Designing single-cell experiments to harvest fluctuation information while rejecting measurement noise

Huy D. Vo
Huy.Vo@colostate.edu

Brian Munsky
Brian.Munsky@colostate.edu

February 18, 2022

Abstract

Measurement error is a complicating factor that reduces or distorts the information contained in any experiment. This problem becomes even more serious in the context of experiments that are specifically designed to quantify single-cell gene expression heterogeneity, where important RNA and protein copy numbers are themselves subject to the inherently random fluctuations of biochemical reactions. It is not clear how measurement noise should be managed, in addition to other experiment design variables such as sampling size and frequency, to ensure that collected data will provide useful insights on gene expression mechanisms of interest. To address these fundamental single-cell experiment design challenges, we propose a computational framework that makes explicit use of measurement error modeling and Fisher Information Matrix-based criteria to decide between experimental methods. This unified approach not only quantifies how different noise characteristics affect uncertainty in parameter estimation, but also enables a systematic approach to design hybrid experiments that combine different measurement modalities.

1 Introduction

Heterogeneity in gene expression at the single-cell level has wide-ranging biological and clinical consequences, from bacterial persistence [1] and viral infections [2] to tumor heterogeneity [3, 4]. Beside genetic and environmental factors, a significant degree of heterogeneity is caused by biochemical noise [5, 6, 7]. Therefore, even genetically identical cells grown in the same experimental conditions may display variability in their response to environmental stimuli. This variability, often termed *intrinsic noise* when it originates within the pathway of interest or *extrinsic noise* when it originates outside the pathway of interest, obscures the underlying mechanisms when looked through the lens of deterministic models and bulk measurements [8]. Yet, this so-called noise can be highly

informative when examined through the lens of single-cell measurements coupled with the mathematical modeling framework of the Chemical Master Equation (CME) [9, 10, 11]. Through this joint experimental and modeling approach, mechanisms of single-cell gene expression can be explained, predicted [12, 13, 8], or even controlled [14, 15, 16].

In this paper, we are concerned with how and when stochastic models based on the CME could be inferred with high confidence from single-cell experiments such as flow cytometry or single-molecule *in situ* hybridization (smFISH), from which datasets are produced in the form of fluorescence intensity histograms [17, 18, 19] or molecular count histograms [20, 21, 22]. These measurements, like any other experimental measurement, can be corrupted by errors from imprecise detection or data processing techniques. In flow cytometry, variations in the binding efficiency, binding specificity, or intensity of individual fluorescent reporters or probes, or the existence of background fluorescence, are unavoidable perturbations that may obscure the true copy number of RNA or protein [12, 23, 24]. While smFISH datasets are generally considered to provide the 'gold standard' for measuring transcriptional heterogeneity, the estimation of molecular copy numbers in single cells depends heavily on image segmentation and spot counting algorithms [25, 26, 27, 28] that involve several threshold parameters, which are set *ad hoc* and typically vary from one expert to another. In addition, the sensitivity of smFISH measurements can be heavily affected by the length of the target mRNAs, the number of probes [29], or the number of hybridization steps [30].

The facts of measurement noise motivate intriguing questions on the design of single-cell experiments. For instance, under what conditions can detailed statistical measurement noise models and cheap single-cell measurements combine to replace accurate, yet more expensive experimental equipment? How approximate or coarse (and therefore fast) can image processing be while still faithfully retaining knowledge about parameters or mechanisms of interest? These questions are becoming more pressing as new advances in fluorescence tagging and microscopy technology are leading to more sophisticated experimental protocols to produce ever increasing data on complex gene expression networks. Addressing these challenges is non-trivial, as it requires the careful consideration of the potentially nonlinear combination of measurement noise and the biological question of interest. This paper advocates for a model-driven approach that has an explicit component of measurement error modeling. Our key idea is to use the Fisher Information Matrix (FIM) to quantify how much information is expected from a given observation. The FIM is the basis for a large set of tools for optimal experiment design in a myriad of science and engineering fields [31, 32, 33, 34, 35, 36]. It has been employed in the study of identifiability and robustness of deterministic ODEs models in system biology [37, 38] as well as for designing optimal bulk measurement experiments [39, 40, 41]. As an early application of the FIM to stochastic modeling of gene expression, Komorowski et al. [42] devised a numerical method to compute the FIM based on the Linear Noise Approximation (LNA) of the CME, which they used to demonstrate the different impacts of time-series, time snapshots, and bulk measurements to pa-

rameter uncertainties. There have been subsequent work on approximating the FIM with moment closure techniques [34, 23], with demonstrable effectiveness on designing optimal optogenetic experiments [43]. Fox and Munsky [44] has recently extended the Finite State Projection (FSP) algorithm [44] to compute the FIM of single-cell observations under the assumption of exact measurements that can deal with the complexities of single-cell gene regulation especially when the gene expression distributions are asymmetrical and multi-modal.

In this work, we develop a new computational approach (Fig. 1) to approximate the FIMs for the *observed* single-cell measurements without making restrictive assumptions on the shape of measurement error distributions. By building on the FSP-based FIM of Fox and Munsky [44], the approach is applicable to stochastic gene circuits with low-copy numbers of gene products. In addition, we introduce a new component that we term the Probabilistic Distortion Operator (PDO) that enables the explicit treatment of measurement noise that could be temporal, discrete, non-symmetric, non-Gaussian and even the result of its own complex stochastic process dynamics. As a consequence, we can quantitatively compare a broader variety of lossy measurements by computing the FIMs associated with these noisy measurements themselves and estimating their impacts on the amount of information in the resulting experimental data, as well as designing hybrid experiments that combine diverse experimental techniques. As we will show below, the proposed extended FIM approach could, in particular, help in deciding when one should choose a more accurate but lower throughput measurement method, an alternative with higher throughput but lower accuracy, or some optimal combination of the two.

2 Results

Many models have been proposed to describe the heterogeneity of gene expression [12, 45, 46, 47, 13, 48, 49], and the experimental investigation of these tools have used many different experimental approaches, each with their own characteristic measurement errors [50]. Our proposed Fisher Information Analysis tools (methods section) are amenable to explore the impact of a very broad intersection of these models and measurement distortion effects.

