Analysis and design of single-cell experiments to harvest fluctuation information while rejecting measurement noise

Huy D. Vo[1], Linda Forero[1], Luis Aguilera[1], and Brian Munsky[1,2,3]

[1]Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, CO, USA.
[2]School of Biomedical Engineering, Colorado State University, Fort Collins, CO, USA.
[3]Corresponding Author. Email: munsky@colostate.edu

April 25, 2023

Abstract

Despite continued technological improvements, measurement errors will always reduce or distort the information that any real experiment can provide to quantify cellular dynamics. This problem becomes even more serious in the context of cell signaling studies that are specifically designed to quantify heterogeneity in single-cell gene regulation, 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 (e.g., sampling size, measurement times, or perturbation levels) to ensure that collected data will provide useful insights on signaling or gene expression mechanisms of interest. To address these fundamental single-cell analysis and experiment design challenges, we propose a computational framework that takes explicit consideration of measurement errors to analyze single-cell observations and Fisher Information Matrix-based criteria to decide between experiments. Using simulations and single-cell experiments for a reporter gene controlled by an HIV promoter construct, we demonstrate how our approach can analyze and redesign experiments

to optimally harvest fluctuation information while mitigating the effects of image distortion.

Keywords: Single-cell, fluorescence in-situ hybridization (FISH), stochastic gene expression, experiment design, model inference, measurement noise, Fisher Information Matrix

1 Introduction

Heterogeneity in signaling and gene expression at the single-cell level has wide-ranging biological and clinical consequences, from bacterial persistence ([46, 26, 27, 16]) and viral infections ([78, 91, 68, 8]) to tumor heterogeneity ([7, 52]). Beside genetic and environmental factors, a significant degree of heterogeneity is caused by biochemical noise ([53, 66, 17, 3]). 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 viewed through the lens of deterministic models and bulk measurements ([57]). 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) ([28, 1, 81]). Through this joint experimental and modeling approach, mechanisms of signaling and single-cell gene expression can be explained, predicted ([58, 59, 57]), or even controlled ([56, 55, 5, 22]).

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 optical microscopy (e.g., single-molecule *in situ* hybridization, smFISH), from which data sets are produced in the form of fluorescence intensity histograms ([36, 48, 93, 50]) or molecular count histograms ([20, 64, 47]). These measurements, like any other experimental technique, can be corrupted by errors arising from imprecise detection or data processing methods. 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 ([58, 70, 86]). While smFISH data sets 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 ([4, 62, 65, 39]) 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 ([76]), or the number of hybridization steps ([92]).

The inescapable fact of measurement noise motivates 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 signaling and gene expression networks. Addressing these challenges is non-trivial, as it requires careful consideration of the potentially nonlinear combination of measurement noise and the biological question of interest.

The key contribution of this study is to provide new model-driven experiment analysis and design approaches (Fig 1) that include explicit consideration of probabilistic measurement errors in single-cell observations. The process begins with one or more hypotheses written in the form of stochastic gene regulation models (Fig 1A) with uncertain guesses for parameters or mechanisms. Predictions from these biophysical models are then coupled with empirically determined or physically estimated statistical models, known as *Probabilistic Distortion Operators* (PDOs), that explicitly estimate the effects of different types of measurement errors that could be temporal, discrete, non-symmetric, non-Gaussian, or even the result of their own stochastic process dynamics (Fig 1B). During model inference, the biophysical and measurement distortion models are simultaneously optimized to best explain the data (Fig 1C, right) and to quantify uncertainty in their respective parameters (Fig 1C, left). During experiment design (Fig 1D), sensitivity analysis is used to estimate which combinations of experimental conditions, sampling procedures, or measurement strategies can be expected to provide the most information to constrain the current set of hypotheses, and the procedure is iterated in subsequent rounds of experimentation and model refinement.

To address the specific question of model-driven single-cell experiment evaluation and design (Fig 1D), we adopt the framework of the Fisher Information

Matrix (FIM). The FIM is the basis for a large set of tools for optimal experiment design in a myriad of science and engineering fields ([63, 18, 13, 69, 12, 89]), and it has been employed in the study of identifiability and robustness of deterministic ODEs models in system biology ([31, 87]) as well as for designing optimal bulk measurement experiments ([19, 25, 43]). 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 parameter uncertainties. There have been subsequent work on approximating the FIM with moment closure techniques ([69, 70]), with demonstrable effectiveness on designing optimal optogenetic experiments([71]). Under an assumption of ideal measurements, Fox and Munsky ([23]) recently extended the Finite State Projection (FSP) algorithm ([23]) to compute a version of the FIM that allows for time-varying and non-linear models that result in discrete, asymmetric, and multi-modal single-cell expression distributions. By extending the FSP-based FIM ([23]) to also account for realistic measurement errors, our new approach can help scientists to decide which combinations of experimental conditions and measurement assays are best suited to reduce parameter uncertainties and differentiate between competing hypotheses.

To verify our proposed approaches, we simulate data for a simple bursting gene expression model under many different types of measurement errors, and we show that the FIM correctly estimates the effects that measurement distortions have on parameter estimation (we explore more complicated models and distortions in the Supplemental Text). To demonstrate the practical use of our approaches, we apply them to analyze single-cell data for the bursting and deactivation of a reporter gene controlled by an HIV promoter construct upon application of triptolide (Trp). We show that the iterative use of FSP to fit distorted experimental data, followed by FIM analysis to design subsequent experiments can lead to the efficient identification of a well-constrained model to explain and predict gene expression.

2 Methods

2.1 Stochastic modeling of gene expression

The expression dynamics of genes or groups of genes in single cells are often modeled by stochastic reaction networks ([54, 28, 1]). 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 $\boldsymbol{x}(t_+) = \boldsymbol{x}(t_-) + \boldsymbol{\nu}_k$, 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ 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 ([71]). However, we focus on intrinsic noise for the current investigation.

2.1.1 Computing the likelihood of single-cell data

Consider a data set $\mathbf{D}$ that consists of N_c independent single-cell measurements $(t_1, \boldsymbol{c}_1), \ldots, (t_{N_c}, \boldsymbol{c}_{N_c})$ where $\boldsymbol{c}_i$ is the vector of molecule counts of cell i measured at time t_i. The likelihood function of the biophysical parameters $\boldsymbol{\theta}$ given the data set $\mathbf{D}$ is given by

$$L(\boldsymbol{\theta}|\mathbf{D}) = \prod_{i=1}^{N_c} p_{\boldsymbol{X}}(t_i, \boldsymbol{c}_i, \boldsymbol{\theta})$$

where $p_{\boldsymbol{X}}(t, \boldsymbol{x}, \boldsymbol{\theta})$ is the probability of observing molecular counts $\boldsymbol{x}$ at time t, obtained by solving the CME (1). Taking the logarithm of both sides, we have the log-likelihood function

$$\log L(\boldsymbol{\theta}|\mathbf{D}) = \sum_{i=1}^{N_c} \log p_{\boldsymbol{X}}(t_i, \boldsymbol{c}_i, \boldsymbol{\theta}), \tag{2}$$

which is mathematically and numerically more convenient to work with.

2.1.2 Sensitivity analysis

Taking the partial derivative of both sides of the CME (1) with respect to parameter θ_ℓ ($\ell = 1, \ldots, d$), we get

$$\frac{d}{dt}\left(\frac{\partial \mathbf{p}_{\mathbf{X}}(t,\theta)}{\partial \theta_\ell}\right) = \frac{\partial \mathbf{A}(t,\theta)}{\partial \theta_\ell}\mathbf{p}_{\mathbf{X}}(t,\theta) + \mathbf{A}(t,\theta)\frac{\partial \mathbf{p}_{\mathbf{X}}(t,\theta)}{\partial \theta_\ell}.$$

When the state space of the Markov process is finite, we can collect the equations above for $\ell = 1, \ldots, d$ along with the CME (1) to form a joint system involving the CME solution $\boldsymbol{p}_{\boldsymbol{X}}(t, \theta)$ and its partial derivatives $s_{\boldsymbol{X},\ell}(t, \boldsymbol{\theta}) \equiv \partial p_{\boldsymbol{X}}(t, \boldsymbol{\theta})/\partial \theta_\ell$, given by

$$\frac{d}{dt}\begin{bmatrix} \mathbf{p}_{\boldsymbol{X}}(t,\theta) \\ \mathbf{s}_{\boldsymbol{X},1}(t,\theta) \\ \vdots \\ \mathbf{s}_{\boldsymbol{X},d}(t,\theta) \end{bmatrix} = \begin{bmatrix} \mathbf{A}(t,\theta) & \mathbf{0} & \ldots & \mathbf{0} \\ \partial_{\theta_1}\mathbf{A}(t,\theta) & \mathbf{A}(t,\theta) & \ldots & \mathbf{0} \\ \vdots & & \ddots & \\ \partial_{\theta_d}\mathbf{A}(t,\theta) & \mathbf{0} & \ldots & \mathbf{A}(t,\theta) \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\boldsymbol{X}}(t,\theta) \\ \mathbf{s}_{\boldsymbol{X},1}(t,\theta) \\ \vdots \\ \mathbf{s}_{\boldsymbol{X},d}(t,\theta) \end{bmatrix}.$$

This forward sensitivity system can be solved numerically with any standard ODE solver. When the state space is infinite, a truncation algorithm based on extending the Finite State Projection ([23]) can be applied to approximately solve the forward sensitivity system (see SI Section 1 for more details).

Knowing the sensitivity of the distribution to parameter changes then allows us to compute the sensitivity of the log-likelihood function with respect to biophysical parameters $\theta_1, \ldots, \theta_d$ using the formula

$$\frac{\partial}{\partial \theta_\ell} \log L(\boldsymbol{\theta}|\mathbf{D}) = \sum_{i=1}^{N_c} \frac{s_{\boldsymbol{X},\ell}(t, \boldsymbol{c}_i, \boldsymbol{\theta})}{p_{\boldsymbol{X}}(t, \boldsymbol{c}_i, \boldsymbol{\theta})},$$

where $s_{\boldsymbol{X},\ell}(t, \boldsymbol{x}, \boldsymbol{\theta}) \equiv \partial p_{\boldsymbol{X}}(t, \boldsymbol{x}, \boldsymbol{\theta})/\partial \theta_\ell$ is the sensitivity for the specific molecular

counts $\boldsymbol{x}$ at time t.

2.1.3 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) $\boldsymbol{p}_{\boldsymbol{Y}}(t)$ of the discrete (continuous) observable measurement $\boldsymbol{Y}(t)$ is related to that of the true copy number distribution via a linear transformation of the form

$$\boldsymbol{p}_{\boldsymbol{Y}}(t, \boldsymbol{\theta}) = \boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}(t, \boldsymbol{\theta})\ \boldsymbol{p}_{\boldsymbol{X}}(t, \boldsymbol{\theta}). \tag{3}$$

Mathematically, $\boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}(t, \boldsymbol{\theta})$ 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

$$\boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}(t, \boldsymbol{\theta})(\boldsymbol{y}, \boldsymbol{x}) := \Pr(\boldsymbol{Y}(t) = \boldsymbol{y} | \boldsymbol{X}(t) = \boldsymbol{x}, \boldsymbol{\theta}), \tag{4}$$

where Pr stands for probability mass if $\boldsymbol{Y}$ is discrete or probability density if $\boldsymbol{Y}$ is continuous. Together, equations (1) and (3) 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 $\boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}$. If the observations $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{N_c}$ in the dataset $\boldsymbol{D}$ are assumed to be distorted according to the PDO $\boldsymbol{C}$, then the log-likelihood function (2) is changed into

$$\log L(\boldsymbol{\theta}|\mathbf{D}) = \sum_{i=1}^{N_c} \log p_{\boldsymbol{Y}}(t_i, \boldsymbol{c}_i, \boldsymbol{\theta}) \tag{5}$$

where $p_{\boldsymbol{Y}}(t, \boldsymbol{y}, \boldsymbol{\theta})$ are point-wise probabilities of the distribution $\boldsymbol{p}_{\boldsymbol{Y}}(t, \boldsymbol{\theta})$ defined in (3).

There are many different ways to specify the PDO, depending on the specifics of the measurement method one wishes 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 Results and SI Section S3.2), or as deterministic binning/aggregation (Results and SI Section S3.2). 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 (SI Section S3.3), or distribution convolution to describe cell segmentation noise (SI Section S3.4). 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.

2.1.4 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 $\boldsymbol{p}_{\boldsymbol{X}}(t, \boldsymbol{\theta})$ and its partial derivatives $\partial_{\theta_j} \boldsymbol{p}_{\boldsymbol{X}}(t, \boldsymbol{\theta})$ with respect to the kinetic parameters ([23]. Using eq. (3), we can transform these into the distribution $\boldsymbol{p}_{\boldsymbol{Y}}(t, \boldsymbol{\theta})$ of $\boldsymbol{Y}(t)$. Furthermore, the sensitivity indices $\boldsymbol{s}^j_{\boldsymbol{Y}}(t, \boldsymbol{\theta}) = \partial_{\theta_j} \boldsymbol{p}_{\boldsymbol{Y}}(t, \boldsymbol{\theta})$, of the observable $\boldsymbol{Y}$ are computable by back-propagating the sensitivities of the noise-free measurement distributions through the PDO,

$$\frac{\partial}{\partial \theta_j} \boldsymbol{p}_{\boldsymbol{Y}}(t, \boldsymbol{\theta}) = \boldsymbol{C}_{\boldsymbol{X} \to \boldsymbol{Y}}(t, \boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \boldsymbol{p}_{\boldsymbol{X}}(t, \boldsymbol{\theta}) + \partial \theta_j \boldsymbol{C}_{\boldsymbol{X} \to \boldsymbol{Y}}(t, \boldsymbol{\theta}) \boldsymbol{p}_{\boldsymbol{X}}(t, \boldsymbol{\theta}). \quad (6)$$

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

$$[\boldsymbol{F}_{\boldsymbol{Y}(t)}(\boldsymbol{\theta})]_{i,j} = \sum_{\boldsymbol{y}} \frac{s^i_{\boldsymbol{Y}}(t, \boldsymbol{y}, \boldsymbol{\theta}) s^j_{\boldsymbol{Y}}(t, \boldsymbol{y}, \boldsymbol{\theta})}{p^{\boldsymbol{Y}}(t, \boldsymbol{y}, \boldsymbol{\theta})}. \quad (7)$$

Details of the numerical approximation can be found in SI Section S3.1. In this formulation, we only need to solve the (usually expensive) sensitivity equations derived from 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.

