation Algorithm [30, 31] is a computational method used in stochastic models of biology and chemical kinetics. It is designed for simulating and modeling

stochastic chemical processes in discrete chemical systems using stoichiometries and associated propensities. This algorithm is valuable for modeling time series behaviors of stochastic processes such as gene regulation and population dynamics, as it enables the modeling of inherent randomness and fluctuations in these processes more accurately. The initial SSA algorithm was developed for well-mixed systems but simulations for spatially varying systems have been developed by creating SSA algorithms for each volume in space and modeling their diffusion [35].

When the number of molecules is large relative to its propensity SSA simulations can become prohibitively expensive to calculate and ordinary differential equations (ODE) offer a computationally efficient alternative for analysis. Researchers have used ODEs to study complex nonlinear systems for cellular processes and gene regulation, exploring a range of behaviors, including bifurcations [36], oscillations [37] and developing synthetic biological modules with good control performance [21] or other interesting behaviors [38, 39].

Despite extensive efforts to control noisy biological processes within cells, the strategies for addressing the control of noisy systems in gene regulation have remained simple. Often, these models rely on ordinary differential equations (ODEs) combined with assumed Gaussian noise to account for measurement variability. Conventional thinking for this approach of control theory suggests that any noise system will always have worse control performance than a deterministic one. In contrast, we hypothesize that consideration of the full probability distributions of system heterogeneities may yield improved control opportunities. While it is possible to robustly control systems despite the presence of noise, the exploration of control techniques utilizing master equation analysis as a foundation for noise-exploiting control theory remains underdeveloped. This concept will be further built upon in the theoretical developments of chapters two and three, while chapter four will discuss practicalities of microscope automation that are necessary before practical implementation of these strategies.

1.0.5 Single cell microscopy

Fluorescent single-cell microscopy is a technique that uses fluorescent labels and specialized microscopes to study chemical species (e.g., DNA, RNA, and proteins) within individual cells, offering high-resolution insights into cellular behavior. These techniques involve labeling specific cellular structures, proteins, or molecules with fluorescent markers, which emit light when exposed to specific wavelengths. This emitted light is captured by a specialized microscope, allowing researchers to visualize and analyze the distribution, behavior, and interactions of these labeled species within a single cell. Single-cell fluorescence microscopy has enhanced the understanding of cellular processes by providing insights into phenomena such as intracellular signaling, protein trafficking, and the central dogma of biology [40].

Single-cell microscopy allows researchers to identify heterogenous populations in a sample of cells. Quantitative data at the single-cell level are important for conducting rigorous statistical analyses. Single-cell microscopy produces datasets that can be statistically analyzed, unveiling valuable insights into their underlying biological and mathematical models based on the heterogeneity of cellular responses. At the single-cell level, it is possible to understand how heterogenous populations of cells evolve over time. The collection of data to describe the evolution populations of cells can often reveal interesting models which predict behavior well.

Single-cell microscopy provides spatial context to cellular events. It allows researchers to precisely locate specific molecules or structures within individual cells and study their sub-cellular distribution, facilitating a better understanding of cellular organization and function. The co-localzation of fluorescent signals reveals associations between biological species, and these signals have been used to provide insights into the central dogma of biology.

1.0.6 Fluorescent labeling

Fluorescent labeling of biological species in combination with high powered microscopes allow researchers to examine and localize various protein and RNA behaviors within a cell. The various labeling techniques such as green fluorescent protein (GFP), MS2, and single molecule fluores-

cence in situ hybridization (smFISH) have enabled a wide range of imaging tools for molecular biologists.

GFP fluorescent cell imaging utilizes green fluorescent protein to label and visualize specific cellular structures or proteins. It has been instrumental in tracking protein localization [41], visualizing gene expression mechanisms, monitoring dynamic processes within living cells [42] and facilitated the investigation of protein-protein interactions [43]. Moreover, it has been indispensable in live-cell imaging, making it possible to capture real-time events and dynamic behaviors within cells. Unfortunately, the long folding time of GFP [44] makes it impossible to use to to track translation dynamics which occur at the minute or sub-minute time scales [45].

MS2 is a fluorescence imaging method to study the dynamics of RNA molecules within living cells. In this technique, an mRNA of interest is tagged with multiple MS2 binding sites, and a fluorescent protein is fused to the MS2 coat protein. When the MS2 sites within the RNA interact with the MS2 coat protein-GFP complex, a fluorescent signal is emitted, enabling researchers to visualize and track the movement, localization, and behavior of RNA molecules in real time.

Single-molecule Fluorescent in situ hybridization (smFISH) imaging is a molecular biology technique for visualizing and pinpointing individual RNA sequences within cells and tissues [46, 47]. By using fluorescently labeled DNA or RNA probes that are complementary to the target sequences, smFISH allows researchers to precisely locate and quantify RNA in its native environment.

Real-time nascent chain tracking, allows scientists to monitor and analyze the creation and movement of nascent chain mRNA in living cells with high precision using fluorescent tags with fast binding affinity [40, 45, 48, 49]. These new imaging techniques have revealed real time behaviors for frame-shifting [50], apoptosis [51], and translation dynamics [52].

The combination of fluorescent labeling tools with new microscopy techniques like highly inclined and laminated optical sheet (HiLo) [53] enable the acquisition of high quality images with less noise. This combination can capture high-quality images with reduced background noise and enabling imaging for longer periods of time or for more axial z-planes with less photo-damage.

HILO's ability to selectively illuminate a thin section of the sample at an angle, combined with the specificity of fluorescent labeling, ensures precise and detailed imaging of cellular processes.

1.0.7 Image processing

Image processing plays an important role in fluorescence microscopy, where image processes can improve image quality and extract scientifically relevant information from images. Through the application of image processing techniques, researchers can efficiently generate data that would otherwise require a multi-step, labor-intensive process. This streamlines the analysis of fluorescence microscopy data.

Convolutional neural nets (CNN's) are a class of deep learning-based tools designed for cell segmentation in images. Cellpose [54] is an implementation of a CNN which identifies masks of cytoplasm and nuclei in cells using a U-Net structure. U-Nets are a type of CNN that are specifically designed for image segmentation. They have a U-shaped architecture, with an encoder path that extracts features from the input image and a decoder path that reconstructs the segmented image from the extracted features.

Spot tracking algorithms used fluorescence microscopy, assist in studying dynamic biological processes at the cellular and molecular levels. Spot tracking algorithms like can track the movement and behavior of these fluorescently labeled entities over time. FISH-quant software has been developed to accelerate the collection of spot count data from FISH experiments [55]. They enable researchers to investigate biological processes such as intracellular transport, protein trafficking, and cell migration. By accurately tracing the trajectories of individual fluorescent spots, researchers can derive insights into the underlying mechanisms of cellular functions.

The combination of microscope automation, high-throughput image analysis, and advanced decision-making enables the statistical analysis of experimental designs using Fisher information. Fisher information is a mathematical concept that can be used to quantify the amount of information that an experiment can provide about a model. Fisher information techniques can be used to design experimental designs that maximize the amount of information that can be obtained from

a given amount of data. When combined with large data acquisitions, this technology has the potential to greatly improve the likelihood of the model given a dataset [56].

The development of fluorescence microscopy techniques and image processing techniques has been critical to the analysis of cellular behavior. The creation of 'smart microscopy' and automation tools that enable the collection of population data statistics from single-cell data can begin to answer how heterogenous behaviors in single-cell experiments is poised to increase the rapidity at which such experiments can inform us about model likelihoods and parameter estimates. Chapter four discusses the development of an automated microscope in more detail, which is heavily motivated by modern developments in FIM for single-cell experiment design under circumstances of noisy experiment al measurements [57].

Chapter 2

Stochastic Control of Static Systems

[1] Synthetic biology seeks to develop modular bio-circuits that combine to produce complex, controllable behaviors. These designs are often subject to noisy fluctuations and uncertainties, and most modern synthetic biology design processes have focused to create robust components to mitigate the noise of gene expression and reduce the heterogeneity of single-cell responses. However, deeper understanding of noise can achieve control goals that would otherwise be impossible. We explore how an "Optogenetic Maxwell Demon" could selectively amplify noise to control multiple cells using single-input-multiple-output (SIMO) feedback. Using data-constrained stochastic model simulations and theory, we show how an appropriately selected stochastic SIMO controller can drive multiple different cells to different user-specified configurations irrespective of initial condition. We explore how controllability depends on cells' regulatory structures, the amount of information available to the controller, and the accuracy of the model used. Our results suggest that gene regulation noise, when combined with optogenetic feedback and non-linear biochemical auto-regulation, can achieve synergy to enable precise control of complex stochastic processes.
Keywords: *synthetic biology, autoregulation, gene regulation noise, optogenetic feedback control, single-input-multiple-output control, Maxwell's Demon*

2.1 Introduction

Synthetic biology seeks to develop and characterize biological circuits and modular components that can be reliably re-engineered, re-assembled, and controlled to produce complex biological behaviors [58]. The design and implementation of modular components as simple functional

[1]Reprinted with permission from May, M (2021, November 18). Exploiting Noise, Non-Linearity, and Feedback for Differential Control of Multiple Synthetic Cells with a Single Optogenetic Input. ACS Synthetic Biology, 10(12), 3396–3410 Copyright 2023 American Chemical Society. https://pubs.acs.org/doi/full/10.1021/acssynbio.1c00341

As first author, MPM was the primary researcher on this paper and was responsible for all simulations and writing of the initial manuscript.

units [59] capable of being assembled to perform desired regulatory needs has yielded powerful capabilities of synthetic cells to perform specific actions in response to specific stimuli [60]. Early advances led to the development of programmed cells that are capable of complex logic like switching, self-regulation, and fast acting control [61]. Genetically engineered switches were originally built in bacteria [62], but have been extended to yeast [63], mammalian cells [64], and even multicellular plants [65]. In turn, these simple switches have led to more complex engineered biological systems and biotechnologies capable of performing tasks like controlling cells to behave as digital displays [4].

Much of the design process for synthetic biology has focused directly on building better biological components, such as creating more sophisticated gene regulatory structures [66–68], introducing new response elements or reporters [40], or introducing more orthogonal cellular signals [69]. Development work on gene regulatory structure has led to substantial advances in phenotype control in plant biology and targeted or modified protein turnover in therapeutics [70, 71]. Improvements to response reporters and the experimental techniques used to analyze cells, especially in the form of fluorescent protein reporters [72,73], real time single-gene MCP-MS2-based transcription elongation assays [74], and nascent chain translation elongation assays [40, 45, 48, 49] has made it possible for cells to transmit their internal states to human observers. These technologies allow for more observation and control, not just at the protein level, but at the gene and RNA levels as well. Advances in cellular signals have also introduced the potential for synthetic regulatory modules in separate cells to communicate with one another and control multicellular population dynamics [75–77]. For example, by considering a simple model for the effect of cellular quorum sensing on cell densities, simple circuits can be tuned to control cell densities [78].

Although the above advances have been developed primarily to control autonomous biological behaviors, these improvements to regulatory structures, response reporters, and signaling capabilities can also provide a framework to allow observers or external electronic devices to monitor cellular environments and dynamically reprogram the cellular logic. When coupled to advances in microfluidics, these capabilities introduce a new part-biology-part-machine (or cyber-organic [79])

paradigm that adds new possibilities for distributed external and internal biological control of synthetic biological circuits [80]. In particular, recent developments in optogenetics [81, 82] have greatly increased the speed and sensitivity by which external signals can be communicated from humans or machines to cells. Using these advances in microfluidics and light-activated gene regulatory elements is rapidly improving the potential to integrate carbon- and silicon-based circuits, which in turn is making hybrid bio-electronic circuits far more powerful than before.

A key challenge to integrating cell-based genetic circuity with electronic control is that an uncountable number chemical species and regulating bodies must diffuse and interact with one another in space and time within each cell. Tractable analysis requires immense simplifications of these infinitely complex and chaotic dynamics, and such simplifications naturally result in large uncertainties that can only be accounted for through the introduction of approximate models with stochastic analyses. One of the greatest challenges to improving externally controlled cellular behaviors is that this 'noise' in gene regulation introduces large amounts of single-cell heterogeneity which must accounted for [83–85]. Under the context of this noise, cellular responses are probabilistic–their distributions may shift gradually and even be statistically predictable under environmental or genetic manipulations [19, 86, 87], but individual cells appear to behave at random with very little information about their instantaneous external environment [88]. Unfortunately, this noise makes it difficult to precisely predict how or when an individual cell will respond to a new environmental stimulus. Although recent work has shown that external feedback can control and reduce cellular heterogeneity within a large population [89], it may seem unlikely that any feedback control strategy could reliably guarantee that specific cells within a population will respond as desired, and independently of their initial conditions.

Most efforts on external feedback control in synthetic biology have focused on the use of chemical or optical inputs to manipulate cell population averages [90] or to control individual cells within a larger population [90]. Such efforts can be classified as single-input-single-output (SISO) or multiple-input-single-output (MISO) control in that they seek only to push cells to a single phenotype. For example, recent experimental and computational studies [91–93] have used periodic

chemical input fluctuations to control one cell or a population of multiple cells to be as close as possible to the same unstable fixed point, a task that is similar to the inverted-pendulum problem in classical control theory. Other recent work has sought to control multiple individual cells, each with their own tailored optogenetic inputs [81], which corresponds to multi-input-multi-output control (MIMO). SISO and MISO control are limited to control only a single cellular response at a time, while MIMO requires advanced hardware such as digital micro-mirror devices that devote a separate input to each individual cell [81, 94]. However, the combination of synthetic biological designs, precise external controls, and quantitative measurements and models of single-cell noise could create new opportunities for single-input-multiple-output (SIMO) control, where multiple individual cells could be controlled to achieve *different phenotypes*, but requiring only a single input signal. In this article, we explore the potential of a noise-enabled SIMO control strategy that is similar to a hypothetical Maxwell's Demon (MD), who watches the random process and identifies short instances in time when specific cells have randomly increased sensitivities to small perturbations [95]. In the context of synthetic biology, we explore how advances in fast fluorescent reporters allow this MD to watch biological responses; how noise breaks the symmetry between identical cells in identical environments; and how advances in optogenetics may enable the MD to drive specific cells toward specific phenotypes, even when each cell always experiences the exact same input signal as every other cell.

Through simulation of multiple models that have been parametrized from existing bulk level optogenetic control experiments, we demonstrate that realizing an optogenetic MD for use in genetic regulation applications requires not only careful consideration of the cellular regulatory systems to be controlled, the fluorescent sensors to be observed, and the optogenetic inputs to be delivered, but it is also necessary to build predictive stochastic models and deterministic SIMO control algorithms to serve as its brain. In this article, we use a combination of simulations and theoretical analyses to explore how existing biological parts could be combined with models in the context of optogenetics to control the gene expression of multiple cells at once, even when both cells receive the same light signal at the same time and have the exact same genetics. We

then introduce a new probabilistic model predictive control (pMPC) strategy that can in principle control multiple cells to different phenotypes, even when only observing a single cell. Finally, we show that controls that are designed and optimized using one approximate model of the biological system can be effective to control a much more complex biological process whose mechanisms are not considered during the controller design.

2.2 Results and Discussion

We start by defining a set of two models, each with a different level of complexity, to describe the dynamics of optogenetically activated T7 polymerase in temporally-varying light conditions. We then use experimental data from Baumschlager et al [23] to independently constrain each of these models to the same data. We then take the simpler of the two models and extend it to include a typical auto-regulation module, and we examine the performance of this extended model under different fluctuating input signals at different frequencies using deterministic and stochastic analyses. We then show how intrinsic noise in the system dynamics can be utilized by a feedback controller to break symmetry in the process and force a system of two cells each to obtain specified phenotypes and independent of initial conditions. We then propose a new probabilistic model predictive controller scheme that is capable to differentially control multiple cells even when only one is directly observed. Finally, we demonstrate in principle that an optogenetic controller that is identified using a coarse-grained simplified model is capable to control behavior of a more complicated system with different and hidden dynamics.

2.2.1 A data-constrained model for light-induced T7 polymerase activation and gene expression

We begin by developing an *unregulated* model (denoted as $\mathcal{M}_U$) to describe the light-induced activation of an optogenetically controlled T7 polymerase as studied in Baumschlager et al [23]. As depicted in Fig. 2.1A, this system contains two light activated T7 domains ($T7_n$ and $T7_c$) that are produced at constant rates k_n and k_c, and which degrade in a first order decay process with rate

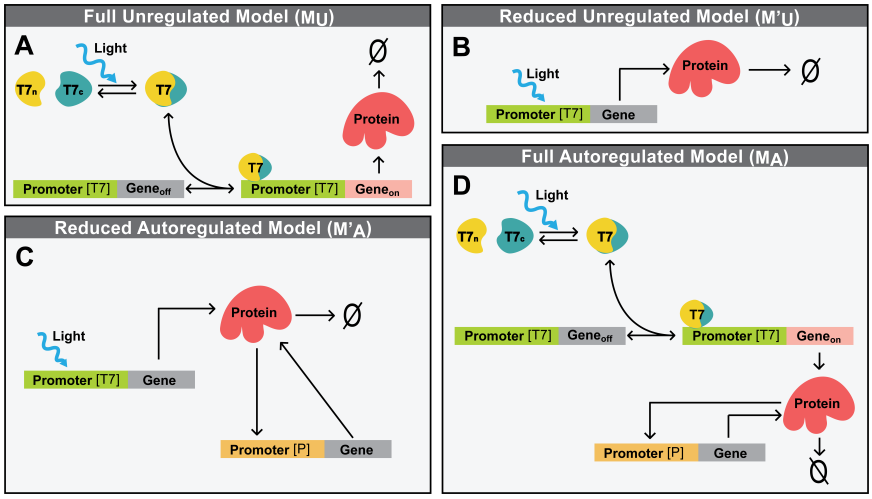

Figure 2.1: Diagrams of models for light-induced gene regulation with and without auto-regulation. (A) The full physical model ($\mathcal{M}_U$) describes the full mechanistic processes of T7 polymerase dimerization and gene activation in addition to protein production and degradation. (B) A simplified model ($\mathcal{M}'_U$) lumps together all T7 dimerization and gene activation dynamics so that protein is produced at a light-controlled rate. (C) The simplified model is extended ($\mathcal{M}'_A$) to include auto-regulation through the addition of a secondary promoter. (D) The full mechanistic model is also extended ($\mathcal{M}_A$) through inclusion of the same auto-regulation promoter.

γ_M. These domains dimerize when subject to light activation leading them to form an active T7 polymerase at a light dependent rate of $u(\phi)$ and the complex dissociates at a rate of k_{i1} and decays at a rate γ_T. The T7 polymerase dimer can bind to, or unbind from, the gene at rates of k_{f2} and k_{i2}, respectively. When bound, active protein production occurs at a rate of k_{f1}, where transcription and translation are lumped into a single event. Proteins are assumed to degrade according a first order rate process with rate γ_P. These interactions are described by the following reactions

$$\mathcal{M}_U = \left\{ \begin{array}{rcl} \emptyset & \leftrightarrow & T7_n, \\ \emptyset & \leftrightarrow & T7_c, \\ T7_n + T7_c & \leftrightarrow & T7\ , \\ T7 & \rightarrow & \emptyset\ , \\ T7 + g_{\text{off}} & \longleftrightarrow & g_{\text{on}}\ , \\ g_{\text{on}} & \longrightarrow & P + g_{\text{on}}\ , \\ P & \longrightarrow & \emptyset\ , \end{array} \right\} \tag{2.1}$$

where the first two bidirectional reactions describe production and decay of $T7_n$ and $T7_c$; the third bidirectional reaction describes light-induced reversible dimerization, where the light induction level is denoted as ϕ and its effect is modeled by the function $u(\phi)$; the fourth unidirectional reaction describes the decay of the T7 dimer; the fifth bidirectional reaction describes the T7 association and dissociation to the gene; and final two unidirectional reactions describes the production and degradation for the resulting protein product, where transcription and translation have been lumped into a single reaction. The rate for each reaction is given directly above or below its respective arrow.

The ODE for $\mathcal{M}_U$ can be written in vector form as

$$\frac{d}{dt}\begin{bmatrix} [T7_n] \\ [T7_c] \\ [T7] \\ [g_{\text{on}}] \\ [P] \end{bmatrix} = \begin{bmatrix} -u(\phi)[T7_n][T7_c] + k_{i1}[T7] + k_n - \gamma_M[T7_n] \\ -[T_7n][T7_c] + k_{i1}[T7] + k_c - \gamma_T[T7_c] \\ u(\phi)[T7_n][T7_c] - k_{i1}[T7] - k_{f2}[T7](g_{\text{total}} - [g_{\text{on}}]) + k_{i2}[g_{\text{on}}] - \gamma_T[T7] \\ k_{f2}[T7](g_{\text{total}} - [g_{\text{on}}]) - k_{i2}[g_{\text{on}}] \\ k_{f1}[g_{\text{on}}] - \gamma_P[P] \end{bmatrix}, \tag{2.2}$$

where we have assumed a fixed number of gene copies $[g_{\text{total}}] = [g_{\text{on}}] + [g_{\text{off}}]$ to remove the variable $[g_{\text{off}}]$ from the equations. Once written in this form, Model $\mathcal{M}_U$ can be integrated numerically for any given set of initial conditions and parameters. The model parameters were then fit to

Table 2.1: Parameters of the full model system.

Parameter Name	Parameter Value	Units
k_n	2.00×10^{-1}	molecules/min
k_c	6.00×10^{-1}	molecules/min
k_{i1}	2.00×10^{2}	min^{-1}
γ_M	5.00×10^{-2}	min^{-1}
γ_T	5.00×10^{-2}	min^{-1}
k_{i2}	5.00×10^{-1}	min^{-1}
k_{f2}	5.00×10^{-1}	molecules/min
k_{f1}	1.42×10^{-2}	min^{-1}
γ_P	2.03×10^{-2}	min^{-1}

the experimental data from Baumschlager et al [23], in which the system was subjected to three different levels of UV radiation:

$$\phi(t) = \begin{cases} 320 \text{ Watts/cm}^2, \text{ for } 0 \leq t < 270 \text{ minutes}, \\ 0 \text{ Watts/cm}^2, \text{ for } 270 \leq t < 570 \text{ minutes}, \\ 20 \text{ Watts/cm}^2, \text{ for } 570 \leq t \text{ minutes}. \end{cases}$$

The parameters of $\mathcal{M}_U$ and $\{u(\phi_1), u(\phi_2), u(\phi_3)\}$ are simultaneously fit to the measured time series fluorescent protein trajectory. This fit suggests u=[0.4060, 0.00, 0.0400] molecules^{-1}min^{-1} when ϕ=[320, 0, 20] Watts/cm^2, respectively. To interpolate for intermediate values of light intensity, the function $u(\phi)$ is then defined as the cubic spline of the three data points and is shown with a red line in Fig. 2.2B. The resulting fit of the model to the data is shown by the red dashed lines in Fig. 2.2A, and the remaining parameters of the model fit are shown in Table 2.1.

Having determined a baseline ODE-based model that yields a good fit to existing experimental data for the system's temporal response, in the next sections we will specify a much simpler, but less accurate, version of this model and use that approximate model to perform stochastic analyses, suggest design modifications, and specify a controller that can drive differential gene expression among two or more cells using a single external input signal.

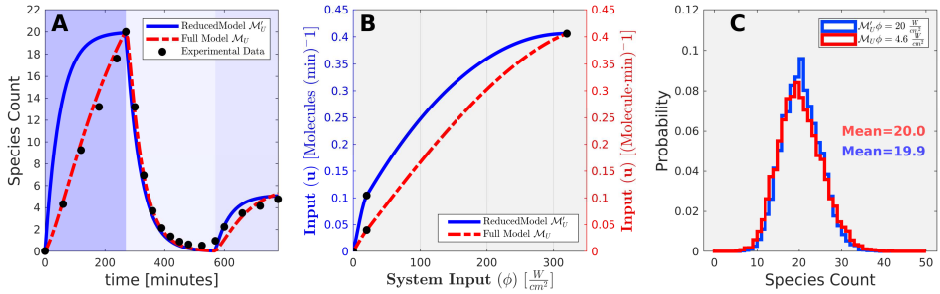

Figure 2.2: Parameter estimation of full and reduced unregulated models, $\mathcal{M}_U$ and $\mathcal{M}'_U$, respectively. (A) Fits of $\mathcal{M}_U$ (red line) and $\mathcal{M}'_U$ (blue line) to experimental data from Baumschlager et al. [23] (black dots).(B) Calibration curves representing the conversion of light [Watts per cm^2] to the associated reaction rates for $\mathcal{M}_U$ (red) and $\mathcal{M}'_U$ (blue). (C) Steady state histograms predicted by models $\mathcal{M}_U$ (red) and $\mathcal{M}'_U$ at inputs that are calibrated to results in an average expression of 20 molecules per cell.

2.2.2 A substantially reduced and approximate model for T7 activation and gene expression

Although we were able to find many good parameter sets so that model $\mathcal{M}_U$ could match the data [23], identification of a single unique 'best' parameter set is infeasible due to the high number of parameters, severe limitations on available experimental data, and sloppiness [96] in the model parameters. Finding a simpler, but better constrained model would not only help to reduce sensitivity to unknown parameters, but could also dramatically reduce computational costs when using that model for design decisions or for the specification of feedback control strategies. To simplify these model reactions and to obtain a more identifiable parameter set, we next propose a simple generalized birth-death model in which the protein production rate is given by $k_0 + u'(\phi)$, where k_0 is the baseline production rate with no light input, and $u'(\phi)$ is a light-dependent control input function. We use the apostrophe notation $(\cdot)'$ to denote use of the reduced model in which the units of the control signal at a given light intensity have been adjusted to molecules per minute. Under this simple rule and assuming linear decay at rate γ_P, the approximate dynamics of $P(t)$ are written simply as:

$$\frac{dP}{dt} = k_0 + u'(\phi) - \gamma_P P. \tag{2.3}$$

In this model, which we denote as $\mathcal{M}'_U$, the parameters k_0 and γ_P and the specific values of $u'(\phi)$ at $\phi \in \{320, 0, 20\}$ Watts/cm^2 are again simultaneously fit to capture the time dynamics of

Table 2.2: Parameters of the reduced model system.

Parameter Name	Parameter Value	units
γ_P	2.03×10^{-2}	1/min
k_0	1.00×10^{-4}	molecules/min

the experimental data [23]. This fit suggests $u'(\phi)$=[0.4060, 0.00, 0.1044] molecules/min when ϕ=[320, 0, 20] Watts/cm^2 respectively, and the cubic spline of these three data points yields the calibration curve $u'(\phi)$ as shown in Fig. 2.2B (blue line). The resulting fit of model $\mathcal{M}'_U$ is shown in the solid blue line Fig. 2.2A, and the remaining parameters for the reduced model are shown in Table 2.2.

We next extended both the original model and the simplified model to include discrete stochastic events for protein production and degradation, as well as for T7 dynamics for the full model. Using the exact same parameter values as for the ODE analysis, we then simulated the model using the Stochastic Simulation Algorithm (SSA, [30]). Fig. 2.2C shows the probability distribution of P sampled using 10^5 independent SSA runs, each simulated to 3000 minutes and only using the last data point of each simulation for either the full model (red) or the simplified model (blue). Both models provide reasonably good, but not identical, matches to the mean protein levels and the overall time scales as shown in Fig. 2.2A, and we observed strong agreement in their stationary distributions. However, it remains to be seen if the differences in architecture and time scales need to be accounted for when we use the simplified model ($\mathcal{M}'_U$) to guide our modification of the gene regulatory system and to design a feedback controller that remains effective even when applied to the more complex original model ($\mathcal{M}_U$).

Table 2.3: Hill function parameters for auto-regulation motif.

Parameter Name	Parameter Value	Units
κ	4.06×10^{-1}	molecules/min
γ	2.03×10^{-2}	min^{-1}
β	20	molecules
α	8	unitless
k_0	1×10^{-4}	molecules/min

2.3 Results

2.3.1 Addition of an autoregulation motif creates light-dependent bistability

Starting from the simplified model, $\mathcal{M}'_U$, we next asked if a common auto-regulation motif could be added to introduce light-dependent bistability for the system and so that we could explore how that added motif would impact the controllability of the overall system. Fig. 2.1C shows a schematic of the new model denoted as $\mathcal{M}'_A$, which has auto-regulation due to the addition of a secondary promoter that is self-activated by protein P to produce more of itself. To incorporate this auto-regulation behavior, a Hill function production rate is assumed with high cooperativity, and the new rate equation becomes:

$$\frac{dP}{dt} = u'(\phi) + k_0 + \kappa \frac{P^\alpha}{P^\alpha + \beta^\alpha} - \gamma P, \tag{2.4}$$

where the first term $u'(\phi)$ corresponds to the control input as a function of light input; the second and third terms correspond to the Hill function activity of the auto-regulation promoter with leakiness k_0, and the final term corresponds to the first order decay of the protein. The Hill parameters (β, α) of $\mathcal{M}'_A$ are chosen in order to exhibit bi-modal behavior in the dynamics of P. All parameters of this auto-regulatory model are shown in Table 2.3.