2.1 Distorted single-cell measurements sample a probability distribution that is a linear transformation of their true molecular count distribution

In the context of single-cell smFISH and flow cytometry experiments, let $\boldsymbol{X}$ be the stochastic process that models the copy numbers of mRNAs and proteins of interest, and $\boldsymbol{Y}$ represents the stochastic process of observed quantities such as the number of spots in single-cell images, or the fluorescence intensities recorded in flow cytometry analysis. Viewing $\boldsymbol{Y}$ as a probabilistic distortion of $\boldsymbol{X}$, we seek to model the conditional probability $P(\boldsymbol{Y}|\boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$. By making appropriate conditional independence assumptions on $\boldsymbol{Y}$ (details in Methods

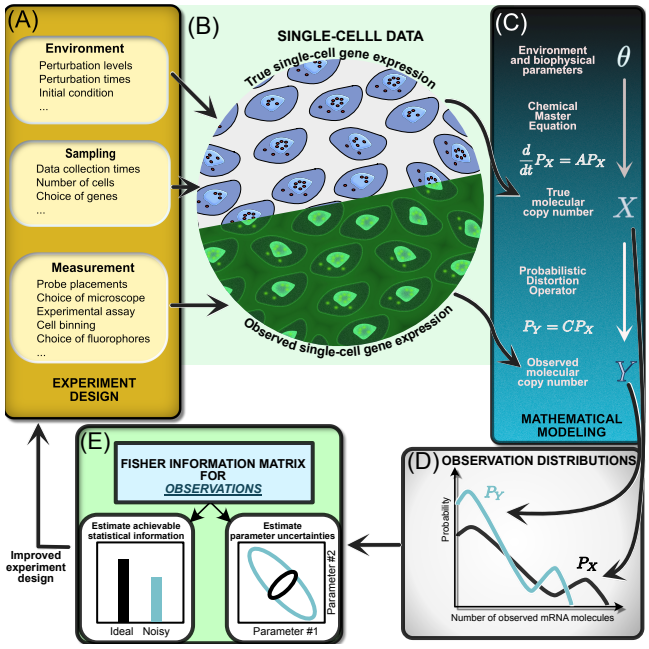

Figure 1: Summary of our iterative approach to designing and analyzing single-cell microscopy experiment. **Box (A):** Experiment design considerations can be roughly divided into three categories: environmental conditions, sampling conditions, and measurement conditions. How these conditions are specified affect the dynamics of single-cell gene expression, as well as how those dynamics is visualized/quantified in the microscope, as illustrated in **Box (B)**. We conceptualize a two steps information flow from the parameters that encode the dynamics (modeled as a random process θ) to the observed process Y (**Box (C)**). The first step is the molecular noise govern by the parameter-dependent Chemical Master Equation, giving rise to fluctuations in single-cell chemical abundances X. The actual values of these quantities, however, are filtered through another noisy channel of the measurement process, which can be modeled by a Probabilistic Distortion Operator that we propose in this paper. The distortion on the distributions of *observed* copy numbers can be computed by extending the classical Finite State Projection algorithm (**Box D**). We can then compute the Fisher Information Matrix of the distorted distributions themselves, allowing us to estimate the increase in parameter uncertainties and the loss in information due to different forms of measurement error (**Box E**). Such estimates can be used to inform new experiment designs in order to alleviate the distorting effects of measurement noise in single-cell data.

section), we deduce that the probability distribution $\boldsymbol{p_Y}(t)$ of a distorted measurement $\boldsymbol{Y}(t)$ is a linear transformation of the probability distribution $\boldsymbol{p_X}(t)$ of the true cellular state $\boldsymbol{X}(t)$. We refer to this linear transformation as the *probabilistic distortion operator* (PDO). This matrix is element-wise the conditional probability density/mass function of $\boldsymbol{Y}(t)$ given $\boldsymbol{X}(t)$. Specifying appropriate statistical distributions for this conditional probabilistic dependence gives us concrete models of $\boldsymbol{Y}$ for different types of measurement distortion. These transformations allow us to obtain the expected distribution of observations from the solution of any chemical master equation (CME) that can be computed either numerically or analytically. In addition, we can 'back-propagate' through the distortion operator to find the derivatives of the observation distributions with respect to model parameters. This allows us to compute the Fisher Information Matrix of the observation variable $\boldsymbol{Y}$ itself under different assumptions on the type and degree of measurement noise. We give concrete examples of our approach for the rest of this section.

2.2 Fisher Information Matrix Analysis reveals how optimal experiment design changes with the measurement approach

We first illustrate our methods using the random telegraph model, one of the simplest, yet most commonly utilized models of bursty gene expression [51, 21, 52, 53, 54, 55]. In this model, a gene is either in the inactive or active state, with transition between states occurring randomly with average rates k_{ON} (to activate the gene) and k_{OFF} (to deactivate the gene). When active, the gene can be transcribed with an average rate k_r to produce mRNA molecules, each of which degrade with an average rate γ. These parameters cannot be measured directly and must be statistically inferred from datasets collected using single-molecule, single-cell experimental methods such as smFISH [21, 56], flow cytometry [18, 19, 24], or live-cell imaging [52, 57]. In this work, we focus only on the former two experimental approaches in which the data consists of independent single-cell measurements taken at different times. Using the telegraph model, we will examine measurement distortion effects corresponding to probe binding inefficiency and spot colocolization (in smFISH experiments) or reporter fluorescence intensity variability (in flow cytometry experiments).

We consider five formulations of measurement distortion matrix (cf.eq. (3) in Methods), corresponding to scenarios in which experimental errors arise from either inefficient mRNA detection, additive false positives, or spatial integration of cell fluorescence intensity. Specifically,

1. The first model concerns with when $\tilde{y}$ is obtained from a "lossy" spot-counting process applied on images taken in an smFISH experiment. We model $y|\boldsymbol{x} = n$ with a binomial distribution $B(p_{detect}, n)$, where each spot has a chance $p_{miss} := 1 - p_{detect}$ of being ignored by the counting algorithm, resulting in underestimation of the true mRNA copy number (a similar model can be found in the context of single-cell RNA sequenc-

ing [55, 58]). We choose $p_{detect} = p_{miss} := 0.5$ in this example, but it is straightforward to generalize to different values (see Supplemental section S3.2). We call this distortion "Missing spots with uniform rate".

2. The second model is a simple variation of the first model in which p_{detect} varies with the number of mRNA molecules n, specifically, $p_{detect} := 1.0/(1.0 + 0.01n)$. This formulation implies that spot detection rate degrades as the number of mRNA molecules in the cell increases, which may correspond to the effect of co-localization and/or image pixelation. We call this model "Missing spots with varying rate".

