Optimal Design of Single-Cell Experiments within Temporally Fluctuating Environments

Zachary R Fox

Inria Saclay Ile-de-France, Palaiseau 91120, France

Institut Pasteur, USR 3756 IP CNRS Paris, 75015 , France

School of Biomedical Engineering, Colorado State University Fort Collins, CO 80523, USA and

zachrfox@gmail.com

Gregor Neuert

Department of Molecular Physiology and Biophysics, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA

Department of Biomedical Engineering, School of Engineering, Vanderbilt University, Nashville, TN 37232, USA

Department of Pharmacology, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA and

gregor.neuert@vanderbilt.edu

Brian Munsky

Department of Chemical and Biological Engineering, Colorado State University Fort Collins, CO 80523, USA

School of Biomedical Engineering, Colorado State University Fort Collins, CO 80523, USA and

brian.munsky@colostate.edu

(Dated: October 20, 2019)

Abstract

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. cerevisae* under time-varying MAPK induction. We validate 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 demonstrate how the FSP-FIM approach can be extended to explore how different measurement times or genetic modifications can help to minimize uncertainty in the sensing of extracellular environments, such as external salinity modulations. This work demonstrates the potential of quantitative models to not only make sense of modern biological data sets, but to close the loop between quantitative modeling and experimental data collection.

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 non-linear 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 uses an assumed (and potentially incorrect) mathematical model of the system to estimate and optimize the value of potential experimental settings. 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 [3–6]. 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, time-varying High Osmolarity Glycerol (HOG) Mitogen Activated Protein Kinase (MAPK) induction of transcription during osmotic stress response in yeast [7–9]. 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-mRNA Fluorescence *in situ* Hybridization (smFISH), where specific fluorescent oligonucleotide probes are hybridized to mRNA of interest in fixed cells [10, 11]. Cells are then imaged and the mRNA abundance in each cell can be counted, either by hand or using automated software such as [12]. 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) is 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 HOG-MAPK pathway in yeast is a model system to study dynamics of signal transduction induced gene regulation in single cells [13–18] and stochastic models of HOG-MAPK activated transcription have been used to predict adaptive transcription responses across yeast cell populations [8, 9, 19]. In particular, previous studies have measured two stress response genes, *STL1* and *CTT1*, and used them to infer the model depicted in Fig. 1a. This calibration and uncertainty quantification process required intense experimental effort to fix and image tens of thousands of cells at more than a dozen time points and for multiple biological replicas as well as intense computational effort for both the processing of the smFISH images and the fitting of stochastic kinetic models to the quantified experimental data. In light of such expenses, we aim to develop methods that can specify experiments that are equally or more informative, yet which could minimize experimental and computational efforts.

Toward this goal, the first part of our current study demonstrates the use of FSP based Fisher information to optimize experiments to minimize the uncertainty in stochastic model parameters for the time varying MAPK-induced gene expression response. In the second part of this study, we expand upon this result to find 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.

BACKGROUND

Finite State Projection models can predict osmotic stress responses in yeast.

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 vari-

ability in gene expression levels. Adequate description of such variable responses can only be achieved through the use of stochastic computational models [20–23].

In this work, we use the chemical master equation framework [24] of stochastic chemical kinetics, which has been the workhorse of stochastic modeling of gene expression, whether through simulated sample paths of its solution via the stochastic simulation algorithm [25], moment approximations [7, 26], or finite state projections (FSP) [27]. 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-based models [6, 9], which can be done with guaranteed errors using the FSP approach [27, 28]. This method truncates a CME into a finite state, continuous time Markov chain, for which the set of ordinary differential equations, $\frac{d\mathbf{p}}{dt} = \mathbf{A}(t)\mathbf{p}$ describes the flow of probability among all of the most likely observable states for the system. Details of the FSP approach to solving chemical kinetic systems are provided in Supplementary Note 1.

For signal-activated transcription in the HOG-MAPK stress response pathway in yeast, an FSP model has been used to fit and predict mRNA distributions at a range of NaCl concentrations [8, 9]. This model of osmotic stress response consists of transitions between four different gene states, shown in Fig. 1a. The probability of a transition from the i^{th} to the j^{th} gene state in the infinitesimal time dt is given by $k_{ij}dt$. Each i^{th} state also has a corresponding mRNA transcription rate, k_{ri}, but the mRNA degrade with rate γ, independent of gene state. Further descriptions and validations of this model are given in Supplementary Note 1 and in [8, 9, 19]. To accurately fit and predict mRNA levels across cell populations, the authors in [8] cross-validated across a number of different potential models with different numbers of gene states and time varying parameters. The most predictive of these was the model shown in Fig. 1a, in which the transition rate from the second gene activation state to the first gene activation state is a function of nuclear MAPK levels, $f(t)$. The nuclear localization of MAPK affects this transition with a threshold function,

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

where α and β set the threshold for $k_{21}(t)$ activation/deactivation. Figure 1b (left) shows the nuclear localization dynamics of MAPK (i.e. $f(t)$) for osmotic stress responses to 0.2M and 0.4M NaCl, with simulated nuclear localization dynamics fit to a model (from [9],