Fig. 2.3A shows bifurcation diagrams for the simplified unregulated and auto-regulated models, $\mathcal{M}'_U$ and $\mathcal{M}'_A$, respectively. These diagrams show that as the light input sweeps *slowly* from zero to 0.5 Watts per cm^2, bifurcation and hysteresis become apparent in the auto-regulated model ($\mathcal{M}'_A$,

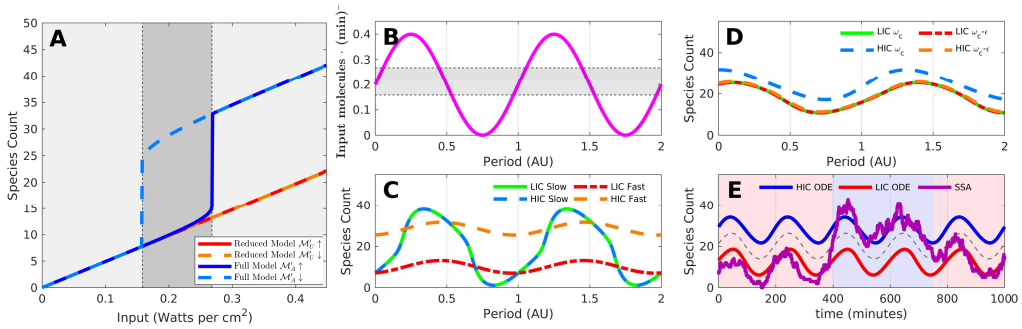

Figure 2.3: Input-output analysis for the simplified model with and without auto-regulation (Models $\mathcal{M}'_U$ and $\mathcal{M}'_A$). (A) Bifurcation (hysteresis) analyses of $\mathcal{M}'_U$ (red shades) and $\mathcal{M}'_A$ (blue shades). Solid (dashed) lines depict the change in the steady state solution as UV intensity is increased (decreased). Bi-modality (i.e., two possible steady states at the same control input) and hysteresis (i.e., different paths of solid and dashed lines) only occur for $\mathcal{M}'_A$. (B) Sinusoidal input used for temporal excitation shown for two periods. (C) Steady state temporal behavior of $\mathcal{M}'_A$ under sinusoidal input shown for two periods. Dashed lines correspond to high initial conditions, and solid lines correspond to low initial conditions. Different colors correspond to different input signal frequencies: fast (0.01 RPM, red/orange), slow (0.001 RPM, blue shades). All plots show steady state temporal response after a transient time of at least two oscillation periods or 3500 minutes, whichever is longer. (D) Same as (C) but with frequencies 0.0043522 RPM (red/orange) and 0.0043525 RPM (blue shades). The $\mathcal{M}'_A$ maintains memory of initial condition provided that input frequency is greater than a critical value (i.e., solid and dashed lines remain distinct). (E) Capability of Model $\mathcal{M}'_A$ to track inputs assuming stochastic fluctuations as analyzed using an extended SSA [30] with extra reactions [97].

light/dark blue), but these effects are not observed in the unregulated model ($\mathcal{M}'_U$, orange/red). For either model, low and high light inputs each result in a single stable point at low or high expression, respectively. For intermediate light inputs, however, two history-dependent stable points coexist for $\mathcal{M}'_A$, and it is possible for two cells to maintain different stable points (or phenotypes) provided that the light intensity remains in the bi-stable region and that the cells begin in the separate basins of attraction for the different stable points.

In the hysteresis plots of Fig. 2.3A, it is assumed that the light input sweeps very slowly so that the response reaches equilibrium at each light level before subsequent changes. However, in the context of feedback control, we are more interested in how cells respond to light fluctuations at faster, transient time scales. Therefore, we next test the stability of input-to-output behaviors under time-varying inputs. We start simulations for two cells with identical parameters, but at different initial conditions (i.e., one at a high initial concentration of 40 molecules per cell and another at a low concentration of 0 molecules per cell), and we subject these both to the same

sinusoidally-varying input signal whose range encompasses both bifurcation points as shown in Fig. 2.3B. Fig. 2.3C shows the steady state trajectories for two pairs of such systems, where dashed lines correspond to trajectories that start at low initial conditions and solid lines correspond to trajectories starting at high initial conditions. In all cases, the system is subject to at least four cycles or 3500 minutes so that transient dynamics have had time to decay, and the time axis is scaled to show the response over two oscillation periods. Fig. 2.3C (blue shades) shows that when the input frequency is slow ($\omega_{slow} = 0.001$ rotations per minute (RPM)), the system loses memory of its initial condition and the trajectories from both initial conditions decay to a single trajectory. However, when the frequency is fast (red and orange trajectories, $\omega_{fast} = 0.01RPM$), the system can maintain memory indefinitely. Fig. 2.3D shows that the cut off frequency for this maintenance of memory is sharp in that memory is possible at a frequency of ω_c (blue shades) but is lost at a slightly slower frequency of $\omega_c - \varepsilon$, where $\omega_c = 0.0043525$ (red/orange) RPM is the critical frequency and $\varepsilon = 3 \times 10^{-7}$ RPM is a small perturbation to that frequency.

2.3.2 Intrinsic noise can drive cells to switch phenotypes

Using deterministic analyses of the bistable model, we have seen that two cells with different initial conditions maintain separate phenotypes as they respond to the same fluctuating input signal. The flip side of this deterministic result is that two fully converged solutions of the same ODE never diverge, such that two cells starting at the exact same initial condition will never express unique phenotypes even when bi-stability is possible. However, low copy numbers of important regulatory molecules (DNA, RNA and proteins) often result in stochastic fluctuations in cellular concentrations (also known as 'intrinsic noise') that dramatically affects both of these results. When added to a bistable deterministic process, noise can drive two cells starting at the same initial phenotype to diverge or even drive two cells to exchange phenotypes by chance over time [98, 99]. With this possibility in mind, we next ask how noise would affect the ability of cells to track a temporally varying input signal. For this, we examined the production and degradation reactions and converted model $\mathcal{M}_A$ to an equivalent discrete stochastic model with the exact same

rate parameters, and we explored how discrete intrinsic noise of stochastic models could be used drive cells to separate phenotypes. To extend the SSA to approximate time-varying inputs, we adopted an approach similar to that in Voliotis et al [97], and added a fast 'null event' reaction that updates the clock and input signal value on a time scale that is much faster (i.e., average of 100 events per period of the input signal) than that of the input signal fluctuations. We then compared the ODE and the SSA analysis of $\mathcal{M}'_U$ under a sinusoidal input with moderate input frequency of $\omega = 5.00 \times 10^{-3}$ RPM $> \omega_c$ for which the ODE trajectories maintain memory of their initial conditions. Although the ODE solutions (smooth lines in Fig. 2.3E) will never converge, the stochastic trajectories (purple fluctuating trajectories) switch occasionally between the two fluctuating phenotypes. In other words, with the addition of noise to the system, each cell can slowly 'forget' its original configuration. Moreover, the probability of switching depends on the transient stochastic state of the process and the frequency and amplitude of external input fluctuations. Previous studies have observed similar effects for how noise creates variation in a population of cells, and past feedback control efforts have sought to counteract this variation to keep all cells at a chosen (and in some cases unstable) phenotype [91–93]. In the next section, we do not try to reduce variability among cells, but rather we seek to exploit the condition- and time-dependent disruption of symmetry to push one cell to a chosen phenotype while forcing another cell or group of cells toward a different chosen cell fate.

2.3.3 Finite State Projection analyses uncover effective strategies to differentially control two cells using a single input

We consider a system of N_c cells, each with identical regulatory mechanisms and parameters, but whose fluctuating protein concentrations at time t are denoted by $\tilde{x}_1(t), \tilde{x}_2(t), \ldots$, which we can arrange into the vector $\tilde{\mathbf{X}}(t) = [\tilde{x}_1(t), \tilde{x}_2(t), \ldots]^T$. Here, the notation $\tilde{(\cdot)}$ denotes that the corresponding quantity (e.g., protein copy number) is the result of a stochastically fluctuating process. Our overall goal is to design a feedback control law to force $\tilde{\mathbf{X}}(t)$ as close as possible toward an arbitrary target state $\hat{\mathbf{T}}$. For an example with two cells, $\hat{\mathbf{T}} = [\hat{T}_1, \hat{T}_2]^T = [30, 10]^T$ would correspond

to having the first cell in the high expression phenotype and the second cell in the low expression phenotype. In general, this fluctuating control signal could depend on measurements of $\tilde{\mathbf{X}}(t)$ and/or the current time according to some as yet to be determined control function $\tilde{u} = u(\tilde{\mathbf{X}}(t), t)$. We define a cost function as the expected squared Euclidean distance between $\tilde{\mathbf{X}}(t)$ and $\hat{\mathbf{T}}$ at steady-state, which can be written as

$$\begin{aligned} J &= \lim_{t\to\infty} \mathbb{E}\{|\tilde{\mathbf{X}}(t) - \hat{\mathbf{T}}|_2^2\} \\ &= \lim_{t\to\infty} \sum_{i,j,\ldots} P(\tilde{x}_1(t) = i, \tilde{x}_2(t) = j, \ldots) \left((i - \hat{T}_1)^2 + (j - \hat{T}_2)^2) + \ldots\right) \\ &= \lim_{t\to\infty} \sum_{i,j,\ldots} P_{ij\ldots}(t) C_{ij\ldots} = \mathbf{C}\mathbf{P}_\infty, \end{aligned} \tag{2.5}$$

where $\mathbf{C}$ is a constant vector of squared Euclidean distances of each state from the target, and $\mathbf{P}_\infty$ is the steady-state probability mass vector (i.e., the stationary probability for each unique value of $\tilde{\mathbf{X}}$).

As described in Methods, the master equation for a finite number of cells subject to a fluctuating state- or time-dependent input signal can be written as

$$\frac{d}{dt}\mathbf{P} = (\mathbf{A}_0 + \mathbf{B}\mathbf{u}'(t))\mathbf{P}, \tag{2.6}$$

where $\mathbf{P} \in \mathbb{R}_{\geq 0}^{n^{N_c N'}}$ is the non-negative probability mass vector for all possible states; $\mathbf{A}_0 \in \mathbb{R}^{n^{N_c N'} \times n^{N_c N'}}$ is an uncoupled infinitesimal generator; $\mathbf{u}'(t) \in \mathbb{R}_{\geq 0}^{n^{N_c N'}}$ is a (potentially time varying) vector of non-negative control inputs with one entry for every distinct state in the system's state space; and $\mathbf{B}$ is a fixed tensor that operates on $\mathbf{u}'$ to adjust the master equation to account for the optogenetic input. Explicit examples for the construction of quantities $\mathbf{A}$ and $\mathbf{B}\mathbf{u}'$ for different control laws are provided in Methods.

At first, we consider the special case where the control signal depends only on the current state at each instant in time. In this case, the vector $\mathbf{u}'$ is constant with respect to time and depends only on the enumeration of the possible states. As such, the infinitesimal generator in Eq. (2.6) reduces

to a time-homogeneous master equation for a standard discrete state Markov process. We note that the control signal $\tilde{u}'(t)$ may still fluctuate due to changes in $\tilde{\mathbf{X}}$ and can be written using the notation $\tilde{u}'(t) = \mathbf{u}_{\tilde{\mathbf{X}}'(t)}$ to denote that the index of the vector $\mathbf{u}'$ is specified by the instantaneous state $\tilde{\mathbf{X}}(t)$.

By changing the specification of $\mathbf{u}'$, we can further simplify this class to consider several different types of state-based control rules. Here we consider three different possibilities: an UnAware Controller (UAC) that has no information about any cells to make its control decision:

$$\mathbf{u}^{\mathrm{UAC}}_{\tilde{\mathbf{X}}(t)} = u^{\mathrm{UAC}} = \text{constant, where } \mathbf{u}^{\mathrm{FAC}} \text{ is the same for every } \tilde{\mathbf{X}}(t)); \tag{2.7}$$

a Fully Aware Controller (FAC) which has complete knowledge of all cells $\tilde{x}_1, \tilde{x}_2, \ldots$ and therefore uses the full state vector to make its control decisions:

$$\mathbf{u}^{\mathrm{FAC}}_{\tilde{\mathbf{X}}(t)} = u^{\mathrm{FAC}}(\tilde{x}_1, \tilde{x}_2, \ldots), \text{ where } \mathbf{u}^{\mathrm{FAC}} \text{ has a unique element for every } \tilde{\mathbf{X}}(t), \tag{2.8}$$

and a Partially Aware Controller (PAC) which uses only information of a single cell, e.g., x_1, to make its control decisions:

$$\mathbf{u}^{\mathrm{PAC}}_{\tilde{\mathbf{X}}(t)} = u^{\mathrm{PAC}}(\tilde{x}_1), \text{ where } \mathbf{u}^{\mathrm{PAC}} \text{ changes only with } \tilde{x}_1(t). \tag{2.9}$$

For our specific case of model $\mathcal{M}'_A$, the FSP truncation size is $n = 50$, and the number of species is $N' = 1$; thus for two cells ($N_c = 2$), the matrix $\mathbf{A}_0 \in \mathbb{R}^{2500\times 2500}$.

With the formal definition of the cost function J from Eq. (3.14) and the CME from Eq. (2.6), we can then compute the gradient of the cost function with respect to the control law as (see derivation in SI):

$$\nabla_{\mathbf{u}}(J) = \mathbf{C}(\mathbf{A}_0 + \mathbf{B}\mathbf{u}')^{-1}\mathbf{B}\mathbf{P}. \tag{2.10}$$

Having specified the cost function gradient with respect to the controller, we run an optimization algorithm (see Methods) to search along the gradient to find local minima for the cost function. We note again that for each of these optogenetic Maxwell's Demon control strategies, every cell expe-

riences equal inputs at every instant in time, although the magnitude of that single input changes in time as the cells fluctuate to different states.

Starting with the unaware (UAC) and fully aware (FAC) controllers, we used the FSP approach and local optimization (see SI for details on the optimization procedure) to find a locally optimal control strategy in the form of the constant u^{UAC} or the two-dimensional scalar field $u^{FAC}(x_1, x_2)$, and with and without the addition of auto-regulation to the model. The resulting optimal controllers are presented in the top row of Fig. 2.4, where the color at each point represents the control input magnitude $u^{xAC}(x_1, x_2)$ as a function of the species quantities x_1 for the first cell and x_2 for the second cell. In the figure, a vector field of white arrows depicts the net direction of probability flow due to the combined action of the internal auto-regulatory effects and the feedback control input $u^{xAC}(x_1, x_2)$ on the system. For practical implementation, the control input $u'(\phi) = u^{xAC}(x_1, x_2)$ would be converted to light intensity through inversion of the calibration curve in Fig. 2.2B (blue line). The middle row of Fig. 2.4 shows the resulting steady state joint distribution of each condition, and in the bottom row of Fig. 2.4 shows the corresponding marginal distributions, with cell one in solid blue lines and cell two in dashed red lines. Fig. 2.4 shows that symmetry is broken only in the case where the cells' genetic design includes auto-regulation *and* the feedback controller contains knowledge of the cells (i.e., the far right row of Fig. 2.4). In this best-case scenario, the two cells are very effectively driven each to their own unique and pre-chosen phenotype, irrespective of their initial conditions. If feedback is included without auto-regulation, the cells' distributions are made tighter at some intermediate value between the target values for cell one and cell two; this results in a slightly better numerical cost value, but it becomes even less likely that both cells will reach their target phenotypes at the same time as compared to the uncontrolled situation (compare first and second row in Fig. 2.4). Conversely, if auto-regulation is included without feedback (i.e., the light level is fixed at some optimal value), the cells exhibit a bimodal distribution with some cells near each target value, but there is no means to control which cell expresses which phenotype, and the cost function is again worse than the case with no auto-regulation (compare first and third rows of Fig. 2.4). These data taken together suggest that auto-regulation and feedback

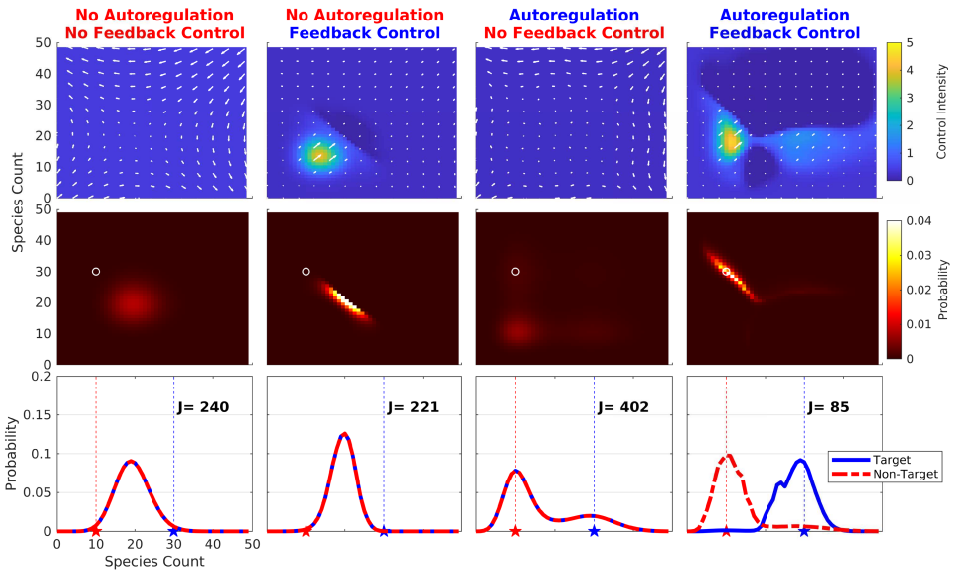

Figure 2.4: Optimized control laws and performance results for cells and different combinations of regulatory structure and control strategy. (Top row) Optimized control input versus cells' states, $\mathbf{u}^{xAC}(x_1, x_2)$ (colormap at far right). White vectors plotted over the control inputs represent the net flow of probability at that point in state space. (Middle row) Resulting joint probability distribution ($\mathbf{P}(x_1, x_2)$, colormap at far right). The target point $T = [30, 10]^T$ is denoted by a white circle. (Bottom row) Corresponding marginal distributions for the cells $\mathbf{P}(x_1)$ (blue) and $\mathbf{P}(x_2)$ (red) and time-averaged cost function, J. Leftmost column shows results with no auto-regulation and a constant control signal. Second column shows the result with no auto-regulation and feedback control strategy. Third column shows the result with auto-regulation and constant control signal (UAC in text). Rightmost column shows the results for auto-regulation and fully aware feedback control (FAC in text).

control, in addition to intrinsic single-cell noise, are all critical to the break symmetry and enable differential control of multiples cells using a single input. Specifically, noise breaks the symmetry of cell behavior and allows cells to switch independently between phenotypes, feedback helps to reinforce this noise and steer cells toward desired phenotypes, and auto-regulation helps stabilize cell behaviors once they have attained their desired phenotypes.

2.3.4 Effective differential control of many cells using a single input is possible, even when observations are limited to a single cell of interest.

We next examined a more general problem to control an arbitrary number of cells simultaneously. In this case, we consider a situation where the controller acts on many cells simultaneously with the goal of steering a single observed cell (x_1) to one state and all remaining unobserved cells

to another state. To handle this problem, we first utilize a partially aware controller (PAC) that observes only $\tilde{x}_1(t)$ and ignores the states of all other cells. Fig. 2.5 compares the results of this simplified controller (middle column) to those of the two-cell controller from the previous section (left column). The resulting control law, $u^{PAC}(x_1)$, is optimized to find the best control signal input for each possible value x_1 of the observed cell, and the top row of Fig. 2.5 shows that this optimal PAC controller depends only on the observed cell (x_1 axis) but is constant with respect to all unobserved cells (x_2 axis). Despite the simplicity of this control strategy and the fact that it requires only knowledge of the instantaneous expression of the single observed cell, the second and third rows of Fig. 2.5 show that the PAC controller effectively breaks symmetry to force the observed cell to a high expression phenotype while most unobserved cells are correctly directed to the low expression phenotypes.

Although the partially aware controller under-performs compared to the fully aware controller for the case of exactly two cells (compare middle and left columns of Fig. 2.5), the advantage of the PAC is that it works equally well, and without any modification, for any arbitrarily large number of unobserved cells (Supplemental Fig. S1). In contrast, to use the FAC for more than two cells requires modification, such as training of a higher rank tensor representation of the control algorithm, or defining a control law based on the mean, median, or some other statistical quantity for the groups of cells to be assigned to each phenotype. Unfortunately, the former high-order tensor approach is computationally intractable using existing methods, and the latter approach rapidly loses performance as the number of cells is increased. For example, when the controller is based on the observation of $\tilde{x}_1$ and the mean of the remaining cells $\{\tilde{x}_2, \tilde{x}_3, \ldots\}$ (FACM, see Supplemental Information), we observe that for any more than a single cell in the second group, the PAC outperforms the FACM (Supplemental Fig. S1, compare FACM and PAC controllers).

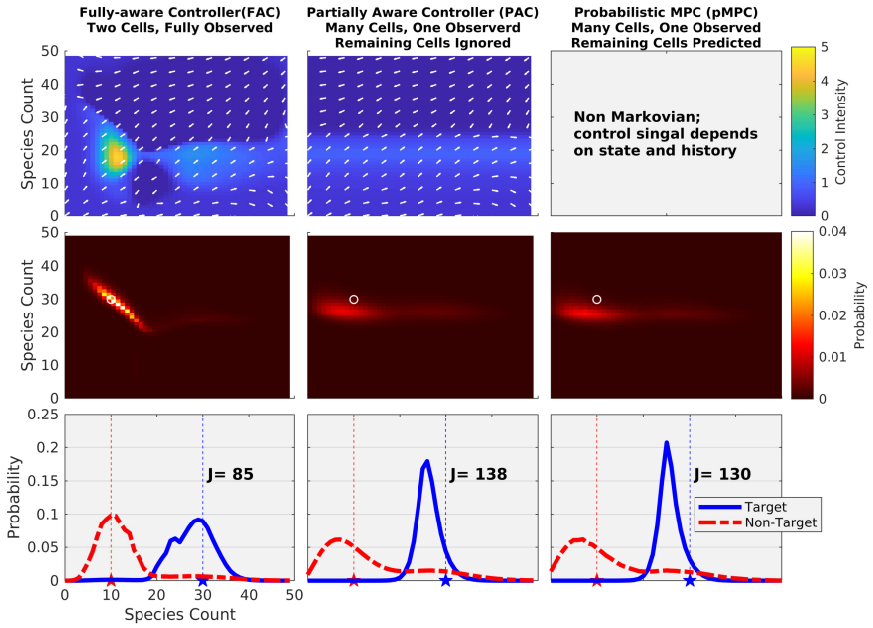

Figure 2.5: Control laws and performance for strategies to control many cells at once using partial knowledge only of a single observed cell. Each row shows: (top) the control law $\mathbf{u}'_{\tilde{\mathbf{X}}} = u^{xAC}(\tilde{x}_1, \tilde{x}_2)$, (middle) the resulting joint probability distribution ($\mathbf{P}(x_1, x_2)$), and (bottom) the marginal distributions for the observed cell $\mathbf{P}(x_1)$ (blue) and remaining cells $\mathbf{P}(x_2)$ (red). The middle column shows results for the partially aware controller (PAC in text), where the control depends only on the single observed cell ($\mathbf{u}'_{\tilde{\mathbf{X}}} = u^{PAC}(\tilde{x}_1)$). The rightmost row shows results for the probabilistic model predictive controller (pMPC), which also applies to an arbitrary number of cells. For the pMPC, the process is non-Markovian in that the controller ($\mathbf{u}'_{\tilde{\mathbf{X}}} = u^{\text{pMPC}}(\tilde{x}_1, \tilde{\mathbf{P}}(x_{i\neq 1}))$ depends not only upon the state of observed $\tilde{x}_1$, but also on the predicted probability vector for the unobserved cells, $\tilde{\mathbf{P}}(x_{i\neq 1})$. To enable comparison to previous two-cell cases, the leftmost column shows the results for the fully aware feedback control (FAC) with only two cells. Color bars are shown to the right of each row.

2.3.5 A probabilistic Model Predictive Controller (pMPC) can improve the control of many cells using a single observer and a single input signal.

In the previous section, the PAC control was based on only the observation of a single observed cell, and had no information about the other cells that it was also seeking to control. However, knowing the history of the input signal (i.e., the light intensity over time in the past), the FSP approach allows for the possibility that the controller can estimate the probability distribution of all non-observed cells. With this possibility in mind, we next explored a new class of controller that could use direct knowledge of the protein expression in the observed cell, the known control input signal at the current time, and a probabilistic model to predict the distribution of expression

in the unmeasured cells. Specifically, we use the FSP approach to integrate our prediction of the probability distribution for the unobserved cells as:

$$\frac{d}{dt}\tilde{\mathbf{P}}_{uo} = \left(\mathbf{A}_0 + \mathbf{B}\tilde{u}'(t)\right)\tilde{\mathbf{P}}_{uo}, \tag{2.11}$$

where $\tilde{\mathbf{P}}_{uo} \in \mathbb{R}^n_{\geq 0}$ is the estimate of the probability mass vector for the protein expression in the unobserved cells, $\mathbf{A}_0 \in \mathbb{R}^{n\times n}$ is the infinitesimal generator for a single cell in the absence of any control input, and the scalar variable $\tilde{u}'(t) > 0$ is the instantaneous input signal that is produced by the controller. We note that the probability mass vector estimate $\tilde{\mathbf{P}}_{uo}$ is the result of a stochastic process that depends upon the full history of the input signal $\tilde{u}'(t)$.

Using this prediction for the unobserved cells, the pMPC controller law can now be defined as

$$u^{\mathrm{pMPC}}(\tilde{x}_1, t) = c_{\tilde{x}_1} + \mathbf{z}_{\tilde{x}_1}\tilde{\mathbf{P}}_{uo}(t) \tag{2.12}$$

or written in vector form for all possible values of $\tilde{x}_1$ as:

$$\tilde{\mathbf{u}}^{\mathrm{pMPC}}(t) = \mathbf{c} + \mathbf{Z}\tilde{\mathbf{P}}_{uo}(t) \tag{2.13}$$

where $\mathbf{c} = [c_0, c_1, \ldots, c_{n-1}]^T$ is a constant vector in $\mathbb{R}^n$, and $\mathbf{Z} \in \mathbb{R}^{n\times n}$ is a matrix of linear weights which adjusts the input based off of the estimated unobserved probability distribution $\tilde{\mathbf{P}}_{uo}$. In our practical implementation, we assume that the controller in Eq. (2.13) is piecewise constant with respect to $\tilde{\mathbf{P}}_{uo}$ over a time step of 0.5 min, but it changes instantaneously with each even that affects the observed cell $\tilde{x}_1$. The weights of $\mathbf{c}$ and $\mathbf{Z}$ are then jointly optimized to minimize the cost function, J. The simple formulation of the control law in Eq. (2.13) admits the possibility for non-achievable negative values of light in order to construct a computationally tractable optimization procedure. However, in testing the controller, this non-physical situation is corrected by saturating negative control signals to zero in the true test of the system (see Supplemental Information). We note that this approximation to allow for negative control signals in the control law specifi-

cation and subsequent correction to saturate these to zero in the control law test suggest that the pMPC controller identified here is sub-optimal. However, despite this non-optimal design, Fig. 2.5 shows that the resulting *non-optimized* pMPC (J=130) controller outperforms the *fully optimized* PAC (J=138), demonstrating that probabilistic model predictions can be used to improve control performance even in the absence of observations for many of the cells under its control. Having succeeded in our main goal to determine if probabilistic predictions could improve control results in principle, we leave the fine tuning of the specific pMPC control strategy to future investigations and more sophisticated control design strategies.

