The Journal of Chemical Physics

Exact and efficient hybrid Monte Carlo algorithm for accelerated Bayesian inference of gene expression models from snapshots of single-cell transcripts

Cite as: J. Chem. Phys. **151**, 024106 (2019); https://doi.org/10.1063/1.5110503
Submitted: 17 May 2019 . Accepted: 14 June 2019 . Published Online: 09 July 2019

Yen Ting Lin, and Nicolas E. Buchler

openaccess

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Non-equilibrium hybrid insertion/extraction through the 4th dimension in grand-canonical simulation
The Journal of Chemical Physics **151**, 021101 (2019); https://doi.org/10.1063/1.5110478

Machine-learned electron correlation model based on correlation energy density at complete basis set limit
The Journal of Chemical Physics **151**, 024104 (2019); https://doi.org/10.1063/1.5100165

A new perspective for nonadiabatic dynamics with phase space mapping models
The Journal of Chemical Physics **151**, 024105 (2019); https://doi.org/10.1063/1.5108736

AIP Publishing

Exact and efficient hybrid Monte Carlo algorithm for accelerated Bayesian inference of gene expression models from snapshots of single-cell transcripts

Cite as: J. Chem. Phys. **151**, 024106 (2019); doi: 10.1063/1.5110503
Submitted: 17 May 2019 • Accepted: 14 June 2019 •
Published Online: 9 July 2019

Yen Ting Lin[1,a),b)] and Nicolas E. Buchler[2]

AFFILIATIONS

[1] Center for Nonlinear Studies and Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
[2] Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, North Carolina 27607, USA

[a)] **Current address:** Information Sciences Group, Computer, Computational and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.
[b)] **Electronic mail:** yentingl@lanl.gov.

ABSTRACT

Single cells exhibit a significant amount of variability in transcript levels, which arises from slow, stochastic transitions between gene expression states. Elucidating the nature of these states and understanding how transition rates are affected by different regulatory mechanisms require state-of-the-art methods to infer underlying models of gene expression from single cell data. A Bayesian approach to statistical inference is the most suitable method for model selection and uncertainty quantification of kinetic parameters using small data sets. However, this approach is impractical because current algorithms are too slow to handle typical models of gene expression. To solve this problem, we first show that time-dependent mRNA distributions of discrete-state models of gene expression are dynamic Poisson mixtures, whose mixing kernels are characterized by a piecewise deterministic Markov process. We combined this analytical result with a kinetic Monte Carlo algorithm to create a hybrid numerical method that accelerates the calculation of time-dependent mRNA distributions by 1000-fold compared to current methods. We then integrated the hybrid algorithm into an existing Monte Carlo sampler to estimate the Bayesian posterior distribution of many different, competing models in a reasonable amount of time. We demonstrate that kinetic parameters can be reasonably constrained for modestly sampled data sets if the model is known *a priori*. If there are many competing models, Bayesian evidence can rigorously quantify the likelihood of a model relative to other models from the data. We demonstrate that Bayesian evidence selects the true model and outperforms approximate metrics typically used for model selection.

I. INTRODUCTION

Gene expression is a biochemical process driven by the chance collisions of molecules, which can result in strong stochastic signatures and cell-to-cell variability in gene dynamics. Advances in single-cell and single-molecule technologies have provided unprecedented resolution on the stochastic dynamics of gene expression.[1] Dynamic assays measure gene expression in living cells either directly via transcript tagging[2–5] or indirectly via fluorescent or luminescent proteins.[6–9] Static assays measure transcript levels in fixed cells either using a cocktail of fluorescently labeled DNA oligos that bind specific transcripts[10,11] or via single-cell RNA sequencing.[12,13] Static assays are popular because they do not require genetic modifications and are easily multiplexed. The disadvantage is that static assays only provide population snapshots of transcripts levels and cannot follow the dynamics of transcription in a single cell through time.

To this end, static assays have relied upon mathematical models to *infer* dynamic properties of gene expression in single cells from the measured snapshot of transcript levels; see Ref. 14 for a review. Inference requires (1) appropriate models of stochastic gene expression, (2) numerical methods to calculate the time-dependent mRNA distribution in a population of cells given any underlying model and associated parameters, and (3) quantifying the likelihood that measured data were sampled from the calculated distribution. We recently developed a Bayesian approach (BayFISH) that uses this likelihood to infer best-fitting parameters from single cell data and quantifies their uncertainty using the posterior distribution.[15,16] Although Bayesian inference is the most complete and rigorous approach, it requires significantly more computation than other approximate methods, e.g., maximum likelihood.

Here, we developed a hybrid numerical method that accelerates the calculation of time-dependent mRNA distributions by 1000-fold compared to standard methods. We integrated this method into BayFISH and, for the first time, one can estimate the Bayesian posterior distribution of many competing models of gene expression in a reasonable amount of time. The Bayesian evidence rigorously quantifies the likelihood of a model relative to other models given the data, and we show that Bayesian evidence selects the true model and outperforms approximate metrics, e.g., Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC), typically used for model selection. Our accelerated Bayesian inference represents a significant advance over existing methods used for inferring gene expression models from snapshots of single cell transcripts.

II. CONNECTING MODELS OF GENE EXPRESSION TO SINGLE CELL DATA

Our inference method uses data from single-molecule RNA Fluorescence *In Situ* Hybridization (smFISH) but could include single cell data from other static assays. The smFISH technique labels transcripts with fluorescent DNA oligos and measures both the number of mature mRNAs (m) and the number of gene loci with high-activity transcription sites (*TS*s); see Fig. 1(a). A typical smFISH data set is a histogram $h = h(\vec{\omega})$, where $\vec{\omega} \in \Omega$ is the set of all possible states (m, TS) that can be measured in a cell.

A broad spectrum of measured gene expression profiles in bacteria and eukaryotes is well-explained by discrete state gene expression models,[17,18] summarized by the following reactions:

$$\begin{aligned}
&\text{Gene state } \; i \xrightarrow{\kappa_{ij}} \text{Gene state } \; j,\\
&\text{Gene state } \; i \xrightarrow{\beta_i} \text{Gene State } \; i + \text{mRNA},\\
&\qquad \text{mRNA} \xrightarrow{\delta} \varnothing.
\end{aligned}$$

In this article, we adopt a two-allele, 3-state model [Fig. 1(b)] as a case study for modeling stochastic gene expression in eukaryotes and for testing our method of accelerated Bayesian inference. We further focus on dynamic smFISH experiments that perturb gene expression (e.g., induction) and then measure mRNA distributions at different times after induction to infer dynamics and kinetic parameters. Induction can change one or more of the model parameters [Fig. 1(c)]. The smFISH data from an induction experiment consist of a joint histogram $h = h(\vec{\omega}, t_\ell)$, where t_ℓ are independent observations made at different times before and after induction. If the changed parameters are unknown *a priori*, then one should evaluate all possible induction models, which leads to a combinatorial explosion in model space. For example, there are $2^8 = 256$ candidate induction models for the 3-state model shown in Fig. 1(b), of which the model shown in Fig. 1(c) is one. In Sec. V, we will consider $2^6 = 64$ candidate models where the same two parameters (δ and β_0) are known *a priori* to not change upon induction.

A likelihood approach is used to connect mathematical models of stochastic gene expression to single cell data. Formally, the likelihood $\mathcal{L}$ is the probability that a candidate model $\mathcal{M}$ and its associated parameter set $\vec{\theta}$ would generate a given set of data (h). The number of parameters (i.e., dimension of $\vec{\theta}$) is determined by the model structure $\mathcal{M}$. Mathematically, the likelihood $\mathcal{L}$ is a function of the joint probability distribution $\mathbb{P}(\vec{\omega}, t_\ell|\vec{\theta}, \mathcal{M})$ of a candidate model $\mathcal{M}$ and its associated parameters $\vec{\theta}$ at discrete observation times,

$$\mathcal{L} = \prod_{t_\ell \in \Phi} \left\{ \mathbb{M}_\ell \cdot \prod_{\vec{\omega} \in \Omega} [\mathbb{P}(\vec{\omega}, t_\ell|\vec{\theta}, \mathcal{M})]^{h(\vec{\omega}, t_\ell)} \right\}, \tag{1}$$

where Φ is the set of observation times and $\mathbb{M}_\ell$ is the multinomial coefficient associated with each $h(\vec{\omega}, t_\ell)$ that arises because the data were not ordered.

In our Bayesian inference work flow [Fig. 1(d)], each candidate model $\mathcal{M}$ in the class of possible models $\{\mathcal{M}\}$ will require a large number ($\geq 10^6$) of Monte Carlo steps where, at each step, numerical simulations calculate the time-dependent mRNA distributions and evaluate the likelihood that different parameter sets $\vec{\theta}$ for that model generated the observed data. Our previous software[15,16] took days to perform the likelihood calculations for one model, which highlights the challenge of using Bayesian inference to evaluate hundreds of models and perform model selection. Below, we develop a hybrid method that both accelerates numerical simulation and likelihood calculations, and (in contrast to standard methods) scales with the number of multicore processors, thus allowing for efficient parallelization.

III. A NOVEL HYBRID METHOD TO CALCULATE THE TIME EVOLUTION OF DISCRETE-STATE MODELS

While exact time-dependent solutions exist for two-state models,[19–21] it is hard to generalize this analysis to models with more states. It is therefore necessary to solve the general time-dependent problem using numerical simulations. There are two classes of numerical procedures to solve the time evolution of a discrete-state model for a given set of parameters. The first class forward-evolves the chemical master equations (CMEs), which are a system of infinitely many coupled ordinary differential equations (ODEs) that describe the joint probability distributions $\mathbb{P}(\vec{\omega}, t)$ as a function of time.[22,23] To be numerically feasible, a truncation scheme (e.g., only consider mRNA levels below a maximum M)