To convert the FIM between parameters defined in linear space to the same parameters in logarithmic space, we apply the transformation:

$$[\boldsymbol{F}_{\boldsymbol{Y}(t)}(\log \boldsymbol{\theta})]_{i,j} = \theta_i \theta_j [\boldsymbol{F}_{\boldsymbol{Y}(t)}(\boldsymbol{\theta})]_{i,j}. \tag{8}$$

2.1.5 Parameter estimation and uncertainty quantification

Models are fit to experimental data in Matlab, (R2021b, [35]) using an iterated combination of the builtin `fminsearch` algorithm (an implementation of the Neldar-Mead simplex method, [44]) to get close to the MLE (i.e., to maximize the likelihood in Eq. 5) followed by a customized version of the Metropolis Hastings (MH) sampling routine, `mhsample` ([14]). All parameter searches were conducted in logarithmic space, and model priors were defined as lognormal with log-means and standard deviations as described in the main text. For the MH sampling proposal function, we used a (symmetric) multivariate Gaussian distribution centered at the current parameter set and with a covariance matrix proportional to the inverse of the Fisher Information matrix (calculated at the MLE parameter set) and scaled by a factor of 0.8, which achieves an approximately 20-50% proposal acceptance rate for all combinations of PDOs and data sets. MH chains were run for 20,000 samples. Convergence was checked by computing the autocorrelation function and verifying that the effective sample size was at least 1000 for every parameter in every MH chain. All data, model construction, FSP analysis, FIM calculation, parameter estimation, and visualization tasks were performed using the Stochastic System Identification Toolkit (SSIT) available at: `https://github.com/MunskyGroup/SSIT` using manuscript specific scripts provided in folder `CommandLine/Vo_et_al_2023`.

2.2 Single-cell Labeling and Imaging

2.2.1 Cell Culture

The experiments presented here were performed on Hela Flp-in H9 cells (H-128). The generation of the H-128 cell line has previously been discussed ([84]). Briefly, Tat expression regulates the MS2X128 cassette-tagged HIV-1 reporter gene in H-128 cells. The HIV-1 reporter consists of the $5'$ and $3'$ long terminal repeats, the polyA sites, the viral promoter, the SD1 and SA7 splice donors, and the Rev-responsive element. Additionally, the MS2 coating protein conjugated

with a green fluorescent protein (MCP-GFP), which binds to MS2 stem loops when transcribed, is expressed persistently by H-128 cells. Cells were cultured in Dulbecco's modified Eagle medium (DMEM, Thermo Fisher Scientific, 11960-044) supplemented with 10% fetal bovine serum (FBS, Atlas Biologicals, F-0050-A), 10 U/mL penicillin/streptomycin (P/S, Thermo Fisher Scientific, 15140122), 1 mM L-glutamine (L-glut, Thermo Fisher Scientific, 25030081), and 150 μg/mL Hygromycin (Gold Biotechnology, H-270-1) in a humidified incubator at 37°C with 5% CO_2.

2.2.2 smiFISH and Microscopy

Single-molecule inexpensive fluorescence *in situ* hybridization (smiFISH) was performed following a protocol previously described ([88, 32]). This technique is known as inexpensive, because the primary probes consist of a region binding the transcript of interest plus a common sequence established as FLAP, which is bound by a complementary FLAP sequence conjugated with a fluorescent dye following a short PCR-cyle (the primary and FLAP-Y-Cy5 probes used in this study were purchased from IDT, see Table 1). To perform smiFISH, H-128 cells were plated on 18 mm cover glasses within a 12-well plate ($\sim 10^5$ cells/well), 24 h before the experiment. Some samples were exposed to 5 μM triptolide (Sigma-Aldrich, 645900) for different incubation periods. Immediately after these drug treatment periods, samples were washed out twice with RNAse free 1XPBS, and fixed in 4% PFA at RT for 10 min, followed by 70% ethanol permeabilization at 4°C for least 1 h. After washing each sample with 150 μL of wash A buffer (Biosearch Technologies, SMF-WA1-60) for 5 min, each cover glass was set on a droplet (cells facing down) consisting of 45 μL hybridization buffer (Biosearch Technologies, SMF-HB1-10) and 1 μL of the duplex smiFISH probes (MS2-transcript-binding probe mix + FLAP-Y-binding region annealed to FLAP-Y-Cy5) in a humidified chamber at 37°C overnight. The following day, samples were placed in a fresh 12-well plate, and the cells (facing up) were incubated twice in wash A buffer at 37°C for 30 min, first alone, and then containing DAPI. Finally, cells were incubated with wash B buffer (Biosearch Technologies, SMF-WB1-20) at RT for 5 min, and then mounted on a 15 μL drop of Vectashield mounting medium (Vector Laboratories, H-1000-10), and sealed with transparent nail polish.

Fluorescent images were acquired with an Olympus IX81 inverted spinning disk confocal (CSU22 head with quad dichroic and additional emission filter

wheel to eliminate spectral crossover) microscope with 60x/1.42 NA oil immersion objective. Confocal z-stacks (0.5 μm step-size, 27 stacks in each channel) were collected. Each field of view was imaged using four high-power diode lasers with rapid (microsecond) switcher (405 nm for DAPI, 488 nm for MS2-MCP-GFP reporter, 561 nm for cytosol marker, and 647 nm for smiFISH MS2-Cy5, exposure time of 100 ms for all except smiFISH channel that was 300 ms) when samples had cytosol marker or three lasers for samples without the marker. The system has differential interference contrast (DIC) optics, built-in correction for spherical aberration for all objectives, and a wide-field Xenon light source. We used an EMCCD camera (iXon Ultra 888, Andor) integrated for image capture using Slide book software (generating 60× images with 160 nm/pixel). The imaging size was set to 624 x 928 pixels^2.

2.2.3 smiFISH image processing

In this study, we employed Python to implement an image processing pipeline comprising three steps: cell segmentation, spot detection, and data management (as previously described in [72]). Nuclear segmentation on the DAPI channel (405nm) was carried out using Cellpose ([82]) with a 70-pixel diameter as an input parameter. Spot quantification for both the MS2-MCP-GFP reporter channel (488 nm) and the smiFISH MS2-Cy5 channel (647 nm) was performed independently, using BIG-FISH software ([34]). The spot quantification procedure employed a voxel-XY of 160 nm, a voxel-Z of 500 nm, and spot radius dimensions of 160 nm and 350 nm in the XY and Z planes, respectively. This process included the following steps: i) the original image was filtered using a Laplacian of Gaussian filter to improve spot detection. ii) Local maxima in the filtered image were then identified. iii) To distinguish between genuine spots and background noise, we implemented both an automated thresholding strategy and a manual approach involving the use of multiple threshold values between 400 and 550. The intensity thresholds (550 for MS2-MCP-GFP and 400 for smiFISH MS2-Cy5) that resulted in the highest number of co-detected spots in both channels were chosen. Additionally, we quantified the average nuclear intensity for each color channel by computing the mean intensity of all pixels within the segmented nuclear region. Data management involved organizing all quantification data into a single dataset, including information on the specific image and cell in which the spots were detected. All image processing codes are available from `https://github.com/MunskyGroup/FISH_Processing`.

3 Results

Many models have been used to capture and predict observations of single-cell heterogeneity in gene expression ([58, 80, 15, 51, 59, 79, 75]). When selecting an experimental assay to parameterize such models, one is faced with several choices, each with its own characteristic measurement errors ([67]). Here, we start by introducing several mathematical forms for probabilistic distortion operators (PDOs) that can quantify these measurement errors. We then use a model and simulated data to show how different measurement errors can affect model identification, and we show how this can be corrected through consideration of the PDO in the estimation process. Next, we show how models and PDOs can be used in the framework of Fisher Information in iterative design of single-cell experiments for efficient identification of predictive models. Finally, we illustrate the practical use of the PDO, model inference, and FIM based experiment design on the experimental investigation of bursting gene expression from a reporter gene controlled by an HIV-1 promoter.

3.1 Distorted single-cell measurements sample a probability distribution that is the image of their true molecular count distribution through a linear operator

Most parameters needed to define single-cell signaling or gene expression models cannot be measured directly or calculated from first principles. Instead, these must be statistically inferred from datasets collected using single-molecule, single-cell experimental methods such as smFISH ([64, 29]), flow cytometry ([48, 50]), or live-cell imaging ([83, 24, 21]). In this work, we focus on the former two experimental approaches in which collected data consists of independent single-cell measurements taken at different times. Here, we consider measurement distortion effects corresponding to probe binding inefficiency and spot detection for smiFISH experiments, and in the supplemental text, we extend this to consider effects of reporter fluorescence intensity variability in flow cytometry experiments (Section S3.1), data binning (Section S3.2), effects of competition with non-specific probe targets (Section S3.3), and effects of segmentation errors (Section S3.4).

We first consider five formulations to define the measurement distortion matrix (cf. eq. (4) in Methods, and illustrated in Fig 2), corresponding to scenarios in which experimental errors arise from either inefficient mRNA detection, additive

false positives, or combinations. Specifically,

1. The first model supposes that $\tilde{y}$ is obtained from a "lossy" spot-counting process applied on images taken in an smFISH experiment. We model $y|\boldsymbol{x} = j$ with a binomial distribution $B(p_{\text{detect}}, j)$, where each spot has a chance $p_{\text{miss}} := 1 - p_{\text{detect}}$ of being ignored by the counting algorithm, resulting in underestimation of the true mRNA copy number. In the context of optical microscopy, such a distortion might result from quantifying spots at a single plane, where p_{detect} might represent the fraction of the imaged section compared to the full volume of the cell, but similar error models have also been proposed in the context of single-cell RNA sequencing ([45, 6]). We call this distortion "Missing Spots". Its PDO, $\mathbf{C}_{\text{MS}}$ is illustrated in Fig 2A for $p_{\text{miss}} := 0.5$ and can be defined:

$$[\mathbf{C}_{MS}]_{i,j} = \binom{j}{i}(1 - p_{\text{miss}})^i p_{\text{miss}}^{j-i}.$$

2. The second model is a simple variation of the first model in which $p_{\text{detect}}(j)$ varies with the number of mRNA molecules j. For example, the specific formulation, $p_{\text{detect}} := 1.0/(1.0 + 0.01j)$ 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 which could cause the under-counting of overlapping spots. We call this model "Missing Spots with Varying Rate", and $\mathbf{C}_{\text{MSVR}}$ is illustrated in Fig 2B and can be expressed:

$$[\mathbf{C}_{\text{MSVR}}]_{i,j} = \binom{j}{i} p_{\text{detect}}(j)^i (1 - p_{\text{detect}}(j))^{j-i}.$$

3. The third model assumes that $\tilde{y}$ is the output of a spot-detection process contaminated by false positives, e.g., due to background fluorescence noise in the image that can appear to be spots. We model these false positives by additive Poisson noise, making $\tilde{y} = j + \tilde{e}$ where $\tilde{e} \sim \text{Poisson}(\lambda)$. We call this model "Poison Noise," and $\mathbf{C}_{\text{PN}}$, which is illustrated in Fig 2C for $\lambda = 10$ and can be expressed:

$$[\mathbf{C}_{\text{PN}}]_{i,j} = \frac{\lambda^{j-i}\exp(\lambda)}{(j-i)!} \text{ for } i \geq j \text{ and otherwise zero}$$

4. The fourth model is a simple extension of the third model in which the

number of detected spots given a true number, j, is a Poisson distribution with a varying mean $\lambda(j)$. Specifically, $\lambda(j) := \max(0, \lambda_0 + \lambda_1 j)$, where λ_0 is Poisson noise, and λ_1 is the Poisson detection rate. We call this model "Poisson Observation", and $\mathbf{C}_{\mathrm{PO}}$ is illustrated in Fig 2D and expressed as:

$$[\mathbf{C}_{\mathrm{PO}}]_{i,j} = \frac{\lambda(j)^i \exp(\lambda(j))}{i!}.$$

5. The fifth model concerns fluorescent intensity integration measurements (such as those used in flow cytometry). We use the model proposed in ([58]), in which $\tilde{y} = \kappa \cdot j + \tilde{\eta}_j + \tilde{\varepsilon}_{BG}$, where $\tilde{\eta}_j \sim \mathcal{N}(0, j\sigma^2_{probe})$ models fluorescent heterogeneity and $\tilde{\varepsilon}_{BG} \sim \mathcal{N}(\mu_{BG}, \sigma^2_{BG})$ is the background noise. The PDO is a hybrid matrix with discrete columns (i.e., its domain are probability vectors over discrete CME states) and continuous rows (its range consists of continuous probability density functions over the range of fluorescence intensities). This PDO, which we label "Integrated Intensity" and is illustrated in Fig 2E for $\mu_{BG} = 200$, $\sigma_{BG} = 400$, $\kappa = \sigma^2_{probe} = 25$, can be expressed:

$$\mathbf{C}_{\mathrm{II}}(y|j) = \frac{1}{\sqrt{2\pi(\sigma^2_{\mathrm{BG}} + j\sigma^2_{\mathrm{probe}})}} \exp\left(\frac{-(y-\kappa j)^2}{2(\sigma^2_{\mathrm{BG}} + j\sigma^2_{\mathrm{probe}})}\right).$$

For intensity measurement with finite resolution, $\mathbf{C}_{\mathrm{II}}$ can be defined over discrete bins (e.g., $(y_0, y_1], (y_1, y_2], \ldots$) by integrating as follows:

$$[\mathbf{C}_{\mathrm{II}}]_{i,j} = \int_{y_{i-1}}^{y_i} \frac{1}{\sqrt{2\pi(\sigma^2_{\mathrm{BG}} + j\sigma^2_{\mathrm{probe}})}} \exp\left(\frac{-(y-\kappa j)^2}{2(\sigma^2_{\mathrm{BG}} + j\sigma^2_{\mathrm{probe}})}\right) dy.$$

We stress that these PDOs are provided as just a few of many possible distortions that could be modeled using 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 (Section S3.3) and another to describe cell merging due to image segmentation errors (Section S3.4).