For a closer look at how the pMPC approach works to control observed and unobserved cells alike, Figs. 2.6A shows an example control input over time, and Fig. 2.6C shows the resulting trajectories over time for the observed cell (blue), the predicted probability distribution for unobserved cells (gray shading), and a representative unobserved cell (red). We reiterate that the controller has no direct knowledge of the red line. From the figure, it is clear that observed cell is well maintained near to its target value with low variability. Moreover, Fig. 2.6C shows that knowledge of the fluctuating input signal is sufficient to yield good predictions of the unobserved cell response (compare red line with dark gray shading), although as expected there are periods of poor predictions when the specific unobserved cell samples the higher or lower tail of the predicted distribution (e.g., at about 1900 minutes for the red curve in Fig. 2.6C). In addition to outperforming the PAC approach in terms of the overall cost function, the pMPC provides additional predictions for when the controller is effective, or when unobserved cells are more likely to escape from their intended phenotype. To illustrate this Figs. 2.6B shows the probability of observing fifteen or less protein molecules in the unobserved cell. When this probability exceeds a 90 percent threshold the control is expected to be effective, and the region is labeled as orange in Figs. 2.6B and Figs. 2.6C. Figs. 2.6D shows the marginal distributions averaged over all times, and Fig. 2.6E shows the marginal distribution only for the periods of time when the probabilistic model predicts its own effective control of the unobserved cells (orange regions). By focusing only on these times identified as successful by the controller, the cost function of the controller substantially decreases

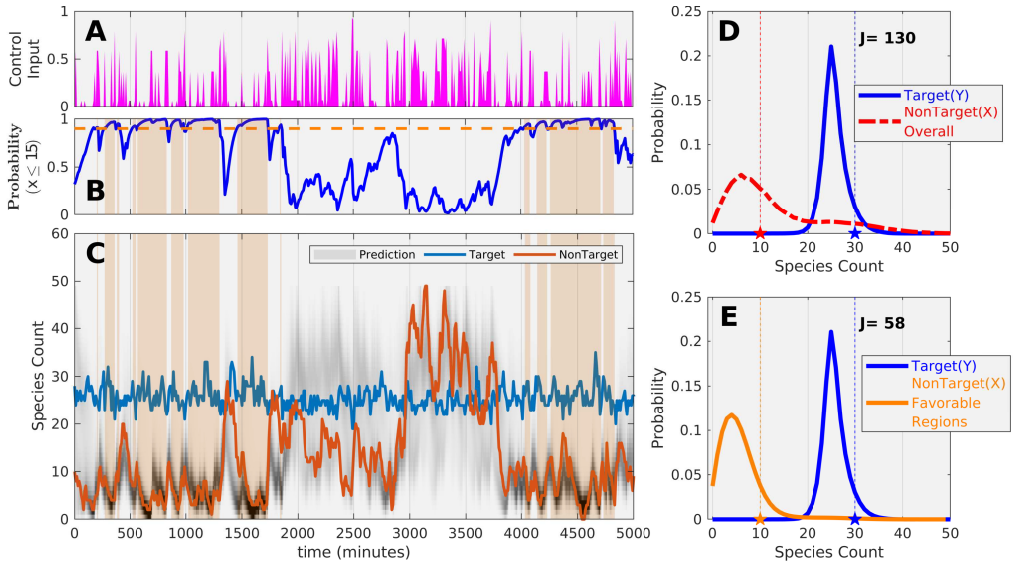

Figure 2.6: Control signal and response versus time for pMPC control. (A) Control signal generated by the pMPC controller. (B) Probability that unobserved cell is 15 or less versus time. When the probability is greater than 0.9 (orange line), then the system is considered to be in a "effective control" state. (C) Predicted transient distributions for unobserved cells (gray), observed cell (blue), and a single unobserved cell (red). Orange regions correspond to effective control times. (D) Time averaged performance of the control law in terms of marginal distributions for the observed call (blue) and unobserved cells (red). (E) Average of the pMPC performance, when considering only effective control periods.

from $J = 130$ to $J = 58$. These results suggest that predicted dynamic information about the unobserved cells can not only be used to improve the quality of the controller, but that the pMPC can also be used to self-assess when control is working well, and when it is not.

2.3.6 Controllers designed using simplified models can be effective to control more complicated processes with hidden mechanisms and dynamics.

We next ask how well could controllers designed using simplified stochastic models work when they are applied to control more complex systems that contain additional hidden states and which have unknown dynamics or time delays. To perform this analysis, we first account for the difference in meaning and units for the input signal $u'(\phi)$ used in the reduced auto-regulation models ($\mathcal{M}'_A$ in Fig. 2.1C) and its analog $u(\phi)$ used in the full auto-regulation model ($\mathcal{M}_A$ in Fig. 2.1D).

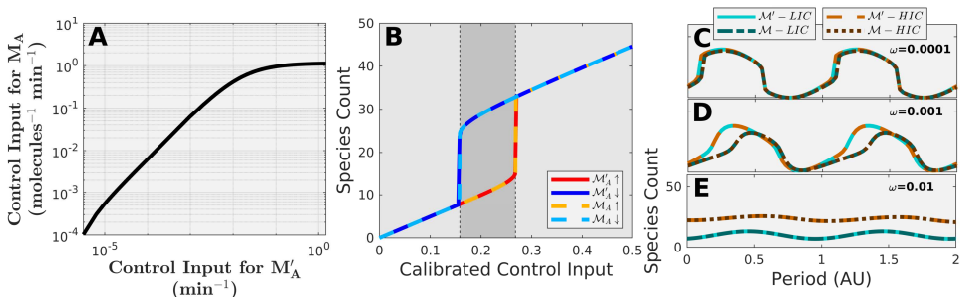

Figure 2.7: Calibration and use of controllers for use with a new, more complex model. (A) Calibration curve identified to match steady state ODE of simple and complex auto-regulation model ($\mathcal{M}'_A$ and $\mathcal{M}_A$ from Fig. 2.1C,D). (B) Steady state analyses show that the calibrated inputs result in similar hysteresis behavior for both $\mathcal{M}'_A$ and $\mathcal{M}_A$. (C) Input-output response analysis shows that $\mathcal{M}_A$ with calibrated inputs closely matches behavior of $\mathcal{M}'_A$ at ultra slow frequencies (0.0001 RPM), but (D) the more complex model begins to lag at slow frequencies (0.001 RPM). (E) At fast frequencies of 0.01 RPM, the complex auto-regulation model $\mathcal{M}_A$ is able to retain memory of its initial conditions and again exhibits similar phenomena compared to the simplified model.

By using steady state ODE analyses of $\mathcal{M}'_A$ and $\mathcal{M}_A$, we quantified the calibration curve to map inputs between the two models as shown in Fig. 2.7A. After calibration of the input signals, we verified that the full and reduced auto-regulation models result in similar bifurcation diagrams as shown in Fig. 2.7B. However, although calibration allows us to match both the quasi-steady (i.e., very slow) and fast fluctuating input responses of the two models (Figs. 2.7C,E, respectively), the temporal responses to slow input frequencies are qualitative and quantitatively different, as can be observed by the different input-to output time lags and amplitudes in Fig. 2.7D.

Having calibrated the controller for the full model to match the response of the simpler model, we then take the UAC, FAC, and PAC controllers from above and apply them directly (i.e., without any further tuning or optimization) to the full mechanistic model with auto-regulation. Despite the differences in temporal behaviors between the two models, the previously identified UAC and FAC controllers still work to break symmetry and drive both cells toward the correct differentiated phenotypes as shown in Fig. 2.8. We note that with further modifications, the control laws derived using the simplified model could certainly be improved for use in the more complex system. However, our primary goal was to explore how well designs made in one context should perform

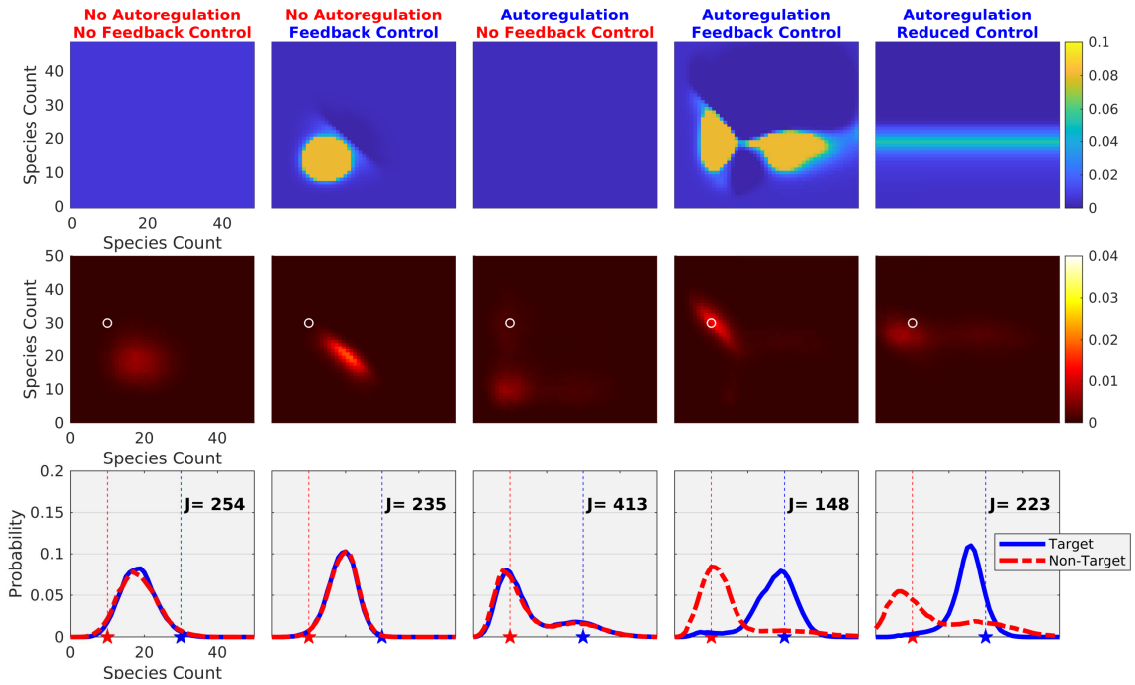

Figure 2.8: Full Models are paired with the calibrated controllers (top row) to solve for the joint probability distributions (middle row) and marginal probability distributions (bottom row). As before both autoregulation and feedback are needed to break symmetry and control fails if either of these are missing (rightmost three columns). The fully aware controller (FAC, fourth column) successfully works to control two cells with the complex dynamics, and the partially aware controller (PAC, rightmost column) successfully can control a single observed cell to one phenotype and an arbitrary number of unobserved cells to another different phenotype.

when used in another different context, and subsequent fine tuning for the complex model is left for future investigations.

Conclusion

The treatment of noise in synthetic biology has largely been centered around the management of noise as a nuisance property that needs to be mitigated or eliminated. Despite improvements to minimizing noise in bio-circuits, noise largely remains a fundamental physical limit due to the combination of very small cell sizes, where single molecular events have increased importance, and increasing complexity of synthetic circuits, where most dynamical influences are unknown or unmeasured. The results here show how a few increasingly common synthetic biology motifs–such

as optogenetic transcription factors, activatable polymerases and auto-regulation promoters–can in principle be combined to form regulatory modules and integrated with new external feedback controllers not only to mitigate intrinsic noise, but even to exploit that noise to achieve new multi-cellular behaviors.

Our work compliments that of previous efforts to develop single- and multiple-input-single-output (SISO and MISO) control for synthetic biology applications [91, 92] that have primarily sought to control one cell at a time or to control an entire population of cells to all reach the same phenotype. Specifically, we have demonstrated new types of optimizable single-input-multiple-output (SIMO) stochastic controllers that rely on the integration of noise, non-linear auto-regulation, and feedback to simultaneously control of multiple cells using a single chemical or optogenetic input. The first of these, the fully aware controller (FAC), assumes full knowledge of each individual cell's behavior and achieves the best control performance. The disadvantage of the FAC is that it requires knowledge of each individual cell (i.e., optical tracking and image processing analyses) and computationally intensive operations both to solve multi-dimensional chemical master equations and to search very high dimensional spaces for optimal controllers. However, a second partially aware controller (PAC) requires only the knowledge of a single cell of interest, yet the PAC can control that cell to one phenotype and drive all others to an alternate phenotype with an accuracy almost equal to the FAC. The advantage of the PAC is that it is very easy to implement and optimize as the dimension of the control law must only consider the single observed cell. The third controller introduces a probabilistic model predictive control (pMPC) strategy that computationally predicts the probability distribution of all non-observed cells based on integration of the chemical master equation under the known history of the applied input signal. Although we envision that similar control strategies may have applications in other fields, such as for autonomous vehicle or smart grid applications, to our knowledge, the proposed pMPC approach is the first example of a hybrid control strategy that predicts and exploits noise and feedback to simultaneously and differentially control multiple identical agents using a single control signal.

In addition to noise, one of the most important challenges in model-driven synthetic biology, is that many important regulatory mechanisms are currently unknown, even for simple biological systems. Similarly, very few parameters are known at any level of certainty, and most of these parameters vary from cell to cell or situation to situation. Moreover, it is already extremely computationally expensive to combine all known mechanisms into a single computational model, and although such models can be useful to reproduce a variety of biological behaviors [100, 101], such whole cell models are far too inefficient to enable the vast numbers of different simulations needed to optimize a design or control strategy. To circumvent these concerns, we demonstrated how a highly simplified phenomenological model could be used to design a controller that could be easily recalibrated using steady state dose-response measurements and then applied directly to a control a more complex system with hidden dynamics and with qualitatively and quantitatively different dynamic response features. Although it is common practice to use simple deterministic models to guide engineering design of modern complex devices, this demonstration in the context of stochastic single-cell processes suggests that there is also hope for similar applications of simple models in synthetic biology.

Although the potential of our computational results remains to be verified through independent experimental investigation, we believe that this numerical demonstration of the potential for a new control paradigm not only opens new possibilities for integrated “cyber organic" approaches in synthetic biology [23, 80–82], but could also offer insight into natural cellular differentiation processes where cellular states are sensed, and control signals are transmitted, by neighboring cells. For example, it has been suggested that stochastic fluctuations in expression lead embryonic stem cells to achieve substantial, and functionally relevant heterogeneity in Nanog expression, where transiently low Nanog expression cells are prone toward differentiation, whereas high Nanog expression cells are less likely to differentiate [102]. As such, it might be interesting to explore the possibility that temporally controlled fluctuations in Nanog transcription factors [103, 104] could selectively direct specific neighboring cells to differentiate while maintaining others in the stem state. Overall, we envision that advancing synthetic biology motifs, especially an increasing diversity of orthogo-

nal transcription factors and promoters [68, 105], improved live cell reporters [40, 106], and faster and more specific optogenetically controlled transcription factors inputs [82], will integrate synergistically with new probabilistic model predictive control analyses to improve future efforts to understand how noise, non-linearity, and feedback combine to drive cell fate decisions in applications ranging from synthetic biofuel and biomaterial production to developmental dynamics or regenerative medicine.

2.4 Methods

2.4.1 Definition of Models in Terms of Stoichiometries and Propensities

To introduce our numerical approaches, consider a generic cell regulatory process that contains N' distinct chemical species that interact with each other through M' different reactions. At any point in time, the current state of the process in a *single* cell can be described by an N'-element vector $\mathbf{x} = [x_1, \ldots, x_{N'}]^T \in \mathcal{X}$, where $\mathcal{X}$ denotes the set of all possible states (e.g., the nonnegative spaces $\mathbb{R}^{N'}_{\geq 0}$ for a continuous process or $\mathbb{Z}^{N'}_{\geq 0}$ for a discrete process). The definition of the full state $\mathbf{X}$ for *multiple* cells is easily concatenated to consider a set of N_c individual cells

$$\mathbf{X} = [x_1^{\text{cell } 1}, \ldots, x_{N'}^{\text{cell } 1}, \ldots, x_1^{\text{cell } N_c}, \ldots, x_{N'}^{\text{cell } N_c}]^T, \tag{2.14}$$

with an appropriate change to the total numbers of species ($N \equiv N_c N'$) and reactions ($M \equiv N_c M'$).

Under the assumption of a well-mixed spatial environment within each cell, one can define the dynamics of such a process by specifying the reaction stoichiometry vector and reaction rate for each μ^{th} reaction [107]. The *stoichiometry vector*, $\mathbf{s}_\mu \in \mathbb{Z}^N$ is the net integer change in molecules after exactly one event of the μ^{th} chemical reaction (i.e., $\mathbf{s}_\mu \equiv \mathbf{X}(\text{after } \mu^{\text{th}} \text{ reaction}) - \mathbf{X}(\text{before } \mu^{\text{th}} \text{ reaction})$). For continuous processes, the *reaction rate*, $f_\mu(\mathbf{X}, \mathbf{\Lambda}, u)$ is a scalar that defines the speed at which the μ^{th} reaction would be expected to occur given the current state $\mathbf{X}(t)$, fixed physical parameters $\mathbf{\Lambda}$ and time- or state-varying control parameter $u(\mathbf{X}, t)$. For

discrete stochastic chemical reactions, the reaction rate is replaced with a *propensity function* $w_\mu(\mathbf{X}, \mathbf{\Lambda}, u)dt$, which describes the probability that a single μ^{th} reaction would occur in the next infinitesimal time step of length dt given $\mathbf{X}(t)$, $\mathbf{\Lambda}$, and $u(\mathbf{X}, t)$. For reduced order models $\mathcal{M}'_U$ and $\mathcal{M}'_A$, we replace u with u' to denote the change in units needed for consistency with the model reduction.

2.4.2 ODE Representation of Models

Using these simple definitions, one can easily write an ordinary differential equation (ODE) to define a deterministic description of the process dynamics as:

$$\frac{d\mathbf{X}}{dt} = \sum_{\mu=1}^{M} \mathbf{s}_\mu f_\mu(\mathbf{X}, \mathbf{\Lambda}, \mathbf{u}), \tag{2.15}$$

$$= \mathbf{S}\mathbf{f}(\mathbf{X}, \mathbf{\Lambda}, u(\mathbf{X}, t)), \tag{2.16}$$

where $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_M] \in \mathbb{Z}^{N\times M}$ is the stoichiometry matrix, and $\mathbf{f}(\mathbf{X}, \mathbf{\Lambda}, u(\mathbf{X}, t)) = [f_1, \ldots, f_M]^T \in \mathbb{R}^M_{\geq 0}$ is the vector of non-negative reaction rates. For any given stoichiometry matrix $\mathbf{S}$, and reaction rate function vector, $\mathbf{f}(\mathbf{X}, \Lambda, u)$, the rate of change of $\mathbf{X}$ described by Eq. (2.15) can be integrated numerically to describe the system dynamics over time.

2.4.3 Discrete Stochastic Representation of Models

For discrete stochastic systems, the specification of the reaction stoichiometry and propensity functions is sufficient to generate individual trajectories of the process using Gillespie's Stochastic Simulation Algorithm (SSA, [30]). Alternatively, one can also use these two properties to uniquely define the Chemical Master Equation (CME, [34]) as:

$$\frac{d}{dt}P(\mathbf{X}) = -\sum_{\mu=1}^{M} w_\mu(\mathbf{X}(t), \mathbf{\Lambda}, u(\mathbf{X}, t))P(\mathbf{X}) + \sum_{\mu=1}^{M} w_\mu(\mathbf{X}-\mathbf{s}_\mu, \mathbf{\Lambda}, u(\mathbf{X}-\mathbf{s}, t))P(\mathbf{X}-\mathbf{s}_\mu). \tag{2.17}$$

For this discrete state description, one can always enumerate all possible states as $\{\mathbf{X}_1, \mathbf{X}_2, \ldots\} \equiv \mathcal{X}$ and define a probability mass vector as the similarly ordered probabilities, $\mathbf{P} \equiv [P(\mathbf{X}_1), P(\mathbf{X}_2), \ldots]^T$. Because the CME in Eq. (2.17) is linear in every term $P(\mathbf{X}_i)$, it is often written in matrix format:

$$\frac{d}{dt}\mathbf{P} = \mathbf{A}\mathbf{P}, \tag{2.18}$$

where the square matrix $\mathbf{A}$ is known as the *infinitesimal generator* and is defined directly from Eq. (2.17) as

$$A_{ij} = \begin{cases} -\sum_{\mu=1}^{M} w_\mu(\mathbf{X}_j, \mathbf{\Lambda}, u(\mathbf{X}_j, t)), & \text{for } i = j, \\ w_\mu(\mathbf{X}_j, \mathbf{\Lambda}, u(\mathbf{X}_j, t)), & \text{for } \mathbf{X}_i = \mathbf{X}_j + \mathbf{s}_\mu \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

We note that the summation of $\mu = 1$ to $M = N_c M'$ accounts for the increase in the number of possible reactions due to the existence of multiple cells. Because each term $\mathbf{X}_i$ refers to a specific enumerated state vector that is fixed in time, in the special case where $u \equiv u(\mathbf{X}_i)$ (i.e., where u depends only on the current state and does not depend *explicitly* on time), the matrix $\mathbf{A}$ is constant with respect to time. For convenience, we can define the control parameter in vector form $\mathbf{u} = [u(\mathbf{X}_1), u(\mathbf{X}_2), \ldots]$. The final CME model with control can be written in simple form by separating the infinitesimal generator into it basal and control induced components as:

$$\frac{d}{dt}\mathbf{P} = (\mathbf{A}_0 + \mathbf{B}\mathbf{u})\mathbf{P}. \tag{2.20}$$

In this formulation, although the dynamics of each identical cell was independent and uncoupled in the basal infinitesimal generator $\mathbf{A}_0$, the added infinitesimal generator from the control input, $\mathbf{Bu}$, can introduce coupling between cells. As an example, consider the fully aware controller (FAC) for two cells. The i^{th} state is written $\mathbf{X}_i = [x_{i1}, x_{i2}]^T$, and the control infinitesimal generator can

be written as:

$$[\mathbf{Bu}]_{ij}^{FAC} = \begin{cases} -2u^{FAC}(x_{i1}, x_{i2}), & \text{for } i = j, \\ u^{FAC}(x_{i1}, x_{i2}), & \text{for } \mathbf{X}_i = \mathbf{X}_j + [1, 0]^T, \\ u^{FAC}(x_{i1}, x_{i2}), & \text{for } \mathbf{X}_i = \mathbf{X}_j + [0, 1]^T, \\ 0, & \text{otherwise.} \end{cases} \tag{2.21}$$

Similarly, for the partially aware controller (FAC) for two cells, the control infinitesimal generator can be written as:

$$[\mathbf{Bu}]_{ij}^{PAC} = \begin{cases} -2u^{PAC}(x_{i1}), & \text{for } i = j, \\ u^{PAC}(x_{i1}), & \text{for } \mathbf{X}_i = \mathbf{X}_j + [1, 0]^T, \\ u^{PAC}(x_{i1}), & \text{for } \mathbf{X}_i = \mathbf{X}_j + [0, 1]^T, \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

As discussed in the main text, in either case, the coupling introduced by the control infinitesimal generator $[\mathbf{Bu}]^{FAC}$ or $[\mathbf{Bu}]^{PAC}$ is sufficient to break symmetry and encourage cells toward desired differential expression phenotypes.

2.4.4 Solution Scheme for Chemical Master Equation

To solve the CME in Eq. (2.20), we use the Finite State Projection (FSP [108]) approach, which truncates the allowable state space for every species and results in a finite dimensional ODE. However, it should be noted that the state space of a single arbitrary chemical species is given by the ordered set $[0, 1, ..., n-1]$ up to some truncation limit n. The state space of multiple species, is enumerated by forming a tuple of all possible species available, each up to a similar maximum number. For a system of N_c cells with N' chemically reacting species, where each species can range up to a maximum of $n-1$ copies per cell, the number of distinct states is $n^{N'N_c}$, which quickly becomes intractable when n, N_c, or N' is large. For this reason, model reductions, simplifications, or approximations are essential, especially when these models are to be used with millions of different parameter sets when searching for optimal control strategies. Further, it is

important to test and verify if control strategies designed and optimized using such simplified models will continue to be effective when applied to more general and more complex systems.

2.4.5 Fitting of models to data

Fits of the reduced model to experimental data was performed numerically by optimizing both the set of model parameters and the calibration variables in unison. Since the parameter fits of $\mathcal{M}'_U$ did not reveal a single set of unique parameters which fit data, the decay rate was calculated by hand by fitting the middle region of Fig. 2.2A and then fitting the parameters after fixing the protein decay rate. Fitting the $\mathcal{M}_U$ to experimental data using calibration was performed by hand since mathematical tools to fit data often yielded poor results by becoming stuck in local minima. These hand fits were also constrained such that the decay rate of the protein is the same decay rate in $\mathcal{M}'_U$.

2.5 Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM124747. The work reported here was partially supported by a National Science Foundation grant (DGE-1450032). Any opinions, findings, conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.6 Author Contribution

BM designed and acquired funding for the study. MPM and BM developed computational and theoretical tools. MPM performed all computational analyses. MPM wrote the manuscript. MPM and BM edited the manuscript.

2.7 Conflicts of Interest

None

2.8 Data Availability

All data and computational codes needed to reproduce the figures in this manuscript are freely available through the GitHub page: https://github.com/MunskyGroup/Michael_May_et_al_2021

Chapter 3

Stochastic Control of Dynamical Systems Under Variations of Scale and Parameter Uncertainty

[2] Previous research has focused to enable robust control performance despite the presence of noise, but understanding how controllers may exploit that noise remains incomplete. Motivated by Maxwells-Demon, we previously proposed a cellular control regime in which the exploitation of stochastic noise can break symmetry between and allow for specific control of multiple cells using a single input signal (i.e., single-input-multiple-output or SIMO control). The current work extends that analysis to include uncertain stochastic systems where system dynamics are are affected by time delays, intrinsic noises, and model uncertainty. We find that noise-exploiting controllers can remain highly effective despite coarse approximations to the model's scale or incorrect estimations or extrinsic noise in key model parameters, and these controllers can even retain performance under substantial observer or actuator time delays. We also demonstrate how SIMO controllers could drive multi-cell systems to follow different trajectories with different phases and frequencies. Together, these findings suggest that noise-exploiting control should be possible even in the practical case where models are aways approximate, where parameters are always uncertain, and where observations are corrupted by errors.

Keywords: Stochastic Control, Gene Regulation, Optogenetics

3.1 Introduction

Uncertain fluctuations, or 'noise,' is a common theme throughout many fields of engineering, and robust control is a frequent concern when attempting to control any human-made system such

[2]Chapter 3 is published on biorxiv as: M. P. May, B. Munsky, 'Exploiting Intrinsic Noise for Heterogeneous Cell Control Under Time Delays and Model Uncertainties', bioRxiv, 2023, https://doi.org/10.1101/2023.10.07.561335

As first author, MPM was the primary researcher on this paper and was responsible for all simulations and writing of the initial manuscript.

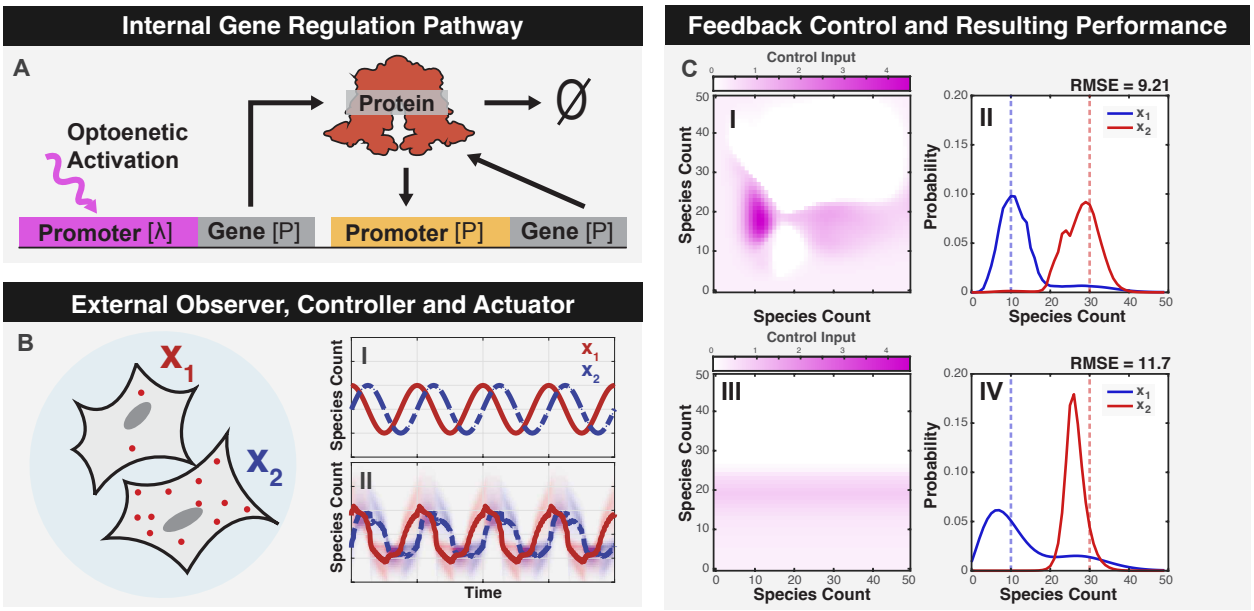

Figure 3.1: Single-input-multiple-output (SIMO) control of multiple cells using a single optogenetic input. (**A**) Schematic of the light-activated genetic system with auto-regulation. (**B**) Diagram of the stochastic SIMO control problem using two optogenetic cells sharing a single input. (**C**) Noise exploiting controllers were optimized to define a fully aware control input (I) and a partially aware control input (III), where the control signal (color scale) depends on the observed expression within the cell or cells (x- and y-axes). (II and IV) Corresponding steady state marginal distributions for the different cells (red and blue) under these controllers demonstrate a clear a break in symmetry. Dashed lines represent the control target objective. Control performance (RMSE) is show above each distribution.

as vehicles [109], chemical processes [110], or biology [21]. Noise arrises from many sources, occasionally from quantum events and thermally-induced fluctuations, but more commonly from unknown or uncharacterized physical processes, and as such, these fluctuations usually cannot be efficiently reproduced or predicted using purely physical or mechanistic models. Rather, statistical models are usually employed where key mechanisms or dynamics are subject to stochastic variations or inputs that act as a proxy for random or un-modeled fluctuations elsewhere in the system. In this context, much effort has been placed on enhancing the robustness of controlled processes, where uncertain or unpredictable variations in internal or external parameters are modeled as intrinsic or extrinsic noise [111]. For many macroscopic, human-engineered systems, the resulting system dynamics can be modeled effectively simply by combining a deterministic (i.e., noise-free) model with additive noise (usually assumed to be Gaussian under arguments based on the central limit theorem) [112], and most current approaches seek to control the system to minimize variations that result from the noisy inputs.

