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

Huy D. Vo
Huy.Vo@colostate.edu

Brian Munsky
Brian.Munsky@colostate.edu

May 11, 2021

**Abstract**

Measurement error is a complicating factor that could reduce or distort the information contained in an experiment. This problem becomes even more serious in the context of experiments to measure single-cell gene expression heterogeneity, in which important quantities such as RNA and protein copy numbers are themselves subjected to the inherent randomness of biochemical reactions. Yet, it is not clear how measurement noise should be managed, in addition to other experiment design variables such as sampling size and frequency, in order to ensure that the collected data provides useful insights on the gene expression mechanism of interest. To address these experiment design challenges, we propose a model-centric framework that makes explicit use of measurement error modeling and Fisher Information Matrix-based criteria to decide between experimental methods. This unified approach not only allows us to see how different noise characteristics affect uncertainty in parameter estimation, but also enables a systematic approach to designing hybrid experiments that combine different measurement methods.

## 1 Introduction

Heterogeneity in gene expression at the single-cell level has wide-ranging biological and clinical consequences, from bacterial persistence [22], viral infection [7], to tumor heterogeneity [6, 26]. Beside genetic and environmental factors, a significant degree of heterogeneity is caused by biochemical noise [27, 38, 2]. Therefore, even genetically identical cells grown in the same experimental conditions may still 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 upstream from the pathway of interest, obscures the underlying mechanisms when looked through the lens of deterministic models and bulk measurements [31]. 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, 15, 1]. Through this joint experimental and modeling approach, mechanisms of single-cell gene expression can be explained, predicted [32, 33, 31], or even controlled [30, 29, 4].

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 and single-molecule *in situ* hybridization (smFISH), from which datasets are produced in the form of fluorescence intensity histograms [17, 24, 25] or molecular count histograms [11, 36, 23]. These measurements, like any other experimental measurement, can be corrupted by errors from imprecise detection or data processing techniques. In flow cytometry, variations in the binding efficiency, binding specificity or intensity of individual fluorescent reporters or probes, or the existence of background fluorescence, are unavoidable perturbations that may obscure the true copy number of RNA or protein [32, 44]. While smFISH datasets are generally considered to provide the 'gold standard' for measuring transcriptional heterogeneity, the estimation of molecular copy numbers in single cells depends heavily on image segmentation and spot counting algorithms [3, 35, 37, 18] that involve several threshold parameters, which are set *ad hoc* and typically vary from one expert to another. In addition, the standard settings of smFISH have been shown to lead to low sensitivity for the detection of short RNAs, yet increasing the number of probes lead to spurious detection due to non-specific binding, especially in the case of short RNA sequences [42]. To add more complexity, camera noise and technical errors also introduce discrepancies between the estimated copy number and the true number of molecules present within the cell [18] (Fig. 1).

The presence of these various sources of measurement error indicates that error modeling is a significant component in analyzing single-cell experimental datasets. In addition, it also begs an even more intriguing question of how error models can be taken into account when designing optimally informative experiments for discrete stochastic gene regulation. For instance, under what conditions can detailed statistical measurement noise models and cheap single-cell measurements combine to replace accurate, yet far more expensive experimental equipment? This paper advocates for a model-driven approach that has an explicit component of measurement error modeling. We compute the Fisher Information Matrix (FIM) of the distribution of *observed* measurements to quantitatively compare different types of measurements as well as designing hybrid experiments that combine diverse measurements. This approach is applicable to stochastic gene circuits in the low copy number regime, with explicit treatment of measurement noise that could be temporal, discrete, non-symmetric, and non-Gaussian. In addition, it allows us to compute the FIM associated with the noisy measurements themselves, thus enabling the comparisons of different types of measurement errors on the amount of information in the resulting experimental data, and therefore to the precisions of models inferred from these data. In particular, it could help decide whether one should choose a more accurate, but lower throughput, measurement method or an alternative with

