Poisson representation: a bridge between discrete and continuous models of stochastic gene regulatory networks

Xinyu Wang[1,†], Youming Li[2,†], Chen Jia[1,*]

[1] Applied and Computational Mathematics Division, Beijing Computational Science Research Center, Beijing 100193, China.

[2] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China.

[†] These authors contribute equally to this work.

[*] Correspondence: chenjia@csrc.ac.cn

Abstract

Stochastic gene expression dynamics can be modeled either discretely or continuously. Previous studies have shown that the mRNA or protein number distributions of some simple discrete and continuous gene expression models are related by Gardiner's Poisson representation. Here we systematically investigate the Poisson representation in complex stochastic gene regulatory networks. We show that when the gene of interest is unregulated, the discrete and continuous descriptions of stochastic gene expression are always related by the Poisson representation, no matter how complex the model is. In addition, using a simple counterexample, we find that the Poisson representation in general fails to link the two descriptions when the gene is regulated. However, for a general stochastic gene regulatory network, we demonstrate that the discrete and continuous models are approximately related by the Poisson representation in the limit of large protein numbers. These theoretical results are further applied to analytically solve many complex gene expression models whose exact distributions are previously unknown.

1 Introduction

Over the past two decades, breakthroughs have been made in the theory [1, 2] and experiments [3–5] of single-cell stochastic gene expression dynamics. The models of stochastic gene expression can be classified into two categories: discrete and continuous ones. Discrete models are those where the mRNA and protein numbers change by discrete integer amounts when reactions occur [1, 6]. In continuous models, the mRNA and protein fluctuations correspond to hops on the real axis rather than on the integer axis [7, 8]. Continuous models of stochastic gene expression can be further classified according to whether the fluctuations of genes are modeled discretely or continuously. In previous studies, those models with the gene fluctuations described in the discrete sense and with the mRNA and protein fluctuations described in the continuous sense [9, 10] are sometimes referred to as hybrid models. However, in this study, we will refer to them as continuous models to emphasize the continuous changes of mRNA and protein abundances.

The Poisson representation was first introduced by Gardiner and Chaturvedi [11–13] and serves as a powerful tool to investigate the properties of chemical master equations. Recently, some studies [14–16] used the Poisson representation to examine the relationship between discrete and continuous gene expression models and found that the gene product number distributions of some simple discrete and continuous models are related by the Poisson representation. It has been also shown [17–19] that various continuous models can be viewed as different macroscopic limits of the discrete models.

The importance of the Poisson representation can be seen from the following two specific examples. In experiments, it was observed that many mRNAs and proteins are produced in the bursty sense with the burst size having a geometric distribution [3, 4]. For such bursty gene expression, it has been shown that in steady state, the gene product number of the discrete model has a negative binomial distribution [20], while for the continuous model, the gene product number has a gamma distribution [8]. The discrete negative binomial distribution and the continuous gamma distribution are related by the Poisson representation and thus the former is also referred to as a Poisson-gamma distribution. The second example is the classical telegraph model of stochastic gene expression which describes the synthesis and degradation of the gene product, as well as the switching of the promoter between an active and an inactive state [21]. The discrete telegraph model can be solved exactly [22] and the steady-state gene product number distribution contains a confluent hypergeometric function. The continuous version of the telegraph model was first solved in [23] and the gene product distribution has a beta distribution in steady state. Interestingly, the discrete hypergeometric-type distribution and the continuous beta distribution are also linked by the Poisson representation, and this beautiful relation has been used in the parameter inference of the telegraph model [24].

While the Poisson representation has been shown to be able to link the copy number distributions of some simple discrete and continuous gene expression models [14–16], it remains unclear whether this result can be generalized to complex systems of gene expression and gene regulation. In particular, most all previous studies on the Poisson representation only focus on the expression of a particular gene that is unregulated. The aim of the present paper is to study to whether and what extent the Poisson representation can connect the discrete and continuous models of a general gene regulatory network. Even when the gene of interest is unregulated, it is still not clear whether the Poisson representation works for a general model with complex promoter switching mechanisms, bursty production of mRNAs and proteins, and upstream cellular drives.

The present paper is organized as follows. In Section 2, we recall the definition of univariate Poisson representation and describe the discrete and continuous expression models in detail for an unregulated gene with complex gene state switching, bursty production of protein, degradation of protein, and upstream cellular drives. In Section 3, we show that for an unregulated gene, the protein number distributions for the discrete and continuous models can always be related by the Poisson representation, no matter how complex the model is. In Section 4, we show that the Poisson representation cannot accurately relate the discrete and continuous models for an autoregulated gene; however, it approximately works in the limit of large protein numbers. The results in Sections 3 and 4 are applied to compute the exact or approximate protein number distributions for many complex gene expression models whose analytical solutions are previously unknown. In Section 5, we recall the definition of multivariate Poisson representation and demonstrate that for a general stochastic gene regulatory network, the joint distribution of protein numbers of the discrete and continuous models can be approximately related by the multivariate Poisson representation when protein numbers are large. We conclude in Section 6.

2 Poisson representation and models

2.1 Poisson representation

We first recall the definition of Poisson representation [11–13]. Let p_n, $n = 0, 1, 2, \cdots$ be the probability mass function of a nonnegative integer-valued discrete random variable. We say that p_n has a *Poisson representation* if it can be represented as the integral form

$$p_n = \int_0^\infty \frac{e^{-x}x^n}{n!}\rho(x)\mathrm{d}x, \quad n = 0, 1, 2, \ldots, \tag{1}$$

where $\rho(x)$, $x \geq 0$ is the probability density function of a nonnegative real-valued random variable. Note that $e^{-x}x^n/n!$ is the probability mass function of a Poisson distribution with mean x. Hence, intuitively, the Poisson representation decomposes a discrete probability distribution p_n into the weights sum of many Poisson distributions with different means $x \geq 0$, with the weights being distributed according to the probability density $\rho(x)$. In what follows, we refer to $\rho(x)$ as the *Poisson kernel* of p_n. If Eq. (1) is satisfied, we say that p_n and $\rho(x)$ are related by the Poisson representation.

Given the discrete distribution p_n, it is usually difficult to find its Poisson kernel $\rho(x)$ directly from Eq. (1). In order to find the Poisson kernel, recall that the generating function $f(z)$ of the discrete distribution p_n and the Laplace transform $\hat{\rho}(z)$ of the Poisson kernel $\rho(x)$ are defined by

$$f(z) = \sum_{n=0}^{\infty} p_n z^n, \quad \hat{\rho}(z) = \int_0^{\infty} \rho(x) e^{-zx} \mathrm{d}x.$$

It is well known that any nonnegative integer-valued discrete distribution is uniquely determined by its generating function and any nonnegative real-valued continuous distribution is uniquely determined by its Laplace transform. In terms of the generating function and Laplace transform, Eq. (1) can be converted into the following equivalent form [15]:

$$f(1-z) = \hat{\rho}(z). \tag{2}$$

Given the discrete distribution p_n, its Poisson kernel $\rho(x)$ can be determined in the following way. First we compute the generating function $f(z)$ of p_n. If $f(1-z)$, as a function of z, is the Laplace transform of a nonnegative real-valued random variable, then taking the inverse Laplace transform with respect to z gives the Poisson kernel $\rho(x)$. We stress that while Eqs. (1) and (2) are equivalent, they have different emphases: the former reveals the hidden structure behind the discrete distribution p_n, while the latter provides a convenient way to compute the Poisson kernel $\rho(x)$. It is easy to check that if the Poisson kernel of p_n exists, then p_n must be super-Poissonian, i.e. its variance must be greater than or equal to its mean, and thus its Fano factor must be greater than or equal to 1 [25].

2.2 Discrete gene expression models

Here we consider a general stochastic gene expression model with complex gene state switching, bursty production of protein, degradation of protein, and upstream cellular drives (Fig. 1(a)). For simplicity, here we do not take genetic feedback loops into account. The extension to a general gene regulatory network will be discussed in Section 5.

We assume that the gene of interest can exist in $L+1$ states, denoted by $G_0, G_1, \ldots, G_L$, and let P denote the corresponding protein. These gene states correspond to different conformational states during chromatin remodeling or different binding states with transcription factors [26]. Most previous studies assume that the gene switches between an active and an inactive state with the former having a higher transcription rate than the latter [22]. However, recent experiments suggest that the gene may transition among multiple states with different transcription rates [27–33]. In addition, we assume that the protein is produced in a bursty manner with an arbitrary burst size distribution. Let n denote the copy number of protein molecules. Then the gene expression dynamics can be described by the following reactions:

$$\begin{gathered} G_i \xrightarrow{k_{ij}(t)} G_j, \quad i,j = 0,1,\ldots,L, \\ G_i \xrightarrow{u_i(t)\theta_k} G_i + kP, \quad k \geq 0, \quad i = 0,1,\ldots,L, \quad P \xrightarrow{d(t)} \varnothing. \end{gathered} \tag{3}$$

Here we allow all reaction rates to be time-dependent. Such time-varying rates can be understood as upstream cellular drives due to different aspects of cellular environment or gene networks inside the cell [14]. The first reaction describes gene state switching with rates $k_{ij}(t)$, the second reaction describes bursty production of protein in all gene states, and

the third reaction describes protein degradation with rate $d(t)$. In particular, in each gene state G_i, protein synthesis occurs at a rate $u_i(t)$ in bursts of a random size sampled from an arbitrary distribution θ_k, $k \geq 0$. This means that in each burst, there is a probability θ_k of producing k protein molecules.

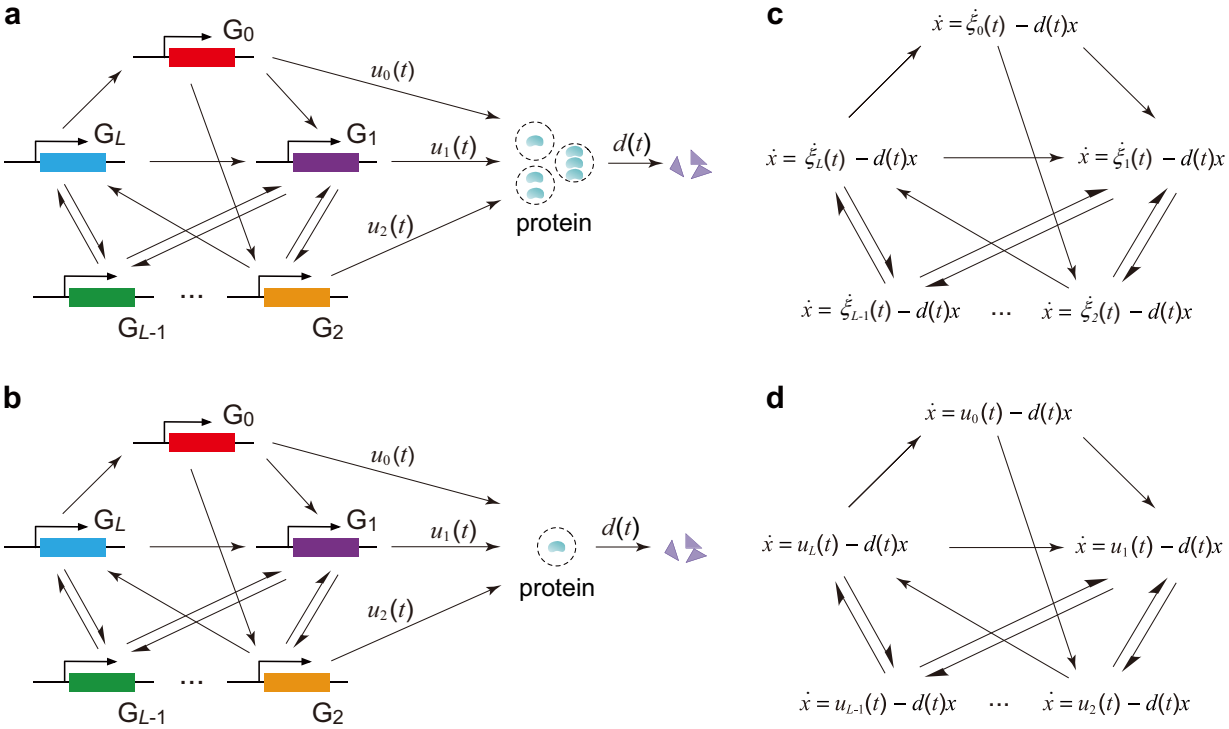


Figure 1: **Models. (a)** Schematic of a gene expression model with complex gene switching mechanism and bursty production of gene product molecules. The gene can exist in $L+1$ states, $G_0, G_1, \ldots, G_L$, and the switching among them can be arbitrary. Different gene states correspond to different burst frequencies of the gene product (some arrows corresponding to gene product synthesis are omitted in the figure). In each gene state, the gene product molecules are produced in a bursty manner. **(b)** Schematic of a gene expression model with complex gene switching mechanism and non-bursty production of gene product molecules. In each gene state, the gene product molecules are produced in a non-bursty manner. **(c)** Schematic of the switching SDE model, which is a continuous model of stochastic gene expression in the case of bursty production. In each gene state G_i, the synthesis and degradation of the gene product is described by the SDE $\dot{x} = \dot{\xi}_i(t) - d(t)x$. The SDE is driven by a compound Poisson process $\xi_i(t)$ which describes stochastic gene expression bursts. **(d)** Schematic of the switching ODE model, which is a continuous model of stochastic gene expression in the case of non-bursty production. In each gene state G_i, the synthesis and degradation of the gene product is described by the ODE $\dot{x} = u_i(t) - d(t)x$.

The microstate of the gene can be represented by an ordered pair (i, n), where $i = 0, 1, \ldots, L$ is the state of the gene and $n \geq 0$ is the number of the protein. Let $p_{i,n}(t)$ denote the probability of having n protein copies when the gene is in state i. Then the stochastic gene expression dynamics can be modeled as a Markov jump process, and the evolution of the Markovian model is governed by the following set of chemical master equations (CMEs):

$$\begin{aligned}\dot{p}_{i,n}(t) = u_i(t)\Big[\sum_{k=0}^{n}\theta_{n-k}p_{i,k}(t) - p_{i,n}(t)\Big] + d(t)\left[(n+1)p_{i,n+1}(t) - np_{i,n}(t)\right] \\ + \Big[\sum_{j\neq i}k_{ji}(t)p_{j,n}(t) - \sum_{j\neq i}k_{ij}(t)p_{i,n}(t)\Big].\end{aligned} \tag{4}$$

Here the first term on the right-hand side describes protein synthesis, the second term describes protein degradation, and the last term describes gene state switching. Since the gene state and protein number are both discrete variables, the above Markovian model will be referred to as the *discrete model* in what follows.

We emphasize that protein synthesis is not always bursty. Both non-bursty and bursty protein expression are commonly observed in naturally occurring systems [34]. When protein synthesis is bursty, the burst size typically has the geometric distribution $\theta_k = p^k(1-p)$, which is supported by experiments [4]. The geometric burst size distribution is due to rapid synthesis of protein from a single short-lived transcript [35, 36]. Some other burst size distributions, such as the Poisson and negative binomial distributions, have also been studied in the literature [37–40].