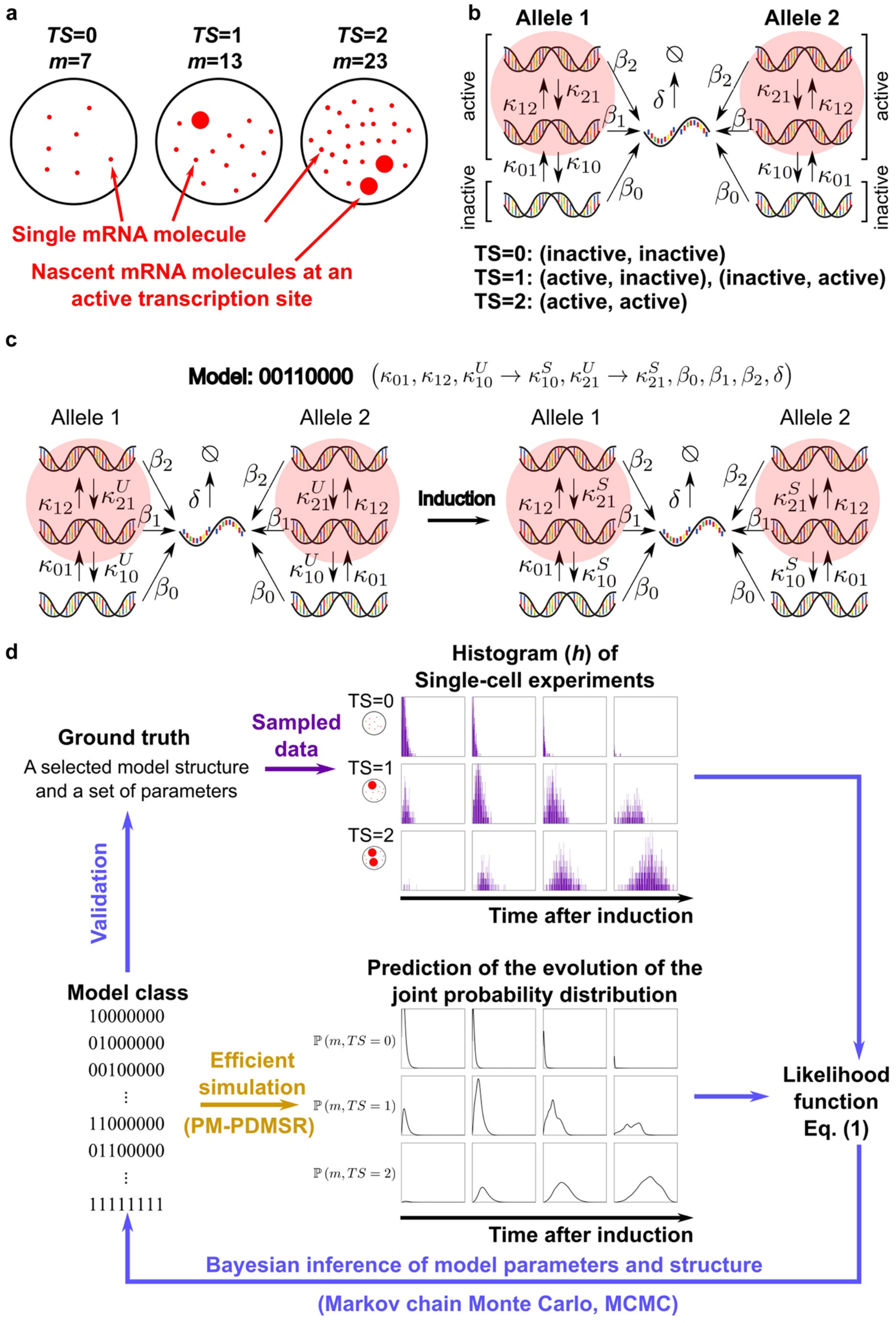
a
TS=0
m=7
TS=1
m=13
TS=2
m=23
Single mRNA molecule
Nascent mRNA molecules at an active transcription site
b
Allele 1
Allele 2
active
inactive
κ_{12} κ_{21} κ_{01} κ_{10}
β_0 β_1 β_2 δ $\emptyset$
TS=0: (inactive, inactive)
TS=1: (active, inactive), (inactive, active)
TS=2: (active, active)
c
Model: 00110000 $(\kappa_{01}, \kappa_{12}, \kappa_{10}^U \to \kappa_{10}^S, \kappa_{21}^U \to \kappa_{21}^S, \beta_0, \beta_1, \beta_2, \delta)$
Allele 1
Allele 2
κ_{12} κ_{21}^U κ_{01} κ_{10}^U
Induction
κ_{12} κ_{21}^S κ_{01} κ_{10}^S
d
Histogram (h) of Single-cell experiments
Ground truth
A selected model structure and a set of parameters
Sampled data
TS=0
TS=1
TS=2
Time after induction
Validation
Prediction of the evolution of the joint probability distribution
Model class
10000000
01000000
00100000
⋮
11000000
01100000
⋮
11111111
Efficient simulation (PM-PDMSR)
$\mathbb{P}(m, TS=0)$
$\mathbb{P}(m, TS=1)$
$\mathbb{P}(m, TS=2)$
Time after induction
Likelihood function Eq. (1)
Bayesian inference of model parameters and structure
(Markov chain Monte Carlo, MCMC)

FIG. 1. Single cell data and models of gene expression. (a) The single-molecule RNA FISH (smFISH) technique provides information on the localization and numbers of mature mRNAs (*m*) in single cells, including clusters of nascent transcripts produced at transcription sites (*TS*s) at active genetic loci. (b) A diploid, two-allele 3-state genetic model where κ_{ij} is the transition rate between genetic states, β_i is the mRNA synthesis rate of each state, and δ is the mRNA degradation rate. (c) Induction changes one or more parameters from an unstimulated (U) to a stimulated value (S). Here, we show one of the many possible induction models $\mathcal{M}$, labeled in binary (00110000). (d) Schematic of the Bayesian inference work flow.

is used to reduce the infinite size of the dynamical system. While this approach delivers accurate estimates of the temporal evolution of the truncated CME, there are two shortcomings. First, the number of ODEs scales as S^2M, where S is the number of genetic states for each allele. The ODE system becomes unwieldy for mammalian cells where the number of observed mRNAs per cell can be $\mathcal{O}(10^3)$.[24–27] Second, the forward integration of the CME requires stiff ODE solvers, which can place demands on memory resources and hinder parallel processing. The second class of numerical procedures utilizes kinetic Monte Carlo methods (e.g., continuous time Markov chain simulation[28–33]) to sample the temporal evolution of the joint probability distribution $\mathbb{P}(\vec{\omega}, t)$. While this approach is computationally less expensive, it comes at the cost of having to sample over many runs to achieve equivalent accuracy to the CME.

In this article, we develop a hybrid simulation method (the Poisson Mixture with a Piecewise Deterministic Markov Switching Rate or PM-PDMSR) which leverages analytical results and the efficiency of the kinetic Monte Carlo method. The key result is that the mRNA distribution can be exactly calculated for any realization (trajectory) of the genetic state, $s(t)$; see Appendix. Once transient initial conditions have burned off ($t \gg \delta^{-1}$), where δ is the mRNA degradation rate, the mRNA (N_{mRNA}) distribution is always Poisson, $\mathbb{P}(N_{\mathrm{mRNA}} = m) = \lambda^m(t)e^{-\lambda(t)}/m!$ with a dynamic rate $\lambda(t)$ satisfying the following piecewise ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda(t) = \beta_{s(t)} - \delta\lambda(t) \tag{2}$$

with an initial condition $\lambda(0) = 0$. Given any trajectory $s(t)$, we can exactly compute the mRNA distribution $\mathbb{P}(m|s(t))$; see Figs. 2(a) and 2(b). Our goal, however, is to determine the joint distribution $\mathbb{P}(\vec{\omega}, t)$, which requires us to generate N_s sample paths of $s(t)$ that cover $\mathbb{P}(s, t)$. The sample paths in the small genetic state space (S^2-dimensional) are efficiently generated using standard kinetic Monte Carlo methods. After accumulating a large number of sample paths N_s generated by the underlying model, the mixture of the Poisson distributions recovers the mRNA distribution via a convolution

$$\widehat{\mathbb{P}}(N_{\mathrm{mRNA}}(t) = m, s(t) = i) = \frac{1}{N_s}\sum_{k=1}^{N_s} \delta_{i,s_k(t)} \frac{\lambda_k^m(t)e^{-\lambda_k(t)}}{m!}, \tag{3}$$

where $\lambda_k(t)$ is the solution of (2) subject to the kth sample path of genetic switching trajectory $s_k(t)$ and $\delta_{i,j}$ is the Kronecker delta [see Figs. 2(c)–2(e)].

A detailed description of the hybrid simulator is given in the Appendix. We evaluated the efficiency of the hybrid simulator relative to the CME in performing a single step of the Bayesian inference

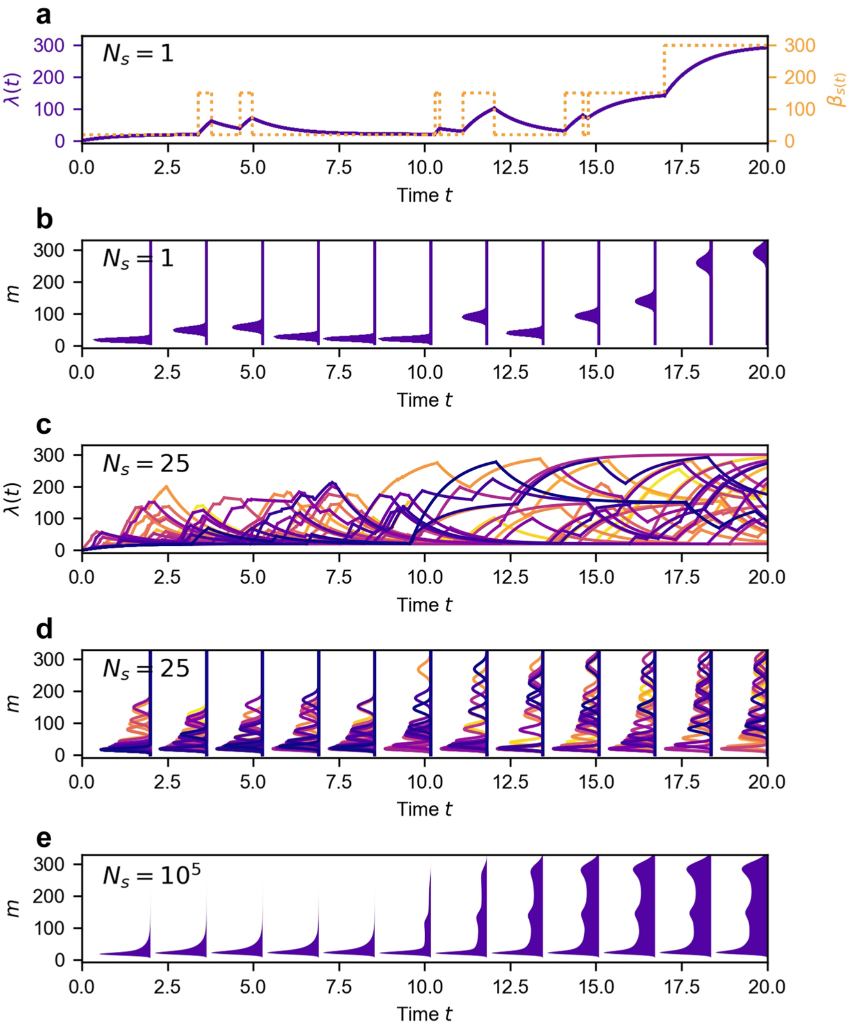

FIG. 2. Hybrid simulation method, PM-PDMSR. For simplicity, we illustrate the principle of PM-PDMSR for a single allele, 3-state model ($\mathcal{M}$ = 00110000). The gene is induced at t = 10. Model parameters: before stimulation $\left(\kappa_{01}, \kappa_{12}, \kappa_{21}^U, \kappa_{10}^U, \beta_0, \beta_1, \beta_2, \delta\right)$ = (0.5, 0.5, 5, 5, 20, 150, 300, 1) and after stimulation, $\kappa_{21}^S = \kappa_{10}^S$ = 0.5. [(a) and (c)] Changing transcription and dynamic rates for N_s = 1 and N_s = 25 sample paths. [(b) and (d)] Poisson mRNA distribution for the sample paths shown in (a) and (c), respectively. (e) Convolution of Poisson mixtures generated from $N_s = 10^5$ sample paths.

