The FEBS Journal

FEBS PRESS science publishing by scientists

STATE-OF-THE-ART REVIEW

Stochastic models for single-cell data: Current challenges and the way forward

Ian S. Hsu and Alan M. Moses

Department of Cell & Systems Biology, University of Toronto, ON, Canada

Keywords
non-Gaussian distribution; parameter fitting; single-cell data; stochastic models; structure identification

Correspondence
A. M. Moses, Department of Cell & Systems Biology, University of Toronto, Toronto ON, Canada M5S3B2
Tel: +1-416-946-3980
E-mail: alan.moses@utoronto.ca

(Received 30 July 2020, revised 22 December 2020, accepted 10 February 2021)

doi:10.1111/febs.15760

Although the quantity and quality of single-cell data have progressed rapidly, making quantitative predictions with single-cell stochastic models remains challenging. The stochastic nature of cellular processes leads to at least three challenges in building models with single-cell data: (a) because variability in single-cell data can be attributed to multiple different sources, it is difficult to rule out conflicting mechanistic models that explain the same data equally well; (b) the distinction between interesting biological variability and experimental variability is sometimes ambiguous; (c) the nonstandard distributions of single-cell data can lead to violations of the assumption of symmetric errors in least-squares fitting. In this review, we first discuss recent studies that overcome some of the challenges or set up a promising direction and then introduce some powerful statistical approaches utilized in these studies. We conclude that applying and developing statistical approaches could lead to further progress in building stochastic models for single-cell data.

Introduction: modeling single-cell data in systems biology

Over the last 20 years, there has been increasing interest in quantifying molecular biology at the single-cell level, ranging from HIV-1 reactivation [1] to bacterial motility [2] to gene expression in single neurons [3]. Unlike approaches that average over bulk populations of cells, single-cell approaches quantify features of large numbers of individual cells. This has led to insight into various biological reaction networks, including those that underlie gene regulation, signaling transduction, and metabolism. For example, studies of key steps in gene expression [4] and large-scale explorations at the single-cell level (e.g., single-cell metabolomics [5]) have been achieved. Properties of cells previously buried under the population average have been discovered: Examples include the fundamental 'noise' in gene expression [4,6,7], mRNA transcription bursting dynamics [8–10], ON/OFF responses (besides the expected graded responses [11–13]), and stochastic nuclear localization dynamics of transcription factors [14,15]. More recently, lessons learned from single cells are being applied to medicine (e.g., personalized cancer therapy [16]).

To facilitate this single-cell research, systems biology tools have been developed. One of the widely used tools is flow cytometry [17], which can quantify properties (e.g., cell sizes, fluorophore concentrations) of many individual cells at a time point (an example of time point measurements [18]). Advanced microfluidic systems and other devices [19–21] are now combined with quantitative fluorescence microscopes and have

Abbreviations
CME, Chemical master equations; HOG pathway, high osmotic glycerol pathway; KL divergence, Kullback–Leibler divergence; OU process, the Ornstein–Uhlenbeck process.

assisted high-throughput screening [22–24] and time-lapse imaging [25]. New imaging techniques [26,27] provide even more detailed information about single cells. More recently, single-cell sequencing [28] has been applied to explore developmental trajectories during embryogenesis [29] and neurogenesis [30], to develop the human cell atlas [31], and to characterize cell fate choice [32] and epigenomic cell-state dynamics [33]. Although various sources of technical variability still need to be addressed [34,35], the single-cell sequencing data will undoubtedly lead to even more discoveries.

Alongside the new experimental data, stochastic modeling and powerful simulations have been developed [36–41]. Models have explained many phenomena, including variability in developmental switching [42–44], incomplete penetrance [45], single-molecule protein distributions [46,47], and bimodality in gene expression [12]. Modeling approaches used in systems biology are diverse (comprehensively reviewed in Ref. [48]). Here, we focus on quantitative stochastic models that can be inferred from single-cell data.

There are many reasons why systems biologists want to construct stochastic models at the single-cell level. First of all, isogenic cells display wide cell-to-cell variations [49]. These variations can be due to low numbers of components in the cellular processes or introduced by some external processes with which the cell interacts [6,7]. Single-cell models can provide support and interpretation for single-cell molecular mechanisms that generate this variability [7,50]. Hence, in the areas where cell-to-cell variability has been appreciated (e.g., gene expression, signaling transduction [51]), single-cell modeling is widely applied, while for other areas, such as metabolism, it is less widely applied. Secondly, advanced single-cell technologies allow precise quantification of cell subpopulations and subcellular structure [52,53], suggesting the possibility to test quantitative predictions. Finally, the single cell is the fundamental unit of an organism. Models at the single-cell level can serve as the basis for multiscale models of more complex processes (e.g., the immune system [54], stem cell differentiation [52], and cancer progression [55]).

Optimizing models to minimize differences between predictions and data (model fitting) does not always improve predictive power. For example, an overly complex model may explain idiosyncrasies of certain datasets but may not be generalizable for predictions that help researchers design new experiments. Hence, improving model predictions for unseen experiments is critical. Although models have been widely applied in systems biology, they are rarely used to make quantitative predictions. For example, engineered single-cell systems, also known as cell circuits, are often designed based on qualitative use of models [56–58], but stochastic models have not been widely used to design these circuits in the context of single-cell variability (examples of notable exceptions include references [59–61]). We believe technical challenges limit the use of stochastic models, so in this review, we shall focus on challenges in developing mechanistic models that quantitatively predict cellular phenomena (Box 1). Throughout, we refer the reader to more comprehensive reviews of the areas of data analysis that we touch upon.

To correctly predict cellular behaviors and provide support for single-cell mechanisms, ruling out models with the wrong molecular interactions (referred to here as model structure) is crucial. The model structure defines a set of equations that describe how components in a mechanism interact with each other. Different model structures may lead to different predictions about cellular processes—for example, different ways to form an integral-feedback mechanism will lead to different steady-state behaviors [64] and dynamics [65]. Ideally, differences in predictions about quantifiable features (e.g., how mean expression level relates to the noise in gene expression, how expression level changes across time) can be used to rule out some models (Fig. 1). In practice, we would like to rule out as many model structures as possible with the data we have, so we can specifically design experiments (and collect more data) to test the remaining models.