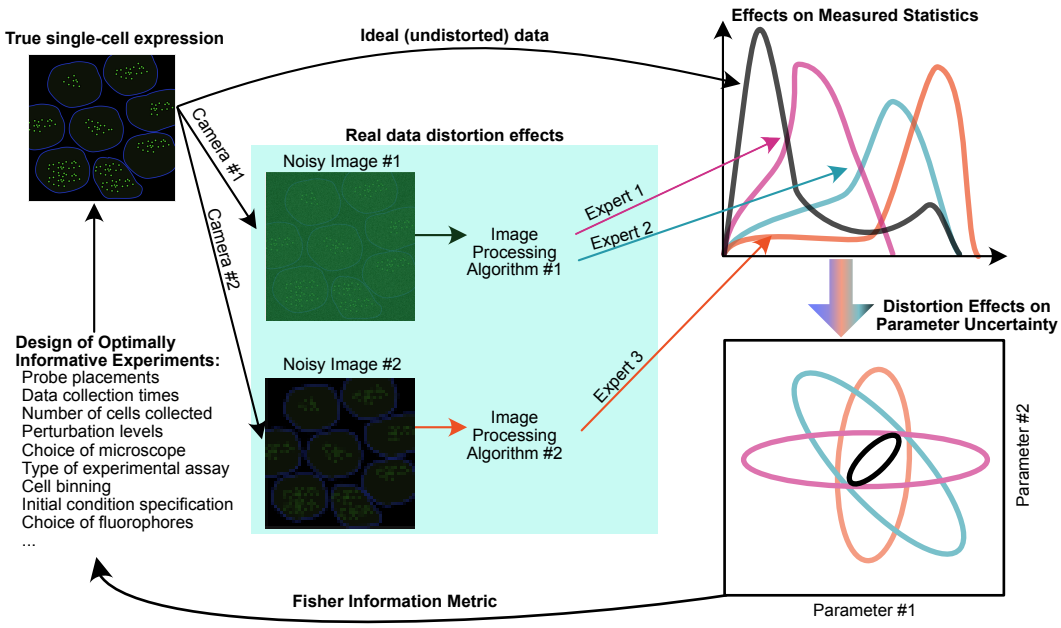


Figure 1: The challenge of interpreting smFISH experiment. The information from the ideal image of a cell could be corrupted by camera noise. These noisy images contain ambiguities that may lead to different counts by experts and/or artificial vision systems. Filtering algorithms could be applied to the noisy image, but the true signal could only be recovered in part. As a consequence, there is much uncertainty in the final estimates of the true copy number of mRNA within the cell.

higher throughput but lower accuracy.

## 2 Results

Many models have been proposed to describe the heterogeneity of gene expression [32, 33, 41], and the experimental investigation of these tools have used many different experimental approaches, each with their own characteristic measurement errors. Our proposed Fisher Information Analysis tools (methods section) are amenable to explore the impact of a very broad intersection of these models and measurement distortion effects. However, in the interest of simplicity, we choose illustrate these methods using the telegraph model, one of the simplest, yet most commonly utilized models of bursty gene expression [34, 36, 43, 40, 41, 21]. In this model, a gene is either in the inactive or active state, with transition between states occurs randomly with average rates $k_{\mathrm{ON}}$ (to activate the gene) and $k_{\mathrm{OFF}}$ (to deactivate the gene). When active, the

gene can be transcribed with an average rate $k_r$ to produce mRNA molecules, each of which degrade with an average rate $\gamma$. These parameters cannot be measured directly and must be statistically inferred from datasets collected using single-molecule, single-cell experimental methods such as smFISH [36, 16], flow cytometry [44], or live-cell imaging [43, 14]. In this work, we focus only on the former two experimental approaches in which the data consists of independent single-cell measurements taken at different times. Using the telegraph model, we will examine measurement distortion effects corresponding to probe binding inefficiency and spot colocolization (in smFISH experiments), and reporter fluorescence intensity variability (in flow cytometry experiments).