When gene expression is non-bursty, the burst size distribution $\theta_k = \delta_{1,k}$ is the point mass at one, where is $\delta_{1,k}$ the Kronecker delta which takes the value of 1 when $k = 1$ and the value of 0 otherwise. In this case, the reactions describing the protein dynamics are given by (Fig. 1(b))

$$\begin{aligned} G_i &\xrightarrow{k_{ij}(t)} G_j, \quad i,j = 0,1,\ldots,L, \\ G_i &\xrightarrow{u_i(t)} G_i + P, \quad i = 0,1,\ldots,L, \quad P \xrightarrow{d(t)} \varnothing, \end{aligned} \tag{5}$$

where $u_i(t)$ is the protein synthesis rate in the gene state G_i. The evolution of the discrete model in the non-bursty case is governed by the following set of CMEs:

$$\begin{aligned} \dot{p}_{i,n}(t) = {} & u_i(t)[p_{i,n-1}(t) - p_{i,n}(t)] + d(t)[(n+1)p_{i,n+1}(t) - np_{i,n}(t)] \\ & + \Big[\sum_{j\neq i} k_{ji}(t)p_{j,n}(t) - \sum_{j\neq i} k_{ij}(t)p_{i,n}(t)\Big]. \end{aligned}$$

2.3 Continuous gene expression models

In many previous papers [8, 18, 40, 41], the gene state is modeled as a discrete variable, while the protein number is modeled as a continuous variable. One important reason for this modeling choice is that in bulk experiments and many single-cell experiments without single-molecule resolution such as flow cytometry, data are often recorded as continuous measurements at a macroscopic scale [18]. The models with discrete gene state and continuous protein abundance are sometimes called hybrid models [42]. However, in the present paper, such kind of models will be referred to as *continuous models* to emphasize that the protein abundance varies continuously.

We have seen that the gene expression dynamics can be modeled discretely as a Markov jump process. We next focus on the continuous counterpart of the discrete model. In the continuous model, the bursty production of protein in each gene state is usually described by a compound Poisson process [40]. This is because in each gene state, transcripts are produced one by one and thus mRNA synthesis can be modeled as a simple Poisson process. Once a transcript is synthesized, it can produce a large number of protein molecules before it is finally degraded. Hence protein synthesis can be modeled as a compound Poisson process. In each gene state G_i, the gene expression dynamics can be modeled continuously by the stochastic differential equation (SDE) [18, 19]

$$\dot{x} = \dot{\xi}_i(t) - d(t)x,$$

where x denotes the protein abundance which varies continuously, the first term on the right-hand side describes bursty protein synthesis with $\xi_i(t)$ being a compound Poisson process with arrival rate $u_i(t)$ and jump distribution $w(x)$, and the second term describes protein degradation with rate $d(t)$. Here $u_i(t)$ is the burst frequency and $w(x)$ is the burst size distribution. Note that in the discrete model, the burst size distribution is θ_k, where $k = 0, 1, 2, \ldots$ varies discretely, while in the continuous model, the burst size distribution should be chosen differently as $w(x)$, where $x \geq 0$ varies continuously. The relationship between θ_k and $w(x)$ will be explained later.

Since the gene can switch between different states, the gene expression dynamics can be modeled continuously as a hybrid switching SDE [18], as illustrated in Fig. 1(c). Let $p_i(x,t)$ denote the probability density of the protein

abundance x when the gene is in state i. Then the evolution of the switching SDE model is governed by the Kolmogorov forward equation

$$\begin{aligned}\partial_t p_i(x,t) = u_i(t)\Big[\int_0^x w(x-x')p_i(x',t)\mathrm{d}x' - p_i(x,t)\Big] + d(t)\partial_x[xp_i(x,t)] \\ + \Big[\sum_{j\neq i} k_{ji}(t)p_j(x,t) - \sum_{j\neq i} k_{ij}(t)p_i(x,t)\Big].\end{aligned} \tag{6}$$

Here the first term on the right-hand side describes protein synthesis, the second term describes protein decay, and the last term describes gene state switching.

The remaining question is how to choose the burst size distribution $w(x)$ of the continuous model. To answer this, recall that in previous papers, the burst size distribution of the discrete model is often chosen to be the geometric distribution $\theta_k = p^k(1-p)$ [20], while the burst size distribution of the continuous model is often chosen to be the exponential distribution $w(x) = e^{-x/B}/B$ [8], where $B = \sum_{k=1}^{\infty} k\theta_k = p/(1-p)$ is the mean burst size of the discrete model. A crucial observation is that the geometric and exponential distributions are related by the Poisson representation, i.e.

$$\theta_k = \int_0^\infty \frac{e^{-x}x^k}{k!} w(x)\mathrm{d}x, \quad k = 0,1,2,\ldots. \tag{7}$$

In other words, the exponential distribution is exactly the Poisson kernel of the geometric distribution. Thus in what follows, *we always assume that θ_k and $w(x)$ are related by the Poisson representation*, even when θ_k is not geometrically distributed.

Note that non-geometric burst size distributions were also studied in some previous papers. In [37], the burst size of the discrete model is assumed to have the Poisson distribution $\theta_n = e^{-\lambda}\lambda^n/n!$. According to our assumption, the corresponding burst size distribution of the continuous model is the point mass at λ, i.e. $w(x) = \delta(x-\lambda)$. In [39], the burst size of the discrete model is assumed to have the negative binomial distribution

$$\theta_n = \frac{(s)_n}{n!} p^n (1-p)^r,$$

where $(s)_n = s(s+1)\cdots(s+n-1)$ is the Pochhammer symbol. In this case, the corresponding burst size distribution of the continuous model is the gamma distribution

$$w(x) = \frac{1}{\Gamma(s)B^s} x^{s-1} e^{-x/B},$$

with $B = p/(1-p)$, since the gamma distribution is the Poisson kernel of the negative binomial distribution.

We next consider the special case where protein synthesis is non-bursty. In this case, in each gene state G_i, we use the following ordinary differential equation (ODE) to model the gene expression dynamics [15, 18]:

$$\dot{x} = u_i(t) - d(t)x,$$

where the two terms on the right-hand side characterize synthesis and degradation of protein. Since the gene can switch between different states, the gene expression dynamics can be modeled continuously as a hybrid switching ODE [43], as depicted in Fig. 1(d). The switching ODE is also called a piecewise deterministic Markov process [44], and this model has been studied extensively in the literature [45–47]. In the non-bursty case, the evolution of the switching ODE model is governed by the Kolmogorov forward equation

$$\partial_t p_i(x,t) = -u_i(t)\partial_x p_i(x,t) + d(t)\partial_x\left[xp_i(x,t)\right] + \Big[\sum_{j\neq i} k_{ji}(t)p_j(x,t) - \sum_{j\neq i} k_{ij}(t)p_i(x,t)\Big]. \tag{8}$$

This switching ODE model is also referred to as "effective upstream drive" in [14], since it integrates the effects of upstream cellular drives with time-dependent synthesis and degradation rates.

In fact, Eq. (8) can also be viewed as a special case of Eq. (6). To see this, recall that in the discrete model, the burst size distribution $\theta_k = \delta_{1,k}$ is the point mass at one when gene expression is non-bursty. The corresponding burst size distribution for the continuous model should be chosen as

$$w(x) = \delta(x) - \delta'(x), \tag{9}$$

where $\delta(x)$ is Dirac's delta function and $\delta'(x)$ is the weak derivative of $\delta(x)$. Under this choice, θ_k and $w(x)$ are related by the Poisson representation since the generating function of the former is $h(z) = \sum_{k=0}^{\infty} \theta_k z^k = z$, while the Laplace transform of the latter is $\hat{w}(z) = \int_{0^-}^{\infty} e^{-zx} w(x) \mathrm{d}x = 1 - z$ [48], where $\int_{0^-}^{\infty}$ stands for $\lim_{\epsilon \to 0^+} \int_{-\epsilon}^{\infty}$. Note that when integrating the delta function, $x = 0$ has to be an interior point of the integration region and hence $\int_0^{\infty}$ in the definition of the Laplace transform should be modified slightly as $\int_{0^-}^{\infty}$. It thus follows from Eq. (2) that $w(x)$ is the Poisson kernel of θ_k. Inserting Eq. (9) into Eq. (6), we immediately obtain Eq. (8).

2.4 Examples of Poisson representation

For some simple gene expression systems, it was already known that the steady-state protein number distributions for the discrete and continuous models are related by the Poisson representation [14, 15]. To help readers gain deeper insights, we next provide several examples.

Example 1. Let G be the gene of interest and let P be the corresponding gene product (mRNA or protein). The simplest gene expression model is given by

$$G \xrightarrow{u} G + P, \quad P \xrightarrow{d} \varnothing,$$

where the gene product is produced in a non-bursty manner. When the system reaches steady state, the gene product number for the discrete Markovian model has the Poisson distribution $p_n^{ss} = s^n e^{-s}/n!$, where $s = u/d$. On the other hand, since gene expression is non-bursty, the corresponding continuous model is simply the ODE $\dot{x} = u - dx$, which has a unique globally asymptotic stable equilibrium $s = u/d$. In other words, the steady-state distribution of the continuous model is the point mass at s, i.e. $p^{ss}(x) = \delta(x - s)$. It is easy to see that the steady-state solutions for the two models are related by the Poisson representation, i.e.

$$p_n^{ss} = \int_0^{\infty} \frac{e^{-x} x^n}{n!} p^{ss}(x) \mathrm{d}x, \quad n = 0, 1, 2, \ldots. \tag{10}$$

Example 2. Consider a gene expression system described by the reactions

$$G \xrightarrow{up^k(1-p)} G + kP, \quad k \geq 0, \quad P \xrightarrow{d} \varnothing, \tag{11}$$

where the gene product is produced in bursts with the burst size having a geometric distribution $\theta_k = p^k(1-p)$. In steady state, it has been shown [20] that the gene product number for the discrete Markovian model has the negative binomial distribution

$$p_n^{ss} = \frac{(s)_n}{n!} p^n (1-p)^s,$$

where $s = u/d$ and $(s)_n = s(s+1)\ldots(s+n-1)$ is the Pochhammer symbol. Since gene expression is bursty, the corresponding continuous model is the SDE $\dot{x} = \dot{\xi}(t) - dx$, with $\xi(t)$ being a compound Poisson process with

burst frequency u and exponential burst size distribution $w(x) = e^{-x/B}/B$. It is well-known [8] that the gene product number for the continuous SDE model has the gamma distribution

$$p^{ss}(x) = \frac{1}{B^s \Gamma(s)} x^{s-1} e^{-x/B}.$$

It is easy to check that the steady-state distributions for the two models are related by the Poisson representation.

In single-cell RNA sequencing experiments, the observed number of mRNA reads usually has a negative binomial distribution [49]. The reaction scheme (11) provides a mechanistic understanding of the negative binomial distribution from the biophysical perspective. However, in the biostatistical and bioinformatic literature, the origin of the negative binomial distribution is often interpreted as follows [50–54]: there is a true number x of transcripts that varies in the population and is distributed as a gamma random variable. The number of observed reads is distributed as a Poisson random variable with mean x. Hence the total number of the reads obeys the Poisson-gamma distribution, i.e. the negative binomial distribution.

Example 3. The classical telegraph model of gene expression is described by the following reactions:

$$G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \quad G^* \xrightarrow{u} G^* + P, \quad P \xrightarrow{d} \varnothing, \tag{12}$$

where the gene switches between an inactive state G and an active state G^*, and the gene product is produced in a non-bursty manner when the gene is active. It has been shown [22] that the steady-state copy number distribution for the discrete model is given by

$$p_n^{ss} = \frac{s^n}{n!} \frac{(\alpha)_n}{(\beta)_n} {}_1F_1(\alpha + n; \beta + n; -s), \tag{13}$$

where $\alpha = a/d$, $\beta = (a+b)/d$, $s = u/d$, and ${}_1F_1(\alpha; \beta; z)$ is the confluent hypergeometric function. Since gene expression is non-bursty, the corresponding continuous model is the switching system of ODEs

$$\begin{array}{ll} \dot{x} = u - dx & \text{(gene is active)}, \\ \quad a \Big\Uparrow \Big\Downarrow b & \\ \dot{x} = -dx & \text{(gene is inactive)}. \end{array}$$

It is also known [9] that the steady-state distribution for the continuous model is the beta distribution

$$p^{ss}(x) = \frac{\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta - \alpha)} s^{1-\beta} x^{\alpha-1} (s - x)^{\beta-\alpha-1}, \quad 0 < x < s.$$

It is easy to check that the generating function $f^{ss}(z)$ of the discrete distribution p_n^{ss} and the Laplace transform $\hat{p}^{ss}(z)$ of the continuous distribution $p^{ss}(x)$ are related by $f^{ss}(1-z) = \hat{p}^{ss}(z) = {}_1F_1(\alpha; \beta; -sz)$. This indicates that the two distributions are related by the Poisson representation. Due to this reason, the steady-state distribution (13) of the telegraph model is also called the Poisson-beta distribution [24, 55–57].

3 Poisson representation for unregulated genes

3.1 Main results

From the above examples, it can be seen that for some simple gene expression systems, even for systems with a geometrical burst size distribution (Example 2) or two-state gene state switching (Example 3), the steady-state copy number distributions for the discrete and continuous models are related by the Poisson representation. A natural question

is whether this result holds for general gene expression systems — does it also hold for systems with an arbitrary burst size distribution, complex multi-state gene switching, and even feedback regulation? In addition, it is also unclear whether the above result holds for the time-dependent dynamic behavior of the system — are the time-dependent distributions for the discrete and continuous models also related by the Poisson representation, even for systems with upstream cellular drive?

In fact, these questions have been partially answered in some previous papers [14–16]. However, the conclusion is not general due to the following reasons. First, all these papers only consider unregulated genes and did not take feedback controls into account. Second, most previous papers [14, 15] only focus on non-bursty gene expression and do not take transcriptional or translational bursting into account; while [16] focuses on bursting, it does not include gene state switching. Third, [15, 16] do not consider upstream cellular drives, while [14] essentially assumes that initially there is no gene product molecules in the cell. In this paper, we will answer the above questions for general gene regulatory networks with complex gene state switching, arbitrary burst size distributions, arbitrary feedback controls, and upstream cellular drives. In addition, we will consider both the steady-state and time-dependent behaviors of the system.

We first focus on the dynamics for unregulated genes, where all the reaction rates are independent of the protein number n. The system can be described by the following reactions:

$$\begin{gathered} G_i \xrightarrow{k_{ij}(t)} G_j, \quad i,j = 0,1,\ldots,L, \\ G_i \xrightarrow{u_i(t)\theta_k} G_i + kP, \quad k \geq 0, \quad i = 0,1,\ldots,L, \quad P \xrightarrow{d(t)} \varnothing. \end{gathered} \tag{14}$$

In particular, if $\theta_k = \delta_{1,k}$ is the point mass at one, then gene expression in each state is non-bursty. The following theorem is the main result of this section (see Appendix A for the detailed proof).

Theorem 1. Consider the gene expression system described by Eq. (14). Suppose that the burst size distributions for the discrete and continuous models are related by the Poisson representation, i.e. Eq. (7) is satisfied. If for each gene state G_i, the initial distribution $p_{i,n}(0)$ of the discrete model and the initial distribution $p_i(x,0)$ of the continuous model are related by the Poisson representation, i.e.

$$p_{i,n}(0) = \int_0^\infty \frac{e^{-x}x^n}{n!} p_i(x,0)\mathrm{d}x, \quad n = 0,1,2,\ldots,$$

then for each gene state G_i and each time $t \geq 0$, the time-dependent distributions of the discrete and continuous models are related by the Poisson representation, i.e.

$$p_{i,n}(t) = \int_0^\infty \frac{e^{-x}x^n}{n!} p_i(x,t)\mathrm{d}x, \quad n = 0,1,2,\ldots.$$

In particular, the steady-state distributions of the discrete and continuous models are also related by the Poisson representation.