Analyses of noise and control theory are equally relevant to understand the basic biological processes of gene transcription regulation [113]or mRNA translation regulation [51]or to modify these processes for practical use [23, 114]. However, at this mesoscopic scale of cellular biology, where transcription factors compete to activate or deactivate individual genes or where mRNA or protein molecules are present at just a few copies (or none at all) per cell, the additive noise model is much less realistic. In this case, Brownian motion, discrete stochastic gene regulation, and noisy mRNA dynamics collectively generate a fundamentally stochastic environment that cells must effectively manage. At this scale, the order or timing of a single reaction event (e.g., the binding of a transcription factor to a promoter) can have dramatic consequences that could last for several cellular generations (e.g., the activation or repression of a gene that promotes unfettered cell growth and differentiation). The cell's drive towards homeostasis requires dealing with the inherently chaotic and noisy processes that reside within it, and despite these challenges, cells generally demonstrate strong capability to survive these noisy processes. When seeking to understand how such systems evolve or react, the central limit theorem and the Gaussian noise may not apply, and

a more detailed statistical analysis of probabilistic behavior is needed uncover hidden properties of cellular control mechanisms.

One emerging field that is particularly dependent on the integration of control and noise is synthetic biology, which aims to develop modular [66] and orthogonal [69] components to sense and manipulate [80] complex logical systems, enabling them to exhibit a wide range of advanced biological behaviors [4]. Advances in optogenetics have enhanced the ability to reliably actuate embedded systems within cells, offering the potential to exert precise temporal and spatial control on cellular components [23, 80, 82, 115]. These developments have facilitated computer-programmable regulation of cellular protein production through external optogenetic inputs and smart microscope techniques [91, 94, 116]. These digital-synthetic actuators enable fine-tuned, computer-modulated control of cellular systems, previously unattainable, with faster response times compared to chemical diffusion [23, 81] . Classical and modern control methods like PID control and model predictive control have been implemented in such systems [117]to control synthetic systems to different stable points.

It has recently been shown that new control techniques that leverage the complete probability distribution information of the system could actually harness the noise of single-cell gene regulation to achieve more complicated control objectives. For example, inspired by the genetic toggle switch from Kobayashi et al. [118], Szymanska et al. [95] showed that noise could be exploited to achieve independent control of multiple cells using a single input, even despite uncertain parameters or time delays due to maturation of fluorescent proteins or limited observation of the regulatory proteins. In May et al. [119], we identified a simplified stochastic model to reproduce data measured in Baumschlager, et al. [23] for the expression from a transcription promoter under the optogenetic control of a UV-activated T7 polymerase (see model in Figure 3.1A, top promoter). We then proposed the addition of a positive auto-regulation (Figure 3.1A, bottom promoter) to help maintain an elevated expression phenotype in the presence of UV excitation, and we demonstrated how a Single-Input-Multiple-Output (SIMO) multicellular controller could control multiple cells to arbitrary phenotypes using only a single input.

This paper extends the analysis of the SIMO multicellular control problem by examining the impact of model uncertainties and fluctuating control objectives on the control performance. These uncertainties include coarse-grained approximations of the system dynamics, errors or extrinsic variations in the system parameter, and time delays between the observation of the cellular dynamics and actuation of the control process. In the following 'Methods' we introduce our formulation of the chemical master equation (CME) to analyze the discrete stochastic distribution of cellular responses; we define multiple controllers and demonstrate the computation of their effects on cell dynamics; and we show how the control law can be optimized to improve performance. In the 'Results' section, we explore the how model approximations, parameter inaccuracies, and time delays affect control performance, and we demonstrate a simple scheme for controlling cells to track a dynamically changing reference signal. Using discrete stochastic models based on the chemical master equation, we demonstrate that combining biochemical noise, nonlinear auto-regulation, and a single optogenetic feedback could control two genetically identical cells with different initial conditions to follow different desired trajectories at different frequencies and phases.

3.2 Methods

In May et al. [119], we developed two models for the description of an optogenetically controlled gene expression system. These first model consisted of six species to describe the light-activated association of two T7 split domains (species 1 and 2) which combine to form an active T7 polymerase (species 3) under optogenetic excitation. The active polymerase could then associate with inactive T7 promoters (species 4), resulting in the formation of an active allele (species 5) that could then transcribe and translated to produce the desired protein product (species 6). The second, much simpler, single-species model was developed by assuming quasi-steady equilibrium for the first five species. Both models were independently parameterized using the same experimental data from Baumschlager et al. [23]. Furthermore, an extension was made to each model to incorporate a secondary self-activated promoter-gene construct, where the expression rate was determined by a Hill function 3.1. Through simulations, we showed that a feedback con-

trol law could be designed to force multiple cells to different and individually chosen equilibrium states using a single optogenetic control signal. We also showed that when this control law was parametrized using the simple model, it could be used effectively to control the behavior of the more complicated system, thus demonstrating that control performance could remain high despite inaccuracies in the model. In the current work, we adopt the simpler of the two models and extend our analyses to consider the effects of additional model inaccuracies, including coarse-grained model approximations, parameter errors, and time delays, and we also explore the possibility that a single optogenetic control signal could drive the system to track temporally-changing reference signals. Section 3.2.1 introduces the model; Section 3.2.2 develops a Master Equation description of the models probabilistic dynamics; Section 3.2.3 defines a control objective and optimizes that metric to obtain a baseline control law; and Sections 3.2.4 and 3.2.5 introduce uncertainties into the model related to the granularity of the model approximation or the introduction of time delays, respectively.

3.2.1 Model

To assess the impact of model approximations on the implementation of noise-enhanced control strategies, we begin with the one-species model proposed by May et al. [119] (Figure. 3.1A). This model comprises two reactions for production and degradation of the key protein. The nonlinear, UV-dependent production is defined by the following equation:

$$\nu_1(x,t) = \kappa \frac{x^\eta}{x^\eta + \beta^\eta} + k_0 + u(UV(t)), \tag{3.1}$$

where x is the instantaneous protein level; κ is the maximum strength of the auto-regulation promoter; β is the concentration at which auto-regulation promoter reaches its half maximal strength; η is the cooperativity in the auto-regulation promoter; k_0 is the leakage rate from both promoters; and $u(UV(t))$ is the UV-dependent strength of the T7 promoter. Feedback enables the external modulation of the light input using the state of the system, thereby controlling the T7 promoter strength as a function of state rather than time and eliminating the explicit time dependance in

Table 3.1: Baseline Model Parameters

Parameter	Value	Units	Meaning
κ	0.406	1/min	maximal production rate
β	20.0	molecules	concentration at half-max activation
η	8.00	unitless	cooperativity factor
k_0	0.0001	1/min	promote leakage rate
γ	0.0203	1/min	degradation/dilution rate
$u(t)$	variable	1/min	applied control signal

$u(UV(t))$. Protein degradation is assumed to be a first order process with rate γ:

$$\nu_2(x) = \gamma x. \tag{3.2}$$

All baseline parameters describing the auto-regulation promoter, κ, η, β, and k_0, and the degradation rate γ are presented in Table 1 [119] and are fixed throughout the current study.

3.2.2 Stochastic analyses of the model

To describe the discrete stochastic behavior of the above model for a population of N_c cells, we define the current state of the system as the tuple of the non-negative numbers of proteins in each cell: $\mathbf{X}_i = [x_1, x_2, \ldots, x_{N_\mathrm{c}}]_i \in \mathbb{Z}_{\geq 0}$, where the index i denotes the enumeration of the state within the countably infinite set of all possible states, i.e., $\mathbf{X}_i \in \mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots\}$. The stoichiometry vector, $\mathbf{s}_\mu$, for reaction number μ is then defined as the change in state following that reaction event (e.g., $\mathbf{X}_i \rightarrow \mathbf{X}_i + \mathbf{s}_\mu$). Specifically, the $2N_c$ possible reactions are defined in pairs corresponding to production ($\mu \in \{1, 3, 5, \ldots\}$) and degradation ($\mu \in \{2, 4, 6, \ldots\}$) as:

$$\begin{aligned}
\mathbf{s}_1 &= \mathbf{e}_1,\ \mathbf{s}_2 = -\mathbf{e}_1, \\
\mathbf{s}_3 &= \mathbf{e}_2,\ \mathbf{s}_4 = -\mathbf{e}_2, \\
&\vdots \\
\mathbf{s}_{2\mathrm{N_c}-1} &= \mathbf{e}_{\mathrm{N_c}},\ \mathbf{s}_{2\mathrm{N_c}} = -\mathbf{e}_{\mathrm{N_c}},
\end{aligned} \tag{3.3}$$

where each $\mathbf{e}_i \in \mathbb{Z}^{\mathrm{N_c}}$ is is a Euclidean vector (i.e., unity for the i^{th} entry and otherwise zero). The corresponding propensity functions are:

$$\begin{aligned} w_1(\mathbf{X}) &= u(\mathbf{X},t) + \kappa \tfrac{x_1^\eta}{x_1^\eta + \beta^\eta} + k_0,\ w_2(\mathbf{X}) = \gamma x_1, \\ w_3(\mathbf{X}) &= u(\mathbf{X},t) + \kappa \tfrac{x_2^\eta}{x_2^\eta + \beta^\eta} + k_0,\ w_4(\mathbf{X}) = \gamma x_2, \\ &\vdots \\ w_{2\mathrm{N_c}-1}(\mathbf{X}) &= u(\mathbf{X},t) + \kappa \tfrac{x_N^\eta}{x_N^\eta + \beta^\eta} + k_0,\ w_{2\mathrm{N_c}}(\mathbf{X}) = \gamma x_N. \end{aligned} \tag{3.4}$$

These definitions of the stoichiometry and propensity functions allow us to implement the Gillespie stochastic simulation algorithm (SSA) [30, 107] to generate representative trajectories of the stochastic process. At each step, two random numbers are generated to determine the time and the type of the next reaction. Given the current state $\mathbf{X}$, the time until the next reaction is distributed according to an exponential random with rate parameter equal to the inverse of the sum of the propensity functions:

$$Pr(\delta t = \tau) = \sum w_i(\mathbf{X}) \exp\left(- \sum \mu = 1^{2N_c - 1} w_\mu(\mathbf{X}) \tau \right), \tag{3.5}$$

and an instance of this random variable, δt, can be sampled from this distribution using the expression:

$$\delta t = -\log\left(r_1 \sum_{\mu=1}^{2N_c-1} w_\mu(\mathbf{X}) \right), \tag{3.6}$$

where r_1 is a uniform random variable between zero and one. The probability for the specific individual reaction R_k to fire from all possible reactions is given by:

$$Pr(R_k) = \frac{w_k}{\sum w_{\mu=1}^{2N_c-1}(\mathbf{X})}, \tag{3.7}$$

and which specific reaction that fires at time $t + \tau$ is a categorical random variable pulled from $Pr(R_k)$. The SSA is then simulated by stepping through time one reaction at a time and updating the state of the system by adding the stoichiometry of the reaction to the state of the system.

However, a more direct analysis of the chemical master equation (CME) is necessary to quantify performance and optimize the controller design. The high dimensional CME is a linear ODE that describes the time-dependent changes in probability mass of all possible states. Using the specified reaction propensities and stoichiometries, the CME can be expressed as:

$$\frac{d}{dt}P(\mathbf{X}_i) = \sum_{\mu=1}^{2\mathrm{N_c}} \left(-w_\mu(\mathbf{X}_i)P(\mathbf{X}_i) + w_\mu(\mathbf{X}_i - \mathbf{s}_\mu)P(\mathbf{X}_i - \mathbf{s}_\mu)\right). \tag{3.8}$$

For convenience, the CME can also be formulated more compactly in matrix-vector form as:

$$\frac{d}{dt}\mathbf{P} = (\mathbf{A}_0 + \mathbf{B}\mathbf{u}^{\mathcal{C}})\mathbf{P}, \tag{3.9}$$

where $\mathbf{P} = [P(\mathbf{X}_1), P(\mathbf{X}_2), \ldots]^T$ is the enumerated probability mass vector for all possible states of the system; $\mathbf{A}_0$ is the infinitesimal generator of the stochastic process due to the autoregulation promoter and degradation events; $\mathbf{u}^{\mathcal{C}} = [u^{\mathcal{C}}(\mathbf{X}_1), u^{\mathcal{C}}(\mathbf{X}_2), \ldots]^T$ is the collection of control inputs associated with each state; and $\mathbf{B}\mathbf{u}^{\mathcal{C}}$ is the contribution that these control inputs make to the infinitesimal generator when included into the feedback process.

More specifically, the zero-control infinitesimal generator, $\mathbf{A}_0$, is constructed according to:

$$[\mathbf{A}_0]_{ij} = \begin{cases} -\sum_{\mu=1}^{2N_\mathrm{c}} w_\mu(\mathbf{X}_j), & \text{for } i = j, \\ w_\mu(\mathbf{X}_j), & \text{for } \mathbf{X}_i = \mathbf{X}_j + \mathbf{s}_\mu \\ 0, & \text{otherwise,} \end{cases} \tag{3.10}$$

and the feedback infinitesimal generator, $\mathbf{Bu}^{\mathcal{C}}$, of the controller is constructed according to

$$[\mathbf{Bu}^{\mathcal{C}}]_{ij} = \begin{cases} -N_{\mathrm{c}} u^{\mathcal{C}}(\mathbf{X}_j), & \text{for } i = j \\ u^{\mathcal{C}}(\mathbf{X}_j), & \text{for } \mathbf{X}_i = \mathbf{X}_j + \mathbf{e}_{i_{\mathrm{c}}}, \text{ and } i_{\mathrm{c}} = 1, \ldots, N_{\mathrm{c}} \\ 0, & \text{otherwise,} \end{cases}, \tag{3.11}$$

where $u^{\mathcal{C}}(\mathbf{X}_j)$ is the specification of the controller, $\mathcal{C}$, in terms of the current instantaneous state, or its partial observations.

For a given controller, the equilibrium distribution of the system ($\mathbf{P}^*$) can be found by solving Eq. 3.9 and is given by:

$$\mathbf{P}^* = \mathrm{null}(\mathbf{A} + \mathbf{Bu}^{\mathcal{C}}). \tag{3.12}$$

In principle, the master equation in Eq. 3.9 could contain an uncountably infinite number of states, and therefore the exact solution as well as the null vector in Eq. 3.12 may not be computable exactly. To address this issue, we first truncate the system at a finite number for each species and then apply a reflecting boundary condition, resulting in a finite dimensional master equation. To assess the time interval over which this truncation is valid, we also solve the Finite State Projection [108] for the same truncation, which allows us to compute an upper bound on the truncation error as a function of time.

3.2.3 Quantification and optimization of control performance

In the SIMO control of stochastic processes, a single input is applied simultaneously to all systems at once, and therefore a control signal that acts beneficially on one cell may destabilize other cells. An effective controller must strike a balance among the desired behaviors of all cells in the system. To quantify overall performance success, we define the *steady state performance error*, J, as the expected steady state Euclidean distance of the process from the specified target state, $\mathbf{T}$:

$$J = \mathbb{E}\{|\mathbf{X} - \mathbf{T}|_2\}. \tag{3.13}$$

The squared score is easily calculated by applying a linear operator to the steady state probability distribution $\mathbf{P}^*$ (Eq. 3.12) as follows:

$$\begin{aligned} J^2 &= \lim_{t\to\infty} \mathbb{E}\{|\mathbf{X}(t) - \mathbf{T}|_2^2\}, \\ &= \sum_{i_1,i_2,\ldots} P^*(x_1 = i_1, x_2 = i_2, \ldots) \left[(i - T_1)^2 + (j - T_2)^2 + \ldots\right], \\ &= \mathbf{C}\mathbf{P}^*, \end{aligned} \tag{3.14}$$

where $\mathbf{C}$ is simply a vector that contains the squared Euclidian distance of each state from the specified target $\mathbf{T}$, i.e., $C_i = |\mathbf{X}_i - \mathbf{T}|_2^2$. As a result of this calculation, J is a non-negative scalar that is zero only if $\mathbf{P}^*$ is a delta distribution located exactly at the target vector $\mathbf{T}$.

We consider two controller designs: the fully aware controller ($\mathbf{u}^{FAC}$) that bases its control signal on simultaneous protein count observations from both cells, and the partially aware controller ($\mathbf{u}^{PAC}$) that relies only on observations from a single cell:

$$\mathbf{u}^{FAC} = u(x_1, x_2), \tag{3.15}$$

$$\mathbf{u}^{PAC} = u(x_2), \tag{3.16}$$

where x_1 and x_2 are discrete integers greater than or equal to zero that represent the instantaneous number of proteins in cell one and cell two. Despite their differences in their observation data, both the FAC and the simpler PAC are optimized to achieve the same goal, namely to drive both cells to their respective set points.

Optimization of $\mathbf{u}^{FAC}$ and $\mathbf{u}^{PAC}$ were performed using a gradient descent method to minimize J. Since the square root is a monotonically increasing function, minimizing J^2 results in the same control as would minimizing J directly. Therefore, we calculate the negative gradient $-\frac{d(J^2)}{d\mathbf{u}^{(\cdot)}}$ and adjust parameters a small step $d\mathbf{u}^{(\cdot)}$ in that direction. For example, the calculation of the gradient

for the FAC controller is given by:

$$\frac{d(J^2)}{d\mathbf{u}^{FAC}} = C\frac{dP}{d\mathbf{u}^{FAC}}, \tag{3.17}$$

where $\frac{dP}{d\mathbf{u}^{FAC}}$ can be solved using general minimized residual calculation of

$$A\frac{dP}{d\mathbf{u}^{FAC}} = B\mathbf{u}^{FAC}. \tag{3.18}$$

In [119], controllers were optimized to minimize J^2 for a single target state $\mathbf{T} = [10, 30]$, but in this work they are extend to different arbitrary target points.

3.2.4 Scaling for system granularity

In realistic applications, models are never exact, but are often chosen as simplifications of known processes. For example, when analyzing discrete stochastic chemical kinetics, it is common to project the CME onto lower-dimensional spaces using finite state projection [108], time scale separations [120], Krylov subspaces [121], principle orthogonal decompositions [56], or other coarse meshes [18, 122]. Similarly, measurements are also always inexact and in many cases may only provide information about relative changes – for example, although fluorescent proteins usually cannot be counted exactly, one may reasonably assume that a cell's total fluorescence intensity varies linearly with the fluorescent protein concentration. To explore how mismatches in the assumed system scale (e.g., arising from model approximations or relative measurements) affect the controllability of the cellular process, we define a granularity parameter ($\alpha = M'/M$) that linearly scales each species' population to increase ($\alpha > 1$) or decrease ($\alpha < 1$), while maintaining the dynamics and general behavior of the model. To apply this granularity parameter, we assume that each propensity function, w_μ (Eqns. 3.4) is rescaled to a different level of discreteness by substituting

$$w'(\mathbf{X}) = w(\mathbf{X}/\alpha). \tag{3.19}$$

For example, the production and degradation of protein in cell one would become:

$$\begin{aligned} w_1'(\mathbf{X}) &= u(\mathbf{X}/\alpha, t) + \kappa \tfrac{(x_1/\alpha)^\eta}{(x_1/\alpha)^\eta + \beta^\eta} + k_0, \\ w_2'(\mathbf{X}) &= \gamma x_1/\alpha. \end{aligned} \tag{3.20}$$

We note that in order to reuse a control law that has been defined for one level granularity and apply it to a system at another level of granularity, the inputs to the controller must also be scaled by $1/\alpha$ before computing the control level assigned to the current state, e.g., $u^{\mathcal{C}} = u^{\mathcal{C}}(\mathbf{X}/\alpha)$. Because identification of the original control formulation, $\mathbf{u}^{FAC}(x_1, x_2)$ and $\mathbf{u}^{PAC}(x_1)$ only considered integer values for (x_1, x_2), control signal values at fractional state values after rescaling $(x_1/\alpha, x_2/\alpha)$ are calculated using 2D cubic interpolation from the control values at the nearest integer state values. Finally, to provide a consistent metric for relative scoring, the definition of the performance score is also adjusted according to scale magnitudes. For example, in the two cell system the new steady state performance error would become:

$$\begin{aligned} (J^2)' &= \textstyle\sum_{i=0}^{M'} \sum_{j=0}^{M'} P^*(x_1 = i, x_2 = j)((i/\alpha - \mathbf{T}_1)^2 + (j/\alpha - \mathbf{T}_2)^2), \\ &= \mathbf{C}'\mathbf{P}^*. \end{aligned} \tag{3.21}$$

We reiterate that the system parameters and control law were defined and fixed using the the base granularity ($\alpha = 1$), and to simulate a practical application where scales may be unknown or variable, these are not recomputed or refit upon changing the system granularity.

3.2.5 Observation and actuation time delays

Delays are inherent to any realistic control system, and in this case delays would be expected to arise due to the time needed for various biochemical reactions such as the formation of complete polymerases, activation of promoters, transcription and transport of mature mRNA, and the translation and maturation of protein [123]. Additional delays would also arise from data analysis, decision making, and actuation dynamics. To investigate the effects of observation or actuation time

delays on control performance, we devised a simple time-delay stochastic simulation algorithm. This algorithm records the state history after each stochastic reaction, enabling reconstruction of the population history of the species and the time delayed control input propensity. Using this information, the time-delayed control signal at time t can be specified as:

$$u_\tau(t) = \begin{cases} 0, & \text{for } t \leq \tau, \\ u^{\mathcal{C}}(x_1(t-\tau), x_2(t-\tau)), & \text{for } t > \tau, \end{cases} \tag{3.22}$$

where τ is the time delay between observation and actuation, $u^{\mathcal{C}}$ is the previously optimized control law (e.g., $\mathbf{u}^{\mathrm{FAC}}$ or $\mathbf{u}^{\mathrm{PAC}}$), and x gathered from the state history. We note that the time delay stochastic process was only simulated using the SSA because to our knowledge an appropriate direct FSP/CME integration procedure has not yet been developed. The Extrande method [97] was used to update the control input at an average frequency of 50 updates per minute, far exceeding the dynamics of the system.

3.2.6 Tracking Time-Varying Trajectories

Because optimizing the control law for a static set point as described in Section 3.2.3 requires differentiation of the CME (Eq. 3.9) with respect to the control signal at each state, this calculation is approaching the limits of current feasibility. Extending these calculations to optimize controllers for a dynamically moving set point is much harder and would likely require intractable numerical simulation or development of new mathematical approaches that are beyond the scope of the current study. To circumvent this challenge, we instead propose a simple alternative in which the controller sweep though a piecewise constant set of controllers each designed for a specific static target point along the desired trajectory.

Our goal is to control the system to follow a specific target trajectory, $\mathbf{T}(t)$. We choose a discrete set of K target points along this trajectory:

$$\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_K\}, \tag{3.23}$$

and for each individual target $\mathbf{T}_i$, we optimize to find a corresponding controller, $\mathbf{u}_i = \mathbf{u}(\mathbf{T}_i)$. Finally, to implement the control at any given time, t, we find the index of the nearest precomputed target state:

$$\hat{i}(t) = \text{argmin}_i \left|\mathbf{T}(t) - \mathbf{T}_i\right|_2, \tag{3.24}$$

and assign that controller. Under this piecewise constant controller law, the full time-varying CME becomes

$$\frac{d}{dt}\mathbf{P} = \mathbf{A} + \mathbf{B}\mathbf{u}_{\hat{i}(t)}(\mathbf{x}). \tag{3.25}$$

3.3 Results

As developed in the Methods section, we focus on a SIMO controller where *all cells receive the same input* at every instant of time. For this, the control signal depends upon the observed state, e.g., $u(t) = \mathbf{u}(\mathbf{x})$ as developed above and illustrated in Fig. 3.1(I,III). These controllers have been optimized to break symmetry so that multiple cells can be controlled to different targets as illustrated in Fig. 3.1(II,IV), and we quantify performance according to the RMSE error introduced in Eq. 3.14. To explore how such controllers may perform in realistic settings where the models are approximate and parameter are unknown or extrinsically variable, we now fix the parameters of those controllers and explore performance robustness to different types of model uncertainties, including parameter errors or variations (Section 3.3.1), incorrect assumptions on system scales (Section 3.3.2) and time delays (Section 3.3.3). Finally, in Section 3.3.4, we extend the control analysis to consider the performance of the controller for tracking variable reference signals with different frequencies and phases.

3.3.1 Stochastic SIMO Optogenetic Control can Remain Effective Despite Small Parameter Errors or Extrinsic Uncertainties

Real stochastic processes always have unknown or uncertain mechanisms and parameter values. Although model structures and parameter estimates can often be obtained through fitting to training data, these estimates will never be perfect due to unavoidable measurement errors or lim-

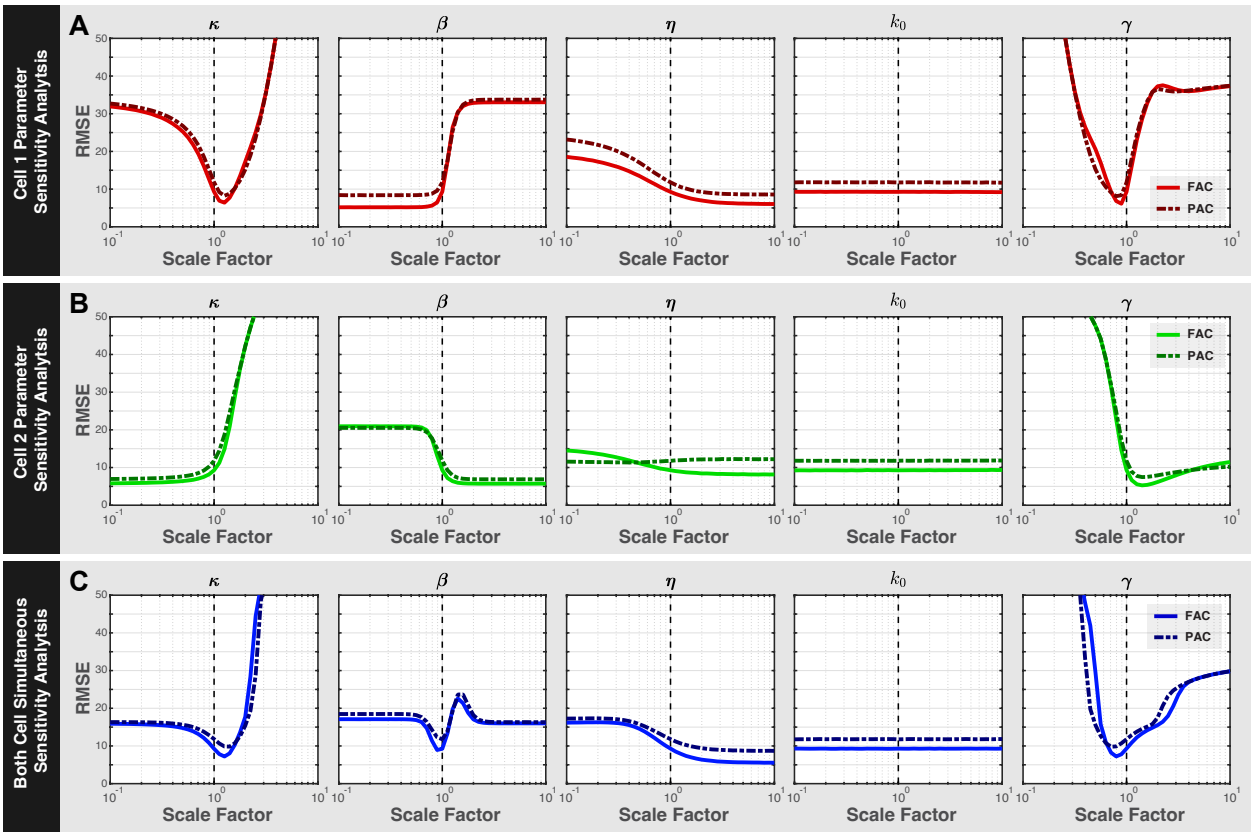

Figure 3.2: Parameter sweeps using the FAC and PAC show a broad range of control performance in cell 1 (**A**), in cell 2 (**B**), and in both cells (**C**). Columns show each parameter in the model, rows show the cell which has its parameter perturbed.

ited data sets. Even with plentiful and precise training data, model and parameter uncertainties are inevitable due to the fact that even genetically identical single-cells exhibit heterogeneity in parameters due to extrinsic variations. To asses the control performance under such parameter errors, we performed sensitivity analysis on each model parameter for each cell, then on both cells simultaneously. For each parameter, $\mathbf{u}^{\mathrm{FAC}}$ (solid lines) and $\mathbf{u}^{\mathrm{PAC}}$(dashed lines) control performance was examined across a parameter perturbation range from one-tenth to ten-fold of the original values. Figure 3.2A shows the results for these parameter sweeps when a single parameter in Cell 1 is varied (all other parameters are fixed), and Figure. 3.2B shows the control performance when a single parameter in Cell 2 is varied. Finally, 3.2C shows the performance change when a single parameter is changed simultaneously in both Cell 1 and Cell 2 at the same time.

Modifications of parameters were found to produce a broad range of effects. For example, increasing β in Cell 1 quickly worsens performance while increasing β in Cell 2 improves performance (compare second column in Figure. 3.2A and 3.2B). In some cases, the effects on performance are not monotonic; for example, increasing κ in Cell 1 (Figure. 3.2A, leftmost column) would be highly advantageous up to a limit after which the control performance degrades rapidly. In other cases (such as for the promoter leakage rate, k_0), the effect of parameter perturbations on performance is insignificant even for relatively large (s=10) perturbations, suggesting that this parameter is not important to control performance. Figure 3.2C shows that even when parameters of both Cell 1 and Cell 2 are jointly changed, these changes could also improve or detract from control performance. In particular, the analysis shows that control performance could be improved by modifying the system design either to increase the auto-regulation promoter strength (κ) or its cooperativity (η) or to decrease the promoter binding constant (β) or the protein degradation rate (γ). We note that co-optimization of both the system and the controller law would allow for further improvements to the performance.