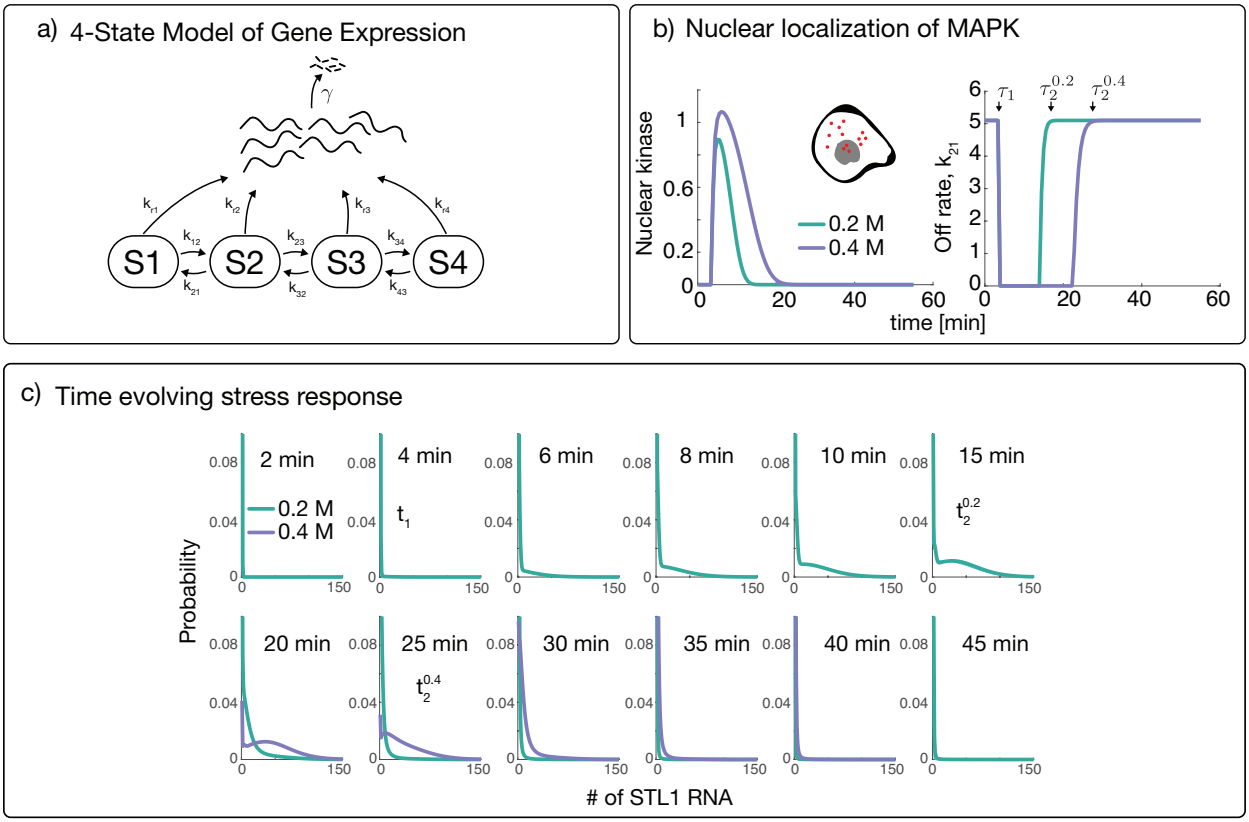

FIG. 1. *Stochastic modeling of osmotic stress response genes in yeast.* (a) Four-state model of gene expression, where each state transcribes mRNA at a different transcription rate, but all mRNA degrade at a single rate γ. (b) The effect of measured MAPK nuclear localization (depicted as red dots in the cell) (left) on the the rate of switching from gene activation state S2 to S1 (right) under 0.2M or 0.4M NaCl osmotic stress. The time at which k_{21} turns off is denoted with τ_1 and is independent of the NaCl level. The time at which k_{23} turns back on is given by τ_2^{NaCl} depending on the level of NACl. (c) Time evolution of the *STL1* RNA in response to the 0.2M and 0.4M NaCl stress.

Supplementary Note 2), and Fig. 1b (right) shows the value of $k_{21}(t)$ for each salinity level. This rate results in a time-varying generator $\mathbf{A}(t)$ for the master equation dynamics (See Supplementary Note 1).

LIKELIHOOD OF SMFISH DATA FOR FSP MODELS

To match FSP model solutions to single-cell data, one needs to compute and maximize the likelihood of the smFISH data given the FSP model [8, 9, 19, 28, 29]. 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 *STL1* transcription consists of four gene states (S1-S4, shown in Fig. 1a), 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:

$$\ell(\mathbf{D}|\boldsymbol{\theta}) = M \prod_{t=t_1}^{t_{N_t}} \prod_{i \in \mathcal{J}_D} p(\mathbf{x}_i^L, t|\boldsymbol{\theta})^{y_i}$$

$$\log \ell(\mathbf{D}|\boldsymbol{\theta}) = \sum_{t=t_1}^{t_{N_t}} \sum_{i \in \mathcal{J}_D} y_i \log(p(\mathbf{x}_i^L, t|\boldsymbol{\theta})) + \log M, \tag{2}$$

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 $p(\mathbf{x}_i^L)$ is the marginalized probability mass of the observable species,

$$p(\mathbf{x}_i^L) = \sum_{\mathbf{x}_j \in \mathbf{x}_i^L} p(\mathbf{x}_j).$$

Neglecting the term $\log M$, which is independent of the model, the summation in Eq. 2 can be rewritten as a product $\mathbf{y} \log \mathbf{p}^L$, where $\mathbf{y} \equiv [y_0, y_1, \ldots]$ is a 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 Eq. 2 with respect to $\boldsymbol{\theta}$ to find the *maximum likelihood estimates* (MLE) of the parameters, $\hat{\boldsymbol{\theta}}$, which will vary depending on each new set of experimental data. We next demonstrate how this likelihood function and the FSP model of the HOG-MAPK system can be used to design optimal smFISH experiments using the FSP-based FIM [6].

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, and which allows one to find experimental settings that can make these uncertainties as small as possible [3, 4, 30–33]. Recently, it has been applied to biological systems to estimate kinetic rate parameters in stochastic gene expression systems [3–6, 34]. In general, the FIM for a single measurement is defined:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left\{ \left(\nabla_{\boldsymbol{\theta}} \log \mathbf{p}(\boldsymbol{\theta})\right)^T \left(\nabla_{\boldsymbol{\theta}} \log \mathbf{p}(\boldsymbol{\theta})\right) \right\}, \tag{3}$$

where $\log \mathbf{p}(\boldsymbol{\theta})$ is the log-likelihood of observing that measurement, and the expectation is taken across 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 Fisher information matrix, i.e.,

$$\sqrt{N_c}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{dist} \mathcal{N}(0, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}). \tag{4}$$

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

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\theta_j}\end{bmatrix} = \begin{bmatrix}\mathbf{A}(t) & \mathbf{0}\\ \mathbf{A}_{\theta_j}(t) & \mathbf{A}(t)\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\theta_j}\end{bmatrix}, \tag{5}$$

where $\mathbf{A}_j = \frac{\partial \mathbf{A}}{\partial \theta_j}$. The FSP-FIM at a single time t is then given by:

$$\mathbf{F}(\boldsymbol{\theta}, t)_{j,k} = \sum_i \frac{1}{p(\mathbf{x}_i; t, \boldsymbol{\theta})} \mathbf{s}^i_{\theta_j}(t)\mathbf{s}^i_{\theta_k}(t), \tag{6}$$

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

$$\mathcal{I}(\boldsymbol{\theta}, \mathbf{t}, \mathbf{c}) = \sum_{l=1}^{N_t} c_l \mathbf{F}(\boldsymbol{\theta}, t = t_l), \tag{7}$$

where $\mathbf{c} = [c_1, c_2, \ldots, c_{N_t}]$ is the number of cells measured at each l^{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_{total}, one can find how many cells should be measured at each time point and which time points should be skipped entirely, (i.e., $c_l = 0$). We next verify the FSP-FIM for this stochastic model with a time-varying parameter, and later find the optimal $\mathbf{c}$ for *STL1* mRNA in yeast cells.

