# Optimal Design of Single-Cell Experiments within Temporally Fluctuating Environments

Zachary Fox, Gregor Neuert, Brian Munsky

Hindawi
Complexity
Volume 2020, Article ID 8536365, 15 pages
https://doi.org/10.1155/2020/8536365

*Research Article*

# Optimal Design of Single-Cell Experiments within Temporally Fluctuating Environments

**Zachary R. Fox,[1,2,3] Gregor Neuert,[4,5,6] and Brian Munsky[3,7]**

[1]*Inria Saclay Ile-de-France, Palaiseau 91120, France*
[2]*Institut Pasteur, USR 3756 IP CNRS, Paris 75015, France*
[3]*School of Biomedical Engineering, Colorado State University, Fort Collins, CO 80523, USA*
[4]*Department of Molecular Physiology and Biophysics, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA*
[5]*Department of Biomedical Engineering, School of Engineering, Vanderbilt University, Nashville, TN 37232, USA*
[6]*Department of Pharmacology, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA*
[7]*Department of Chemical and Biological Engineering, Colorado State University Fort Collins, CO 80523, USA*

Correspondence should be addressed to Brian Munsky; munsky@colostate.edu

Received 20 October 2019; Accepted 12 February 2020; Published 13 June 2020

Guest Editor: George V. Popescu



Modern biological experiments are becoming increasingly complex, and designing these experiments to yield the greatest possible quantitative insight is an open challenge. Increasingly, computational models of complex stochastic biological systems are being used to understand and predict biological behaviors or to infer biological parameters. Such quantitative analyses can also help to improve experiment designs for particular goals, such as to learn more about specific model mechanisms or to reduce prediction errors in certain situations. A classic approach to experiment design is to use the Fisher information matrix (FIM), which quantifies the expected information a particular experiment will reveal about model parameters. The finite state projection-based FIM (FSP-FIM) was recently developed to compute the FIM for discrete stochastic gene regulatory systems, whose complex response distributions do not satisfy standard assumptions of Gaussian variations. In this work, we develop the FSP-FIM analysis for a stochastic model of stress response genes in *S. cerevisiae* under time-varying MAPK induction. We verify this FSP-FIM analysis and use it to optimize the number of cells that should be quantified at particular times to learn as much as possible about the model parameters. We then extend the FSP-FIM approach to explore how different measurement times or genetic modifications help to minimize uncertainty in the sensing of extracellular environments, and we experimentally validate the FSP-FIM to rank single-cell experiments for their abilities to minimize estimation uncertainty of NaCl concentrations during yeast osmotic shock. This work demonstrates the potential of quantitative models to not only make sense of modern biological datasets but to close the loop between quantitative modeling and experimental data collection.

## 1. Introduction

The standard approach to design experiments has been to rely entirely on expert knowledge and intuition. However, as experimental investigations become more complex and seek to examine systems with more subtle nonlinear interactions, it becomes much harder to improve experimental designs using intuition alone. This issue has become especially relevant in modern single-cell-single-molecule investigations of gene regulatory processes. Performing such powerful, yet complicated, experiments involves the selection from among a large number of possible experimental designs, and it is often not clear which designs will provide the most relevant information. A systematic approach to solve this problem is model-driven experiment design, in which one combines existing knowledge or experience to form an assumed (and partially incorrect) mathematical model of the system to estimate and optimize the value of potential experimental settings. In practice, such preliminary models would be defined by existing data taken in simpler or more general settings such as inexpensive bulk experiments or would be estimated from literature values conducted on

similar genes, pathways, or organisms. When parameter or model structures are uncertain, these could be described according to a prior distribution, and experiments would need to be selected according to which performs best on average across the many possible model/parameter combinations.

In recent years, model-driven experiment design has gained traction for biological models of gene expression, whether in the Bayesian setting [1] or using Fisher information for deterministic models [2], and even in the stochastic, single-cell setting [3–7]. Despite the promise and active development of model-driven experiment design from the theoretical perspective, more general, yet biologically inspired, approaches are needed to make these methods suitable for the experimental community at large. In this work, we apply model-driven experiment design to an experimentally validated model of stochastic transcription that is activated by time-varying high osmolarity glycerol (HOG) mitogen-activated protein kinase (MAPK) induction in yeast [8–10]. To demonstrate a concrete and practical application of model-driven experiment design, we find the optimal *measurement schedule* (i.e., when measurements ought to be taken) and the appropriate *number of individual cells* to be measured at each time point.

In our computational analyses, we consider the experimental technique of single-molecule mRNA fluorescence *in situ* hybridization (smFISH), where specific fluorescent oligonucleotide probes are hybridized to mRNA of interest in fixed cells [11, 12]. Cells are then imaged, and the mRNA abundance in each cell is counted, either by hand or using automated software such as [13]. Such counting can be a cumbersome process, but little thought has been given typically to how many cells should be measured and analyzed at each time. Furthermore, when a dynamic response is under investigation, the specific times at which measurements should be taken (i.e., the times after induction at which cells should be fixed and analyzed) are also unclear. In this work, we use the newly developed finite state projection-based Fisher information matrix (FSP-FIM, [6]) to optimize these experimental quantities for osmotic stress response genes in yeast.

The first part of our current study introduces a discrete stochastic model to analyze time-varying MAPK-induced gene expression response in yeast and then demonstrates the use of FSP-based Fisher information to optimize experiments to minimize the uncertainty in model parameters. In the second part of this study, we expand upon this result to find and experimentally verify the optimal smFISH measurement times and cell numbers to minimize uncertainty about unknown environmental inputs (e.g., salt concentrations) to which the cells are subjected. In this way, we are presenting a new methodology by which one can optimally examine behaviors of natural cells to obtain accurate estimations of environmental changes.

## 2. Background

Gene regulation is the process by which small molecules, chromatin regulators, and general and gene-specific transcription factors interact to regulate the transcription of DNA into RNA and the translation of mRNA into proteins. Even within populations of genetically identical cells, these single-molecule processes are stochastic and give rise to cell-to-cell variability in gene expression levels. Adequate descriptions of such variable responses can only be achieved through the use of stochastic computational models [14–17]. In the following sections, we first introduce a nonequilibrium discrete stochastic model of HOG1-MAPK-induced gene expression, and we then discuss how this model can be analyzed and compared to data using finite state project analyses. All analysis codes are available at https://github.com/MunskyGroup/Fox_Complexity_2020.

*2.1. Discrete Stochastic Model of HOG1-MAPK-Induced Gene Expression.* To motivate and demonstrate our new approach, we focus our examination on the dynamics of the HOG1-MAPK pathway in yeast, which is a model system to study osmotic stress driven dynamics of signal transduction and gene regulation in single cells [18–23]. Discrete stochastic models of HOG1-MAPK-activated transcription have been used successfully to predict the variability in adaptive transcription responses across yeast cell populations [9, 10, 24]. In particular, the authors in [9] used smFISH data to fit and cross validate a number of different potential models with different numbers of gene states and time-varying parameters. They found that dynamics of two stress response genes, *STL*1 and *CTT*1, could each be described accurately by the model depicted in Figure 1(a).

In brief, the model [9] consists of transitions between four different gene states (S1, S2, S3, and S4). The probability of a transition from the $i^{\text{th}}$ to the $j^{\text{th}}$ gene state in the infinitesimal time d$t$ is given by the propensity function, $k_{\text{ij}}\text{d}t$. Most of the rates $\{k_{\text{ij}}\}$ are constant in time, except for the transition from S2 to S1, which is controlled by the time-varying level of the HOG1-MAPK signal in the nucleus, $f(t)$. The resulting time-varying rate $k_{21}$ is defined using a linear threshold function:

$$k_{21}(t) = \max[0, \alpha - \beta f(t)], \tag{1}$$

where $\alpha$ and $\beta$ set the threshold for $k_{21}(t)$ activation/deactivation. The function $f(t)$ was calibrated at several NaCl concentrations by fitting the HOG1-MAPK nuclear localization signals as measured using a yellow fluorescence protein reporter [10]. Figure 1(b) shows $f(t)$ for osmotic stress responses to 0.2 M and 0.4 M NaCl, and Figure 1(c) shows the corresponding values of $k_{21}(t)$. In addition to the state transition rates, each $i^{\text{th}}$ state also has a corresponding mRNA transcription rate, $k_{\text{ri}}$. All mRNA molecules degrade with rate $\gamma$, independent of gene state. Further descriptions and validations of this model are given in Supplementary Note 1 and in [9, 10, 24]. All experimentally determined parameters for the *STL*1 and *CTT*1 transcription regulation models are provided in Supplemental Table S1, and experimentally determined parameters for the HOG1-MAPK signal model are listed in Supplemental Table S2 [10].