The above theorem shows that for unregulated genes, if the burst size distributions and the initial distributions for the discrete and continuous models are related by the Poisson representation, then both the time-dependent and steady-state distributions for the two models are also related by the Poisson representation, no matter how complex the model is.

Theorem 1 has many applications. Over the past two decades, there have been numerous studies focusing on the analytical solutions of various discrete and continuous gene expression models. However, there are still many models whose closed-form solutions are still unknown. Let $p_n(t) = \sum_{i=0}^L p_{i,n}(t)$ and $p(x,t) = \sum_{i=0}^L p_i(x,t)$ denote the

protein number distributions for the discrete and continuous models, respectively. Since the distributions for the two models are related by the Poisson representation, once the exact solution for one model is known, we can use the Poisson representation to recover the exact solution for the other model. Specifically, according to Theorem 1, the generating function $f(z,t)$ of the discrete distribution $p_n(t)$ and the Laplace transform $\hat{p}(z,t)$ of the continuous distribution $p(x,t)$ are linked by $\hat{p}(z,t) = f(1-z,t)$. Once the generating function of the discrete model is known, the Laplace transform of the continuous model is also known. Then taking the inverse Laplace transform of $f(1-z,t)$ with respect to z gives the protein number distribution of the continuous model. On the other hand, once the Laplace transform $\hat{p}(z,t)$ of the continuous model is known, the generating function of the discrete model is given by $f(z,t) = \hat{p}(1-z,t)$. Finally, taking derivatives of $f(z,t)$ at $z=0$ gives the protein number distribution of the discrete model, i.e.

$$p_n(t) = \frac{1}{n!}\frac{\partial^n}{\partial z^n}\bigg|_{z=0} f(z,t). \tag{15}$$

Next we apply the above idea to compute the steady-state and time-dependent distributions for some complex gene expression models. All the exact solutions derived below are novel and have not been obtained in the literature.

3.2 Steady-state solution for a multi-state bursty model

Consider a multi-state gene expression system described by the following reactions:

$$\begin{aligned} G_i &\xrightarrow{k_{ij}} G_j, \quad i,j = 0,1,\ldots,L, \\ G_0 &\xrightarrow{u_0 p^k(1-p)} G_0 + kP, \quad k \geq 1, \quad P \xrightarrow{1} \varnothing. \end{aligned} \tag{16}$$

Here we assume that the gene can switch between $L+1$ different states $G_0, G_1, \ldots, G_L$. The gene has a unique active state G_0 and multiple inactive states $G_1, \ldots, G_L$. The gene product P (mRNA or protein) is produced in a bursty manner when the gene is active with the burst size having a geometric distribution, while it is not produced when the gene is inactive. Without loss of generality, we assume that the gene product is degraded with rate $d = 1$. In fact, the discrete model of the above system has been analytically solved in steady-state conditions [58, 59]. However, thus far, the exact steady-state solution for the corresponding continuous model is still unknown.

The consideration behind this model is as follows. Note that in the two-state model (12), the durations of the active and inactive states both have an exponential distribution. This assumption is reasonable for bacterial [60] and yeast [61] cells. However, recent studies have shown that in mammalian cells, the inactive periods for many genes may have a non-exponential distribution with a nonzero peak [27, 28], since the activation of the promoter is a complex multi-step biochemical process due to chromatin remodeling and the binding and release of transcription factors. This indicates that the gene dynamics in the inactive period may contain multiple exponential steps and in sum, the gene would undergo a multiple-state switching process.

Next we compute the steady-state gene product number distribution of the continuous model (see Fig. 2(a) for an illustration), which is a switching SDE. Let $K = (k_{ij})_{(L+1)\times(L+1)}$ be the generator matrix of gene state switching, and let H be the $L \times L$ matrix obtained by removing the first row and the first column of K. We assume that K is irreducible, which means that for any pair of gene states G_i and G_j, there exists a transition path

$$G_i = G_{i_1} \to G_{i_2} \to G_{i_3} \cdots \to G_{i_n} = G_j,$$

from G_i to G_j with positive switching rates, i.e. $k_{i_1 i_2} k_{i_2 i_3} \cdots k_{i_{n-1} i_n} > 0$. Since K is a generator matrix, one eigenvalue of K must be zero. Let $\alpha_1, \ldots, \alpha_L$ denote all the nonzero eigenvalues of $-K$ and let $\beta_1, \ldots, \beta_L$ denote all the eigenvalues of $-H$. It follows from the Perron-Frobenius theorem and the irreducibility of K that α_i and β_i all

have positive real parts [15]. According to [58, 59], the generating function of the steady-state distribution p_n^{ss} of the discrete model is given by

$$f^{ss}(z) = {}_{L+1}F_L\left(\gamma_1, \ldots, \gamma_{L+1}; \alpha_1, \ldots, \alpha_L; B(z-1)\right),$$

where $B = p/(1-p)$ is the mean burst size and $\gamma_1, \ldots, \gamma_{L+1}$ are constants satisfying the equations

$$\sigma_k\left(\gamma_1, \ldots, \gamma_{L+1}\right) = \sigma_k\left(\alpha_1, \cdots, \alpha_L\right) + r\sigma_{k-1}\left(\beta_1, \cdots, \beta_L\right), \quad k = 1, \ldots, L+1,$$

with σ_k being the kth elementary symmetric polynomials defined by

$$\sigma_k\left(\alpha_1, \ldots, \alpha_L\right) = \sum_{1 \le i_1 < \cdots < i_k \le L} \alpha_{i_1}\alpha_{i_2}\ldots\alpha_{i_k}.$$

From Theorem 1, the steady-state gene product number distribution $p^{ss}(x)$ of the continuous model has the Laplace transform $\hat{p}^{ss}(z) = f^{ss}(1-z)$. Taking the inverse Laplace transform and using [62, Eq. 3.36.1.1], we obtain

$$p^{ss}(x) = \sum_{j=1}^{L+1} B^{-\gamma_j} x^{\gamma_j - 1} \Gamma\left[\begin{array}{c} (\beta_L), (\gamma_{L+1})' - \gamma_j \\ (\gamma_{L+1}), (\beta_L) - \gamma_j \end{array}\right] {}_LF_L\left(\gamma_j - (\beta_L) + 1; \gamma_j - (\gamma_{L+1})' + 1; -\frac{x}{B}\right), \tag{17}$$

where

$$(\beta_L) = \beta_1, \ldots, \beta_L, \quad \Gamma\left[\begin{array}{c} a_1, \ldots, a_m \\ b_1, \ldots, b_n \end{array}\right] = \frac{\prod_{l=1}^m \Gamma\left(a_l\right)}{\prod_{l=1}^n \Gamma\left(b_l\right)},$$

$$(\gamma_L)' - \gamma_j = \gamma_1 - \gamma_j, \ldots, \gamma_{j-1} - \gamma_j, \gamma_{j+1} - \gamma_j, \ldots, \gamma_L - \gamma_j.$$

In the special case of $L = 1$, the system reduces to a two-state model and the above result coincides with the one obtained in [18]. To test our analytical solution, we compare it with the numerical solution computed via the stochastic simulation algorithm (SSA) [63] in the case of $L = 2$ under a set of biologically relevant parameters (Fig. 2(b)). Clearly, the copy number distribution is bimodal and the exact solution coincides perfectly with the numerical one.

3.3 Time-dependent solution for a two-state bursty model

Consider a two-state gene expression system described by the following reactions:

$$G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \quad G^* \xrightarrow{up^k(1-p)} G^* + kP, \quad k \ge 1, \quad P \xrightarrow{d} \varnothing,$$

where the gene switches between an active state G and an inactive state G^*, and the gene product is produced in a bursty manner when the gene is active with the burst size having a geometric distribution. In fact, the discrete model of the above system has been analytically solved in time [59, 64]. However, thus far, the exact time-dependent solution for the corresponding continuous model is still unknown.

We then focus on the time-dependent gene product number distribution of the continuous model (see Fig. 2(c) for an illustration), which is a switching SDE. The generating function of the time-dependent distribution $p_n(t)$ of the discrete model has been obtained in [64] and is given by

$$\begin{aligned} f(z,t) = {} & \psi_0(ze^{-dt}){}_2F_1\left(\alpha_1, \alpha_2; \beta; B(z-1)\right) \\ & + \psi_1(ze^{-dt})(B(z-1))^{1-\beta}{}_2F_1(1-s-\alpha_1, 1-s-\alpha_2; 2-\beta; B(z-1)), \end{aligned} \tag{18}$$

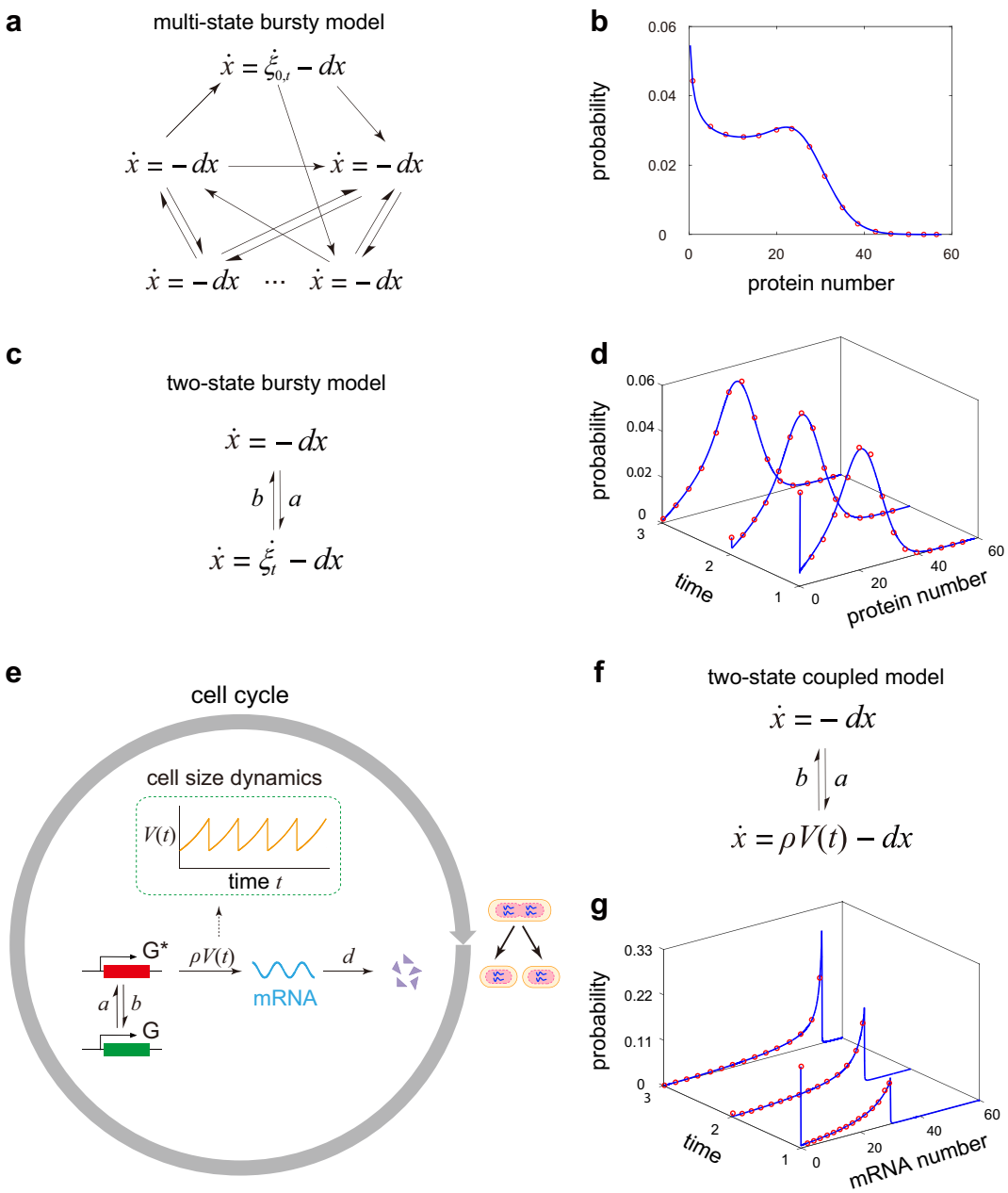


Figure 2: **Analytical solutions for three continuous models of stochastic gene expression**. **(a)** A multi-state bursty model of gene expression. The switching between multiple states can be arbitrary, and in each gene state, the gene product is produced in a bursty manner with the burst size having an exponential distribution. **(b)** Comparison between the exact steady-state distribution of the continuous model given by Eq. (17) (blue curves) and the numerical one obtained from the SSA (red circles) for the model illustrated in (a). The model parameters are chosen as $L = 2, u_0 = 30, B = 1, d = 1, k_{12} = 0.35, k_{13} = 0.35, k_{21} = 0.8, k_{23} = 0.35, k_{31} = 0.8, k_{32} = 0.35$. **(c)** A two-state bursty model of gene expression. In the active state, the gene product is produced in a bursty manner with the burst size having an exponential distribution. **(d)** Comparison between the exact time-dependent distributions of the continuous model given by Eq. (18) (blue curves) and the numerical ones obtained from the SSA (red circles) for the model illustrated in (c). The parameters are chosen as $u = 30, B = 1, d = 1, a = 3.4, b = 0.5$. The time points are chosen to be $t = 1, 2, 3$. **(e)** Schematic of a coupled dynamic model of gene expression and cell size. Within each cell cycle, the cell size grows exponentially and we use an extended telegraph model to describe gene expression fluctuations with the synthesis rate scaling with cell size. **(f)** The coupled model of gene expression and cell size within each cell cycle. **(g)** Comparison between the exact time-dependent distributions of the continuous model given by Eq. (20) (blue curves) and the numerical ones obtained from the SSA (red circles) for the model illustrated in (e). The parameters are chosen as $d = 1, g = 0.1, \rho = 30, V_b = 1, a = 2.85, b = 0.5, t = 1, 2, 3$. In (d) and (g), the initial gene product number is zero and the gene initially starts from the inactive state.

where $B = p/(1-p)$ is the mean burst size and

$$\begin{aligned}\psi_0(z) &= \frac{1}{1-\beta}[sB(z-1)_2F_1(1-s-\alpha_1, 2-s-\alpha_2; 1-\beta; B(z-1)) \\ &\quad + (1-\beta)[1-B(z-1)]_2F_1(1-s-\alpha_1, 1-s-\alpha_2; 1-\beta; B(z-1))], \\ \psi_1(z) &= \frac{s^2(B(z-1))^\beta(1-B(z-1))^s}{\beta(\beta-1)}{}_2F_1(\alpha_1, \alpha_2; \beta+1; B(z-1)),\end{aligned}$$

with the parameters α_1, α_2, β, and s given by

$$\alpha_1 + \alpha_2 = \frac{a+b+u}{d}, \quad \alpha_1\alpha_2 = \frac{au}{d^2}, \quad \beta = \frac{a+b}{d}, \quad s = \frac{u}{d}.$$

From Theorem 1, the time-dependent gene product number distribution $p(x,t)$ of the continuous model has the Laplace transform $\hat{p}(z,t) = f(1-z,t)$. Taking the inverse Laplace transform of $f(1-z,t)$ gives the distribution of the continuous model, i.e.

$$p(x,t) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{-iT}^{iT} f(1-z,t)e^{zx}\mathrm{d}z.$$

Numerically, we may replace z by iz in the Laplace transform to obtain the Fourier transform of $p(x,t)$ and then take the inverse Fourier transform to recover $p(x,t)$. Fig. 2(d) compares this exact solution with the SSA. Clearly, they are in perfect agreement with each other.