RESULTS

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 Eqn. 5. In our model, the temporal addition of osmotic shock causes nuclear translocation of HOG-MAPK, according to the time-varying function in Eq. 1.

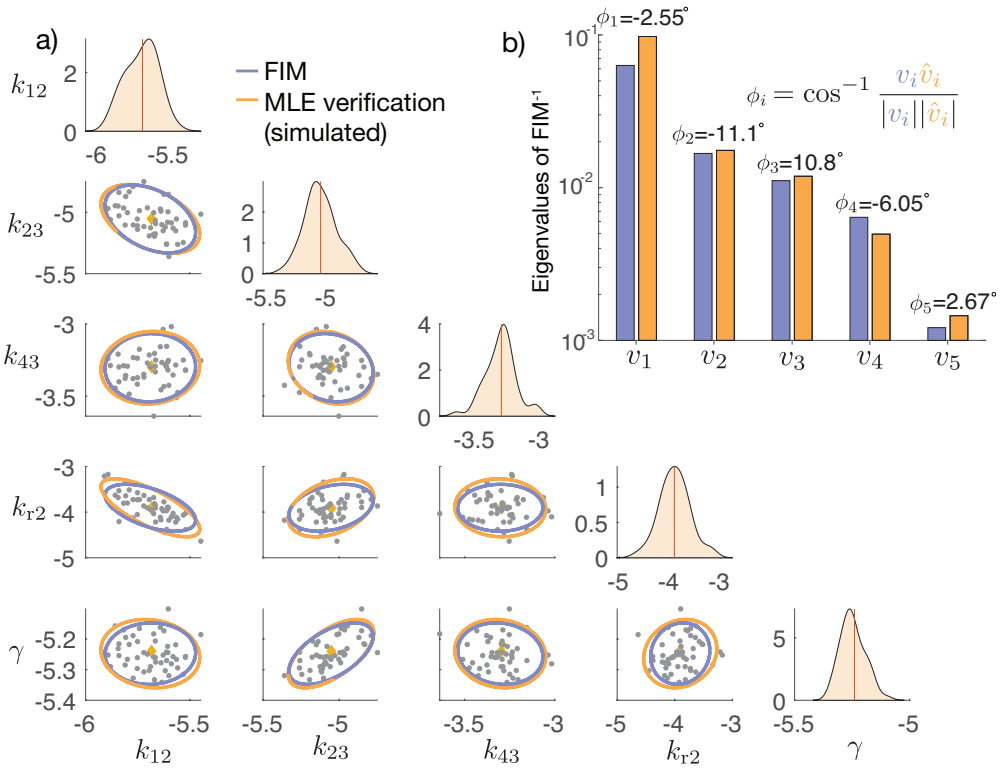

FIG. 2. *Verification of the FSP-FIM for the time-varying HOG-MAPK model.*(a) Scatter plots and density plots of the spread of MLE estimates for 50 simulated data sets for a subset of model parameters. All parameters are shown in logarithmic scale. The ellipses show the 95% CI for the inverse of the FIM (purple) and covariance of scatter plot (orange). The yellow dots indicate the parameters at which the FIM and simulated data sets 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 (ϕ_i) are shown in degrees.

Model parameters simultaneously fit to experimentally measured 0.2M and 0.4M *STL1* mRNA were adopted from [9] and used as a reference set of parameters (yellow dots in Fig. 2a and S1), which we define as $\boldsymbol{\theta}^*$. These reference parameters were used to generate 50 unique and independent simulated data sets, and each n^{th} simulated data set was fit to find the parameter set, $\hat{\boldsymbol{\theta}}_n$, that maximizes the likelihood for that simulated data set. This process was repeated for two different experiment designs, including the original intuitive design from [9] (results shown in Fig. 2) and an optimized design discussed below (results shown in Fig. 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 $\{\hat{\boldsymbol{\theta}}_n\}$ and are shown as yellow dots in Figs. 2 and S1. Using the asymptotic normality of the maximum likelihood estimator and its relationship to the FIM (Eq. 4), 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 Figs. 2a 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, Σ_{MLE}, in Figs. 2b and S1b. For further validation, we noted that the principle directions of the ellipses in Figs. 2a and S1a also match for the FIM and MLE analyses, as quantified by the angle between the paired FIM and Σ_{MLE} eigenvectors (Figs. 2b and S1b). For comparison, the angles between rank-matched eigenvectors of the FIM and Σ_{MLE} were all less than 12°, whereas non rank-matched eigenvectors were all greater than 79.9°. With the FSP-FIM verified for the HOG-MAPK model, we next explore how the FIM can be used to optimally allocate the number of cells to measure at each time after osmotic shock.

Designing optimal measurements for the HOG-MAPK pathway in *S. cerevisae*

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 0.2M and 0.4M data in [9] (non-spatial model) to establish a baseline parameter set that is experimentally realistic. These baseline 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 [34], 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,

$$\mathbf{c}_{\text{opt}} = \arg\max_{\mathbf{c}} |\mathcal{I}(\mathbf{c};\boldsymbol{\theta})|_{D_s} \text{ such that } \sum_{l=1}^{N_t} c_l = 1, \tag{8}$$

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 (Eqn. 7). The fraction of cells to be measured at each time point, $\mathbf{c}$ 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 0.2M NaCl input and compares these fractions to the experimentally measured number of cells from [9]. While each available time point was allocated a non-zero 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 data sets of 1,000 cells each and found the MLE estimates for each sub-sampled data set. We compared the spread of these MLE estimates to the inverse of the optimized FIM, shown in Fig. S1.