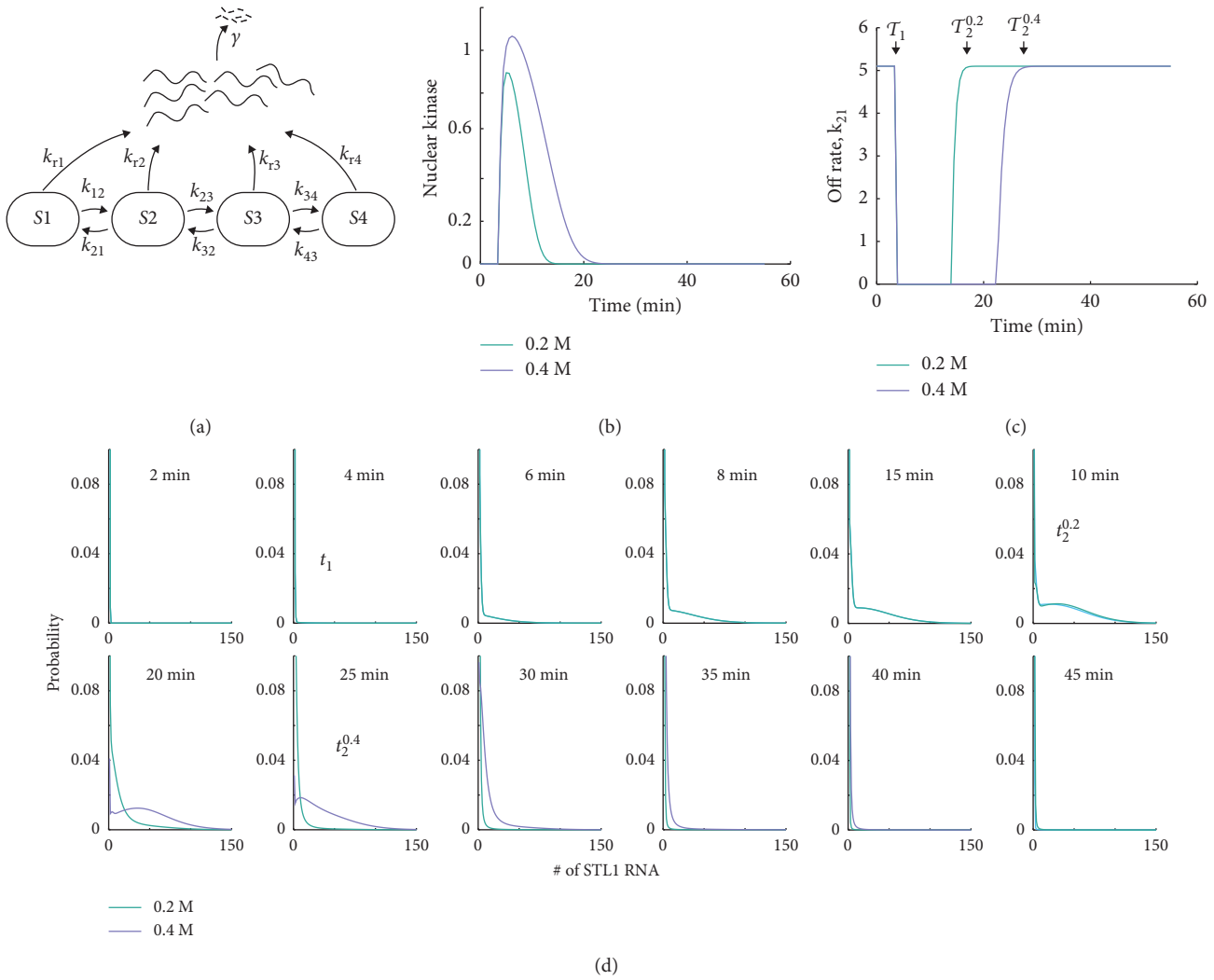


Figure 1: Stochastic modeling of osmotic stress response genes in yeast. (a) Four-state model of gene expression, where each state transcribes mRNA at a different transcription rate, but each mRNA degrades at a single rate $\gamma$. (b) Time-varying MAPK nuclear localization signal. (c) The rate of switching from gene activation state S2 to S1 (right) under 0.2 M or 0.4 M NaCl osmotic stress. The time at which $k_{21}$ turns off is denoted with $\tau_1$ and is independent of the NaCl level. The time at which $k_{23}$ turns back on is given by $\tau_{\text{NaCl}}$ depending on the level of NaCl. (d) Time evolution of the STL1 mRNA in response to the 0.2 M and 0.4 M NaCl stress. Model and parameters from [10] are summarized in Supplementary Notes I and II and Supplementary Tables I and II.

*2.2. The Finite State Projection Analysis of Stochastic Gene Expression.* To analyze the model described above, we apply the chemical master equation (CME) framework of stochastic chemical kinetics [25]. Combining the time-varying and constant state transition rates $\{k_{ij}\}$, transcription rates $\{k_{ri}\}$, and degradation rate $\gamma$ from above, the CME can be written in matrix form as a linear ordinary differential equation, $d\mathbf{p}/dt = \mathbf{A}(t)\mathbf{p}$, where the time-varying matrix $\mathbf{A}(t)$ is known as the infinitesimal generator (see Supplementary Note 1). The CME has been the workhorse of stochastic modeling of gene expression, and it is usually analyzed using simulated sample paths of its solution via the stochastic simulation algorithm [26] or with moment approximations [8, 27]. Alternatively, the CME can also be solved with guaranteed errors using the FSP approach [28, 29], which reduces the full CME only to describe the flow of probability among the most likely observable states of the system. Details of the FSP approach to solving chemical kinetic systems are provided in Supplementary Note 1. Application of the FSP analysis to the model (Figure 1(a)) with dynamic Hog1 (Figure 1(b)) modulates time-varying rates $k_{21}$ (Figure 1(c)) and predicts time-evolving probability distributions at 0.2 M and 0.4 M NaCl, as shown in Figure 1(d) [10].

*2.3. Likelihood of smFISH Data for FSP Models.* Recently, it has come to light that for some systems, it is critical to consider the full distribution of biomolecules across cellular populations when fitting CME models [6, 10]. To match CME model solutions to single-cell smFISH data, one needs to compute and maximize the likelihood of the data given the CME model [9, 10, 24, 30]. Fortunately, the FSP approach allows for computation of the likelihood with

guaranteed accuracy bounds [28]. We assume that measurements at each time point $\mathbf{t} \equiv [t_1, t_2, \ldots, t_{N_t}]$ are independent, as justified by the fact that fixation of cells for measurement precludes temporal cell-to-cell correlations. Measurements of $N_c$ cells can be concatenated into a matrix $\mathbf{D}_t \equiv [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{N_c}]_t$ of the observable mRNA species at each measurement time $t$.

The likelihood of making the independent observations for all $N_c$ measured cells is the product of the probabilities of observing each cell's measured state. For most gene expression models, however, states are only partially observable, and we define the observed state $\mathbf{x}_i^L$ as the marginalization (or lumping) over all full states $\{\mathbf{x}_j\}_i$ that are indistinguishable from $\mathbf{x}_i$ based on the observation. For example, the model of *STL*1 transcription consists of four gene states (S1–S4, shown in Figure 1(a)), which are unobserved, and the measured number of mRNA, which is observed. If we let index $i$ denote the number of mRNA, then the observed state $\mathbf{x}_i^L$ would lump together the full states (S1, $i$), (S2, $i$), (S3, $i$), and (S4, $i$). We next define $y_i$ as the number of experimental cells that match $\mathbf{x}_i^L$ at time $t$. Under these definitions, the likelihood of the observed data (and its logarithm) given the model can be written as

$$\begin{aligned} \ell(\mathbf{D};\boldsymbol{\theta}) &= M \prod_{t=t_1}^{t_{N_t}} \prod_{i \in \mathcal{J}_D} p\left(\mathbf{x}_i^L; t, \boldsymbol{\theta}\right)^{y_i}, \\ \log \ell(\mathbf{D};\boldsymbol{\theta}) &= \sum_{t=t_1}^{t_{N_t}} \sum_{i \in \mathcal{J}_D} y_i \log\left(p\left(\mathbf{x}_i^L; t, \boldsymbol{\theta}\right)\right) + \log M, \end{aligned} \tag{2}$$

where $\mathcal{J}_D$ is the set of states observed in the data, $M$ is a combinatorial prefactor (i.e., from a multinomial distribution) that comes from the arbitrary reordering of measured data, and $p(\mathbf{x}_i^L; t, \boldsymbol{\theta})$ is the marginalized probability mass of the observable species:

$$p\left(\mathbf{x}_i^L; t, \boldsymbol{\theta}\right) = \sum_{\mathbf{x}_j \in \mathbf{x}_i^L} p\left(\mathbf{x}_j; t, \boldsymbol{\theta}\right). \tag{3}$$

The vector of model parameters is denoted by $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots]$. Neglecting the term $\log M$, which is independent of the model, the summation in equation (2) can be rewritten as a product $\mathbf{y} \log \mathbf{p}^L$, where $\mathbf{y} \equiv [y_0, y_1, \ldots]$ is the vector of the binned data and $\mathbf{p}^L = [p(\mathbf{x}_0^L), p(\mathbf{x}_1^L), \ldots]^{\mathrm{T}}$ is the corresponding marginalized probability mass vector. One may then maximize equation (2) with respect to $\boldsymbol{\theta}$ to find the *maximum likelihood estimate* (MLE) of the parameters, $\widehat{\boldsymbol{\theta}}$, which will vary depending on each new set of experimental data. We next demonstrate how this likelihood function and the FSP model of the HOG1-MAPK-induced gene expression system can be used to design optimal smFISH experiments using the FSP-based Fisher information matrix [6].

## 3. Results

*3.1. The Finite State Projection-Based Fisher Information for Models of Signal-Activated Stochastic Gene Expression.* The Fisher information matrix (FIM) is a common tool in engineering and statistics to estimate parameter uncertainties prior to collecting data, which allows one to find experimental settings that can make these uncertainties as small as possible [3, 4, 31–34]. Recently, it has been applied to biological systems to estimate kinetic rate parameters in stochastic gene expression systems [3–6, 35]. In general, the FIM for a single measurement is defined as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left\{\left(\nabla_{\boldsymbol{\theta}} \log \mathbf{p}(\boldsymbol{\theta})\right)^T \left(\nabla_{\boldsymbol{\theta}} \log \mathbf{p}(\boldsymbol{\theta})\right)\right\}, \tag{4}$$

where the vector $\log \mathbf{p}(\boldsymbol{\theta})$ contains the log-probabilities of each potential observation and the expectation is taken over the probability distribution of states $\mathbf{p}(\boldsymbol{\theta})$ assuming the specific parameter set $\boldsymbol{\theta}$. As the number of measurements, $N_c$, is increased such that maximum likelihood estimates (MLE) of parameters are unbiased, the distribution of MLE estimates is known to approach a multivariate Gaussian distribution with a covariance given by the inverse of the FIM, i.e.,

$$\sqrt{N_c}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \xrightarrow{\text{dist}} \mathcal{N}\left(0, \mathcal{I}\left(\boldsymbol{\theta}^*\right)^{-1}\right). \tag{5}$$

In [6], we developed the FSP-based Fisher information matrix (FSP-FIM), which allows one to use the FSP solution $\mathbf{p}(t)$, and its sensitivity $\mathbf{s}_{\theta_j} \equiv \mathrm{d}\mathbf{p}/\mathrm{d}\theta_j$, to find the FIM for stochastic gene expression systems. For a general FSP model, the dynamics of the sensitivity to each $j^{\text{th}}$ kinetic parameter $\mathrm{d}\mathbf{p}/\mathrm{d}\theta_j$ can be calculated according to

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} \mathbf{p} \\ \mathbf{s}_{\theta_j} \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t) & 0 \\ \mathbf{A}_{\theta_j}(t) & \mathbf{A}(t) \end{bmatrix}\begin{bmatrix} \mathbf{p} \\ \mathbf{s}_{\theta_j} \end{bmatrix}, \tag{6}$$

where $\mathbf{A}_{\theta_j} = \partial \mathbf{A}/\partial \theta_j$. Solving equation (6) requires integrating a coupled set of ODEs that is twice as large as the original FSP system. The FSP-FIM at a single time $t$ is then given by

$$\mathbf{F}(\boldsymbol{\theta}, t)_{j,k} = \sum_i \frac{1}{p\left(\mathbf{x}_i; t, \boldsymbol{\theta}\right)} \mathbf{s}_{\theta_j}^i(t) \mathbf{s}_{\theta_k}^i(t), \tag{7}$$

where the summation is taken over all states $\{\mathbf{x}_i\}$ included in the FSP analysis (or over all observed states $\{\mathbf{x}_i^L\}$ in the case of lumped observations). We note that the FSP computation of the FIM should be computationally tractable for problems for which the FSP solution itself is tractable. However, since the size of the FSP sensitivity matrix (equation (6)) scales exponentially with the number of species, practical applications of the presented formulation of the FSP-FIM are currently restricted to models that have, or can be reduced to have, three or fewer distinct chemical species.