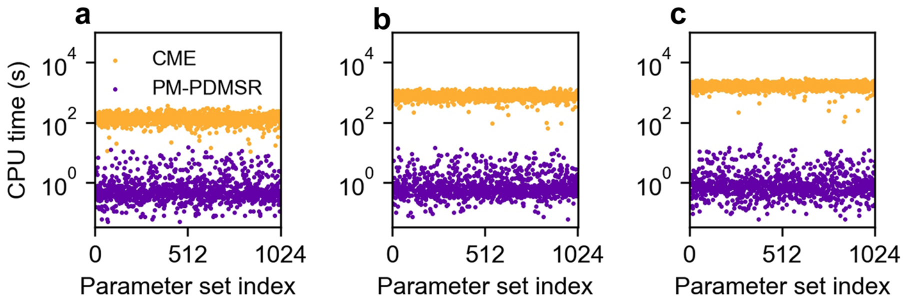

FIG. 3. Efficiency of the hybrid method relative to the CME method. We measured the time for each method to complete one step of Bayesian inference, i.e., calculate joint distribution and evaluate the likelihood. This comparison was performed for increasingly complex model classes: (a) 2-state, (b) 3-state, and (c) 4-state models of gene expression. Each model class was evaluated for 1024 different parameters along with associated data sets; see Appendix for details.

work flow, i.e., simulate the joint distribution $\mathbb{P}(\vec{\omega}, t)$ and calculate the likelihood $\mathcal{L}$ that this joint distribution produced a given data set (h). We benchmarked the simulators on diverse classes of discrete-state models, parameter sets, and data sets; see Fig. 3. The hybrid simulator is up to 10^3 more efficient for models with increased genetic states, $S = 3$ and 4. The efficiency gain of the hybrid simulator originates from the fact that $\mathbb{P}(m|s(t))$ is solved exactly in mRNA space (and is independent of the size of M) and that $\mathbb{P}(s, t)$ is sampled efficiently in genetic-state space via kinetic Monte Carlo techniques. The accelerated hybrid method achieved equivalent accuracy to the CME; see Fig. S1 of the supplementary material. Finally, we tested the scaling of efficiency of different simulators on a modern multicore workstation, which can execute up to 64 parallel threads. We found that the hybrid method runs well in parallel, i.e., the total time needed for a fixed computational task distributed over T threads scales as $1/T$. Surprisingly, the CME method exhibited stiff scaling such that the total time stayed constant and did not decrease with increasing threads; see Fig. S2 of the supplementary material and Sec. VI.

IV. BAYESIAN INFERENCE AND UNCERTAINTY QUANTIFICATION OF MODEL PARAMETERS

Equipped with an efficient simulator of the time-dependent joint probability distribution and likelihood calculation for any model and parameter set, we first turned our attention to uncertainty quantification of model parameters $\vec{\theta}$ for a fixed model $\mathcal{M}$. Given a likelihood, Bayesian inference uses the Bayes formula to update any prior beliefs $\mathbb{P}(\vec{\theta}|\mathcal{M})$ and calculate the posterior distribution $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$ of parameters $\vec{\theta}$ given the data h and a fixed model $\mathcal{M}$,

$$\mathbb{P}(\vec{\theta}|h, \mathcal{M}) = \frac{\mathbb{P}(h|\vec{\theta}, \mathcal{M})\mathbb{P}(\vec{\theta}|\mathcal{M})}{\mathbb{P}(h|\mathcal{M})} = \frac{\mathcal{L} \cdot \mathbb{P}(\vec{\theta}|\mathcal{M})}{\mathbb{P}(h|\mathcal{M})}. \quad (4)$$

As done previously, we resorted to Markov chain Monte Carlo (MCMC) with a Metropolis–Hastings (MH) sampler to estimate the posterior distribution $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$; see Appendix and Ref. 15. We assumed that the prior $\mathbb{P}(\vec{\theta}|\mathcal{M})$ is log-uniform. At each MCMC step, the MH sampler randomly proposes a nearby parameter set and computes the ratio of the posterior probability $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$ relative to that of the current parameter set and probabilistically accepts or rejects the proposal with a prescribed criterion that only depends on the ratio of the likelihood values. The denominator $\mathbb{P}(h|\mathcal{M})$ in (4) is identical for any parameter set $\vec{\theta}$ and cancels during the calculation of the ratio.

We benchmarked our approach on two synthetic data sets that were generated by sampling ($N = 100$ or 1000 cells per time point for a total of 4 time points) from a two-allele, 3-state induction model, where the induction stimulus decreased the downward transition rates; see Methods. Here, the model was known *a priori* and our goal was to infer the kinetic parameters and perform uncertainty quantification by comparing their posterior distributions [Figs. 4(a) and 4(b)] to the ground truth (GT) parameters used to generate the sampled synthetic data set (Fig. S3 of the supplementary material). Our method constrained the posterior parameter distribution around the ground truth, and a 10-fold increase in the number of sampled cells dramatically reduced uncertainty in the inferred parameters. This observation holds true for a synthetic data set generated by a different two-allele, 3-state induction model; see Fig. S4 of the supplementary material.

Fitted models in systems biology often exhibit "sloppiness," where the goodness of the fit remains unchanged when several parameters are perturbed in a coordinated direction. Such directions in the parameter space, called *eigenparameters*, are the principle components of the likelihood function in the high-dimensional parameter space.[34] A common way to visualize the eigenparameters is to project the high-dimensional posteriors to the subspace spanned by any of the two bare parameters;[34,35] see Figs. S5 and S6 of the supplementary material. For example, our results show that simultaneously increasing the ON rate and OFF rate (and, thus, leaving mean transcript levels unchanged) results in a similar goodness of the fit. We also show that the posterior distribution is far from the asymptotic Gaussian limit, even when the number of samples N per time point is as large as 10^3. In this non-Gaussian regime, it is necessary to consider the full posterior distributions for parameter uncertainty quantification, in contrast to heuristic approaches that consider only the covariance matrix of the posterior chain.[36]

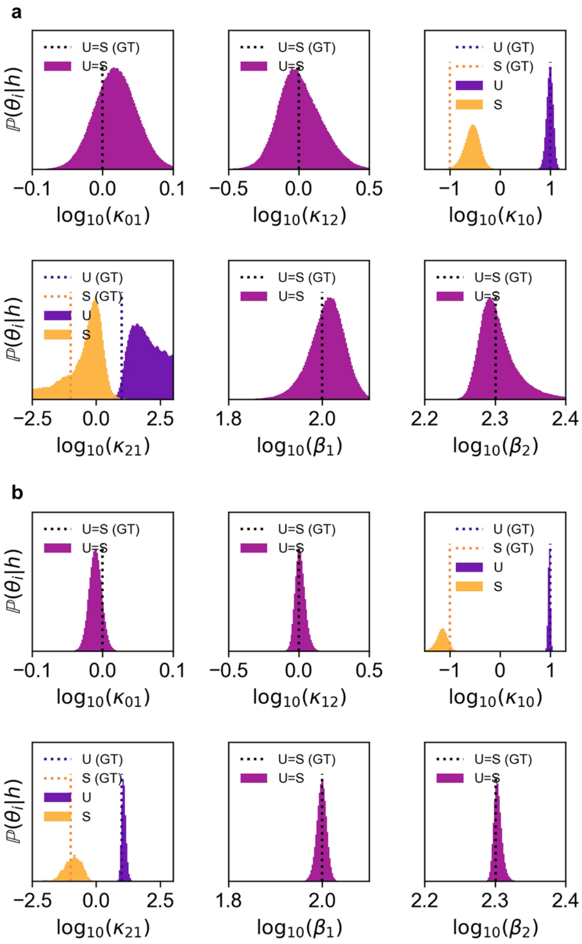

FIG. 4. Parameter inference and uncertainty quantification using the Bayesian posterior distribution. We benchmarked the hybrid method by running Bayesian inference on a synthetic data set sampled (*N* cells at 4 different time points) from a known model $\mathcal{M}$ = 00110000 and "ground-truth" (GT) parameter set. Posterior distributions for (a) *N* = 100 and (b) 1000 cells per time point. These posterior distributions were projected into six, separate one-dimensional parameter spaces. However, projections of the posterior into two-dimensional parameter spaces are useful because they illustrate sloppy modes in the parameter fitting; see Figs. S5 and S6 of the supplementary material. The joint probability distributions corresponding to the best-fit parameters are shown in Fig. S3 of the supplementary material.

V. MODEL SELECTION USING THE FULL BAYESIAN FRAMEWORK

Knowing that our method of accelerated Bayesian inference can reliably constrain the kinetic parameters for a given model, we turned our attention to the harder problem of model selection. The goal was to identify the correct model from 64 possible types of two-allele, 3-state induction models given the same synthetic data set in Fig. 4, which was sampled from a ground-truth model and its parameters. We reduced the number of candidate models from 256 to 64 by keeping β_0 and δ constant upon induction, i.e., always 0 in the binary notation such that $\mathcal{M}$ = xxxx0xx0. Our choice of noninducible parameters stems from our previous work that used a Bayesian approach to infer models of gene expression in stimulated neurons.[15,16] It was known that the mRNA degradation rate (δ) did not change, and our analysis showed that the inferred basal transcription rate (β_0) did not change upon stimulation. We therefore chose to keep these parameters unchanging to mimic our previous case study. Bayesian analysis naturally provides a quantitative measure of the likelihood of any model $\mathcal{M}$, i.e., the probability of the model to reproduce the experimentally observed data *h*. The measure, referred to as the marginalized likelihood or *evidence*,[37,38] is the denominator of (4),

$$\mathbb{P}(h|\mathcal{M}) = \int \underbrace{\mathbb{P}(h|\vec{\theta},\mathcal{M})}_{\mathcal{L}}\,\underbrace{\mathbb{P}(\vec{\theta}|\mathcal{M})}_{\text{Prior}}\,d\vec{\theta}. \qquad (5)$$

The evidence is simply the probability that a model $\mathcal{M}$ produced data *h* and is equal to the sum of the probabilities of the model (i.e., likelihood) over all sets of parameters that could have produced the data. The evidence is a convolution of the likelihood with the prior $\mathbb{P}(\vec{\theta}|\mathcal{M})$, which quantifies the belief regarding the initial parameter distributions. The dimensionality of $\vec{\theta}$ does not have to be identical for two different models, and this prior inherently penalizes models with too many parameters; see Sec. VI. The complexity of each model $\mathcal{M}$ increases with the total number of 1's in the binary notation because there will be two values (before induction and after induction) to be inferred for each inducible parameter.

The evidence for a model $\mathcal{M}$ is not calculated during the MCMC sampling of the posterior distribution and has to be computed separately. Computing the evidence is a sophisticated problem,[39–42] and we adopted an Importance Sampler of the Harmonic Mean Estimator (IS-HME) proposed by Robert and Wraith,[43] which resamples the posterior distribution estimated by the MCMC to compute the evidence of each model; see Appendix. We first carried out the MCMC calculations of posterior distributions for each of the 64 possible types of two-allele, 3-state induction models for the data sets described in Fig. 4 and Fig. S4 of the supplementary material. We then used IS-HME to compute the evidence of each model given the underlying data set. We compared the IS-HME evidence to maximum likelihood metrics used for model selection, such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).[15] Both BIC and AIC are approximations to the Bayesian evidence and become equivalent in the limit of large sample sizes; see Sec. VI.

Our results demonstrate that the IS-HME evidence $\mathbb{P}(h|\mathcal{M})$ of the ground truth model dominates over other models (≥95%) when using Bayesian inference on the larger data set (*N* = 1000 cells sampled per time point); see Fig. 5. The BIC approximation also selected the ground-truth model (although incorrect models exhibited significant probabilities, e.g., >5%), whereas the AIC failed to select the correct model. When the sample size dropped to *N* = 100 cells per time point, even IS-HME evidence could not reliably select the ground-truth model with this underpowered data set.