3. The third model concerns with when $\tilde{y}$ is the output of a spot-detection process contaminated by false positives. We model this false positives by additive Poisson noise, making $\tilde{y} = \tilde{n} + \tilde{e}$ where $\tilde{e} \sim \text{Poisson}(\lambda)$. We choose $\lambda := 10$ for illustration. We call this model "Additive Poisson noise".

4. The fourth model is a composition of two noise models from above, where the mRNA copy number is first distorted by the "lossy" spot-counting process and a signal "enhancement" process that introduces false positive noise. The resulting PDO is the composition of the operators of the two constituent noise models. The order by which these noisy channels are applied is important, as different orders of multiplications of the constituent distortion operators have different results (for linear operators, $A.B \neq B.A$ in general).

5. The fifth model concerns fluorescent intensity integration measurements (such as those used in flow cytometry). We use the model proposed in [12], in which $\tilde{y} = \kappa \cdot \tilde{n}_{RNA} + \eta_{\tilde{n}_{RNA}} + \tilde{\varepsilon}_{BG}$, where $\tilde{\eta}_{\tilde{n}_{RNA}} \sim \mathcal{N}(0, \tilde{n}_{RNA}\sigma^2_{probe})$ models fluorescent heterogenity and $\tilde{\varepsilon}_{BG} \sim \mathcal{N}(\mu_{BG}, \sigma^2_{BG})$ is the background noise. For all simulations, we use $\mu_{BG} = 200$, $\sigma_{BG} = 400$, $\kappa = \sigma^2_{probe} = 25$.

We stress that these are provided as just a few of many possible distortions that cold be modeled with the proposed framework–other, more complex distortion operators are discussed in Supplemental Text S3, including one where a secondary CME is employed to model the uncertain chemical kinetics of the measurement process, and a model to describe cell merging due to image segmentation errors (Supplemental section S3.5).

To demonstrate how measurement distortions affect parameter identification, and why explicit measurement error modeling is necessary, we use the bursting gene expression model to simulate 1,000 datasets of mRNA expressions distorted by the "Missing spots with varying rate" effect above. Each dataset consists of five batches of 1,000 independent single-cell measurements that are collected at five equally-spaced time points $j\Delta t, j = 1, 2, 3, 4, 5$ with $\Delta t = 30$ minute. We considered two methods to fit the telegraph model to these data based on the Maximum Likelihood Estimator (MLE): one in which the likelihood function ignores measurement noise, and one where measurement noise

modeled by the PDO is incorporated into the likelihood function (see Supplemental text S2 for their formulations). If one fails to account for measurement uncertainty, these fits produce strongly biased estimates for the RNA production and degradation rates as seen in Fig. 2 (red). Because this bias is inherent to the measurement technique, it cannot be corrected simply by averaging over more experiments. On the other hand, using the distortion correction method, Fig. 2 (dark green) shows that the inaccurate spot counting procedure can be corrected by explicitly accounting for measurement uncertainty in the modeling phase. Most importantly, we found that the FIM-estimated covariance matrix and the sample covariance matrix of the MLE-PDO fits differ by a relative error of only 0.12 (in Frobenius norm).

Having demonstrated the close proximity between our FIM computation and MLE uncertainty, we now ask whether the sampling period Δt can be tuned to increase information in experimental data modeled with the five measurement distortion models above. Recall that our experimental set-up is such that measurements could be placed at five uniformly-spaced time points $t_j = j\Delta t, \;\; j = 1, 2, 3, 4, 5$, with the sampling period Δt in minutes, and that at each time point we collect an equal number n of single-cell measurements, chosen as $n := 1000$. We find the optimal sampling period Δt for each measurement method, and compare the most informative design that can be achieved for each class. Here, we define 'optimal' in terms of the determinant of the Fisher Information Matrix, the so-called D-optimality criterion, whose inverse estimates the volume of the parameter uncertainty ellipsoid for maximum likelihood estimation [59].

Fig. 3A shows the information volume of the four kinds of noisy measurement described above, in addition to the ideal noise-free smFISH, at different sampling rates. We can observe that every probabilistic distortion to the measurement decreases the information volume (but to different extents), and that each measurement method results in a different optimal sampling rate (see Supplemental Table S1 and Fig. 3B). In Fig. 3C-D, we plot the three-sigma (i.e., 99.7%) confidence ellipsoids of the asymptotic distribution of MLEs projected on the $\log_{10}(k_{ON}) - \log_{10}(k_{OFF})$ plane and the $\log_{10}(k_r) - \log_{10}(\gamma)$ plane.

In addition to the distortion effects discussed above, the extended FIM analysis can also account for much more complex situations. For example, it could be used to compute optimal sampling period when the data is distorted by image segmentation errors of varying degree (Supplemental section S3.5 and Supplemental Figure S6), or to find the optimal probe concentration under noisy probe binding kinetics in smFISH (Supplemental section S3.4 and Supplemental Figure S4).

2.3 Combining different measurement methods the "right" way results in tremendous reduction in parameter uncertainties

While measurement noise tends to incur a loss in information (and increase in uncertainty), as illustrated above, measurement methods with complemen-