Comparing Figs. S1 with Fig. 2 illustrates the increase in information of the optimal 0.2M experiment compared to the intuitively designed experiment from [9]. 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 original intuitive design.

Figure 4 next compares the D_s-optimality criteria for the optimal (solid horizontal lines) and intuitive ([9], dashed horizontal lines) experiment designs to 1,000 randomly designed experiments for the 0.2M (black) and 0.4M (gray) 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 multinomial distribution with equal

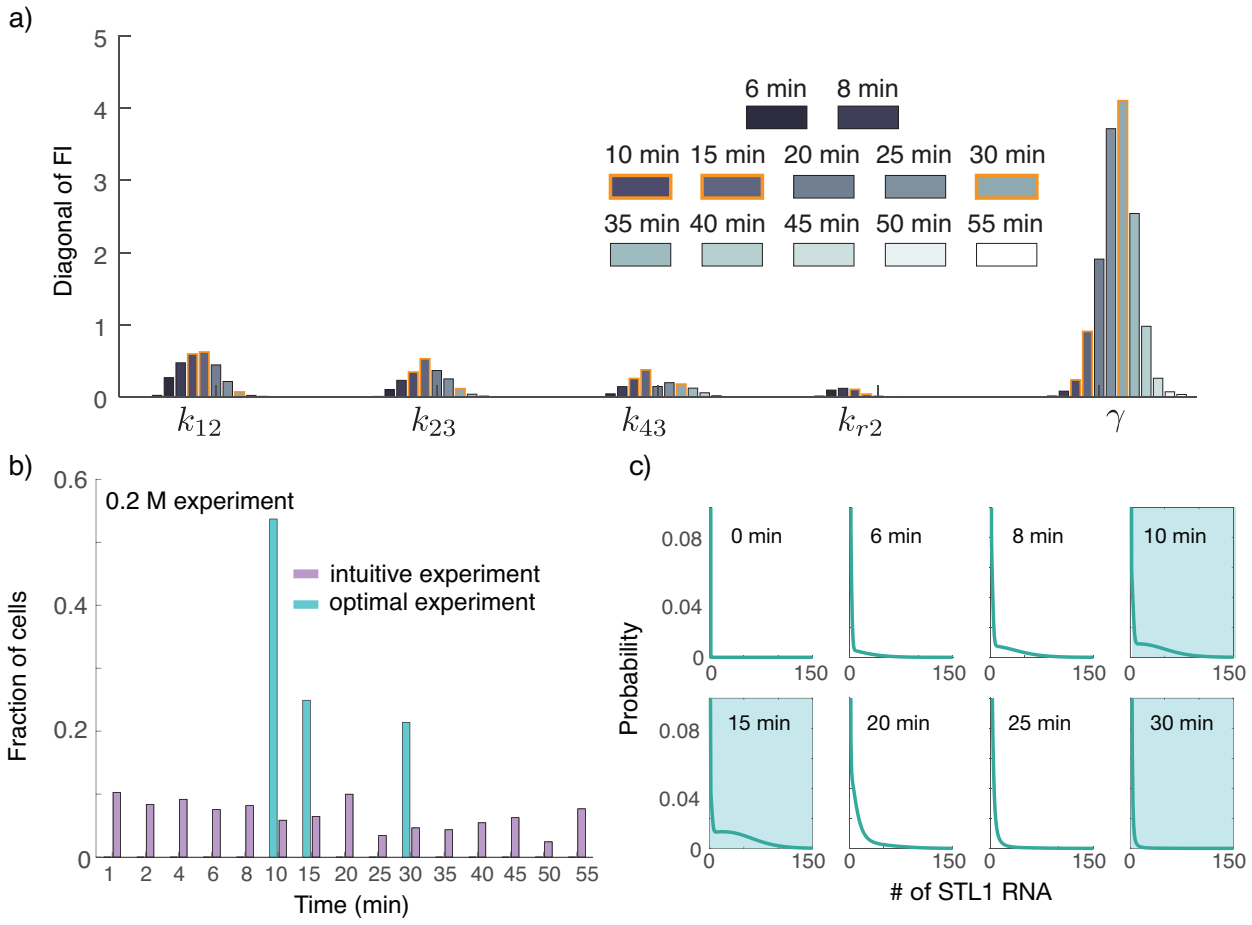

FIG. 3. *Optimizing the allocation of cell measurements at different time points.* (a) Diagonal entries of the Fisher information at different measurement times. The optimal measurement times $t = [10, 15, 30]$ minutes are highlighted in orange. (b) Comparison of optimal fractions cells to measure (blue) 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 [9]. (c) Probability distributions of *STL1* mRNA at several of measurement times. The blue boxes denote the time points of optimal measurements.

probability for each time point. Figures 4a-b show that the intuitive experiment is more informative than most random experiments, but is still substantially less informative than the optimal experiment. To explore the importance of knowing the exact process input dynamics prior to designing the experiment, we next asked how well an experiment design optimized for a 0.2M osmotic shock would do to estimate parameters using an 0.4M experi-