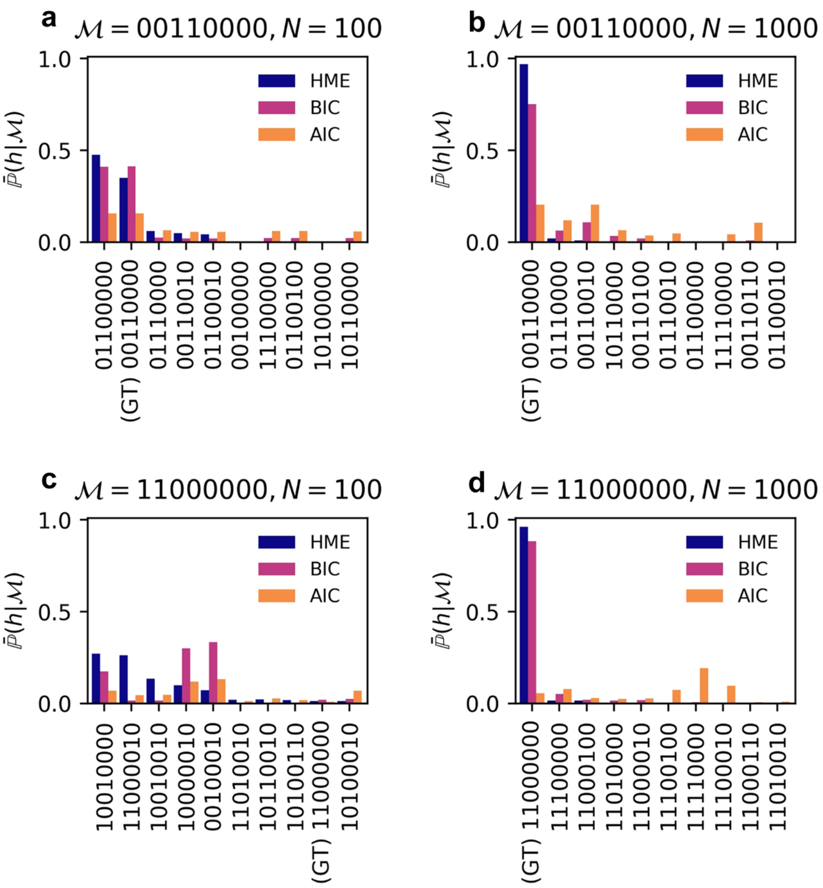

FIG. 5. Model selection using Bayesian evidence. We plot the IS-HME, BIC, and AIC evidence metrics of the top 10 models, ordered by decreasing the IS-HME score. Model selection was performed on data sampled at two densities (N = 100 and 1000 cells per time point at 4 different time points) for two different ground-truth models ($\mathcal{M}$ = 0011000 and $\mathcal{M}$ = 1100000). The parameter posterior distributions of the top 10 models for the ground-truth model ($\mathcal{M}$ = 0011000 and $\mathcal{M}$ = 11000000) are presented in Figs. S7–S10 of the supplementary material.

VI. DISCUSSION

Piecewise-deterministic Markov processes (PDMPs) have become a useful, coarse-grained description of stochastic gene dynamics, where the underlying discrete variable $s(t)$ captures the stochastic dynamics of gene states and the continuous variable $\lambda(t)$ captures the first moment of downstream gene products.[44–50] The key insight of our manuscript was proving that the time-dependent mRNA *distribution* of any underlying $s(t)$ is asymptotically a Poisson distribution with a rate $\lambda(t)$ and that the time-dependent joint probability distributions of discrete-state models are dynamic Poisson mixtures, whose mixing kernels are characterized by a PDMP. This significantly expands upon a related framework, which only considered the stationary distribution of discrete-state models.[51] More generally, our analysis helps bridge a gap between mechanistic discrete-state models and statistical models used in single cell analysis. For example, Wang *et al.* recently proposed a statistical model of gene expression which postulated that mRNA distributions are Poisson mixtures,[52] and our work justifies this assumption.

We used our insight to develop a hybrid method that calculates the time-dependent joint distribution more efficiently than standard numerical methods that forward-integrate the Chemical Master Equation (CME). The efficiency arises because our method analytically solves the mRNA distribution and rapidly samples many path $s(t)$ of discrete-switching events using a kinetic Monte Carlo algorithm. We benchmarked the hybrid method and showed that it is $\mathcal{O}(10^3)$ more efficient than previous methods that directly integrate the CME. Furthermore, the hybrid method runs efficiently in parallel on a multicore processor than it does on a single processor. The stiff CME integrators ran more slowly in parallel, and this sublinear scaling persisted for different integrators. We suspect that the slowdown arises from the competing memory demands of stiff CME integrators running on a multicore processor. While there is room to improve stiff integration and parallelization, we note that current approaches are fundamentally limited compared to the hybrid method because they must integrate the CME for a large number of mRNA states, e.g., 0–1000 mRNAs per cell.

We incorporated the hybrid algorithm into BayFISH and were able, for the first time, to use a full Bayesian framework for model selection and uncertainty quantification of parameters from single-cell smFISH data. We adopted the Bayesian framework for model selection because it naturally quantifies "Occam's factor"[37,40] and, thus, avoids overfitting. For example, the top models based on Bayesian evidence are not the most complex models with the largest number of parameters that change upon induction, e.g., $\mathcal{M}$ = 11110110; see Fig. 5. The evidence resists overfitting because when the dimensionality of parameter space is high, the value of a uniform probability density of the prior parameter distribution $\mathbb{P}(\vec{\theta}|\mathcal{M})$ in (5) is small due to normalization. Thus, Bayesian evidence will favor a model that is complex enough to have a large likelihood but not too complex to decrease the prior parameter density.

We note that when the data sample size is large such that the posterior distributions $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$ can be approximated by a multivariate normal distribution, the logarithm of the evidence converges asymptotically to the Schwarz index (commonly known as the Bayesian Information Criterion, BIC).[38,40] A similar asymptotic criterion is the Akaike Information Criterion (AIC),[53] which aims to minimize the information loss measured by the Kullback–Leibler divergence of the theoretically predicted joint probability distribution from the sampled distribution. Our results show that the posterior distribution estimated from modest data sets can deviate from multivariate normal distributions (see Figs. S5 and S6 of the supplementary material), which suggests that AIC and BIC can underperform in model selection, relative to Bayesian evidence. Here, we benchmarked the ability of Bayesian evidence, BIC and AIC metrics, to select the correct model from synthetic data sets generated by a ground-truth model and parameters. Figure 5 shows that while the BIC (but not AIC) ranked models similar to Bayesian evidence for the larger data set (N = 1000 cells per time point), BIC requires an even larger sample size to confidently converge to the correct model. This is an important issue because most biology labs are ill-equipped to generate and analyze large smFISH data sets, and their sample sizes are typically N = 100–1000 cells per time point. Our work demonstrates why Bayesian inference should be used for modestly sampled data sets. We show that N = 100 cells per time point is sufficient for parameter inference and uncertainty quantification if one has high confidence in the underlying model; see Fig. 4. However, if the goal of the smFISH experiments is model selection, then these smaller data sets are underpowered and the experimentalist needs to increase data sampling by at least 10-fold; see Figs. 5(a) and 5(c). Here, we only considered one round of experiments followed by Bayesian inference, but multiple cycles of data collection and analysis are becoming the norm. Our framework quantifies certainty in both models and parameters using Bayesian evidence and posterior distributions. Future work can complete the data collection and analysis cycle by using the evidence and posterior distributions to rationally dictate the next round of experiments,[54] i.e., different sampling times and densities, which are most informative for constraining models and/or parameters.

In this article, we adopted a Markov chain Monte Carlo algorithm with Metropolis-Hastings sampling (i.e., MCMC-MH) to compute the posterior distributions of the model parameters.[15,16,36] However, there is room to further improve the speed of Bayesian inference. First, Hamiltonian Monte Carlo (HMC) algorithms are more efficient at sampling posterior distributions in high dimensional parameter spaces because they use local sensitivity, i.e., the partial derivatives of the likelihoods with respect to the model parameters.[55–57] Second, although PM-PDMSR is efficient at generating sample paths in the space (s, λ), evaluating the convolution to calculate the joint distribution (3) is the rate-limiting step in the likelihood calculation. Thus, transforming the experimental data h into the mixing kernel of the Poisson mixtures $\rho_s(\lambda)$ would accelerate Bayesian inference. Third, one could use low-order moments of PM-PDMSR and experimental data to formulate a sufficient statistics for likelihood-free approximate Bayesian computation,[35] thus replacing the explicit calculation of the likelihood $\mathcal{L}$. Finally, our proposed PM-PDMSR provides an order-of-magnitudes more efficient algorithm to evaluate the likelihoods compared to CME calculations. In this article, we demonstrated that such an efficiency gain makes expensive Bayesian calculations feasible. If one has a large enough data set such that the posterior distribution is in the Gaussian limit (e.g., $N = 10^4$ cells per time points), then model selection could be achieved by the asymptotic BIC, which only needs the maximum likelihood. In this regime, however, PM-PDMSR is still $\mathcal{O}(10^3)$ more efficient and scalable at estimating the maximum likelihood of complex models when compared to CME methods.

SUPPLEMENTARY MATERIAL

See supplementary material figures for Figs. S1–S10. Figure S1—Accuracy of PM-PDMSR: The likelihood calculated by using PM-PDMSR ($\mathcal{L}_{\text{PM-PDMSR}}$) is plotted against the likelihood calculated using CME integration ($\mathcal{L}_{\text{CME}}$). The summary error $\langle \varepsilon \rangle$ is computed according to Eq. (A23) in the manuscript. Figure S2—Scaling analysis of CME and PM-PDMSR runtimes: We parallelized both CME and PM-PDMSR using Python's `multiprocessing` module. Simultaneously, $\{1, 2, 4, 8, 16, 32\}$ cores are utilized to process the same batch (1024) of synthetic data for each of the 2-, 3-, and 4-state models described in Appendix 2a. We report the average time per thread to process each data set (panels A and C) and the total time of all threads to process the entire batch (panels B and D). PM-PDMSR described in Fig. 3 is parallel (total time ~ 1/number of cores utilized simultaneously), whereas the CME suffers from a stiff scaling such that multiprocessing on a computer—even if it is equipped with multiple CPUs—is not significant more efficient than running a single thread. The machine we used for this benchmarking is equipped with 32 cores and can process simultaneously 64 threads (two Intel© Xeon© CPU E5-2698 v3 at 2.30 GHz). Figure S3—Joint probability distribution of the best-fit parameters: The joint distribution of the best-fit parameters to synthetic data sets with N = 100 (panel A) and 1000 (panel B) cells per time point, $\mathcal{M}$ = 00110000. Figure S4—Parameter inference and uncertainty quantification using the Bayesian posterior distribution: We benchmarked the hybrid method by running Bayesian inference on a synthetic data set sampled (N cells at 4 different time points) from a known model $\mathcal{M}$ = 11000000 and "ground-truth" (GT) parameter set. Posterior distributions (panels A and B) and joint distribution of best-fit parameters (panels C and D) for N = 100 and 1000 cells per time point, respectively. Figure S5—Two-dimensional projection of the posterior distribution (N = 1000): The posterior parameter distribution projected into two-dimensional parameter space for the ground truth model $\mathcal{M}$ = 00110000 and N = 1000 cells. Even with N = 1000, the posterior distribution in some of the dimensions is still far from the Gaussian asymptotic limit. Figure S6—Two-dimensional projection of the posterior distribution ($N = 100$): The posterior parameter distribution projected into two-dimensional parameter space for the ground truth model $\mathcal{M} = 00110000$ and $N = 100$ cells. The posterior distribution is far from the Gaussian asymptotic limit. Figure S7—Posteriors of top 10 models ($\mathcal{M} = 00110000$ and $N = 100$): The posterior parameter distribution of the top 10 performing models in Fig. 5(a) inferred for $N = 100$ cell data set, ordered by the evidence calculated from IS-HME (top: best-performing model). The ground truth model $\mathcal{M} = 00110000$ was ranked as the second best explanatory model for this synthetic data set (see Fig. S3A of the supplementary material). Figure S8—Posteriors of top 10 models ($\mathcal{M} = 00110000$ and $N = 1000$): The posterior parameter distribution of the top 10 performing models in Fig. 5(b) inferred for $N = 1000$ cell data set, ordered by the evidence calculated from IS-HME (top: best-performing model). The ground truth model $\mathcal{M} = 00110000$ was ranked as the best explanatory model for this synthetic data set (see Fig. S3B of the supplementary material). Figure S9—Posteriors of top 10 models ($\mathcal{M} = 11000000$ and $N = 100$): The posterior parameter distribution of the top 10 performing models in Fig. 5(c) inferred for $N = 100$ cell data set, ordered by the evidence calculated from IS-HME (top: best-performing model). The ground truth model $\mathcal{M} = 11000000$ was ranked as the best explanatory model for this synthetic data set (see Fig. S4C of the supplementary material). Figure S10—Posteriors of top 10 models ($\mathcal{M} = 11000000$ and $N = 1000$): The posterior parameter distribution of the top 10 performing models in Fig. 5(d) inferred for $N = 1000$ cell data set, ordered by the evidence calculated from IS-HME (top: best-performing model). The ground truth model $\mathcal{M} = 11000000$ was ranked as the best explanatory model for this synthetic data set (see Fig. S4D of the supplementary material).