In addition to the model structure, the equations in a model are associated with parameters, which are numbers that determine the quantitative properties of the interactions (such as strength or rate). Choosing optimal values of these parameters, usually by parameter fitting (also known as estimation), is necessary before making successful quantitative predictions [66–68]. One common technique of parameter fitting is maximum-likelihood estimation [48], which refers to the method that searches for the parameter values that maximize the likelihood function [48]. Conventional tools for parameter fitting minimize the sum-of-squared differences between predictions and observations (least-squares fitting), which is equivalent to maximum-likelihood estimation only if the distributions of errors are assumed to be Gaussian [69].

Challenges in building stochastic models with single-cell data

There are at least three major challenges in using stochastic single-cell models. First, because the number

Box 1. Qualitative vs. quantitative use of stochastic models, using gene expression as an example

The key components involved in gene expression are the DNA, mRNA, and protein. Their numbers can be generated by the following random processes:

$$\text{DNA} \xrightarrow{k_1} \text{DNA} + \text{mRNA}$$

$$\text{mRNA} \xrightarrow{k_2} \text{mRNA} + \text{protein}$$

$$\text{mRNA} \xrightarrow{k_3} \emptyset$$

$$\text{protein} \xrightarrow{k_4} \emptyset$$

where rates represent the probabilities of each reaction to occur in a given amount of time, k_1 and k_2 represent the transcription rate and translation rate, respectively, and k_3 and k_4 represent the degradation rates of mRNA and protein, respectively. If we interpret the model qualitatively, at equilibrium, we predict a simple positive relationship between mRNA and protein. The distributions of mRNA and protein numbers can range from Poisson to heavy-tailed (negative binomial or geometric) depending on the parameters [62,63]. These predictions give us intuition about the problem but do not strongly constrain any particular set of observations, limiting the falsifiability.

On the other hand, if we interpret the model quantitatively, and we can estimate k_1, k_2, k_3, and k_4 for a given gene, we obtain specific predictions about the mean, variance, and shapes of the distributions, as well as a quantitative prediction about the relationship between mRNA and protein numbers. We can then compare the predictions to the experimental data and quantify how accurate the models are. In addition, we can identify which genes violate the model and study them further. These models allow quantitative comparison between experimental data and can be used for biological discovery in a different way than theoretical models that only determine what phenomena are possible. It is in this sense that we believe quantitative models offer greater falsifiability.

We believe these issues are particularly important to consider for stochastic models (that include variability) because experimental variability can be confused or misinterpreted in the context of the variability predicted by the model.

of possible model structures is large, there may be many models that can explain data equally well [70]. Cellular mechanisms usually interact with other components and systems that are not well-described in the model, leading to extra variability [71] that makes ruling out models difficult. Second, when we observe stochastic cellular processes, we necessarily measure both experimental and biological variability. In other words, it is often ambiguous where the experimental noise ends and (interesting) biological fluctuations begin. This is a particular difficulty in time-series analysis. For example, the (interesting) stochastic oscillations are hard to separate from random fluctuations. Finally, even when we know the model structure, parameter fitting is a challenge because random motions of small numbers of components in cellular processes are a fundamental source of variability [6] and lead to discrete, heavy-tailed [62], and sometimes bimodal [12] distributions. The non-Gaussian distributions may lead to nonsymmetric fitting errors, violating a key assumption in the least-squares method or in the likelihood functions that assume normality. Violations of these assumptions have been shown to lead to poor model predictions [72].

For each of these challenges, we will discuss some recent research that we believe greatly improves our ability to overcome it, or, at least, points us in a promising direction. Throughout, we refer the reader to more comprehensive reviews to obtain more context for the topics we discuss. Taken together, this recent work suggests that progress can be made by exploring and applying more powerful statistical approaches to extract mechanistic information from the inevitably noisy biological systems.

Ruling out models using correlations between components: the case of mRNA and protein numbers in *E. coli*

Cellular mechanisms are usually embedded in a larger system (e.g., organelles like mitochondria [73], regulatory networks like signaling pathways [56], or the whole cell [74]). The unknown or unaccounted components lead to biological fluctuations in single-cell observations. Although they are not random noise, these fluctuations appear random in our observations because we do not know how to control or predict them. This presents a great challenge when modeling a mechanism