3.4 Time-dependent solution for a coupled dynamic model of gene expression and cell size

Thus far, all the analytically solutions obtained are based on models without upstream cellular drive, i.e. models whose reaction rates are time-independent. Next we consider a complex gene expression model with time-dependent synthesis rate. The model couples gene expression dynamics to cell size dynamics and cell cycle events [65–68], and is described as follows (see Fig. 2(e) for an illustration).

1) Let T denote the cell cycle duration and let $V(t)$ denote the cell volume at time t. We assume that cell volume grows exponentially within each cell cycle, i.e. $V(t) = V_b e^{gt}$ for any $0 \le t \le T$, where V_b is the cell volume at birth and g is the growth rate. The exponential growth of cell volume is commonly observed in various cell types [69]. For simplicity, we assume that the doubling time T and the growth rate g do not involve any stochasticity.

2) In each cell cycle, we use the following two-state model to describe gene expression dynamics:

$$G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b} G, \quad G^* \xrightarrow{\rho V(t)} G^* + P, \quad P \xrightarrow{d} \varnothing, \tag{19}$$

where the gene switches between an active state G and an inactive state G^*, and the gene product is produced when the gene is active with the synthesis rate scaling with cell volume. In fission yeast, mammalian cells, and plant cells, there is evidence that the mean mRNA or protein number for many genes scales with cell size in order to maintain approximately constant concentrations, a phenomenon known as concentration homeostasis [69, 70]. This is due to a coordination of the synthesis rate with cell volume, as assumed in Eq. (19).

3) At cell division, the mother cell is divided into two daughter cells. The volumes of the two daughter cells are assumed to be the same and exactly one half of the volume before division (of course there is some stochasticity in the partitioning of cell size [71, 72] which we are here ignoring). Moreover, we assume that each gene product molecule has probability $1/2$ of being allocated to each daughter cell.

Recently, the discrete model of the above system has been solved analytically and the full time-dependent distributions of the gene product number across cell cycles have been obtained [68]. However, the time-dependent solution for the corresponding continuous model is still unknown. For simplicity, here we only focus on the gene

expression dynamics within a cell cycle under arbitrary initial conditions. Note that the reaction scheme (19) is a special case of our general model (14) with the synthesis rate being time-dependent. According to [68], the generating function of the gene product number distribution of the discrete model within a cell cycle is given by

$$f(z,t) = L_0(z,t) f_0(e^{-dt}z, 0) + L_1(z,t) f_1(e^{-dt}z, 0), \quad t \in [0,T], \tag{20}$$

where $f_i(z,0) = \sum_{n=0}^{\infty} p_{i,n}(0) z^n$, $i = 0, 1$ are the generating functions at $t = 0$ which can be determined by the initial conditions, and the functions L_i, $i = 0, 1$ are given by

$$\begin{aligned} L_0(z,t) = \Big[&{}_1F_1(1+\alpha_1-\alpha_2; 1-\alpha_2; ue^{-dt}(z-1)) {}_1F_1(\alpha_1; \alpha_2; ue^{gt}(z-1)) + \frac{auz}{\alpha_2(\alpha_2-1)} e^{-rt} \\ &\times {}_1F_1(1+\alpha_1; 1+\alpha_2; ue^{-dt}(z-1)) {}_1F_1(1+\alpha_1-\alpha_2; 2-\alpha_2; ue^{gt}(z-1)) \Big] e^{-ue^{-dt}(z-1)}, \\ L_1(z,t) = \Big[&{}_1F_1(\alpha_1-\alpha_2; 1-\alpha_2; ue^{-dt}(z-1)) {}_1F_1(\alpha_1; \alpha_2; ue^{gt}(z-1)) - \frac{(\alpha_2-\alpha_1)u(z-1)}{\alpha_2(\alpha_2-1)} e^{-rt} \\ &\times {}_1F_1(\alpha_1; 1+\alpha_2; ue^{-dt}(z-1)) {}_1F_1(1+\alpha_1-\alpha_2; 2-\alpha_2; ue^{gt}(z-1)) \Big] e^{-ue^{-dt}(z-1)}, \end{aligned}$$

with the parameters r, α_1, α_2, and u given by

$$r = a + b - g, \quad \alpha_1 = \frac{a}{d+g}, \quad \alpha_2 = \frac{a+b}{d+g}, \quad u = \frac{\rho V_b}{d+g}.$$

Note that when the growth rate $g = 0$, the synthesis rate $\rho V(t) = \rho V_b$ is a constant and the time-dependent solution given in Eq. (20) coincides with that of the classical telegraph model [23, 73].

We then focus on the time-dependent gene product number distribution of the continuous model (see Fig. 2(f) for an illustration), which is a switching SDE. From Theorem 1, the time-dependent distribution $p(x,t)$ of the continuous model within a cell cycle has the Laplace transform $\hat{p}(z,t) = f(1-z,t)$. Fig. 2(g) shows that the analytical solution agrees perfectly with the SSA. Here we assume that initially there are no gene product molecules in the cell and the gene is inactive. This mimics the situation where the gene has been silenced by some repressor over a period of time such that all gene product molecules have been removed via degradation; at time $t = 0$, the repressor is removed and we study how gene expression recovers.

4 Poisson representation for autoregulated genes

4.1 A counterexample

We have seen that for unregulated genes, the gene product number distributions of the discrete and continuous models are always related by the Poisson representation. A natural question is whether this also holds for regulated genes. To answer this, we focus on a multi-state model of autoregulatory gene expression, which is described by the following reactions:

$$\begin{gathered} G_i \xrightarrow{k_{ij}(n,t)} G_j, \quad i,j = 0,1,\ldots,L, \\ G_i \xrightarrow{u_i(n,t)} G_i + P, \quad i = 0,1,\ldots,L, \quad P \xrightarrow{d(n,t)} \varnothing, \end{gathered} \tag{21}$$

where $G_0, G_1, \ldots, G_L$ are $L+1$ different gene states and P is the corresponding protein. Note that these reactions are the same as those in Eq. (5). Due to autoregulation, following [74], we allow all rate constants to depend on the protein number n. If the gene is unregulated, then all rate constants are independent of n. For simplicity, here we do not take protein bursting into account. However, the results can be generalized easily to the case of bursty protein synthesis [75].

Next we consider the discrete and continuous models for the autoregulatory system. Let $p_{i,n}(t)$ denote the probability of having n protein copies when the gene is in state i. The evolution of the discrete model is governed by the following set of CMEs:

$$\begin{aligned}\dot{p}_{i,n}(t) = &[u_i(n-1,t)p_{i,n-1}(t) - u_i(n,t)p_{i,n}(t)] + [d(n+1,t)(n+1)p_{i,n+1}(t) - d(n,t)np_{i,n}(t)] \\ &+ \Big[\sum_{j\neq i} k_{ji}(n,t)p_{j,n}(t) - \sum_{j\neq i} k_{ij}(n,t)p_{i,n}(t)\Big],\end{aligned} \tag{22}$$

where the three terms on the right-hand side represent protein synthesis, protein degradation, and gene state switching, respectively. For the continuous model, in each gene state G_i, we use the following ODE to model the gene expression dynamics:

$$\dot{x} = u_i(x,t) - d(x,t)x,$$

where the two terms on the right-hand side characterize synthesis and degradation of the protein. Since the gene can switch among different states, the gene expression dynamics can be modeled continuously as a hybrid switching system of ODEs. Let $p_i(x,t)$ denote the probability density of the protein abundance x when the gene is in state i. The evolution of the switching ODE model is governed by the Kolmogorov forward equation

$$\begin{aligned}\partial_t p_i(x,t) = &-\partial_x\left[u_i(x,t)p_i(x,t)\right] + \partial_x\left[d(x,t)xp_i(x,t)\right] \\ &+ \Big[\sum_{j\neq i} k_{ji}(x,t)p_j(x,t) - \sum_{j\neq i} k_{ij}(x,t)p_i(x,t)\Big],\end{aligned} \tag{23}$$

where the three terms on the right-hand side describe protein synthesis, protein degradation, and gene state switching, respectively.

Unfortunately, for regulated genes, the protein number distributions for the discrete and continuous models in general fail to be related by the Poisson representation. This can be seen from the following example.

Example 4. Consider a negative autoregulatory gene circuit (see Fig. 3(a) for an illustration) with the following reactions [76, 77]:

$$G \xrightarrow{a} G^*, \quad G^* \xrightarrow{b+\nu n} G, \quad G^* \xrightarrow{u} G^* + P, \quad P \xrightarrow{d} \varnothing, \tag{24}$$

where G and G^* denote the inactive and active states of the gene and P denotes the corresponding protein. Due to negative autoregulation, the switching rate from G^* to G depends on the protein number n. Here a and b are the spontaneous switching rates between the two gene states and ν characterizes the strength of negative feedback. It follows from [76, 77] that the steady-state generating function of the protein number distribution for the discrete model is given by

$$f^{ss}(z) = \frac{{}_1F_1\left(\alpha;\beta;w(z-z_0)\right)}{{}_1F_1\left(\alpha;\beta;w(1-z_0)\right)},$$

where

$$\alpha = \frac{a}{d}, \quad \beta = \frac{a+b}{d+\nu} + \frac{\nu u}{(d+\nu)^2}, \quad w = \frac{u}{d+\nu}, \quad z_0 = \frac{d}{d+\nu}.$$

Since the generating function and the Poisson kernel are related by $\hat{\rho}^{ss}(z) = f^{ss}(1-z)$ (see Eq. (2)), it is easy to check that the Poisson kernel of the protein number distribution is given by [62, Eq. 3.33.1.1]

$$\rho^{ss}(x) = \frac{\Gamma(\beta)w^{1-\beta}}{\Gamma(\alpha)\Gamma(\beta-\alpha)\,{}_1F_1\left(\alpha;\beta;w(1-z_0)\right)} x^{\alpha-1}(w-x)^{\beta-\alpha-1}e^{(1-z_0)x}, \quad x < w. \tag{25}$$

On the other hand, the steady-state protein number distribution of the corresponding continuous model has been computed in [18] and is given by

$$p^{ss}(x) = \frac{\Gamma(\tilde{\beta})\tilde{w}^{1-\tilde{\beta}}}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta}-\tilde{\alpha})_1F_1(\tilde{\alpha};\tilde{\beta};\tilde{w}\lambda_0)} x^{\tilde{\alpha}-1}(\tilde{w}-x)^{\tilde{\beta}-\tilde{\alpha}-1}e^{\lambda_0 x}, \quad x < \tilde{w}, \tag{26}$$

where

$$\tilde{\alpha} = \frac{a}{d}, \quad \tilde{\beta} = \frac{a+b}{d} + \frac{\nu u}{d^2}, \quad \tilde{w} = \frac{u}{d}, \quad \lambda_0 = \frac{\nu}{d}.$$

Eqs. (25) and (26) look similar. Comparing these two equations in detail, we find that $\rho^{ss}(x) \neq p^{ss}(x)$, which means the steady-state solutions for the discrete and continuous models are not related by the Poisson representation. However, when the feedback strength is much smaller than the degradation rate, i.e. $\nu \ll d$, we have

$$\tilde{\alpha} \approx \alpha, \quad \tilde{\beta} \approx \beta, \quad \tilde{w} \approx w, \quad 1 - z_0 \approx \lambda_0.$$

In this case, the steady-state solutions for the two models are approximately related by the Poisson representation. We emphasize that the condition $\nu \ll d$ is not very strong and is satisfied for a wide range of biological systems. To see this, note that the feedback contribution νn to the gene switching rate from G^* to G usually has the same order as the spontaneous contribution b. This suggests that b/ν should have the same order as the mean protein number in the active gene state, i.e. $\langle n \rangle_{\text{active}} = u/d$. Hence d/ν should have the same order as u/b. Recent single-cell experiments on transcriptional bursting of mammalian cells have shown that u/b is relatively large. In [27], the authors monitored protein expression in mouse fibroblasts using live-cell imaging and found that the values of a, b, and u for different genes are typically on the order of 0.01/min, 0.1/min, and 1/min, respectively (see Figs. 1(D), 1(E), and S8 of [27] for details). These experimental results imply that the condition $\nu \ll d$ is indeed satisfied in many biological systems.

From the above example, we have seen that when the gene is regulated, the solutions for the discrete and continuous models are not generally related by the Poisson representation. However, the Poisson representation can still build a bridge between the two models in some special cases. Next we will show that in the limit of large protein numbers, the distributions for the discrete and continuous models are approximately related by the Poisson representation.

4.2 Main results

To discuss the limit case of large protein numbers, we introduce a new parameter K which stands for the system size, with $K \to \infty$ corresponding to a macroscopic scale. In previous papers, the system size K is often understood to be the cell volume [5] or the typical protein number in the active gene state [78, 79]. With this new parameter, the limit of large protein numbers corresponds to the limit of $K \to \infty$. Since the protein number is mainly controlled by the synthesis rate, it is natural to assume that the protein synthesis rate is proportional to K. Specifically, we assume that the synthesis rates in all gene states are proportional to K, while the other rate parameters are all functions of the protein concentration, i.e.

$$u_i(n,t) = K\bar{u}_i(y,t), \quad k_{ij}(n,t) = \bar{k}_{ij}(y,t), \quad d(n,t) = \bar{d}(y,t), \tag{27}$$

where $y = n/K$ is the protein concentration, and $\bar{u}_i(y,t), \bar{k}_{ij}(y,t)$, and $\bar{d}(y,t)$ are functions independent of K. We will next prove that the protein number distributions for the discrete and continuous models are related by the Poisson representation when $K \gg 1$.

To this end, we first focus on the dynamics of the protein concentration. Let $\bar{p}_i(y,t)$ denote the probability density of the protein concentration $y = n/K$ when the gene is in state i. We next derive the dynamic equation of the protein

concentration using the discrete model. When $K \gg 1$, the probability density $\bar{p}_i(y,t)$ of the protein concentration and the probability distribution $p_{i,n}(t)$ of the protein number should be related by [18]

$$Kp_{i,n}(t) \approx \bar{p}_i\left(\frac{n}{K},t\right). \tag{28}$$

Multiplying K on both sides of Eq. (22) and applying Eq. (27), we obtain

$$\begin{aligned} K\partial_t p_{i,n}(t) = K^2 &\left[\bar{u}_i\left(\frac{n-1}{K},t\right)p_{i,n-1}(t) - \bar{u}_i\left(\frac{n}{K},t\right)p_{i,n}(t)\right] \\ &+ K\left[\bar{d}\left(\frac{n+1}{K},t\right)(n+1)p_{i,n+1}(t) - \bar{d}\left(\frac{n}{K},t\right)np_{i,n}(t)\right] \\ &+ K\Big[\sum_{j\neq i}\bar{k}_{ji}\left(\frac{n}{K},t\right)p_{j,n}(t) - \sum_{j\neq i}\bar{k}_{ij}\left(\frac{n}{K},t\right)p_{i,n}(t)\Big]. \end{aligned} \tag{29}$$

Taking the limit of $K \to \infty$ in the above equation and applying Eq. (28), we find that the evolution of the protein concentration is governed by

$$\begin{aligned} \partial_t\bar{p}_i(y,t) = &-\partial_y\left[\bar{u}_i(y,t)\bar{p}_i(y,t)\right] + \partial_y[\bar{d}(y,t)y\bar{p}_i(y,t)] \\ &+ \Big[\sum_{j\neq i}\bar{k}_{ji}(y,t)\bar{p}_j(y,t) - \sum_{j\neq i}\bar{k}_{ij}(y,t)\bar{p}_i(y,t)\Big]. \end{aligned} \tag{30}$$

Since $\bar{u}_i(y,t), \bar{k}_{ij}(y,t)$, and $\bar{d}(y,t)$ are all independent of K, the protein concentration distribution $\bar{p}_i(y,t)$ is also independent of K when $K \gg 1$.