The FIM for a sequence of measurements taken independently (e.g., for smFISH data) at times $\mathbf{t} = [t_1, t_2, \ldots, t_{N_t}]$ can be calculated as the sum across the measurement times:

$$\mathcal{I}(\boldsymbol{\theta}, \mathbf{t}, \mathbf{c}) = \sum_{l=1}^{N_t} c_l \mathbf{F}\left(\boldsymbol{\theta}, t = t_l\right), \tag{8}$$

where $\mathbf{c} = [c_1, c_2, \ldots, c_{N_t}]$ is the number of cells measured at each $l^{\text{th}}$ measurement time. For smFISH experiments, the vector $\mathbf{c}$ plays an important role in the design of the study. By optimizing over all vectors $\mathbf{c}$ that sum to $N_{\text{total}}$, one can find how many cells should be measured at each time point and which time points should be skipped entirely (i.e., $c_l = 0$).

In the next section, we verify the FSP-FIM for this stochastic model with a time-varying parameter and later find the optimal **c** for *STL*1 mRNA in yeast cells.

*3.2. The FSP-FIM Can Quantify Experimental Information for Stochastic Gene Expression under Time-Varying Inputs.* Our work in [6] was limited to models of stochastic gene expression that had piecewise constant reaction rates. Here, we extend this to time-varying reaction rates that affect the promoter switching in the system and which lead to time-varying $\mathbf{A}(t)$ in equation (6). For example, in the model depicted in Figure 1(a), the temporal addition of osmotic shock causes nuclear translocation of HOG1-MAPK, according to the time-varying function in equation (1).

Model parameters simultaneously fit to experimentally measured 0.2 M and 0.4 M *STL*1 mRNA were adopted from [10] and used as a reference set of parameters (yellow dots in Figure 2(a) and S1), which we define as $\boldsymbol{\theta}^*$. These reference parameters were used to generate 50 unique and independent simulated datasets, and each $n^{\text{th}}$ simulated dataset was fit to find the parameter set, $\widehat{\boldsymbol{\theta}}_n$, that maximizes the likelihood for that simulated dataset. This process was repeated for two different experiment designs, including the original intuitive design from [10] (results shown in Figure 2) and an optimized design discussed below (results shown in Figure S1). To ease the computational burden of this fitting, the four parameters with the smallest sensitivities and largest uncertainties (i.e., those parameters that had the least effect on the model predictions and which were most difficult to identify) were fixed at their baseline values. The resulting MLE estimates for the remaining five parameters were collected into a set of $\{\widehat{\boldsymbol{\theta}}_n\}$ and are shown as yellow dots in Figures 2 and S1. Using the asymptotic normality of the maximum likelihood estimator and its relationship to the FIM (equation (5)), we then compared the 95% confidence intervals (CIs) of the inverse of the Fisher information (i.e., the Cramér–Rao bound) to those of the MLE estimates (compare the purple and orange ellipses in Figures 2(a) and S1a). We also compared the eigenvalues of the inverse of the Fisher information, $\{v_i\}$, to the correspondingly ranked eigenvalues of the covariance matrix of MLE estimates, $\Sigma_{\text{MLE}}$, in Figures 2(b) and S1b. For further validation, we noted that the principle directions of the ellipses in Figures 2(a) and S1a also match for the FIM and MLE analyses, as quantified by the angle between the paired FIM and $\Sigma_{\text{MLE}}$ eigenvectors (Figures 2(b) and S1b). For comparison, the angles between rank-matched eigenvectors of the FIM and $\Sigma_{\text{MLE}}$ were all less than 12°, whereas non-rank-matched eigenvectors were all greater than 79.9°. With the FSP-FIM verified for the HOG1-MAPK-induced gene expression model, we next explore how the FSP-FIM can be used to optimally allocate the number of cells to measure at each time after osmotic shock.

*3.3. Designing Optimal Measurements for the HOG1-MAPK Pathway in S. cerevisiae.* To explore the use of the FSP-FIM for experiment design in a realistic context of MAPK-activated gene expression, we again utilize simulated time-course smFISH data for the osmotic stress response in yeast.

We start with a known set of underlying model parameters that were taken from simultaneous fits to 0.2 M and 0.4 M data in [10] (nonspatial model) to establish a baseline parameter set that is experimentally realistic. These parameters are then used to optimize the allocation of measurements at different time points $t = [1, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$ minutes after NaCl induction. Specifically, we ask what fraction of the total number of cells should be measured at each time to maximize the information about a specific subset of important model parameters. We use a specific experiment design objective criteria referred to as $D_s$-optimality, which corresponds to minimizing the expected volume of the parameter space uncertainty for the specific parameters of interest [35] and which is found by maximizing the product of the eigenvalues of the FIM for those same parameters.

Mathematically, our goal is to find the optimal cell measurement allocation:

$$\mathbf{c}_{\text{opt}} = \arg\max_{c} |\mathcal{I}(\mathbf{c};\boldsymbol{\theta})|_{D_s} \text{such that} \sum_{l=1}^{N_t} c_l = 1, \tag{9}$$

where $c_l$ is the fraction of total measurements to be allocated at $t = t_l$, and the metric $|\mathcal{I}(\mathbf{c};\boldsymbol{\theta})|_{D_s}$ refers to the product of the eigenvalues for the total FIM (equation (8)). The fraction of cells to be measured at each time point, **c**, was optimized using a greedy search, in which single-cell measurements were chosen one at a time according to which time point predicted the greatest improvement in the optimization criteria (see Supplementary Note 3 for more information).

To illustrate our approach, we first allocated cell measurements according to $D_s$-optimality as found through this greedy search. Figure 3 shows the optimal fraction of cells to be measured at each time following a 0.2 M NaCl input and compares these fractions to the experimentally measured number of cells from [10]. While each available time point was allocated a nonzero fraction of measurements, three time points at $t = [10, 15, 30]$ minutes were vastly more informative than the other potential time points. To verify this result, we simulated 50 datasets of 1,000 cells each and found the MLE estimates for each subsampled dataset. We compared the spread of these MLE estimates to the inverse of the optimized FIM, shown in Figure S1.

Comparing Figure S1 with Figure 2 illustrates the extent by which the design of optimal measurement times for a 0.2 M NaCl experiment can increase information collection and reduce parameter uncertainties compared to the intuitive measurment design from [10]. In addition to providing much higher Fisher information, the optimal experiment requires measurement of only three time points compared to the 16 time points that were measured in the original experiment. Furthermore, we note that the FIM prediction of the MLE uncertainty is more accurate for the simpler optimal design, which is likely related to our observation that MLE estimates converge more easily for the optimized experiment design than they do for the original intuitive design.

Figure 4 next compares the $D_s$-optimality criteria for the optimal (solid horizontal lines) and intuitive ([10], dashed horizontal lines) experiment designs to 1,000 randomly designed experiments for the 0.2 M (black) and 0.4 M (gray)