3.1.1 Measurement noise introduces bias and uncertainty into model identification

To illustrate the impact of measurement distortion on parameter estimation and our use of the PDO formalism to mitigate these effects, we begin with an analysis of the random telegraph model (Fig 3A), one of the simplest, yet most commonly utilized models of bursty gene expression ([61, 64, 83, 73, 74, 45]). 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 degrades with an average rate γ.

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 mRNA expression data where each cell is distorted by the MSVR effect above (PDO reproduced in Fig 3B). Each data set 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 $\Delta t = 30$ minutes. 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 3C,D (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 3C,D (dark green) shows that the inaccurate spot counting procedure can be corrected by explicitly accounting for measurement uncertainty in the modeling phase. Quantitatively (see Table 2), the MLE fits that incorporate noise modeling have lower bias (in terms of relative root-mean-squared errors) for all four parameters compared to the uncorrected fits.

Figures 3C,D also demonstrate that the Fisher Information Matrix computed for the noisy measurement (see eq. (7)) produces a close approximation to the covariance matrix of the MLE (compare magenta and black ellipses). This suggests that the FIM can provide a relatively inexpensive means to quantify the magnitude and direction of parameter uncertainty, a fact that will be helpful in

designing experiments that can reduce this uncertainty, as we will explore next.

3.1.2 Fisher Information Matrix Analysis reveals how optimal experiment design can change in response to different measurement distortions

Having demonstrated the close proximity between our FIM computation and MLE uncertainty for simulated analyses of the bursting gene expression model, we next ask whether the sampling period Δt could be tuned to increase information but using the same number of measurements. Recall that our simulated experimental set-up is such that measurements could be placed at five uniformly-spaced time points $t_j = j\Delta t, \quad j = 1, 2, 3, 4, 5$, 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 distortion (MS, MSVR, PN, PO, and II), 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 ([2]). Figures 4A,B shows the information volume to the five kinds of noisy measurement described above (Fig 2), in addition to the ideal noise-free smiFISH, at different sampling rates. We 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. In Fig 4C-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.

3.2 FIM and PDO Analysis of Experimental Measurements for HIV-1 Promoter Bursting Kinetics

To provide a concrete example for the use of the FIM and PDO in practice, we performed single-cell measurements to quantify the relative measurement distortion between different single-mRNA labeling strategies in HeLa (H-128) cells (see Methods). We expressed a transcription reporter gene with 128 repeats of the MS2 hairpin, and we simultaneously used both MCP-GFP (green labels in Fig 5A) and smiFISH MS2-Cy5 (magenta labels in Fig 5A) to target the MS2 repeats. In each cell, both approaches detect similar patterns for the number

and spatial locations of mRNA within the nuclei (Fig 5B-G). For our particular choice of image processing algorithm ([72, 82, 34]) and intensity threshold for spot detection (see Methods) and a two-pixel (x,y,z) Euclidean distance threshold for co-localization detection, after analyzing 135 cells in steady-state conditions, we found that 46.9% of mRNA spots (15,258 out of 32,526 total) were detected in both channels (e.g., spots denoted by white triangles in Fig 5A). However, many spots (21.3%, 6,913 spots) are detected only using smiFISH (e.g., those denoted with magenta triangles) and 31.8% (10,355 spots) are only detected using the MS2-MCP labels (e.g., green triangles). Due to differences in label chemistry, background fluorescence, and image analysis errors, the quantified distribution of mRNA expression depends heavily upon which assay is utilized (compare Figs 5C-E to 5F-H). For example, we observed that analyses of cells that have fewer than 10 spots in the smiFISH channel frequently result in large numbers of spurious spots when the same cells are analyzed in the MCP-GFP channel (black markers on left limit of Fig 5C and high density at zero in Figs 5D,E). These differences in measurement quantification support the need for consideration of measurement distortions in subsequent analysis.

3.2.1 PDO measurement noise parameters can be calibrated using single-cell experiments with multiple measurement modalities

To demonstrate the parameterization and selection of a PDO for these data, we measured expression using smiFISH or MCP-GFP spot counts as well as with total integrated fluorescence intensity in the FISH or GFP channels. These measurements were collected for 135 cells at t=0 and 62 cells at t=300 min after transcription deactivation by 5μM Trp. We then defined the detected smiFISH spots as the "true" measurements $\{x_1, \ldots, x_{N_c}\}$ and considered the MCP-GFP spot counts, FISH intensity, and GFP intensity values as three different groups of distorted observations $\{y_1, \ldots, y_{N_c}\}$. We then assumed three possible PDO formulations for each measurement modality, including the simple "Poisson Observation" (PO) and "Integrated Intensity" (II) PDOs from above, as well as an extended "Spurious Gaussian" (SG) PDO that is formulated starting with the II model but then extended to allow for a random fraction (f) of cells with a true spot count of 10 or fewer cells to be miscounted as a random observation drawn from a Gaussian ($i \sim \mathcal{N}(\mu, \sigma^2)$) distribution. For all PDOs, we assumed that distorted quantities would be rounded to their nearest non-negative integer value (e.g., negative values would be rounded up to zero). For a each PDO and

its corresponding parameter set $\mathbf{\Lambda}$, we calculate the corresponding PDO, $\mathbf{C}(\mathbf{\Lambda})$, and the log-likelihood to observe $\{y_i\}$ given $\{x_i\}$ is computed as

$$\log L(\{y_i\}|\mathbf{C}, \{x_i\}) = \sum_{i=1}^{N_c} \log[\mathbf{C}(\mathbf{\Lambda})]_{i=y_i, j=x_i}.$$

We then independently maximize this likelihood function for each combination of the three different distorted data sets (MCP-GFP spots, FISH intensity, GFP intensity) and for the three PDO formulations (PO, II, SG). For each data set, we finally select the PDO formulation that results in lowest Bayesian Information Criteria (BIC $\equiv k \log(N_c) - 2 \log L$, where k is the number of parameters in the PDO) for the t=0 and t=300 min data (Fig 6B). We also verified in all cases that the PDO selection also maximized the likelihood of the predictions for held-out data at t=18 min after Tpt treatment (Fig 6B). Upon fitting and selection based on either BIC or cross-validation, we found that MCP-GFP spot count data and GFP intensity distortions were best represented by the "Spurious Gaussian" PDO ($\mathbf{C}_{\mathrm{SG}}$), while the FISH intensity measurements were best represented by the "Integrated Intensity" PDO ($\mathbf{C}_{\mathrm{II}}$).

Figures 5C,F show the contours of the corresponding PDOs for the MCP-GFP spot and FISH intensity data, respectively, and Figs 5D,E,G,H show the PDF and CDF for the "true" smiFISH mRNA count data data at t=0 (blue lines) compared to the distorted data (red lines). From the figures, we find the distortion model does a good job to calibrate between the total mRNA and MCP-GFP- or smiFISH-detected spot counts (compare blue and green lines in Figs 5D,E,G,H and the BIC values for (0,300) min in Fig 5B). We next verified that the PDO remains constant by showing that the same models and same parameters also accurately reproduce the difference in total mRNA and MCP-GFP or smiFISH measurements at a held out time of 18 min after application of 5 μM Trp (see Fig 6, and BIC values for 18 min in Fig 5B).

3.2.2 The PDO allows estimation of predictive bursting gene expression model parameters from distorted data

We next asked if using the estimated PDO while fitting the MCP-GFP spot data or the total FISH or GFP intensity data would enable the identification of appropriate model parameters to predict the smiFISH mRNA counts. Based on previous observations ([21]), we proposed a 3-state bursting gene expression model (Fig 7A) where each of two alleles can occupy one of three states: (S1=OFF)

where no transcription can occur; (S2=Poised) where the promoter is ready to begin transcription; and (S3=Active) where transcripts are produced in rapid bursts of mRNA expression. The model contains six parameters: k_{ON} and k_{OFF} are the promoter transition rates between the OFF and Poised states; ω and k_{Ex} are the burst frequency and burst exit rates; and k_r and γ are the transcription and mRNA degradation rates, respectively. Based on previous observation ([21]) that triptolide (Trp) represses transcription after an average of 5-10 min needed for diffusion of Trp to the promoter and completion of nascent mRNA elongation and processing, we model the Trp response as a complete inactivation of transcription ($\omega \to 0$) that occurs at t=5 min.

To specify a prior guess for the model parameters, we considered published values from ([21]), where we used live-cell imaging of ON and Poised transcription sites to determine the burst frequency $\omega \approx 0.2$ min^{-1} and that k_r and k_{Ex} are too fast to be estimated independently, but are related by a burst size of $k_r/k_{\mathrm{Ex}} = \beta \approx 7.1$ mRNA/burst. Also, based on observations in ([21]) that transcription sites remain in the OFF or Poised/Active states for long periods of 200 minutes or more, we assumed that k_{ON} and k_{OFF} would be too slow to estimate except under much longer experiments. We therefore estimated $k_{\mathrm{ON}} = k_{\mathrm{OFF}} = 10^{-4}$ min^{-1}, but we sought only to estimate the relative rate for k_{OFF}. We estimated a typical mammalian mRNA half-life of 60 min yielding a rate $\gamma = 5.8 \times 10^{-3}$ min^{-1}. With these baseline values as rough estimates, we assumed a log-normal prior distribution with a standard log deviation of one order of magnitude from the literature-based values for parameters $\{\omega, \beta, \gamma\}$ and two orders of magnitude for the more approximate value for $k_{\mathrm{OFF}}\}$. These initial parameter guesses and prior uncertainties are summarized in Table 3(column 3).

Having specified parameter priors, we then took single-cell data for 135 cells at steady state (t=0) and 62 cells at t=300 min after application of 5μM Trp, and we applied the Metropolis-Hasting (MH) algorithm (20,000 samples) to estimate rates and uncertainties for the four free parameters. As above, we assumed that the total spot count analysis provided the "true" spot count, and we considered four estimation problems using either the "true" smiFISH mRNA counts, the MCP-GFP spot count data, the FISH intensity data, or the GFP intensity data, each using the empirically estimated distortion operators from before (e.g., Figs 5C,F,I for the MCP-GFP spots, smiFISH intensity, and GFP intensity data, respectively). Table 3 presents the MLE parameter values after this initial stage, and Fig 7B (left two columns) compare the resulting fits of the model to the data. Figure 7B (right columns) shows scatter plots of parameter

uncertainties for the parameter identification using either the smiFISH data (top row) or with the three distortion measurements using the empirical distortion operator (bottom three rows). From Figs 7B and Table 3, one can see that all fits do a reasonable job to match their intended data, and when fit using the PDO parameters, model predictions based only on the smiFISH and MCP-GFP data match well to the total spot count analysis (Fig 7B, middle column).

3.2.3 The PDO-corrected FIM accurately estimates parameter uncertainty after experimental analysis of the bursting gene expression model

We next used the maximum likelihood models when fit to the 0 and 300 min data to compute the FIM for the corrected MCP-GFP and smiFISH measurements. Because the parameters cover multiple orders of magnitude, we transform the FIM to consider the parameters in log10 space (Eq 8). To adjust the FIM to consider the case where there is a prior on the parameters, we add the inverse of the prior covariance (in logspace) to the calculated FIM: $\mathbf{FIM}_{\text{total}} = \mathbf{FIM} + \Sigma_{\text{prior}}^{-1}$. Given that we only collected a single experimental data set for each time point, it is not possible to directly compare the FIM to the spread of MLE estimates like we could for simulated data. However, in Fig 7B (right columns) for the smiFISH mRNA count and PDO-corrected MCP-GFP spots/intensity or smiFISH intensity data, we can compare the uncertainty predicted by the FIM analysis to the posterior uncertainty of our parameters given our data. This comparison shows that in most cases the FIM does an excellent job to estimate the direction and magnitude of parameter uncertainties (compare overlapping black and magenta ellipses in Figs 7B and see Table 3 for direct comparison of estimated standard deviations). However, it is important to note an exception where the FIM prediction does not match the MH analysis for the largest distortion (GFP intensity measurements). In this case, the FIM predicted variance is much larger than for the other cases (note the change in scales in Fig 7B (bottom right)), but the posterior found by the MH analysis is clearly non-Gaussian. As we will see in the next section, the reason for this failure is likely that the (0,300) min experiment design with this distortion provides insufficient information to identify the model.