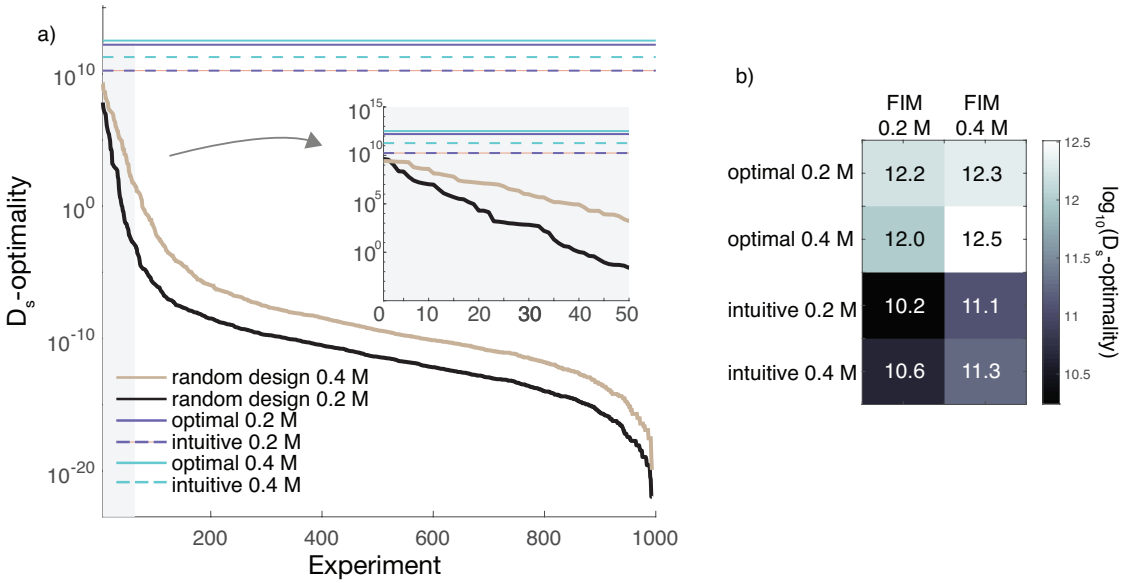

FIG. 4. *Information gained by performing optimal experiments compared to actual experiments* (a) D_s-optimality for optimal design using three time points compared to the intuitive experiment design made using 16 time points (purple, 0.2M and blue, 0.4M). Dashed lines represent intuitive experiment designs. Randomly designed experiments with 0.2 M and 0.4 M NaCl are shown in black and gray. 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-optimality for different experiment designs (y-axis) computed using the Fisher information for either the 0.2 M perturbation or the 0.4 M NaCl perturbation.

ment and vice-versa. Figure 4b shows that the simpler optimal experiment designs perform better than the intuitive designs in all cases, even when the design was found assuming a different environmental condition.

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.

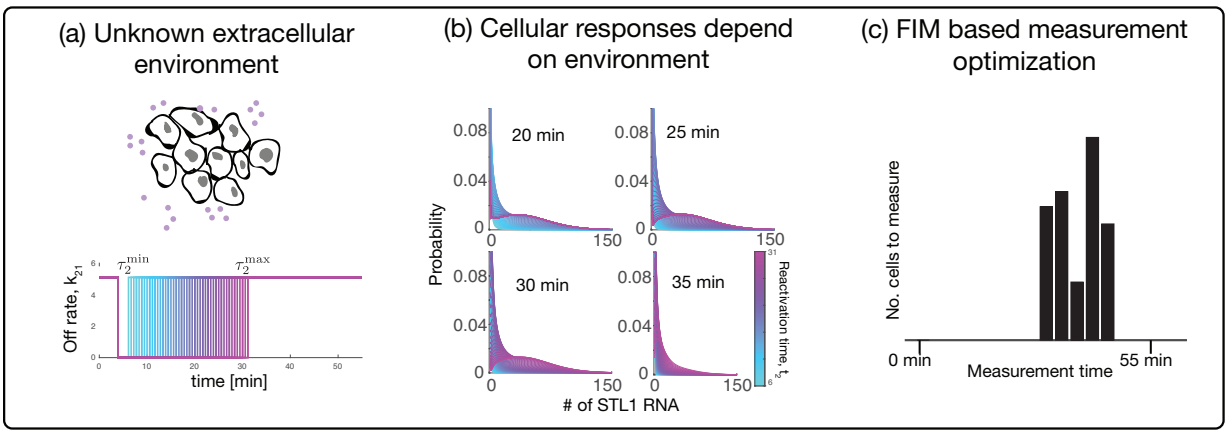

FIG. 5. *Overview of optimal design for biosensing experiments for the osmotic stress response in yeast.* (a) Unknown salt concentrations (purple dots) in the environment give rise to different reactivation times, τ_2, which affect the gene expression in the model through the rate k_{21}. These different reactivation times cause downstream *STL1* expression dynamics to behave differently as shown in panel (b). (c) Different responses can be used to resolve experiments that reduce the uncertainty in τ_2.

Toward this end, we assume a known and parametrized model (i.e., the model defined above, which was identified previously in [9]), 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 at $t = 0$ – 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.

In the HOG-MAPK transcription model, extracellular osmolarity ultimately affects stress response gene transcription levels through the time-varying parameter $k_{21}(t)$ in Eq. 1, and Fig. 1b shows the effect 0.2M and 0.4M salt concentrations on k_{21} activation. Higher salt concentrations delay the time at which $k_{21}(t)$ returns to its nonzero value, and the function in Eq. 1 is well-approximated by a the sum of three Heaviside step functions, $u(t - \tau_i)$ as:

$$k_{21}(t) = k_{21}^0 \left(u(t) - u(t - \tau_1) + u(t - \tau_2)\right), \tag{9}$$

where τ_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, [8, 9]), and τ_2 is the variable time it takes for the nuclear kinase to drop back below that threshold. In practice, the threshold-crossing time, τ_2, is directly related to the salt concentration experienced by the cell under reasonable salinity levels. This relationship is shown in Fig. 1b and 5b, where a 0.2M NaCl input exhibits a shorter τ_2 than does a 0.4M input. For our analyses, we assume a prior uncertainty such that time τ_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 τ_2 (and therefore the concentration of NaCl).