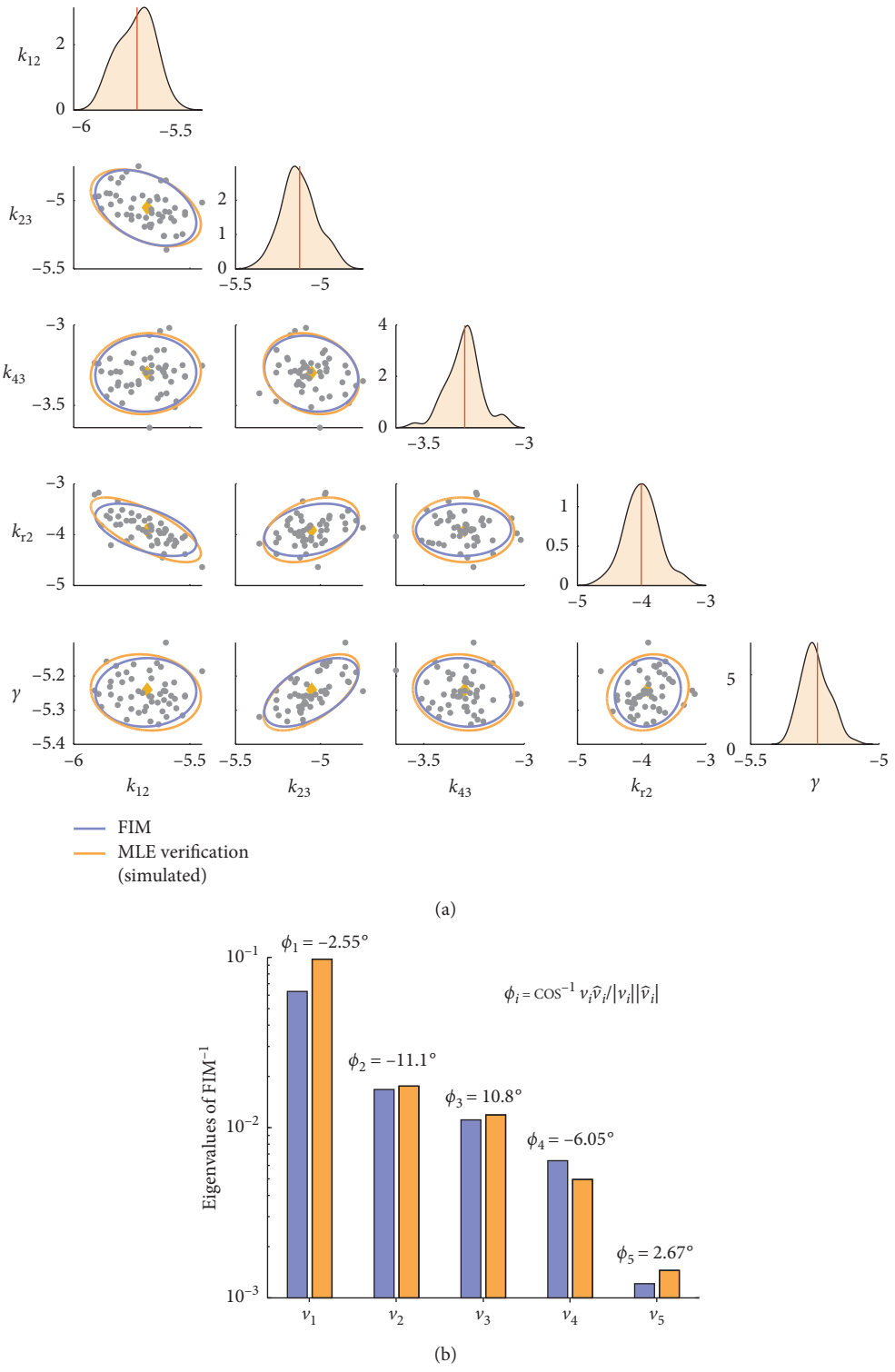


Figure 2: Verification of the FSP-FIM for the time-varying HOG1-MAPK model. (a) Marginal parameter histograms (top panels) and joint scatter plots (gray dots) for the MLE parameter estimates from 50 simulated datasets and for a subset of model parameters. All parameters are shown in logarithmic scale. The ellipses show the 95% CI for the inverse of the FIM (purple) and Gaussian approximation of MLE scatter plot (orange). The yellow dots indicate the "true" parameters at which the FIM and simulated datasets were generated. (b) Rank-paired eigenvalues ($v_i$) for the covariance of MLE estimates (orange) and inverse of the FIM (blue). The angles between corresponding rank-paired eigenvectors ($\phi_i$) are shown in degrees.

conditions. To generate these random experiment designs, we selected a random subset of the measurement times and allocated the total 1,000 cells among chosen time points using a multinomial distribution with equal probability for each time point. Figure 4(a) shows that the intuitive experiment is more informative than most random

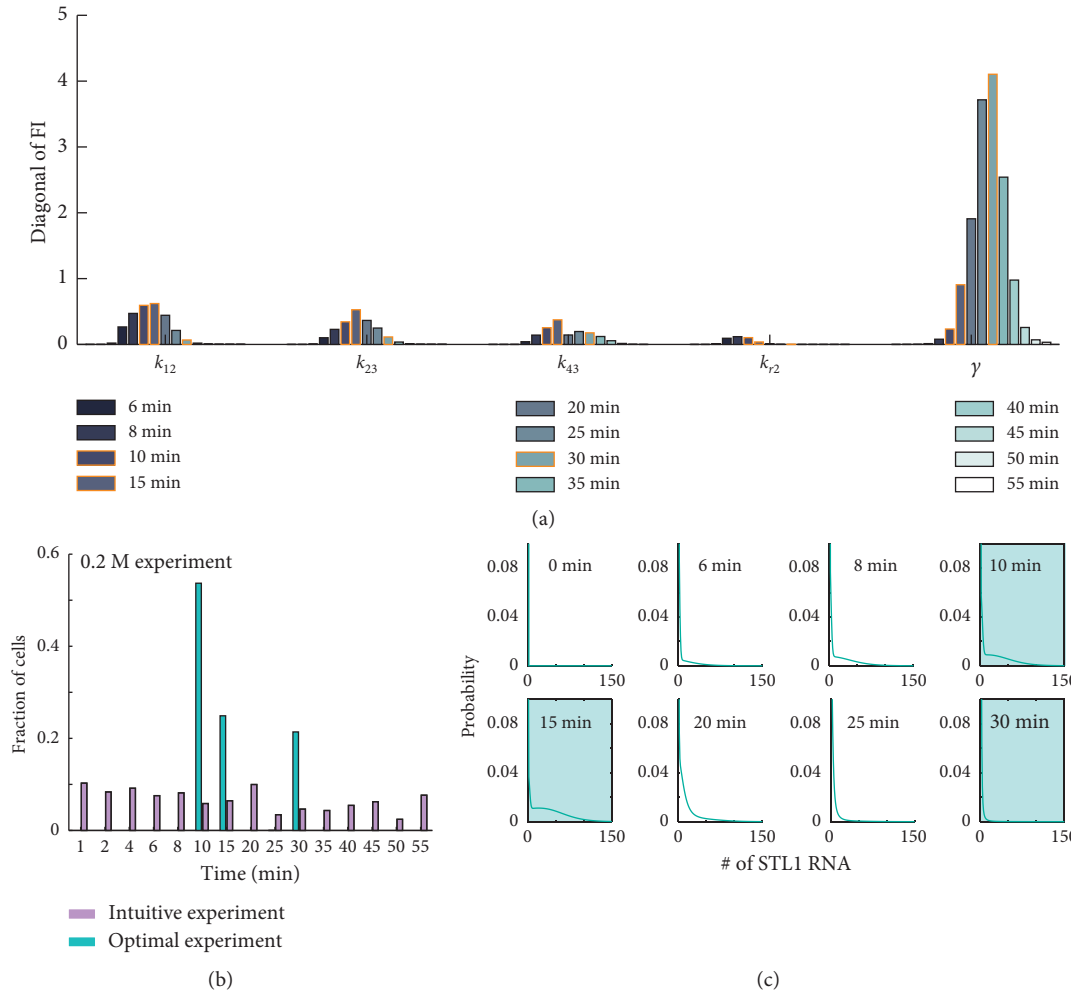


Figure 3: Optimizing the allocation of cell measurements at different time points. (a) Diagonal entries of the Fisher information at different measurement times. The optimal measurement times $t = [10, 15, 30]$ minutes are highlighted in orange. (b) Comparison of optimal fractions of cells to measure (blue) at different time points determined by the FSP-FIM compared to experimentally measured numbers of cells at 0.2 M NaCl (purple) from our work in [10]. (c) Probability distributions of *STL*1 mRNA at several of measurement times. The blue boxes denote the time points of optimal measurements.

experiments but is still substantially less informative than the optimal experiment.

In many practical applications, a scientist would be unlikely to have precise *a priori* knowledge of model parameters prior to conducting experiments. Rather, they would have some estimate of these parameters, such as rough knowledge of appropriate time scales or existing data from another type of experiment. Such estimates could come from previous analyses of the system response to simpler experimental conditions, for measurements taken on slightly different cell lines or organisms, or considering results from different genes in related regulatory pathways. To explore the importance of knowing the exact process parameters or input dynamics prior to designing the experiment, we asked how well an experiment design optimized using parameters from one gene at a given level osmotic shock (e.g., *STL*1 at 0.2 M NaCl) would do to estimate parameters for another gene in a different osmotic shock condition (e.g., *CTT*1 at 0.4 M NaCl). Figure 4(b) demonstrates the impact of such mismatched experiment designs, where each row corresponds to a different intuitive or optimized experiment design (i.e., a specific allocation of cells to be measured at each time), and each column corresponds to a specific gene and specific osmotic shock condition to which that design could be applied. In all cases, the much simpler FIM-based optimal experiment designs perform as well or better than the more difficult intuitive designs, even when these FIM designs were computed assuming different environmental conditions and assuming genes whose parameters differ considerably from one another (see Supplemental Tables 1 and 2 for parameter sets). In other words, these results suggest that if one can compute a simple yet optimal experiment design based on one well-analyzed gene in a previously studied environmental condition, then that design may be equally effective when applied to new investigations for related genes in similar biological contexts.

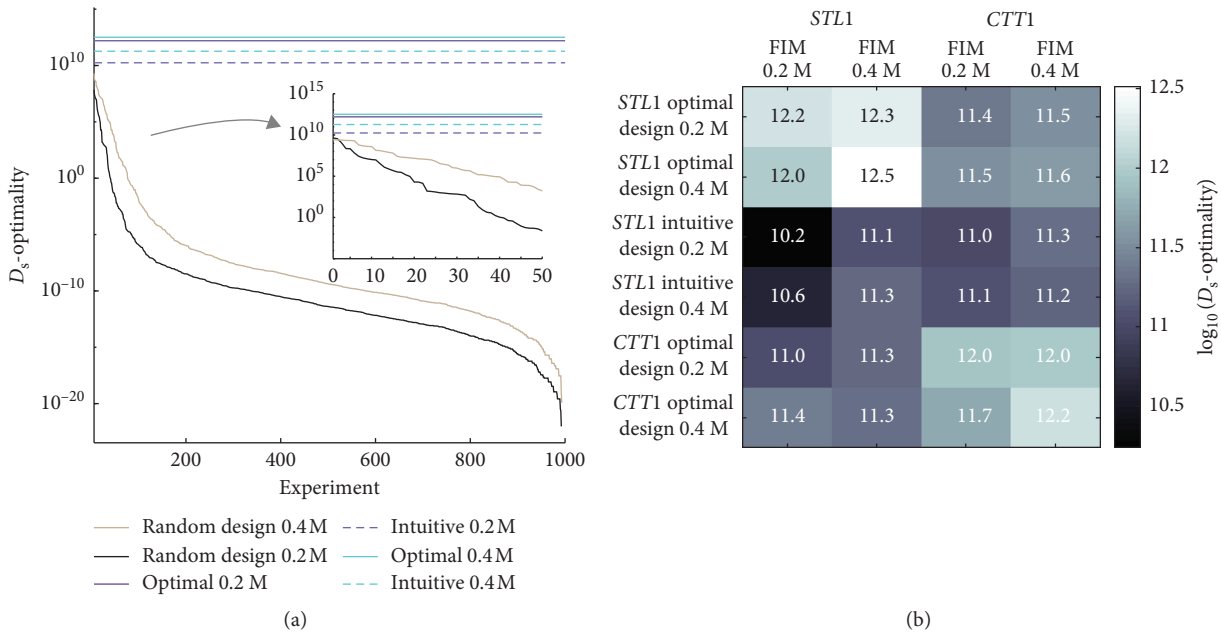


FIGURE 4: Information gained by performing optimal experiments compared to actual experiments. (a) $D_s$-optimality for optimal design using three time points compared to the intuitive experiment designs made using 16 time points is shown with horizontal lines (purple, 0.2 M, and blue, 0.4 M). Solid horizontal lines denote the optimal designs and dashed lines represent intuitive experiment designs. Randomly designed experiments with 0.2 M and 0.4 M NaCl are shown in black and orange. For the random experiments, the time points were selected by sampling them from the experimental measurement times, and then a random number of measurements were assigned to each selected time point. The inset shows the first 50 randomly designed experiments. (b) The $D_s$-metric for different experiment designs (different rows) when applied to different genes or different experimental levels of osmotic shock (different columns). Lighter shades (higher $D_s$-metrics) indicate experimental designs that are more suitable to identify parameters.