3.2.4 PDO-corrected FIM analysis accurately ranks designs for most informative transcription-repression experiment

To demonstrate the practical use of FIM for experiment design, we next asked what design for a Trp-based transcription repression experiment would be best to improve our model of mRNA expression identified in Fig 7. We restricted the set of possible experiments to the previous data set (135 cells at t_1=0 min and 62 cells at t_2=300 min) plus an additional set of 100 cells at a new time t_3 after Trp application, where t_3 could be any time in the allowable set $t_3 \in [0, 6, 12, 18, \ldots, 1200]$ min.

We drew 20 random parameter samples from the previous 20,000-sample MH chains that were estimated using data at t_1=0 and t_2=300 min for each data type. We computed the FIM for each parameter set and for every potential choice for t_3. Figure 8A shows the determinant of the expected covariance of MLE parameters ($\Sigma \approx \mathbf{FIM}^{-1}$, defined in log10 parameter space) versus t_3 assuming direct observation of smiFISH mRNA (left) or distorted observations of MCP-GFP spots, FISH Intensity, or GFP intensity (right). As expected, distortion always increases expected uncertainty (compare Figs 8A(left) to the other columns). We also find that the optimal time for the experiment can be highly dependent on the particular assay. In particular, larger distortions require earlier sampling times to prevent mRNA expression from falling below the noise introduced by the distortion.

Although a complete validation of these designs would require a host of experiments that are beyond the scope of the current investigation, we analyze a new data set of 96 cells taken at $t_3 = 18$ min to compare experiment designs with time combinations of (0,18) min, (0,300) min, and (0,18,300) min. Figure 8B shows FIM predictions of uncertainty using parameters from the original fit to the (0,300) min data (magenta) compared to the uncertainty estimate using parameters fit to the final data set with all three time points (cyan). The original and final FIM are in agreement for all combinations of experiment designs (different groups of magenta and cyan bars) and for direct observations of smiFISH mRNA (left) or for any of the distorted data sets. Moreover, upon using the MH to estimate the posterior parameter distributions, Fig 8C shows that both the original and the final FIM correctly predict the trend of parameter uncertainties for each of the different experiment combinations. For example, the FIM correctly predicts the result that a set of (135,62) cells at (0,300) min is more informative than the larger set of (135,96) cells at (0,18) min if one uses the

undistorted data (far left), but the opposite is true if one considers the distortion due to measurements of GFP intensity (far right). Finally, Fig 8D shows the variance in each parameter's estimation for every measurement type as predicted by the FIM analysis (horizontal axis) and verified using 20,000 MH samples of the posterior (vertical axis). The results show a clear correlation between the predicted and measured uncertainty under either the initial (t=0,300, magenta) or the final (t=0,18,300, cyan) data sets. The analysis also correctly predicts which parameters are well-identified under which measurements. For example, the FIM analysis correctly predicts that the degradation rate γ (triangles) is well identified using just t=0,300 min data for the undistorted data, the MCP-GFP spot data, or the FISH intensity data, but requires the t=18 min data to be identified using the GFP intensity data. In other words, the FIM analysis now explains the failure to identify well-constrained parameters that we observed at the end of the previous section, and correctly suggests that this failure can be substantially ameliorated with an additional measurement of at 62 cells at 18 min.

4 Discussion

Parameter estimation is a major step in constructing quantitative models for all physical or biological processes, and many such models for gene regulation and cell signaling are now being inferred from quantitative single-cell imaging experiments. Such measurements are subject to errors, where different labels or different image processing can yield different measurement values (e.g., see Figs 5 and 6). We have demonstrated that these experimental distortions can be mathematically described using a general class of probability transition kernels that we dub *Probability Distortion Operators* (PDOs, Figs 2,5B,E), and that these PDOs lead to changes in the estimated parameters that are reflected not only in the magnitudes but also in the directions of their uncertainties (Fig 4, 7 and 8). We have introduced a new computational framework to systematically account for these noise effects and provide a first step toward integrating PDOs into the interpretation of single-cell experiments (Figs 5-8). Our results indicate that an appropriate statistical analysis coupled with a careful tuning of experimental design variables can meaningfully compensate for measurement noise in the data. For example, our results indicate that, when used iteratively with small sets of experimental data (e.g., less than a couple hundred cells at only two points in

time), FIM analysis can correctly predict which subsequent experiments are most likely to be informative, and which are unlikely to provide additional insight into model parameters (Fig 8). Insight provided by such an integration of models and experiments could allow for better allocation of experimental resources first by helping to eliminate estimation biases that are due to experimental noise and second by helping to identify specific experimental conditions that are less prone to be impacted by those measurement artifacts.

Although the FIM is a classical tool for optimal experiment design that has been used extensively in myriad areas of science and engineering ([63, 18, 13, 69, 12, 89]), it has not seen widespread adoption in the investigation of biological processes, in part because biological processes are heavily subject to heterogeneities that are not accounted for in traditional FIM analyses. However, there has been some progress to extend these tools to the context of gene expression modeling; for example, ([42]) proposes a method to approximate the FIM for single-cell experiment data based on the Linear Noise Approximation (LNA). Alternatively, the FIM has been approximated by using moment closure techniques ([69, 70]). These approaches work well in the case of high or moderate molecular copy numbers, but they break down when applied to systems with low molecular copy numbers ([23]), and it is not clear how or if 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.

The current investigation provides a few examples and preliminary experimental data to illustrate the broad potential of the FIM and PDO formulations to improve the interpretation and design of single-cell experiments. However, these are limited at present, and there is much that remains to be done to elucidate the full capabilities for these new techniques. For example, not only is additional experimental testing needed to validate the use of FIM-based methods for a broader range of experiment designs and imaging conditions, but the current analyses also need further development (i) to allow for more complex definitions

of PDOs, (ii) to expand the use of FIM and PDO analyses to situations where prior knowledge and models are unavailable or limited, and (iii) to extend the FIM analyses for use in important tasks of model reduction or model selection. A few of these limitations and future directions are discussed as follows.

The current investigation uses high- and low-fidelity calibration experiments to parameterize three different PDOs for different measurement modalities, selects the PDO that minimizes the Bayesian Information Criteria in each case, and confirms that this selection also led to the best prediction of held out data. However, our search over possible PDOs was far from exhaustive, and it is almost certain that more accurate PDOs could be found. Adjustments are easily made so that PDOs can capture many different aspects of experimental error or so that they can be applied to many different types of models. For example, through experimental analysis of mRNA counts using different modalities (e.g., smiFISH and MCP-GFP labeling), we demonstrated how one could construct the PDO based on empirical measurements (Fig 8). Similarly, one could examine different microscopes, different cameras, different laser intensities, different image processing pipelines, or any of a number of permutations to compare and quantify differences between high-fidelity and noisy measurements of signaling or gene expression phenomena. In the current work, we have relied on specific parameterized statistical distributions (e.g., binomial or Poisson as examined in Figs 3,4,5 or data binning or categorization in Section S3.2) or mechanistic distortion models (e.g., integrated fluorescence intensity in Section S3.1, stochastic binding kinetics in Section S3.3, or segmentation errors in Section S3.4) to formulate the PDO. However, one could also envision parameter-free statistical methods based on recent probabilistic machine learning methods (e.g., normalizing flows ([41]) for modeling the conditional distribution $p(y|x)$ of a noisy label y given a vector x of features. In either case, one would need only to calibrate the PDO once for each combination of labeling, microscopy, and imaging techniques and then one could to apply that PDO to many different biological processes, models, parameter sets or experiment designs. For simplicity, we have assumed that calibration data is available and that the PDO is constant in time. However, the formulations of the PDO and FIM (Eqs. 3-7) are sufficiently general such that these requirements can be relaxed. With these relaxations, the FIM could be used either to guide experiment designs to aid the simultaneous identification of both a parameterized PDO and the gene regulation model itself, or provide clear guidance that such distortions lead to degeneracy in the FIM and indicate which non-identifiable parameter combinations (i.e., the null space of the distorted

FIM) are most in need of model reduction.

An important limitation of any model-based experiment design approach is that to make predictions, one must have some prior knowledge about the system under investigation. In the experimental example above, we used insight gathered in a previous biological context (i.e., live-cell analyses of nascent nascent transcription sites from [21]) to guess some model parameters (i.e., the transcription burst sizes and frequencies), and we used general knowledge of the cell line to guess other parameters (e.g., the half-life for mammalian mRNA). For many single-cell optical microscopy investigations, such information is available in advance, due to the fact that one must choose which genes or pathways to investigate before designing smFISH probes or modifying cells and promoters to express the MS2-MCP reporters, and this choice is typically based on experience or previous investigations in the literature (e.g., on analyses of related genes or pathways, in different environmental conditions, or with other exploratory experimental techniques). However, for earlier stage exploratory investigations, where such prior knowledge may not be available, one may need to collect some preliminary data before building models (e.g., collect data for a small handful of time points). In this case, computing the FIM after fitting to the first round of experiments can help to elucidate which parameter sets of the models are well-identified (i.e., vectors in parameter space corresponding to large eigenvalues of the FIM after the *initial* experiment), which can be improved with different experiment designs (i.e., vectors in parameter space that correspond to larger eigenvalues of the FIM for *different* experiments), and which cannot be identified for any experiment (i.e., vectors in parameter space that lie in the null space of the FIM no matter what experiment is considered).

Although we have only considered one model in the main text for the current investigation, the insight provided by the FIM could also be utilized to analyze multiple candidate model structures, an important task that has previously been explored under the assumption that smFISH yields exact measurements of mRNA content [59, 37, 40]. As discussed above, the FIM is useful to identify and prune highly-uncertain parameter combinations or remove unidentifiable model mechanisms. For example, using the distorted FISH intensity measurements at t=(0,300), the FIM analysis clearly shows that the burst frequency ($\log_{10}\omega$) and burst size ($\log_{10}\beta$) are jointly uncertain along the negative diagonal (Fig 7B, rightmost column), meaning that the average total production rate and standard deviation ($\omega\beta \approx 1.04 \pm 0.16 s^{-1}$) could be well constrained, while the individual parameters $\beta = 5.0 \pm 5.8$ and $\omega = 1.1 \pm 2.8$ could not be independently

identified. A similar observation holds using the MCP-GFP spot count data. In both cases, without the constraint of the prior, this uncertainty would extend to the limit where β is much less than one, at which point the proposed 3-state model with parameters $[k_{\mathrm{ON}}, k_{\mathrm{OFF}}, \omega, \beta, \gamma]$ reduces to an equivalent 2-state model with parameters $[k_{\mathrm{ON}}, k_{\mathrm{OFF}}, k_{\mathrm{r}} \approx \omega\beta, \gamma]$. Indeed, Fig S11 shows that this simpler model (Fig S11A) provides nearly (but not quite) as good fits to the all data types when estimated from the t=(0,300) min data (Fig S11B, left columns, compare fit likelihood and BIC values to Fig 7B), but these fits have much less parameter uncertainty (Fig S11B, right columns). Moreover, for the smiFISH mRNA counts, the MCP-GFP spot counts and the MCP-GFP intensity data, the simpler model led to *better* predictions for the held out data at t =18 min (Fig S11B, middle columns compare prediction likelihood values to Fig 7B). Moving forward, if one's goal were to differentiate further between these or other competing hypotheses for model mechanisms, one could use FIM-based experiment design to suggest conditions that promise strong uncertainty reduction for several competing models at once. For example, Fig S12 shows the variance reduction predicted for various possible experiment designs for the reduced model. Comparing FIM analyses of potential experiment designs to the results to the original model in Fig 8, we find that for all distortions, experiments that reduce uncertainty for one model should also be effective for the other. However, the purpose of the current study is only to introduce the FIM+PDO formulation, and a complete analysis of the use of FIM insight for model selection is left for future investigations.

Whether one starts with initial parameter and model structures guesses from previous experimentation or based on a preliminary round of experiments, subsequent model identification is most effective when pursued as an iterative endeavor, requiring evaluation of uncertainty and model-driven experiment design at each stage. For example, in our analysis of the HIV-1 reporter gene, initial data at 0 and 300 min after Tpt revealed that the practical identifiability of parameters depended heavily on which measurement was used. Assuming ideal measurements (i.e., using smiFISH), Fig 8B shows that the model with (n=4) free parameters could be identified to an log10-uncertainty volume of $\det(\mathbf{FIM}^{-1}) = 1.88 \times 10^{-11}$ if we include the prior or $\det(\mathbf{FIM}^{-1}) = 1.91 \times 10^{-11}$ if we do not include the prior. However, using total FISH intensity, the model was much less certain at $\det(\mathbf{FIM}^{-1}) = 6.39 \times 10^{-9}$ with the prior or $\det(\mathbf{FIM}^{-1}) = 1.43 \times 10^{-6}$ without the prior. Since the determinant of the n-parameter FIM scales with the number of cells according to: $\det(\alpha\mathbf{FIM}) = \alpha^n \det(\mathbf{FIM})$, it would take a

factor of $\alpha = (1.43 \times 10^{-6}/1.91 \times 10^{-11})^{1/4} = 16.6$ times as many cells using FISH-intensity to achieve the same accuracy as with the smiFISH mRNA data.[1] Moreover, we also see that the optimal subsequent experiment design also depends heavily on which measurement modality is used. For smiFISH experiments that are assumed to be free from observation noise, the optimal next time point is very late (t=696 min), whereas for the distorted observations, measurements should be taken much earlier (e.g., at t=138 min for MCP-GFP spots, at t=0 min for FISH intensity, and at t=42 min for GFP intensity). Furthermore, in the worse case, choosing the next experiment based on an incorrect assumption for the PDO could lead to waste of experimental efforts – e.g., using the long time as suggested by the smiFISH analysis would be almost entirely worthless if used with one of the other measurements assays. The current study, which is meant only to introduce the FIM and its use for experiment design, has limited availability of experimental data (i.e., one replica with three time points) and only for an artificial HIV-1 reporter construct. A full examination for the use of FIM in iterative single-cell experiment design for endogenous gene regulatory pathways is ongoing and will be described in future publications.