3.3.2 Controllers trained using one assumed level of granularity can remain effective at other levels of granularity.

We next explored how the granularity of the system would affect the differential control of the multi-cell system. Specifically, we define the level of granularity using the parameter α as described in Section 3.2.4 and which affects the propensity functions and scoring according to Eqns. 3.19-3.21. Effectively, larger α corresponds to systems where larger numbers of individual molecules are needed to achieve the same concentration (e.g., larger volumes), while smaller α corresponds to situations where smaller population sizes can achieve that concentration (e.g., smaller volumes). We previously optimized the controllers $\mathbf{u}^{\mathrm{FAC}}$ and $\mathbf{u}^{\mathrm{PAC}}$ based on a the default assumption that $\alpha = 1$, and we wished to know what would be the consequences if this same controller were to be applied to a system that has a different granularity.

To explore this tradeoff, control performance scores for the $\mathbf{u}^{\mathrm{FAC}}$ and $\mathbf{u}^{\mathrm{PAC}}$ controllers were calculated at different levels of granularity (α) between 0.2 and 2.0. Figure 3.3 (A - F) shows the joint probability distributions (left plots) and marginal probability distributions (right plots) of the system at a low granularity ($\alpha = 0.2$), the original granularity ($\alpha = 1.0$), and at an increased granularity ($\alpha = 2.0$).

Figure 3.3G shows the trend of the performance versus alpha for both the FAC (solid cyan line) and the PAC (solid magenta line) controller. This improvement in performance appears to approach a small value as the granularity goes to infinity, but since the size of the FSP increases with the square of the system size, systems much larger than α=2 (where $\mathbf{A}_0 \in \mathbb{R}^{10^4 \times 10^4}$) become more difficult to calculate using master equation techniques. To bypass this limit in the FSP, sixteen SSA simulations were used to sample the CME of a system with a much larger volume of α=100. Each SSA was run for 5×10^7 minutes and only the last 4×10^7 minutes were sampled to estimate the stationary distribution and calculate the performance score. The performance score estimates of this high granularity SSA using the FAC and PAC were 2.03 and 5.24 respectively, which are plotted as dashed lines in Figure. 3.3G. Although it is unclear if further performance improvements could be obtained with further increases to the system scale, for all cases considered

so far, we found that both controllers monotonically improved with increased α and that $\mathbf{u}^{\mathrm{FAC}}$ always outperforms the $\mathbf{u}^{\mathrm{PAC}}$.

The effect that granularity has on the control performance depends on two competing phenomena: First, because the controller was optimized for one level of granularity ($\alpha = 1$), one might expect that the controller would become worse if the granularity were incorrect. However, at large granularity, we find that the opposite is true - the control performance actually improves when applied to an incorrect model. The reason for this surprising result is that relative amplitude of stochastic fluctuations (i.e., the standard deviation divided by the mean) in a chemical process decreases with the inverse square root of the process scale [124]. In other words, the process becomes more predictable and therefore more controllable. At the extreme as the system size increases, the dynamics converge towards a deterministic process, except for certain exceptional initial conditions lying on manifolds that would separate different steady state behaviors [125].

However, this improvement in the steady state performance does not come without a cost. Although higher granularity reduces noise and makes it easier to maintain desired states once they have been achieved, noise is necessary to break symmetry between the two cells' dynamics in order to achieve those states in the first place. This tradeoff is illustrated in Fig. 3.3(H), which plots the control performance over time after changing the control goal to exchange the low- and high-target cells with one another. From the figure, we can see although steady state control performance improves for higher values of α, the time taken to reach that steady state performance increases with α, suggesting that larger-volume system may become more susceptible to times delays or less able to track variable reference trajectories.

3.3.3 Heterogeneous control can remain effective despite moderate time delays

In general, feedback control can only be effective if one can quickly make measurements, compute adjustments to the control signal, and implement the needed changes within an appropriate amount of time relative to the characteristic timescale of the system. As the time required for

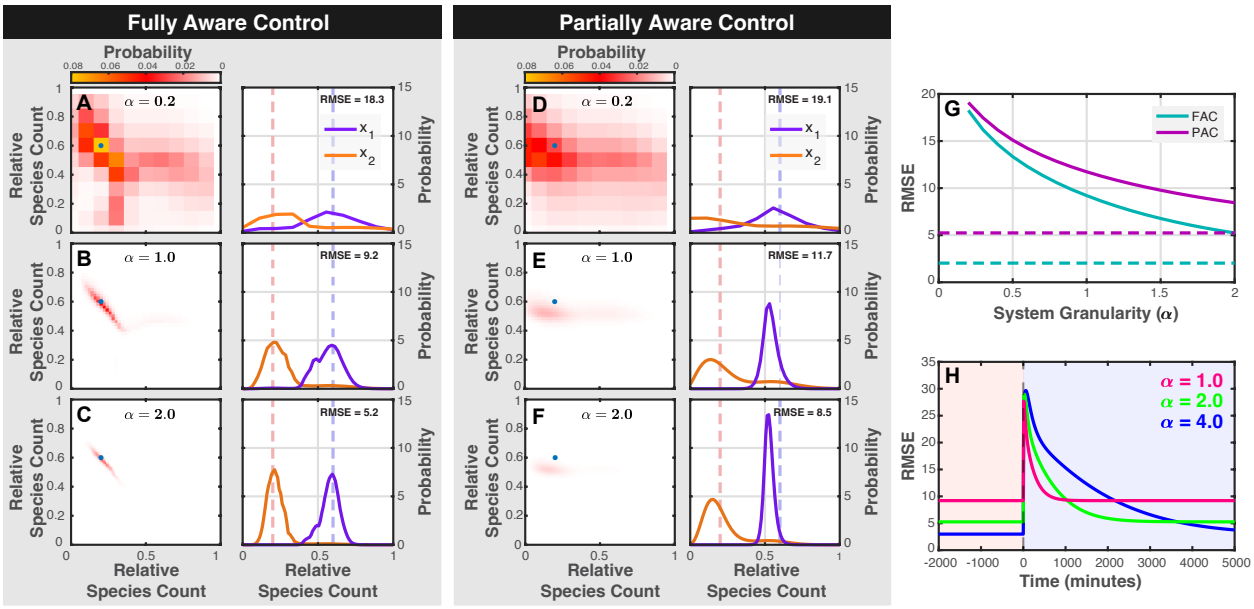

Figure 3.3: Systems with increased granularity are less noisy and have better control performance. (A-C) Joint (left) and marginal (right) distributions for the FAC shows increased control performance and tighter distributions as α increases from 0.2 (top row) to 2.0 (bottom row) . (D-F) Joint (left) and marginal (left) distributions using the PAC controller at different levels of granularity.

any of these steps increases, control performance will be degraded, perhaps even leading to large fluctuations or instability. To explore how time delays affect the noise-enhanced controllers $\mathbf{u}^{\mathrm{FAC}}$ and $\mathbf{u}^{\mathrm{PAC}}$, we generated large sets of time-delayed stochastic simulations (see Section 3.2.5) for different lengths of the time delay. Each SSA was sub-sampled for 1000 times over 10000 minutes of simulation time after a burnin period of 10000 minutes.

Figure 3.4 shows the joint distributions (left) and marginal distributions (right) at varying levels of time-delay, with panels A-C showing results for the FAC controller and panels D-F showing results for the PAC controller. Figure 3.4G summarizes these results by plotting the score of both controllers versus the time delay. From the figures, it is clear that performance is rapidly degraded as the delay approaches and then exceeds the characteristic time ($\tau_c = 1/\gamma = 49$ min) set by the degradation rate of the process (yellow dashed line). At very small time delays ($\tau < 3.4$ min), the FAC outperforms the PAC but at moderate time delays ($\tau > 3.4$ min) the PAC outperforms the FAC.

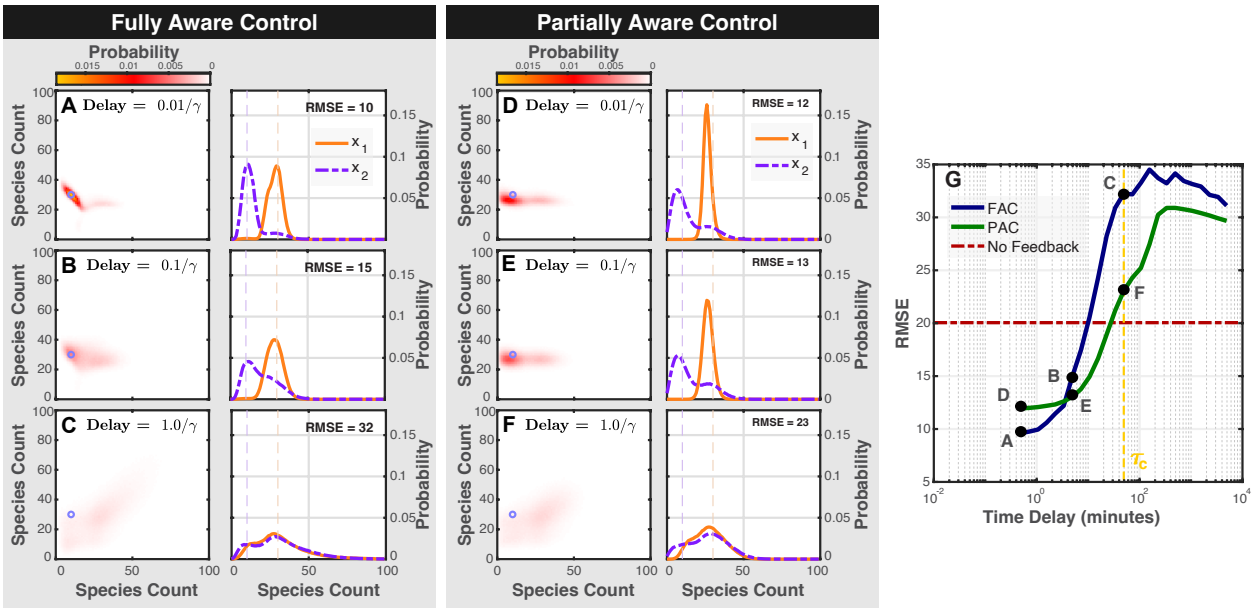

Figure 3.4: Effects of time delay on control performance. (A-C) Joint (left, color scale shown at top) and marginal distributions (right) of the controlled system at different levels of time delay using the FAC controller. The target state, $\mathbf{T}$ is denoted by a small circles on the left panels and dashed lines on the right panels. The steady state RMSE is shown in each case. (D-F) Same as (A-C) but for the PAC controller. (G) RMSE control performance versus time delay for both FAC (blue) and PAC (green). Letters A-F correspond to panels A-F. Dashed red line corresponds to optimal performance with no feedback (i.e., constant input). Dashed yellow line corresponds to characteristic system time, $\tau_c = 1/\gamma$.

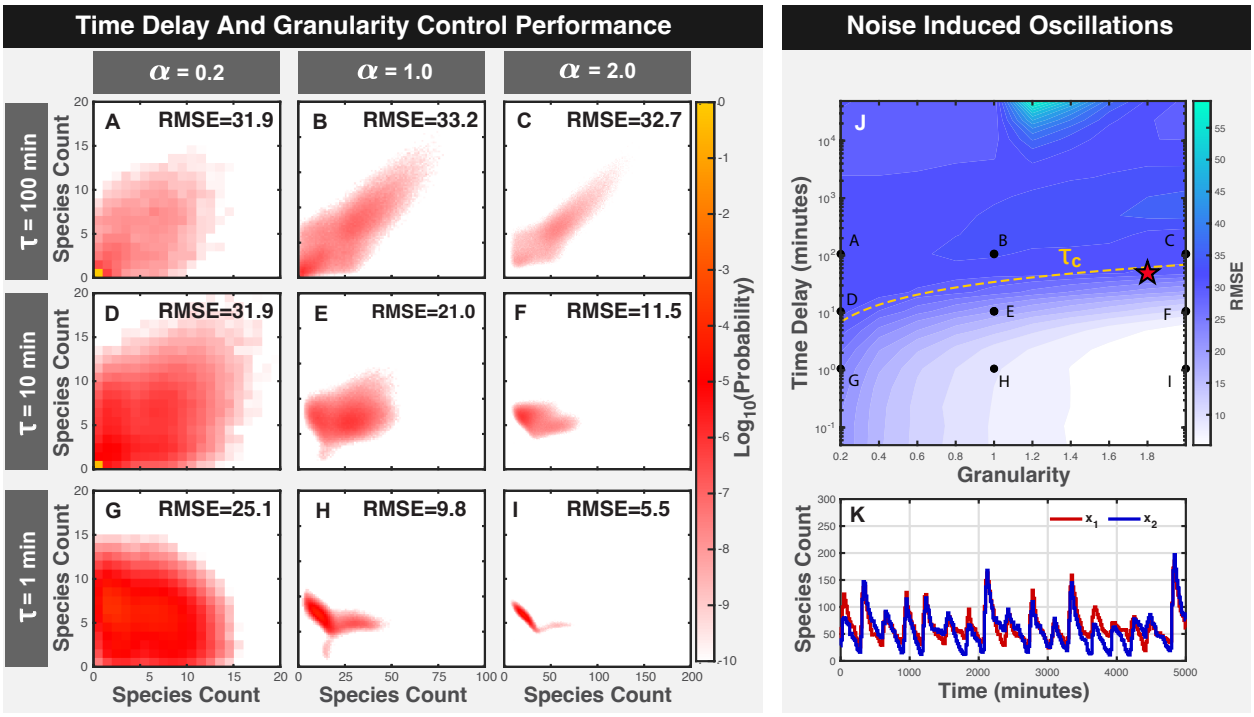

Figure 3.5: Joint effects of granularity and time delay on control performance. (A-I) Joint probability distributions at different combinations of τ (rows) and α (columns). Overall RMSE shown at top. (J) Heat-map of control performance versus (τ, α). Points A-I correspond to panels A-I. Dashed yellow line shows characteristic time $\tau_c = \alpha/\gamma/$. (K) Controlled system trajectory for $\alpha = 4.0$ and $\tau =$ min (denoted by red star in panel J).

At high time delays, both controllers lose their asymmetry, resulting in significantly worse performances ($RMSE = 32$ and 23) and even perform worse than under a simple constant control input without feedback (depicted by a horizontal red line in Fig. 3.4G).

As discussed in the previous section, increasing the granularity of the system α reduces the randomness to improve the steady state performance but at the cost of slowing down the controller response. To explore the joint effects of time delays and granularity, Fig, 3.5J shows the FAC steady state control performance as a function of the time delay and system granularity (τ, α) using thirty-two stochastic simulations simulated to steady state at each combination. Control performance errors measured by RMSE improved as α increased when the delays were small ($\tau = 1$ min and $\tau = 10$ min) but not when $\tau = 100$ min.

Recalling that under the system granularity rescaling (Eq. 3.20), as α changes, the effective degradation rate scales according to $\gamma' = \gamma/\alpha$. Therefore, the critical limit for the delays should

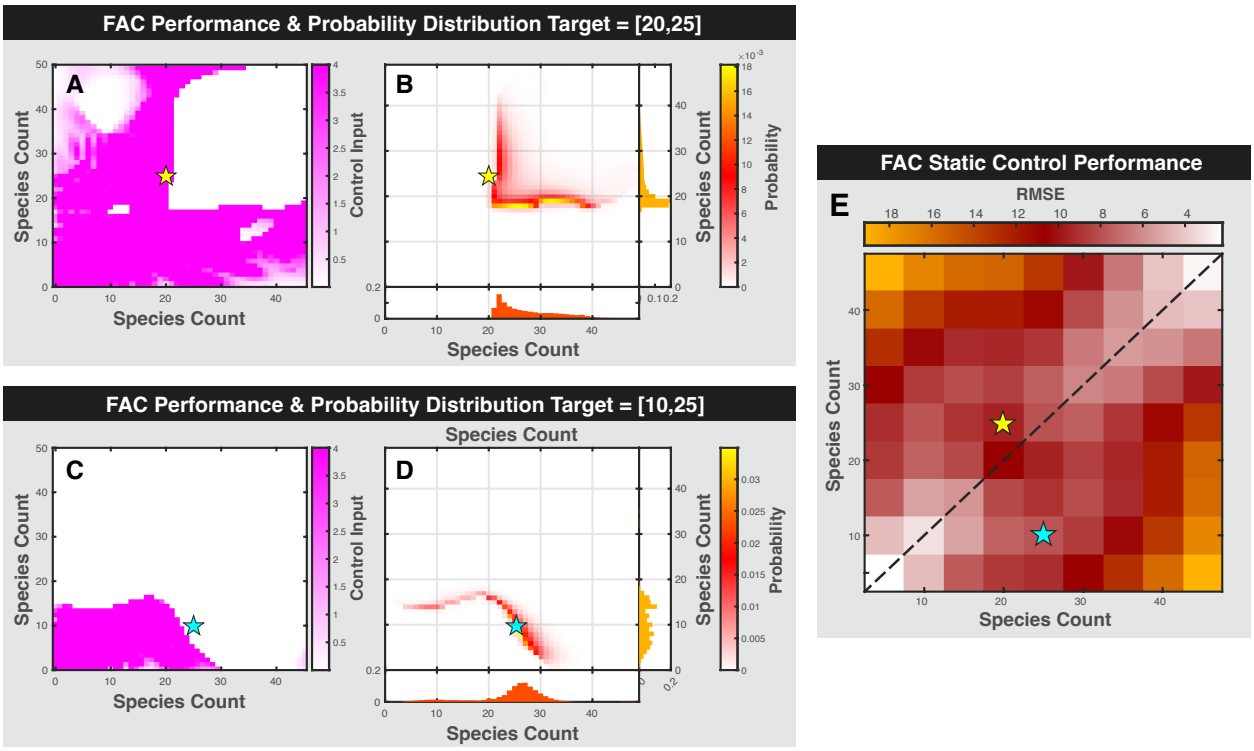

Figure 3.6: FAC control laws and performance for different target points. (A) Optimized control input for target $\mathbf{T} = [20, 25]$, with target point denoted by star. (B) Corresponding steady state response distribution. Marginal distributions for x_1 and x_2 on right and below. (C,D) Same as (A,B) but for $\mathbf{T} = [10, 25]$. (E) FAC control performance (RMSE) versus targets $\mathbf{T} = [T_1, T_2]$. Stars correspond to target points in panels A-D. Dashed diagonal shows line of symmetry.

also change with the system scale according to $\tau_c = 1\alpha/\gamma$. Figure3.5(J) depicts this characteristic line and shows that as the time delay approaches and then exceeds this level, the steady state performance becomes dramatically worse.

Finally, to understand the model of failure at these longer delays, it is interesting to examine trajectories induced by the controller just below this characteristic delay. For parameter set denoted by the red star in Figure. 3.5(J), Figure. 3.5(K) shows the controlled response after a long burn in period to achieve steady state. In this case, the application of feedback after a delay leads to a strong oscillatory behavior and worse performance than that achieved without any control at all. This observation further stresses the importance of considering time delays when designing such controllers.

3.3.4 With noise-induced control, a single input can drive multiple cells to follow different temporal trajectories.

We next optimized the FAC controller and calculated its performance for the two-cell system over a two dimensional domain of discrete target points $\mathbf{T}_i = [T_{i1}, T_{i2}]$ between between five and forty-five. For example, Fig. 3.6 (A) shows the optimized FAC control input of the system when the target is $\mathbf{T} = [20, 25]$, and Fig. 3.6 (B) shows the corresponding steady state probability distribution of the controlled system. Figure 3.6 (C and D) show the the same thing but for a different target state of $\mathbf{T} = [10, 25]$. Figure 3.6(E) shows the overall steady state FAC control performance over the entire domain of static set points, and illustrates that some regions are easier to attain than others.

We developed a method (Section 3.2.6) to control the system dynamics to follow a predefined path $\mathbf{T}(t)$ by alternating between 32 different pre-computed controllers along each path. We considered three representative pathways, including an in-sync reference point (Fig. 3.7B1) where

$$T_1^{\mathrm{B}}(t) = 20\sin(2\pi ft) + 10, \text{ and } T_2^{\mathrm{B}}(t) = 20\sin(2\pi ft) + 10;$$

a phase lagged reference point (Fig. 3.7C1) where

$$T_1^{\mathrm{C}}(t) = 20\sin(2\pi ft) + 10, \text{ and } T_2^{\mathrm{C}}(t) = 20\cos(2\pi ft) + 10;$$

and a frequency separated reference point (Fig. 3.7D1) where

$$T_1^{\mathrm{D}}(t) = 20\sin(2\pi ft) + 10, \text{ and } T_2^{\mathrm{D}}(t) = 20\cos(2\pi ft) + 10.$$

For each reference signal, the driving frequency was defined as $f = 10^{-4}$ cycles per minute.

All FSP simulations were calculated under the time varying control law, and Figs. 3.7(B2, C2, and D2) show the corresponding response distributions (shading) and median responses (lines) for two cells x_1 and x_2 in red and blue, respectively. Regions with purple shading depict the

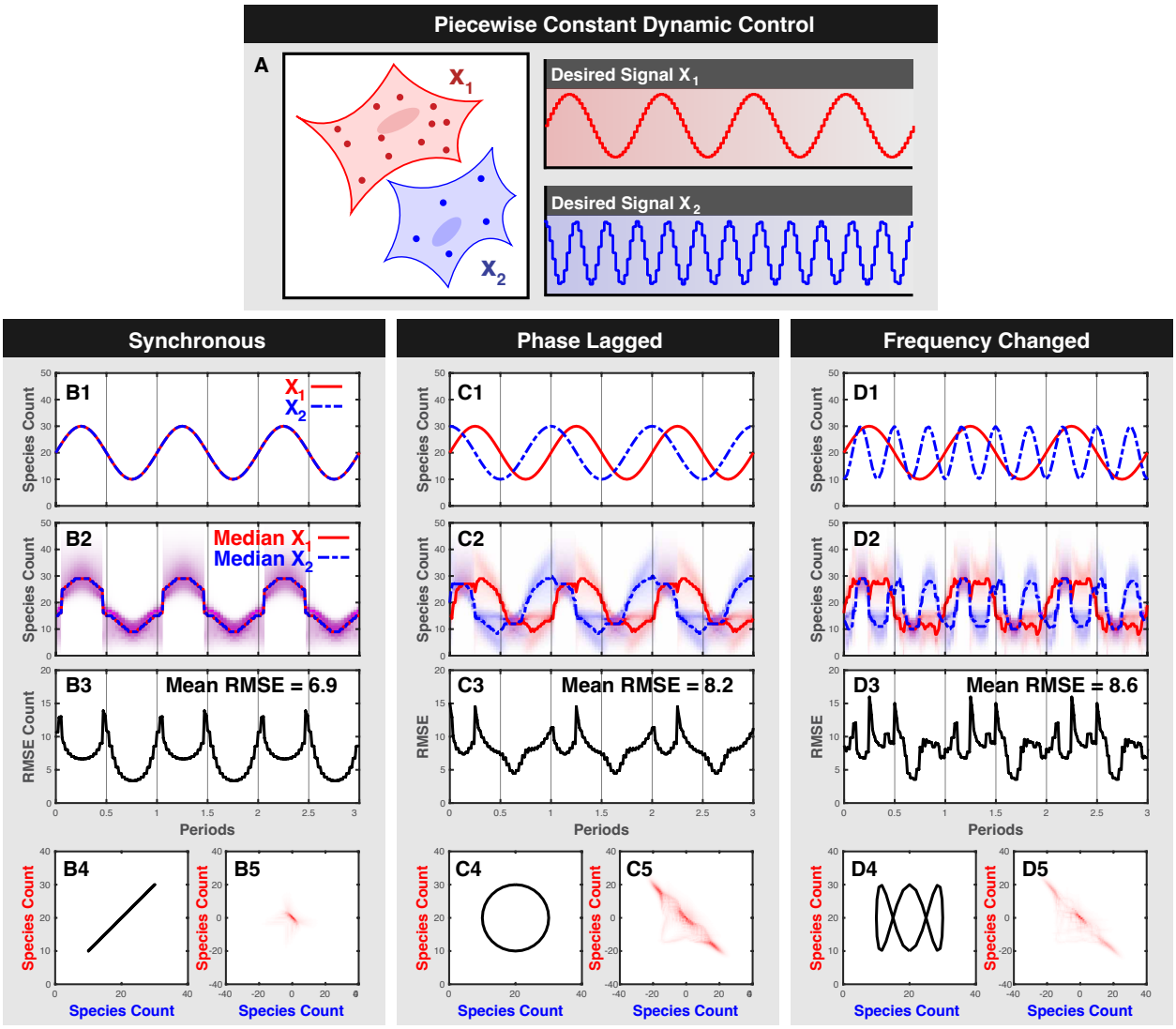

Figure 3.7: Tracking time-varying reference signal. (A) Schematic of SIMO control to drive two cells to follow different trajectories. (B1) Reference signal for x_1 (red) and x_2 (blue). (B2) Controlled response. Distributions shown in shading. Median shown in lines. Three periods are shown after decay of transient dynamics. (B3) RMSE performance over time. (B4) Phase space of reference signal. (B5) Time-averaged distribution of tracking error. (C1-C5) Same as (B1-B5) but for phase-lagged reference signal. (D1-D5) Same as (B1-B5) but for reference signal with two different frequencies and phases.

exact overlap of red and blue, when both cells have the same distribution of response. From Figs. 3.7(B2, C2, and D2), we observe that the SIMO control system can effectively drive the system to follow all three input trajectories, including when the two reference trajectories have different phases and frequencies (panel D2). To quantify the overall performance, Figs. 3.7(B3, C3, and D3) show the root mean squared error as a function of time, and Figs. 3.7(B4, C4, and D4) show the time-averaged error distribution for the system response relative to the time varying target. From these figures, we observe that all trajectories result in short RMSE spikes during short transient periods when the controller passes through the regions of poor control (i.e., through regions found in Fig. 3.6E to have high RMSE errors). Overall, the simulations show that the synchronous control performed best with an average RSME of 6.9. When the system is driven with a phase-lag, the average RSME of the score increases to 8.2, and when driven at a different frequency, the average RMSE goes up to 8.6.

In general, a given control system cannot effectively track a reference signal that changes faster that the system's natural time scale. Since it was shown that increasing granularity (α) improves steady state control performance (Fig. 3.4G) but also lengthens this time scale (Fig. 3.4H), one should expect that these competing effects of granularity would also affect the types and speeds of signals to which the system can respond. To examine how driving frequency and system scale affects control performance, 64 SSA trajectories of the phase-separated dynamic controller were simulated over a two dimensional domain of points (α, f) for 32 cycles after reaching steady state. Figure 3.8(A) shows the desired reference signal for the system, and Fig. 3.8(B) shows the mean of the system response when the frequency is $f = 10^{-4}$ cycles/min and the granularity is $\alpha = 4.0$. Figure 3.8(C) shows the corresponding control performance over normalized periodic time as α is held at 4.0 and f is increased from $f = 10^{-4}$ cycles/min to $f = 10^{-2}$ cycles/min and $f = 10^{-0}$ cycles/min leading to average control performances of 6.2, 16.1, and 15.0 respectively. Figure 3.8D,E extends this analysis to plot the average control performance over different combinations of α and f. The characteristic frequency (given by the predicted τ_c^{-1} over a range of α) is plotted as the yellow dashed line.