Recall that for the continuous model, the evolution of the protein number is described by Eq. (23). Interestingly, comparing Eq. (30) with Eq. (23), we find that the probability density $\bar{p}_i(y,t)$ of the protein concentration and the probability density $p_i(x,t)$ of the protein number are related by

$$\bar{p}_i(y,t) = Kp_i(Ky,t). \tag{31}$$

To proceed, let $f_i(z,t) = \sum_{n=0}^{\infty} p_{i,n}(t)z^n$ be the generating function of the discrete distribution $p_{i,n}(t)$. When $K \gg 1$, applying the two scaling relations given in Eqs. (28) and (31), we obtain

$$\begin{aligned} f_i\left(1-\frac{z}{K},t\right) &= \sum_{n=0}^{\infty} p_{i,n}(t)e^{n\log\left(1-\frac{z}{K}\right)} \approx \sum_{n=0}^{\infty} p_{i,n}(t)e^{-z\frac{n}{K}} \\ &\approx \sum_{n=0}^{\infty}\bar{p}_i\left(\frac{n}{K},t\right)e^{-z\frac{n}{K}}\frac{1}{K} \approx \int_0^{\infty}\bar{p}_i(y,t)e^{-zy}\mathrm{d}y \\ &= \int_0^{\infty} p_i(x,t)e^{-z\frac{x}{K}}\mathrm{d}x = \hat{p}_i\left(\frac{z}{K},t\right), \end{aligned} \tag{32}$$

where we have used the fact that $\log(1+x) \approx x$ when $|x| \ll 1$ and the fact that a Riemann sum converges to a Riemann integral as the partition size tends to zero. This shows that the generating function of the discrete distribution $p_{i,n}(t)$ and the Laplace transform of the continuous distribution $p_i(x,t)$ are approximately equal when $K \gg 1$. Hence, for autoregulatory systems, the time-dependent solutions of the discrete and continuous models are approximately related by the Poisson representation in the limit of large protein numbers, i.e.

$$p_{i,n}(t) \approx \int_0^{\infty}\frac{e^{-x}x^n}{n!}p_i(x,t)\mathrm{d}x. \tag{33}$$

4.3 Numerical validation

We have shown that in the limit of large protein numbers, the solutions for the discrete and continuous models for an autoregulatory system are approximately related by the Poisson representation. In other words, the Poisson kernel of the protein number distribution for the discrete model should be close to the distribution for the continuous model. We emphasize that this result is not true when the mean protein number is small. To test our results, we revisit the negative autoregulatory gene circuit described by Eq. (24). Here we assume that the protein synthesis rate u is proportional to the system size K, i.e. $u = K\bar{u}$. Since the other rate parameters are all functions of the protein concentration, the feedback strength ν should scale with K as $\nu = \bar{\nu}/K$. In the limit of large protein numbers, i.e. $K \gg 1$, it is easy to verify that the parameters given in Eqs. (25) and (26) satisfy

$$\tilde{\alpha} \approx \alpha, \quad \tilde{\beta} \approx \beta, \quad \tilde{w} \approx w, \quad 1 - z_0 \approx \lambda_0.$$

This shows that the Poisson kernel of the protein number distribution for the discrete model is indeed close to the distribution for the continuous model.

Fig. 3(b) compares the Poisson kernel of the discrete distribution given in Eq. (25) and the continuous distribution given in Eq. (26) as K increases. Clearly, the two distributions differ significantly when K is small, while they are in perfect agreement when K is large, as expected. In addition, Fig. 3(c) illustrates the Hellinger distance H between the two distributions as a function of K. Interestingly, we find that $\log_{10} H$ is approximately a linear function of $\log_{10} K$ with slope $\kappa \approx -1$, which implies that the Hellinger distance decays at the speed of $1/K$. This again demonstrates that the Poisson representation becomes valid when protein numbers are large.

4.4 Two-state model with complex feedback regulation

We then apply our results to a complex autoregulatory gene expression model whose analytical solution has not been obtained previously. According to our results, if the exact solution of the continuous model $p(x,t)$ is known, then we can use Poisson representation with Poisson kernel $p(x,t)$ to construct an approximate solution of the corresponding discrete model when protein numbers are large. Similarly, if the exact solution of the discrete model $p_n(t)$ is known, then the Poisson kernel of $p_n(t)$ serves as an approximate solution of the corresponding continuous model in the limit of large protein numbers.

Here we consider the following autoregulatory gene circuit with general feedback regulation:

$$G \xrightarrow{a(n)} G^*, \quad G^* \xrightarrow{b(n)} G, \quad G^* \xrightarrow{u(n)} G^* + P, \quad P \xrightarrow{d(n)} \varnothing, \tag{34}$$

where the protein is produced in a non-bursty manner and all rate constants depend on the protein number n. The protein number distribution for this system has not been solved analytically before. Next we will first derive the exact steady-state distribution of the continuous model by generalizing the result in [47], and then obtain the approximate steady-state distribution of the discrete model by using the Poisson representation.

We first obtain the analytical solution of the continuous model. The autoregulatory gene circuit can be modeled continuously as the following switching ODE:

$$\begin{array}{ll} \dot{x} = u(x) - d(x)x & \text{(gene is active)}, \\ \quad a(x) \Big\uparrow \Big\downarrow b(x) & \\ \dot{x} = -d(x)x & \text{(gene is inactive)}. \end{array}$$

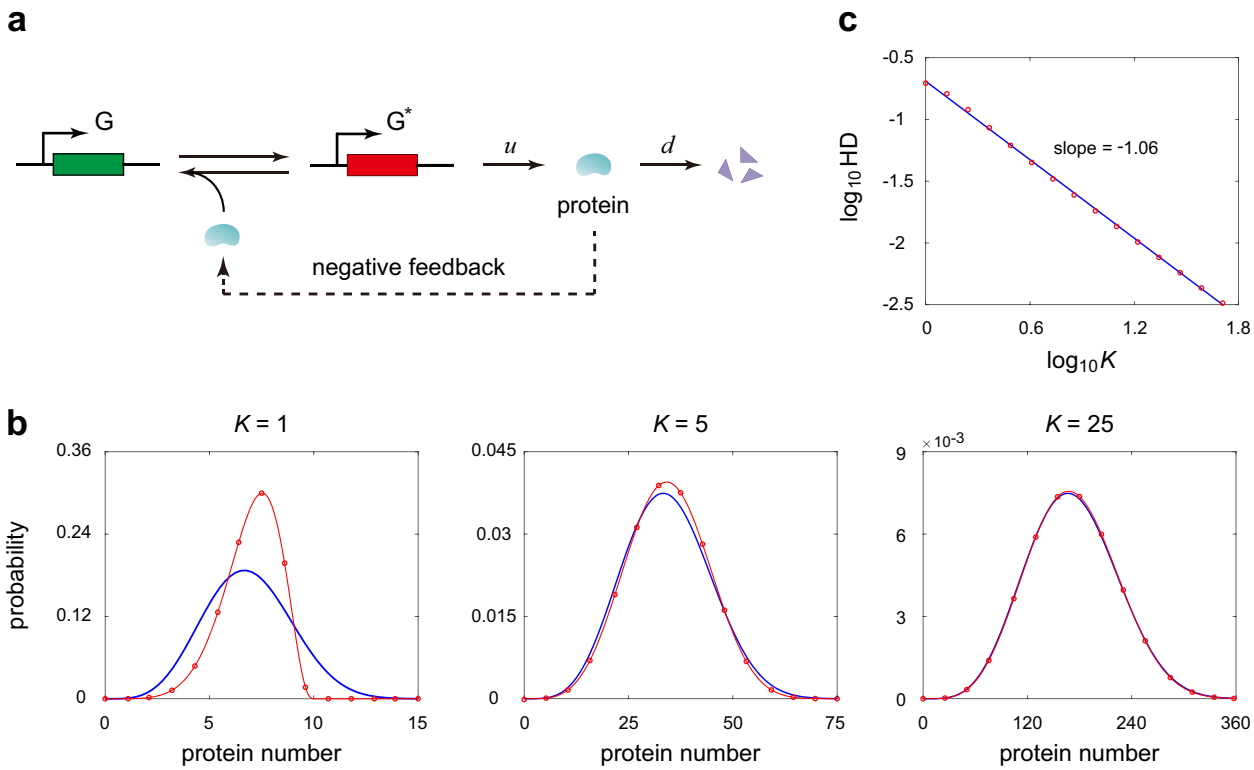


Figure 3: **Poisson representation for a negative feedback loop**. **(a)** A negative autoregulatory gene circuit, where the binding of the protein to the gene inhibits the expression of itself. **(b)** Comparison of the Poisson kernel $\rho^{ss}(x)$ of the discrete model given by Eq. (25) (red circles and red curves) and the exact steady-state distribution $p^{ss}(x)$ of the continuous model given by Eq. (26) (blue curves) as the system size K increases. The parameters are chosen as $a = 5, b = 2, d = 1, \bar{u} = 25, \bar{\nu} = 1.5$. **(c)** The logarithm of the Hellinger distance (HD) between $\rho^{ss}(x)$ and $p^{ss}(x)$ is approximately a linear function of the logarithm of K with slope -1.

For simplicity, we assume that the ODE $\dot{x} = u(x) - d(x)x$ has an unique globally asymptotic stable equilibrium x^* with $x \in [0, \infty]$. Let $p_0(x,t)$ and $p_1(x,t)$ denote the probability densities of the protein number when the gene is in the inactive and active states, respectively. Then the evolution of $p_0(x,t)$ and $p_1(x,t)$ is governed by

$$\begin{aligned} \partial_t p_0 &= \partial_x[d(x)xp_0] - a(x)p_0 + b(x)p_1, \\ \partial_t p_1 &= \partial_x[d(x)xp_1 - u(x)p_1] + a(x)p_0 - b(x)p_1. \end{aligned} \tag{35}$$

When the system reaches the steady state, summing the two equations in Eq. (35) yields

$$\frac{\mathrm{d}}{\mathrm{d}x}\left\{[u(x) - d(x)x]p_1^{ss}(x) - d(x)xp_0^{ss}(x)\right\} = 0.$$

This shows that

$$[u(x) - d(x)x]p_1^{ss}(x) - d(x)xp_0^{ss}(x) = M, \tag{36}$$

where M is a constant. Since $p_0^{ss}(x)$ and $p_1^{ss}(x)$ vanish at infinity, we must have $M = 0$. Hence we can define a function

$$\psi(x) := d(x)xp_0^{ss}(x) = [u(x) - d(x)x]p_1^{ss}(x).$$

It then follows from Eq. (35) that

$$\psi'(x) = \left[\frac{a(x)}{d(x)x} - \frac{b(x)}{u(x) - d(x)x}\right]\psi(x).$$

Integrating the above equation yields

$$\psi(x) = \tilde{C} \exp\left\{ \int_{x_0}^{x} \left[\frac{a(\tau)}{d(\tau)\tau} - \frac{b(\tau)}{u(\tau) - d(\tau)\tau} \right] \mathrm{d}\tau \right\},$$

where $0 < x_0 < x^*$ is arbitrarily chosen and $\tilde{C}$ is a normalization constant. Then the steady-state protein number distribution of the continuous model can be solved analytically as

$$\begin{aligned} p^{ss}(x) &= p_0^{ss}(x) + p_1^{ss}(x) \\ &= \tilde{C} \left(\frac{1}{d(x)x} + \frac{1}{u(x) - d(x)x} \right) \exp\left\{ \int_{x_0}^{x} \left[\frac{a(\tau)}{d(\tau)\tau} - \frac{b(\tau)}{u(\tau) - d(\tau)\tau} \right] \mathrm{d}\tau \right\}, \quad 0 < x < x^*. \end{aligned} \tag{37}$$

According to our results, the Poisson representation links the discrete and continuous models in the limit of large protein numbers. Hence when protein numbers are large, the steady-state protein number distribution for the discrete model has the following generating function:

$$f^{ss}(z) = \sum_{n=0}^{\infty} p_n^{ss} z^n \approx \int_0^{\infty} p^{ss}(x) e^{-(1-z)x} \mathrm{d}x, \tag{38}$$

where $p^{ss}(x)$ is given by Eq. (37). Taking the derivatives of $f^{ss}(z)$ at $z = 0$ gives the approximate steady-state distribution p_n^{ss} of the discrete model.

To test the accuracy of this solution, we consider a positive autoregulatory feedback loop with the switching rates having the functional forms of $a(n) = a_0 + a_1 n^2$ and $b(n) = b$. This corresponds to the case where the switching from the inactive to the active state requires the cooperative binding of two protein molecules (Fig. 4(a)). In addition, we assume that the synthesis rate $u(n) = u$ is independent of n, and the protein decays via the Michaelis-Menten propensity function $d(n) = \frac{d_{max} n}{k_m + n}$ with d_{max} being the maximum degradation rate [64]. This describes enzyme-catalyzed protein degradation. In this case, the steady-state distribution of the continuous model is given by (see Appendix B for details)

$$\begin{aligned} p^{ss}(x) = C &\left[\frac{1}{d(x)} + \frac{1}{u - d(x)} \right] x^{a_0/d_{max}} (x - \alpha_1)^{bA/d_{max}} (x - \alpha_2)^{B/d_{max}} \\ &\exp\left[\frac{k_m}{d_{max} x} \left(-a_0 + a_1 x^2 + \frac{1}{2} a_1 x^3 \right) \right], \end{aligned} \tag{39}$$

where C is a normalization constant and the parameters α_1, α_2, A, and B are given by

$$\alpha_1 + \alpha_2 = \frac{u}{d_{max}}, \quad \alpha_1 \alpha_2 = -\frac{u k_m}{d_{max}}, \quad A = -\frac{k_m + \alpha_1}{\alpha_2 - \alpha_1}, \quad B = \frac{d_{max} + \alpha_2}{\alpha_2 - \alpha_1}.$$

Finally, we introduce the system size K and assume that the parameters u, a_1, and k_m scale with K as

$$u = K\bar{u}, \quad a_1 = \bar{a}_1 / K^2, \quad k_m = \bar{k}_m K,$$

where $\bar{u}$, $\bar{a}_1$, and $\bar{k}_m$ are constants independent of K. Then Eq. (38) gives the approximate steady-state distribution of the discrete model when $K \gg 1$. Fig. 4(b) compares this approximate solution with the numerical one obtained using the finite-state projection algorithm (FSP) [80]. When performing FSP, we truncate the state space at a large integer N and solve the truncated master equation numerically. The truncation size is chosen as $N = 3K\bar{u}/d$, which guarantees that the probability that the protein number is outside the truncated size is very small and practically can always be ignored. It is clear that the approximate solution deviates from the numerical one for small K, but it becomes increasingly accurate as K becomes larger.