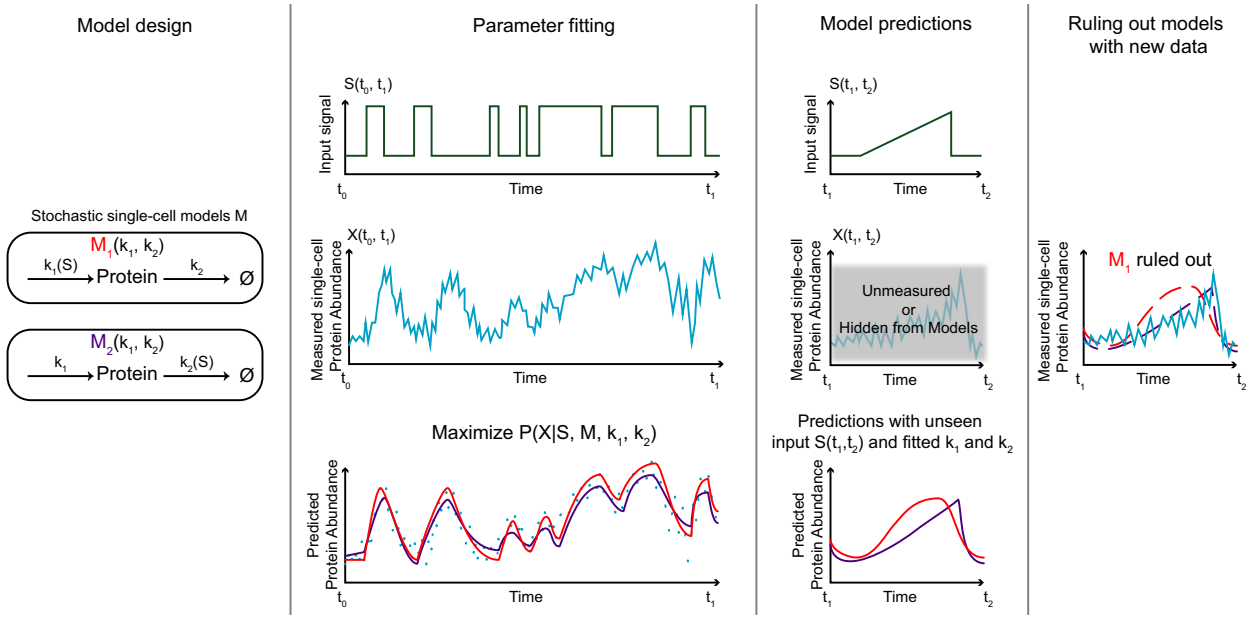

Fig. 1. Models that fit current data equally well can be ruled out through differences in model predictions. In this example, two models of gene expression, M_1 and M_2, are connected to the input signal S through different parameters: protein synthesis rate k_1 is assumed to be a function of S In M_1, and degradation rate k_2 is assumed as a function of S in M_2. The exemplary single-cell trajectory of protein abundance X (middle plots) is assumed as the response to the pattern of the input signal (top plots). Parameter values would be estimated by maximizing the probability of observing the single-cell trajectory X given the input signal S, the models, and a set of parameter values, P(XI S, M, k_1, k_2). Two models may fit the data equally well and show similar solutions (bottom plot, solid lines) with minimal distances from X (blue dots). However, they may make different predictions about how the cell responds to a different input signal type. In that case, the model that makes a worse prediction is ruled out.

quantitatively: These variations can be lumped together with fundamental stochastic processes included in a model, leading to conflicting interpretation of the model parameters. To illustrate this point, here we consider the mechanism of gene expression (Box 1).

Gene expression can be measured through the numbers of mRNA and protein with techniques like fluorescence in situ hybridization and fluorescence imaging with single-molecule sensitivity [75]. Steady-state gene expression shows large cell-to-cell variability in the numbers of mRNA and protein. One of the sources of variation in gene expression is the low numbers of components that fluctuate leading to so-called bursting dynamics [6,62]. If we simulate gene expression with a simple stochastic model (Box 1), the steady-state distribution of protein number will be heavy-tailed and has been shown as a Gamma distribution under some assumptions [62]. This simple model does not include any connection to biological fluctuations (formally, k_1, k_2, k_3, and k_4 are not functions of any other cellular components, for example, transcription factor concentrations [6], available ribosomes [60], cell cycle [76]). Nevertheless, it can still fit observations resulting from natural fluctuations because natural fluctuations will also lead to heavy-tailed distributions for protein numbers [62], and be conflated with fundamental noise due to low numbers of components. For example, a previous study showed that distributions of protein numbers for 137 highly expressed genes in *Escherichia coli* fit well to the simple model, but it would not be reasonable to explain their large cell-to-cell variations as the result of low component numbers [75].

We often do not know how connecting a model parameter to natural biological fluctuations affects model predictions for new, and model interpretation can become ambiguous. For example, although the model above (Box 1) predicts a positive correlation between cellular mRNA and protein numbers during equilibrium, single-cell mRNA and protein numbers of the 137 highly expressed genes showed no correlation [75]. The authors suggested that the lack of correlation could result from a biological fluctuation in k_1. In other words, their data, which represent a snapshot of the cells, might capture a long-time average of the number of proteins in a cell, while the number of mRNAs produced these proteins has already changed. However, we could also consider another model with a biological fluctuation in the translation rate k_2 (e.g., variability in the available ribosomes). This model would also predict no

correlation, and we cannot tell which one of the two models is more accurate.

In a promising approach to this challenge, Hilfinger *et al*. [77] proved that the correlation coefficient between the numbers of protein p and mRNA m (denoted as ρ_{pm}) depends on the mechanisms. For example, $\rho_{xs}$is equal to the ratio of noises $\frac{CV_p}{CV_m}$ (CV means coefficient of variation) when extrinsic noise only affects the transcription rate $k_1$and is equal to the time-averaged ratio of noises $\frac{\tau_p}{\tau_p+\tau_m}\frac{CV_p}{CV_m}$ (τ means the lifetime of protein or mRNA) when extrinsic noise only affects the translate rate k_2. To test this, Hilfinger *et al*. [77] used the same data from 137 highly expressed genes and found that the correlation coefficients of each gene are independent of the noise ratio $\frac{CV_p}{CV_m}$. The data are also inconsistent with the prediction of a model that explains the lack of correlation by invoking a much shorter mRNA lifetime: if the mRNA number fluctuates rapidly (large CV_m) and is not correlated with the protein number (small ρ_{pm}), then there should be a relatively small variation in protein number (small CV_p) because many mRNA bursts should not be transmitted to protein bursts. This model also predicts $\rho_{pm}=\frac{CV_p}{CV_m}$, which is not what has been observed. Therefore, the short lifetime of mRNA cannot explain the lack of correlation, and a broad class of models is now rejected. A similar argument combined with reported mRNA lifetimes leads to a different prediction in the presence of translation rate variability, and it also cannot explain the lack of correlation. Therefore, another broad class of models that connect the translation rate to natural fluctuations is rejected. Eventually, Hilfinger *et al*. showed that a class of model structures that include an antagonistic, noise-canceling interaction from protein to mRNA is consistent with both observations [77], shedding light on further hypothesis testing for the mechanism of gene expression.

This example highlights a general need for a more powerful statistical approach to analyze predictions of stochastic models based on a wider variety of statistical summaries. Taken together, we suggest that the development of diagnostic statistical summaries can yield surprising insight into what types of models are possible. This approach seems to have great promise for ruling out model structures.

Gaussian processes provide an objective statistical method for testing model predictions about stochastic single-cell trajectories

The time dynamics of cellular components provide a rich amount of information about biological functions, including intercellular signaling [78], homeostasis [79,80], circadian rhythm [81], cell cycle [82], etc. Oscillatory or periodic dynamics have been known to be involved in biological functions at the single-cell level and can be found in many signaling pathways [83–85]. However, more and more studies have shown that cellular processes can also carry out signaling functions by stochastically switching between damped oscillations (periodic dynamics that disappear) and noisy fluctuations [84,86,87]. Similarly, a class of transcription factors has been identified to encode environmental or intracellular stresses through aperiodic stochastic pulsatile nuclear localization [14] and has been linked to cell-state decisions [88,89].