*3.4. Using the FSP-FIM to Design Optimal Biosensor Measurements.* Thus far, and throughout our previous work in [6], we have sought to find the optimal set of experiments to reduce uncertainty in the estimates of *model parameters*. In this section, we discuss how the FSP-FIM allows for the optimization of experiment designs to address a more general problem of inferring *environmental variables* from cellular responses. Toward this end, we assume a known and parametrized model (i.e., the model defined above, which was identified previously in [10]), but which is now subject to unknown environmental influences. We explore what would be the optimal experimental measurements to take to characterize these influences. Specifically, we ask how many cells should be measured using smFISH, and at what times, to determine the specific concentration of NaCl to which the cells have been subjected—or, equivalently, we ask what experiments would be best suited to measure the effective stress induction level caused by addition of an unknown solution to the cells.

Recall from above that in the HOG1-MAPK transcription model, extracellular osmolarity ultimately affects stress response gene transcription levels through the time-varying parameter $k_{21}(t)$ (equation (1)) as illustrated in Figure 1(c) for 0.2 M and 0.4 M salt concentrations. Higher salt concentrations delay the time at which $k_{21}(t)$ returns to its nonzero value. The function in equation (1) can be coarsely approximated by the sum of three Heaviside step functions, $u(t-\tau_i)$ as

$$k_{21}(t) = k_{21}^0\left(u(t) - u(t-\tau_1) + u(t-\tau_2)\right), \tag{10}$$

where $\tau_1$ is the fixed delay of the time it takes for nuclear kinase levels to reach the $k_{21}$ deactivation threshold (about 1 minute or less, [9, 10]) and $\tau_2$ is the variable time it takes for the nuclear kinase to drop back below that threshold. In practice, the threshold-crossing time, $\tau_2$, should be directly related to the salt concentration experienced by the cell under reasonable salinity levels. This relationship is shown in Figures 1(b), 1(c), and 5(b), where a 0.2 M NaCl input exhibits a shorter $\tau_2$ than does a 0.4 M input. For our analyses, we assume a prior uncertainty such that time $\tau_2$ can be any value uniformly distributed between $\tau_2^{\min} = 6$ and $\tau_2^{\max} = 31$ minutes, and our goal is to find the experiment that best reduces the posterior uncertainty in $\tau_2$ (and therefore could provide an estimate for the concentration of NaCl).

To reformulate the FSP-FIM to estimate uncertainty in $\tau_2$ given our model, the first step is to compute the sensitivity of the distribution of mRNA abundance to changes in the variable $\tau_2$ using equation (5), in which $\mathbf{A}_{\theta_j}(t)$ is replaced with $\mathbf{A}_{\tau_2}(t) = \partial\mathbf{A}/\partial\tau_2$ as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\tau_2}\end{bmatrix} = \begin{bmatrix}\mathbf{A}(t) & 0\\ \mathbf{A}_{\tau_2}(t) & \mathbf{A}(t)\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\tau_2}\end{bmatrix}. \tag{11}$$

As $k_{21}(t)$ is the only parameter in $\mathbf{A}$ that depends explicitly on $\tau_2$, all entries of $\partial\mathbf{A}/\partial\tau_2$ are zero except for those which depend on $k_{21}(t)$, and

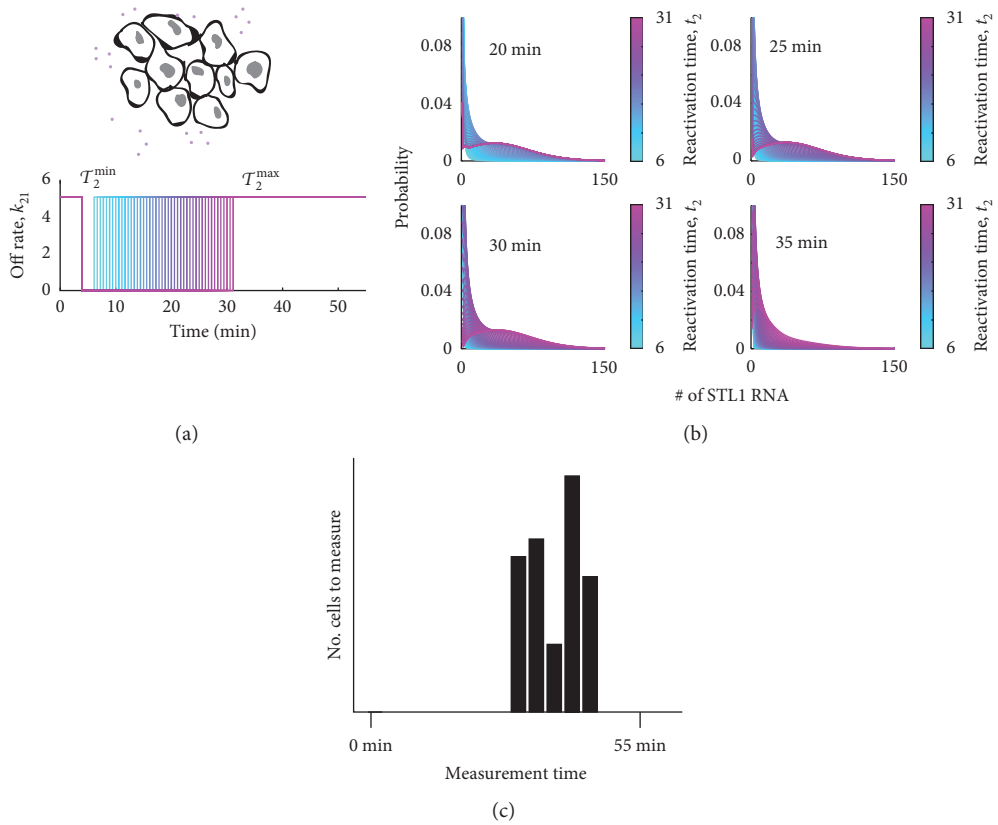


FIGURE 5: Overview of optimal design for biosensing experiments for the osmotic stress response in yeast. (a) Unknown salt concentrations (purple dots) in the environment give rise to different reactivation times, $\tau_2$, which affect the gene expression in the model through the rate $k_{21}$. These different reactivation times cause downstream *STL*1 expression dynamics to behave differently as shown in (b). (c) Different responses can be used to resolve experiments that reduce the uncertainty in $\tau_2$.

$$\mathbf{A}_{\tau_2}(t) = \frac{\partial \mathbf{A}}{\partial k_{21}} \frac{\partial k_{21}}{\partial \tau_2} = \mathbf{A}_{k_{21}} k_{21}^0 \delta(\tau_2), \tag{12}$$

and therefore $\mathbf{A}_{\tau_2} = \partial \mathbf{A}/\partial \tau_2$ is nonzero only at $t = \tau_2$. Using this fact, the equation for the sensitivity dynamics is uncoupled from the FSP dynamics for $t \neq \tau_2$ and can be written simply as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{s}_{\tau_2} = \begin{cases} 0 \text{ for } t < \tau_2 \text{ with } \mathbf{s}(0) = 0, \\ \mathbf{A}(t)\mathbf{s}_{\tau_2} \text{ for } t > \tau_2 \text{ with } \mathbf{s}_{\tau_2}(\tau_2) = k_{21}^0 \mathbf{A}_{k_{21}} \mathbf{p}(\tau_2). \end{cases} \tag{13}$$

If the Fisher information at each measurement time is written into a vector $\mathbf{f} = [f_1, f_2, \ldots, f_{N_t}]$ (noting that the Fisher information at any time $t_l$ is the scalar quantity, $f_l$) and the number of measurements per time point is the vector, $\mathbf{c} = [c_1, c_2, \ldots, c_{N_t}]$, then the total information for a given value of $\tau_2$ can be computed as the dot product of these two vectors:

$$\mathcal{I}(\tau_2) = \sum_{l=1}^{N_t} c_l f_l = \mathbf{c}^T \mathbf{f}. \tag{14}$$

Our goal is to find an experiment that is optimal to determine the value of $\tau_2$, given an assumed prior that $\tau_2$ is sampled from a uniform distribution between $\tau_2^{\min}$ and $\tau_2^{\max}$. To find the experiment $\mathbf{c}_{\text{opt}}$ that will reduce our posterior uncertainty in $\tau_2$, we integrate the inverse of the FIM in equation (14) over the prior uncertainty in $\tau_2$:

$$\begin{aligned} \mathbf{c}_{\text{opt}} &= \underset{\mathbf{c}, \sum c_l = 1}{\arg\min} \int_{\tau_2^{\min}}^{\tau_2^{\max}} \frac{1}{\tau_2^{\max} - \tau_2^{\min}} \mathcal{I}^{-1}(\mathbf{c}; \tau_2 = \tau, \boldsymbol{\theta}) \mathrm{d}\tau, \\ &= \underset{\mathbf{c}, \sum c_l = 1}{\arg\min} \int_{\tau_2^{\min}}^{\tau_2^{\max}} \mathcal{I}^{-1}(\mathbf{c}; \tau_2 = \tau, \boldsymbol{\theta}) \mathrm{d}\tau. \end{aligned} \tag{15}$$

For later convenience, we define the integral in equation (15) (i.e., the objective function of the minimization) by the symbol $\mathcal{J}$, which corresponds to the expected uncertainty about the value of $\tau_2$ for a given $\mathbf{c}$.