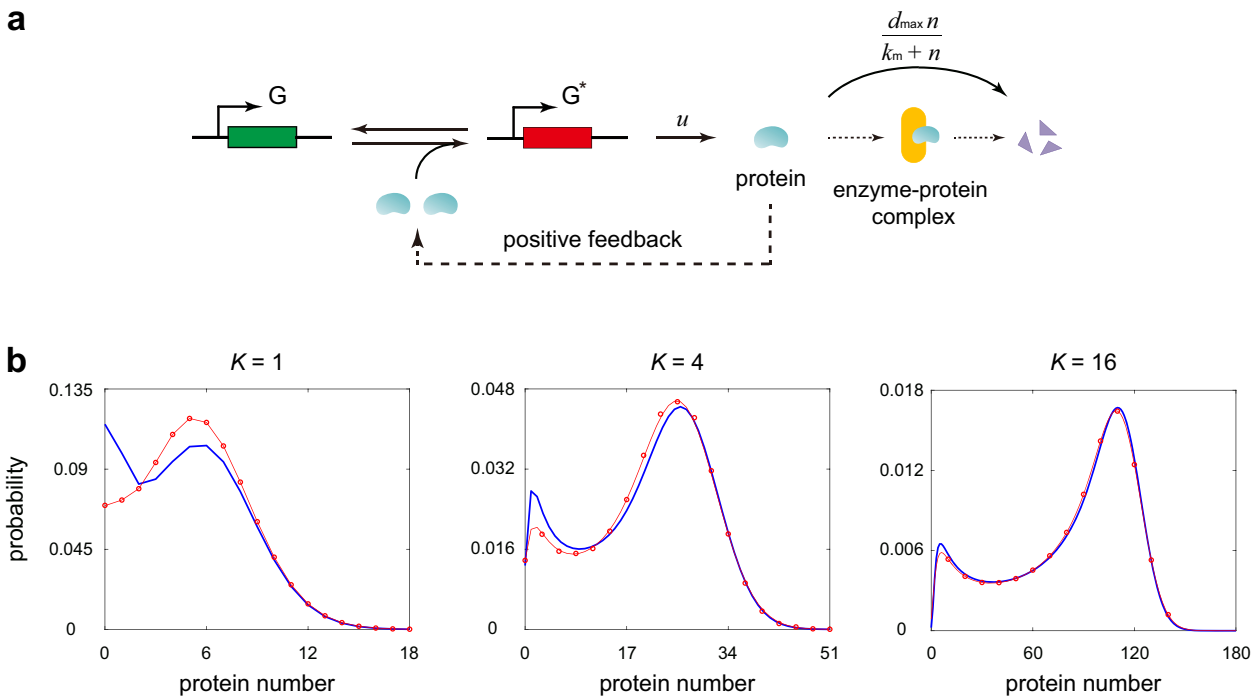


Figure 4: **Poisson representation for a positive autoregulatory gene circuit with cooperative binding and enzyme-catalyzed degradation. (a)** A positive autoregulatory gene circuit. Two protein molecules cooperatively bind to the gene to activate its own expression. The degradation of the protein is enzyme-catalyzed with the rate having the Michaelis-Menten form. **(b)** Comparison of the approximate theoretical distribution of the discrete model given by Eqs. (38) and (39) (red circles and red curves) and the numerical one computed computed using FSP (blue curves). When computing the theoretical distribution of the discrete model, we used the discrete Fourier transform algorithm proposed in [81]. This algorithm can recover the distribution directly from the generating function without calculating the higher order derivatives as in Eq. (15), which is very time-consuming and numerically unstable. The parameters are chosen as $a_0 = 1$, $\bar{a}_1 = 0.2$, $b = 2$, $\bar{u} = 15$, $d_{max} = 2.25$, $\bar{k}_m = 1$.

A phenomenon that deserves special attention is that under enzyme-catalyzed protein degradation, the system is capable of producing a bimodal distribution with two non-zero peaks (middle and right panels in Fig. 4(b)). Note that here protein molecules cannot be produced when the gene is in the inactive state. If the protein is degraded via first-order kinetics with constant rate, then previous studies have shown that when bimodality occurs, the bimodal protein number distribution must have both a zero mode and a non-zero mode, and it can never produce two non-zero modes, even when cooperative protein binding to the gene is taken into account [82]. This suggests that the two non-zero peaks in the protein number distribution observed in Fig. 4(b) is due to Michaelis-Menten-type degradation, rather than due to cooperative protein binding.

5 Generalization to general gene regulatory networks

5.1 Multivariate Poisson representation

In previous sections, we have discussed the Poisson representation for systems with only one gene. Here we will extend the results to a general gene regulatory network, which may involve multiple genes. To this end, we recall the definition of multivariate Poisson representation.

Let $p_{n_1,n_2,\ldots,n_J}$ be the joint probability distribution of a nonnegative integer-valued discrete random vector. We

say that $p_{n_1,n_2,\ldots,n_J}$ has a Poisson representation if it can be represented in the following form:

$$p_{n_1,n_2,\ldots,n_J} = \int_{\mathbb{R}_+^J} \frac{e^{-(x_1+x_2+\cdots+x_J)} x_1^{n_1} x_2^{n_2} \ldots x_J^{n_J}}{n_1! n_2! \ldots n_J!} \rho(x_1, x_2, \ldots, x_J) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_J, \tag{40}$$

where $\rho(x_1, x_2, \ldots, x_J)$ is the joint probability density of a nonnegative continuous random vector. In what follows, we refer to $\rho(x_1, x_2, \ldots, x_J)$ as the Poisson kernel of $p_{n_1,n_2,\ldots,n_J}$. If Eq. (40) is satisfied, we also say that $p_{n_1,n_2,\ldots,n_J}$ and $\rho(x_1, x_2, \ldots, x_J)$ are related by the Poisson representation.

The generating function of the discrete distribution $p_{n_1,n_2,\ldots,n_J}$ is defined by

$$f(z_1, z_2, \ldots, z_J) = \sum_{n_1,n_2,\ldots,n_J=0}^{\infty} p_{n_1,n_2,\ldots,n_J} z_1^{n_1} z_2^{n_2} \ldots z_J^{n_J},$$

and the Laplace transform of the Poisson kernel $\rho(x_1, x_2, \ldots, x_J)$ is defined by

$$\hat{\rho}(z_1, z_2, \ldots, z_J) = \int_{\mathbb{R}_+^J} e^{-(z_1x_1+z_2x_2+\cdots+z_Jx_J)} \rho(x_1, x_2, \ldots, x_J) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_J.$$

Similarly to the univariate case, in terms of the generating function and Laplace transform, Eq. (40) can be converted into the following equivalent form:

$$f(1 - z_1, 1 - z_2, \ldots, 1 - z_J) = \hat{\rho}(z_1, z_2, \ldots, z_J). \tag{41}$$

5.2 Model of gene regulatory networks

Here we consider the stochastic model of a general gene regulatory network involving protein synthesis, protein degradation, gene state switching, and complex feedback controls. Specifically, we assume that the network involves J distinct genes. Each gene can exist in two states: the inactive state G_j and the active state G_j^*. The protein associated with the jth gene is denoted by P_j, and the corresponding protein number is denoted by n_j. Then the network can be described by the following reactions:

$$\begin{gathered} G_j \xrightarrow{k_j^0(n_1,\ldots,n_J,t)} G_j^*, \quad G_j^* \xrightarrow{k_j^1(n_1,\ldots,n_J,t)} G_j, \\ G_j \xrightarrow{u_j^0(t)} G_j + P_j, \quad G_j^* \xrightarrow{u_j^1(t)} G_j^* + P_j, \quad P_j \xrightarrow{d_j(t)} \varnothing, \quad j = 1, 2, \ldots, J, \end{gathered} \tag{42}$$

where the first two reactions describe gene state switching, the middle two reactions describe protein synthesis in the two gene states, and the last reaction describes protein degradation. Here we allow all rate parameters to depend on time t. Due to complex gene regulation mechanisms, the switching rates k_j^0 and k_j^1 may depend on the numbers of all proteins. For example, if the ith gene activates (inhibits) the jth gene, then k_j^0 (k_j^1) must be an increasing function of n_i since the binding of protein P_i to the promoter of gene G_j activates (inhibits) the expression of this gene.

First we focus on the discrete stochastic model of the gene regulatory network. The microstate of the system can be represented by an ordered $2J$-tuple $(i_1, \ldots, i_J, n_1, \ldots, n_J)$, where i_j is the state of the jth gene with $i_j = 0, 1$ corresponding to the inactive and active states, respectively, and n_j is the number of protein P_j. Let $p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)$ denote the probability of having n_j copies of P_j when the jth gene is in state i_j, for each $j = 1, \cdots, J$. Then the stochastic gene expression dynamics can be modeled discretely by a Markov jump process, and the evolution of the

discrete model is governed by the following set of CMEs:

$$\begin{aligned}\dot{p}_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t) = &\sum_{j=1}^{J} d_j(t)\left[(n_j+1)p_{i_1,\ldots,i_J,n_1,\ldots,n_{j-1},n_j+1,n_{j+1},\ldots,n_J}(t) - n_j p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)\right] \\ &+\sum_{j=1}^{J} u_j^{i_j}(t)\left[p_{i_1,\ldots,i_J,n_1,\ldots,n_{j-1},n_j-1,n_{j+1},\ldots,n_J}(t) - p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)\right] \\ &+\sum_{j=1}^{J}\Big[k_j^{1-i_j}(n_1,\ldots,n_J,t)p_{i_1,\ldots,i_{j-1},1-i_j,i_{j+1},\ldots,i_J,n_1,\ldots,n_J}(t) \\ &-k_j^{i_j}(n_1,\ldots,n_J,t)p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)\Big].\end{aligned} \tag{43}$$

Next we consider the continuous model of the network. In each gene state $(i_1,\ldots,i_J)$, the evolution of the network can be modeled by the following set of ODEs:

$$\begin{cases}\dot{x}_1 = u_1^{i_1}(t) - d_1(t)x_1, \\ \dot{x}_2 = u_2^{i_2}(t) - d_2(t)x_2, \\ \ldots \\ \dot{x}_J = u_J^{i_J}(t) - d_J(t)x_J,\end{cases}$$

where x_j denotes the abundance of protein P_j, and the two terms on the right-hand side describe synthesis and degradation of all proteins involved in the network. Since the genes can switch between different states, the whole gene expression dynamics can be modeled continuously as a switching system of ODEs. Let $p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t)$ denote the joint probability density of the numbers of all proteins in the gene state $(i_1,\ldots,i_J)$. Then the evolution of the continuous model is governed by the Kolmogorov forward equation

$$\begin{aligned}\partial_t p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t) = &\sum_{j=1}^{J} d_j(t)\partial_{x_j}[x_j p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t)] - \sum_{j=1}^{J} u_j^{i_j}(t)\partial_{x_j} p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t) \\ &+\sum_{j=1}^{J}\Big[k_j^{1-i_j}(x_1,\ldots,x_J,t)p_{i_1,\ldots,i_{j-1},1-i_j,i_{j+1},\ldots,i_J}(x_1,\ldots,x_J,t) \\ &- k_j^{i_j}(x_1,\ldots,x_J,t)p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t)\Big].\end{aligned} \tag{44}$$

Similarly to autoregulatory systems, we next prove that the joint distributions of the discrete and continuous models are approximately related by the multivariate Poisson representation when the numbers of all proteins are large. To this end, we introduce the system size K with $K \to \infty$ corresponding to a macroscopic scale. Moreover, we assume that the synthesis rates of all proteins scale with K as $u_j^{i_j}(t) = K\bar{u}_j^{i_j}(t)$, with $\bar{u}_j^{i_j}(t)$ being independent of K, and we assume that the switching rates for all genes are functions of the protein concentrations, i.e.

$$k_j^{i_j}(n_1,\ldots,n_J,t) = \bar{k}_j^{i_j}(y_1,\ldots,y_J,t), \tag{45}$$

where $y_j = n_j/K$ is the concentration of P_j, and $\bar{k}_j^{i_j}(y_1,\ldots,y_J,t)$ are functions independent of K.

Next we derive the dynamic equation of protein concentrations from the discrete model. Let $\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t)$ denote the joint probability density of the concentrations of all proteins in the gene state $(i_1,\ldots,i_J)$. When $K \gg 1$, the probability density $\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t)$ of protein concentrations and the probability distribution $p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)$ of protein numbers (for the discrete model) should be related by

$$K^J p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t) \approx \bar{p}_{i_1,\ldots,i_J}\left(\frac{n_1}{K},\ldots,\frac{n_J}{K},t\right). \tag{46}$$

Applying this relation and taking $K \to \infty$ in Eq. (43), we find that the evolution of protein concentrations is governed by

$$\begin{aligned}\partial_t \bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t) = &\sum_{j=1}^{J} d_j(t)\partial_{y_j}[y_j \bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t)] - \sum_{j=1}^{J} \bar{u}_j^{i_j}(t)\partial_{y_j}\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t) \\ &+ \sum_{j=1}^{J}\Big[\bar{k}_j^{1-i_j}(y_1,\ldots,y_J,t)\bar{p}_{i_1,\ldots,i_{j-1},1-i_j,i_{j+1},\ldots,i_J}(y_1,\ldots,y_J,t) \\ &- \bar{k}_j^{i_j}(y_1,\ldots,y_J,t)\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t)\Big].\end{aligned} \tag{47}$$

Recall that the dynamics of the continuous model of the network is described by Eq. (44). Comparing Eq. (47) with Eq. (44), it is easy to see that the probability density $\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t)$ of protein concentrations and the probability density $p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t)$ of protein numbers (for the continuous model) are related by

$$\bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t) = K^J p_{i_1,\ldots,i_J}(Ky_1,\ldots,Ky_J,t). \tag{48}$$

To proceed, let $f_{i_1,\ldots,i_J}(z_1,\ldots,z_J,t) = \sum_{n_1,\ldots,n_J=0}^{\infty} p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t) z_1^{n_1}\ldots z_J^{n_J}$ be the generating function of the discrete distribution $p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)$. When $K \gg 1$, applying Eqs. (46) and (48), we obtain

$$\begin{aligned} f_{i_1,\ldots,i_J}\left(1-\frac{z_1}{K},\ldots,1-\frac{z_J}{K},t\right) &= \sum_{n_1,\ldots,n_J} p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t) e^{\sum_{j=1}^{J} n_j \log\left(1-\frac{z_j}{K}\right)} \\ &\approx \sum_{n_1,\ldots,n_J} \bar{p}_{i_1,\ldots,i_J}\left(\frac{n_1}{K},\ldots,\frac{n_J}{K},t\right) e^{-\sum_{j=1}^{J} z_j \frac{n_j}{K}} \frac{1}{K^J} \\ &\approx \int_{\mathbb{R}_+^J} \bar{p}_{i_1,\ldots,i_J}(y_1,\ldots,y_J,t) e^{-\sum_{j=1}^{J} z_j y_j} \mathrm{d}y_1\ldots\mathrm{d}y_J \\ &= \int_{\mathbb{R}_+^J} p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t) e^{-\sum_{j=1}^{J} z_j \frac{x_j}{K}} \mathrm{d}x_1\ldots\mathrm{d}x_J \\ &= \hat{p}_{i_1,\ldots,i_J}\left(\frac{z_1}{K},\ldots,\frac{z_J}{K},t\right), \end{aligned} \tag{49}$$

where again we have used the fact that a Riemann sum converges to a Riemann integral as the partition size tends to zero. This clearly shows that when $K \gg 1$, the generating function of the discrete distribution $p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t)$ and the Laplace transform of the continuous distribution $p_{i_1,\ldots,i_J}(x_1,\ldots,x_J,t)$ are approximately equal. Hence the solutions of the discrete and continuous models are approximately related by the multivariate Poisson representation in the limit of large protein numbers, i.e.

$$p_{i_1,\ldots,i_J,n_1,\ldots,n_J}(t) \approx \int_{\mathbb{R}_+^J} \frac{e^{-\sum_{j=1}^{J} x_j} \prod_{j=1}^{J} x_j^{n_j}}{\prod_{j=1}^{J} n_j!} p_{i_1,,\ldots,i_J}(x_1,\ldots,x_J,t)\mathrm{d}x_1\ldots\mathrm{d}x_J.$$

5.3 Numerical validation

We then apply our results to a specific gene regulatory network. Here we consider a toggle switch (see Fig. 5(a) for illustration), where two genes repress the expression of each other [83]. The toggle switch can be described by the following reactions:

$$\begin{gathered} G_1 \xrightarrow{a_1} G_1^*, \quad G_1^* \xrightarrow{b_1 n_2} G_1, \quad G_2 \xrightarrow{a_2} G_2^*, \quad G_2^* \xrightarrow{b_2 n_1} G_2, \\ G_1^* \xrightarrow{u_1} G_1^* + P_1, \quad G_2^* \xrightarrow{u_2} G_2^* + P_2, \\ P_1 \xrightarrow{d_1} \varnothing, \quad P_2 \xrightarrow{d_2} \varnothing, \end{gathered}$$

where protein P_2 (P_1) binds to gene G_1 (G_2) and inhibits its expression. The first four reactions describe state switching of the two genes, the next two reactions describe synthesis of the two proteins, and the last two reactions describe degradation of the two proteins.