Deterministic models have been constructed to explain how stochastic dynamics are generated [90,91]. The general explanation is that the systems are capable of generating periodic dynamics in a small region of parameter space and are pushed into a region of non-periodic dynamics by noises (or fluctuations). This qualitative difference between two adjacent regions in parameter space is called bifurcation [92] and has been found in models of many different biological functions.

Although the deterministic models explain the phenomena qualitatively, fitting them with data is challenging. The moment when a trajectory (time-series data) switches from damped oscillations to noisy fluctuations, for example, is hard to pin down because of measurement error and limitations of signal to noise ratio in single-cell data. If the two states of the trajectory cannot be reliably separated, then parameter values could be incorrectly determined or biased. One approach employs stochastic models of a biochemical reaction network to estimate parameter values [93] and perform other analyses [18,94]. The benefit of this approach is that the models provide mechanistic interpretability. However, because biological oscillations usually result from a combination of feedback loops with a time delay [95], dynamical features like periodicity are usually not represented by one single parameter. Hence, this approach usually does not provide a straightforward test for identification of oscillations or other dynamical features [96].

A promising alternative approach is to fit stochastic models that explicitly model the noise in the data. The noise around a steady state in the chemical master equations (CME, Box 2) can be approximated by a Gaussian process [96]. Hence, Gaussian processes regression models [97] are an appealing candidate for time-series data, and they have been used to determine if peaks in a given time series are damped oscillations or noisy fluctuations [96]. Gaussian processes describe

Box 2. The Chemical Master Equation

Because of the stochasticity, modeling single-cell gene expression predicts gene expression in a large number of cells as an ensemble. The CME has been introduced to describe stochastic biological reactions [50]. The CME is a set of linear differential equations that explicitly demonstrate how the probability density vectors of every chemical component affect each other over time. The CME can be written as

$$\frac{d}{dt}p(\mathbf{n},t)=\sum_{i=1}^{m}\left(\underbrace{-p(\mathbf{n},t)\boldsymbol{f_i}(\mathbf{n})}_{\text{flow out from }\mathbf{n}}+\underbrace{p(\mathbf{n}-d_i,t)\boldsymbol{f_i}(\mathbf{n}-d_i)}_{\text{flow into }\mathbf{n}}\right)$$

where $\mathbf{n}$ is a vector of non-negative integers representing copy numbers as the state of each component, $p(\mathbf{n},t)$ mean the probability for all the component to be in the states represented by $\mathbf{n}$ at time t, $m$is the number of reactions in the model, $f_i$is the propensity function of reaction $i$as a function of every component, and $d_i$is a vector of jump sizes of each component when reaction $i$occurs. The first term describes the flow out of state, and the second term describes the flow into state $\mathbf{n}$.

The gene expression model in Box 1 is a case of the CME with the numbers of DNA, mRNA, and protein $\mathbf{n}=[n_d\ \ n_m\ \ n_p]$, number of reactions $m=4$, jump sizes of DNA, mRNA, and protein of each reaction $d_1=[0\ \ 1\ \ 0]$, $d_2=[0\ \ 0\ \ 1]$, $d_3=[0\ \ -1\ \ 0]$, $d_4=[0\ \ 0\ \ -1]$, and propensity functions $f_1=k_1n_d$,$f_2=k_2n_m$, $f_3=k_3n_m$, $f_4=k_4n_p$, where k_i is the reaction rate of reaction i.

With the widely used stochastic simulation algorithm (SSA) [50], researchers can simulate an ensemble of cells for t equal to an arbitrarily long time and compare the simulated distribution to the experimental distribution. However, the non-normal distribution of both the simulated and experimental distributions violates the assumptions of conventional statistical tools for comparison. Estimating the parameters (k_i) in this model by matching the mean and variance to simulations is not likely to produce unbiased estimates in general because these are not sufficient statistics for non-Gaussian distributions produced by the CME.

how the covariance changes as a function of the difference between two time points (referred to as the 'kernel' in the Gaussian Process literature). For example, the covariance may drop exponentially, so two close time points covary tightly while distanced time points vary a lot. This approach provides a more realistic description of random fluctuation than white noise, which assumes independent samples for each time point. However, a key limitation of this approach is the loss of mechanistic interpretability. It also assumes that fitting error is Gaussian and that the dynamics are unimodal (limitations that could in principle be addressed using generalizations such as the recently proposed infinite Gaussian process mixture model [98]).

The Gaussian process model for steady-state fluctuations in the CME leads to two kernels to fit either random fluctuations or damped oscillations [96]. The forms of the kernels are

$$K_{\mathrm{OU}}=\exp(-\tau)$$

$$K_{\mathrm{OUosc}}=\exp(-\tau)\cos(\tau)$$

where τ is the diffenence between two time points. K_{OU} defines a multivariate Ornstein–Uhlenbeck (OU) process that has been widely used for modeling noisy fluctuations [99], and K_{OUosc} define oscillations damped by the OU process. The cosine function in K_{OUosc} suggests periodic peaks of covariance between time points, so K_{OUosc} describes damped oscillation as a process returning to equilibrium: When a cellular process is displaced from its equilibrium by some noises, a damped oscillation follows if some negative feedback loops or time delay reactions transform the displacement into restoring forces (e.g., feedbacks between the rates of different ion fluxes restore electric potential energy [100]).

The two kernels can fit time-series data by standard maximum-likelihood methods [96], and the log-likelihood ratio (LLR) of the data under the two kernels can be computed. For example, this approach was used to quantify HES5 protein fluctuations, which is a transcription factor regulating neurogenesis [89]. Based on the linear stability analysis of a deterministic model, the genetic circuit of HES5 in neural progenitors was predicted to lie on the boundary of bifurcation and linked to dynamics and fate decision. However, because of the noisy nature of the HES5 dynamics, deterministic models cannot be directly compared to data. By using a Gaussian process, they showed that the dynamics of HES5 are more periodic in differentiating cells than progenitors and linked the HES5 dynamics to the fate of neural progenitors during embryogenesis.