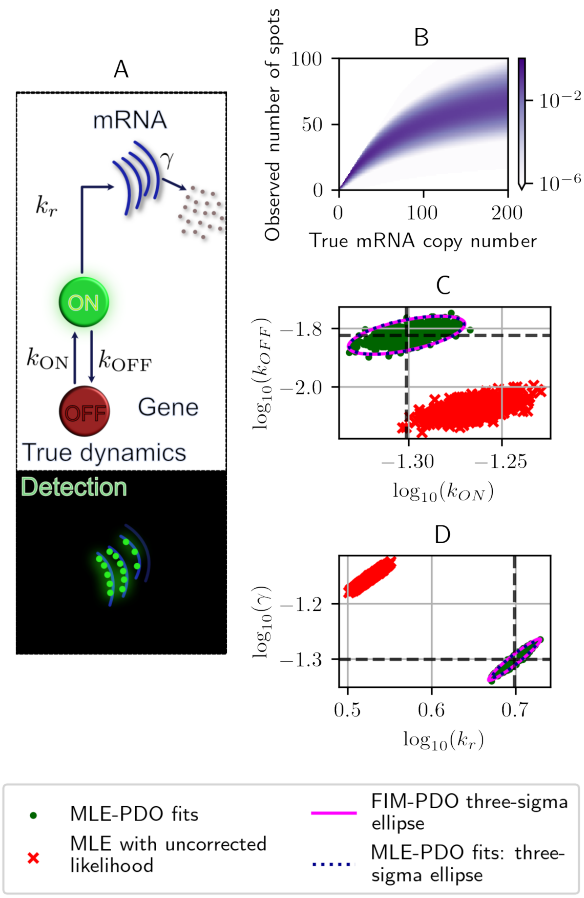

Figure 2: Measurement-distorted counts of single-cell mRNAs produced from telegraphic gene expression dynamics. **(A)**: Schematic of the stochastic telegraph gene expression model. Parameter values: gene activation rate $k_{ON} = 0.015$ events/minute, gene deactivation rate $k_{OFF} = 0.05$ events/minute, transcription rate $k_r = 5$ molecules/minute, degradation rate $\gamma = 0.05$ molecules/minute. **(B)**: A submatrix of the PDO for missing spots with varying rates, restricted to the domain $\{0, \ldots, 200\} \times \{0, \ldots, 100\}$. **(C) and (D)**: Maximum likelihood fits to validate FIM-based uncertainty quantification for observed mRNA distributions under distortion model shown in (B). We simulated $1,000$ datasets and perform maximum likelihood fits to these datasets using either a likelihood function that ignores measurement noise (red, labeled 'MLE with uncorrected likelihood'), or one whose measurement noise is corrected by incorporating the PDO (dark green, labeled 'MLE-PDO fits'). Also displayed are the three-sigma confidence ellipses computed by the FIM-PDO approach or from sample covariance matrix of the corrected MLE fits. Panel (C) shows the results in $\log_{10}(k_{OFF}) - \log_{10}(k_{ON})$ plane, while panel (D) shows them in the $\log_{10}(k_r) - \log_{10}(\gamma)$ plane. The intersection of the horizontal and vertical lines mark the location of the true data-generating parameters.

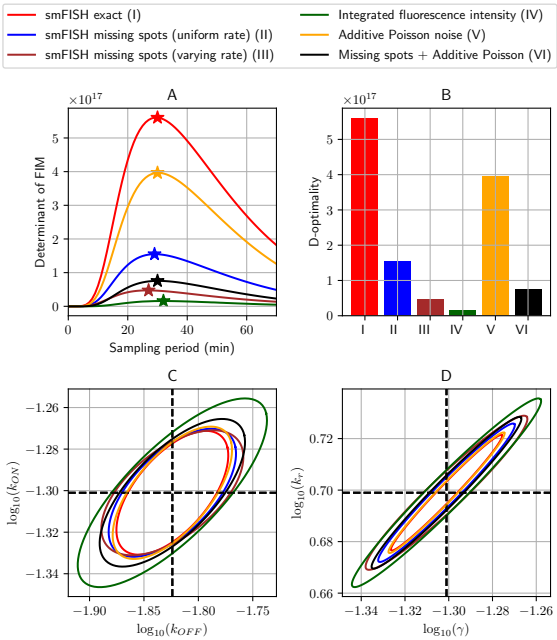

Figure 3: Optimizing sampling rate for bursting gene expression with different types of imprecise measurement. **(A)**: Comparison of D-optimality criteria in single-cell experiments with different types of measurement noise and at different sampling intervals (Δt). In this settings, independent measurements are collected at five equally-spaced time points $k\Delta t,\ k = 1, 2, 3, 4, 5$ with 1000 measurements placed at each timepoint. The $\star$ symbol marks the optimal point of each curve and the bar charts in **(B)** visualizes the relative differences in D-optimal achievable by the measurement methods at their respective optimal sampling rates. **(C) and (D)**: The three-sigma confidence ellipses projected onto the $\log_{10}(k_{ON}) - \log_{10}(k_{OFF})$ and $\log_{10}(k_r) - \log_{10}(\gamma)$ planes. These ellipses are computed from inverting the FIMs of different measurement noise conditions at their optimal sampling rates.

tary distortion characteristics may be combined. Such "wisdom of the *diverse* crowd" can sometimes drastically improve information compared to using solely a single measurement/perturbation method, as recently demonstrated for deterministic models of cell signaling [60]. Our FIM analysis enables us to test this idea in the context of stochastic gene regulation, as we will show in the following example. We revisit the toggle switch example of [44], which is based on experiments originally reported in [61]. The model consists of two species, which we call S_1 and S_2 for simplicity, that repress the production of each other (see Supplemental Table S3 for reactions and propensities). Populations of isogenic cells are assumed to start from the equilibrium distribution of the toggle-switch model with no UV signal. Then a constant pulse of UV is applied to the cells with constant intensity $u := 10$ Joules/m^2 for two hours. We want to find the optimal time (after the initiation of the UV pulse) to collect a batch of cells from which to estimate model parameters. We assume that microscopy or labeling concerns allow for fluorescent probes to label and visualize for each cell either species S_1 or species S_2, but not both simultaneously. In other words, each cell measurement can only report the abundance of a single chemical species within it. Let $\boldsymbol{C}_1$ be the probabilistic distortion operator for S_1-only measurements. We can easily see that

$$\boldsymbol{C}_1(z; (x_1, x_2)) = \delta_{x_1}(z).$$

In other words, the application of $\boldsymbol{C}_1$ on the full joint distribution of both species $\boldsymbol{p}(t)$ at time t will result in the marginal distribution $\boldsymbol{p}_1(t) = \boldsymbol{C}_1\boldsymbol{p}(t)$ of species S_1 at time t. Similarly, we can formulate the probabilistic distortion operator $\boldsymbol{C}_2$ of the S_2-only measurements, which is equivalent to the marginalization of the joint distribution $\boldsymbol{p}(t)$ into the distribution $\boldsymbol{p}_2(t)$ of species S_2. Using the Fisher Information Matrix (FIM) computation introduced above, we can compute the FIM associated with a S_1-only measurement at time t, which we denote by $\boldsymbol{F}_1(t)$ and similarly the single-measurement FIM $\boldsymbol{F}_2(t)$ for a single S_2-only measurement.