To reformulate the FSP-FIM to estimate uncertainty in τ_2 given our model, the first step is to compute the sensitivity of the distribution of mRNA abundance to changes in the variable τ_2 using Eqn. 5, in which $\mathbf{A}_{\theta_j}(t)$ is replaced with $\mathbf{A}_{\tau_2}(t) = \frac{\partial \mathbf{A}}{\partial \tau_2}$ as follows:

$$\frac{d}{dt}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\tau_2}\end{bmatrix} = \begin{bmatrix}\mathbf{A}(t) & \mathbf{0}\\ \mathbf{A}_{\tau_2}(t) & \mathbf{A}(t)\end{bmatrix}\begin{bmatrix}\mathbf{p}\\ \mathbf{s}_{\tau_2}\end{bmatrix}. \tag{10}$$

As $k_{21}(t)$ is the only parameter in $\mathbf{A}$ that depends explicitly on τ_2, all entries of $\frac{\partial \mathbf{A}}{\partial \tau_2}$ are zero except for those which depend on $k_{21}(t)$, and

$$\mathbf{A}_{\tau_2}(t) = \frac{\partial \mathbf{A}}{\partial k_{21}}\frac{\partial k_{21}}{\partial \tau_2} = \mathbf{A}_{k_{21}} k_{21}^0 \delta(\tau_2), \tag{11}$$

and therefore $\mathbf{A}_{\tau_2} = \frac{\partial \mathbf{A}}{\partial \tau_2}$ is non-zero 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:

$$\frac{d}{dt}\mathbf{s}_{\tau_2} = \begin{cases}\mathbf{0} \text{ for } t < \tau_2 \text{ with } \mathbf{s}(0) = \mathbf{0}\\ \mathbf{A}(t)\mathbf{s}_{\tau_2} \text{ for } t > \tau_2 \text{ with } \mathbf{s}_{\tau_2}(\tau_2) = k_{21}^0 \mathbf{A}_{k_{21}}\mathbf{p}(\tau_2)\end{cases}. \tag{12}$$

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 τ_2 can be computed as the dot product of these two vectors,

$$\mathcal{I}(\tau_2) = \sum_{l=1}^{N_t} c_l f_l = \mathbf{c}^T \mathbf{f}. \tag{13}$$

Our goal is to find an experiment that is optimal to determine the value of τ_2, given an assumed prior that τ_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 τ_2, we integrate the inverse of the FIM in Eq. 13 over the prior uncertainty in τ_2,

$$\mathbf{c}_{\text{opt}} = \underset{\mathbf{c}, \sum c_l = 1}{\arg\min} \int_{\tau_2^{\min}}^{\tau_2^{\max}} \frac{1}{\tau_2^{\max} - \tau_2^{\min}} \mathcal{I}^{-1}(\mathbf{c}; \tau_2 = \tau, \boldsymbol{\theta}) d\tau \tag{14}$$

$$= \underset{\mathbf{c}, \sum c_l = 1}{\arg\min} \int_{\tau_2^{\min}}^{\tau_2^{\max}} \mathcal{I}^{-1}(\mathbf{c}; \tau_2 = \tau, \boldsymbol{\theta}) d\tau. \tag{15}$$

For later convenience, we define the integral in Eq. 14 (i.e., the objective function of the minimization) by the symbol $\mathcal{J}$, which corresponds to the expected uncertainty about the value of τ_2 for a given $\mathbf{c}$.

Next, we apply the greedy search from above to solve the minimization problem in Eqn. 15 to find the experiment design $\mathbf{c}_{\text{opt}}$ that minimizes the estimation error of τ_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 right), 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 τ_2, which is shown by the orange bars in Fig. 6. To verify these estimates, we then chose 64 uniformly spaced values of τ_2, which we denote as the set $\{\tau_2^{\text{true}}\}$, and for each τ_2^{true}, we simulated 50 random data sets of 1,000 cells distributed according to the specified experiment designs. For each of the 64×50 simulated data sets, we then determined the value τ_2^{MLE} between $\tau_2^{\min}$ and $\tau_2^{\max}$ that maximized the likelihood of the simulated data according to Eq. 2. The root mean squared estimate (RMSE) error over all random values of τ_2^{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 τ_2 than did the uniform or random experimental designs. A slightly