For the model under consideration to fit the HIV-1 promoter with the 128XMS2 stem-loop cassette, computing the CME solution took an average of 1.3s in Matlab on a 2019 MacBook Pro (2.6 GHz 6-Core Intel Core i7), computing the FIM took an average of 40s (39.94s to solve the sensitivity to all parameters and 0.06s to compute the FIM), and running the MH for 20,000 samples took 26,000s, thus the FIM estimates uncertainty roughly 650 times faster than the MH. For use in experiment design, the cost savings provided by the FIM can be much higher. Because one can reuse pre-computed sensitivities, the computational cost is only 0.06s for each new experiment design (e.g., to explore different numbers of cells, different PDOs, or different time point selections). In contrast, using a traditional approach of generating and then fitting simulated data (e.g., to generate MLE scatter plots as shown Figs 3C, 3D or S5), if one liberally assumed that an experiment could be evaluated using only N_s =100 samples and that adequate fits could be achieved using only N_f =100 parameter guesses (usually far more function evaluations are needed), then evaluating each new experiment design would require one CME solution (1.3s) to generate data

[1]We note that total intensity is much easier to compute than spot detection and could in principle be measured at lower microscope resolution or even with flow cytometry. Depending upon available equipment, collecting 16.6 times as many cells could potentially be achieved at a lower overall experimental cost!

and $1.3 \cdot N_s \cdot N_f = 13{,}000$s to fit those data (217,000 times longer than the FIM approach). Given a priori uncertainty in the model, in practice one would need to redo both the FIM analysis (40s + 0.6s per experiment) or the simulation-based analysis (13,000s per experiment) for many different parameter sets or model structures (e.g., we show results from 20 parameter combinations in Figs 8A and S12A), and the savings provided by the FIM becomes even more important.

Finally, although the presented approach is versatile and can in principle be applied to any stochastic gene regulatory network, its practical use depends 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 ([61, 77, 85, 33, 10, 90]). 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 ([80, 15, 51, 59, 79, 75]). For example, SI Section S4 analyzes a model of the nonlinear genetic toggle switch, and the FIM and PDO analysis is used to ask which species should be measured and for how many cells in order to best identify model parameters. As another example, SI Section S5 analyzes a spatial stochastic model with a four-state gene expression model under time varying MAPK activation signal and nucleus to cytoplasmic transport[57]. Admittedly, given the complexity of gene regulatory networks in single cells, there will always be stochastic gene regulatory models whose direct CME solutions are beyond reach. Nevertheless, continued advancements in computational algorithms ([38, 9, 29, 49, 11, 30, 60]) 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.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Conceptualization: BM. Theory and computational modeling: HDV and BM. Performed experiments/collected data: LSFQ. Image processing: LU. Writing: HDV, LSFQ, LUA, and BM. Resources, supervision, and funding acquisition: BM.

Funding

HDV, LUA and BM were supported by National Institutes of Health (R35 GM124747). LSFQ and BM were also supported by the NSF (1941870).

Acknowledgments

We thank members of Munsky group for valuable comments on the manuscript, especially Eric Ron for suggesting the investigation of cell segmentation noise.

Data Availability Statement

Python code for the simulation and FIM analysis of the two-state bursting gene expression model is available at:

https://github.com/voduchuy/NoisyMeasurementFIM

Python code for processing single-cell images is available at:

https://github.com/MunskyGroup/FISH_Processing

Matlab code for analyzing Tryptolide data using the three-state model is available at:

https://github.com/MunskyGroup/SSIT/,

where all data and codes for fitting and visualization of results are available in the folder "CommandLine/Vo_et_al_2023".

References

[1] David F. Anderson and Thomas G. Kurtz. Continuous Time Markov Chain Models for Chemical Reaction Networks. In Heinz Koeppl, Gianluca Setti, Mario di Bernardo, and Douglas Densmore, editors, *Design and Analysis of*

Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology, pages 3–42. Springer, New York, NY, 2011.

[2] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford Statistical Science Series. Oxford University Press, Oxford, New York, November 1992.

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

[4] Mona Batish, Arjun Raj, and Sanjay Tyagi. Single Molecule Imaging of RNA In Situ. In Jeffrey E. Gerst, editor, *RNA Detection and Visualization: Methods and Protocols*, Methods in Molecular Biology, pages 3–13. Humana Press, Totowa, NJ, 2011.

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

[6] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 39(8):1008–1016, August 2021.

[7] Amy Brock, Hannah Chang, and Sui Huang. Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. *Nature Reviews Genetics*, 10(5):336–342, May 2009.

[8] Youfang Cao, Xue Lei, Ruy M. Ribeiro, Alan S. Perelson, and Jie Liang. Probabilistic control of HIV latency and transactivation by the Tat gene circuit. *Proceedings of the National Academy of Sciences*, November 2018.

[9] Youfang Cao, Anna Terebus, and Jie Liang. State Space Truncation with Quantified Errors for Accurate Solutions to Discrete Chemical Master Equation. *Bulletin of Mathematical Biology*, 78(4):617–661, April 2016.

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

[11] Thomas A. Catanach, Huy D. Vo, and Brian Munsky. Bayesian inference of Stochastic Reaction Networks using Multifidelity Sequential Tempered Markov Chain Monte Carlo. *International Journal for Uncertainty Quantification*, 10(6), 2020.

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

[13] Probal Chaudhuri and Per A. Mykland. Nonlinear Experiments: Optimal Design and Inference Based on Likelihood. *Journal of the American Statistical Association*, 88(422):538–546, June 1993.

[14] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[15] Tyler M. Earnest, Elijah Roberts, Michael Assaf, Karin Dahmen, and Zaida Luthey-Schulten. DNA looping increases the range of bistability in a stochastic model of thelacgenetic switch. *Physical Biology*, 10(2):026002, February 2013.

[16] Imane El Meouche, Yik Siu, and Mary J. Dunlop. Stochastic expression of a multiple antibiotic resistance activator confers transient resistance in single cells. *Scientific Reports*, 6(1):19538, May 2016.

[17] Hana El-Samad and Jonathan S. Weissman. Noise rules. *Nature*, 480(7376):188–189, December 2011.

[18] A. F. Emery and Aleksey V. Nenarokomov. Optimal experiment design. *Measurement Science and Technology*, 9(6):864–876, June 1998.

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

[20] Andrea M. Femino, Fredric S. Fay, Kevin Fogarty, and Robert H. Singer. Visualization of Single RNA Transcripts in Situ. *Science*, 280(5363):585–590, April 1998. Publisher: American Association for the Advancement of Science Section: Report.

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

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

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

[24] Takashi Fukaya, Bomyi Lim, and Michael Levine. Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–368, July 2016.

[25] Kapil G. Gadkar, Rudiyanto Gunawan, and Francis J. Doyle. Iterative approach to model identification of biological networks. *BMC Bioinformatics*, 6(1):155, June 2005.

[26] Javier Garcia-Bernardo and Mary J. Dunlop. Tunable Stochastic Pulsing in the Escherichia coli Multiple Antibiotic Resistance Network from Interlinked Positive and Negative Feedback Loops. *PLOS Computational Biology*, 9(9):e1003229, September 2013.

[27] Javier Garcia-Bernardo and Mary J. Dunlop. Noise and Low-Level Dynamics Can Coordinate Multicomponent Bet Hedging Mechanisms. *Biophysical Journal*, 108(1):184–193, January 2015.

[28] Daniel T. Gillespie. Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, May 2007.

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

[30] Ankit Gupta, Christoph Schwab, and Mustafa Khammash. DeepCME: A deep learning framework for computing solution statistics of the chemical master equation. *PLOS Computational Biology*, 17(12):e1009623, December 2021.

[31] Ryan N Gutenkunst, Joshua J Waterfall, Fergal P Casey, Kevin S Brown, Christopher R Myers, and James P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3(10):1871–1878, 2007.

[32] Gal Haimovich and Jeffrey E Gerst. Single-molecule fluorescence in situ hybridization (smfish) for rna detection in adherent animal cells. *Bioprotocol*, 8(21):e3070–e3070, 2018.

[33] Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics*, 152(14):144106, April 2020.

[34] Arthur Imbert, Wei Ouyang, Adham Safieddine, Emeline Coleno, Christophe Zimmer, Edouard Bertrand, Thomas Walter, and Florian Mueller. Fishquant v2: a scalable and modular tool for smfish image analysis. *RNA*, pages rna–079073, 2022.

[35] The MathWorks Inc. Matlab version: 9.13.0 (r2022b). 2022.

[36] Farren J. Isaacs, Jeff Hasty, Charles R. Cantor, and J. J. Collins. Prediction and measurement of an autoregulatory genetic module. *Proceedings of the National Academy of Sciences*, 100(13):7714–7719, June 2003. Publisher: National Academy of Sciences Section: Biological Sciences.

[37] Daniel Kalb, Huy D Vo, Samantha Adikari, Elizabeth Hong-Geller, Brian Munsky, and James Werner. Visualization and modeling of inhibition of il-1β and tnf-α mrna transcription at the single-cell level. *Scientific Reports*, 11(1):1–12, 2021.

[38] Vladimir Kazeev, Mustafa Khammash, Michael Nip, and Christoph Schwab. Direct Solution of the Chemical Master Equation Using Quantized Tensor Trains. *PLOS Computational Biology*, 10(3):e1003359, March 2014.

[39] Benjamin Kesler, Guoliang Li, Alexander Thiemicke, Rohit Venkat, and Gregor Neuert. Automated cell boundary and 3D nuclear segmentation of

cells in suspension. *Scientific Reports*, 9(1):10237, July 2019. Number: 1 Publisher: Nature Publishing Group.

[40] Zeliha Kilic, Max Schweiger, Camille Moyer, Douglas Shepherd, and Steve Pressé. Gene expression model inference from snapshot rna data using bayesian non-parametrics. *Nature Computational Science*, pages 1–10, 2023.

[41] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

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

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

[44] Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

[45] Anton J. M. Larsson, Per Johnsson, Michael Hagemann-Jensen, Leonard Hartmanis, Omid R. Faridani, Björn Reinius, Åsa Segerstolpe, Chloe M. Rivera, Bing Ren, and Rickard Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, January 2019.

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

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

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

[49] Yen Ting Lin and Nicolas E. Buchler. 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, July 2019.

[50] Joanna Lipinski-Kruszka, Jacob Stewart-Ornstein, Michael W. Chevalier, and Hana El-Samad. Using Dynamic Noise Propagation to Infer Causal Regulatory Relationships in Biochemical Networks. *ACS Synthetic Biology*, 4(3):258–264, March 2015.

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

[52] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334, May 2012. Number: 5 Publisher: Nature Publishing Group.

[53] Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, February 1997. Publisher: National Academy of Sciences Section: Biological Sciences.

[54] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of applied probability*, 4(3):413–478, 1967.

[55] Andreas Milias-Argeitis, Marc Rullan, Stephanie K. Aoki, Peter Buchmann, and Mustafa Khammash. Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth. *Nature Communications*, 7(1):12546, August 2016. Number: 1 Publisher: Nature Publishing Group.

[56] Andreas Milias-Argeitis, Sean Summers, Jacob Stewart-Ornstein, Ignacio Zuleta, David Pincus, Hana El-Samad, Mustafa Khammash, and John Lygeros. In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology*, 29(12):1114–1116, December 2011. Number: 12 Publisher: Nature Publishing Group.

[57] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive

models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, July 2018. Publisher: National Academy of Sciences Section: Biological Sciences.

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

[59] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander van Oudenaarden. Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science*, 339(6119):584–587, February 2013. Publisher: American Association for the Advancement of Science Section: Report.

[60] Kaan Öcal, Michael U. Gutmann, Guido Sanguinetti, and Ramon Grima. Inference and Uncertainty Quantification of Stochastic Gene Expression via Synthetic Models. *bioRxiv*, January 2022.

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

[62] Sethuramasundaram Pitchiaya, Laurie A. Heinicke, Thomas C. Custer, and Nils G. Walter. Single Molecule Fluorescence Approaches Shed Light on Intracellular RNAs. *Chemical Reviews*, 114(6):3224–3265, March 2014. Publisher: American Chemical Society.

[63] Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120, July 1985.

[64] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology*, 4(10):e309, September 2006. Publisher: Public Library of Science.

[65] Arjun Raj and Sanjay Tyagi. Chapter 17 - Detection of Individual Endogenous RNA Transcripts In Situ Using Multiple Singly Labeled Probes. In Nils G. Walter, editor, *Methods in Enzymology*, volume 472 of *Single Molecule Tools: Fluorescence Based Approaches, Part A*, pages 365–386. Academic Press, January 2010.

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

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

[68] Igor M. Rouzine, Brandon S. Razooky, and Leor S. Weinberger. Stochastic variability in HIV affects viral eradication. *Proceedings of the National Academy of Sciences*, 111(37):13251–13252, September 2014.

[69] Jakob Ruess and John Lygeros. 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*, pages 2703–2708, December 2013.

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

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

[72] Adham Safieddine, Emeline Coleno, Frederic Lionneton, Abdel-Meneem Traboulsi, Soha Salloum, Charles-Henri Lecellier, Thierry Gostan, Virginie Georget, Cédric Hassen-Khodja, Arthur Imbert, et al. Ht-smfish: a cost-effective and flexible workflow for high-throughput single-molecule rna imaging. *Nature Protocols*, pages 1–31, 2022.

[73] Alvaro Sanchez and Ido Golding. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(6163):1188–1193, December 2013.