We compute the FIM associated with experiment designs where a batch of cells are collected at a time t, in which n_1 single-cell partial observations are made for counting S_1 and n_2 cells are measured for S_2. Assuming that the expression within each *separately measured* cell is independent from the others, the FIM associated with the experiment design outlined above is given by

$$\boldsymbol{F}(n_1, n_2, t) = n_1\boldsymbol{F}_1(t) + n_2\boldsymbol{F}_2(t).$$

Fig. 4 compares the determinant of the FIM associated with using different combination schemes, including extreme schemes where only S_1 ($n_1 = 1000$, $n_2 = 0$) or S_2 ($n_1 = 0$, $n_2 = 1000$) are measured and half-and-half mixtures with $n_1 = n_2 = 500$ and $n_1 = n_2 = 1000$. The computed optimal sampling times associated with these combination schemes and their relative informational values are visualized in Fig. 4D (also see Supplemental Table S4 for precise numerical values). The ideal case where 1000 joint observations can be made is also displayed for comparison. The advantage of combining measurements from different probe designs is profound, where the combined measurement approach

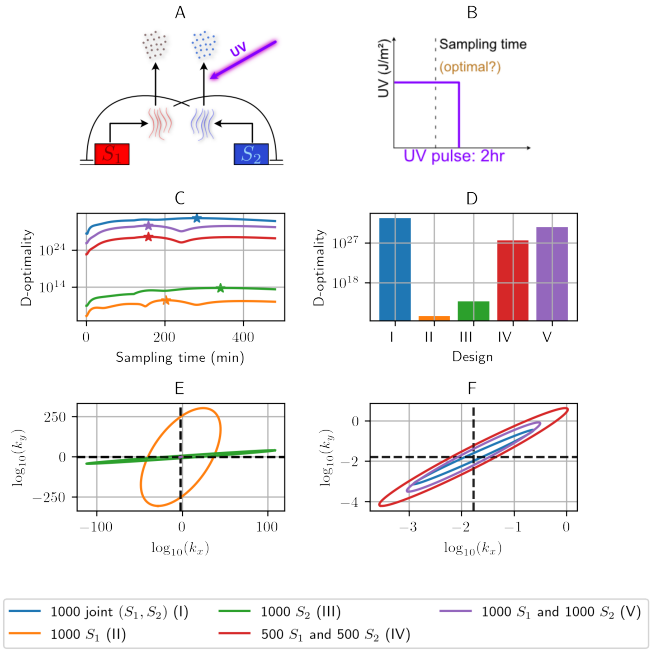

Figure 4: Pure vs mixture of measurement types experiments for estimating toggle switch parameters. **(A)**: Schematic of the toggle switch model (see also Supplemental Table S3 for full list of reactions and propensity functions). **(B)**: Schematic of the experiment design problem. **(C)**: D-optimality across sampling time for experiment designs that use different compositions of partial observations. We compare extreme designs in which all cells are probed for only X (red) or Y (orange) against a mixed design in which the measurements are divided into equal halves. The full-joint observation is also displayed for comparison. **(D)**: D-optimality achievable at the optimal sampling times for the different compositions of measurements. **(E)**: FIM-estimated confidence ellipses on the k_x-k_y plane in log10 space associated with the composition schemes. **(F)** displays a zooming in of panel (E) to reveal the ellipses for the 1000 joint (X, Y) scheme and the combination schemes where both X reporters and Y reporters are used.

results in several orders of magnitude larger determinant for the FIM than using either measurement method alone (Figs. 4A and 4B). The FIM-estimated confidence ellipse associated with the mixture design is on par with that from the full joint observations, while those associated with using only one measurement type are many orders of magnitude larger (Fig. 4C).

3 Discussion

Parameter estimation is a major step in constructing quantitative models for all physical and biological processes. We have demonstrated that experimental noise can potentially distort single-cell measurements of stochastic gene expression (Fig. 2), leading to changes in the estimated parameters that are reflected not only in the magnitudes but also in the directions of their uncertainties (Fig. 3). We have introduced a new computational framework to systematically account for these noise effects, and provide a first step on how this could be integrated into the design of single-cell experiments. Our results indicate that an appropriate statistical analysis coupled with a careful tuning of experimental design variables can meaningfully compensate for even a large degree of measurement noise in the data, and in some conditions such combination of experiment design and model can extract almost as much information using corrupted or partial data as from noise-free or full-dimension measurements (Fig. 4).

Although the FIM is a classical tool for optimal experiment design, which has been used extensively in myriad areas of science and engineering [31, 32, 33, 34, 35, 36], it has not seen widespread adoption in biological investigations. In the context of gene expression modeling, Komorowski et al. [42] proposes a method to approximate the FIM for single-cell experiment data based on the Linear Noise Approximation (LNA). Alternatively, the FIM could be approximated by using moment closure techniques [34, 23]. These approaches work well in the case of high molecular copy numbers, but they break down when applied to systems with low molecular copy numbers [44]. Moreover, it is not clear if or how such approximations can be modified to consider measurement noise and data processing noise that are non-additive, asymmetric, or non-Gaussian as is the case for many biological distortions. To circumvent these issues, our alternative framework directly analyses the probability distributions of the noisy measurements. Explicitly modeling the conditional probability of the observation given the true cell state allows us to express the observation distribution as a linear transformation of the true process distribution that is computable using the finite state projection. This leads to a systematic way to develop composite experimental designs that combine measurements at different fidelity and throughput levels to maximize information given a budgetary constraint.

Adjustments are easily made to the formulations proposed here for the probabilistic distortion operators of smFISH measurements in order to capture more realistic experimental errors. For example, one could construct the PDO based on empirical samples of the joint distribution of the high-fidelity measurements and noisy measurements. Such approach can benefit from progress in recent

probabilistic machine learning methods (e.g. normalizing flows) for modeling the conditional distribution $p(y|x)$ of a noisy label y given a vector x of features. In such case, we only need to estimate the PDO once for each combination of labeling, microscopy, and imaging techniques.

This new approach is versatile and can in principle be applied to any stochastic gene regulatory network, but it does depend on the ability to compute a reasonable approximation to the solution of the chemical master equation (CME) as well as its partial derivatives with respect to model parameters. Fortunately, there are now many relevant stochastic gene expression models for which exact or approximate analytical expressions for the CME solution are available [51, 62, 63, 64, 65, 66]. Furthermore, the FSP and similar approaches have been used successfully to solve the CME for many non-linear and time-inhomogeneous regulatory models for which closed form solutions do not exist [45, 46, 47, 13, 48, 49]. Admittedly, given the complexity of gene regulatory networks in single cells, there are always stochastic gene regulatory models whose direct CME solutions are beyond reach. Nevertheless, continued advancements in computational algorithms [67, 68, 56, 69, 70, 71] are enlarging the set of tractable CME models, which in turns can help accelerate the cycling between data acquisition, model identification, and optimal experiment design in single-cell studies.