## 2.1 Ignoring measurement noise in modeling hampers accurate learning of parameter values

In principle, naively fitting mathematical models to single-cell datasets without proper appreciation for the measurement noise can lead to severe error. Here we provide a simple example based on the telegraph model to illustrate our point, consider a non-equilibrium stochastic bursting gene expression model (see table 1 for specific model and parameters). At some time, a random sample of 1000 cells has a true mRNA distribution as shown in Fig. 2A(red), but if each true mRNA molecule in the cell has a 50 percent chance of being missed by the spot-counting technique, the emeasured probability distribution for the same experiment would be substantially distorted as shown in Fig. 2A (orange). Now suppose that one performs this experiment $1,600$ times and finds the best possible CME model fit to each experiment, e.g., by maximizing strictly bounded likelihood estimates [12]. 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. 2B (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 methods introduced below, Fig. 2B (orange) shows that even the inaccurate spot counting procedure can be effectively corrected by explicitly accounting for measurement uncertainty in the modeling phase.

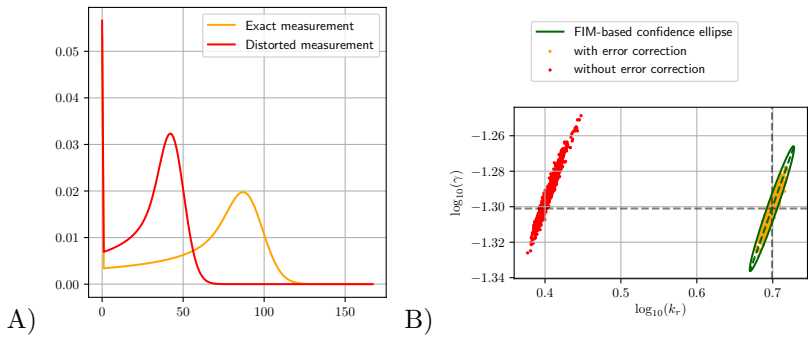


Figure 2: (A) The true mRNA distribution at time $t := 60$ min and its various distortions due to error in image processing. See the main text for specific model parameters. (B) Results of MLEs using 1600 randomly generated datasets that are perturbed by random spot misses. The intersection of the dotted straight lines represents the position in parameter space of the true data generating process. Failing to model measurement uncertainty can lead to biased parameter estimation, regardless of the number of experiments conducted.

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

In the context of single-cell smFISH and flow cytometry experiments, let $\boldsymbol{X}$ be the stochastic process that models the copy numbers of mRNAs and proteins of interest, and $\boldsymbol{Y}$ represents the stochastic process of observed quantities such as the number of spots in single-cell images, or the fluorescence intensities recorded in flow cytometry datasets. Viewing $\boldsymbol{Y}$ as a probabilistic distortion of $\boldsymbol{X}$, we seek to model the conditional probability $P(\boldsymbol{Y}|\boldsymbol{X})$ of $\boldsymbol{Y}$ given $\boldsymbol{X}$. By making appropriate conditional independence assumptions on $\boldsymbol{Y}$ (details in Methods section), we deduce that the probability distribution $\boldsymbol{p}^{\boldsymbol{Y}}(t)$ of a spatially distorted measurement $\boldsymbol{Y}(t)$ is a linear transformation of the probability distribution $\boldsymbol{p}^{\boldsymbol{X}}(t)$ of the true cellular state $\boldsymbol{X}(t)$. We refer to this linear transformation as the *probabilistic distortion operator*. This matrix is element-wise the condition probability density/mass function of $\boldsymbol{Y}(t)$ given $\boldsymbol{X}(t)$. Specifiying appropriate statistical distributions for this conditional probabilistic dependence gives us concrete models of $\boldsymbol{Y}$ for different types of measurement distortion. These transformations allow us to obtain the distribution of observations from the solution of the chemical master equation (CME) that can be obtained either numerically or analytically, as well as its sensitivities with respect to model parameters. This allows us to compute the Fisher Information Matrix of the observation variable $\boldsymbol{Y}$ itself under different assumptions on the type and degree of measurement noise. We give concrete examples of our approach for the rest of this section.