[74] Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E. Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription Factors Modulate c-Fos Transcriptional Bursts. *Cell Reports*, 8(1):75–83, July 2014.

[75] Leonardo A. Sepúlveda, Heng Xu, Jing Zhang, Mengyu Wang, and Ido Golding. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*, March 2016.

[76] Douglas P. Shepherd, Nan Li, Sofiya N. Micheva-Viteva, Brian Munsky, Elizabeth Hong-Geller, and James H. Werner. Counting Small RNA in Pathogenic Bacteria. *Analytical Chemistry*, 85(10):4938–4943, May 2013. Publisher: American Chemical Society.

[77] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[78] Abhyudai Singh, Brandon Razooky, Chris D. Cox, Michael L. Simpson, and Leor S. Weinberger. Transcriptional Bursting from the HIV-1 Promoter Is a Significant Source of Stochastic Noise in HIV-1 Gene Expression. *Biophysical Journal*, 98(8):L32–L34, April 2010.

[79] Samuel O Skinner, Heng Xu, Sonal Nagarkar-Jaiswal, Pablo R Freire, Thomas P Zwaka, and Ido Golding. Single-cell analysis of transcription kinetics across the cell cycle. *eLife*, 5:e12175, January 2016.

[80] Ron Skupsky, John C. Burnett, Jonathan E. Foley, David V. Schaffer, and Adam P. Arkin. HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency. *PLOS Computational Biology*, 6(9):e1000952, September 2010.

[81] Jacob Stewart-Ornstein and Hana El-Samad. Chapter 5 - Stochastic Modeling of Cellular Networks. In Anand R. Asthagiri and Adam P. Arkin, editors, *Methods in Cell Biology*, volume 110 of *Computational Methods in Cell Biology*, pages 111–137. Academic Press, January 2012.

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

[83] David M. Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science*, 332(6028):472–474, April 2011.

[84] Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, et al. A single-molecule view of transcription reveals convoys of rna polymerases and multi-scale bursting. *Nature communications*, 7(1):1–14, 2016.

[85] Philipp Thomas and Ramon Grima. Approximate probability distributions of the master equation. *Physical Review E*, 92(1):012120, July 2015.

[86] Simone Tiberi, Mark Walsh, Massimo Cavallaro, Daniel Hebenstreit, and Bärbel Finkenstädt. Bayesian inference on stochastic gene transcription from flow cytometry data. *Bioinformatics*, 34(17):i647–i655, September 2018.

[87] Mark K. Transtrum, Benjamin B. MacHta, and James P. Sethna. Why are nonlinear fits to data so challenging? *Physical Review Letters*, 104(6), 2010.

[88] Nikolay Tsanov, Aubin Samacoits, Racha Chouaib, Abdel-Meneem Traboulsi, Thierry Gostan, Christian Weber, Christophe Zimmer, Kazem Zibara, Thomas Walter, Marion Peter, et al. smifish and fish-quant–a flexible single rna detection approach with super-resolution capability. *Nucleic acids research*, 44(22):e165–e165, 2016.

[89] Milad R. Vahid, Bernard Hanzon, and Raimund J. Ober. Effect of Pixelation on the Parameter Estimation of Single Molecule Trajectories. *arXiv:1912.07182 [eess]*, December 2019. arXiv: 1912.07182.

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

[91] Ariel D. Weinberger and Leor S. Weinberger. Stochastic Fate Selection in HIV-Infected Patients. *Cell*, 155(3):497–499, October 2013.

[92] Justin C. Wheat, Yehonatan Sella, Michael Willcockson, Arthur I. Skoultchi, Aviv Bergman, Robert H. Singer, and Ulrich Steidl. Single-molecule imaging of transcription dynamics in somatic stem cells. *Nature*, 583(7816):431–436, July 2020.

[93] Ignacio A Zuleta, Andrés Aranda-Díaz, Hao Li, and Hana El-Samad. Dynamic characterization of growth and gene expression using high-throughput automated flow cytometry. *Nature Methods*, 11(4):443–448, April 2014.

Tables

Table 1: **smiFISH probe sequences.** Each primary probe has an added common FLAP-Y binding sequence (`TTACACTCGGACCTCGTCGACATGCATT`)

Probe name	Sequence	Type
MS2-P1	`ATCCGTTCAAAGGCCTATTGGTCCTTTGC`	Primary
MS2-P2	`ATCCTCATGTACTAGCTTCCGAGTAATCT`	Primary
MS2-P3	`AGCTTGGGTTATTACTCCAAGATCACCGC`	Primary
MS2-P4	`TCCTGATAGGCTGTACTCATGCCTAC`	Primary
MS2-P5	`GGGTAATCATTCTAGTGATATGATTCTGTGCC`	Primary
MS2-P6	`GATCATACCGTATTCGTGTATGATTACATGGG`	Primary
MS2-P7	`TAATTGTGCGGTCGCTGACTGATACTTCTA`	Primary
MS2-P8	`TATTTCTCTCTGATACGCTGCGTACTCG`	Primary
MS2-P9	`ACCCTAATGGTGTTTACAAATGGTGGTAGTCC`	Primary
MS2-P10	`TGTTATTCTAATCCGTCACTATTGTTGACGGG`	Primary
MS2-P11	`AAGCCTTACTGATGCTTCCGGTCCATT`	Primary
MS2-P12	`ATCCTCATGTACTAGAGGCTCGGTACTC`	Primary
FLAP Y-Cy5	`AATGCATGTCGACGAGGTCCGAGTGTAA`	Secondary

Parameter	True	MLE fits Mean (Std)		MLE fits rel. RMSE	
		Corrected	Uncorrected	Corrected	Uncorrected
$\log_{10}(k_{OFF})$	-1.30	-1.30 (1.02×10^{-2})	-1.27 (1.15×10^{-02})	7.87×10^{-3}	2.83×10^{-2}
$\log_{10}(k_{ON})$	-1.82	-1.82 (2.16×10^{-2})	-2.07 (2.30×10^{-2})	1.18×10^{-2}	1.37×10^{-1}
$\log_{10}(k_r)$	0.699	0.698 (8.73×10^{-3})	0.526 (9.19×10^{-3})	1.25×10^{-2}	2.48×10^{-1}
$\log_{10}(\gamma)$	-1.30	-1.30 (1.06×10^{-2})	-1.15 (10^{-2})	8.18×10^{-3}	1.14×10^{-1}

Table 2: Performance of the maximum likelihood estimator (MLE) for estimating bursting transcription kinetic parameters. The third and fourth columns compare the mean and standard deviation of fits with and without PDO correction (labeled Corrected and Uncorrected, respectively). The final two columns compare the relative root-mean-squared errors (RMSEs) of these fits. For a quantity of interest q and its n estimated values $\hat{q}_i$, $i = 1, \ldots, n$, we define the relative RMSE as $\frac{\sqrt{\sum_{i=1}^{n}(\hat{q}_i - q)^2}}{|q|}$.

Table 3: **Parameter priors, MLE estimates, Uncertainties upon Initial Fit**. Initial estimates after fitting to data at t =(0,300) min. MH results are from a chain of 20,000 samples. All parameter values and standard deviations values are shown in $\log_{10}$.

Parameter	Quantity	Prior	smiFISH	MCP-GFP Spots	FISH Intens	GFP Intens
k_{OFF} (s^{-1})	$\log_{10}(\lambda)$	-4.0	-4.215	-3.852	-3.840	-4.619
	$\sigma_{p/MH}$	2.0	0.091	0.083	0.069	1.160
	σ_{FIM}	2.0	0.071	0.060	0.070	0.308
ω (s^{-1})	$\log_{10}(\lambda)$	-0.70	-2.173	-0.357	-0.505	-1.205
	$\sigma_{p/MH}$	1.0	0.117	0.462	0.559	0.752
	σ_{FIM}	1.0	0.077	0.356	0.475	0.335
β (s^{-1})	$\log_{10}(\lambda)$	0.85	2.110	0.710	0.524	1.262
	$\sigma_{p/MH}$	1.0	0.104	0.451	0.556	0.428
	σ_{FIM}	1.0	0.069	0.379	0.438	0.307
γ (s^{-1})	$\log_{10}(\lambda)$	-2.2	-2.191	-1.965	-2.283	-2.036
	$\sigma_{p/MH}$	1.0	0.031	0.043	0.053	0.695
	σ_{FIM}	1.0	0.034	0.028	0.053	0.078

Table 4: **Parameter priors, MLE estimates, Uncertainties upon Final Fit**. Final estimates after fitting to data at t =(0,18,300) min. MH results are from a chain of 20,000 samples. All parameter values and standard deviations values are shown in $\log_{10}$.

Parameter	Quantity	Prior	smiFISH	MCP-GFP Spots	FISH Intens	GFP Intens
k_{OFF} (s^{-1})	$\log_{10}(\lambda)$	-4.0	-4.245	-3.734	-3.711	-4.761
	$\sigma_{p/MH}$	2.0	0.066	0.050	0.063	0.962
	σ_{FIM}	2.0	0.066	0.045	0.060	0.506
ω (s^{-1})	$\log_{10}(\lambda)$	-0.70	-2.318	-0.105	-0.891	-0.591
	$\sigma_{p/MH}$	1.0	0.066	0.478	0.533	0.396
	σ_{FIM}	1.0	0.071	0.368	0.373	0.247
β (s^{-1})	$\log_{10}(\lambda)$	0.85	2.194	0.453	0.834	1.467
	$\sigma_{p/MH}$	1.0	0.059	0.476	0.522	0.317
	σ_{FIM}	1.0	0.057	0.427	0.454	0.227
γ (s^{-1})	$\log_{10}(\lambda)$	-2.2	-2.200	-1.980	-2.322	-1.186
	$\sigma_{p/MH}$	1.0	0.035	0.034	0.061	0.101
	σ_{FIM}	1.0	0.036	0.028	0.062	0.072

Figure captions

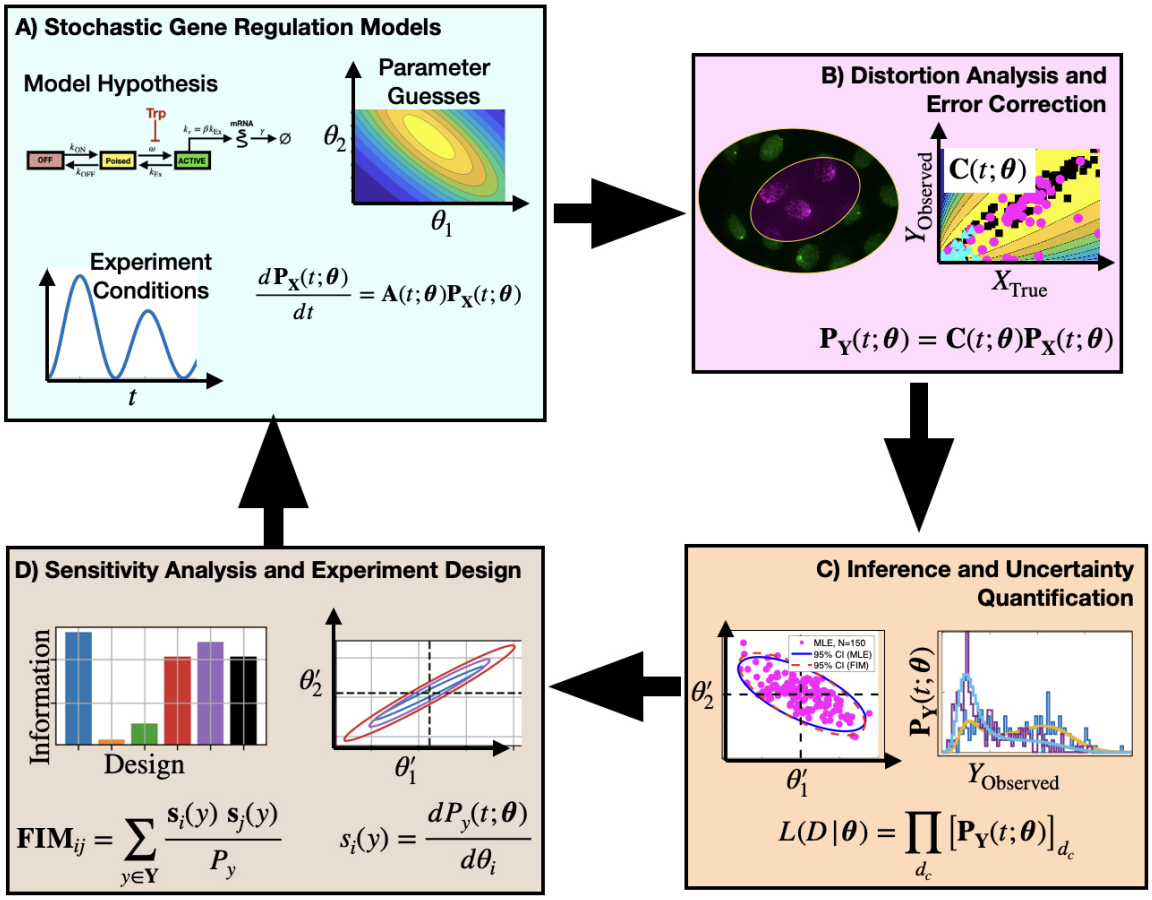


Figure 1: **Proposed approach to analyze and design single-cell experiments under measurement distortion. Box (A):** One or more mechanistic model hypotheses are proposed to describe gene regulation processes and are associated with prior parameter guesses. These models are combined with experiment design considerations (e.g., environmental, sampling, and measurement conditions) and the chemical master equation framework is used to predict statistics of single-cell gene expression. **Box B**: Empirical data and physical models are used to define a probabilistic distortion operator **C** to estimate how measurement errors (e.g., labeling inefficiencies, resolution limitations, or image processing errors) affect observations. **Box C**: Searches over model and PDO parameter space are conducted to identify model and parameter combinations that maximize the likelihood of the observed (i.e., distorted) data. **Box D**: Sensitivity analysis and the Fisher Information Matrix are computed and used to analyze the information content for different mechanistic models, distortion effects, or experiment designs. By estimating parameter uncertainties in the context of different forms of measurement distortion, subsequent experiment designs can help to alleviate measurement noise distortion effects (return to **Boxes A**, **B**, and **C**).