ACKNOWLEDGMENTS

Y.T.L. was supported by the Center for Nonlinear Studies, Los Alamos National Laboratory. N.E.B. was supported by a grant from the National Institutes of Health (Grant No. 1R01GM127614-01). The authors thank Nicolas Hengartner, Marian Anghel, and Ben Callahan for informative discussions and critical comments.

APPENDIX: THEORETICAL RESULTS, NUMERICAL SIMULATIONS, AND METHODS OF STATISTICAL INFERENCE

1. Poisson mixture with piecewise deterministic Markov switching rates

We illustrate our central theoretical results using a single-allele model. However, the results generalize to multiple-allele models because the states of a multiple-allele model can be relabeled as internal states of a single-allele model.

a. Central theoretical result I

Given a trajectory of genetic state $s(t)$, the total number of mRNA, $N_{\text{mRNA}}(t)$, is the sum of two variables $N^{\text{initial}}_{\text{mRNA}}(t)$ and $N^{\text{new}}_{\text{mRNA}}(t)$. $N^{\text{initial}}_{\text{mRNA}}(t)$ describes the number of initial mRNAs that remain at time t. The probability distribution of $N^{\text{initial}}_{\text{mRNA}}(t)$ is a binomial mixture weighted by the initial mRNA distribution $\mathbb{P}_{m,0} := \mathbb{P}\left(N^{\text{initial}}_{\text{mRNA}}(t=0) = n\right)$,

$$\mathbb{P}\left(N^{\text{initial}}_{\text{mRNA}}(t) = n\right) = \sum_{m=0}^{\infty} \mathbb{P}_{m,0} \binom{m}{n} \left(1 - e^{-\delta t}\right)^{m-n} e^{-n\delta t} \Theta\left(m - n + \frac{1}{2}\right). \tag{A1}$$

Here, $\Theta(\cdot)$ is the Heaviside step function. $N^{\text{new}}_{\text{mRNA}}(t)$ describes the number of new mRNAs that were synthesized after $t > 0$ and still remain at time t. The probability distribution of $N^{\text{new}}_{\text{mRNA}}(t)$ is a Poisson distribution with rate $\lambda(t)$,

$$\mathbb{P}\left(N^{\text{new}}_{\text{mRNA}}(t) = m\right) = \frac{\lambda^m(t) e^{-\lambda(t)}}{m!}, \tag{A2}$$

where $\lambda(t)$ satisfies equation $\dot{\lambda}(t) = \beta_{s(t)} - \delta\lambda(t)$ with an initial condition $\lambda(0) = 0$.

Proof We denote the probability of the total number of mRNA N_{mRNA} at time t by $\mathbb{P}_m(t)$. For a given trajectory of the genetic state $s(t)$, the temporal evolution satisfies the chemical master equation (CME)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}_m(t) = -\left(\beta_{s(t)+\delta m}\right)\mathbb{P}_m(t) + \beta_{s(t)}\mathbb{P}_{m-1}(t) + \delta(m+1)\mathbb{P}_{m+1}(t), \tag{A3}$$

for all $m \in \mathbb{Z}_{\geq 0}$ and with a boundary condition $\mathbb{P}_{-1} = 0$. We prove central theoretical result I by using the probability generating function defined by

$$\mathcal{G}(z,t) := \sum_{m=0}^{\infty} z^m \mathbb{P}_m(t), \tag{A4}$$

where $\mathbb{P}_m(t) = \partial_z^m \mathcal{G}(z,t)/m!$. We first solve for $\mathcal{G}(z,t)$ over an interval of time when $s(t)$ is constant. We will then extend our analysis to include piecewise intervals of time with different value of constant s, similar to $s(t)$ generated by a genetic state model. To begin, we apply the operator ∂_t to $\mathcal{G}(z,t)$ and use (A3) to obtain the partial differential equation (PDE),

$$\partial_t \mathcal{G}(z,t) = \delta(1-z)\partial_z \mathcal{G}(z,t) + \beta_s(z-1)\mathcal{G}(z,t). \tag{A5}$$

This linear PDE can be solved using the method of characteristics,[58] and the general solution is

$$\mathcal{G}(z,t) = \left[\sum_{m=0}^{\infty}\left(1 + (z-1)e^{-\delta t}\right)^m \mathbb{P}_{m,0}\right] \exp\left(\frac{\beta_s}{\delta}(z-1)\left(1 - e^{-\delta t}\right)\right), \tag{A6}$$

where $\mathbb{P}_{m,0} := \mathbb{P}_{m,0}(t=0)$ is the initial mRNA distribution of the system. For reasons that will become apparent below, we label the initial-distribution-dependent part by $\mathcal{G}^{\text{init}}(z,t)$ and the rest of the terms by $\mathcal{G}^{\text{new}}(z,t)$,

$$\mathcal{G}^{\text{init}}(z,t) = \sum_{m=0}^{\infty}\left(1 + (z-1)e^{-\delta t}\right)^m \mathbb{P}_{m,0}, \tag{A7a}$$

$$\mathcal{G}^{\text{new}}(z,t) = \exp\left(\frac{\beta_s}{\delta}(z-1)\left(1 - e^{-\delta t}\right)\right). \tag{A7b}$$

The above solution applies to a constant $s(t)$. We now consider a piecewise-constant trajectory for any genetic state $s(t)$. To specify the discrete state and the switching events, we label $s(t)$ by the ordered pairs (t_i, s_i) for $i = 0, 1, \ldots, N$ before an observation time t. The gene starts with a state s_0 at $t_0 := 0$, switches to s_1 at t_1, etc., until the final switching event to s_N at time t_N. Our aim is to compute the generating function $\mathcal{G}(z, t)$ after N switching events $(t \geq t_N)$.

The solution $\mathcal{G}(z, t, |t \leq t_1)$ is identical to (A6) with $s = s_0$ before the first switching event. At $t = t_1$, the generating function is

$$\mathcal{G}(z, t_1) = \mathcal{G}^{\text{init}}(z, t_1) \times \mathcal{G}^{\text{new}}(z, t_1). \tag{A8}$$

Note that after t_1 and before t_2, the genetic state changes to s_1 and only the transcription rate in (A3) changes from β_{s_0} to β_{s_1}. The initial condition of the generating function of this period $(t_1 \leq t \leq t_2)$ is precisely $\mathcal{G}(z, t_1)$ in the above equation. Matching the "initial condition" for $\mathcal{G}(z, t_1)$, we arrive at

$$\begin{aligned}\mathcal{G}(z, t|t_1 \leq t \leq t_2) = \mathcal{G}^{\text{init}}(z, t) \cdot \exp\Bigg\{\frac{z-1}{\delta}\Big[\beta_{s_0}\left(1 - e^{-\delta t_1}\right)e^{-\delta(t-t_1)} \\ + \beta_{s_1}\left(1 - e^{-\delta(t-t_1)}\right)\Big]\Bigg\}.\end{aligned} \tag{A9}$$

We iterate this procedure for each piecewise "episode" until $t = t_N$,

$$\begin{aligned}\mathcal{G}(z, t_N) &= \mathcal{G}^{\text{init}}(z, t_N) \\ &\times \exp\left[\frac{z-1}{\delta}\sum_{n=1}^{N}\beta_{s_{n-1}}\left(1 - e^{-\delta(t_n - t_{n-1})}\right)e^{-\delta(t_N - t_n)}\right],\end{aligned} \tag{A10}$$

and for $t \geq t_N$,

$$\begin{aligned}\mathcal{G}(z, t|t \geq t_N) &= \mathcal{G}^{\text{init}}(z, t) \times \mathcal{G}^{\text{new}}(z, t), \\ \mathcal{G}^{\text{new}}(z, t) &:= \exp\left[\psi\left(\{t_n\}_{n=1}^{N}\right)\right], \\ \psi\left(\{t_n\}_{n=1}^{N}\right) &:= \frac{z-1}{\delta}\Bigg[\sum_{n=1}^{N}\beta_{s_{n-1}}\left(1 - e^{-\delta(t_n - t_{n-1})}\right) \\ &\quad \times e^{-\delta(t - t_N)} + \beta_{s_N}\left(1 - e^{-\delta(t-t_N)}\right)\Bigg].\end{aligned} \tag{A11}$$

The total solution $\mathcal{G}(z, t|t \geq t_N)$ is factorized into two terms, $\mathcal{G}^{\text{init}}(z, t)$ and $\mathcal{G}^{\text{new}}(z, t)$, for any N and t. The probability generating function of the sum of two independent random variables is the product of the generating functions of the random variables. This hints that we can define two random variables, X_1 and X_2, which have generating functions $\mathcal{G}^{\text{init}}(z, t)$ and $\mathcal{G}^{\text{new}}(z, t)$, respectively.

Our next task is to identify variables X_1 and X_2 and their probability distributions. For X_1, we expand $\mathcal{G}^{\text{init}}(z, t)$ to arrive at

$$\mathcal{G}^{\text{init}}(z, t) \equiv \sum_{m=0}^{\infty}\mathbb{P}_{m,0}\sum_{n=0}^{\infty}\binom{m}{n}\left(1 - e^{-\delta t}\right)^{m-n}z^n e^{-n\delta t}. \tag{A12}$$

Recall that the generating function of a Binomial(m, p) distribution is

$$[1 - p + pz]^m = \sum_{n=0}^{\infty}\binom{m}{n}(1-p)^{m-n}z^n p^n. \tag{A13}$$

The probability distribution of X_1 is therefore identified to be a binomial mixture with a temporally decaying parameter $p = \exp(-\delta t)$ and a mixing kernel defined by the initial distribution $\mathbb{P}_{m,0}$. The physical meaning of $X_1(t)$ is the number of initial mRNA molecules that remain at time t, i.e., $N_{\text{mRNA}}^{\text{initial}}(t)$. These mRNA molecules can only degrade with the decay rate δ. Each of the mRNA decays independently and, at time t, there is a probability $\exp(-\delta t)$ that a specific mRNA has not degraded. Importantly, when $t \gg 1/\delta$, this distribution will be concentrated at $m = 0$ (see Corollary I).