## 2.3 Fisher Information Matrix accurately estimates how different single-cell measurement noises induce model parametrization uncertainty

We consider two formulations of measurement distortion matrix (cf.eq. (3) in Methods), corresponding to three different scenarios in which experimental errors arise from either lossy spot counting or spatial integration cell fluorescence intensity. Fig. 3 visualizes the image distortion kernels for these situations as nonlinear functions from the joint domain of $\boldsymbol{x}$ and $y$ to the positive real numbers.

The first model concerns with when $\tilde{y}$ is obtained from an ineffecient mRNA detection process that may be caused by an error in the spot counting algorithm when applied on images taken in an smFISH experiment. We model $y|\boldsymbol{x} = n$ with a binomial distribution $B(p_{miss}, n)$, where each spot has a chance $p_{miss}$ of being ignored by the couting algorithm, resulting in uderestimation of the true mRNA copy number. In this model, the bias $E(y - n_{RNA}) = np_{miss}$ is a linear function of $n_{RNA}$.

The second model conerns fluorescent intensity in flow cytometry measurements. We use the model proposed in [32], in which $\tilde{y} = \kappa \cdot \tilde{n}_{RNA} + \eta_{\tilde{n}_{RNA}} + \tilde{\varepsilon}_{BG}$, where $\tilde{\eta}_{\tilde{n}_{RNA}} \sim \mathcal{N}(0, \tilde{n}_{RNA}\sigma_{probe})$ models fluorescent heterogenity and $\tilde{\varepsilon}_{BG} \sim \mathcal{N}(\mu_{BG}, \sigma_{BG})$ is the background noise.

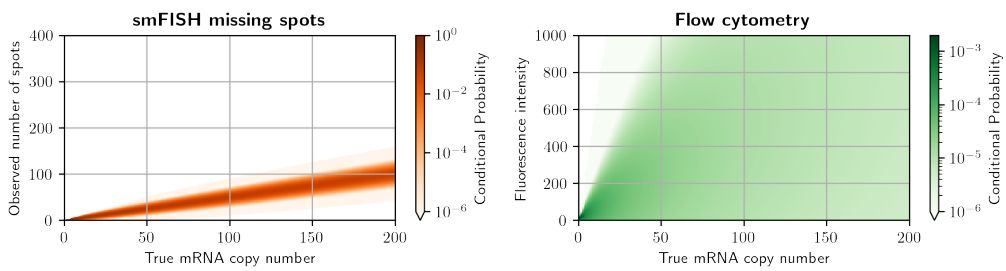


Figure 3: The probabilistic distortion operators. Left: an smFISH experiment where spots can randomly go missing. Right: flow cytometry measurements where only a perturbed version of the total fluorescence is recorded per cell. In these heatmaps, the color at point $(\tilde{x}, \tilde{y})$ is the conditional probability mass/-density of the measurement $y$ having value $\tilde{y}$ given that the true copy number $x$ has value $\tilde{x}$.

To elucidate how these measurement distortions affect parameter identification, we consider three types of experiment design. These are either noise-free smFISH, or the two noisy measurements previously described.

Consider an experimental set-up where 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 $\Delta t$ in minutes. For these classes, we place at each time point an equal number $n$ of single-cell measurements, chosen as $n := 1000$ for smFISH

| Parameter | Value | Interpretation |
|---|---|---|
| $k_{ON}$ | 0.015 | Gene activation rate (per minute) |
| $k_{OFF}$ | 0.05 | Gene deactivation rate (per minute) |
| $k_r$ | 5 | mRNA production rate (molecules per minute) |
| $\gamma$ | 0.05 | mRNA degradation rate (molecules per minute) |

Table 1: Parameters of the bursting gene expression model that were used in the simulations.

and $n := 5,000$ for flow cytometry. These choices of $n$ reflect the fact that flow cytometry has higher throughput than smFISH, albeit with lower precision per measurement. We find the optimal sampling period $\Delta t$ for each class, and compare the most informative design each class can achieve. Here, we define 'optimal' in terms of the determinant of the Fisher Information Matrix, the so-called D-optimality criterion. We compute the Fisher Information Matrices for the gene expression model using kinetic rate constants in table 1. Fig. 4A shows the information volume of the four kinds of noisy measurement described above, in addition to the ideal noise-free smFISH, at different sampling rates. We can observe that probabilistic distortions in measurements decrease the information volume, and that each measurement method requires a different optimal sampling rate. In Fig. 4B-C, we plot the three-sigma (i.e., 99.7%) confidence ellipsoids of the asymptotic distribution of MLEs projected on the $k_{ON} - k_{OFF}$ plane and the $k_r - \gamma$ plane.