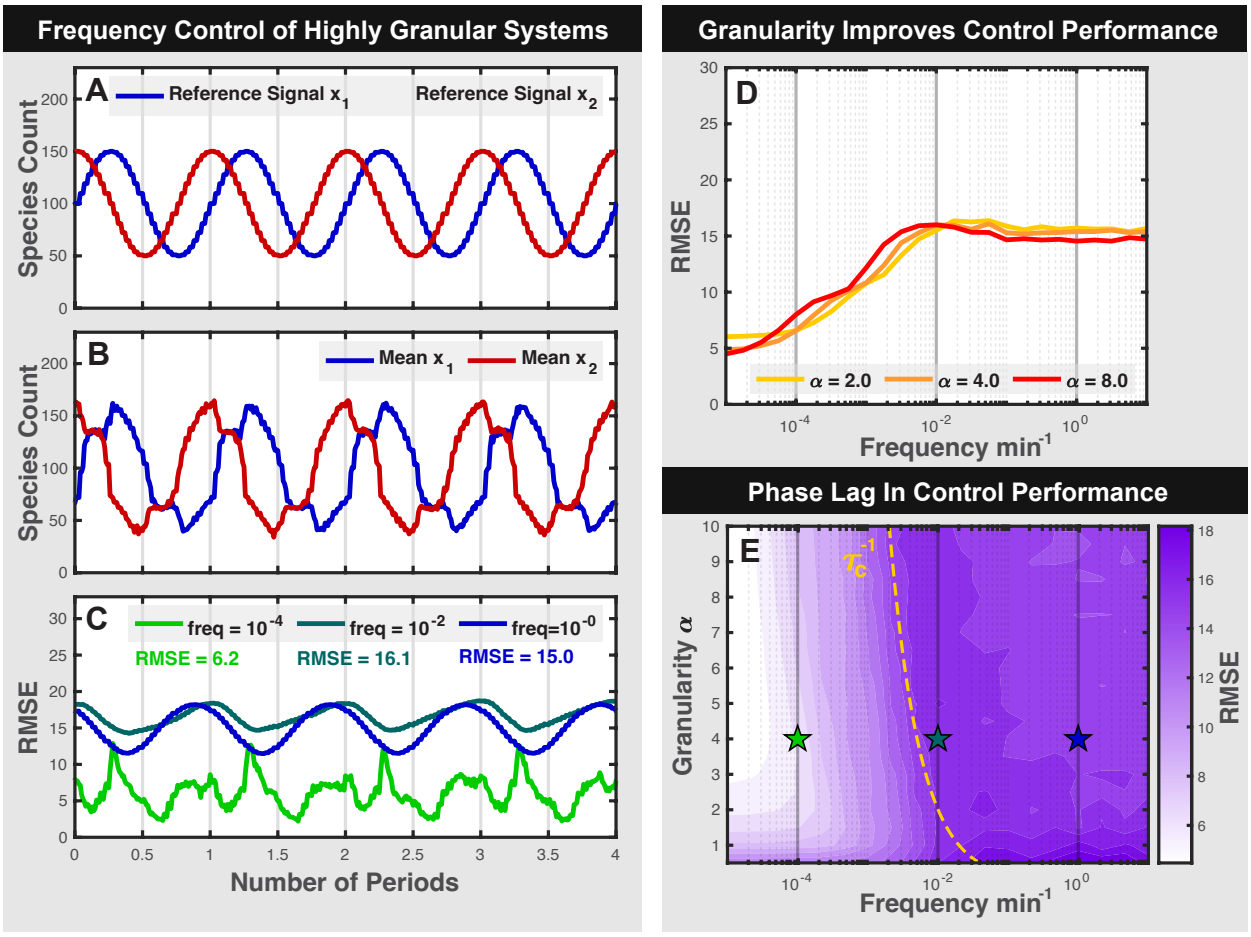

Figure 3.8: Tracking a time-varying signal at different frequencies and system scales. Control performance analyzed over a domain of f, α pairs show worst control performance at moderate frequencies near $1e-2$ due to phase lag. Tracking reference signals when $\alpha = 5$ span a range between 50 and 150 species (A). Stochastic simulations driven using a phased lagged controller at $\alpha = 5$ and low frequency show tighter control compared to $\alpha = 1$ (B). Systems driven at moderate frequency show worse control performance than high frequency or low frequency (C). Control performance only of phase-lagged system only improves with increasing α and low f.

In particular, the heat map of control performance over (f, α) shows that the worst performance occurs at moderate f near $f = 10^{-2}$ cycles/min and that control performance is best when granularity is high and frequency is low 3.8(E). More specifically, we find that strong performance was attainable only if the driving frequency was kept lower than a characteristic frequency $f_c \equiv 1/\tau_c = \gamma/\alpha$, which is denoted by the dashed line in Fig. 3.8E.

3.4 Conclusion

Noise, whether it arises from inherently stochastic processes or from unknown or unmodeled interactions, can play a critical role in the performance of feedback control. For the field of synthetic biology, this noise has typically been avoided and genetic systems have primarily been engineered to be as robust as possible to these uncertain fluctuations. In contrast, many natural cellular processes exist and thrive in settings where single-molecule events such as gene activation leads to large relative fluctuations, and where response heterogeneity is unavoidable. Key results from [119] showed that certain controllers could exploit this noise to achieve objectives that would not be possible in a deterministic setting. With such controllers, a single regulatory signal, such as an optogenetic input, could drive two or more two genetically identical cells to different, arbiltrarilly chosen fates using just a single input signal and irrespective of the cells' initial conditions.

The effectiveness of any model-based controller depends upon the accuracy of the model with which that controller has been optimized, and as the real system deviates from its idealized model, the control performance will naturally be affected. In this work, we explore the effects of several such deviations, including uncertainties or errors in parameters, mismatches in assumptions for systems scale, and times delays.

Regarding parameter values, our perturbation analyses (Fig. 3.2) showed that control performance is strongly affected by the system parameters. We found that whether parameters were incorrect in all cells (e.g., due to systematic errors in the model) or for just one cell at a time (e.g., due to extrinsic noise in the cells themselves) affected the control performance in different way. In most cases, small changes could be tolerated, whereas large changes, especially to certain key

parameters, could be catastrophic. Moreover, we found that there can be room to improve control performance by adjusting system parameters, suggesting that joint optimization of the controller with the system itself could lead to even stronger performance. Armed with such insight into which parameters are the most sensitive and which can safely be ignored, one could in principle focus measurement efforts to more precisely quantify the critical parameters and focus design efforts to reduce variability in key aspects of modular parts.

The size of the system also plays an important role in its control performance. For a fixed concentration, as the volume of a chemical reaction system increases, it becomes less noisy, and its dynamics approach that of deterministic process. At this limit, symmetry can no longer be broken, and feedback control cannot independently drive different cells to different fates. On the other hand, deleterious fluctuations also become smaller, so larger systems can more easily maintained maintain their desired phenotypes. Overall, we have shown (Fig. 3.3A-G) that the removal of noise through system granularity led to better steady state control performance, but such systems were found to take a much longer time to achieve steady state (Fig. 3.3H). Interestingly, this result could have implications on the malleability of cells at different stages of their growth cycle, where differentiation of smaller cells (e.g., those immediately after division) may be more susceptible to control signals, while larger cells (e.g., mature cells that have already established their phenotypes) may be relatively impervious to external signals.

Although our objective was to determine how well control performance would be maintained under different system sizes, we were surprised to find that controllers designed at one level of granularity (e.g., α =1) worked surprisingly well to control systems at much larger granularities. From a practical perspective, the ability to analyze a model at one system scale and then effectively apply it to another could be highly beneficial. Since the computation time of the FSP solution to the CME grows with the square of the number of states, this can cause an explosion in computational requirements for large systems. Our results suggest a promising alternative in which one could learn or optimize a controller using FSP analyses for a computationally feasible number of states and later apply them to larger systems that cannot be solved using current techniques.

Time delay analysis (Fig. 3.4) showed that increasing time delay decreased control performance. However, we also found that not every controller was equally affected by time delays. In particular, we found that at intermediate and larger time delays, a partially aware controller that has less information can outperform a fully aware controller (Fig. 3.4G). We believe this is happening because a controller with more information can afford to be more aggressive to implement its control, and time delay can cause this aggression to backfire.

The control of cells to two slowly changing dynamic reference signals using a single global input by the use of a noise-exploiting controller showed good control performance for a variety of signals. These analysis could be extended to include faster frequency by numerical calculation or by alternative error-probability adjustments.

Chapter 4

Microscopy Automation

[3] A combination of microscopy automation, high-throughput image processing pipelines, and decision-making algorithms is needed to improve the gathering of high quality data at a fast rate, and is needed to accelerate the analysis of heterogenous cell populations. Image processing algorithms to reproducibly determine spot counts (e.g., to count the number of mRNA of a given species within each cell as using smFISH) are integral in determining the likelihood of gene regulation models and to correctly selecting the right model for a given biological process. We demonstrate a new integrated pipeline to automate the image collection including: (i) quickly search in two-dimensions to find fields of view with cells of desired phenotypes, (ii) targeted collection of three-dimensional image data for these chosen fields of view, and (iii) streamlined processing of the collected images for rapid segmentation, spot detection and tracking, and cell/spot phenotype quantification.

4.1 Introduction

Fluorescent labeling encompasses a range of techniques that have become important tools in molecular biology for identifying protein and mRNA behavior in cells. Fluorescent labels using green fluorescent protein (GFP), the MS2-MCP RNA tagging, and single-molecule fluorescence in situ hybridization (smFISH) enable researchers to label DNA, RNA or protein molecules in cells. When exposed to light of a particular wavelength using a laser, these markers emit fluorescent signals, which can be captured by specialized microscopes. Fluorescent imaging enables the localization and movement of important biological molecules for the study of gene expres-

[3]Chapter 4 is in preparation for publication and describes a collaborative effort with Dr. Luis Aguilera (Image processing and microscope emulation) in Brian Munsky's laboratory and Dr. Tatsuya Morisaki (Microscope construction and control) in Tim Staseviche's laboratory. MPM's contributions to this section are in the development of control software to automate the microscopy, image processing tools to process images as they are taken, and decision rules to adjust microscope settings as needed during the process.

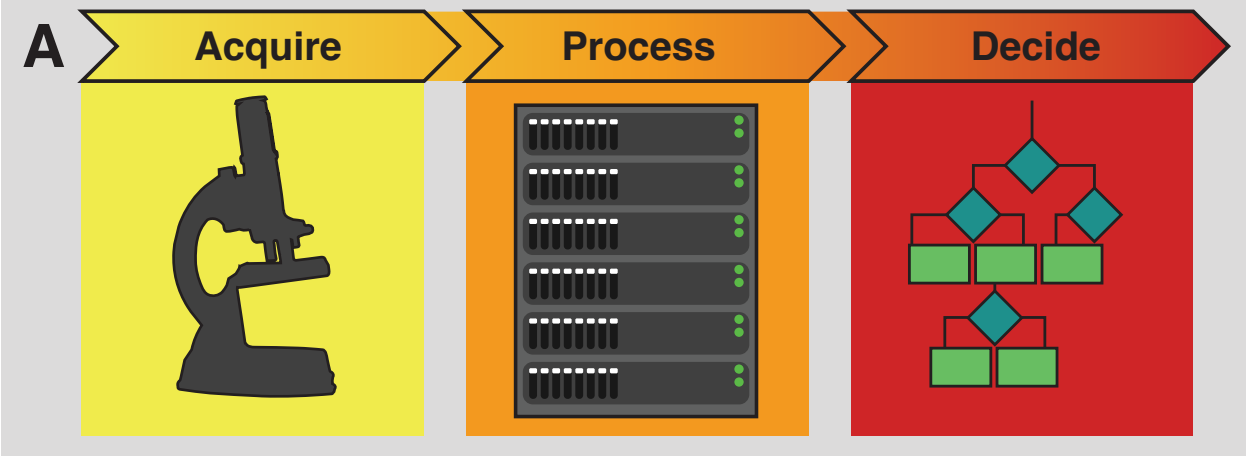

Figure 4.1: Schematic of the high level Acquire-Process-Decide process. High level automation builds upon mid level automation by acquiring acquiring large datasets, processing them quickly using distributed machines, and making decisions about the next acquisition depending upon the results of the processed data.

sion, the investigation of cellular interactions, and the visualization of dynamic processes in living organisms.

Single-molecule fluorescence in situ hybridization (smFISH) imaging offers several advantages, including the ability to pinpoint and quantify individual RNA molecules within cells or tissues. This high-resolution technique enables the study of gene expression, spatial organization, and co-localization of transcripts, providing insights into the molecular mechanisms underlying various biological processes. By utilizing specific RNA tags and fluorescent proteins, MS2 fluorescence microscopy enables researchers to gain insights into the real-time dynamics of gene expression and mRNA transport processes at a single-molecule level [126, 127].

Fluorescent labels and magnified optics enable researchers to collect data at the single-molecule level, to show important localization behaviors within cells that reveal protein and mRNA function. By tracking and visualizing the spatial distribution and movement of these molecules, key details about cellular dynamics and interactions can be determined. This approach also allows for the investigation of sub-cellular individual protein behavior at a fine scale, but also facilitating the analysis of protein behavior over cellular populations.

Image processing of fluorescent microscopy slides is a key step in the analysis of biological specimens and cellular structures. This process involves a series of computational techniques de-

signed to enhance the quality of raw microscope images, extract relevant information, and generate meaningful insights. Simple image processing can encompass tasks such as noise reduction, and contrast enhancement to improve image quality but more advanced types of image processing like deconvolution or spot detection also exists. Additionally, segmentation of cell nuclei and cytoplasm from images, enables the data collection of cell-specific data. Further quantitative analysis, like measuring fluorescence intensity or tracking object movement over time, can be conducted to derive valuable data for analysis.

Highly inclined and laminated optical sheet (HiLo) microscopy is an imaging technique that enhances signal to noise ratios by illuminating the sample at an angle [53]. More specifically, it achieves optical sectioning and improved contrast by selectively illuminating the sample with a thin plane of inclined light, enabling capture of images with less noise background but at the cost of needing multiple image stacks along the z direction. To acquire these three dimensional images, stacks of images must be acquired as the sample moves up and down perpendicular to the imaging plane. Advances in software have enabled the automation of acquisitions of datasets without the need for understanding low level hardware details.

Recent advancements in microscopy automation have enabled new ways to acquire and analyze microscopy data using software tools to automate [128–131] and identify cells [54, 132]. Automation can be used to automatically acquire images at different focal planes, time points, or wavelengths. This can be used to create images of cells over time, to track cell movement over time, or to study the dynamics of cellular processes. This not only reduces human intervention but also enhances the type of data collected and the reproducibility of experiments. While there are different automation tools for image acquisition, there is a need for new tools that can acquire data, process images, and make decisions on cell phenotypes in a simple and efficient manner. Additionally, using high throughput methods to process data increases the response speed of the microscope.

Leveraging microscopy for accelerated data acquisition, coupled with post-processing pipelines that extract scientific insights from experiments and enable data-driven decision-making, repre-

sents a robust approach to automate high-level cell detection and data processing. The incorporation of high-throughput image and data processing ensures rapid system response times, facilitating efficient and real-time analysis in research and experimentation. These tools enable the gathering of data while ensuring that each field of view meets specified criteria for image quality.

Emulating microscopy by generating simulated images offers several benefits, including reducing the material costs and acquisition time required to prepare for collection of real data. This accelerates development and facilitates the creation of accurate imaging protocols. Most importantly, an accurate simulator of cellular images allows researchers to test and optimize image analysis algorithms on a controlled dataset with known ground truth, which can then be used to test different strategies and improve the accuracy and efficiency of subsequent image processing tasks in real experiments.

4.2 Methods

4.2.1 Levels of automation increase the level of abstraction for acquiring datasets

Low level automation was implemented on a HiLo microscope which could interface with hardware using a custom device manager and through an open source microscopy software called Micromanager [129]. Micromanager serves as a device manager that enable the control of associated hardware via commands sent over serial ports in order to actuate physical hardware that control the physical aspects of microscope.

Mid-level automation in the microscopy system was achieved through the utilization of Micromanager and Pyromanager [128, 133], a Python API designed for Python interaction with Micromanager. Micromanager provides a framework for acquiring image datasets, by building upon the low-level hardware automation. Pycromanager is a Python interface for interacting with Micromanager through Python, with extensibility though callback functions. Mid-Level automations were developed using multi-dimensional acquisitions through Pycromanager which could acquire images without needing low level automation details.

High-level automation of the microscope was achieved by connecting multiple mid-level automation processes with data processing and decision making capabilities which alter the settings of the next acquisition. This integration formed an Acquire-Process-Decide pipeline, enabling the system to make informed decisions based on the results of analyzed data, and to perform higher levels of abstraction than the mid-level automation. An example of such a pipeline would be to "image z stacks of all found cells in a gridded region that have at least 3 cells", or to "find 20 cells in a region and image a movie of each cell". Such processes would require an initial acquisition to determine potential locations, post-processing images to find the specified number of cells, deciding to make a new acquisition to image positions which only had 3 or more cells in them, and then acquiring the final dataset. Finally, computational time of the data processing was accelerated using multiple machines in order to speed up the entire Acquire-Process-Decide process time.

4.2.2 High Throughput Image Processing Enables Quick Acquisition of Cell Population Statistics

The development of image processing pipelines was expedited through the utilization of libraries containing image processing steps which could be linked together to form an image processing pipeline and a repository of pre-built pipelines to correct images and gather relevant data from image datasets. These resources accelerated the implementation of complex image analysis workflows, and when combined with distributed computing it enables efficient extraction of spot count data and cell masks from microscopy data.

Data post-processors used image pipelines and image computations to extract relevant information from image datasets. A library of post processors to find image statistics, determine cell masks, count the number of spots per cell, count the number of cells per image, and identify large bright puncta corresponding to transcription sites were developed to enable the collection of processed image data.

Microscopy image emulators were developed to simulate the acquisition of image data without the need for a microscope. These emulations work by cutting out images of real cells from real

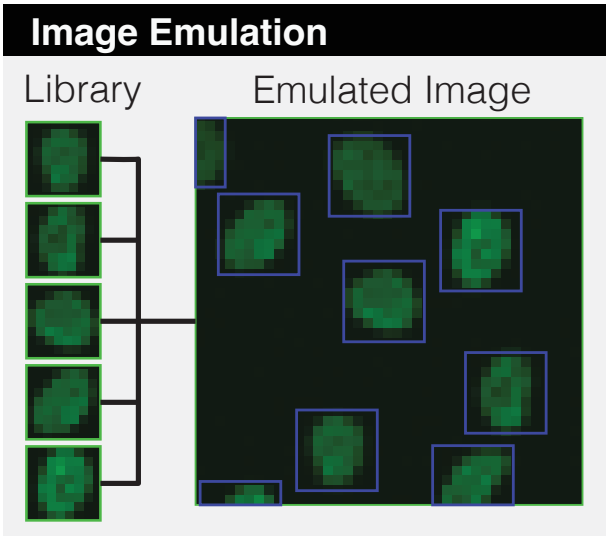

Figure 4.2: Image emulations using nuclei fluorscence data. Pseudorandom image emulation was performed by cutting up segmented nuclei and cytoplasms of real cell images and saving them to a library. The components are randomly selected and pseudo-randomly placed into the canvas without overlap when a new image canvas is emulated.

experimental images and pasting them pseudo-randomly into a blank canvas without overlap (Figure 4.2). Once the loop of the acquisition is started, smaller images are cut out of the image canvas depending on the position of the virtual stage. New canvases are made on each acquisition and data acquisition with the same settings yields different images depending on the outcome of the pseudo-random emulation. Image emulators were important to create reliable synthetic data for cell detection to evaluate cell detection performance metrics from known emulation ground truth data.

4.2.3 Acquire Process Decide Pipelines Enables High Level Automation and Decision Making

An 'acquisition ticket' was made to describe all variables and functions needed to perform a specific type of multi dimensional microscopy image acquisition, like defining positions to image, image exposure times, laser intensities, and more. Acquisition tickets were used with Pycromanager to perform multi dimensional acquisitions to acquire images. Acquisition tickets provide a convenient and efficient way to automate complex microscopy experiments, enabling researchers

to focus on designing and analyzing their experiments rather than programming detailed acquisition sequences.

A high level automation 'grid search' protocol was developed to find cells in a large grid and then perform image collection on images with at least three cells in them. In more detail, this involves acquiring an initial set of images on a grid, using Cellpose cell detection to determine the number of cells in the ROI, reject ROI with fewer than the accepted number of cells, and then acquire either z-stacks or a time series movies of accepted fields of view. Image processing using machine learning was performed on distributed machines to speed up decision making times. A similar protocol was developed which used mean intensity as a faster yet less accurate metric to determining cells counts.

A 'find cells' protocol was developed to image a large unconnected grid and to image the ROI with the most cells in them until a desired number of cells were found. More specifically, the program started by acquiring proposal images of ROI over a large area, counting the number of cells in each ROI using a cell detection algorithm, sorting the sequence of locations based on the number of cells found in the ROI and imaging fields of views with the most cells first until the desired number of cells were imaged and finishing the acquisition when the desired number of cells was found.

Automation pipelines were analyzed on (1) emulated images, (2) on a fixed slide of fluorescently labeled DUSP1 mRNA in Dexamethazone stimulated HeLa cells, and (3) on a fixed slide of fluorescently labeled H128 HeLa cells with bright red puncta corresponding to transcription sites. For the DUSP1 slide, DUSP1 exons were fluorescently labeled with CY5 (red), while GAPDH exons were labeled with CY3 (green). In the case of the H128 HeLa cells, slides were labeled with three fluorphores: MCP-GFP(Blue), GAPDH exons Cy3 (green). [134]. The fluorescently labeled H128 HeLa were chosen for their bright puncta.

An image puncta detector was created to identify highly active transcription sites which were visually identified as a very bright puncta in the image. The Laplacian of Gaussians (LoG) is an image processing technique used to detect and identify spots within an image. It operates by

convolving the image with a Gaussian smoothing filter to suppress noise and enhance the features of interest. Then, the Laplacian operator is applied to highlight regions of rapid intensity change within the smoothed image. The puncta detector score of the entire image was given by the maximal activation of the LoG in the image.

Sensitivity and specificity were used to analyze various image quantifications like machine learned cell count, mean intensity, and LoG puncta score. Sensitivity (true positive rate) measures the accuracy of identifying the number of cells in a positive image, while specificity (true negative rate) assesses the test's ability to correctly rule out negative samples. We used sensitivity to analyze the accuracy of various image quantifications. Ground truth was established through manual cell counting or from data taken from the emulator.

We developed a method for finding the sharpest image in a z-stack by capturing images in the z-dimension and selecting the one with the highest relative sharpness. This calculation uses a sharpness kernel to measure the detail in each pixel of the image, and the overall sharpness of the image is calculated as the sum of the convolved detail image. This enabled the imaging of sharp images.

4.3 Results

The 'grid search' protocol utilized an image emulator to capture sixty-four images in an eight by eight grid. Cellpose nuclei detection was applied with a diameter of 250 and a flow-threshold of 0.4, recording cell counts in each ROI. Images with three or more nuclei were accepted and are displayed in Figure 4.3(A)(green boxes). Acceptance ratios and computation times are provided below. Figure 4.3(B) shows an example of the image processing to mask a z-stack of emulated nuclei. Figure 4.3(C) shows plots of actual cell counts from the image emulator's ground truth data. True positives (green circles), false positives (red circles), true negatives (black x), and false negatives (red circles) were plotted by comparing estimated and true cell counts.

Examination of the accepted fields of view for the image emulator shows good match with expectations. Fields of views with three or more cells were generally accepted with decent ac-

4.3.1 Validations of Automated Image Search with known ground truth

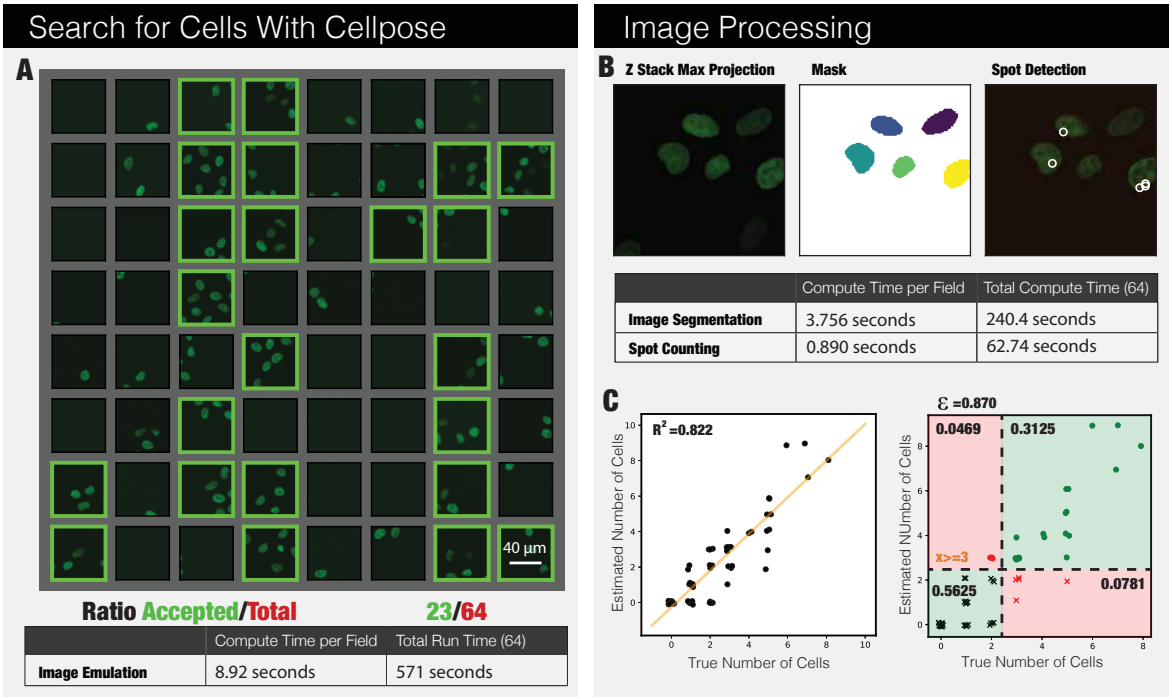

Figure 4.3: Automated data acquisitions using the image emulator. An eight by eight grid of images was acquired using the 'grid search' procedure using an image emulator that replaces acquired images with emulated ones. (A). Images which were believed to contain three or more nuclei using Cellpose were highlighted in green boxes, and an acceptance ratio was measured to be twenty-three out of sixty-four total images. (B) Images of Cellpose nuclei masks show good match with expectation, but missing a dim nuclei in the bottom right edge. (C) Correlations ($R^2 = 0.822$) and sensitivity ($\epsilon = 0.870$) suggest accurate determination of the number of nuclei.

curacy and a few exceptions. Under the acceptance criteria of three of more cells, an acceptance ratio of 23/64 was seen in the emulated images. Scatter plots of detected cell count versus ground truth cell count show a good correlation with an R^2 value of 0.870 which suggests that the true number of cells and the detected number of cells match well. Z-stack images were acquired from positions that were detected to contain three or more cells Figure 4.3(B). Image masks created by Cellpose were generally accurate, but missed nuclei can be seen in the bottom right of the Z-Stack Figure 4.3(B,left) which does not appear in the mask. System sensitivity and specificity for the emulation was calculated using estimated cell count as the data, and true cell counts as the ground truth. An accuracy rate was calculated by taking the total number of true positives divided by

the number of true or false positives. The sensitivity of 0.870 suggests that the accuracy of true positives is high.

4.3.2 Demonstration of automated search to find and quantify smFish measurements of DUSP1 in HeLa Cells

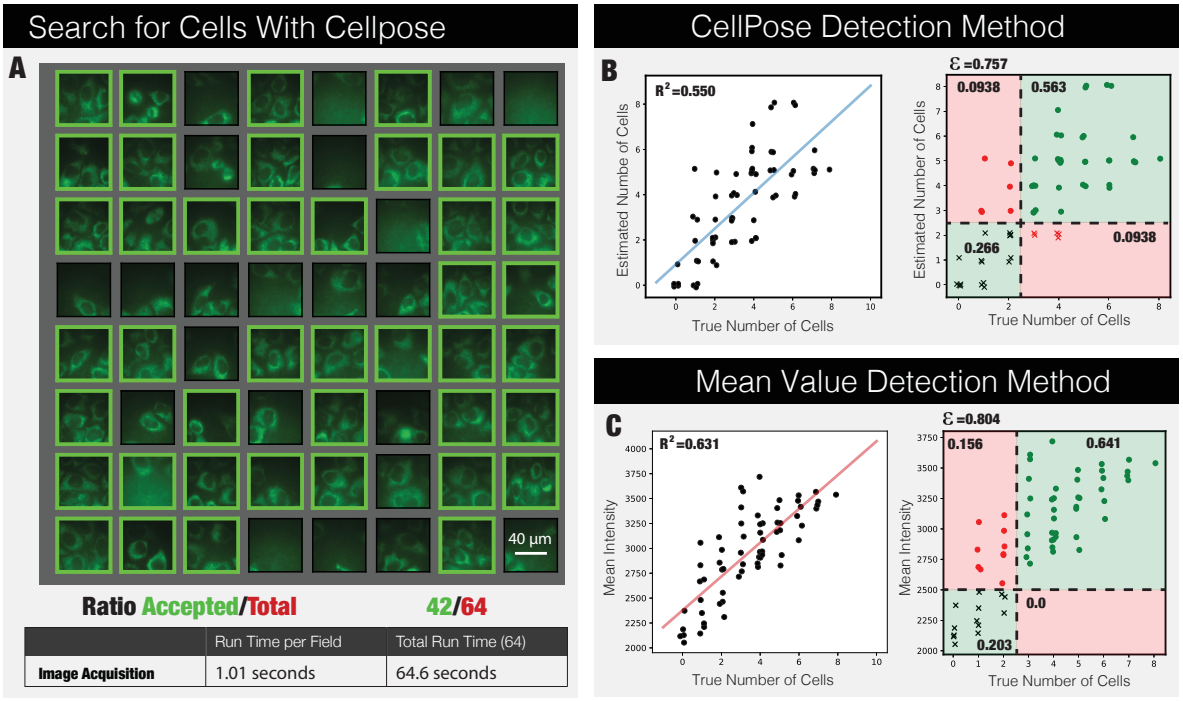

Figure 4.4: Automated data acquisitions of fluorescently labeled mRNA. An eight by eight grid of images was acquired using the 'grid search' procedure using smFISH stained cytoplasmic GAPDH exons. (A) Images which were believed to contain three or more cells using the Cellpose cytoplasm model were labeled in green. Image acceptance ratios (42/64) and acquisition times are shown in the bottom. (B) Correlations ($R^2 = 0.550$) and sensitivity ($\epsilon = 0.757$) of the Cellpose detection method can be seen. (C) Correlations ($R^2 = 0.631$) and sensitivity ($\epsilon = 0.804$) of the mean intensity detection method show similar accuracy and sensitivity to Cellpose for this set of images.