Because the model is generic, the data are not limited to gene expression. We, for example, used the same Gaussian process regression model to quantify the pulsatile nuclear localization dynamics of Crz1

before and after calcium bursts [101]. More Crz1 pulses were found after calcium bursts than before the bursts, and we proposed a time delay model to explain the link between the two dynamics. One prediction of the model is that Crz1 dynamics are damped oscillations after calcium bursts and are noisy fluctuation before the burst; moreover, larger calcium bursts are followed by more periodic Crz1 dynamics. Using the LLR statistic, we found a trend of periodicity in Crz1 dynamics that increased with preceding calcium burst sizes, providing direct evidence for the model from the time-series data.

In these examples, Gaussian processes are fit to single-cell stochastic time-series data to distinguish oscillating models from stochastic fluctuations. In the future, perhaps other mechanistic models can be connected to the parameters of Gaussian Process kernels, and so model structure and parameters may be inferred directly from time-series [102].

Using the whole distribution of data to fit stochastic models helps make precise predictions

A single-cell model with a correct structure sometimes fails to make a correct prediction even after its parameters are fitted to a rich amount of data [72]. One reason is that the statistical tools for fitting parameters (such as the Gaussian processes described in the previous section) assume normally distributed errors. However, as discussed above, distributions in isogenic populations are often non-Gaussian. Ideally, analytic forms of the data distribution as a function of the parameters could be derived from the CME (Box 2) or other stochastic models. In that case, parameter fitting could proceed using standard approaches such as maximum-likelihood estimations. In practice, explicit forms of the likelihood are rarely obtained even for simple models.

One approach to estimate model parameters is by optimizing parameter values to maximize the likelihood of observing the measured moments [103]. Although sometimes the first two or higher moments are sufficient for fitting a simple model to non-Gaussian distributions [104,105], for a model with many unknown parameters or high dimensional multi-modal distributions, a handful of moments are not expected to constrain the parameters [72].

Another approach is by optimizing parameter values to fit the entire (non-Gaussian) distributions obtained through simulations. Examples include measuring differences between distributions through information-theoretic metrics [106] or approximating the whole distributions with finite numbers of every chemical species (known as finite-state space systems [107]). Compared to the moment-based approach, this approach is more computationally expensive because it requires a large amount of stochastic simulation or, in the case of finite-state space systems, high-dimensional matrix exponentiation [106,108]. Major approximation methods are comprehensively reviewed in reference [108]. When feasible, we think this approach can improve model predictions and will discuss two recent applications of this approach.

A general method to quantify the difference between model predictions obtained through stochastic simulations and observations is needed. Kullback–Leibler (KL) divergence can quantify distance with the whole continuous distributions and is 0 if and only if two distributions are identical [109]. The KL divergence defined in the continuous case is

$$D(P|Q) = \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} dx \geq 0,$$

where P is the experimental distribution, and Q is the normalized simulations of the CME (e.g., molecule numbers divided by cell sizes to get concentration x [62]). It is well known that minimizing KL divergence is mathematically equivalent to maximizing the likelihood, assuming accurate enough simulation to obtain Q. Unfortunately, the infinity and negative infinity in the formula prevents a straightforward approximation of KL divergence through binning [110]. Hence, specialized estimators of KL divergence must be used in practice [111].

We used KL divergence between the simulated and experimental distributions to fit a stochastic model representing the high osmotic glycerol (HOG) pathway [12] to data from flow cytometry [112]. This system shows bimodal single-cell gene expression [12] and is therefore poorly described by the mean and variance. We tested the capacity of the model to explain the mutations we made in the components of the signaling pathway. Interestingly, the model could reproduce the quantitative effects of four out of seven mutations, suggesting that fitting stochastic models based on the KL divergence may be a promising approach when nonstandard distributions are observed.

Although the CME can describe stochastic cellular processes in general, analytic approximations to the CME are only available in simple cases and the algorithms for obtaining numerical solutions can be another challenge in the simulation-based paradigm for parameter fitting. In some studies, models are designed to describe highly skewed distributions and require a large number of simulation runs to estimate

the predicted distribution. Widely used algorithms like Gillespie's SSA can miss the long tail of an asymmetrical distribution if the number of simulations is not large enough [107]. To overcome this problem, Munsky *et al.*, 2018, used the finite state projection (FSP) algorithm to approximate the probability of the whole distribution [107] before they fitted parameters. The goal of their FSP algorithm is to approximate a system of infinite states (i.e., no upper bound of components' numbers) with finite states. In their algorithm, the numbers of states expand until the sum of every state's probability density is above one minus an allowable error, for example, $1–10^{-6}$ (where the real sum of the probability density of all infinite states is 1 by definition [107]). This way, they can approximate the small probability density of the long tail.

Again, using gene expression responses of the HOG pathway [72] as a model system, single-cell time series of mRNA copy number in the nucleus and cytoplasm was fit with the whole distribution calculated via FSP. The authors reported a precise prediction of their gene regulation model, which contains 13 nonspatial or 15 spatial parameters to predict nascent RNA [72]. In contrast, by fitting the same gene regulation model with the moment-based approach [103], the authors showed that the moment-based approach led to predictions wrong by four orders of magnitude [72].

Conclusion

The stochastic processes quantified at the single-cell level revealed exciting and puzzling phenomena that challenged our previous understanding about cellular mechanisms, including gene regulatory networks [12,66,113] and signal transduction pathways [83,91,114]. Although advanced experimental techniques assist researchers in collecting an ample amount of data, modeling single-cell stochastic dynamics remains challenging. The challenges include weak falsifiability of models, ambiguous boundaries between damped oscillations and noisy fluctuations on time trajectories, and violation in the assumptions of conventional tools for parameter fitting. We reviewed recent research that addresses each challenge. One common theme among these is the application of statistical theories that provide a clearer framework to interpret data, enhance the power to distinguish similar dynamics, and increase the precision of parameter fitting, hence improving the falsifiability of models.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

ISH conceived, wrote and edited the manuscript, made the figure and graphical abstract. AMM provided guidance, wrote and edited the manuscript.