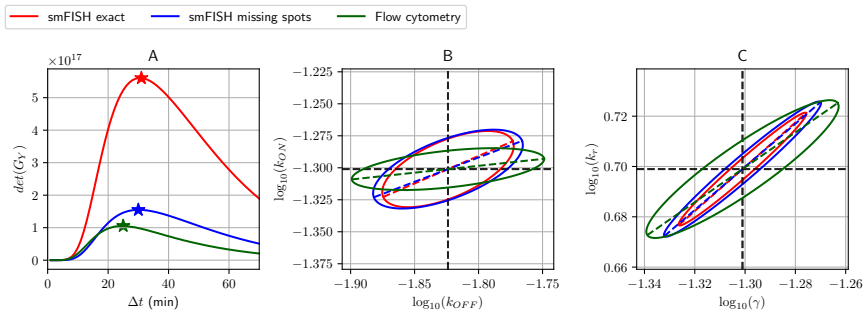


Figure 4: Optimizing sampling rate for bursting gene expression with different types of imprecise measurement. **(A)**: Comparison of D-optimality criteria ($\det(G_Y)$) in single-cell experiments with different types of measurement noise and at different sampling intervals ($\Delta t$). In this settings, independent measurements are collected at five equally-spaced time points $k\Delta t,\ k = 1, 2, 3, 4, 5$ with 1000 measurements placed at each timepoint for smFISH variants and 2000 measurements per time point for flow cytometry. The $\star$ symbol marks the optimal point of each curve. **(B) and (C)**: The three-sigma confidence ellipses projected onto the $k_{ON} - k_{OFF}$ and $k_r - \gamma$ planes. These ellipses are computed from inverting the FIMs of different measurement noise conditions at their optimal sampling rates.

## 2.4 The new FIM based approach allows us to find the maximally informative mixture of measurement types under material constraints

From the previous section, we see that different types of measurements result in different uncertainty directions in parameter space. This means that diferrent measurements have complementary strengths in reducing uncertainty along specific directions in the parameter space. This prompted us to ask whether one can use a combination of these different measurements to improve parameter estimation even further.

Motivated by this question, we consider the experiment design described in the previous section. For exact smFISH measurement with 1000 measurements placed at each time point, the previous computations told us that the optimal sampling period is $\Delta t_{opt} = 30$ minutes. We now slightly extended the design constraints to allow for mixing exact smFISH and flow-cytometry measurements at each time point. For simplicity, we require that the same composition of smFISH and flow-cytometry to be used at all time points, and assume an exchange rate (in terms of material cost or processing time) of one smFISH observation for five flow cytometry observations. The constraint we put on the experiment design is that any composition $(n_{\text{smFISH}}, n_{\text{flow-cytometry}})$ of these two measure-

ments must not exceed the cost of 1000 smFISH measurements, that is,

$$n_{\text{smFISH}} + \frac{1}{5} \cdot n_{\text{flow-cytometry}} \leq 1000.$$

We continue to use D-optimality as our design objective. Fig.5A visualizes the deterimant of the Fisher Information Matrix for all combinations of measurements that satisfy the constraint. Interestingly, the optimal composition is 756 smFISH samples and 1220 flow-cytometry samples per timepoint, which results in D-optimality criterion value of about $6 \times 10^{17}$, which is superior to the optimal values achieved by using either smFISH ($5.6 \times 10^{17}$) or flow-cytometry alone ($10^{17}$). In addition, Figs.5B-C visualize the uncertainty directions associated with each measurement allocation strategy, showing how diversifying measurment methods leads to better parameter constraints. In particular, the optimal mixture is able use the strength of flow-cytometry in constraining the burst frequency $k_{\text{ON}}$ (Fig.5B), yet it is also able to utilize the strength of smFISH in constraining mRNA sysnthesis and degradation rates (Fig.5C).