The total mRNA is $N(t) = X_1(t) + X_2(t) = N_{\text{mRNA}}^{\text{initial}}(t) + X_2(t)$, so $X_2(t)$ is identified to be the number of new mRNA molecules that were synthesized *after* $t = 0$ but which have not degraded at time t. We refer to this variable as $N_{\text{mRNA}}^{\text{new}}(t)$. The square bracket of $\mathcal{G}^{\text{new}}(z, t)$ in (A11) is the piecewise solution $\lambda(t)$ of the following ODE for a given genetic trajectory $s(t)$:

$$\frac{\text{d}}{\text{d}t}\lambda(t) = \beta_{s(t)} - \delta\lambda(t), \quad \text{and } \lambda(0) = 0. \tag{A14}$$

We now expand $\mathcal{G}^{\text{new}}(z, t)$,

$$\begin{aligned}\mathcal{G}^{\text{new}}(z, t) = \exp[(z-1)\lambda(t)] &= e^{-\lambda(t)}\sum_{m=0}^{\infty}\frac{z^m\lambda^m(t)}{m!} \\ &= \sum_{m=0}^{\infty}z^m\frac{e^{-\lambda(t)}\lambda^m(t)}{m!} = \sum_{m=0}^{\infty}z^m q_m(\lambda(t)),\end{aligned} \tag{A15}$$

where $q_m(\lambda(t))$ is the probability density function of a Poisson distribution with rate $\lambda(t)$, as in (A2). □

Corollary I. The transient time scale for the initial distribution is $\mathcal{O}(1/\delta)$. When $t \gg \mathcal{O}(1/\delta)$, the mRNA distribution converges to a Poisson with a dynamic rate parameter $\lambda(t)$.

Proof Physically, the time scale of degradation of each initially populated mRNA is $1/\delta$, so for a time scale which is much longer than this, the initial distribution will be fully degraded. Mathematically, the probability that the initial mRNA molecules have not fully decayed is

$$\mathbb{P}\left(N_{\text{mRNA}}^{\text{initial}}(t) > 0\right) = 1 - \mathbb{P}\left(N_{\text{mRNA}}^{\text{initial}}(t) = 0\right)1 - \sum_{m=1}^{\infty}\mathbb{P}_{m,0}\left(1 - e^{-\delta t}\right)^m. \tag{A16}$$

In the asymptotic limit $t \gg 1/\delta$, $\exp(-\delta t) \ll 1$ so

$$\begin{aligned}\mathbb{P}\left(N_{\text{mRNA}}^{\text{initial}}(t) > 0\right) &= 1 - \sum_{m=0}^{\infty}\mathbb{P}_{m,0}\left(1 - me^{-\delta t}\right)\left[1 + \mathcal{O}\left(e^{-\delta t}\right)\right] \\ &= \left\langle N_{\text{mRNA}}^{\text{initial}}(0)\right\rangle e^{-\delta t}\left[1 + \mathcal{O}\left(e^{-\delta t}\right)\right].\end{aligned} \tag{A17}$$

Here, $\left\langle N_{\text{mRNA}}^{\text{initial}}(0)\right\rangle$ is the first moment of the initial distribution. $\mathbb{P}\left(N_{\text{mRNA}}^{\text{initial}}(t) > 0\right)$ decays exponentially fast, and we can bind this probability to be smaller than ε when $t > \delta^{-1}\log\left(\left\langle N_{\text{mRNA}}^{\text{initial}}(0)\right\rangle/\varepsilon\right)$. □

b. Central theoretical result II

At long times $t \gg 1/\delta$, the mRNA distribution asymptotically converges to a Poisson mixture regardless of the initial mRNA distribution and genetic switching trajectory $s(t)$,

$$\mathbb{P}(N_{\text{mRNA}}(t) = m, s(t) = i) = \int_0^\infty \rho_i(\lambda, t) \frac{\lambda^m e^{-\lambda}}{m!} \mathrm{d}\lambda, \tag{A18}$$

where the joint probability density $\rho_i(\lambda, t)$ satisfies the forward Kolmogorov equation

$$\partial_t \rho_i = -\partial_\lambda [(\beta_i - \delta\lambda)\rho_i] + \sum_{j \neq i} (\kappa_{ji}\rho_j - \kappa_{ij}\rho_i). \tag{A19}$$

The initial condition for $\rho_i(\lambda, t = 0)$ is defined by

$$\rho_i(\lambda, t = 0) := \delta(\lambda) \mathbb{P}(s(t = 0) = i), \tag{A20}$$

where $\delta(\lambda)$ is the Dirac delta distribution at $\lambda = 0$.

Proof The solution of (A14) subject to random switching events of $s(t)$ is a random process. Formally, $\lambda(t)$ *and* the discrete switching states $s(t)$ jointly comprise a piecewise deterministic Markov process (PDMP).[59–61] The forward Kolmogorov equation describing the temporal evolution of the joint probability distribution is (A19).[47,49] Therefore,

$$\mathbb{P}(N_{\text{mRNA}}(t) = m, s(t) = i) \\ = \int_0^\infty \mathbb{P}(N_{\text{mRNA}}(t) = m | \lambda(t) = \ell, s(t) = i) \rho_i(\ell, t) \mathrm{d}\ell. \tag{A21}$$

We then use central theoretical result I and corollary I to show that $\mathbb{P}(N_{\text{mRNA}}(t) = m|\lambda(t) = \ell, s(t) = i) = \ell^m e^{-\ell}/m!$ asymptotically when $t \gg 1/\delta$ to complete the proof. □

c. Efficient numerical method for sampling $\rho_i(\lambda, t)$

Because λ is continuous, solving the forward Kolmogorov equation (A19) is as complex as solving the full CME, both of which are infinite dimensional systems. Instead, we used an efficient kinetic Monte Carlo simulation[44,49] to generate a large number of sample paths to estimate the asymptotic joint distribution $\mathbb{P}(N_{\text{mRNA}}(t) = m|\lambda(t) = \ell, s(t) = i)$ when $t \gg 1/\delta$ using (A18). The pseudocode of this algorithm is shown in Algorithm 1.

ALGORITHM 1. An efficient kinetic Monte Carlo algorithm which generates exact sample paths of the piecewise deterministic Markov process $[\lambda(t), s(t)]$.

```
Require: Initial state $\lambda(t = 0) = 0$ and $s(t) = s_0$. Kinetic rate $\kappa_{ij}$ (switching rates from discrete state $i \rightarrow j$), $\beta_k$ (transcription rates),
    and $\delta$ (degradation rate). $N$ discrete observation times $\mathcal{T} := \{t_\ell\}_{\ell=1}^{N}$.
Ensure: An exact sample path of the random process $(\lambda(t), s(t))$ at $N$ discrete times $\mathcal{T}$.
$t \leftarrow 0, \lambda \leftarrow 0, s \leftarrow s_0$                                        ▷ Initiate system time and state
for $t_{\text{observation}}$ in $\mathcal{T}$ do
    while $t < t_{\text{observation}}$ do
        $\kappa \leftarrow \sum_i \kappa_{si}$                                   ▷ Compute the total propensity of switching
        $u \leftarrow \text{Unif}(0, 1)$
        $\Delta t \leftarrow -\kappa^{-1} \log(u)$                                 ▷ Sample the random advanced time
        if $t + \Delta_t < t_{\text{observation}}$ then                  ▷ A switching event occurs before $t_{\text{observation}}$
            $c_0 \leftarrow 0, c_i \leftarrow \sum_{j-1}^{i} \kappa_{sj}$ for $i \in \{1, 2, \ldots S\}$                ▷ Sample the switching events
            $k \leftarrow 0$
            $w \leftarrow c_S \times \text{Unif}(0, 1)$
            while $w > \kappa_{c_k}$ do
                $k \leftarrow k + 1$
            end while
            $\lambda \leftarrow \beta_s/\delta + (\lambda - \beta_s/\delta) \exp(-\delta\Delta t), s \leftarrow k$        ▷ Update system state
        else                                          ▷ No switching event occurs before $t_{\text{observation}}$
            $\Delta t \leftarrow t_{\text{observation}} - t$
            $\lambda \leftarrow \beta_s/\delta + (\lambda - \beta_s/\delta) \exp(-\delta\Delta t)$                ▷ Update system state
        end if
        $t \leftarrow t + \Delta t$                                               ▷ Update system time
    end while
    Output the system state $(\lambda, s)$ at the observation time $t_{\text{observation}}$
end for
```

d. Computing the joint distribution from sample paths

To compute the time-dependent joint distribution of genetic states and mRNAs, we generated N_s sample paths $\{\lambda_k(t), s_k(t)\}_{k=1}^{N_s}$ with Algorithm 1. We then used (A18) to estimate $\rho_i(\lambda)$,

$$\widehat{\mathbb{P}}(N_{\text{mRNA}}(t) = m, s(t) = i) = \frac{1}{N_s}\sum_{k=1}^{N_s}\delta_{i,s_k(t)}\frac{\lambda_k^m(t)e^{-\lambda_k(t)}}{m!}, \tag{A22}$$

where $\delta_{i,j}$ is the Kronecker delta function which is equal to 1 if $i = j$ and 0 otherwise. We determined that $N_s \equiv 10^5$ is a sufficient number of sample paths to estimate the same joint distribution obtained by forward-integrating the CME. We refer to our method as the Poisson Mixture with a Piecewise Deterministic Markov Switching Rate (PM-PDMSR).

The goal was to compute the joint distribution of genetic states and mRNAs before and after induction. Similar to the situation in many experimental systems,[62,63] the joint distribution is at stationarity before induction. Upon induction, we assume that some model parameters are changed, which results in the time-evolution of the joint probability distribution $\mathbb{P}(N_{\text{mRNA}}(t) = m, s(t) = i)$ toward a new stationary state. We label the kinetic rates (κ_{ij} and β_k) before and after induction by U (Unstimulated) and S (Stimulated). To use PM-PDMSR to estimate the stationary distribution before induction, we first solved for the marginal stationary distribution of the genetic state p_i^*, where $0 = \sum_j(\kappa_{ij}^U - \kappa_{ji}^U)p_i^*$, $i = 1, 2, \ldots, S$ for an S-state model. We initiated $N_s p_i^*$ sample paths in PM-PDMSR at $\lambda = \beta_i/\delta$ and state $s = i$ and ran for $t = 10/\delta$ so that the Poisson mixture relaxes to stationarity. Upon induction at $t = 10/\delta$, we changed model parameters $\kappa_{ij}^U \to \kappa_{ij}^S$ and $\beta_k^U \to \beta_k^S$ for $i, j, k \in \{1, 2, \ldots S\}$ and continued simulating the temporal evolution of the joint probability distribution after induction using PM-PDMSR. This is valid because our previous proofs and arguments for (A18) apply even when the kinetic rate constants change upon induction.

2. Testing the speed and accuracy of the simulators

We benchmarked the efficiency of PM-PDMSR vs the "gold-standard" simulator, i.e., forward-integration of a truncated CME.[22,23] Both simulators were embedded into BayFISH and evaluated on their ability to perform a single Monte Carlo step, i.e., simulate the time-dependent joint distribution of a model and its parameters, and to calculate the likelihood of a synthetic data set generated by the same model and parameters. This single-step benchmarking was performed for 1024 diverse models and parameters across two-allele, discrete-state models of increasing complexity (2-state, 3-state, and 4-state induction models).

a. Generating diverse models, parameters, and synthetic data sets

Each model had a randomly chosen subset of parameters that were one value $(\kappa_{ij}^U, \beta_k^U)$ before induction from $t = 0$ to $t = 20$ and with different parameters $(\kappa_{ij}^S, \beta_k^S)$ after induction from $t = 20$ to $t = 22$. The genetic switching rates of parameters κ_{ij} ranged between 10^{-2} and 10^2, and the transcription rates β_k range between 0 and 200. For each instance of a model, we randomly generated the switching rate constants in the logarithmic space ($\log_{10}\kappa_{ij} \sim \text{Unif}(-2, 2)$ if $|i - j| = 1$) and transcription rate constants in linear space ($\beta_k \sim \text{Unif}(0, 200)$). Note that not all parameters were random or changed upon induction: We fixed $\beta_0 = 0$ and the mRNA degradation rate $\delta = 1$, and we constrained $\beta_i \le \beta_j$ if $i < j$. For the purpose of benchmarking algorithm speed and efficiency, we considered a complex induction model $\mathcal{M}$ for each discrete model class. The corresponding induction model is $\mathcal{M} = 11010$ (2-state), $\mathcal{M} = 11110110$ (3-state), and $\mathcal{M} = 11111101110$ (4-state).