Reference

1 Kutsch O, Benveniste EN, Shaw GM & Levy DN (2002) Direct and quantitative single-cell analysis of human immunodeficiency virus type 1 reactivation from latency. *J Virol* **76**, 8776–8786.
2 Vaknin A & Berg HC (2004) Single-cell FRET imaging of phosphatase activity in the *Escherichia coli* chemotaxis system. *Proc Natl Acad Sci USA* **101**, 17072–17077.
3 Hinkle D, Glanzer J, Sarabi A, Pajunen T, Zielinski J, Belt B, Miyashiro K, McIntosh T & Eberwine J (2004) Single neurons as experimental systems in molecular biology. *Prog Neurogibol* **72**, 129–142.
4 Rosenfeld N (2005) Gene regulation at the single-cell level. *Science* **307**, 1962–1965.
5 Heinemann M & Zenobi R (2011) Single cell metabolomics. *Curr Opin Biotechnol* **22**, 26–31.
6 Elowitz MB (2002) Stochastic gene expression in a single cell. *Science* **297**, 1183–1186.
7 Raj A & van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226.
8 Corrigan AM, Tunnacliffe E, Cannon D & Chubb JR (2016) A continuum model of transcriptional bursting. *eLife* **5**, e13051.
9 Pedraza JM & Paulsson J (2008) Effects of molecular memory and bursting on fluctuations in gene expression. *Science* **319**, 339–343.
10 Golding I, Paulsson J, Zawilski SM & Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036.
11 Munsky B & Neuert G (2015) From analog to digital models of gene regulation. *Phys Biol* **12**, 045004.
12 Pelet S, Rudolf F, Nadal-Ribelles M, de Nadal E, Posas F & Peter M (2011) Transient activation of the HOG MAPK pathway regulates bimodal gene expression. *Science* **332**, 732–735.
13 Veening J-W, Igoshin OA, Eijlander RT, Nijland R, Hamoen LW & Kuipers OP (2008) Transient heterogeneity in extracellular protease production by *Bacillus subtilis*. *Mol Syst Biol* **4**, 184.
14 Dalal CK, Cai L, Lin Y, Rahbar K & Elowitz MB (2014) Pulsatile dynamics in the yeast proteome. *Curr Biol* **24**, 2189–2194.
15 Levine J, Lin Y & Elowitz M (2013) Functional roles of pulsing in genetic circuits. *Science* **342**, 1193–1200.
16 Tian Q, Price ND & Hood L (2012) Systems cancer medicine: towards realization of predictive, preventive,

personalized and participatory (P4) medicine. *J Intern Med* **271**, 111–121.
17 McKinnon KM (2018) Flow cytometry: an overview. *Curr Protoc Immunol* **120**, 5.1.1–5.1.11.
18 Komorowski M, Costa MJ, Rand DA & Stumpf MPH (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc Natl Acad Sci USA* **108**, 8645–8650
19 Hansen AS, Hao N & O'Shea EK (2015) High-throughput microfluidics to control and measure signaling dynamics in single yeast cells. *Nat Protoc* **10**, 1181–1197.
20 Liu P & Mathies RA (2009) Integrated microfluidic systems for high-performance genetic analysis. *Trends Biotechnol* **27**, 572–581.
21 Lindström S & Andersson-Svahn H (2010) Overview of single-cell analyses: microdevices and applications. *Lab Chip* **10**, 3363–3372.
22 Hinojos CAD, Sharp ZD & Mancini MA (2005) Molecular dynamics and nuclear receptor function. *Trends Endocrinol Metab* **16**, 12–18.
23 Simpson JC, Cetin C, Erfle H, Joggerst B, Liebel U, Ellenberg J & Pepperkok R (2007) An RNAi screening platform to identify secretion machinery in mammalian cells. *J Biotechnol* **129**, 352–365.
24 Snijder B, Sacher R, Rämö P, Liberali P, Mench K, Wolfrum N, Burleigh L, Scott CC, Verheije MH, Mercer J *et al.* (2012) Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol Syst Biol* **8**, 579.
25 Muzzey D & van Oudenaarden A (2009) Quantitative time-lapse fluorescence microscopy in single cells. *Annu Rev Cell Dev Biol* **25**, 301–327.
26 Lubeck E & Cai L (2012) Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat Methods* **9**, 743–748.
27 Liu Z & Keller PJ (2016) Emerging imaging and genomic tools for developmental systems biology. *Dev Cell* **36**, 597–610.
28 Ayyaz A, Kumar S, Sangiorgi B, Ghoshal B, Gosio J, Ouladan S, Fink M, Barutcu S, Trcka D, Shen J *et al.* (2019) Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125.
29 Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A & Schier AF (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131.
30 Zywitza V, Misios A, Bunatyan L, Willnow TE & Rajewsky N (2018) Single-cell transcriptomics characterizes cell types in the subventricular zone and uncovers molecular defects impairing adult neurogenesis. *Cell Rep* **25**, 2457–2469.e8.
31 Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M *et al.* (2017) The human cell atlas. *eLife* **6**, e27041.
32 Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H & Grimes HL (2016) Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702.
33 Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J & Bock C (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* **10**, 1386–1397.
34 Stegle O, Teichmann SA & Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145.
35 Hicks SC, Townes FW, Teng M & Irizarry RA (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578.
36 Mendes P, Hoops S, Sahle S, Gauges R, Dada J & Kummer U (2009) Computational modeling of biochemical networks using COPASI. In Systems Biology (Maly IV, ed), pp. 17–59.Humana Press, Totowa, NJ.
37 Blinov ML, Schaff JC, Ruebenacker O, Wei X, Vasilescu D, Gao F, Morgan F, Ye L, Lakshminarayana A, Moraru II *et al.* (2014) Pathway commons at virtual cell: use of pathway data for mathematical modeling. *Bioinformatics* **30**, 292–294.
38 Resasco DC, Gao F, Morgan F, Novak IL, Schaff JC & Slepchenko BM (2012) Virtual cell: computational tools for modeling in cell biology. *Wiley Interdiscip Rev Syst Biol Med* **4**, 129–140.
39 Tang MY, Proctor CJ, Woulfe J & Gray DA (2010) Experimental and computational analysis of polyglutamine-mediated cytotoxicity. *PLoS Comput Biol* **6**, e1000944.
40 Maarleveld TR, Olivier BG & Bruggeman FJ (2013) StochPy: a comprehensive, user-friendly tool for simulating stochastic biological processes. *PLoS One* **8**, e79345.
41 de Vargas RL & Claassen M (2015) Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. *Curr Opin Biotechnol* **34**, 9–15.
42 Hasty J, Isaacs F, Dolnik M, McMillen D & Collins JJ (2001) Designer gene networks: towards fundamental cellular control. *Chaos Interdiscip J Nonlinear Sci* **11**, 207.
43 Cao Y, Lu H-M & Liang J (2008) Stochastic probability landscape model for switching efficiency, robustness, and differential threshold for induction of genetic circuit in phage λ. In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 611–614.