Next, we apply the greedy search from above to solve the minimization problem in equation (15) to find the experiment design $\mathbf{c}_{\text{opt}}$ that minimizes the estimation error of $\tau_2$. Figure 6 shows examples of seven different experiments to accomplish this task, ranked according to the FSP-FIM value $\mathcal{J}$ from most informative (top left) to least informative (bottom left), but all using the same number of measured cells. For each experiment, the FSP-FIM was used to estimate the posterior uncertainty (i.e., expected standard deviation) in the estimation of $\tau_2$, which is shown by the orange bars in

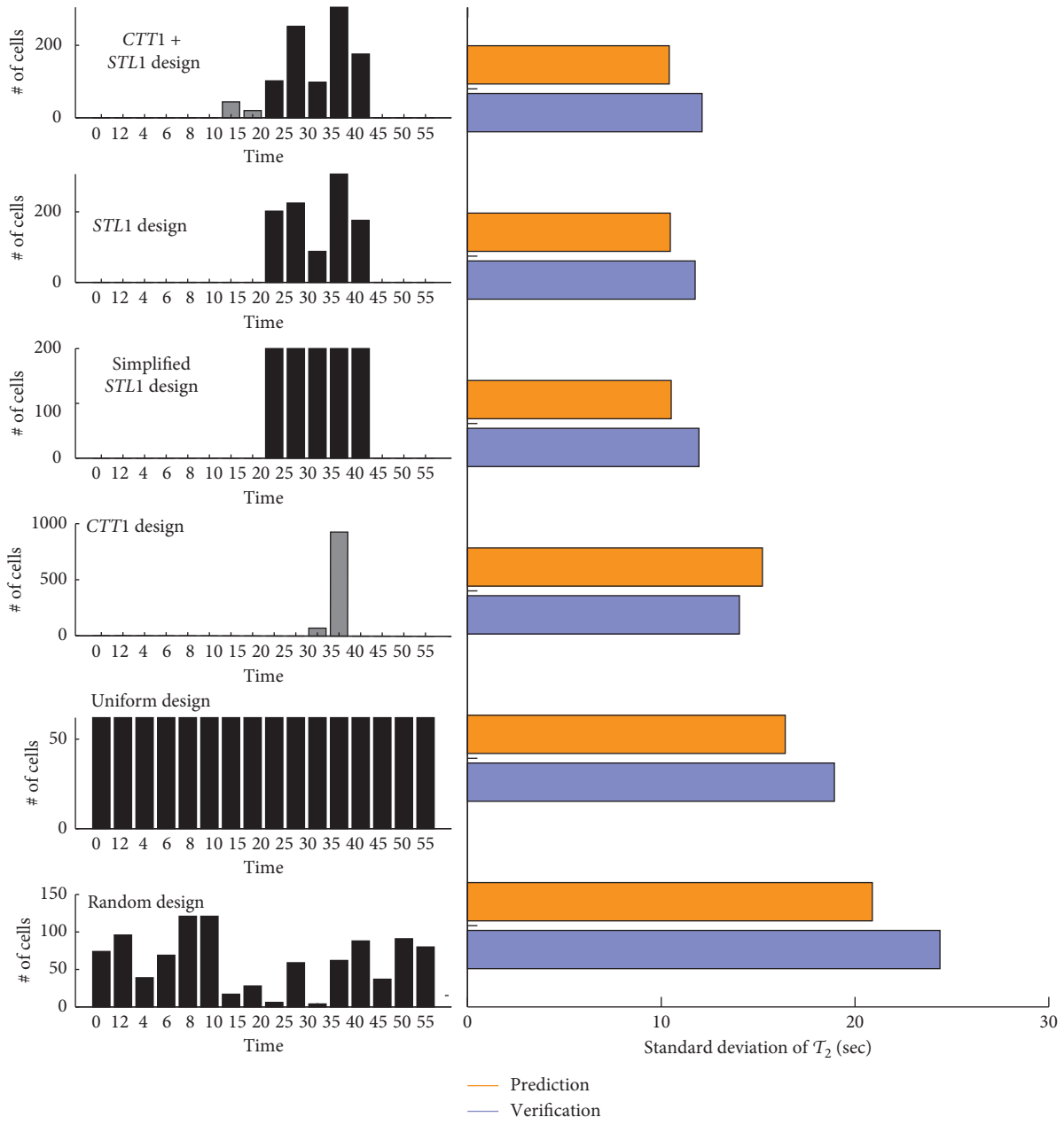


FIGURE 6: Verification of the uncertainty in $\tau_2$ for different experiment designs. The left panel shows various experiment designs, where the sum of the bars (i.e., the total number of measurements) is 1,000. Gray bars represent the measurements of *CTT*1 and black bars *STL*1. The right panel shows the value of the objective function in equation (15) for each experiment design in orange, and the RMSE values for verification are shown in purple.

Figure 6. To verify these estimates, we then chose 64 uniformly spaced values of $\tau_2$, which we denote as the set $\{\tau_2^{\text{true}}\}$, and for each $\tau_2^{\text{true}}$, we simulated 50 random datasets of 1,000 cells distributed according to the specified experiment designs. For each of the $64 \times 50$ simulated datasets, we then determined the value $\tau_2^{\text{MLE}}$ between $\tau_2^{\text{min}}$ and $\tau_2^{\text{max}}$ that maximized the likelihood of the simulated data according to equation (2). The root mean squared estimate (RMSE) error over all random values of $\tau_2^{\text{true}}$ and estimates, $\sqrt{\langle(\tau_2^{\text{MLE}} - \tau_2^{\text{true}})^2\rangle}$, was then computed for each of the six different experiment designs. Figure 6 shows that the FIM-based estimation of uncertainty and the actual MLE-based uncertainty are in excellent agreement for all experiments (compare purple and orange bars). Moreover, it is clear that the optimal design selected by the FIM analysis performed much better to estimate $\tau_2$ than did the uniform or random experimental designs. A slightly simplified design, which uses the same time points as the optimal, but with equal numbers of measurements at each time, performed nearly as well as the optimal design.

The set of experiment designs shown in Figure 6 includes the best design that only uses *STL*1 (second from top), the best design that uses only *CTT*1 (fourth from top), and the best design that uses some cells with *CTT*1 and some with

*STL*1 (top design). To find the best experiment design for measurement of two different genes, we assumed that at each time, either *STL*1 mRNA *or CTT*1 mRNA (but not both) could be measured, corresponding to using smFISH oligonucleotides for either *STL*1 or *CTT*1. To determine which gene should be measured at each time, we compute the Fisher information for *CTT*1 and *STL*1 for every measurement time and averaged this value over the range of $\tau_2$. For each measurement time $t_l$, the gene is selected that has the higher average Fisher information for $\tau_2$. The number of cells per measurement time was then optimized as before, except the choice to measure *CTT*1 or *STL*1 was based on which mRNA had the larger Fisher information (equation (14)) at that specific point in time. The best *STL*1-only experiment design was found to yield uncertainty of 10.5 seconds (standard deviation); the best *CTT*1-only experiment was found to yield an uncertainty of 15.2 seconds and the best mixed *STL*1/*CTT*1 experiment design was found to yield an uncertainty of 10.4 seconds. In other words, for this case, the *STL*1 gene was found to be much more informative of the environmental condition than was *CTT*1, and the use of both *STL*1 and *CTT*1 provides only minimal improvement beyond the use of *STL*1 alone. We note that although measurement times in the optimized experiment design were restricted to a resolution of five minutes or more, the value of $\tau_2$ could be estimated with an error of only 10 seconds, corresponding to a roughly 30-fold improvement of temporal resolution beyond the allowable sampling rate.

*3.5. Experimental Validation for FSP-FIM-Based Designs of Biosensor Measurements.* To experimentally validate our FSP-FIM-based approach to design optimal measurement times, we next examined experimental smFISH data taken for the *STL*1 and *CTT*1 genes at different times following yeast osmotic shock [10]. These data include a total of 535–4808 cells measured at each of 16 time points following osmotic shocks of 0.2 M or 0.4 M NaCl. We asked how well could we identify the concentration of the osmotic shock from the experimental data using only 75 individual cells per experiment. We again proposed the six different potential experiments depicted in Figure 6, including the optimal *STL*1 and *CTT*1 design, the optimal *STL*1 design, the simplified *STL*1 design with 15 cells for each of the optimal five time points, the optimal *CTT*1 design, the uniform *STL*1 design, and the random STL1 design. For each design, we created 1,000 different experimental replica datasets, each consisting of 100 cells randomly chosen from the original data. For each replica dataset, we then used the CME model (Supplementary Note 1) with a parametrized form of the HOG1-MAPK nuclear localization signal (Supplementary Note 2) to find the NaCl concentration that maximizes the likelihood of the data given the model.

Figure 7 shows the resulting histograms for the estimated NaCl concentrations for each of the six experiment designs, when the cells were actually subjected to experimental osmotic shocks of 0.2 M NaCl (Figure 7(a)) or 0.4 M NaCl (Figure 7(c)). From Figures 7(a) and 7(c), it is clear that the FSP analysis provides an accurate estimate for the level of the osmotic shock input using a relatively small number of cells, despite the fact that producing such estimates was not an intended use of the model in its original formulation or parameter inference [9, 10]. Figures 7(b) and 7(d) show the uncertainty (standard deviation) in the experimental estimate of NaCl concentration (light bars), when cells are collected according to the six specific experiment designs, and compare these results to the FSP-FIM uncertainty estimates (dark bars) using the simplified step input function (equation (10)). With the exception of the suboptimal *CTT*1-only design, the close matches between the relative trends of the variance in experimental estimation of NaCl and the variance predicted by the FSP-FIM analysis with the approximated step-function input give further experimental validation that the FSP-FIM approach can be used to choose more informative experiment designs, even in cases where the FSP analyses use inexact assumptions for model kinetics. The single discrepancy in trends led us to more closely examine the model and experimental data for *CTT*1 expression at the 35-minute time point that dominates the *CTT*1-only design. By examining Supplemental Figure S7 from [10], we found that this specific combination of *CTT*1 at 35 minutes following 0.4 M NaCl osmotic shock showed a greater discrepancy between model and data than any of the other 63 combinations of 16 times, two genes, and two conditions, yet it is unclear if that difference was an artifact of the experiment or an actual transient effect that only affected that specific combination of gene, time, and environmental condition.