The synthetic data of each model with its parameters were generated by running 1000 independent trajectories of standard continuous-time Markov chain simulation. We collected the statistics of the trajectories at four discrete times $t_\ell = 20, 20.5, 21$, and 22 ($N = 1000$ cells per time point). The measured allele activity state TS was marginalized: we define $TS = 1$ when its internal state is $s > 0$. The synthetic data set therefore consists of a histogram at discrete times t_ℓ, $h(m, TS, t_\ell)$, which is the number of trajectories with m mRNAs and TS active transcription sites (which can be 0, 1, or 2 for a two-allele system) at time t_ℓ. For each model class (2-, 3-, and 4-state models), we repeated the process 1024 times to test diverse parameter combinations.

b. CME simulators

Given an induction model and associated parameters, we forward propagated the CME using the same parameter values used to create the synthetic data sets in Sec. IV. We truncated the number of mRNA at 500 (i.e., there is no transcription event once the system reaches $N_{\text{mRNA}} = 500$) with an absolute error tolerance of 10^{-5}. The truncation number M was motivated by data sets in animals, showing that mRNA populations can be as large as $\mathcal{O}(500)$.[24–27] We tested different software platforms, including Matlab, Python (SciPy), and a research software ACME,[23] to forward integrate the same stiff CME. Python's stiff integrator (using backward differentiation formula, BDF) outperformed other integrators and software platforms. Thus, Python (with SciPy) was chosen to be the platform for direct integration of the CME in the following analysis.

c. Comparison of the PM-PDMSR and CME simulators

We incorporated these PM-PDMSR and CME simulators into modified BayFISH software to evaluate their speed and accuracy for one Monte Carlo step. Both algorithms were implemented in c++ and compiled using Intel's icc compiler. All PM-PDMSR and CME simulations were carried out on the same machine with Intel© Xeon© CPU E5-2695 v3 at 2.30 GHz. We computed the joint distributions of the models and parameter sets used to create the synthetic data sets. These joint distributions and corresponding synthetic data ($h(m, TS, t_\ell)$) were then used to compute the likelihood $\mathcal{L}$ of the generated data $h(m, n_{TS}, t_\ell)$ using (1). The execution time of a Monte Carlo step for each simulator for each model class of the generated parameter set was recorded and presented in Fig. 3. We also compared the accuracy of the calculated likelihoods of each synthetic data set. We compared the average error of the PM-PDMSR likelihood relative to the more accurate CME likelihood. We define the average error by

$$\langle \varepsilon \rangle := \left\langle \left| \frac{\mathcal{L}_{\text{PM-PDMRS}} - \mathcal{L}_{\text{CME}}}{\mathcal{L}_{\text{CME}}} \right| \right\rangle. \tag{A23}$$

The relative accuracy of CME vs PM-PDMSR is presented in the supplementary material.

To test the parallelization of each simulator, we simultaneously ran 32 simulations on a 32-core machine (two Intel© Xeon© CPU E5-2698 v3 at 2.30 GHz) and recorded the execution time; see Fig. 3. PM-PDMSR on 32 parallel threads takes the same amount of time as running a single thread. By contrast, the CME on 32 parallel threads takes 32 times longer than a single thread; see Fig. S2 of the supplementary material. This suboptimal scaling holds true on the multiple machines that we tested, and our algorithms leverage Python's subprocesses functionality (Pool). We suspect that the slowdown in the CME is due to the high memory demand of the BDF integrator. Results of a more detailed scaling analysis with different numbers of parallel threads are presented in Fig. S2 of the supplementary material.

3. Bayesian statistical inference for model parameters and structure

a. Synthetic data

We synthesized data sets to test if Bayesian statistical inference could identify the ground truth (of the model parameter values and the model structure.) We chose two 3-state ground-truth models for data synthesis: (1) an ON-rate induction model, $\kappa_{01}^{U} = 1 \rightarrow \kappa_{01}^{S} = 12$, $\kappa_{12}^{U} = 0.25 \rightarrow \kappa_{12}^{S} = 20$, $\kappa_{10}^{U} = \kappa_{10}^{S} = 3$, $\kappa_{21}^{U} = \kappa_{21}^{S} = 10$, $\beta_0 = 0$, $\beta_1 = 25$, and $\beta_2 = 300$; and (2) an OFF-rate induction model, $\kappa_{01}^{U} = \kappa_{01}^{S} = 1$, $\kappa_{12}^{U} = \kappa_{12}^{S} = 1$, $\kappa_{10}^{U} = 10 \rightarrow \kappa_{10}^{S} = 0.1$, $\kappa_{21}^{U} = 10 \rightarrow \kappa_{21}^{S} = 0.1$, $\beta_0 = 0$, $\beta_1 = 100$, and $\beta_2 = 200$. The protein degradation rate constant is $\delta = 1$ by choosing the time scale of the model. We relaxed the models from $t = 0$ to $t = 20$ and sampled the system at $t = 20$, 20.5, 21, and 22. The observation time scale was motivated by our recent experimental procedure.[15,16] For each model, we synthesized by sampling 100 and 1000 synthetic data at each of the discrete sampled times from the joint probability distribution. The data consist of the sampled and discrete number of mRNA and whether the gene is active. Again, the genetic space s is marginalized that we defined $s > 0$ is an active allele ($TS = 1$) and otherwise inactive ($TS = 0$).

b. Bayesian analysis

The model class we considered for Bayesian inference is the set of two-allele, 3-state models with $\beta_0^{U} = \beta_0^{S} = 0$. The rest of the parameters are free parameters, but depending on the model structure, some of the perturbed (S) parameters may be constrained to the unperturbed (U) value. Combinatorially, there are in total $2^6 = 64$ models we considered, as there are six biophysical parameters $\vec{\theta} := (\kappa_{01}, \kappa_{12}, \kappa_{21}, \kappa_{10}, \beta_1, \beta_2)$.

We adopted a plain Markov chain Monte Carlo algorithm to sample the posterior distribution $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$, using the Metropolis-Hastings sampler.[64–66] Specifically, we perform random jumps in the logarithm space of the parameters ($\log \kappa_{ij}$ and $\log \beta_k$). The jump kernel was chosen to be uniformly distributed $\text{Unif}(-D, D)$, where the metaparameter D (diffusivity) globally regulates how wide the isotropic diffusion kernel is. For each model structure, we adjusted the metaparameter D such that the acceptance rate of the Metropolis-Hastings sampler was between 0.2 and 0.3.[67] We assumed that our prior is uniform in the logarithm space $\log \theta_i$.

Before running the MCMC samplers, we randomized 400 initial guesses of the model parameters and forward evolved the MCMC for 5×10^4 iterations each chain. Most of the chains converged to a unique parameter region, and the likelihood value in this region was significantly higher than the few chains trapped in (presumably) local maxima. We independently initiated 32 MCMC chains with different random initial speeds from the parameter values that maximized the likelihood in the previous test runs. We collected a total length (the sum of the length of all 32 chains, $>10^7$) to accurately approximate the posterior distribution $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$.

c. Computing the evidence $\mathbb{P}(h|\mathcal{M})$ from the posterior distributions $\mathbb{P}(\vec{\theta}|h, \mathcal{M})$

As described in Ref. 43, the evidence $\mathbb{P}(h|\mathcal{M})$ is computed by the algebraic identity

$$\mathbb{P}(h|\mathcal{M}) = \left[\int \frac{\phi(\vec{\theta}')\mathbb{P}(\vec{\theta}'|h, \mathcal{M})}{\mathbb{P}(h|\vec{\theta}', M)\mathbb{P}(\vec{\theta}'|\mathcal{M})} \mathrm{d}\vec{\theta}' \right]^{-1}, \tag{A24}$$

with an importance sampler $\phi(\vec{\theta}')$ satisfying the normalization condition

$$\int \phi(\vec{\theta}')\mathrm{d}\vec{\theta}' = 1. \tag{A25}$$

In this work, the posterior distributions were exclusively unimodal. Therefore, we chose the importance sampler to be proportional to an indicator function on an ellipsoid located at the high posterior density region. We first ranked the posterior chain by their likelihood and selected the top 20% parameter sets to construct the importance sampler. We performed a principle component analysis on the selected samples to compute the mean $\bar{\theta}_k$, variance σ_k^2, and eigenvector $\hat{e}_k$, $k = 1, 2, \ldots, 6$, in the six-dimensional parameter space. We used the eigenvalues and eigenvectors to construct an ellipsoid centering at the mean and with the axis length proportional to the eigenvalues along with the eigenvectors,

$$\mathcal{E} := \left\{ \vec{\theta} \middle| \sum_{k,j=1}^{6} \frac{\left(R_{k,j}\theta_j - R_{k,j}\bar{\theta}_j\right)^2}{\alpha\sigma_k^2} < 1 \right\}, \tag{A26}$$

where $R_{i,j} := (\hat{e}_i)_j$ is the linear transformation onto the eigenbasis. We tuned the metaparameter α such that there were precisely 20% of the points of the posterior chains inside the ellipsoid $\mathcal{E}$. These samples were then used to compute the evidence.

One must also specify the prior distribution $\mathbb{P}(\vec{\theta}|\mathcal{M})$ to compute the evidence. For simplicity, we imposed a uniform prior in the logarithm space of the parameters, bound by $(10^{-16}, 10^{4})$. $\mathbb{P}(\vec{\theta}|\mathcal{M})$is $1/V$, where V is the bounded volume in the parameter space. Let the posterior chains to be $\{\vec{\theta}_k\}_{k=1}^{N_P}$, where N_P is the total number of samples in the posterior chains. Then, the marginalized likelihood is estimated computed by the Monte Carlo sampler,

$$\widehat{\mathbb{P}}(h|\mathcal{M}) = \left[\sum_{k=1}^{N_P} \frac{1}{\hat{\mathbb{P}}(h|\vec{\theta}_k,\mathcal{M})}\mathbf{1}_{\vec{\theta}_k\in\mathcal{E}}\right], \tag{A27}$$

where $\mathbf{1}_{\vec{\theta}_k\in\mathcal{E}}$ is the characteristic function which is equal to 1 if $\vec{\theta}_k$ is in the ellipsoid $\mathcal{E}$ and 0 otherwise. In Fig. 5, we presented a normalized probability among all the 64 linear 3-state models we considered,

$$\bar{\mathbb{P}}(h|\mathcal{M}_i) := \frac{\widehat{\mathbb{P}}(h|\mathcal{M}_i)}{\sum_{j=1}^{64}\widehat{\mathbb{P}}(h|\mathcal{M}_j)}, \tag{A28}$$

which is reported in Fig. 5.

d. Estimating the evidence $\mathbb{P}(h|\mathcal{M})$ *from the Bayesian information criterion and Akaike information criterion*

The Schwarz index is an asymptotic result of the Bayesian evidence $\mathbb{P}(h|\mathcal{M})$ when the sample size is large.[68] Given a model $\mathcal{M}_i$, the Bayesian Information Criterion (BIC) is defined to be twice of the its Schwarz index,

$$\mathrm{BIC}(\mathcal{M}_i) := -2\mathcal{L}_i^{\max} + m_i \log N, \tag{A29}$$

where $\mathcal{L}_i^{\max}$ and m_i are the maximum likelihood and the number of free parameters of model $\mathcal{M}_i$, respectively, and N is the sample size (the number of data). Thus, to estimate the normalized probability $\bar{\mathbb{P}}$ using BIC, we use

$$\bar{\mathbb{P}}_{\mathrm{BIC}}(h|\mathcal{M}_i) := \frac{\exp\left[-\frac{1}{2}\mathrm{BIC}(\mathcal{M}_i)\right]}{\sum_{j=1}^{64}\exp\left[-\frac{1}{2}\mathrm{BIC}(\mathcal{M}_j)\right]}. \tag{A30}$$

We remark that the calculation of BIC only involves estimating the maximum likelihood $\mathcal{L}_i^{\max}$ of each model $\mathcal{M}_i$ and not the full posterior distribution $\mathbb{P}(\vec{\theta}|\mathcal{M}_i)$.