44 Tian T & Burrage K (2004) Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage λ. *J Theor Biol* **227**, 229–237.
45 Binder BJ, Landman KA, Newgreen DF & Ross JV (2015) Incomplete penetrance: the role of stochasticity in developmental cell colonization. *J Theor Biol* **380**, 309–314.
46 Kim K-Y, Lepzelter D & Wang J (2007) Single molecule dynamics and statistical fluctuations of gene regulatory networks: a repressilator. *J Chem Phys* **126**, 034702.
47 Chemla YR, Moffitt JR & Bustamante C (2008) Exact solutions for kinetic models of macromolecular dynamics. *J Phys Chem B* **112**, 6025–6044.
48 Loskot P, Atitey K & Mihaylova L (2019) Comprehensive review of models and methods for inferences in bio-chemical reaction networks. *Front Genet* **10**, 549.
49 Balázsi G, van Oudenaarden A & Collins JJ (2011) Cellular decision making and biological noise: from microbes to mammals. *Cell* **144**, 910–925.
50 Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* **58**, 35–55.
51 Longo D & Hasty J (2006) Dynamics of single-cell gene expression. *Mol Syst Biol* **2**, 64.
52 Wu J & Tzanakakis ES (2013) Deconstructing stem cell population heterogeneity: single-cell analysis and modeling approaches. *Biotechnol Adv* **31**, 1047–1062.
53 Rajagopal V, Holmes WR & Lee PVS (2018) Computational modeling of single-cell mechanics and cytoskeletal mechanobiology. *Wiley Interdiscip Rev Syst Biol Med* **10**, e1407.
54 Germain RN, Meier-Schellersheim M, Nita-Lazar A & Fraser IDC (2011) Systems biology in immunology: a computational modeling perspective. *Annu Rev Immunol* **29**, 527–585.
55 Cortesi M, Liverani C, Mercatali L, Ibrahim T, Giordano E (2020) Computational models to explore the complexity of the epithelial to mesenchymal transition in cancer. *WIREs Syst Biol Med* **12**, e1488.
56 Agrawal S, Archer C & Schaffer DV (2009) Computational models of the notch network elucidate mechanisms of context-dependent signaling. *PLOS Comput Biol* **5**, e1000390.
57 Hasty J, McMillen D, Isaacs F & Collins JJ (2001) Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet* **2**, 268–279.
58 Ribeiro AS, Dai X & Yli-Harja O (2009) Variability of the distribution of differentiation pathway choices regulated by a multipotent delayed stochastic switch. *J Theor Biol* **260**, 66–76.
59 Potvin-Trottier L, Lord ND, Vinnicombe G & Paulsson J (2016) Synchronous long-term oscillations in a synthetic gene circuit. *Nature* **538**, 514–517.
60 Raveh A, Margaliot M, Sontag ED & Tuller T (2016) A model for competition for ribosomes in the cell. *J R Soc Interface* **13**, 20151062.
61 Chait R, Ruess J, Bergmiller T, Tkačik G & Guet CC (2017) Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nat Commun* **8**, 1535.
62 Friedman N, Cai L & Xie XS (2006) Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* **97**, 168302.
63 Paulsson J & Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* **84**, 5447–5450.
64 Drengstig T, Jolma IW, Ni XY, Thorsen K, Xu XM & Ruoff P (2012) A basic set of homeostatic controller motifs. *Biophys J* **103**, 2000–2010.
65 Muzzey D, Gómez-Uribe CA, Mettetal JT & van Oudenaarden A (2009) A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell* **138**, 160–171.
66 Atkinson MR, Savageau MA, Myers JT & Ninfa AJ (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell* **113**, 597–607.
67 Zechner C, Unger M, Pelet S, Peter M & Koeppl H (2014) Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat Methods* **11**, 197–202.
68 Cui J, Kaandorp JA, Ositelu OO, Beaudry V, Knight A, Nanfack YF & Cunningham KW (2009) Simulating calcium influx and free calcium concentrations in yeast. *Cell Calcium* **45**, 123–132.
69 Moses A (2017) Statistical Modeling and Machine Learning for Molecular Biology. CRC Press, Boca Raton, Florida.
70 Mélykúti B, August E, Papachristodoulou A & El-Samad H (2010) Discriminating between rival biochemical network models: three approaches to optimal experiment design. *BMC Syst Biol* **4**, 38.
71 Llamosi A, Gonzalez-Vargas AM, Versari C, Cinquemani E, Ferrari-Trecate G, Hersen P & Batt G (2016) What population reveals about individual cell identity: single-cell parameter estimation of models of gene expression in yeast. *PLoS Comput Biol* **12**, e1004706.
72 Munsky B, Li G, Fox ZR, Shepherd DP & Neuert G (2018) Distribution shapes govern the discovery of predictive models for gene regulation. *Proc Natl Acad Sci USA* **115**, 7533–7538.
73 Malina C, Larsson C & Nielsen J (2018) Yeast mitochondria: an overview of mitochondrial biology and the potential of mitochondrial systems biology. *FEMS Yeast Res* **18**.