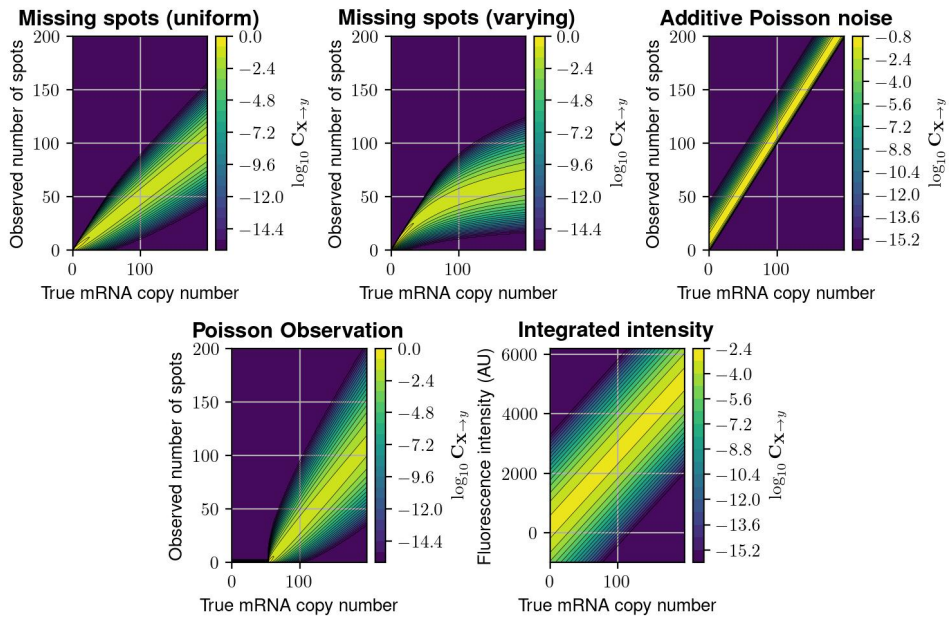


Figure 2: **Probabilistic distortion operators**. **(A)** The "Missing Spots" (MS) distortion model, where spots can randomly go missing. **(B)** The "Missing Spots with Variable Rate" (MSVR) distortion, where the probability of missing a spot increases with spot density. **(C)** The "Poisson Noise" (PN) model, where false positive spots are added to the counted number of spots. **(D)** A "Poisson Observation" (PO) model, where the detected spots follow a Poisson distribution with mean proportional to the true number. **(E)** The "Integrated Intensity" (II) model, where only a perturbed version of the total fluorescence is recorded per cell. In each heat map, the color at point $(\tilde{x}, \tilde{y})$ is the conditional probability mass/density of the measurement y having value $\tilde{y}$ given that the true copy number x has value $\tilde{x}$.

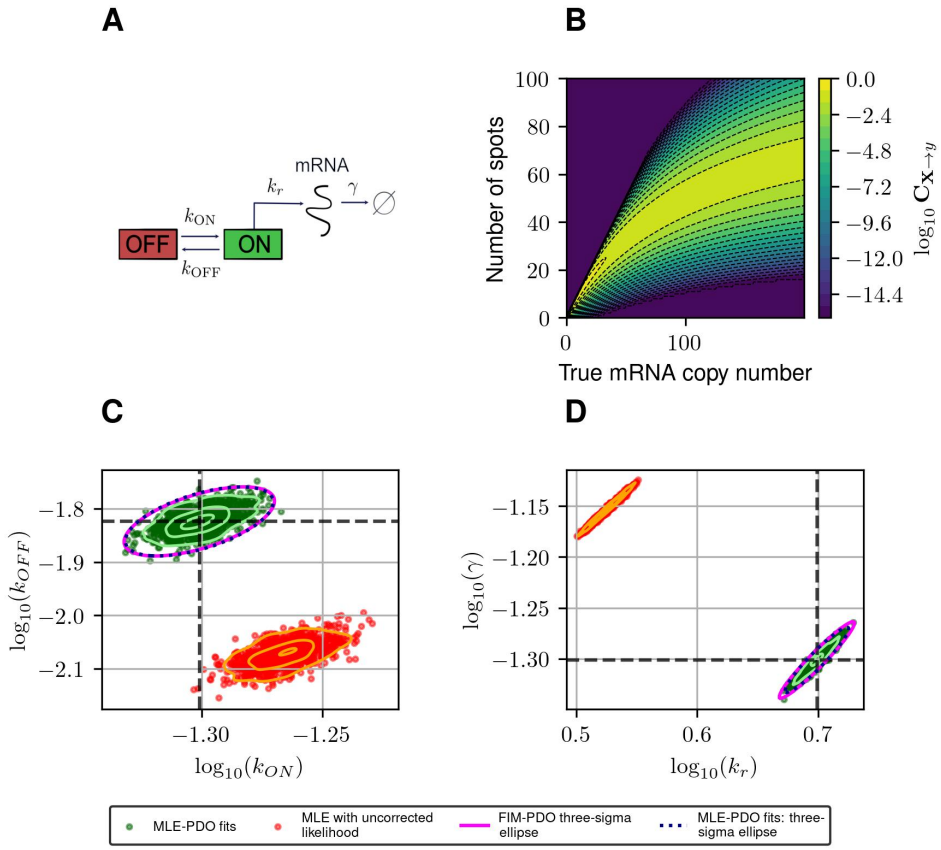


Figure 3: **Estimating and correcting for how measurement distortion affects model identification.** **(A)**: Schematic of the random telegraph gene expression model. Parameter values: gene activation rate $k_{ON} = 0.015$ events/minute, gene deactivation rate $k_{OFF} = 0.05$ events/minute, transcription 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'). The estimated density contours (delineating 10, 50, and 90 -percentile regions) of the fits are superimposed in light shades. 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 thick horizontal and vertical lines marks the location of the true data-generating parameters. See Table 2 for a quantitative comparison between the uncorrected MLE and MLE-PDO.

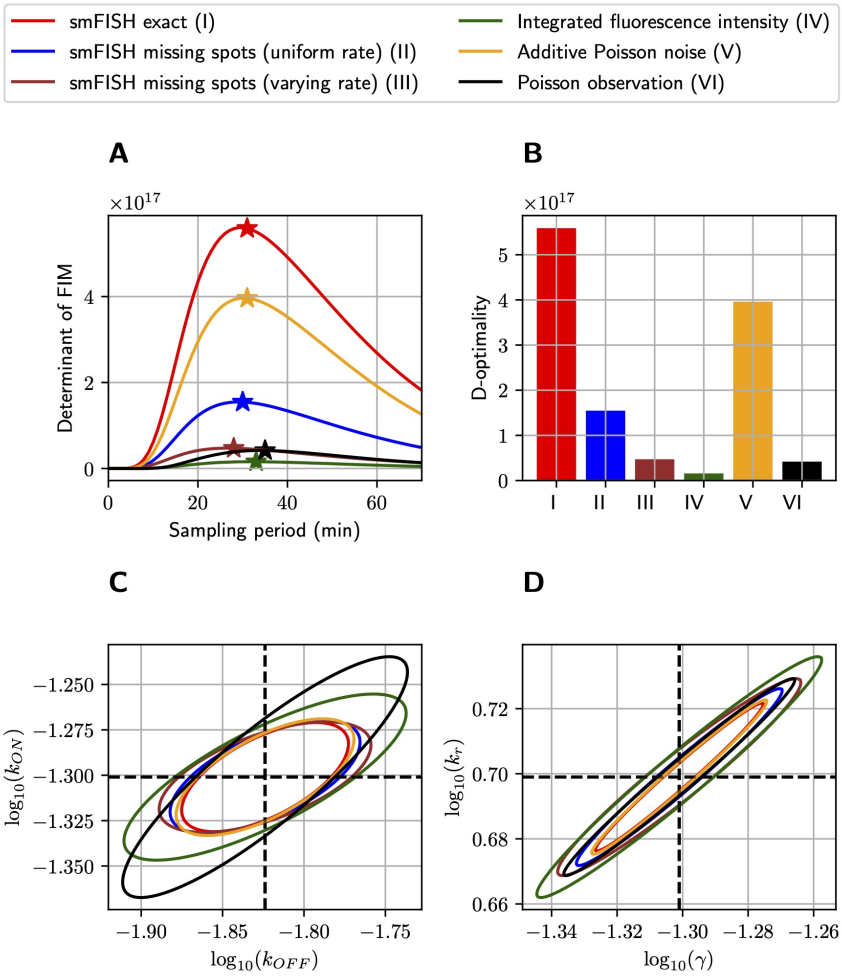
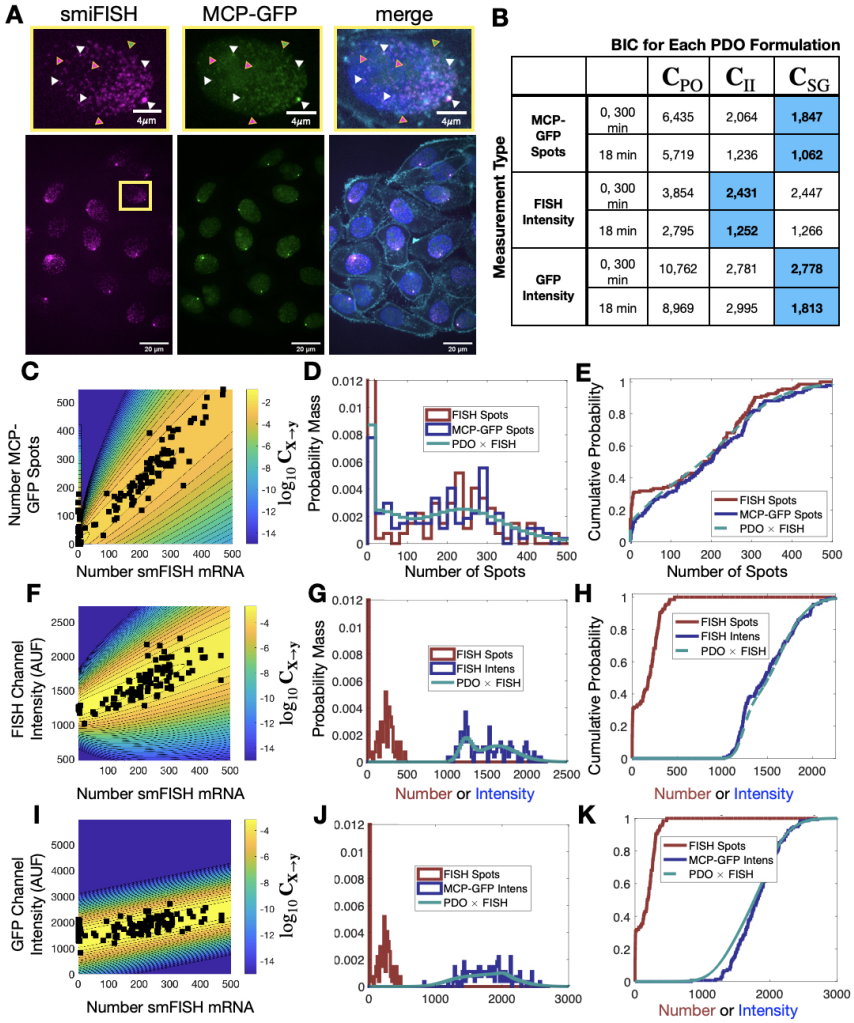


Figure 4: **Optimizing experiment sampling rate under different measurement distortion effects. (A)**: Comparison of D-optimality criteria in single-cell experiments with different types of measurement noise and at different sampling intervals (Δt). In this settings, independent measurements are collected at five equally-spaced time points $k\Delta t,\ k = 1, 2, 3, 4, 5$ with 1000 measurements placed at each time-point. 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 by inverting the FIMs of different measurement noise conditions at their optimal sampling rates. All analyses use the model and parameters from Fig 3.



Measurement Type		C_{PO}	C_{II}	C_{SG}
MCP-GFP Spots	0, 300 min	6,435	2,064	**1,847**
	18 min	5,719	1,236	**1,062**
FISH Intensity	0, 300 min	3,854	**2,431**	2,447
	18 min	2,795	**1,252**	1,266
GFP Intensity	0, 300 min	10,762	2,781	**2,778**
	18 min	8,969	2,995	**1,813**

Figure 5: **Effect of labeling strategy on single-cell mRNA expression quantification.** **A**) Example image of a single-cell population expressing a reporter gene controlled by an HIV-1 promoter and containing 128XMS2 stem-loop cassette, in which the mRNA was simultaneously measured using MCP-GFP labeling (green) and with smiFISH probes against MS2-Cy5 (magenta). A higher resolution image of the indicated single nucleus is shown at the top; merged image on the right includes DAPI and MemBrite cell surface stain 543/560 to denote the cell nucleus and plasma membrane, respectively. Triangles denote example mRNA that are detected in both channels (white, 34.8%), only in the MCP-GFP channel (green, 37.5%) and only in smiFISH (MS2-Cy5) channel (magenta, 27.7%). **B**) BIC of different combinations of PDO (columns) and measurement type (rows) given an assumed 'true' measurement of smiFISH mRNA. In all cases, PDO parameters are chosen to maximize likelihood for $t =$ 0 and 300 min data, and $t =$18 min is predicted without changing parameters. Blue shading denotes PDO selection is identical if based on BIC for (0,300) min data or prediction of 18 min data. **C**) Scatter plot of the spot count using MCP-GFP versus using smiFISH for data collected at $t =$0 min (black squares). Shading and contour lines denote the levels of the PDO ($\log_{10} \boldsymbol{C}$) determined empirically from the data. **D,E**) Empirical probability mass (D, bin size = 20) and cumulative distributions (E, no binning) for smiFISH spot count (red) and MCP-GFP spot count (blue). Predicted MCP-GFP spot distributions using the smiFISH spot measurements and the estimated PDO are shown in green. **F-H**) Same as (C-E) but for the total integrated intensity measurement of the smiFISH channel. **I-K**) Same as (C-E) but for the total integrated intensity measurement of the GFP channel.