4 Methods

4.1 Stochastic modeling of gene expression

The expression dynamics of gene or groups of genes in single cells are often modeled by stochastic reaction networks. For these, the time-varying molecular copy numbers in single cells are treated as a Markov jump process $\{\boldsymbol{X}(t)\}_{t\geq 0}$ whose sample paths $\boldsymbol{x}(t) = (x_1(t), \ldots, x_N(t))$ take place in discrete multi-dimensional space, where $x_i(t)$ is the integer count of species i at time t. Each jump in this process corresponds to the occurrence of one of the reaction events R_k $(k = 1, \ldots, M)$, which brings the cell from the state $\boldsymbol{x}(t_-)$ right before event time t to a new state $\boldsymbol{x}(t_-) + \boldsymbol{\nu}_k$ where $\boldsymbol{\nu}_k$ is the stoichiometry vector corresponding to the k-th reaction. The probabilistic rate at which each reaction occurs is characterized by its propensity (or reaction intensity) function, $\alpha_k(t, \boldsymbol{x}|\boldsymbol{\theta})$. The vector $\boldsymbol{\theta}$ is a d-dimensional vector of *model parameters*. Intuitively, we interpret $\alpha_k(t, \boldsymbol{x}|\boldsymbol{\theta})\Delta t$ as the probability for reaction k to fire during the waiting interval $[t, t+\Delta t)$ for a sufficiently small waiting time Δt. The probability distributions $\boldsymbol{p}_{\boldsymbol{X}}(t; \boldsymbol{\theta})$ of single-cell gene product copy numbers model what is often termed *intrinsic noise* in gene expression. These distributions are the solution of the chemical master equation (CME)

$$\frac{d}{dt}\boldsymbol{p}_{\boldsymbol{X}}(t; \boldsymbol{\theta}) = \boldsymbol{A}(t, \boldsymbol{\theta})\boldsymbol{p}_{\boldsymbol{X}}(t; \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{A}(t; \boldsymbol{\theta})$ is the infinitesimal generator of the Markov process described above (see Supplemental Text S1 for detailed definition of $\mathbf{A}(t, \boldsymbol{\theta})$). Extrinsic noise can be modeled by assuming a probabilistic variation for the model parameters, then integrate (1) over that distribution [43]. However, we focus on intrinsic noise only for the current investigation.

4.2 Modeling distortion of measurements

Let $\boldsymbol{y}(t)$ be the multivariate measurement made on a single cell at time t, such as the discrete number of spots in an smFISH experiment, or the total fluorescence intensity, such as from a flow cytometry experiment. Because of random measurement noise, $\boldsymbol{y}(t)$ is the realization of

a random vector $\boldsymbol{Y}(t)$ that is the result of a random distortion of the true process $\boldsymbol{X}(t)$. The probability mass (density) vector (function) $\boldsymbol{p_Y}(t)$ of the discrete (continuous) observable measurement $\boldsymbol{Y}(t)$ is is related to that of the true copy number distribution via a linear transformation of the form

$$\boldsymbol{p_Y}(t;\boldsymbol{\theta}) = \boldsymbol{C_{X\to Y}}(t;\boldsymbol{\theta})\ \boldsymbol{p_X}(t;\boldsymbol{\theta}). \tag{2}$$

Mathematically, $\boldsymbol{C_{X\to Y}}(t;\boldsymbol{\theta})$ functions as a Markov kernel, and we shall call it the *probabilistic distortion operator* (PDO) to emphasize the context in which it arises. Considered as a matrix whose rows are indexed by all possible observations $\boldsymbol{y}$ and whose columns are indexed by the CME states $\boldsymbol{x}$, it is given entry-wise as

$$\boldsymbol{C_{X\to Y}}(t;\boldsymbol{\theta})(\boldsymbol{y},\boldsymbol{x}) := \Pr(\boldsymbol{Y}(t)=\boldsymbol{y}|\boldsymbol{X}(t)=\boldsymbol{x},\boldsymbol{\theta}), \tag{3}$$

where Pr stands for probability mass if $\boldsymbol{Y}$ is discrete or probability density if $\boldsymbol{Y}$ is continuous. Together, equations (1) and (2) describe single-cell measurements as the time-varying outputs of a linear dynamical system on the space of probability distributions on the lattice of N-dimensional discrete copy number vectors. The output matrix of this dynamical system is the PDO $\boldsymbol{C_{X\to Y}}$.

There are different ways to specify the PDO, depending on the specifics of the measurement method we want to model. In this paper, we demonstrate various examples where the PDO is formulated as probabilistic models that use simple distributions as building blocks (see section 2.2.2 in main text and sections S3.2 in SI), or as deterministic binning/aggregation (section 2.2.3 main text and sections S3.3, S4, S5 in SI). One could even use a secondary CME to model the uncertain chemical kinetics of the measurement process, such as the random time needed to achieve chemical fixation in smFISH experiments or the dropout and amplification of mRNA that occurs during single-cell RNAseq experiments (Supplemental section S3.4), or distribution convolution to describe cell segmentation noise (Supplemental section S3.5). Despite the variety of ways these measurement noise models can be derived, they all lead to the same mathematical object (i.e., the PDO), which allows computation of the FIM associated with that particular noisy single-cell observation approach, and the effects of all PDO can be analyzed using the same computational procedure as we describe next.

4.3 Computation of the Fisher Information Matrix for distorted experimental measurements

In practice, when closed form solutions to the CME do not exist, a forward sensitivity analysis using an extension of the finite state projection algorithm can be used to evaluate the probability distribution $\boldsymbol{p_X}(t;\boldsymbol{\theta})$ and its partial derivatives $\partial_{\theta_j}\boldsymbol{p_X}(t;\boldsymbol{\theta})$ with respect to the kinetic parameters [44]. Using eq. (2), we can transform these into the distribution $\boldsymbol{p_Y}(t;\boldsymbol{\theta})$ of $\boldsymbol{Y}(t)$. Furthermore, the sensitivity indices $\boldsymbol{s}_j^{\boldsymbol{Y}}(t;\boldsymbol{\theta}) = \partial_{\theta_j}\boldsymbol{p_Y}(t;\boldsymbol{\theta})$, of the observable $\boldsymbol{Y}$ are computable by back-propagating the sensitivities of the noise-free measurement distributions through the PDO

$$\frac{\partial}{\partial\theta_j}\boldsymbol{p_Y}(t;\boldsymbol{\theta}) = \boldsymbol{C_{X\to Y}}(t;\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\boldsymbol{p_X}(t;\boldsymbol{\theta}) + \partial\theta_j\boldsymbol{C_{X\to Y}}(t;\boldsymbol{\theta})\boldsymbol{p_X}(t;\boldsymbol{\theta}). \tag{4}$$