74 Yu R & Nielsen J (2019) Big data in yeast systems biology. *FEMS Yeast Res* **19**, foz070.
75 Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, Emili A & Xie XS (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single. *Cells* **329**, 8.
76 Cookson NA, Cookson SW, Tsimring LS & Hasty J (2010) Cell cycle-dependent variations in protein concentration. *Nucleic Acids Res* **38**, 2676–2681.
77 Hilfinger A, Norman TM & Paulsson J (2016) Exploiting natural fluctuations to identify kinetic mechanisms in sparsely characterized systems. *Cell Syst* **2**, 251–259.
78 Nandagopal N, Santat LA, LeBon L, Sprinzak D, Bronner ME & Elowitz MB (2018) Dynamic ligand discrimination in the notch signaling pathway. *Cell* **172**, 869–880.e19.
79 Law NC, Marinelli I, Bertram R, Corbin KL, Schildmeyer C & Nunemaker CS (2020) Chronic stimulation induces adaptive potassium channel activity that restores calcium oscillations in pancreatic islets *in vitro*. *Am J Physiol-Endocrinol Metab* **318**, E554–E563.
80 Ang J & McMillen DR (2013) Physical constraints on biological integral control design for homeostasis and sensory adaptation. *Biophys J* **104**, 505–515.
81 Schmal C, Reimann P & Staiger D (2013) A circadian clock-regulated toggle switch explains AtGRP7 and AtGRP8 oscillations in *Arabidopsis thaliana*. *PLOS Comput Biol* **9**, e1002986.
82 Ahmadian M, Tyson JJ, Peccoud J & Cao Y (2020) A hybrid stochastic model of the budding yeast cell cycle. *Npj Syst Biol Appl* **6**, 1–10.
83 Hoffmann A, Levchenko A, Scott ML & Baltimore D (2002) The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298**, 1241–1245.
84 Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, Milo R, Sigal A, Dekel E, Yarnitzky T, Liron Y, Polak P, Lahav G *et al.* (2006) Oscillations and variability in the p53 system. *Mol Syst Biol* **2**, 2006.0033.
85 Sneyd J, Han JM, Wang L, Chen J, Yang X, Tanimura A, Sanderson MJ, Kirk V & Yule DI (2017) On the dynamical structure of calcium oscillations. *Proc Natl Acad Sci USA* **114**, 1456–1461.
86 Bonev B, Stanley P & Papalopulu N (2012) MicroRNA-9 modulates Hes1 ultradian oscillations by forming a double-negative feedback loop. *Cell Rep* **2**, 10–18.
87 Imayoshi I, Isomura A, Harima Y, Kawaguchi K, Kori H, Miyachi H, Fujiwara T, Ishidate F & Kageyama R (2013) Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. *Science* **342**, 1203–1208.
88 Kroll JR, Tsiaxiras J & van Zon JS (2020) Variability in β-catenin pulse dynamics in a stochastic cell fate decision in *C. elegans*. *Dev Biol* **461**, 110–123.
89 Manning CS, Biga V, Boyd J, Kursawe J, Ymisson B, Spiller DG, Sanderson CM, Galla T, Rattray M & Papalopulu N (2019) Quantitative single-cell live imaging links HES5 dynamics with cell-state and fate in murine neurogenesis. *Nat Commun* **10**, 2835.
90 Wang C, Liu H & Zhou J (2019) Oscillatory dynamics of p53 genetic network induced by feedback loops and time delays. *IEEE Trans NanoBioscience* **18**, 611–621.
91 Martinez-Corral R, Raimundez E, Lin Y, Elowitz MB & Garcia-Ojalvo J (2018) Self-amplifying pulsatile protein dynamics without positive feedback. *Cell Syst* **7**, 453–462.e1.
92 Strogatz SH (2018) Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. CRC Press, Boca Raton, Florida.
93 Bronstein L, Zechner C & Koeppl H (2015) Bayesian inference of reaction kinetics from single-cell recordings across a heterogeneous cell population. *Methods* **85**, 22–35.
94 Elf J (2003) Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res* **13**, 2475–2484.
95 Novák B & Tyson JJ (2008) Design principles of biochemical oscillators. *Nat Rev Mol Cell Biol* **9**, 981–991.
96 Phillips NE, Manning C, Papalopulu N & Rattray M (2017) Identifying stochastic oscillations in single-cell live imaging time series using Gaussian processes. *PLoS Comput Biol* **13**, e1005479.
97 Rasmussen CE (2004) Gaussian Processes in Machine Learning. In Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures (Bousquet O, von Luxburg U & Rätsch G, eds), pp. 63–71.Springer, Berlin, Heidelberg.
98 McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE & Engelhardt BE (2018) Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput Biol* **14**, e1005896.
99 Gillespie DT (1996) Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. *Phys Rev E* **54**, 2084–2091.
100 Morris C & Lecar H (1981) Voltage oscillations in the barnacle giant muscle fiber. *Biophys J* **35**, 193–213.
101 Hsu IS, Strome B, Plotnikov S & Moses AM (2019) A noisy analog-to-digital converter connects cytosolic calcium bursts to transcription factor nuclear localization pulses in yeast. *G3* **9**, 561–570.
102 Hofmann T, Schölkopf B & Smola AJ (2008) Kernel methods in machine learning. *Ann Stat* **36**, 1171–1220.

103 Ruess J, Milias-Argeitis A & Lygeros J (2013) Designing experiments to understand the variability in biochemical reaction networks. *J R Soc Interface* **10**, 20130588.
104 Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J & Koeppl H (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci USA* **109**, 8340–8345.
105 Bronstein L & Koeppl H (2018) A variational approach to moment-closure approximations for the kinetics of biomolecular reaction networks. *J Chem Phys* **148**, 014105.
106 Smadbeck P & Kaznessis YN (2013) A closure scheme for chemical master equations. *Proc Natl Acad Sci USA* **110**, 14261–14265.
107 Munsky B & Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys* **124**, 044104.
108 Schnoerr D, Sanguinetti G & Grima R (2017) Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *J Phys Math Theor* **50**, 093001.
109 Kullback S & Leibler RA (1951) On information and sufficiency. *Ann Math Stat* **22**, 79–86.
110 Perez-Cruz F (2008) Kullback-Leibler divergence estimation of continuous distributions. In 2008 IEEE International Symposium on Information Theory, pp. 1666–1670.IEEE, Toronto, ON.
111 Zhao P & Lai L (2020) Analysis of K nearest neighbor KL divergence estimation for continuous distributions. In IEEE International Symposium on Information Theory (ISIT), pp. 2562-2567, Los Angeles, CA, USA.
112 Strome B, Hsu IS, Li Cheong Man M, Zarin T, Nguyen Ba A & Moses AM (2018) Short linear motifs in intrinsically disordered regions modulate HOG signaling capacity. *BMC Syst Biol* **12**, 75.
113 Santillán M, Mackey MC & Zeron ES (2007) Origin of bistability in the lac Operon. *Biophys J* **92**, 3830–3842.
114 Hersen P, McClean MN, Mahadevan L & Ramanathan S (2008) Signal processing by the HOG MAP kinase pathway. *Proc Natl Acad Sci USA* **105**, 7165–7170.