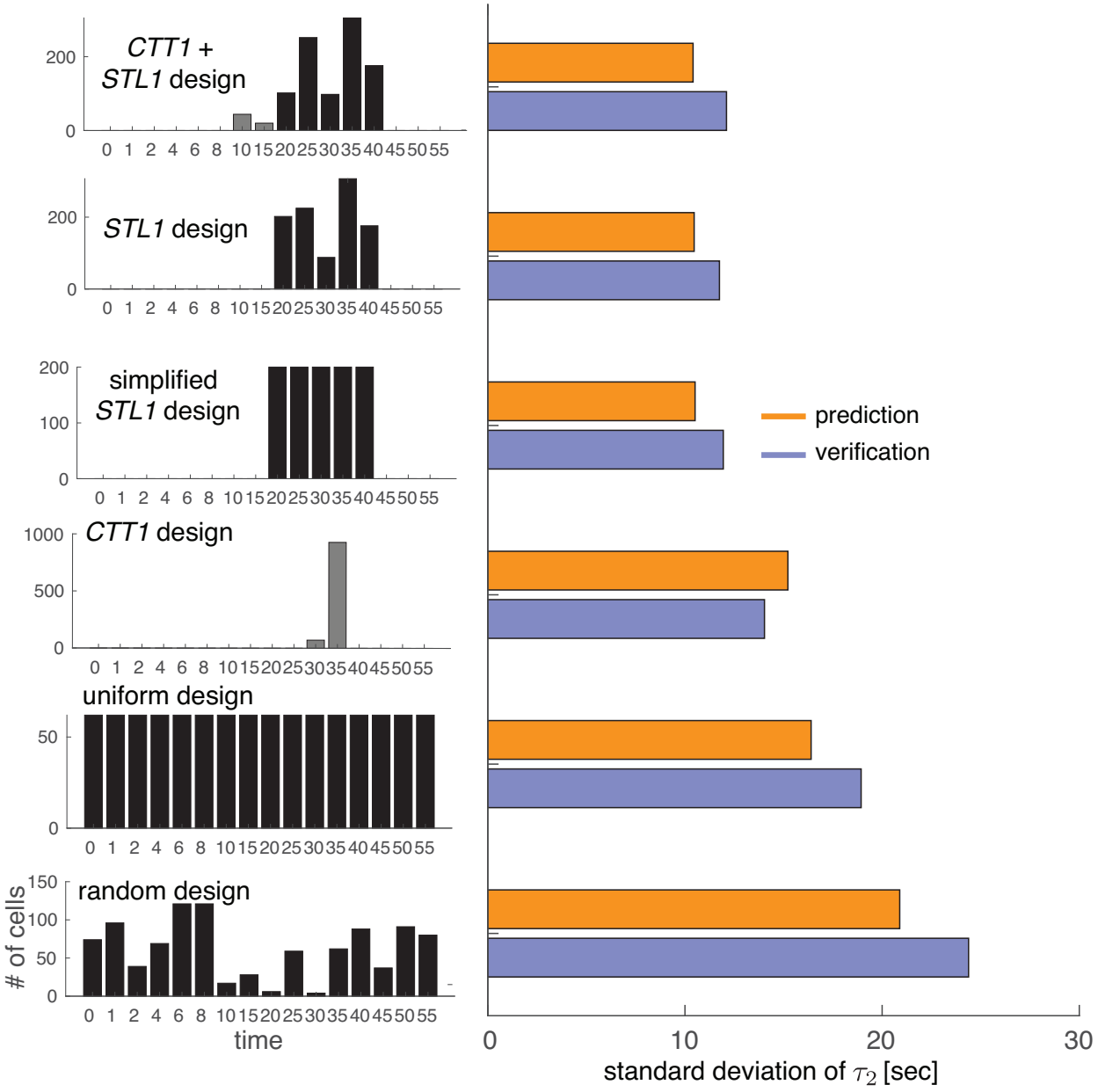

FIG. 6. *Verification of the uncertainty in τ_2 for different experiment designs.* The left panel shows various experiment designs, where the sum of the bars (i.e., the total number of measurements) is 1,000. Gray bars represent the measurements of *CTT1* and black bars *STL1*. The right panel shows the value of the objective function in Eq. 14 for each experiment design in orange, and the MSE values for verification are shown in purple.

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 Fig. 6 includes the best design that only uses *STL1* (second from top), the best design that uses only *CTT1* (fourth from top), and the best designs that uses some cells with *CTT1* and some with *STL1* (top design). To find the best experiment design for measurement of two different genes, we assumed that at each time, either *STL1* mRNA *or* *CTT1* mRNA (but not both) could be measured, corresponding to using smFISH oligonucleotides for either *STL1* or *CTT1*. To determine which gene should be measured at each time, we compute the Fisher information for *CTT1* and *STL1* for every measurement time and averaged this value over the range of τ_2. For each measurement time t_l, the gene is selected that has the higher average Fisher information for τ_2. The number of cells per measurement time were then optimized as before, except the choice to measure *CTT1* or *STL1* was based on which mRNA had the larger Fisher information (Eq. 13) at that specific point in time. The best *STL1*-only experiment design was found to yield uncertainty of 10.5 seconds (standard deviation); the best *CTT1*-only experiment was found to yield an uncertainty of 15.2 seconds and the best mixed *STL1*/*CTT1* experiment design was found to yield an uncertainty of 10.4 seconds. In other words, for this case the *STL1* gene was found to be much more informative of the environmental condition than was *CTT1*, and the use of both *STL1* and *CTT1* provides only minimal improvement beyond the use of *STL1* 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 τ_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.

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 [8, 9]. 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 STL1 gene only required three time

points. Moreover, these three time points ($t = [10, 15, 30]$ minutes, highlighted by blue in Fig. 3b) are at biologically meaningful time points. At $t = 10$ and 15 minutes, the system is increasing to maximal expression, and the probability of measuring a cell with elevated of RNA 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 Fig. 3a, 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 information about γ. Because γ 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., γ 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 [9], 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 performed 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 (Fig. 4a). 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 Fig. 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 compared to assumptions used in the experiments design (Fig. 4b), a fact that allows for effective experiment designs despite inaccurate prior assumptions.

In addition to suggesting optimal experiments to identify model parameters, we showed that the FSP-FIM combined with an existing model could be used to design optimal exper-

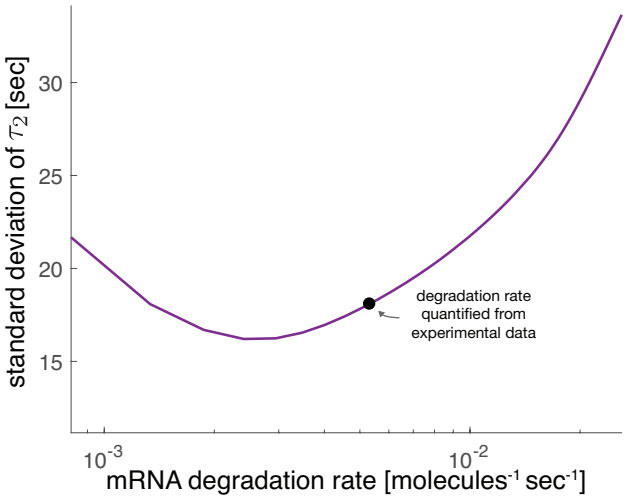

FIG. 7. *Optimal mRNA degradation rates to reduce uncertainty about the extracellular environment.* Uncertainty in the time at which the *STL1* gene turns off, τ_2, as a function of mRNA degradation rate (purple). The black dot corresponds to the degradation rate that was quantified from experimental data.

iments to learn about fluctuating extracellular environments (Figs. 5 and 6). 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, Fig. 7 shows the expected uncertainty in τ_2 as a function of the degradation rate of the *STL1* 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 *STL1* degradation rate to most accurately determine the extracellular fluctuations would be 2.4×10^{-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 [9]. This result is consistent with our earlier finding that the faster degrading *STL1* mRNA is a much better determinant of the HOG1 dynamics than is the slower-degrading *CTT1* 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 to 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 [35–37], and recent work in [38] 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 creates 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 [39–41] or optogenetic control [42–44]. Many such experiments involve stochastic bursting behaviors at the mRNA or protein level [7–9, 43], 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.

DATA AVAILABILITY

All data and codes associated with this article will be made available upon acceptance of the article at: `https://github.com/MunskyGroup/fox_et_al_complexity_2019`.

ACKNOWLEDGEMENTS

ZRF and BEM 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 National Institutes of Health [DP2 GM11484901, R01GM115892] and Vanderbilt Startup Funds. The presented analyses used the computational resources of the W M Keck High Performance Compute Cluster supported under a W M Keck Foundation Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

[1] J. Liepe, S. Filippi, M. Komorowski, and M. P. H. Stumpf, PLOS Computational Biology **9**, 1 (2013).

[2] J. F. Apgar, D. K. Witmer, F. M. White, and B. Tidor, Molecular BioSystems **6**, 1890 (2010).

[3] J. Ruess, A. Milias-Argeitis, and J. Lygeros, Journal of The Royal Society Interface **10** (2013).