Then, the Fisher Information Matrix (FIM) $\boldsymbol{F_{Y(t)}}(\boldsymbol{\theta})$ of the noisy measurements $\boldsymbol{Y}(t)$ at time t is computed by

$$[\boldsymbol{F_{Y(t)}}(\boldsymbol{\theta})]_{i,j} = \sum_{\boldsymbol{y}} \frac{s_i^{\boldsymbol{Y}}(t,\boldsymbol{y};\boldsymbol{\theta})s_j^{\boldsymbol{Y}}(t,\boldsymbol{y};\boldsymbol{\theta})}{p^{\boldsymbol{Y}}(t,\boldsymbol{y};\boldsymbol{\theta})}. \tag{5}$$

Details of the numerical approximation can be found in supplemental text S3.1. In this formulation, we only need to solve the (usually expensive) sensitivity equations derived form the CME once, then apply relatively quick linear algebra operations to find the FIM corresponding to any new PDO for specific microscope, fixation protocol, or probe designs. For improved efficiency, we have implemented a parallel Finite State Projection solver in C++ [72] for solving the sensitivity equations. The computation of the distorted distributions and FIMs are done in Python using the NumPy library.

Acknowledgments

HV and BM were supported by National Institutes of Health (R35 GM124747). We thank members of Munsky group for valuable comments on the manuscript, especially Dr. Linda Forero and Eric Ron for suggesting the investigation of cell segmentation noise.

References

[1] Lewis K (2010) Persister Cells. *Annual Review of Microbiology* 64(1):357–372.

[2] Cao Y, Lei X, Ribeiro RM, Perelson AS, Liang J (2018) Probabilistic control of HIV latency and transactivation by the Tat gene circuit. *Proceedings of the National Academy of Sciences*.

[3] Brock A, Chang H, Huang S (2009) Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics* 10(5):336–342.

[4] Marusyk A, Almendro V, Polyak K (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* 12(5):323–334. Number: 5 Publisher: Nature Publishing Group.

[5] McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences* 94(3):814–819. Publisher: National Academy of Sciences Section: Biological Sciences.

[6] Raj A, van Oudenaarden A (2008) Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* 135(2):216–226.

[7] Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular Decision Making and Biological Noise: From Microbes to Mammals. *Cell* 144(6):910–925.

[8] Munsky B, Li G, Fox ZR, Shepherd DP, Neuert G (2018) Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences* 115(29):7533–7538. Publisher: National Academy of Sciences Section: Biological Sciences.

[9] McQuarrie DA (1967) Stochastic Approach to Chemical Kinetics. 4:89.

[10] Gillespie DT (2007) Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry* 58(1):35–55.

[11] Anderson DF, Kurtz TG (2011) Continuous Time Markov Chain Models for Chemical Reaction Networks in *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, eds. Koeppl H, Setti G, di Bernardo M, Densmore D. (Springer, New York, NY), pp. 3–42.

[12] Munsky B, Trinh B, Khammash M (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology* 5:318.

[13] Neuert G, et al. (2013) Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science* 339(6119):584–587. Publisher: American Association for the Advancement of Science Section: Report.

[14] Milias-Argeitis A, et al. (2011) In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology* 29(12):1114–1116. Number: 12 Publisher: Nature Publishing Group.

[15] Milias-Argeitis A, Rullan M, Aoki SK, Buchmann P, Khammash M (2016) Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth. *Nature Communications* 7(1):12546. Number: 1 Publisher: Nature Publishing Group.

[16] Benzinger D, Khammash M (2018) Pulsatile inputs achieve tunable attenuation of gene expression variability and graded multi-gene regulation. *Nature Communications* 9(1):3521. Number: 1 Publisher: Nature Publishing Group.

[17] Isaacs FJ, Hasty J, Cantor CR, Collins JJ (2003) Prediction and measurement of an autoregulatory genetic module. *Proceedings of the National Academy of Sciences* 100(13):7714–7719. Publisher: National Academy of Sciences Section: Biological Sciences.

[18] Lim HN, van Oudenaarden A (2007) A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nature Genetics* 39(2):269–275. Number: 2 Publisher: Nature Publishing Group.

[19] Lipinski-Kruszka J, Stewart-Ornstein J, Chevalier MW, El-Samad H (2015) Using Dynamic Noise Propagation to Infer Causal Regulatory Relationships in Biochemical Networks. *ACS Synthetic Biology* 4(3):258–264.

[20] Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of Single RNA Transcripts in Situ. *Science* 280(5363):585–590. Publisher: American Association for the Advancement of Science Section: Report.

[21] Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* 4(10):e309. Publisher: Public Library of Science.

[22] Li G, Neuert G (2019) Multiplex RNA single molecule FISH of inducible mRNAs in single yeast cells. *Scientific Data* 6(1):94. Number: 1 Publisher: Nature Publishing Group.

[23] Ruess J, Milias-Argeitis A, Lygeros J (2013) Designing experiments to understand the variability in biochemical reaction networks. *Journal of The Royal Society Interface* 10(88):20130588. Publisher: Royal Society.

[24] Tiberi S, Walsh M, Cavallaro M, Hebenstreit D, Finkenstädt B (2018) Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics* 34(17):i647–i655.

[25] Batish M, Raj A, Tyagi S (2011) Single Molecule Imaging of RNA In Situ in *RNA Detection and Visualization: Methods and Protocols*, Methods in Molecular Biology, ed. Gerst JE. (Humana Press, Totowa, NJ), pp. 3–13.

[26] Pitchiaya S, Heinicke LA, Custer TC, Walter NG (2014) Single Molecule Fluorescence Approaches Shed Light on Intracellular RNAs. *Chemical Reviews* 114(6):3224–3265. Publisher: American Chemical Society.

[27] Raj A, Tyagi S (2010) Chapter 17 - Detection of Individual Endogenous RNA Transcripts In Situ Using Multiple Singly Labeled Probes in *Methods in Enzymology*, Single Molecule Tools: Fluorescence Based Approaches, Part A, ed. Walter NG. (Academic Press) Vol. 472, pp. 365–386.

[28] Kesler B, Li G, Thiemicke A, Venkat R, Neuert G (2019) Automated cell boundary and 3D nuclear segmentation of cells in suspension. *Scientific Reports* 9(1):10237. Number: 1 Publisher: Nature Publishing Group.

[29] Shepherd DP, et al. (2013) Counting Small RNA in Pathogenic Bacteria. *Analytical Chemistry* 85(10):4938–4943. Publisher: American Chemical Society.

[30] Wheat JC, et al. (2020) Single-molecule imaging of transcription dynamics in somatic stem cells. *Nature* 583(7816):431–436.