Thus, our new FIM-based analysis not only confirms our qualitative intuition about the benefit of diversifying our experimental approaches, but also gives us a systematic way to potentially maximizing the scientific value of our experiment under budgetary or material constraints.

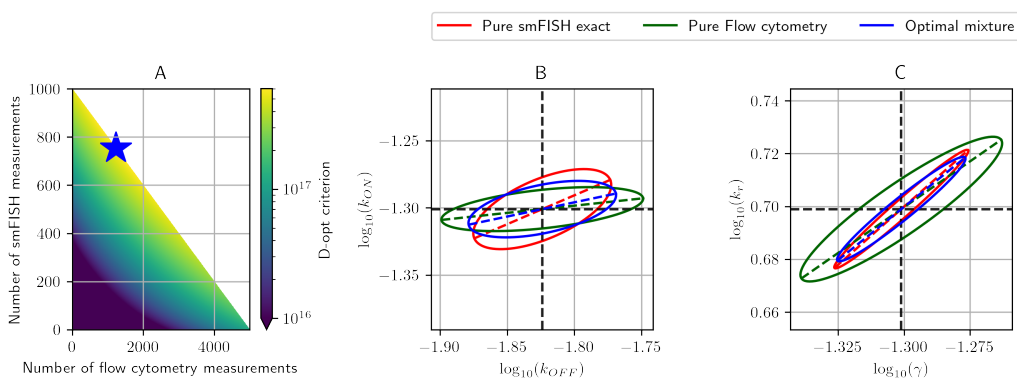


Figure 5: Budgeting smFISH and flow-cytometry measurements for maximal information. **(A)**: Determinant of the Fisher Information Matrix for different combinations of flow cytometry and smFISH measurements at 45 minutes. The $\star$ marker at 756 smFISH and 1220 flow cytometry measurements indicates the optimal allocation in terms of D-optimality criterion. **(B) and (C)**: The three-sigma confidence ellipses projected onto the $k_{ON} - k_{OFF}$ and $k_r - \gamma$ planes for three allocation strategies: all smFISH, all flow-cytometry, and D-optimal mixture marked in panel (A).

# 3 Methods

## 3.1 Stochastic modeling of gene expression

The expression dynamics of a gene or a group of genes in single cells could be modeled by a stochastic reaction network. 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 molecular count of species $i$ at time $t$ (which is an integer). Each jump in this process corresponds to the occurence of one of the reaction events $R_k$ ($k = 1, \ldots, M$), which brings the cell from the state $\boldsymbol{x}(t_-)$ right before event time $t$ to a new state $\boldsymbol{x}(t_-) + \boldsymbol{\nu}_k$ where $\boldsymbol{\nu}_k$ is the stoichiometry vector corresponding to the $k$-th reaction. The probabilistic rates at which reactions occur are characterized by the propensity (or intensity) functions $\alpha_k(t, \boldsymbol{x}|\boldsymbol{\theta})$. The vector $\boldsymbol{\theta}$ is a $d$-dimensional vector of *model parameters*. Intuitively, we interpret $\alpha_k(t, \boldsymbol{x}|\boldsymbol{\theta})\Delta t$ as the probability for reaction $k$ to fire after the waiting interval $[t, t + \Delta t)$ and with a sufficiently small waiting time $\Delta t$. The probability distributions $\boldsymbol{p}^{\boldsymbol{X}}(t; \boldsymbol{\theta})$ of single-cell gene products 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.