To validate our results, we assume that the synthesis rates scale with the system size K as $u_j = \bar{u}_j K, j = 1, 2$ and the feedback strengths scale with K as $b_j = \bar{b}_j / K, j = 1, 2$, where we regard $\bar{u}_i$, d_i,a_i, and $\bar{b}_i$ as constants. For simplicity, we focus on the steady-state protein number distributions of the discrete and continuous models, where the continuous model can be described by the following switching system of ODEs:

$$
\begin{array}{ccccc}
(G_1 \text{ inactive}, G_2 \text{ inactive}), & \begin{cases} \dot{x}_1 = -d_1 x_1 \\ \dot{x}_2 = -d_2 x_2 \end{cases} & \underset{b_1 x_2}{\overset{a_1}{\rightleftharpoons}} & \begin{cases} \dot{x}_1 = u_1 - d_1 x_1 \\ \dot{x}_2 = -d_2 x_2 \end{cases} & (G_1 \text{ active}, G_2 \text{ inactive}), \\
 & b_2 x_1 \Big\uparrow\Big\downarrow a_2 & & b_2 x_1 \Big\uparrow\Big\downarrow a_2 & \\
(G_1 \text{ inactive}, G_2 \text{ active}), & \begin{cases} \dot{x}_1 = -d_1 x_1 \\ \dot{x}_2 = u_2 - d_2 x_2 \end{cases} & \underset{b_1 x_2}{\overset{a_1}{\rightleftharpoons}} & \begin{cases} \dot{x}_1 = u_1 - d_1 x_1 \\ \dot{x}_2 = u_2 - d_2 x_2 \end{cases} & (G_1 \text{ active}, G_2 \text{ active}).
\end{array}
$$

According to our results, the steady-state distributions, $p^{ss}_{i_1,i_2,n_1,n_2}$ and $p^{ss}_{i_1,i_2}(x_1, x_2)$, of the discrete and continuous models are approximately related by the two-dimensional Poisson representation when $K \gg 1$, i.e.

$$
p^{ss}_{i_1,i_2,n_1,n_2} \approx \int_{\mathbb{R}^2_+} \frac{e^{-x_1-x_2} x_1^{n_1} x_2^{n_2}}{n_1! n_2!} p^{ss}_{i_1,i_2}(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2. \tag{50}
$$

To test this, we compare the steady-state approximate solution of the discrete model generated from the Poisson representation with the numerical one obtained from the SSA. In the Poisson representation, the Poisson kernel is chosen to be the steady-state solution of the continuous model obtained from the SSA. For convenience of comparison, here we only focus on the marginal distribution of protein P_2 (see Fig. 5(b)). It is clear that the Poisson representation fails to connect the discrete and continuous models when protein numbers are small, but it indeed links the two models when protein numbers are large.

6 Conclusions and discussion

In this paper, we systematically explore when the Poisson representation can be used to connect the distributions of the discrete and continuous models for stochastic gene regulatory networks. We found that if the gene of interest is unregulated, then the mRNA or protein number distributions for the discrete and continuous models are always related by the Poisson representation, whenever (i) the initial distributions for the two models are related by the Poisson representation and (ii) the burst size distributions for the two models are also related by the Poisson representation. This holds for any unregulated system with complex promoter switching mechanisms, bursty production of the gene product, and even upstream cellular drives. If the distributions for the two models are related by the Poisson representation and the exact distribution for one model is known, then we can use this relation to recover the exact distribution for the other model. This technique was then used to obtain some new analytical results including (i) the steady-state solution for a multi-state bursty model, (ii) the time-dependent solution for a two-state bursty model, and (iii) the time-dependent solution for a coupled dynamic model of gene expression and cell size.

We next examined whether the Poisson representation also plays a crucial role for regulated systems. For an autoregulated gene, we showed by a simple example that the protein number distributions for the discrete and continuous models are not exactly related by the Poisson representation. However, by introducing a macroscopic scaling parameter, we found that the discrete and continuous models are approximately related by the Poisson representation in the limit

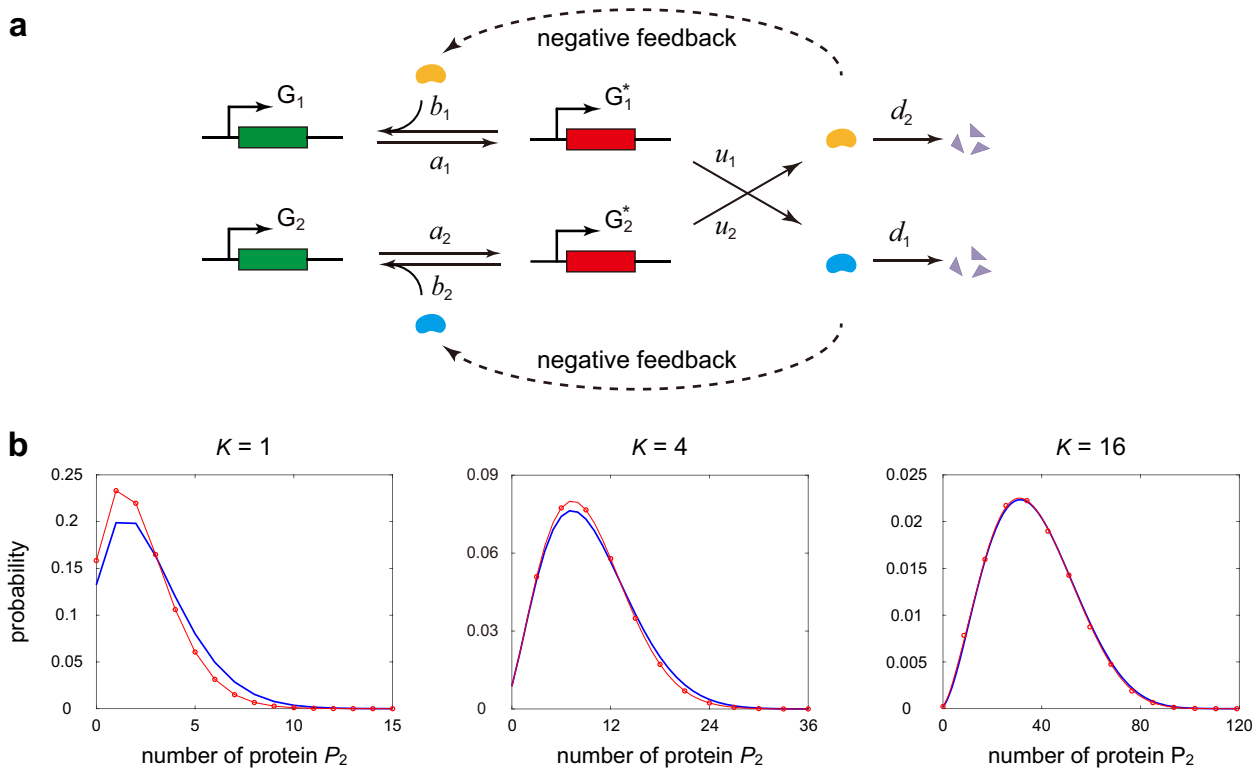


Figure 5: **Poisson representation for a toggle switch.** **(a)** Schematics of a toggle switch. The protein associated with each gene binds to and inhibits the expression of the other gene. **(b)** Comparison of the steady-state distribution of the discrete model computed using the SSA (blue curves) and the approximate one given by Eq. (50) (red circles and red curves). When computing the approximate distribution, we first calculate the steady-state distribution of the continuous model using the SSA and then construct the approximate distribution using the Poisson representation. The parameters are chosen as $a_1 = 2, b_1 = 0.01, a_2 = 3, b_2 = 2.5, \bar{u}_1 = 2, \bar{u}_2 = 6, d_1 = 1, d_2 = 1$.

of large protein numbers. As an application, we derived the approximate distribution for the discrete model of an autoregulatory gene circuit with complex feedback regulation. More interestingly, we generalized these results to a complex regulatory networks with multiple genes. Similarly, we revealed that for an arbitrary gene regulatory network, the joint protein number distributions for the discrete and continuous models are approximately related by the Poisson representation in the limit of large protein numbers.

In summary, the present work shows that the Poisson representation serves as a powerful tool to investigate the dynamic properties of stochastic gene regulatory networks, but it should be used with caution when the molecule numbers of some proteins are very small. We anticipate that the results of the present paper can be extended to non-Markovian [84] and delayed [85] models of gene regulatory networks. It would be also interesting to study whether the concept of Poisson representation and the results of the present paper can be generalized to spatially resolved gene expression models [86].

Data Availability Statement

The data that supports the findings of this study are available within the article.

Acknowledgements

The authors thank Dr. Bingjie Wu for valuable suggestions. C. J. acknowledges support from National Natural Science Foundation of China with grant No. U1930402, No. 12271020, and No. 12131005.

Appendix A The proof of Theorem 1

For the gene expression model described by Eq. (14), the probability distribution $p_{i,n}(t)$ of the discrete model satisfies the set of CMEs

$$\begin{aligned}\dot{p}_{i,n}(t) = u_i(t)\left(\sum_{k=0}^{n}\theta_{n-k}p_{i,k}(t) - p_{i,n}(t)\right) + d(t)\left[(n+1)p_{i,n+1}(t) - np_{i,n}(t)\right] \\ + \left[\sum_{j\neq i}k_{ji}(t)p_{j,n}(t) - \sum_{j\neq i}k_{ij}(t)p_{i,n}(t)\right], \quad i = 0,1,\ldots,L.\end{aligned} \tag{51}$$

To proceed, let $f_i(z,t) = \sum_{n=0}^{\infty} p_{i,n}(t)z^n$ denote the generating function of $p_{i,n}(t)$ and let $h(z) = \sum_{n=0}^{\infty}\theta_n z^n$ denote the generating function of the burst size distribution θ_n of the discrete model. Note that

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{k=0}^{n}\theta_{n-k}p_{i,k}(t)z^n &= \sum_{n=0}^{\infty}\sum_{k=0}^{n}\theta_{n-k}z^{n-k}p_{i,k}(t)z^k = \sum_{k=0}^{\infty}\sum_{n=k}^{\infty}\theta_{n-k}z^{n-k}p_{i,k}(t)z^k \\ &= \sum_{k=0}^{\infty}\left(\sum_{n=0}^{\infty}\theta_n z^n\right)p_{i,k}(t)z^k = h(z)f_i(z,t).\end{aligned}$$

Multiplying z^n on both sides of Eq. (51) and then summing over n, we obtain

$$\partial_t f_i(z,t) = u_i(t)\left[h(z)-1\right]f_i(z,t) + d(t)(1-z)\partial_z f_i(z,t) + \left[\sum_{j\neq i}k_{ji}(t)f_j(z,t) - \sum_{j\neq i}k_{ij}(t)f_i(z,t)\right], \tag{52}$$

subject to the initial condition

$$f_i(z,0) = \sum_{n=0}^{\infty}p_{i,n}(0)z^n, \quad i = 0,1,\ldots,L.$$

From Eq. (2), the generating function $f_i(z,t)$ and the Poisson kernel $\rho_i(x,t)$ of the discrete distribution $p_{i,n}(t)$ are related by $f_i(1-z,t) = \hat{\rho}_i(z,t)$, where $\hat{\rho}_i(z,t) = \int_0^{\infty}\rho_i(x,t)e^{-zx}\mathrm{d}x$ denotes the Laplace transform of $\rho_i(x,t)$. It then follows from Eq. (52) that the evolution of $\hat{\rho}_i(z,t)$ is given by

$$\partial_t\hat{\rho}_i(z,t) = u_i(t)[h(1-z)-1]\hat{\rho}_i(z,t) - d(t)z\partial_z\hat{\rho}_i(z,t) + \left[\sum_{j\neq i}k_{ji}(t)\hat{\rho}_j(z,t) - \sum_{j\neq i}k_{ij}(t)\hat{\rho}_i(z,t)\right], \tag{53}$$

subject to the initial conditions

$$\hat{\rho}_i(z,0) = f_i(1-z,0) = \sum_{n=0}^{\infty}p_{i,n}(0)(1-z)^n, \quad i = 0,1,\ldots,L.$$

On the other hand, the probability distribution $p_i(x,t)$ of the continuous model satisfies the Kolmogorov forward equation

$$\begin{aligned}\partial_t p_i(x,t) = u_i(t)\left[\int_0^x w(x-x')p_i(x',t)\mathrm{d}x' - p_i(x,t)\right] + d(t)\partial_x[xp_i(x,t)] \\ + \left[\sum_{j\neq i}k_{ji}(t)p_j(x,t) - \sum_{j\neq i}k_{ij}(t)p_i(x,t)\right], \quad i = 0,1,\ldots,L,\end{aligned} \tag{54}$$

where $w(x)$ is the burst size distribution of the continuous model. Recall that the burst size distributions of the discrete and continuous models are related by $h(1-z) = \hat{w}(z)$, where $\hat{w}(z) = \int_0^\infty w(x)e^{-zx}\mathrm{d}x$ denotes the Laplace transform of $w(x)$. Taking the Laplace transform with respect to x on both sides of Eq. (54), it is easy to check that the evolution of $\hat{p}_i(z,t) = \int_0^\infty p_i(x,t)e^{-zx}\mathrm{d}x$ is governed by

$$\partial_t \hat{p}_i(z,t) = u_i(t)[\hat{w}(z)-1]\hat{p}_i(z,t) - d(t)z\partial_z(\hat{p}_i(z,t)) + \left[\sum_{j\neq i} k_{ji}(t)\hat{p}_j(z,t) - \sum_{j\neq i} k_{ij}(t)\hat{p}_i(z,t)\right], \tag{55}$$

subject to the initial conditions

$$\hat{p}_i(z,0) = \int_0^\infty p_i(x,0)e^{-zx}\mathrm{d}x, \quad i = 0,1,\ldots,L.$$