Akaike Information Criterion (AIC) is another commonly adopted matrix information criterion. The motivation of AIC is to minimize the information loss, measured by the Kullback–Leibler divergence (KL) of the prediction from the data. It is derived[53] as

$$\mathrm{AIC}(\mathcal{M}_i) := -2\mathcal{L}_i^{\max} + 2m_i + 2\frac{m_i + m_i^2}{N - m_i - 1}. \tag{A31}$$

Thus, the (normalized) evidence calculated by the AIC is

$$\bar{\mathbb{P}}_{\mathrm{AIC}}(h|\mathcal{M}_i) := \frac{\exp\left[-\frac{1}{2}\mathrm{AIC}(\mathcal{M}_i)\right]}{\sum_{j=1}^{64}\exp\left[-\frac{1}{2}\mathrm{AIC}(\mathcal{M}_j)\right]}. \tag{A32}$$

We remark that the negative logarithm of the likelihood function (1) converges to $N \times \mathrm{KL}\big(h(\omega)/N \| \mathbb{P}(\omega|\mathcal{M}_i,\vec{\theta})\big)$ only when $N \gg 1$ such that the multinomial coefficient $\mathbb{M}_\ell$ can be expanded by the Stirling approximation. In most biological cases, the sample size is far from this regime, and the Kullback–Leibler divergence is a poor choice to approximate the likelihood function (1).

REFERENCES

[1] X. Pichon, M. Lagha, F. Mueller, and E. Bertrand, "A growing toolbox to image gene expression in single cells: Sensitive approaches for demanding challenges," Mol. Cell **71**, 468–480 (2018).

[2] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, "Real-time kinetics of gene activity in individual bacteria," Cell **123**, 1025–1036 (2005).

[3] D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer, "Real-time observation of transcription initiation and elongation on an endogenous yeast gene," Science **332**, 475–478 (2011).

[4] A. M. Corrigan, E. Tunnacliffe, D. Cannon, and J. R. Chubb, "A continuum model of transcriptional bursting," eLife **5**, e13051 (2016).

[5] T. Fukaya, B. Lim, and M. Levine, "Enhancer Control of transcriptional bursting," Cell **166**, 358–368 (2016).

[6] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, "Regulation of noise in the expression of a single gene," Nat. Genet. **31**, 69–73 (2002).

[7] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," Science **297**, 1183–1186 (2002).

[8] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, "Mammalian genes are transcribed with widely different bursting kinetics," Science **332**, 472–474 (2011).

[9] D. Nicolas, B. Zoller, D. M. Suter, and F. Naef, "Modulation of transcriptional burst frequency by histone acetylation," Proc. Natl. Acad. Sci. U. S. A. **115**, 7153–7158 (2018).

[10] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, "Imaging individual mRNA molecules using multiple singly labeled probes," Nat. Methods **5**, 877–879 (2008).

[11] D. Zenklusen, D. R. Larson, and R. H. Singer, "Single-RNA counting reveals alternative modes of gene expression in yeast," Nat. Struct. Mol. Biol. **15**, 1263–1271 (2008).

[12] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, "mRNA-Seq whole-transcriptome analysis of a single cell," Nat. Methods **6**, 377–382 (2009).

[13] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," Exp. Mol. Med. **50**, 96 (2018).

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

[15] M. Gómez-Schiavon, L.-F. Chen, A. E. West, and N. E. Buchler, "BayFISH: Bayesian inference of transcription dynamics from population snapshots of single-molecule rna fish in single cells," Genome Biol. **18**, 164 (2017).

[16] L. F. Chen, Y. T. Lin, D. A. Gallegos, M. F. Hazlett, M. Gomez-Schiavon, M. G. Yang, B. Kalmeta, A. S. Zhou, L. Holtzman, C. A. Gersbach, J. Grandl, N. E. Buchler, and A. E. West, "Enhancer histone acetylation modulates transcriptional bursting dynamics of neuronal activity-inducible genes," Cell Rep. **26**, 1174–1188 (2019).

[17] A. Raj and A. van Oudenaarden, "Nature, nurture, or chance: Stochastic gene expression and its consequences," Cell **135**, 216–226 (2008).

[18] D. Nicolas, N. E. Phillips, and F. Naef, "What shapes eukaryotic transcriptional bursting?," Mol. BioSyst. **13**, 1280–1290 (2017).

[19] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," Proc. Natl. Acad. Sci. U. S. A. **105**, 17256–17261 (2008).

[20] N. Kumar, T. Platini, and R. V. Kulkarni, "Exact distributions for stochastic gene expression models with bursting and feedback," Phys. Rev. Lett. **113**, 268105 (2014).

[21] S. Tiberi, M. Walsh, M. Cavallaro, D. Hebenstreit, and B. Finkenstadt, "Bayesian inference on stochastic gene transcription from flow cytometry data," Bioinformatics **34**, i647–i655 (2018).

[22] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," J. Chem. Phys. **124**, 044104 (2006).

[23] Y. Cao, A. Terebus, and J. Liang, "Accurate chemical master equation solution using multi-finite buffers," Multiscale Model. Simul. **14**, 923–963 (2016).

[24] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, "Stochastic mRNA synthesis in mammalian cells," PLoS Biol. **4**, e309 (2006).

[25] S. C. Little, M. Tikhonov, and T. Gregor, "Precise developmental gene expression arises from globally stochastic transcriptional activity," Cell **154**, 789–800 (2013).

[26] K. Bahar Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg, and S. Itzkovitz, "Bursty gene expression in the intact mammalian liver," Mol. Cell **58**, 147–156 (2015).

[27] S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding, "Single-cell analysis of transcription kinetics across the cell cycle," eLife **5**, e12175 (2016).

[28] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," J. Comput. Phys. **22**, 403–434 (1976).

[29] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," J. Phys. Chem. **81**, 2340–2361 (1977).

[30] H. Kuwahara and I. Mura, "An efficient and exact stochastic simulation method to analyze rare events in biochemical systems," J. Chem. Phys. **129**, 165101 (2008).

[31] D. T. Gillespie, M. Roh, and L. R. Petzold, "Refining the weighted stochastic simulation algorithm," J. Chem. Phys. **130**, 174103 (2009).

[32] M. K. Roh, B. J. Daigle, D. T. Gillespie, and L. R. Petzold, "State-dependent doubly weighted stochastic simulation algorithm for automatic characterization of stochastic biochemical rare events," J. Chem. Phys. **135**, 234108 (2011).

[33] B. J. Daigle, M. K. Roh, L. R. Petzold, and J. Niemi, "Accelerated maximum likelihood parameter estimation for stochastic biochemical systems," BMC Bioinf. **13**, 68 (2012).

[34] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," PLoS Comput. Biol. **3**, e189 (2007).

[35] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," J. R. Soc., Interface **6**, 187–202 (2009).

[36] B. Munsky, G. Li, Z. R. Fox, D. P. Shepherd, and G. Neuert, "Distribution shapes govern the discovery of predictive models for gene regulation," Proc. Natl. Acad. Sci. U. S. A. **115**, 7533–7538 (2018).

[37] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2005).

[38] D. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (OUP Oxford, 2006).

[39] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," Biometrika **82**, 711–732 (1995).

[40] R. E. Kass and A. E. Raftery, "Bayes factors," J. Am. Stat. Assoc. **90**, 773–795 (1995).

[41] M. D. Weinberg, "Computing the Bayes factor from a Markov chain Monte Carlo simulation of the posterior distribution," Bayesian Anal. **7**, 737–770 (2012).

[42] A. Pajor, "Estimating the marginal likelihood using the arithmetic mean identity," Bayesian Anal. **12**, 261–287 (2017).

[43] C. P. Robert and D. Wraith, "Computational methods for Bayesian model choice," AIP Conf. Proc. **1193**, 251–262 (2009).

[44] P. Bokes, J. R. King, A. T. A. Wood, and M. Loose, "Transcriptional bursting diversifies the behaviour of a toggle switch: Hybrid simulation of stochastic gene expression," Bull. Math. Biol. **75**, 351–371 (2013).

[45] Y. T. Lin and C. R. Doering, "Gene expression dynamics with stochastic bursts: Construction and exact results for a coarse-grained model," Phys. Rev. E **93**, 022409 (2016).

[46] Y. T. Lin and T. Galla, "Bursting noise in gene expression dynamics: Linking microscopic and mesoscopic models," J. R. Soc., Interface **13**, 20150772 (2016).

[47] P. G. Hufton, Y. T. Lin, T. Galla, and A. J. McKane, "Intrinsic noise in systems with switching environments," Phys. Rev. E **93**, 052119 (2016).

[48] P. C. Bressloff, "Stochastic switching in biology: From genotype to phenotype," J. Phys. A: Math. Theor. **50**, 133001 (2017).

[49] Y. T. Lin, P. G. Hufton, E. J. Lee, and D. A. Potoyan, "A stochastic and dynamical view of pluripotency in mouse embryonic stem cells," PLoS Comput. Biol. **14**, e1006000 (2018).

[50] Y. T. Lin and N. E. Buchler, "Efficient analysis of stochastic gene dynamics in the non-adiabatic regime using piecewise deterministic Markov processes," J. R. Soc., Interface **15**, 20170804 (2018).

[51] U. Herbach, "Stochastic gene expression with a multistate promoter: Breaking down exact distributions," SIAM J. Appl. Math. **79**(3), 1007–1029 (2019).

[52] J. Wang, M. Huang, E. Torre, H. Dueck, S. Shaffer, J. Murray, A. Raj, M. Li, and N. R. Zhang, "Gene expression distribution deconvolution in single-cell RNA sequencing," Proc. Natl. Acad. Sci. U. S. A. **115**, E6437–E6446 (2018).

[53] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa (Springer New York, New York, NY, 1998), pp. 199–213.

[54] K. Murphy, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012).

[55] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," Phys. Lett. B **195**, 216–222 (1987).

[56] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo* (Chapman & Hall, 2011), Vol. 2, p. 2.

[57] M. Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," preprint arXiv:1701.02434 (2017).

[58] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, 2007).

[59] M. H. A. Davis, "Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models," J. R. Stat. Soc. Ser. B **46**, 353–388 (1984).

[60] I. Bena, "Dichotomous Markov noise: Exact results for out-of-equilibrium systems," Int. J. Mod. Phys. B **20**, 2825–2888 (2006).

[61] A. Faggionato, D. Gabrielli, and M. Ribezzi Crivellari, "Non-equilibrium thermodynamics of piecewise deterministic Markov processes," J. Stat. Phys. **137**, 259 (2009).

[62] G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden, "Systematic identification of signal-activated stochastic gene regulation," Science **339**, 584–587 (2013).

[63] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq, "Transcription factors modulate c-Fos transcriptional bursts," Cell Rep. **8**, 75–83 (2014).

[64] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," J. Chem. Phys. **21**, 1087–1092 (1953).

[65] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," Biometrika **57**, 97–109 (1970).

[66] M. Newman and G. Barkema, in *Monte Carlo Methods in Statistical Physics* (Oxford University Press, New York, USA, 1999), Chaps. 1–4.

[67] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling of random walk metropolis algorithms," Ann. Appl. Probab. **7**, 110–120 (1997).

[68] G. Schwarz, "Estimating the dimension of a model," Ann. Stat. **6**, 461–464 (1978).