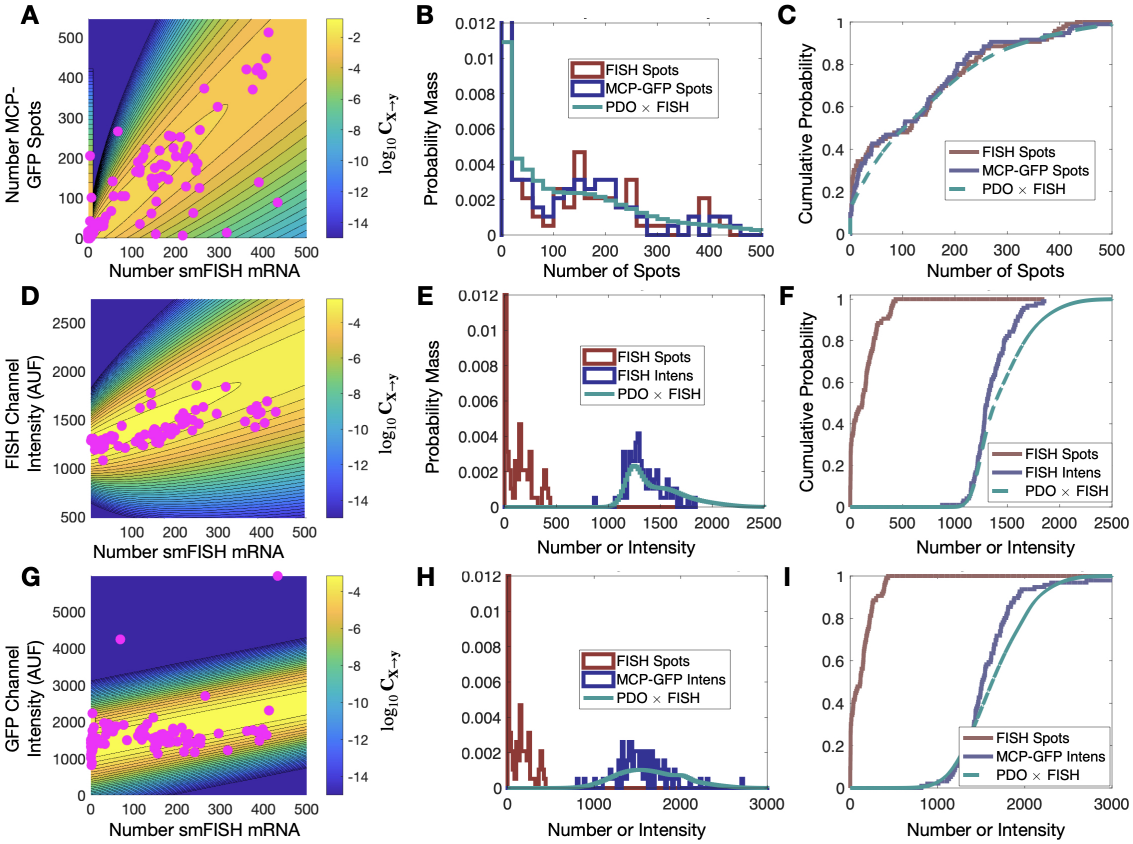


Figure 6: **Validation of PDO on held out data. A**) Scatter plot of smiFISH spot counts and MCP-GFP spot counts for data collected at t =18 min and contour of PDO determined from data at $t = (0, 300)$ min. **B,C**) PDF (bin size = 20) and CDF for smiFISH mRNA detection (red) or with MCP-GFP (blue) and MCP-GFP-based PDO-prediction of total mRNA (green). **D-F**) Same format as (B-D) but for measurements of the total smiFISH fluorescence intensity per cell.**G-I**) Same format as (B-D) but for measurements of the total GFP fluorescence intensity per cell.

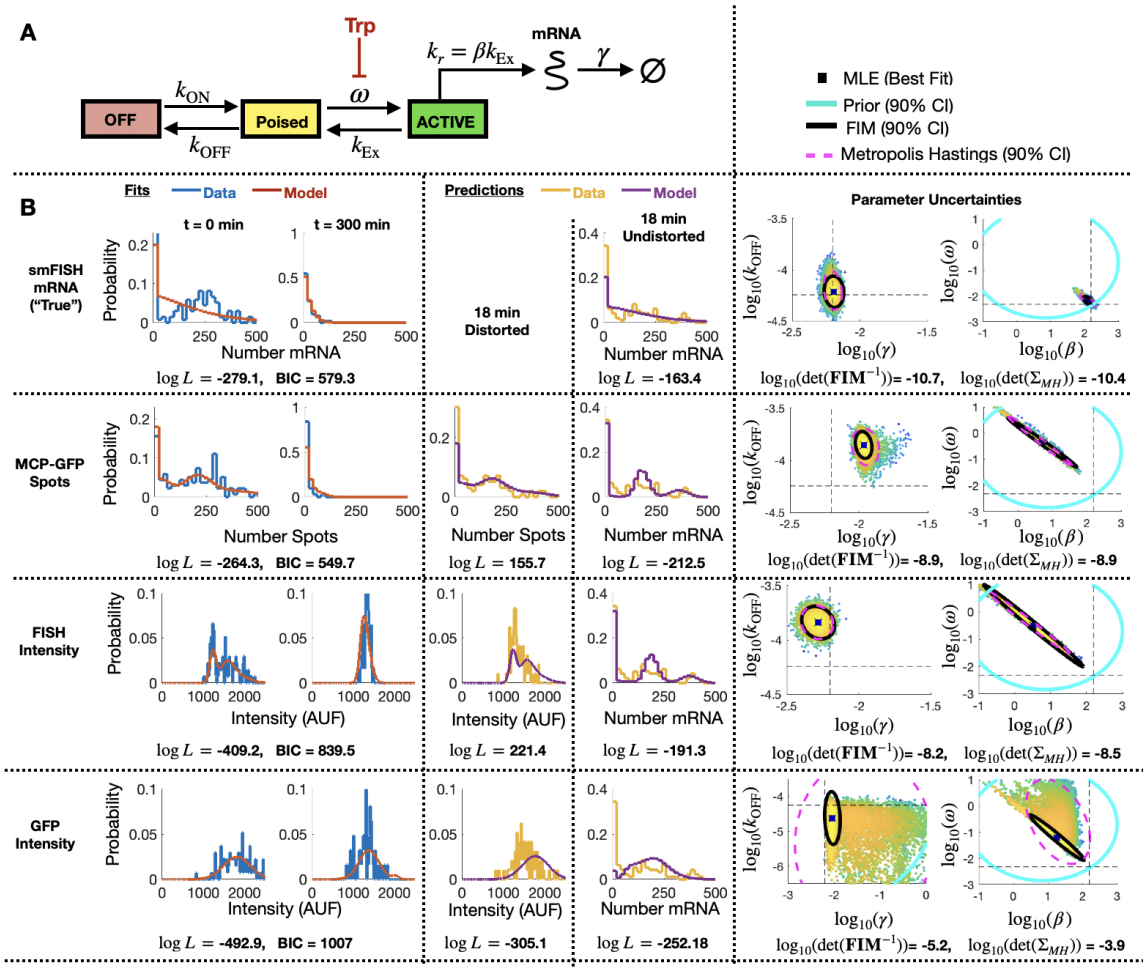


Figure 7: **Identification of stochastic model for MS2X128 cassette-tagged HIV-1 reporter gene. (A)** Schematic of the 3-state bursting gene expression model. **(B)** Results for model fitting, prediction, and uncertainty quantification for measurements based smiFISH spots (top row), MCP-GFP spots (row 2), total FISH intensities (row 3) and GFP intensities (row 4). Left two columns show the measured and model-fitted probability mass vectors (PMV) at 0 and 300 min after 5μM Tpt. Third column shows the model-predicted and measured PMV for the corresponding (distorted) measurement modality at 18 min after 5μM Tpt. Fourth column shows the model prediction without distortion and measured PMV for the smFISH mRNA count at 18 min after 5μM Tpt. All histograms use a bin size of 20. Log-likelihood values for all model-data comparisons (and BIC values for fitting cases, $k = 4$ parameters, N = 197 cells) are computed without binning and are shown below the corresponding histograms. Right two columns show joint parameter uncertainty for model estimation using data for 0 and 300 min after 5μM Tpt. In each case, the 90% CI for prior is shown in cyan; Metropolis Hastings samples (N = 20,000) are shown in dots; 90% CI for posterior is shown in dashed magenta; and FIM-based estimate of 90% CI is shown in black. Horizontal and vertical dashed black lines denote the "true" parameters and are defined as the MLE when using fit to the smFISH counts and using all time points. Determinant of inverse FIM and covariance of MH samples is shown below each pair of uncertainty plots (both use log base 10).

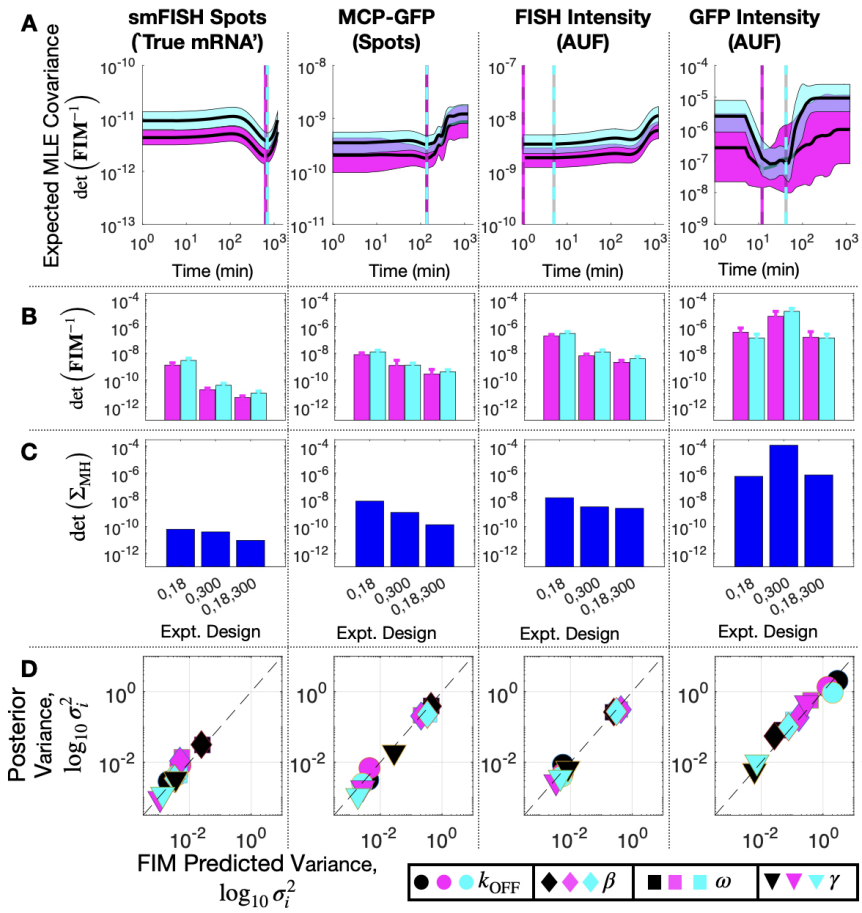


Figure 8: **Design of Subsequent Experiment for MS2X128 cassette-tagged HIV-1 reporter gene.** **(A)** Expected volume of uncertainty ($\det(\mathbf{FIM}^{-1})$) versus time of third measurement assuming 100 cells and measurement of: (left to right) smiFISH mRNA, MCP-GFP spots, FISH intensity, or GFP intensity. Solid lines and shading denote mean ± SD for 20 parameter sets selected from MH chains after fitting initial data (magenta, t =(0,300) min) or final data (cyan, t =(0,18,300) min). Cyan and magenta vertical lines denote the optimal design for the third experiment time assuming the corresponding parameter values. **(B)** Expected volume of MLE uncertainty ($\det(\mathbf{FIM}^{-1})$) for different sets of experiment times and measurement modalities and averaged over 20 parameters sets sampled from the MH chains for initial fit (magenta) or final parameter estimates (cyan). **(C)** Volume of MLE uncertainty ($\det(\Sigma_{\mathrm{MH}})$ estimated from MH analysis in the same experiment designs as **B**. **(D)** Posterior variance versus FIM prediction of variance for each parameter (symbol key at bottom right), for each measurement modality (different columns) and for analyses based on different sets of data: t =(0,18) min (black), t =(0,300) min (magenta), or t =(0,18,300) min (cyan). All MH analyses contain 20,000 samples. Measurements include 135 cells at $t = 0$, 96 at $t = 18$, and 62 at $t = 300$. Parameter uncertainties defined in log base 10 for all panels.