## 4. Discussion

The methods developed in this work present a principled, model-driven approach to allocate how many snapshot single-cell measurements should be taken at each time during analysis of a time-varying stochastic gene regulation system. We demonstrate and verify these theories on a well-established model of osmotic stress response in yeast cells, which is activated upon the nuclear localization of phosphorylated HOG1 [9, 10]. For this system, we showed how to optimally allocate the number of cells measured at each time so as to maximize the information about a subset of model parameters. We found that the optimal experiment design to estimate model parameters for the *STL*1 gene only required three time points. Moreover, these three time points ($t = [10, 15, 30]$ minutes, highlighted by blue in Figure 3(b)) are at biologically meaningful time points. At $t = 10$ and 15 minutes, the system is increasing to maximal expression, and the probability to measure a cell with elevated mRNA content is high, which helps reduce uncertainty about the parameters in the model that control maximal expression. Similarly, at the final experiment time of $t = 30$ minutes, the system is starting to shut down gene expression, and therefore this time is valuable to learn about the time scale of deactivation in the system as well as the mRNA degradation rate. These effects are clearly illustrated in Figure 3(a), which shows that times $t = 10$ and $t = 15$ minutes provide the most information about parameters $k_{12}$, $k_{23}$, and $k_{43}$, whereas measurements at $t = 30$ minutes provide the most

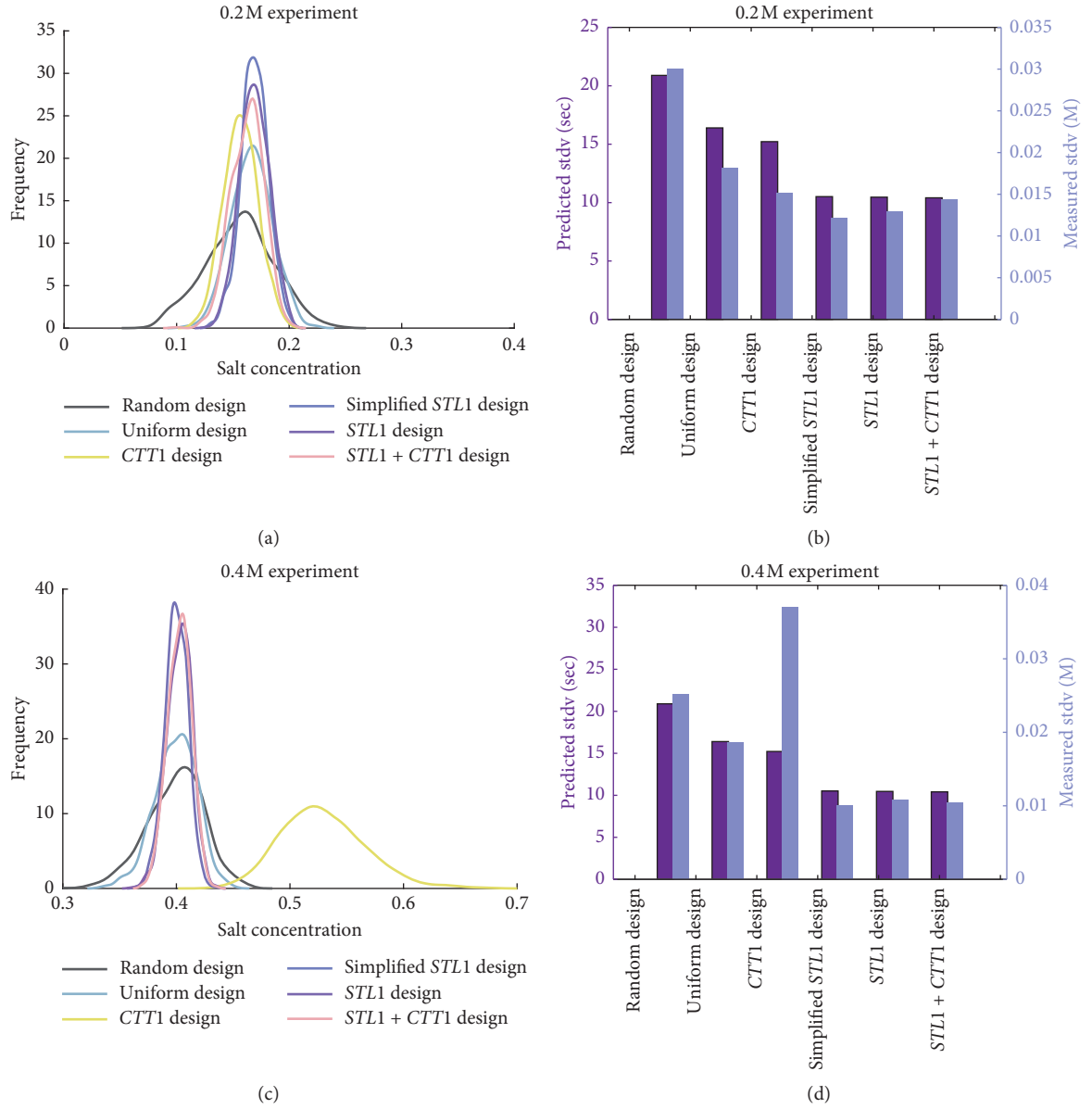


Figure 7: Experimental validation of FSP-FIM-based design for optimal biosensor measurements. (a) Distribution of FSP-based MLE estimates for NaCl concentration using the six experimental designs from Figure 6. Each distribution comes from 1,000 replicas of 75 cells per replica spread out over the possible 16 time points. Replica data were sampled randomly from published experimental data [10] that contain two or three biological replicas and 535–4808 cells per time point. The true experimentally applied level of osmotic shock was 0.2 M NaCl. (b) The MLE estimation standard deviation for each experiment design applied to a dataset taken at 0.2 M NaCl (blue). These deviations are compared to FSP-FIM deviation predictions using a piecewise constant model for HOG1 nuclear localization (purple). (c, d) Same as (a, b) but for a true NaCl concentration of 0.4 M.

information about $\gamma$. Because $\gamma$ is the easiest parameter to estimate (e.g., its information is greater), not as many cells are needed at $t = 30$ minutes to constrain that parameter. Similarly, because $k_{r2}$ is the most difficult parameter to estimate (e.g., it has the lowest information across all experiments) and because $t = 10$ minutes is one of the few time points to provide information about $k_{r2}$, the optimal experimental design selects a large number of cells at the time $t = 10$ minutes. This analysis demonstrates that the optimal experiment design can change depending upon which parameters are most important to determine (e.g., $\gamma$ or $k_{r2}$ in this case), a fact that we expect will be important to consider in future experiment designs.

Because we constrained all potential experiment designs to be within the subset of experiments performed in our previous work [10], we are able to compare the information of optimal experiment designs to intuitive designs that have actually been performed. We found that while the intuitive experiments were almost always better than could be expected by random chance, they still provided several orders of magnitude lower Fisher information than would be possible with optimal experiments (Figure 4(a)). Moreover,

in our analyses, we found that optimal designs could require far fewer time points than those designed by intuition (e.g., only three time points were needed in Figure 3), and therefore these designs can be much easier and less expensive to conduct. We also found that utility of optimal experiment designs could be relatively insensitive to variation in the experimental conditions or the specific model parameters used for the experiment design. For example, we found that experiments optimized for one gene at one level of osmotic shock were still at least as good—and in most cases better—than intuitive designs, even when conducted using different genes and at a different level of osmotic shock (Figure 4(b)). In practice, this fact would allow for effective experiment designs despite inaccurate prior assumptions.

In addition to suggesting optimal experiments to identify model parameters, we showed that the FSP approach could be used to infer parameters of fluctuating extracellular environments from single-cell data and that the FSP-FIM combined with an existing model could be used to design optimal experiments to improve this inference (Figures 5 and 6). We experimentally verified this potential by examining many small sets of single-cell smFISH measurements for different genes and different measurement times, and we showed that an FSP-FIM analysis could correctly rank which experiment designs would give the best estimates of osmotic shock environmental conditions. Along a very similar line of reasoning, one can also adapt the FSP-FIM analysis to learn what biological design parameters would be optimal to reduce uncertainty in the estimate of important environmental variables. For example, Figure 8 shows the expected uncertainty in $\tau_2$ as a function of the degradation rate of the *STL*1 gene assuming that 50 cells could be measured at each experimental measurement time $t = [1, 2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]$ minutes using the smFISH approach. We found that the best choice for *STL*1 degradation rate to most accurately determine the extracellular fluctuations would be $2.4 \times 10^{-3}$ mRNA/min, which is about half of the experimentally determined value of $5.3 \times 10^{-3} \pm 5.9 \times 10^{-5}$ from [10]. This result is consistent with our earlier finding that the faster degrading *STL*1 mRNA is a much better determinant of the HOG1 dynamics than the slower-degrading *CTT*1 mRNA and suggests that other less stable mRNA could be more effective still. We expect that similar, future applications of the FSP-based Fisher information will be valuable in other systems and synthetic biology contexts where scientists seek to explore how different cellular properties affect the transmission of information between cells or from cells to human observers. Indeed, similar ideas have been explored recently using classical information theory in [36–39], and recent work in [7, 40] has noted the close relationship between Fisher information and the channel capacity of biochemical signaling networks.

We expect that computing optimal experiment designs for time-varying stochastic gene expression will create opportunities that could extend well beyond the examples presented in this work. Modern experimental systems are making it much easier for scientists and engineers to precisely perturb cellular environments using chemical induction [41–43] or optogenetic control [44–46]. Many such experiments involve stochastic bursting behaviors at the mRNA or protein level [8–10, 45], and precise optimal experiment design will be crucial to understand the properties of stochastic variations in such systems. A related field that is also likely to benefit from such approaches is biomolecular image processing and feedback control, for which one may need to decide in real time which measurements to make and in what conditions.

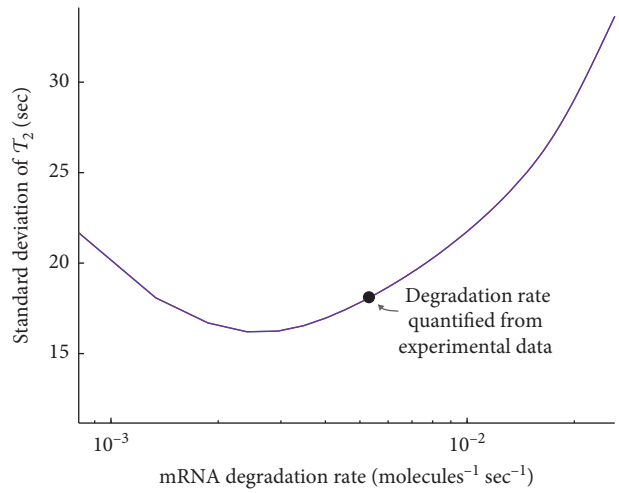


Figure 8: Optimal mRNA degradation rates to reduce uncertainty about the extracellular environment. Uncertainty in the time at which the *STL*1 gene turns off, $\tau_2$, as a function of mRNA degradation rate (purple). The black dot corresponds to the degradation rate that was quantified from experimental data.