## 3.2 Modeling distortion of measurements

Let $\boldsymbol{y}(t)$ be the multivariate measurement made on a single cell at time $t$, such as the discrete number of spots in an smFISH experiment, or the continuous fluorescence intensity 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 is related to that of the true copy number distribution via a linear transformation of the form

$$\boldsymbol{p}^{\boldsymbol{Y}}(t; \boldsymbol{\theta}) = \boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}\boldsymbol{p}^{\boldsymbol{X}}(t; \boldsymbol{\theta}). \tag{2}$$

We call $\boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}$ the *probabilistic distortion operator* (PDO). 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

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

where Pr stands for probability mass if $\boldsymbol{Y}$ is discrete and probability density if $\boldsymbol{Y}$ is continuous. Together, equations (1) and (2) describe single-cell measurements as the time-varying outputs of a linear dynamical system on the space of

probability distributions on the lattice of $N$-dimensional discrete copy number vectors. The output matrix of this dynamical system is the PDO $\boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}$.

### 3.3 Computation of the Fisher Information Matrix for distorted experimental measurements

In practice, when closed form solutions to the CME do not exist, 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 [13]. Using eq. (2), we can transform these into the distribution $\boldsymbol{p}^{\boldsymbol{Y}}(t;\boldsymbol{\theta})$ of $\boldsymbol{Y}(t)$. Furthermore, the sensitivity indices $\partial_{\theta_j}\boldsymbol{p}^{\boldsymbol{Y}}(t;\boldsymbol{\theta})$, of the observable $\boldsymbol{Y}$ are likewise computable by

$$\boldsymbol{s}_j^{\boldsymbol{Y}}(t;\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_j}\boldsymbol{p}^{\boldsymbol{Y}}(t;\boldsymbol{\theta}) = \boldsymbol{C}_{\boldsymbol{X}\to\boldsymbol{Y}}(t)\frac{\partial}{\partial\theta_j}\boldsymbol{p}^{\boldsymbol{X}}(t;\boldsymbol{\theta}). \tag{4}$$

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

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

In this formulation, we only need to solve the (usually expensive) sensitivity equations derived form the CME only once, then apply relatively quick linear algebra operations to find the FIM corresponding to any given probabilistically calibrated microscopes or probe design.

## 4 Discussion

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

The FIM is a classical tool for optimal experiment design that is extensively used in different areas of science and engineering [5, 9, 20, 10, 8, 45]. In the context of gene expression modeling, Komorowski et al. [19] proposes a method to approximate the FIM for single-cell experiment data based on the Linear Noise Approximation (LNA). Alternatively, the FIM could be approximated

by using moment closure techniques [39]. These approaches work well in the case of high molecular copy number, but they break down when applied to systems with low molecular copy numbers [13]. In addition, it is not clear how these approximations perform when considering measurement noise and data processing noise that are non-additive, asymmetric, and non-Gaussian. To circumvent these issues, we propose an alternative framework that aims at the direct analysis of the probability distribution of the noisy measurements. By explicitly modeling the conditional probability of the observation given the true cell state, we express the oservation distribution as a linear transformation of the true process distribution that could be computed with the finite state projection. Furthermore, this approach provides a systematic way to develop composite experimental designs that combine measurements at different fidelity and throughput levels to maximize information given a budgetary constraint.

Although the formulations proposed here for the probabilistic distortion operators of smFISH measurements are derived from somewhat idealistic assumptions on the physics of the underlying measurements, adjustments can be made to capture more realistic experimental errors. For example one could construct the PDO based on empirical data that samples from the joint distribution of the high-fidelity measurements and noisy measurements. Such an approach is expected to benefit from progress in supervised machine learning literature, which has the similar goal of approximating the conditional distribution $p(y|x)$ of a noisy label $y$ given a vector $x$ of features. In such case, we only need to estimate the PDO once for each combination of labeling, microscopy, and imaging techniques.


### Acknowledgement

HV and BM were supported by National Institutes of Health (R35 GM124747).


## 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] 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.

[3] 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.

[4] 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.

[5] G. E. P. Box and H. L. Lucas. Design of Experiments in Non-Linear Situations. *Biometrika*, 46(1/2):77–90, 1959.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] Zachary Fox, Gregor Neuert, and Brian Munsky. Finite state projection based bounds to compare chemical master equation models using single-cell data. *The Journal of Chemical Physics*, 145(7), August 2016.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Donald A. McQuarrie. Stochastic Approach to Chemical Kinetics. 4:89, 1967.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.