The 'grid search' automation protocol was evaluated on a fixed sample of HeLa cells with Cy3 labeled GAPDH exons on an eight by eight loose grid of images for a total of sixty-four images. Since these data are not emulated, ground truth for this data were created from cell counts measured by eye. Cells at the edge of the image were considered in the image if more than half the cell was believed to be in the ROI. Image acceptance ratios and runtimes for the system to acquire-

process-decide were gathered for the real system and shown on the bottom. Images were accepted using two different methods and compared later, either using Cellpose, or using a whole image mean intensity greater than 2500. These metrics were used to estimate if three or more cells were in the image. Correlations, sensitivities and specificities for the Cellpose detection method, or the mean value detection method are shown in figure (Figure 4.4(B) and Figure 4.4(C)) respectively. Although correlations using the real data decreased, explanation might be using a human to create a ground truth for erratic data.

Examinations of real ROI show cytoplasmic images instead of nuclear images obtained with the image emulation. Despite the machine learning model's adjustment to focus on cytoplasm instead of nuclei, it is apparent that the models exhibited lower accuracy compared to their performance with nuclear stains in the emulation dataset. This discrepancy can potentially be attributed to the presence of more background noise in the real data, occasional bright artifacts that were not cells, and increased cell density, all of which render image processing more error-prone. An R^2 value of 0.550 and a sensitivity of 0.757 corroborate the less precise identifications in real ROI images, in stark contrast to the ϵ value of 0.870 achieved in the emulation dataset. Alternatively, given the higher cell density in real data compared to emulated data, it may be advisable to employ more stringent criteria for selecting ROI in high-density images. Notably, the total image acquisition and analysis time for 64 images was 64.6 seconds for an average of 1.01 seconds per image.

We assessed two methods for their capacity to estimate images containing at least three cells: the Cellpose method and the Mean Intensity method. An image was considered to have three or more cells if a specific measurement surpassed a particular threshold. For Cellpose, this threshold was set at three detected cells, while the Mean Intensity method employed a threshold of 2500 average pixel intensity.

Scatter plots depicting the performance of each detection method revealed an R^2 value of 0.550 for the Cellpose detection method and an R^2 value of 0.631 for the mean intensity method. Sensitivity analysis indicated that the mean intensity detection method exhibited higher sensitivity than the Cellpose method ($\epsilon = 0.804$ versus $\epsilon = 0.757$). Despite using the same image dataset, the mean

intensity method detected more true positives (0.828) compared to the Cellpose method (0.641). It is worth considering that one possible reason for the Cellpose method's suboptimal performance could be related to the need for further training to enhance detection performance.

4.3.3 Cellpose is a reliable indicator for cell detection

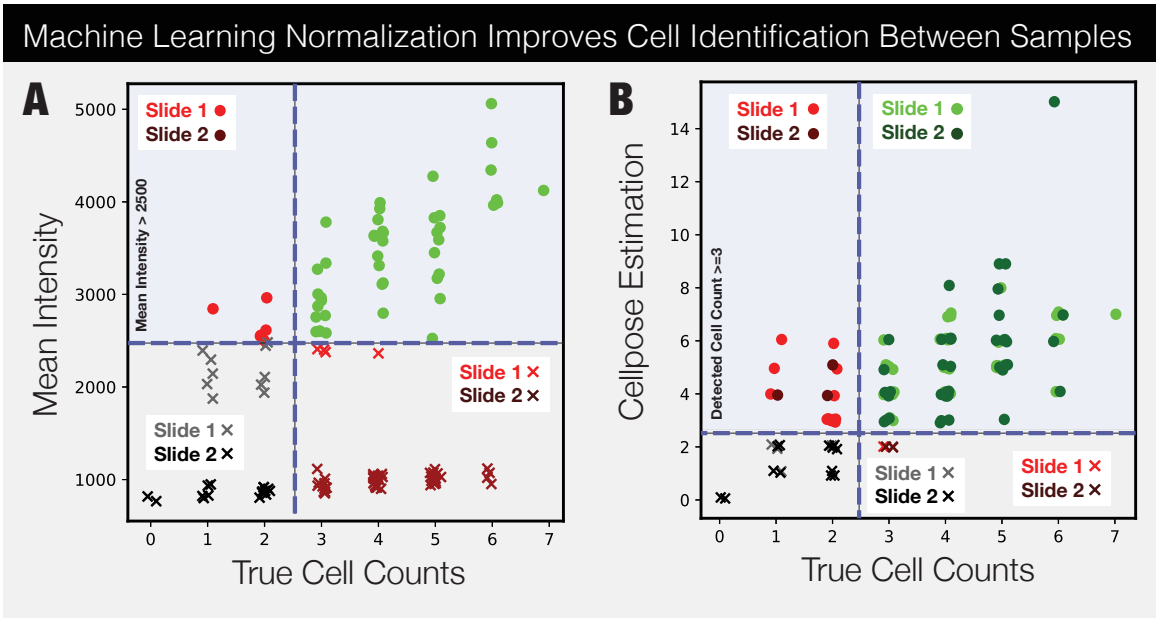

Figure 4.5: Median image processing on two slides. The mean intensity method and the Cellpose identification method were compared using grid searches on two different slides with the same imaging conditions. (A) The mean intensity method was used to determine which regions of interest (ROIs) to keep for re-imaging. Images were predicted to have three or more cells if the median intensity was greater than 2500. Scatter plots of slide one data and slide two data show large discrepancy between the two slides. (B) The same images were then analyzed using Cellpose. Scatter plots of slide one and slide took look much more uniform.

The 'grid search' automation was evaluated on two different slides with fluorescently labeled GAPDH. Images from these two datasets were analyzed for the presence of at least three cells using either the mean intensity method or the Cellpose method. Scatter plots of the mean intensity versus true intensity were made for both slides (Figure 4.5(A)), and using Cellpose (Figure 4.5).

Scatter plots of the mean intensity versus true intensity show a large difference in intensity values between the two slides (Figure 4.5), with an average intensity near 3000 for slide one and

an average near 1000 for slide two. This large difference in mean intensity between two slides makes the 2500 cuttoff for determining cells in an image work well for slide one but not for slide two. As a result, no positives are detected in slide two using this method and the rate of false negatives becomes 0.75.

Scatter plots of mean intensity versus true intensity reveal a large difference in intensity values between the two slides (Figure 4.5), with an average intensity of around 3000 for slide one and around 1000 for slide two. This large difference in mean intensity makes the 2500 threshold for determining cells in an image may work well for slide one but not for slide two. Therefore mean intensity is not a good indicator over different samples, although corrections through renormalization for each slide could correct this concern. On the other hand, the strong overlap in the scatter plot in Figure 4.5(B) shows that Cellpose is still able to accurately segment cells in both slides, regardless of the difference in average intensity. This can be confirmed with an average sensitivity of $\epsilon = 0.860$ and a detection rate of 0.719 between the two slides was measured by taking the radio of true positives versus the total number of cells. This makes Cellpose a more reliable 'off-the-shelf' method for cell segmentation across a variety of samples despite being substantially slower and marginally less accurate for a single slide.

4.3.4 Puncta identification leads to data capture of targeted phenotypes.

The 'grid search' automation protocol was evaluated on a fixed sample of H128 cells with bright transcription sites on an eight by eight loose grid of images for a total of sixty-four images. Each image was labeled by eye to determine the number of bright spots in the image Figure 4.6(A) and each image was analyzed for puncta using the LoG of the image to detect spots in the image. Plots of LoG detected versus true puncta counts were made using an inner standard deviation of two and an outer standard deviation of seven. True positives, false positives (green and red circles), true negatives and false negatives (black crosses and red crosses) were made Figure 4.6(B).

Images were estimated to contain at least one puncta if the maximum LoG score in the red channel exceeded 500. A sensitivity analysis of the maximum LoG scores demonstrated a high

sensitivity of 0.955, indicating that the max LoG score is a robust true-positive indicator for puncta identification (Figure 4.6(B)). However, it should be noted that max LoG is not a good estimate to count the number of puncta, as it represents only the highest activation within the entire image and is only meant to determine if there is one or more puncta in the image. This limitation results in a weak correlation ($R^2 = 0.158$) between the number of puncta and max LoG, despite the high sensitivity.

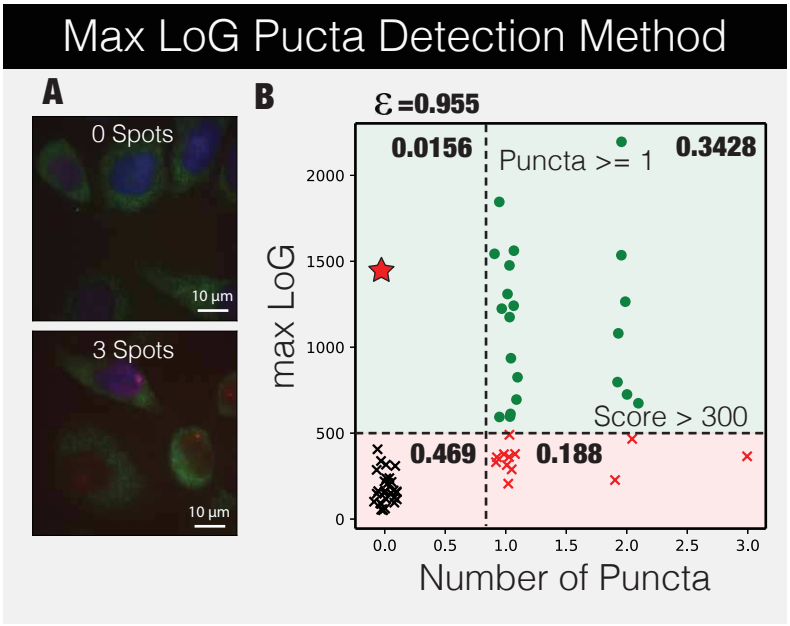

Figure 4.6: Puncta detection using Laplacian of Gaussians. A 'grid search' protocol was analyzed using MS2 labeling of transcription sites, over an eight by eight grid of images for sixty-four total images. Images were analysed using a transcription site finder that implemented the max LoG of the image to identify the presence of any bright puncta. (A) Images with (bottom) and without (top) puncta show a bright spot in the red channel of the image. (B) Sensitivity analysis of the transcription site finder to determine at least one puncta in the image shows a sensitivity of $\epsilon = 0.955$.

4.3.5 Data Collection times scales linearly with the number of targeted fields of view

Image collection times were gathered over a range of ROI between one and 961 under three different numbers of Z stacks. Collection times are plotted against the number of ROI on a log

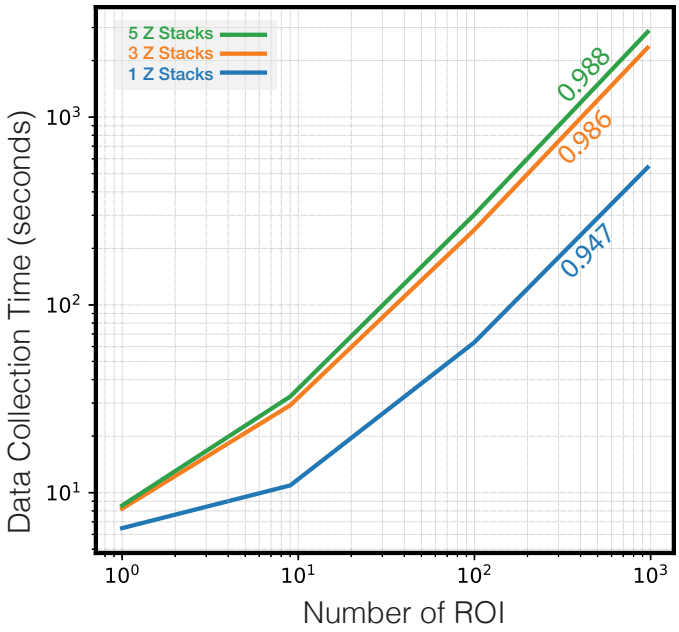

Figure 4.7: Figure of time scales to collect image data. Image collection times were measured for varying numbers of ROI positions from one to 961 and plotted on a logarithmic scale. The slope of the line in log space determines the scaling factor of the system when the number of ROIs is large. Slopes of $m = 0.947$, $m = 0.986$, and $m = 0.988$ were found for one z-stack, three z-stacks, and five z-stacks, respectively. These results indicate that the scaling factor of the system is near 1, regardless of the number of z-stacks.

log scale and the scaling factor was found for each by taking the slope of the linear portion of the non-linear curve near 10^3 in log space (Figure 4.7).

The scale factors of the one z-stack, three-z stack and five z-stack log scale data acquisitions were found to be $m = 0.947$, $m = 0.986$, and $m = 0.988$ respectively. Since all slopes of the scale factors are close to one, the time taken to complete an acquisition scales linearly with the number of ROI. Slopes of acquisition times near one ROI show slopes of 0.238, 0.577, and 0.608 for one z-stack, three z-stacks and five z-stacks respectively.

Table 4.1: Table of initial imaging, processing, and collection times for an automated search.

	Slide 1	Slide 2	Slide 3
Initial Imaging (seconds)	215.	233.	213.
Processing (seconds)	234.	295.	292.
Secondary Imaging (seconds)	840. (70 images)	1050. (87 images)	858. (72 images)
Mean Spot Count	149.1	191.2	130.12
Standard Deviation Spot Count	88.8	153	131

Initial imaging times, processing times, and collection times were gathered using a 'find cells' procedure which finds 100 cells. Specifically, this method entailed capturing a grid of images across a predefined area, sorting these images based on the estimated number of cells present, and then revisiting images until at least 100 estimated cells had been imaged. This process prioritized images with the highest estimated cell counts. Images were acquired in a single channel and using 7 z-stacks. The average initial imaging time amounted to 223.25 seconds for the initial acquisition of 400 images (Table 4.1). Cell detection, executed using Cellpose, averaged 281.75 seconds for 400 images. Image collection times varied between 840 seconds and 1050 seconds depending on the number of images needed to complete the acquisition. It's noteworthy that, despite the initial request for 100 cells, all final collection acquisitions were achieved with fewer than 100 images.

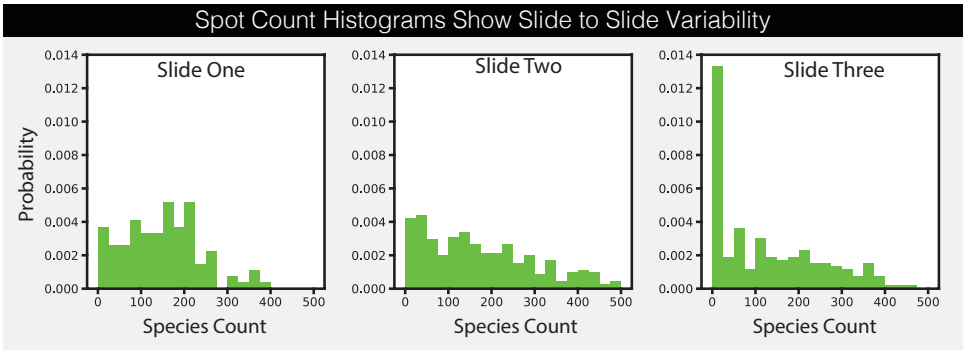

Figure 4.8: Slide to slide histograms of spot count frequency. Image processing pipelines were used to accelerate spot counting of cells over three different slides using Cellpose cytoplasmic masks and an LoG spot counting methods. Histograms of spot counts per cell are shown in each subplot and variability between distributions can be observed. Slide three shows a large amount of cells with zero spots detected with respect to slide one and two.

Image processing pipelines were used to calculate histograms of spot count frequencies per cell over three different real samples. These samples were masked using Cellpose and spots were detected using an LoG method that identifies positions within images that contained spots above an activation threshold of 300 using a inner and outer gaussian σ of two and eight respectively. Spots in images were localized to individual cells using the masks and the total spot counts within cells were gathered for each cell in each image. These data was used to create histograms of spot count frequency. Spot count histograms between samples showed variability between samples.

Figure 4.8(right) has a large number of cells with zero spots within them when compared to Slide one and Slide two.

4.4 Conclusion

Automation is a key technology for the development of smart microscopy tools that can automatically optimize experiments. By automating complex acquisition processes that rely on multiple individual data acquisitions, data processing, and decision making in a streamlined manner, we can search for and image cells more efficiently and effectively. Here, we demonstrated our ability to create complex acquisition processes that rely on multiple individual data acquisitions, data processing, and decision making in a streamlined manner to search and image cells with desired properties and phenotypes. This is a step towards the development of smart microscopy tools that can automatically optimize experiments and acquire large processed datasets and images from microscopy slides.

While we observed high sensitivity in image identification, none of the methods were perfect. While many misidentifications were due to unusual features in the real data (e.g., strange cell morphologies, bright artifacts), some may have been attributed to poor machine learning settings that were not trained on these specific images. These identification imperfections can lead to the capture of incorrect data, which still requires some human intervention to clean up if the error rates are not low enough.

To create complex optimizations, we developed a library of acquisitions, data processors, and decisions that could be strung together. This approach allowed us to modularize the automation process and make it easier to manage. While it is possible to create software that can do high-level automation, it can come at the cost of increased complexity, making it difficult to maintain and add new features.

Automated acquisitions will allow scientists to collect more data, more quickly and more accurately. Smart microscopes could be used to perform complex experiments that involve multiple steps or that require continuous monitoring. By automating experiments and making it easier to

design and perform new experiments, smart microscopes could accelerate the discovery of models through brute force data acquisition.

Automation will also enable scientists to design experiments to develop a smart microscope that can automate the optimization of experiment settings and suggest better experiments that maximize information. Recent mathematical tools suggest that poor image acquisition settings that decrease information can be represented by a distortion matrix that increases the space of likely parameters [57]. Developing Fisher information software could rapidly determine how settings can increase information content in a fast and automated way. These advances could accelerate the discovery of models while minimizing time and reagent costs.

Chapter 5

Conclusion

This work aimed to demonstrate the performance of noise-exploiting control and to develop automation tools to begin to answer questions about the behavior of cell populations. Motivated by synthetic biology, biological circuits and optogenetics provide powerful tools for precise control of biological systems, but noise in these systems can undermine desired control performance.

Noise in control theory was initially addressed through the implementation of 'robust' control techniques. In this area of control theory external disturbances, uncertainties in system parameters, and sensor measurement noise could severely impact system performance. While robust control methods have traditionally been effective in managing stochastic systems in spite of noise, there are opportunities to develop noise-exploiting control methods that enable new control techniques that were previously thought to be impossible.

In the introduction, we showed that conventional thinking about the control of stochastic systems has limitations, and that master equation analysis provides opportunities to create a noise-exploiting controller that works with the noise rather than against it. We also showed that the development of microscopy automation could enable analyses of cell populations and experiment information.

In chapter two, we developed and optimized a control theory problem for the simultaneous control of two similar systems with a single input to two different probability distributions. We found not only one controller that could achieve this, but multiple controllers with different levels of observation and prediction which could break symmetry between two cell systems. We also determined that autoregulation is a key component that contributes to the high performance of these controllers.

In chapter three, we extended the analysis to include parameter uncertainties, time delays, intrinsic noise, and a moving reference point. Extending noise-exploiting control theory to regulate

moving reference points at a fixed frequency despite the noise demonstrates its ability in achieving precise control.

In chapter four, we developed automated microscopy tools to gather large datasets of single-cell gene expression data, enabling us to begin answering fundamental questions about the development, function, and heterogeneity of cell populations. This required high level automation tools which could acquire, process, and make decisions.

5.1 Future opportunities

Noise-exploiting control theory offers a promising method to harness and manipulate the inherent variability of gene regulation to our advantage. By designing control strategies that embrace and leverage biological noise, we can potentially enhance the precision and robustness of synthetic biological circuits. Future opportunities in this domain may involve the design of noise-tolerant genetic circuits, and the creation of new synthetic control motifs with enhanced capabilities.

The presence of noise poses a significant challenge when creating large synthetic systems, often requiring the incorporation of complex control motifs to mitigate its effects. Rather than relying on intricate control mechanisms to combat noise, these controllers embrace and harness noise as a resource for system improvement.

The development of control theory that challenges conventional wisdom regarding the impact of noise on control system performance represents a future possibility in the field of control engineering. Historically, conventional thinking has often deemed noise as a hindrance, detrimental to the precision and stability of control systems. However, the research in chapters two and three suggests that noise exploiting control could be used to develop new techniques for the control of such systems.

While our initial focus was on solving a control problem involving two cells and a single control signal, it's important to recognize that our solution extends more broadly. In a general sense, we've effectively addressed a control problem that pertains to the manipulation of two targets using just one input, while harnessing the inherent noise within the system to achieve our objectives. This

broader perspective underscores the versatility and potential of noise-exploiting control strategies, as they can be applied to diverse scenarios beyond the specific context of our initial problem.

Biological systems are of interest to control engineers because they exhibit robustness and adaptability despite the inherent stochasticity and noise associated with molecular processes. By applying noise-exploiting control principles to biological contexts, we may gain deeper insights into how living organisms finely tune and utilize noise for their advantage. This interplay could uncover new methods for understanding living systems.

Automating single-cell microscopy data acquisition has emerged as a powerful tool for revealing the heterogeneous responses within cell populations. By leveraging advanced microscopy automation, researchers can efficiently capture a wealth of data at the single-cell level, allowing for a comprehensive exploration of cell population behaviors.

Using Fisher information with large datasets through "smart" software applications could enable the optimization of experiment information in an automated and efficient manner. By analyzing large datasets, these software-driven approaches can identify important time points, experiment variables and settings, thereby guiding experimental setups to maximize information gain while minimizing resource consumption. This synergy between statistical theory and smart software not only streamlines the research process but also holds the potential to accelerate discoveries.

In more detail, automations could be acquired using different experimental concentrations, time-points and imaging settings to find outcomes which make the largest Fisher information matrix possible. Some work has already been done to analyze how exposure times influence these measurements using a distortion operator which related accepted settings to distorted settings.

Bibliography

[1] Baojun Wang, Richard I Kitney, Nicolas Joly, and Martin Buck. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature communications*, 2(1):508, 2011.

[2] Sang-Woo Han and Yasuo Yoshikuni. Microbiome engineering for sustainable agriculture: using synthetic biology to enhance nitrogen metabolism in plant-associated microbes. *Current Opinion in Microbiology*, 68:102172, 2022.

[3] D Ryan Georgianna and Stephen P Mayfield. Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature*, 488(7411):329–335, 2012.

[4] Jonghyeon Shin, Shuyi Zhang, Bryan S Der, Alec AK Nielsen, and Christopher A Voigt. Programming escherichia coli to function as a digital display. *Molecular Systems Biology*, 16:1–12, 2020.

[5] Christopher V Rao. Expanding the synthetic biology toolbox: engineering orthogonal regulators of gene expression. *Current opinion in biotechnology*, 23(5):689–694, 2012.

[6] Timothy S Jones, Samuel MD Oliveira, Chris J Myers, Christopher A Voigt, and Douglas Densmore. Genetic circuit design automation with cello 2.0. *Nature protocols*, 17(4):1097–1113, 2022.

[7] Ally Huang, Peter Q Nguyen, Jessica C Stark, Melissa K Takahashi, Nina Donghia, Tom Ferrante, Aaron J Dy, Karen J Hsu, Rachel S Dubner, Keith Pardee, et al. Biobits™ explorer: A modular synthetic biology education kit. *Science advances*, 4(8):eaat5105, 2018.

[8] Edgar Perea, Ignacio Grossmann, Erik Ydstie, and Turaj Tahmassebi. Dynamic modeling and classical control theory for supply chain management. *Computers & Chemical Engineering*, 24(2-7):1143–1149, 2000.

[9] James D Blight, R Lane Dailey, and Dagfinn Gangsaas. Practical control law design for aircraft using multivariable techniques. *International Journal of Control*, 59(1):93–137, 1994.

[10] Mark W Spong and Masayuki Fujita. Control in robotics. *IEEE Control Systems Society*, 10(2):1–25, 2011.

[11] Timothy Frei and Mustafa Khammash. Adaptive circuits in synthetic biology. *Current Opinion in Systems Biology*, 28:100399, 2021.

[12] Victoria Hsiao, Anandh Swaminathan, and Richard M Murray. Control theory for synthetic biology: recent advances in system characterization, control design, and controller implementation for synthetic biology. *IEEE Control Systems Magazine*, 38(3):32–62, 2018.

[13] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[14] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996.

[15] Kemin Zhou and John Comstock Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

[16] Peter Dorato. A historical review of robust control. *IEEE Control Systems Magazine*, 7(2):44–47, 1987.

[17] Xin Qi, Murti V Salapaka, Petros G Voulgaris, and Mustafa Khammash. Structured optimal and robust control with multiple criteria: A convex solution. *IEEE Transactions on Automatic Control*, 49(10):1623–1640, 2004.

[18] Brian Munsky and Mustafa 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, 2008.

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

[20] Katherine A Riccione, Robert P Smith, Anna J Lee, and Lingchong You. A synthetic biology approach to understanding cellular information processing. *ACS synthetic biology*, 1(9):389–402, 2012.

[21] Ania Ariadna Baetica, Yoke Peng Leong, and Richard M. Murray. Guidelines for designing the antithetic feedback motif. *Physical Biology*, 17, 9 2020.

[22] Shaobin Guo and Richard M Murray. Construction of incoherent feedforward loop circuits in a cell-free system and in cells. *ACS synthetic biology*, 8(3):606–610, 2019.

[23] 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*, 6:2157–2167, 2017.

[24] Michael B. Sheets, Nathan Tague, and Mary J. Dunlop. An optogenetic toolkit for light-inducible antibiotic resistance. *Nature Communications*, 14, 12 2023.

[25] Dayu Lin, Maureen P Boyle, Piotr Dollar, Hyosang Lee, ES Lein, Pietro Perona, and David J Anderson. Functional identification of an aggression locus in the mouse hypothalamus. *Nature*, 470(7333):221–226, 2011.

[26] Guy Soffer, James M Perry, and Steve CC Shih. Real-time optogenetics system for controlling gene expression using a model-based design. *Analytical Chemistry*, 93(6):3181–3188, 2021.

[27] Dominik Niopek, Dirk Benzinger, Julia Roensch, Thomas Draebing, Pierre Wehler, Roland Eils, and Barbara Di Ventura. Engineering light-inducible nuclear localization signals for precise spatiotemporal control of protein dynamics in living cells. *Nature communications*, 5(1):4404, 2014.

[28] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[29] Michael B. Elowitz Gürol M. Süel, Rajan P. Kulkarni, Jonathan Dworkin, Jordi Garcia-Ojalvo. Tunability and Noise Dependence in Differentiation Dynamics. *Science*, 315(March):1716–1719, 2007.

[30] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81:2340–2361, 1977.

[31] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[32] Ramon Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of chemical physics*, 136(15), 2012.

[33] Abhyudai Singh and Joao Pedro Hespanha. Lognormal moment closures for biochemical reactions. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 2063–2068. IEEE, 2006.

[34] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of applied probability*, 4:413–478, 1967.

[35] Matthias Jeschke, Alfred Park, Roland Ewald, Richard Fujimoto, and Adelinde M Uhrmacher. Parallel and distributed spatial simulation of chemical reactions. In *2008 22nd Workshop on Principles of Advanced and Distributed Simulation*, pages 51–59. IEEE, 2008.

[36] Christoph Schmal, Peter Reimann, and Dorothee Staiger. A circadian clock-regulated toggle switch explains atgrp7 and atgrp8 oscillations in arabidopsis thaliana. *PLoS Computational Biology*, 9(3):e1002986, 2013.

[37] Stefan Müller, Josef Hofbauer, Lukas Endler, Christoph Flamm, Stefanie Widder, and Peter Schuster. A generalized model of the repressilator. *Journal of mathematical biology*, 53:905–937, 2006.

[38] Jiliang Hu, Daniel R Amor, Matthieu Barbier, Guy Bunin, and Jeff Gore. Emergent phases of ecological diversity and dynamics mapped in microcosms. *Science*, 378(6615):85–89, 2022.

[39] Anna Posfai, Thibaud Taillefumier, and Ned S Wingreen. Metabolic trade-offs promote diversity in a model ecosystem. *Physical review letters*, 118(2):028103, 2017.

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

[41] SR Kain, M Adams, A Kondepudi, TT Yang, WW Ward, and P Kitts. Green fluorescent protein as a reporter of gene expression and protein localization. *Biotechniques*, 19(4):650–655, 1995.

[42] David Ehrhardt. Gfp technology for live cell imaging. *Current Opinion in Plant Biology*, 6(6):622–628, 2003.

[43] Myoung Sup Shim, April Nettesheim, Joshua Hirt, and Paloma B Liton. The autophagic protein lc3 translocates to the nucleus and localizes in the nucleolus associated to nufip1 in response to cyclic mechanical stress. *Autophagy*, 16(7):1248–1261, 2020.

[44] Patrick J Macdonald, Yan Chen, and Joachim D Mueller. Chromophore maturation and fluorescence fluctuation spectroscopy of fluorescent proteins in a cell-free expression system. *Analytical biochemistry*, 421(1):291–298, 2012.