## Data Availability

All data and codes associated with this article are available at https://github.com/MunskyGroup/Fox_Complexity_2020.

## Disclosure

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

ZRF and BM were supported by National Institutes of Health (R35 GM124747). ZRF was also supported by the Agence Nationale de la Recherche (ANR-18-CE91-0002, CyberCircuits). GN was supported by the National Institutes of Health (DP2 GM11484901 and R01GM115892) and Vanderbilt Startup Funds. The presented analyses used the computational resources of the WM Keck High Performance Computing Cluster supported under a WM Keck Foundation Award.

## Supplementary Materials

Supplementary note 1: stochastic model of yeast stress response. Supplementary note 2: nuclear localization of HOG-MAPK. Supplementary note 3: optimization of cell measurements. Table I: HOG-MAPK model parameters. Table II: HOG-signaling model parameters. Figure 1: verification of the FSP-FIM for the time-varying HOG-MAPK model. (*Supplementary Materials*)

## References


[1] J. Liepe, S. Filippi, M. Komorowski, and M. P. H. Stumpf, "Maximizing the Information Content of Experiments in Systems Biology," *PLoS Computational Biology*, vol. 9, no. 1, Article ID e1002888, 2013.

[2] J. F. Apgar, D. K. Witmer, F. M. White, and B. Tidor, "Sloppy models, parameter uncertainty, and the role of experimental design," *Molecular BioSystems*, vol. 6, no. 10, p. 1890, 2010.

[3] J. Ruess, A. Milias-Argeitis, and J. Lygeros, "Designing experiments to understand the variability in biochemical reaction networks," *Journal of The Royal Society Interface*, vol. 10, no. 88, Article ID 20130588, 2013.

[4] M. Komorowski, M. J. Costa, D. A. Rand, and M. P. H. Stumpf, "Sensitivity, robustness, and identifiability in stochastic chemical kinetics models," *Proceedings of the National Academy of Sciences*, vol. 108, no. 21, pp. 8645–8650, 2011.

[5] C. Zimmer, "Experimental design for stochastic models of nonlinear signaling pathways using an interval-wise linear noise approximation and state estimation," *PLoS One*, vol. 11, no. 9, Article ID e0159902, 2016.

[6] Z. R. Fox and B. Munsky, "The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments," *PLoS Computational Biology*, vol. 15, no. 1, Article ID e1006365, 2019.

[7] V. Singh and I. Nemenman, "Universal properties of concentration sensing in large ligand-receptor networks," *Physical Review Letters*, vol. 124, no. 2, Article ID 028101, 2020.

[8] C. Zechner, J. Ruess, P. Krenn et al., "Moment-based inference predicts bimodality in transient gene expression," *Proceedings of the National Academy of Sciences*, vol. 109, no. 21, pp. 8340–8345, 2012.

[9] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, "Systematic identification of signal-activated stochastic gene regulation," *Science*, vol. 339, no. 6119, pp. 584–587, 2013.

[10] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, "Distribution shapes govern the discovery of predictive models for gene regulation," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. 7533–7538, 2018.

[11] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, "Imaging individual mRNA molecules using multiple singly labeled probes," *Nature Methods*, vol. 5, no. 10, pp. 877–879, 2008.

[12] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, "Visualization of single RNA transcripts in situ," *Science*, vol. 280, no. 5363, pp. 585–590, 1998.

[13] N. Tsanov, A. Samacoits, R. Chouaib et al., "smiFISH and FISH-quant—a flexible single RNA detection approach with super-resolution capability," *Nucleic Acids Research*, vol. 44, no. 22, p. e165, 2016.

[14] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl, "Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings," *Nature Methods*, vol. 11, no. 2, pp. 197–202, 2014.

[15] R. M. Kumar, P. Cahan, A. K. Shalek et al., "Deconstructing transcriptional heterogeneity in pluripotent stem cells," *Nature*, vol. 516, no. 7529, pp. 56–61, 2014.

[16] L. S. Weinberger, J. C. Burnett, J. E. Toettcher, A. P. Arkin, and D. V. Schaffer, "Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 tat fluctuations drive phenotypic diversity," *Cell*, vol. 122, no. 2, pp. 169–182, 2005.

[17] B. Munsky, G. Neuert, and A. van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science*, vol. 336, no. 6078, pp. 183–187, 2012.

[18] H. Sharifian, F. Lampert, K. Stojanovski et al., "Parallel feedback loops control the basal activity of the HOG MAPK signaling cascade," *Integrative Biology*, vol. 7, no. 4, pp. 412–422, 2015.

[19] E. Klipp, B. Nordlander, R. Krüger, P. Gennemark, and S. Hohmann, "Integrative model of the response of yeast to osmotic shock," *Nature Biotechnology*, vol. 23, no. 8, pp. 975–982, 2005.

[20] B. Schoeberl, C. Eichler-Jonsson, E. D. Gilles, and G. Müller, "Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors," *Nature Biotechnology*, vol. 20, no. 4, pp. 370–375, 2002.

[21] D. Muzzey, C. A. Gómez-Uribe, J. T. Mettetal et al., "A systems-level analysis of perfect adaptation in yeast osmoregulation," *Journal of End-to-End-Testing*, vol. 138, no. 1, pp. 160–171, 2009.

[22] H. Saito and F. Posas, "Response to hyperosmotic stress," *Genetics*, vol. 192, no. 2, pp. 289–318, 2012.

[23] S. Pelet, F. Rudolf, M. Nadal-Ribelles, E. de Nadal, F. Posas, and M. Peter, "Transient activation of the HOG MAPK pathway regulates bimodal gene expression," *Science*, vol. 332, no. 6030, pp. 732–735, 2011.

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

[25] N. G. Van Kampen and N. Godfried, *Stochastic Processes in Physics and Chemistry*, Elsevier, Amsterdam, Netherlands, 1992.

[26] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[27] A. Singh and J. P. Hespanha, "Approximate moment dynamics for chemically reacting systems," *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 414–418, 2011.

[28] Z. Fox, G. Neuert, and B. Munsky, "Finite state projection based bounds to compare chemical master equation models using single-cell data," *The Journal of Chemical Physics*, vol. 145, no. 7, Article ID 074101, 2016.

[29] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *The Journal of Chemical Physics*, vol. 124, no. 4, Article ID 044104, 2006.

[30] M. Gomez-Schiavon, L.-F. Chen, A. E. West, and N. E. Buchler, "BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells," *Genome Biology*, vol. 18, no. 1, p. 164, 2017.

[31] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1993.
[32] G. Casella and R. L. Berger, *Statistical Inference*, Wadsworth and Brooks/Cole, Pacific Grove, CA, USA, 1990.
[33] C. Kreutz and J. Timmer, "Systems biology: experimental design," *FEBS Journal*, vol. 276, no. 4, pp. 923–942, 2009.
[34] B. Steiert, A. Raue, J. Timmer, and C. Kreutz, "Experimental design for parameter estimation of gene regulatory networks," *PLoS One*, vol. 7, no. 7, Article ID e40052, 2012.
[35] J. Ruess, F. Parise, A. Milias-Argeitis, M. Khammash, and J. Lygeros, "Iterative experiment design guides the characterization of a light-inducible gene expression circuit," *Proceedings of the National Academy of Sciences*, vol. 112, no. 26, pp. 8148–8153, 2015.
[36] R. Cheong, A. Rhee, C. J. Wang, I. Nemenman, and A. Levchenko, "Information transduction capacity of noisy biochemical signaling networks," *Science*, vol. 334, no. 6054, pp. 354–358, 2011.
[37] R. Suderman, J. A. Bachman, A. Smith, P. K. Sorger, and E. J. Deeds, "Fundamental trade-offs between information flow in single cells and cellular populations," *Proceedings of the National Academy of Sciences*, vol. 114, no. 22, pp. 5755–5760, 2017.
[38] J. Selimkhanov, B. Taylor, J. Yao et al., "Accurate information transmission through dynamic biochemical signaling networks," *Science*, vol. 346, no. 6215, pp. 1370–1373, 2014.
[39] G. Tkačik and A. M. Walczak, "Information transmission in genetic regulatory networks: a review," *Journal of Physics. Condensed Matter: An Institute of Physics Journal*, vol. 23, no. 15, Article ID 153102, 2011.
[40] T. Jetka, K. Nienałtowski, S. Filippi, M. P. H. Stumpf, and M. Komorowski, "An information-theoretic framework for deciphering pleiotropic and noisy biochemical signaling," *Nature Communications*, vol. 9, no. 1, p. 4591, 2018.
[41] A. H. Ng, T. H. Nguyen, M. Gómez-Schiavon et al., "Modular and tunable biological feedback control using a de novo protein switch," *Nature*, vol. 572, no. 7768, pp. 265–269, 2019.
[42] A. Thiemicke, H. Jashnsaz, G. Li, and G. Neuert, "Generating kinetic environments to study dynamic cellular processes in single cells," *Scientific Reports*, vol. 9, no. 1, Article ID 10129, 2019.
[43] J.-B. Lugagne, S. S. Carrillo, M. Kirch, A. Köhler, G. Batt, and P. Hersen, "Balancing a genetic toggle switch by real-time feedback control and periodic forcing," *Nature Communications*, vol. 8, no. 1, p. 1671, 2017.
[44] R. Chait, J. Ruess, T. Bergmiller, G. Tkačik, and C. C. Guet, "Shaping bacterial population behavior through computer-interfaced control of individual cells," *Nature Communications*, vol. 8, p. 2557, 2017.
[45] M. Rullan, D. Benzinger, G. W. Schmidt, A. Milias-Argeitis, and M. Khammash, "An optogenetic platform for real-time, single-cell interrogation of stochastic transcriptional regulation," *Molecular Cell*, vol. 70, no. 4, pp. 745–756, 2018.
[46] S. M. Castillo-Hair, E. A. Baerman, M. Fujita, O. A. Igoshin, and J. J. Tabor, "Optogenetic control of Bacillus subtilis gene expression," *Nature Communications*, vol. 10, no. 1, p. 3099, 2019.