[31] Pronzato L, Walter E (1985) Robust experiment design via stochastic approximation. *Mathematical Biosciences* 75(1):103–120.

[32] Emery AF, Nenarokomov AV (1998) Optimal experiment design. *Measurement Science and Technology* 9(6):864–876.

[33] Chaudhuri P, Mykland PA (1993) Nonlinear Experiments: Optimal Design and Inference Based on Likelihood. *Journal of the American Statistical Association* 88(422):538–546.

[34] Ruess J, Lygeros J (2013) Identifying stochastic biochemical networks from single-cell population experiments: A comparison of approaches based on the Fisher information in *52nd IEEE Conference on Decision and Control*. pp. 2703–2708.

[35] Chao J, Sally Ward E, Ober RJ (2016) Fisher information theory for parameter estimation in single molecule microscopy: tutorial. *Journal of the Optical Society of America A* 33(7):B36.

[36] Vahid MR, Hanzon B, Ober RJ (2019) Effect of Pixelation on the Parameter Estimation of Single Molecule Trajectories. *arXiv:1912.07182 [eess]*. arXiv: 1912.07182.

[37] Gutenkunst RN, et al. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology* 3(10):1871–1878.

[38] Transtrum MK, MacHta BB, Sethna JP (2010) Why are nonlinear fits to data so challenging? *Physical Review Letters* 104(6).

[39] Faller D, Klingmüller U, Timmer J (2003) Simulation Methods for Optimal Experimental Design in Systems Biology. *SIMULATION* 79(12):717–725.

[40] Gadkar KG, Gunawan R, Doyle FJ (2005) Iterative approach to model identification of biological networks. *BMC Bioinformatics* 6(1):155.

[41] Kreutz C, Timmer J (2009) Systems biology: Experimental design. *The FEBS Journal* 276(4):923–942.

[42] Komorowski M, Costa MJ, Rand DA, Stumpf MPH (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences* 108(21):8645–8650. Publisher: National Academy of Sciences Section: Biological Sciences.

[43] Ruess J, Parise F, Milias-Argeitis A, Khammash M, Lygeros J (2015) Iterative experiment design guides the characterization of a light-inducible gene expression circuit. *Proceedings of the National Academy of Sciences* 112(26):8148–8153.

[44] Fox ZR, Munsky B (2019) The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments. *PLOS Computational Biology* 15(1):e1006365. Publisher: Public Library of Science.

[45] Skupsky R, Burnett JC, Foley JE, Schaffer DV, Arkin AP (2010) HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency. *PLOS Computational Biology* 6(9):e1000952.

[46] Earnest TM, Roberts E, Assaf M, Dahmen K, Luthey-Schulten Z (2013) DNA looping increases the range of bistability in a stochastic model of thelacgenetic switch. *Physical Biology* 10(2):026002.

[47] Lou C, Stanton B, Chen YJ, Munsky B, Voigt CA (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology* 30(11):1137–1142.

[48] Skinner SO, et al. (2016) Single-cell analysis of transcription kinetics across the cell cycle. *eLife* 5:e12175.

[49] Sepúlveda LA, Xu H, Zhang J, Wang M, Golding I (2016) Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*.

[50] Raj A, van Oudenaarden A (2009) Single-Molecule Approaches to Stochastic Gene Expression. *Annual Review of Biophysics* 38(1):255–270.

[51] Peccoud J, Ycart B (1995) Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology* 48(2):222–234.

[52] Suter DM, et al. (2011) Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* 332(6028):472–474.

[53] Sanchez A, Golding I (2013) Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science* 342(6163):1188–1193.

[54] Senecal A, et al. (2014) Transcription Factors Modulate c-Fos Transcriptional Bursts. *Cell Reports* 8(1):75–83.

[55] Larsson AJM, et al. (2019) Genomic encoding of transcriptional burst kinetics. *Nature* 565(7738):251–254.

[56] Gómez-Schiavon M, Chen LF, West AE, Buchler NE (2017) BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. *Genome Biology* 18(1):164.

[57] Fukaya T, Lim B, Levine M (2016) Enhancer Control of Transcriptional Bursting. *Cell* 166(2):358–368.

[58] Breda J, Zavolan M, van Nimwegen E (2021) Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology* 39(8):1008–1016.

[59] Atkinson AC, Donev AN (1992) *Optimum Experimental Designs*, Oxford Statistical Science Series. (Oxford University Press, Oxford, New York).

[60] Jashnsaz H, et al. (2020) Diverse Cell Stimulation Kinetics Identify Predictive Signal Transduction Models. *iScience* 23(10):101565.

[61] Kobayashi H, et al. (2004) Programmable cells: Interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences* 101(22):8414–8419.

[62] Singh A, Bokes P (2012) Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal* 103(5):1087–1096.

[63] Thomas P, Grima R (2015) Approximate probability distributions of the master equation. *Physical Review E* 92(1):012120.

[64] Ham L, Schnoerr D, Brackston RD, Stumpf MPH (2020) Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics* 152(14):144106.

[65] Cao Z, Grima R (2020) Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences* 117(9):4682–4692.

[66] Vastola JJ (2021) Solving the chemical master equation for monomolecular reaction systems and beyond: A Doi-Peliti path integral view. *Journal of Mathematical Biology* 83(5):48.

[67] Kazeev V, Khammash M, Nip M, Schwab C (2014) Direct Solution of the Chemical Master Equation Using Quantized Tensor Trains. *PLOS Computational Biology* 10(3):e1003359.

[68] Cao Y, Terebus A, Liang J (2016) State Space Truncation with Quantified Errors for Accurate Solutions to Discrete Chemical Master Equation. *Bulletin of Mathematical Biology* 78(4):617–661.

[69] Lin YT, Buchler NE (2019) Exact and efficient hybrid Monte Carlo algorithm for accelerated Bayesian inference of gene expression models from snapshots of single-cell transcripts. *The Journal of Chemical Physics* 151(2):024106.

[70] Catanach TA, Vo HD, Munsky B (2020) Bayesian inference of Stochastic Reaction Networks using Multifidelity Sequential Tempered Markov Chain Monte Carlo. *International Journal for Uncertainty Quantification* 10(6).

[71] Gupta A, Schwab C, Khammash M (2021) DeepCME: A deep learning framework for computing solution statistics of the chemical master equation. *PLOS Computational Biology* 17(12):e1009623.

[72] Vo H, Munsky B (2020) A parallel implementation of the finite state projection algorithm for the solution of the chemical master equation. *bioRxiv : the preprint server for biology.*