[45] Bin Wu, Carolina Eliscovich, Young J. Yoon, and Robert H. Singer. Translation dynamics of single mRNAs in live cells and neurons. *Science*, 352(6292):1430–1435, 2016.

[46] Andrea M Femino, Fredric S Fay, Kevin Fogarty, and Robert H Singer. Visualization of single rna transcripts in situ. *Science*, 280(5363):585–590, 1998.

[47] Samuel O Skinner, Leonardo A Sepúlveda, Heng Xu, and Ido Golding. Measuring mrna copy number in individual escherichia coli cells using single-molecule fluorescent in situ hybridization. *Nature protocols*, 8(6):1100–1113, 2013.

[48] Xavier Pichon, Amandine Bastide, Adham Safieddine, Racha Chouaib, Aubin Samacoits, Eugenia Basyuk, Marion Peter, Florian Mueller, and Edouard Bertrand. Visualization of single endogenous polysomes reveals the dynamics of translation in live human cells. *Journal of Cell Biology*, 214:769–781, 2016.

[49] Chong Wang, Boran Han, Ruobo Zhou, and Xiaowei Zhuang. Real-time imaging of translation on single mrna transcripts in live cells. *Cell*, 165:990–1001, 2016.

[50] Kenneth Lyon, Luis U Aguilera, Tatsuya Morisaki, Brian Munsky, and Timothy J Stasevich. Live-cell single rna imaging reveals bursts of translational frameshifting. *Molecular Cell*, 75(1):172–183, 2019.

[51] Charlotte A. Cialek, Gabriel Galindo, Tatsuya Morisaki, Ning Zhao, Taiowa A. Montgomery, and Timothy J. Stasevich. Imaging translational control by argonaute with single-molecule resolution in live cells. *Nature Communications*, 13, 12 2022.

[52] Amanda Koch, Luis Aguilera, Tatsuya Morisaki, Brian Munsky, and Timothy J Stasevich. Quantifying the dynamics of ires and cap translation with single-molecule resolution in live cells. *Nature structural & molecular biology*, 27(12):1095–1104, 2020.

[53] Makio Tokunaga, Naoko Imamoto, and Kumiko Sakata-Sogawa. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nature methods*, 5(2):159–161, 2008.

[54] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

[55] Arthur Imbert, Wei Ouyang, Adham Safieddine, Emeline Coleno, Christophe Zimmer, Edouard Bertrand, Thomas Walter, and Florian Mueller. Fish-quant v2: a scalable and modular tool for smfish image analysis. *Rna*, 28(6):786–795, 2022.

[56] Huy D Vo, Zachary Fox, Ania Baetica, and Brian Munsky. Bayesian estimation for stochastic gene expression using multifidelity models. *The Journal of Physical Chemistry B*, 123(10):2217–2234, 2019.

[57] Huy D Vo, Linda S Forero-Quintero, Luis U Aguilera, and Brian Munsky. Analysis and design of single-cell experiments to harvest fluctuation information while rejecting measurement noise. *Frontiers in Cell and Developmental Biology*, 11:1133994, 2023.

[58] Jeff Hasty, David McMillen, and J J Collins. Engineered gene circuits. *Nature*, 420:224 EP –, nov 2002.

[59] Tom Knight. Idempotent vector design for standard assembly of biobricks. *MIT Libraries*, pages 1–11, 2003.

[60] Seung Hoon Jang, Ji Won Cha, Nam Soo Han, and Ki Jun Jeong. Development of bicistronic expression system for the enhanced and reliable production of recombinant proteins in leuconostoc citreum. *Scientific Reports*, 8:1–11, 2018.

[61] Filippo Menolascina, Gianfranco Fiore, Emanuele Orabona, Luca De Stefano, Mike Ferry, Jeff Hasty, Mario di Bernardo, and Diego di Bernardo. In-vivo real-time control of protein expression from endogenous and synthetic gene networks. *PLOS Computational Biology*, 10:1–14, 2014.

[62] Timothy Gardner, Charles Cantor, and James Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403:339–342, 2000.

[63] Min Wu, Ri Qi Su, Xiaohui Li, Tom Ellis, Ying Cheng Lai, and Xiao Wang. Engineering of regulated stochastic cell fate determination. *Proceedings of the National Academy of Sciences of the United States of America*, 110:10610–10615, 2013.

[64] Beat P Kramer, Alessandro Usseglio Viretta, Marie Daoud-El Baba, Dominique Aubel, Wilfried Weber, and Martin Fussenegger. An engineered epigenetic transgene switch in mammalian cells. *Nature biotechnology*, 22:867–870, 2004.

[65] Katherine A Schaumberg, Mauricio S Antunes, Tessema K Kassaw, Wenlong Xu, Christopher S Zalewski, June I Medford, and Ashok Prasad. Quantitative characterization of genetic parts and circuits for plant synthetic biology. *Nature Methods*, 13:94–100, 2016.

[66] Andrew H. Ng, Taylor H. Nguyen, Mariana Gómez-Schiavon, Galen Dods, Robert A. Langan, Scott E. Boyken, Jennifer A. Samson, Lucas M. Waldburger, John E. Dueber, David Baker, and Hana El-Samad. Modular and tunable biological feedback control using a de novo protein switch, 2019.

[67] Tae Seok Moon, Chunbo Lou, Alvin Tamsir, Brynne C Stanton, and Christopher A Voigt. Genetic programs constructed from layered logic gates in single cells. *Nature*, 491:249 EP –, 10 2012.

[68] Thomas M. Groseclose, Ronald E. Rondon, Zachary D. Herde, Carlos A. Aldrete, and Corey J. Wilson. Engineered systems of inducible anti-repressors for the next generation of biological programming. *Nature Communications*, 11:1–15, 2020.

[69] Chang C. Liu, Michael C. Jewett, Jason W. Chin, and Chris A. Voigt. Toward an orthogonal central dogma. *Nature Chemical Biology*, 14:103–106, 2018.

[70] Tessema K. Kassaw, Alberto J. Donayre-Torres, Mauricio S. Antunes, Kevin J. Morey, and June I. Medford. Engineering synthetic regulatory circuits in plants. *Plant Science*, 273:13–22, 2018.

[71] Florian Lienert, Jason J Lohmueller, Abhishek Garg, and Pamela A Silver. Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nature reviews Molecular cell biology*, 15:95–107, 2014.

[72] Jonathan W. Young, James C.W. Locke, Alphan Altinok, Nitzan Rosenfeld, Tigran Bacarian, Peter S. Swain, Eric Mjolsness, and Michael B. Elowitz. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature Protocols*, 7:80–88, 2012.

[73] Nathan M. Belliveau, Stephanie L. Barnes, William T. Ireland, Daniel L. Jones, Michael J. Sweredoski, Annie Moradian, Sonja Hess, Justin B. Kinney, and Rob Phillips. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E4796–E4805, 2018.

[74] Edouard Bertrand, Pascal Chartrand, Matthias Schaefer, Shailesh M Shenoy, Robert H Singer, and Roy M Long. Localization of ash1 mrna particles in living yeast. *Molecular cell*, 2:437–445, 1998.

[75] Wentao Kong, David R. Meldgin, James J. Collins, and Ting Lu. Designing microbial consortia with defined social interactions. *Nature Chemical Biology*, 14:821–829, 2018.

[76] Michael J Liao, M Omar Din, Lev Tsimring, and Jeff Hasty. Rock-paper-scissors: Engineered population dynamics increase genetic stability michaelcrease genetic stability. *Science*, 365:1045–1049, 2019.

[77] Gianfranco Fiore, Antoni Matyjaszkiewicz, Fabio Annunziata, Claire Grierson, Nigel J. Savery, Lucia Marucci, and Mario Di Bernardo. In-silico analysis and implementation of a multicellular feedback control strategy in a synthetic bacterial consortium. *ACS Synthetic Biology*, 6:507–517, 2017.

[78] Matthew Jemielita, Ned S. Wingreen, and Bonnie L. Bassler. Quorum sensing controls vibrio cholerae multicellular aggregate formation. *eLife*, 7:1–25, 2018.

[79] M. Khammash, M. Di Bernardo, and D. Di Bernardo. Cybergenetics: Theory and methods for genetic control system. *Proceedings of the IEEE Conference on Decision and Control*, 2019-Decem:916–926, 2019.

[80] Michael B. Sheets, Wilson W. Wong, and Mary J. Dunlop. Light-inducible recombinases for bacterial optogenetics. *ACS Synthetic Biology*, 9:227–235, 2020.

[81] 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*, 70:745–756, 2018.

[82] Susan Y. Chen, Lindsey C. Osimiri, Michael Chevalier, Lukasz J. Bugaj, Taylor H. Nguyen, R. A. Greenstein, Andrew H. Ng, Jacob Stewart-Ornstein, Lauren T. Neves, and Hana El-Samad. Optogenetic control reveals differential promoter interpretation of transcription factor nuclear translocation dynamics. *Cell Systems*, 11:336–353.e24, 2020.

[83] Supravat Dey and Abhyudai Singh. Propagation of stochastic gene expression in the presence of decoys. *Proceedings of the IEEE Conference on Decision and Control*, 2020-Decem:5873–5878, 2020.

[84] Manuel Razo-Mejia, Sarah Marzen, Griffin Chure, Rachel Taubman, Muir Morrison, and Rob Phillips. First-principles prediction of the information processing capacity of a simple genetic circuit. *Physical Review E*, 102(2):022404, 2020.

[85] Tomislav Plesa, Guy-Bart Stan, Thomas E. Ouldridge, and Alex Dack. Integral feedback in synthetic biology: Negative-equilibrium catastrophe. pages 1–27, 2021.

[86] Brian Munsky, Gregor Neuert, and Alexander Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336:183–187, 2012.

[87] Daniel L. Jones, Robert C. Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346:1533–1536, 2014.

[88] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information transduction capacity of noisy biochemical signaling networks. *science*, 334:354–358, 2011.

[89] Taek Kang, Tyler Quarton, Chance M. Nowak, Kristina Ehrhardt, Abhyudai Singh, Yi Li, and Leonidas Bleris. Robust filtering and noise suppression in intragenic mirna-mediated host regulation. *iScience*, 23, 2020.

[90] Remy Chait, Jakob Ruess, Tobias Bergmiller, Gašper Tkačik, and Cualin C Guet. Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature communications*, 8:1–11, 2017.

[91] Jean Baptiste Lugagne, Sebastián Sosa Carrillo, Melanie Kirch, Agnes Köhler, Gregory Batt, and Pascal Hersen. Balancing a genetic toggle switch by real-time feedback control and periodic forcing. *Nature Communications*, 8:1–7, 2017.

[92] Agostino Guarino, Davide Fiore, Davide Salzano, and Mario Di Bernardo. Balancing cell populations endowed with a synthetic toggle switch via adaptive pulsatile feedback control. *ACS Synthetic Biology*, 9:793–803, 2020.

[93] Davide Fiore, Agostino Guarino, and Mario Di Bernardo. Analysis and control of genetic toggle switches subject to periodic multi-input stimulation. *IEEE Control Systems Letters*, 3:278–283, 2019.

[94] Zachary R Fox, Steven Fletcher, Achille Fraisse, Chetan Aditya, and Sebastián Sosa. Micromator: Open and flexible software for reactive microscopy. *bioRxiv*, pages 1–9, 2021.

[95] Paulina Szymańska, Nicola Gritti, Johannes M. Keegstra, Mohammad Soltani, and Brian Munsky. Using noise to control heterogeneity of isogenic populations in homogenous environments. *Physical Biology*, 12, 2015.

[96] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3:1871–1878, 2007.

[97] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G. Bowsher. Stochastic simulation of biomolecular networks in dynamic environments. *PLoS Computational Biology*, 12, 6 2016.

[98] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149:1633–1648, 1998.

[99] Paul C. Bressloff. Stochastic switching in biology: From genotype to phenotype. *Journal of Physics A: Mathematical and Theoretical*, 50, 2017.

[100] Peter L. Freddolino and Saeed Tavazoie. The dawn of virtual cell biology. *Cell*, 150:248–250, 2012.

[101] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. MacKlin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150:389–401, 2012.

[102] Tibor Kalmar, Chea Lim, Penelope Hayward, Silvia Muñoz-Descalzo, Jennifer Nichols, Jordi Garcia-Ojalvo, and Alfonso Martinez Arias. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7:33–36, 2009.

[103] Hiroshi Ochiai, Takeshi Sugawara, Tetsushi Sakuma, and Takashi Yamamoto. Stochastic promoter activation affects nanog expression variability in mouse embryonic stem cells. *Scientific Reports*, 4:1–9, 2014.

[104] Shuai Gong, Qiuhui Li, Collene R. Jeter, Qingxia Fan, Dean G. Tang, and Bigang Liu. Regulation of nanog in cancer cells. *Molecular Carcinogenesis*, 54:679–687, 2015.

[105] Robert A. Langan, Scott E. Boyken, Andrew H. Ng, Jennifer A. Samson, Galen Dods, Alexandra M. Westbrook, Taylor H. Nguyen, Marc J. Lajoie, Zibo Chen, Stephanie Berger, Vikram Khipple Mulligan, John E. Dueber, Walter R.P. Novak, Hana El-Samad, and David Baker. De novo design of bioactive protein switches. *Nature*, 572:205–210, 2019.

[106] Bin Shao, Jayan Rammohan, Daniel A. Anderson, Nina Alperovich, David Ross, and Christopher A. Voigt. Single-cell measurement of plasmid copy number and promoter activity. *Nature Communications*, 12, 2021.

[107] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188:404–425, 1992.

[108] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124, 2006.

[109] Shuai Liang, Bin Xu, and Jinrui Ren. Kalman-filter-based robust control for hypersonic flight vehicle with measurement noises. *Aerospace Science and Technology*, 112, 5 2021.

[110] Sergio Lucia, Tiago Finkler, and Sebastian Engell. Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. *Journal of Process Control*, 23:1306–1319, 2013.

[111] SP Bhattacharyya. Robust control under parametric uncertainty: An overview and recent results. *Annual Reviews in Control*, 44:45–77, 2017.

[112] Yu Feng and Hongda Sun. Robust optimal control for discrete-time lti systems over multiple additive white gaussian noise channels. *IEEE Transactions on Automatic Control*, 2022.

[113] Mengyi Sun and Jianzhi Zhang. Allele-specific single-cell rna sequencing reveals different architectures of intrinsic and extrinsic gene expression noises. *Nucleic Acids Research*, 48:533–547, 1 2020.

[114] Sara Dionisi, Karol Piera, Armin Baumschlager, and Mustafa Khammash. Implementation of a novel optogenetic tool in mammalian cells based on a split t7 rna polymerase. *ACS Synthetic Biology*, 11:2650–2661, 8 2022.

[115] Gabriele Lillacci, Yaakov Benenson, and Mustafa Khammash. Synthetic control systems for high performance gene expression in mammalian cells. *Nucleic acids research*, 46:9855–9863, 2018.

[116] Armin Baumschlager and Mustafa Khammash. Synthetic biological approaches for optogenetics and tools for transcriptional light-control in bacteria. *Advanced Biology*, 5:2000256, 2021.

[117] Maurice Filo, Sant Kumar, and Mustafa Khammash. A hierarchy of biomolecular proportional-integral-derivative feedback controllers for robust perfect adaptation and dynamic performance. *Nature Communications*, 13, 12 2022.

[118] Hideki Kobayashi, Mads Kærn, Michihiro Araki, Kristy Chung, Timothy S Gardner, Charles R Cantor, and James J Collins. Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences*, 101:8414–8419, 2004.

[119] Michael May and Brian Munsky. Exploiting noise , nonlinearity , and feedback to differentially control multiple synthetic cells with a single optogenetic input. pages 1–28, 2021.

[120] Slaven Peleš, Brian Munsky, and Mustafa Khammash. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *The Journal of chemical physics*, 125(20), 2006.

[121] Shev MacNamara, Alberto M Bersani, Kevin Burrage, and Roger B Sidje. Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation. *The Journal of chemical physics*, 129(9), 2008.

[122] José Juan Tapia, James R Faeder, and Brian Munsky. Adaptive coarse-graining for transient and quasi-equilibrium analyses of stochastic gene regulation. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5361–5366. IEEE, 2012.

[123] Xiaodong Cai. Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of chemical physics*, 126(12), 2007.

[124] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[125] Michael Strasser, Fabian J. Theis, and Carsten Marr. Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophysical Journal*, 102:19–29, 1 2012.

[126] Evelina Tutucci, Maria Vera, Jeetayu Biswas, Jennifer Garcia, Roy Parker, and Robert H Singer. An improved ms2 system for accurate reporting of the mrna life cycle. *Nature methods*, 15(1):81–89, 2018.

[127] Maria Vera, Evelina Tutucci, and Robert H Singer. Imaging single mrna molecules in mammalian cells using an optimized ms2-mcp system. *Imaging Gene Expression: Methods and Protocols*, pages 3–20, 2019.

[128] Henry Pinkard, Nico Stuurman, Ivan E Ivanov, Nicholas M Anthony, Wei Ouyang, Bin Li, Bin Yang, Mark A Tsuchida, Bryant Chhun, Grace Zhang, et al. Pycro-manager: open-source software for customized and reproducible microscope control. *Nature methods*, 18(3):226–228, 2021.

[129] Nico Stuurman, Nenad Amdodaj, and Ron Vale. μmanager: open source software for light microscope imaging. *Microscopy Today*, 15(3):42–43, 2007.

[130] Zachary R Fox, Steven Fletcher, Achille Fraisse, Chetan Aditya, Sebastián Sosa-Carrillo, Julienne Petit, Sébastien Gilles, François Bertaux, Jakob Ruess, and Gregory Batt. Enabling reactive microscopy with micromator. *Nature Communications*, 13(1):2199, 2022.

[131] Angel Carro, Manuel Perez-Martinez, Joaquim Soriano, David G Pisano, and Diego Megias. imsrc: converting a standard automated microscope into an intelligent screening platform. *Scientific reports*, 5(1):10502, 2015.

[132] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022.

[133] Arthur Edelstein, Nenad Amodaj, Karl Hoover, Ron Vale, and Nico Stuurman. Computer control of microscopes using μmanager. *Current protocols in molecular biology*, 92(1):14–20, 2010.

[134] Linda S Forero-Quintero, William Raymond, Tetsuya Handa, Matthew N Saxton, Tatsuya Morisaki, Hiroshi Kimura, Edouard Bertrand, Brian Munsky, and Timothy J Stasevich. Live-cell imaging reveals the spatiotemporal organization of endogenous rna polymerase ii phosphorylation at a single gene. *Nature Communications*, 12(1):3158, 2021.

Appendix A

Supplemental Figures

This supplemental information contains additional details from Chapter one about the specification and optimization of model-controller pairs as well as supplemental figures to support the minor results that are discussed in the main manuscript.

A.0.1 Formulation and Optimization of Feedback Control Designs

The general tensor form of the FSP under state control can be written:

$$\frac{\partial P^i}{\partial t} = ([\mathbf{A}_0]^i_j + [\mathbf{Bu}]^i_j)P^j. \tag{A.1}$$

For the partial observation situation where not all states can be observed, we reformulate the FSP analysis as:

$$\frac{\partial P^i}{\partial t} = (A^i_j + B^i_{jm}M^m_\alpha v^\alpha)P^j, \tag{A.2}$$

where the indices $\{i, j, m\}$ refer to states in the Markov chain, and index α refers to the distinctly observable subset of those states that define the control inputs; $\mathbf{B}$ is the controller tensor as in the main text; and $\mathbf{v}$ the vector of control signals associated with each distinctly observable state. The control scheme used (i.e., FAC, PAC, or UAC) is defined $\mathbf{u} = \mathbf{Mv}$, where $\mathbf{M}$ is a lifting operator which takes the controller as defined on the observable state space $\mathbf{v}$ and lifts it into the proper dimensions to multiply with the control tensor. For the FAC control scheme, the number of elements in $\mathbf{v}$ matches the total number of states (i.e., $\mathbf{u} = \mathbf{v}$) meaning that $\mathbf{M}$ simply an identity matrix. For less than complete observation as in the PAC or UAC schemes, a single element of $\mathbf{v}$ can influence multiple parts of the state space, and $\mathbf{M}$ is a tall rectangular matrix that contains only ones or zeros, and $\mathbf{u} = \mathbf{Mv}$ describes the light activity in each state, while the lower dimensional $\mathbf{v}$ describes the smaller set of unique values of the control input for distinct observable states. With

the UAC, the $\mathbf{M}$ matrix is represented by the $\mathbf{N} \times 1$ matrix filled with only ones and $\mathbf{v}$ is a scalar quantity.

Deriving how $\mathbf{P}$ changes with small changes in $\mathbf{v}$ at steady state gives

$$\frac{\partial \dot{P}^i}{\partial u^n} = 0^i_n = \partial_n((A^i_j + B^i_{jm} M^m_k v^k) P^j), \tag{A.3}$$

$$0^i_n = (B^i_{jm} M^m_k \delta^k_n) P^j + ((A^i_j + B^i_{jm} M^m_k v^k) P^j_n, \tag{A.4}$$

$$\frac{\partial P^j}{\partial u^n} = -[(A + BMv)^{-1}]^j_i B^i_{ko} M^o_n P^k. \tag{A.5}$$

Finally, plugging this into the definition for the objective score (Eq. 5 in main text), we have:

$$\frac{\partial J}{\partial v^n} = \frac{\partial (C_j P^j)}{\partial v^n} = C_j P^j_{,n} = -C_j [(A + BMv)^{-1}]^j_i B^i_{ko} M^o_n P^k. \tag{A.6}$$

With this expression in hand, $\mathbf{v}$ can be optimized by starting at an appropriate $\mathbf{P}$ and changing $\mathbf{v}$ in the direction of the negative gradient $(-J_{,n})$, and then updating the new $\mathbf{P}$. The process continues iteratively until convergence to a local minimum.

A.0.2 Model Predictive Control

The probabilistic model predictive control (PMPC) uses a simple linear machine to generate light inputs, $\mathbf{u}(t)$, according to

$$\mathbf{u}(t) = \max\{\mathbf{0}, \mathbf{c} + \mathbf{Z}\mathbf{P}^i_{\mathrm{nt}}(t)\}, \tag{A.7}$$

where $\tilde{\mathbf{P}}_{\mathrm{nt}}$ denotes the random probability distribution of the unobserved cells given the history of $\mathbf{u}(\tau)$ for $\tau \in (0, t)$; the vector $\mathbf{c}$ provides the bias of the control signal with one entry for every possible value of the observed cell's value; and $\mathbf{Z}$ is a matrix which takes $\mathbf{P}_{\mathrm{nt}}$ and outputs adjustments to the controller based on the probabilistic predictions for the unobserved cells. In this formulation, each row of $\mathbf{c}$ and $\mathbf{Z}$ represents the deterministic control bias and probabilistic correction for each state of the observed cell, given the history of the control signal. A heuristic

optimization was performed on the system to tune the entries of $\mathbf{c}$ and $\mathbf{Z}$ and then simulating the process for long time periods. In circumstances where the proposed controller produces a non-physical negative control signals, the control signal is set to zero. Figure A.1(A and B) shows the weights of $\mathbf{c}$ and $\mathbf{Z}$ after joint optimization was performed on the system.

A.0.3 Distribution of Objective Scores

Analyses presented in the main text present the objective score, J in terms of the expected Euclidean distance from the target state (Eq. 5 in main text). Certain controllers yield high variability in this distance over time in a single trajectory or if sampled for many different cells at a specific time point. To analyze this variability, SSA simulations were performed using each set of controllers and the score at each time point was measured. Histograms of score for time-series trajectories using the FAC, PAC, and pMPC show a long tail distribution of the score (Fig. A.1C). These data taken together suggest that the score is often much lower than the expected value, but the average performance, J, is dominated by rare moments in time where the performance is poor, causing a large temporary increase in score and raising the average despite overall good performance. Small changes to the highly weighted tails can yield better average scores while returning remarkably similar distributions.

A.0.4 Quantification of Controller Performance for Multiple Cells

We considered a set of five controllers, including the FAC, PAC, and pMPC controllers already described in the main text in addition to two adhoc controller that use (i) the mean of the FAC controller over all cells to be driven to the second target state (MFAC):

$$u_{\text{MFAC}}(\{x_1, x_2, \ldots\}, x_N) =< u_{\text{FAC}}(x_1, x_{i>1}) >, \tag{A.8}$$

and (2) the FAC controller evaluated at the mean of all unobserved cells (FACM):

$$u_{\text{FACM}}(\{x_1, x_2, \ldots\}, x_N) = u_{\text{FAC}}(x_1, < x_{i>1} >). \tag{A.9}$$

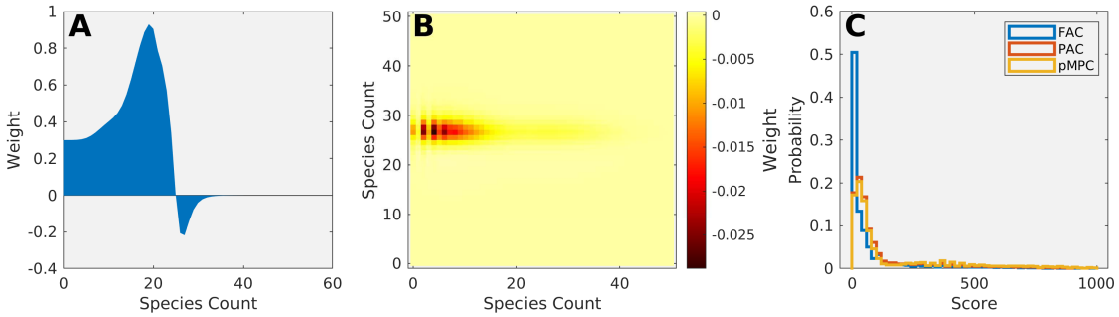

Figure A.1: Visualization of pMPC Control Law. (A) Weights of **c** show that the pMPC tends to increase the controller when the observed cell is below 20, but tends to decrease the control signal when $\mathbf{P}_n t$ is weighted above 20 and turns off when the observed cell is above 30. (Z) Weights of Z show that the pMPC optimization preferred to weight the control input down when both the observed cell was above twenty five and the unobserved cell was near ten. (C) Distribution of scores obtained during time trajectory show that score over time is a heavy tailed distribution. Although the probability of a high score is low, the score value itself tends to be vary large and can increase variability in simulations as well as attributing large differences in score for similar looking distributions.

To quantify the performance of each controller under varying numbers of cells, we generated 32 independent simulations over 99,000 minutes (following a 1,000 minute burn in period) while sampling every 100 minutes using two, four, eight, and sixteen cells in the second group. The objective score for each simulation was computed by averaging J over its corresponding trajectory. The lines depicted shown in Fig. A.2A show the median of the 32 independent objective scores (colored line) as well as the 25th and 75th percentile (shaded regions) versus the number of cells. These data in Fig. A.2 show that the performance of the FAC, MFAC, and FACM outperform the UAC, but this performance decreases as the number of cells increases. However, the PAC and pMPC performance is independent from the number of cells. These data taken together suggest that the PAC and pMPC are both better controllers once the number of cells is greater than or equal to two, and attempts to extend the FAC to work on summary statistics from many cells does not appear to be effective. Improved FAC controllers would be possible using higher order tensor control formulations (as opposed to summary statistics described here), but the computational complexity for the design of such control algorithms is currently prohibitive and left to future work.

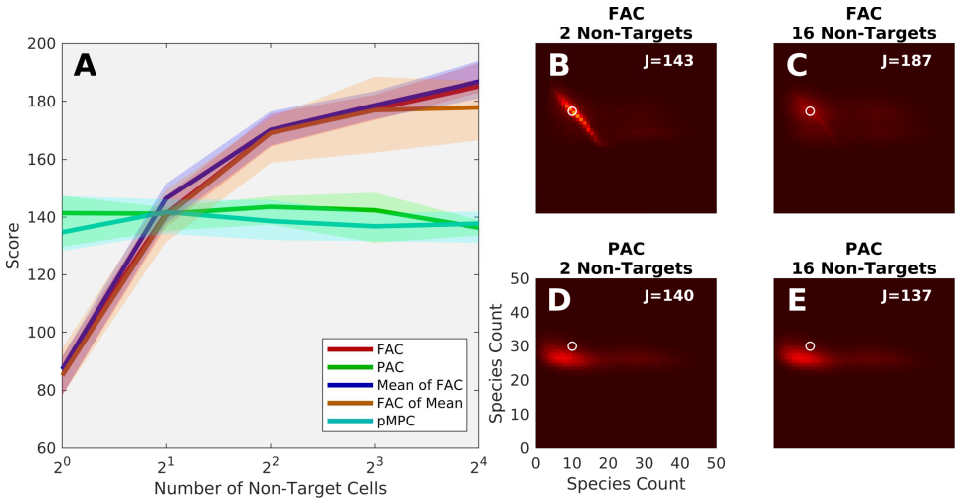

Figure A.2: Performance of stochastic controllers using varying numbers of cells. (A) median score of 32 simulations using a set of five controllers in colored lines, with 25% and 75% quartiles shown in the color-shaded region. (B and C) FAC controller joint distribution of two cells chosen from a set of two (B) or sixteen (C) cells total, shows rapid degradation of performance when more cells are considered. (D and E) PAC controller joint distribution of two cells chosen from a set of two (D) or sixteen (E) cells total shows no change in the joint distribution.