[4] M. Komorowski, M. J. Costa, D. A. Rand, and M. P. H. Stumpf, Proceedings of the National Academy of Sciences of the United States of America **108**, 8645 (2011).

[5] C. Zimmer, PloS One **11**, e0159902 (2016).

[6] Z. R. Fox and B. Munsky, PLoS computational biology **15**, e1006365 (2019).

[7] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, Proceedings of the National Academy of Sciences **109**, 8340 (2012).

[8] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, Science **339**, 584 (2013).

[9] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, Proceedings of the National Academy of Sciences of the United States of America **163**, 201804060 (2018).

[10] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, Nature Methods **5**, 877 (2008).

[11] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, Science **280**, 585 (1998).

[12] N. Tsanov, A. Samacoits, R. Chouaib, A.-M. Traboulsi, T. Gostan, C. Weber, C. Zimmer, K. Zibara, T. Walter, M. Peter, E. Bertrand, and F. Mueller, Nucleic Acids Research **44**, e165 (2016), http://oup.prod.sis.lan/nar/article-pdf/44/22/e165/25365000/gkw784.pdf.

[13] Sharifian, Hoda, Lampert, Fabienne, Stojanovski, Klement, Regot, Sergi, Vaga, Stefania, Buser, Raymond, Lee, Sung Sik, Koeppl, Heinz, Posas, Francesc, Pelet, Serge, and Peter, Matthias, Integrative Biology : Quantitative Biosciences from Nano to Macro **7**, 412 (2015).

[14] Klipp, Edda, Nordlander, Bodil, Krüger, Roland, Gennemark, Peter, and Hohmann, Stefan, Nature Biotechnology **23**, 975 (2005).

[15] B. Schoeberl, C. Eichler-Jonsson, E. D. Gilles, and G. Müller, Nature biotechnology **20**, 370 (2002).

[16] Muzzey, Dale, Gómez-Uribe, Carlos A, Mettetal, Jerome T, and van Oudenaarden, Alexander, Cell **138**, 160 (2009).

[17] Saito, Haruo and Posas, Francesc, Genetics **192**, 289 (2012).

[18] S. Pelet, F. Rudolf, M. Nadal-Ribelles, E. de Nadal, F. Posas, and M. Peter, Science (New York, N.Y.) **332**, 732 (2011).

[19] B. Munsky, Z. Fox, and G. Neuert, Methods **85**, 12 (2015).

[20] Zechner, Christoph, Unger, Michael, Pelet, Serge, Peter, Matthias, and Koeppl, Heinz, Nature methods **11**, 197 (2014).

[21] R. M. Kumar, P. Cahan, A. K. Shalek, R. Satija, A. J. DaleyKeyser, H. Li, J. Zhang, K. Pardee, D. Gennert, J. J. Trombetta, T. C. Ferrante, A. Regev, G. Q. Daley, and J. J. Collins, Nature **516**, 56 (2014).

[22] L. S. Weinberger, J. C. Burnett, J. E. Toettcher, A. P. Arkin, and D. V. Schaffer, Cell **122**, 169 (2005).

[23] B. Munsky, G. Neuert, and A. van Oudenaarden, Science **336**, 183 (2012).

[24] N. G. Van Kampen and N. Godfried, *Stochastic processes in physics and chemistry* (Elsevier, 1992).

[25] D. T. Gillespie, The Journal of Physical Chemistry **81**, 2340 (1977).

[26] A. Singh and J. P. Hespanha, IEEE Transactions on Automatic Control **56**, 414 (2011).

[27] B. Munsky and M. Khammash, The Journal of Chemical Physics **124**, 044104 (2006).

[28] Z. Fox, G. Neuert, and B. Munsky, Journal of Chemical Physics **145** (2016).

[29] M. Gomez-Schiavon, L.-F. Chen, A. E. West, and N. E. Buchler, Genome biology **18**, 164 (2017).

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

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

[32] C. Kreutz and J. Timmer, The FEBS Journal **276**, 923 (2009).

[33] B. Steiert, A. Raue, J. Timmer, and C. Kreutz, PloS One **7**, e40052 (2012).

[34] J. Ruess, F. Parise, A. Milias-Argeitis, M. Khammash, and J. Lygeros, Proceedings of the National Academy of Sciences of the United States of America **112**, 8148 (2015).

[35] R. Cheong, A. Rhee, C. J. Wang, I. Nemenman, and A. Levchenko, Science (New York, N.Y.) **334**, 354 (2011).

[36] R. Suderman, J. A. Bachman, A. Smith, P. K. Sorger, and E. J. Deeds, Proceedings of the National Academy of Sciences of the United States of America **114**, 5755 (2017).

[37] J. Selimkhanov, B. Taylor, J. Yao, A. Pilko, J. Albeck, A. Hoffmann, L. Tsimring, and R. Wollman, Science (New York, N.Y.) **346**, 1370 (2014).

[38] T. Jetka, K. Nienałtowski, S. Filippi, M. P. H. Stumpf, and M. Komorowski, Nature Communications **9**, 4591 (2018).

[39] Ng, Andrew H, Nguyen, Taylor H, Gomez-Schiavon, Mariana, Dods, Galen, Langan, Robert A, Boyken, Scott E, Samson, Jennifer A, Waldburger, Lucas M, Dueber, John E, Baker, David, and El-Samad, Hana, Nature **572**, 265 (2019).

[40] A. Thiemicke, H. Jashnsaz, G. Li, and G. Neuert, Scientific reports **9**, 10129 (2019).

[41] J.-B. Lugagne, S. Sosa Carrillo, M. Kirch, A. Köhler, G. Batt, and P. Hersen, Nature Communications **8**, 1671 (2017).

[42] R. Chait, J. Ruess, T. Bergmiller, G. Tkačik, and C. C. Guet, Nature Communications **8**, 2557 (2017).

[43] M. Rullan, D. Benzinger, G. W. Schmidt, A. Milias-Argeitis, and M. Khammash, Molecular Cell **70**, 745 (2018).

[44] S. M. Castillo-Hair, E. A. Baerman, M. Fujita, O. A. Igoshin, and J. J. Tabor, Nature Communications **10**, 3099 (2019).