Here we assume that for each gene state G_i, the initial distributions of the discrete and continuous models are related by the Poisson representation, i.e.

$$p_{i,n}(0) = \int_0^\infty \frac{e^{-x}x^n}{n!} p_i(x,0)\mathrm{d}x, \quad n = 0,1,2,\ldots,$$

which means that $\hat{\rho}_i(z,0) = \hat{p}_i(z,0)$. Comparing Eqs. (53) and (55), it is easy to see that $\hat{\rho}_i(z,t)$ and $\hat{p}_i(z,t)$ satisfy the same initial value problem. For the initial value problem of a linear system of first-order partial differential equations, the well-known Cauchy-Kovalevskaya theorem [87] states that if the initial conditions of the system of partial differential equations are restricted on a hypersurface and if the coefficients and initial conditions are sufficiently smooth, then there exists a unique solution in the hypersurface. This guarantees the existence and uniqueness of the solution of Eq. (55), which shows that

$$\hat{\rho}_i(z,t) = \hat{p}_i(z,t), \quad \forall z \geq 0,\ t \geq 0,\ i = 0,1,\ldots,L.$$

Finally, taking the inverse Laplace transform yields

$$\rho_i(x,t) = p_i(x,t), \quad \forall x \geq 0,\ t \geq 0,\ i = 0,1,\ldots,L.$$

Appendix B Derivation of Eq. (39)

Recall that for autoregulatory gene circuit whose reactions are given by Eq. (34), the steady-state distribution for the continuous model is given by

$$p^{ss}(x) = p_0^{ss}(x) + p_1^{ss}(x) = \left(\frac{\tilde{C}}{d(x)x} + \frac{\tilde{C}}{u(x)-d(x)x}\right)\exp\left\{\int_{x_0}^x \left[\frac{a(\tau)}{d(\tau)\tau} - \frac{b(\tau)}{u(\tau)-d(\tau)\tau}\right]\mathrm{d}\tau\right\}.$$

Next we compute the integral in the above equation when

$$a(x) = a_0 + a_1x^2, \quad b(x) = b, \quad u(x) = u, \quad d(x) = \frac{d_{max}x}{k_m + x}.$$

First, it is easy to check that

$$\begin{aligned}\int_{x_0}^x \frac{a(\tau)}{d(\tau)\tau}\mathrm{d}\tau &= \int_{x_0}^x \frac{a_0 + a_1\tau^2}{\frac{d_{max}\tau^2}{k_m+\tau}}\mathrm{d}\tau = \int_{x_0}^x \left[\frac{(k_m+\tau)a_0}{d_{max}\tau^2} + \frac{(k_m+\tau)a_1}{d_{max}}\right]\mathrm{d}\tau \\ &= \frac{k_m a_0}{d_{max}}\left(\frac{1}{x_0} - \frac{1}{x}\right) + \frac{a_0}{d_{max}}\log\frac{x}{x_0} + \frac{a_1}{d_{max}}\left[k_m(x-x_0) + \frac{1}{2}(x-x_0)^2\right].\end{aligned} \tag{56}$$

To proceed, note that

$$\frac{1}{u-\frac{d_{max}\tau^2}{k_m+\tau}}=\frac{k_m+\tau}{uk_m+u\tau-d_{max}\tau^2}=-\frac{1}{d_{max}}\frac{\tau+k_m}{\tau^2-\frac{u}{d_{max}}\tau-\frac{uk_m}{d_{max}}}=-\frac{1}{d_{max}}\left[\frac{A}{\tau-\alpha_1}+\frac{B}{\tau-\alpha_2}\right],$$

where α_1, α_2, A, and B are four constants given by

$$\alpha_1+\alpha_2=\frac{u}{d_{max}},\quad \alpha_1\alpha_2=-\frac{uk_m}{d_{max}},\quad A=\frac{-k_m-\alpha_1}{\alpha_2-\alpha_1},\quad B=\frac{k_m+\alpha_2}{\alpha_2-\alpha_1}.$$

Then we have

$$\begin{aligned}\int_{x_0}^{x}\frac{b(\tau)}{u(\tau)-d(\tau)\tau}\mathrm{d}\tau&=\int_{x_0}^{x}\frac{b}{u-\frac{d_{max}\tau^2}{k_m+\tau}}\mathrm{d}\tau=b\int_{x_0}^{x}\frac{\tau+d_1}{uk_m+u\tau-d_{max}\tau^2}\mathrm{d}\tau\\&=-\frac{b}{d_{max}}\left[\log\left(\frac{x-\alpha_1}{x_0-\alpha_1}\right)^A+\log\left(\frac{x-\alpha_2}{x_0-\alpha_2}\right)^B\right].\end{aligned}\tag{57}$$

Inserting Eqs. (56) and (57) into Eq. (37) in the main text, we finally obtain Eq. (39).

References

[1] Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2**, 157–175 (2005).

[2] Schnoerr, D., Sanguinetti, G. & Grima, R. Approximation and inference methods for stochastic biochemical kinetics — a tutorial review. *J. Phys. A: Math. Theor.* **50**, 093001 (2017).

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

[4] Cai, L., Friedman, N. & Xie, X. S. Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).

[5] Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *science* **329**, 533–538 (2010).

[6] Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proc. Natl. Acad. Sci. USA* **105**, 17256–17261 (2008).

[7] Rosenfeld, N., Elowitz, M. B. & Alon, U. Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.* **323**, 785–793 (2002).

[8] Friedman, N., Cai, L. & Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 168302 (2006).

[9] Karmakar, R. & Bose, I. Graded and binary responses in stochastic gene expression. *Phys. Biol.* **1**, 197 (2004).

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

[11] Gardiner, C. W. & Chaturvedi, S. The poisson representation. I. A new technique for chemical master equations. *J. Stat. Phys.* **17**, 429–468 (1977).

[12] Chaturvedi, S. & Gardiner, C. The poisson representation. II Two-time correlation functions. *J. Stat. Phys.* **18**, 501–522 (1978).

[13] Kast, S. CW Gardiner: Handbook of Stochastic Methods, 2nd edn., Springer, Berlin, 1997, 29 Abb., 442 Seiten, ISBN 3-540-61634-9, Softcover 68, 00 DM (1998).

[14] Dattani, J. & Barahona, M. Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization. *J. R. Soc. Interface* **14**, 20160833 (2017).

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

[16] Gorin, G. & Pachter, L. Modeling bursty transcription and splicing with the chemical master equation. *Biophys. J.* **121**, 1056–1069 (2022).

[17] Jia, C., Zhang, M. Q. & Hong, Q. Emergent Lévy behavior in single-cell stochastic gene expression. *Phys. Rev. E* **96**, 040402(R) (2017).

[18] Jia, C., Yin, G. G., Zhang, M. Q. *et al.* Single-cell stochastic gene expression kinetics with coupled positive-plus-negative feedback. *Phys. Rev. E* **100**, 052406 (2019).

[19] Chen, X. & Jia, C. Limit theorems for generalized density-dependent Markov chains and bursty stochastic gene regulatory networks. *J. Math. Biol.* **80**, 959–994 (2020).

[20] Paulsson, J. & Ehrenberg, M. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys. Rev. Lett.* **84**, 5447 (2000).

[21] Ko, M. S. A stochastic model for gene induction. *J. Theor. Biol.* **153**, 181–194 (1991).

[22] Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).

[23] Iyer-Biswas, S., Hayot, F. & Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).

[24] Vu, T. N. *et al.* Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**, 2128–2135 (2016).

[25] Iyer-Biswas, S. & Jayaprakash, C. Mixed Poisson distributions in exact solutions of stochastic autoregulation models. *Phys. Rev. E* **90**, 052712 (2014).

[26] Zhang, J. & Zhou, T. Promoter-mediated transcriptional dynamics. *Biophys. J.* **106**, 479–488 (2014).

[27] Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *science* **332**, 472–474 (2011).

[28] Harper, C. V. *et al.* Dynamic analysis of stochastic transcription cycles. *PLoS Biol.* **9**, e1000607 (2011).

[29] Neuert, G. *et al.* Systematic identification of signal-activated stochastic gene regulation. *Science* **339**, 584–587 (2013).

[30] Senecal, A. *et al.* Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep.* **8**, 75–83 (2014).

[31] Corrigan, A. M., Tunnacliffe, E., Cannon, D. & Chubb, J. R. A continuum model of transcriptional bursting. *Elife* **5**, e13051 (2016).

[32] Sepúlveda, L. A., Xu, H., Zhang, J., Wang, M. & Golding, I. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science* **351**, 1218–1222 (2016).

[33] Featherstone, K. *et al.* Spatially coordinated dynamic gene transcription in living pituitary tissue. *Elife* **5**, e08494 (2016).

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

[35] Bokes, P., King, J. R., Wood, A. T. & Loose, M. Multiscale stochastic modelling of gene expression. *J. Math. Biol.* **65**, 493–520 (2012).

[36] Jia, C. Simplification of Markov chains with infinite state space and the mathematical theory of random gene expression bursts. *Phys. Rev. E* **96**, 032402 (2017).

[37] Elgart, V., Jia, T., Fenley, A. T. & Kulkarni, R. Connecting protein and mRNA burst distributions for stochastic models of gene expression. *Phys. Biol.* **8**, 046001 (2011).

[38] Schwabe, A., Rybakova, K. N. & Bruggeman, F. J. Transcription stochasticity of complex gene regulation models. *Biophys. J.* **103**, 1152–1161 (2012).

[39] Kuwahara, H., Arold, S. T. & Gao, X. Beyond initiation-limited translational bursting: the effects of burst size distributions on the stability of gene expression. *Integr. Biol.* **7**, 1622–1632 (2015).

[40] Jędrak, J. & Ochab-Marcinek, A. Time-dependent solutions for a stochastic model of gene expression with molecule production in the form of a compound poisson process. *Phys. Rev. E* **94**, 032401 (2016).

[41] Mackey, M. C., Tyran-Kaminska, M. & Yvinec, R. Dynamic behavior of stochastic gene expression models in the presence of bursting. *SIAM J. Appl. Math.* **73**, 1830–1852 (2013).

[42] Holehouse, J., Cao, Z. & Grima, R. Stochastic modeling of autoregulatory genetic feedback loops: A review and comparative study. *Biophys. J.* **118**, 1517–1525 (2020).

[43] Yin, G. G. & Zhang, Q. *Continuous-time Markov chains and applications: a singular perturbation approach*, vol. 37 (Springer, 2012).

[44] Davis, M. H. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the*

Royal Statistical Society: Series B (Methodological) **46**, 353–376 (1984).

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

[46] Kurasov, P., Lück, A., Mugnolo, D. & Wolf, V. Stochastic hybrid models of gene regulatory networks–a pde approach. *Math. Biosci.* **305**, 170–177 (2018).

[47] Kurasov, P., Mugnolo, D. & Wolf, V. Analytic solutions for stochastic hybrid models of gene regulatory networks. *J. Math. Biol.* **82**, 1–29 (2021).

[48] Makris, N. The fractional derivative of the Dirac delta function and additional results on the inverse laplace transform of irrational functions. *Fractal and Fractional* **5**, 18 (2021).

[49] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

[50] Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* **10** (2011).

[51] Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).

[52] Anjum, A. *et al.* Identification of differentially expressed genes in rna-seq data of arabidopsis thaliana: A compound distribution approach. *J. Comput. Biol.* **23**, 239–247 (2016).

[53] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 1–15 (2019).

[54] Ahlmann-Eltze, C. & Huber, W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701–5702 (2020).

[55] Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, 1–12 (2013).

[56] Mou, T., Deng, W., Gu, F., Pawitan, Y. & Vu, T. N. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Front. Genet.* **10**, 1331 (2020).

[57] Chen, L., Zhu, C. & Jiao, F. A generalized moment-based method for estimating parameters of stochastic gene transcription. *Math. Biosci.* **345**, 108780 (2022).

[58] Chen, M. *et al.* Exact distributions for stochastic gene expression models with arbitrary promoter architecture and translational bursting. *Phys. Rev. E* **105**, 014405 (2022).

[59] Jia, C. & Li, Y. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. *bioRxiv* https://doi.org/10.1101/2022.01.05.475050 (2022).

[60] So, L.-h. *et al.* General properties of transcriptional time series in Escherichia coli. *Nat. Genet.* **43**, 554–560 (2011).

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

[62] Prudnikov, A., Brychkov, Y. A. & Marichev, O. *Integrals and series, Vol. 5: Inverse laplace transforms* (Gordon and Breach Science Publishers, 1992).

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

[64] Cao, Z. & Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 1–15 (2018).

[65] Cao, Z. & Grima, R. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci. USA* **117**, 4682–4692 (2020).

[66] Thomas, P. & Shahrezaei, V. Coordination of gene expression noise with cell size: analytical results for agent-based models of growing cell populations. *J. R. Soc. Interface* **18**, 20210274 (2021).

[67] Jia, C., Singh, A. & Grima, R. Concentration fluctuations in growing and dividing cells: Insights into the emergence of concentration homeostasis. *PLoS Comput. Biol.* **18**, e1010574 (2022).

[68] Jia, C. & Grima, R. Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model. *Iscience* **26** (2023).

[69] Lin, J. & Amir, A. Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.* **9**, 1–11 (2018).

[70] Berry, S. & Pelkmans, L. Mechanisms of cellular mRNA transcript homeostasis. *Trends Cell Biol.* (2022).

[71] Jia, C., Singh, A. & Grima, R. Cell size distribution of lineage data: analytic results and parameter inference. *iScience* **24**, 102220 (2021).

[72] Jia, C., Singh, A. & Grima, R. Characterizing non-exponential growth and bimodal cell size distributions in fission yeast: An analytical approach. *PLoS Comput. Biol.* **18**, e1009793 (2022).

[73] Jiao, F., Tang, M. & Yu, J. Distribution profiles and their dynamic transition in stochastic gene transcription. *J. Differ. Equations* **254**, 3307–3328 (2013).

[74] Wang, Z., Zhang, Z. & Zhou, T. Exact distributions for stochastic models of gene expression with arbitrary regulation. *Science China Mathematics* **63**, 485–500 (2020).

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

[76] Hornos, J. E. *et al.* Self-regulating gene: an exact solution. *Phys Rev E* **72**, 051907 (2005).

[77] Huang, L., Yuan, Z., Liu, P. & Zhou, T. Feedback-induced counterintuitive correlations of gene expression noise with bursting kinetics. *Physical Review E* **90**, 052702 (2014).

[78] Assaf, M., Roberts, E. & Luthey-Schulten, Z. Determining the stability of genetic switches: explicitly accounting for mRNA noise. *Phys. Rev. Lett.* **106**, 248102 (2011).

[79] Lv, C., Li, X., Li, F. & Li, T. Constructing the energy landscape for genetic switching system driven by intrinsic noise. *PLoS one* **9**, e88167 (2014).

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

[81] Bokes, P., King, J. R., Wood, A. T. & Loose, M. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *J. Math. Biol.* **64**, 829–854 (2012).

[82] Jia, C. & Grima, R. Small protein number effects in stochastic models of autoregulated bursty gene expression. *J. Chem. Phys.* **152**, 084115 (2020).

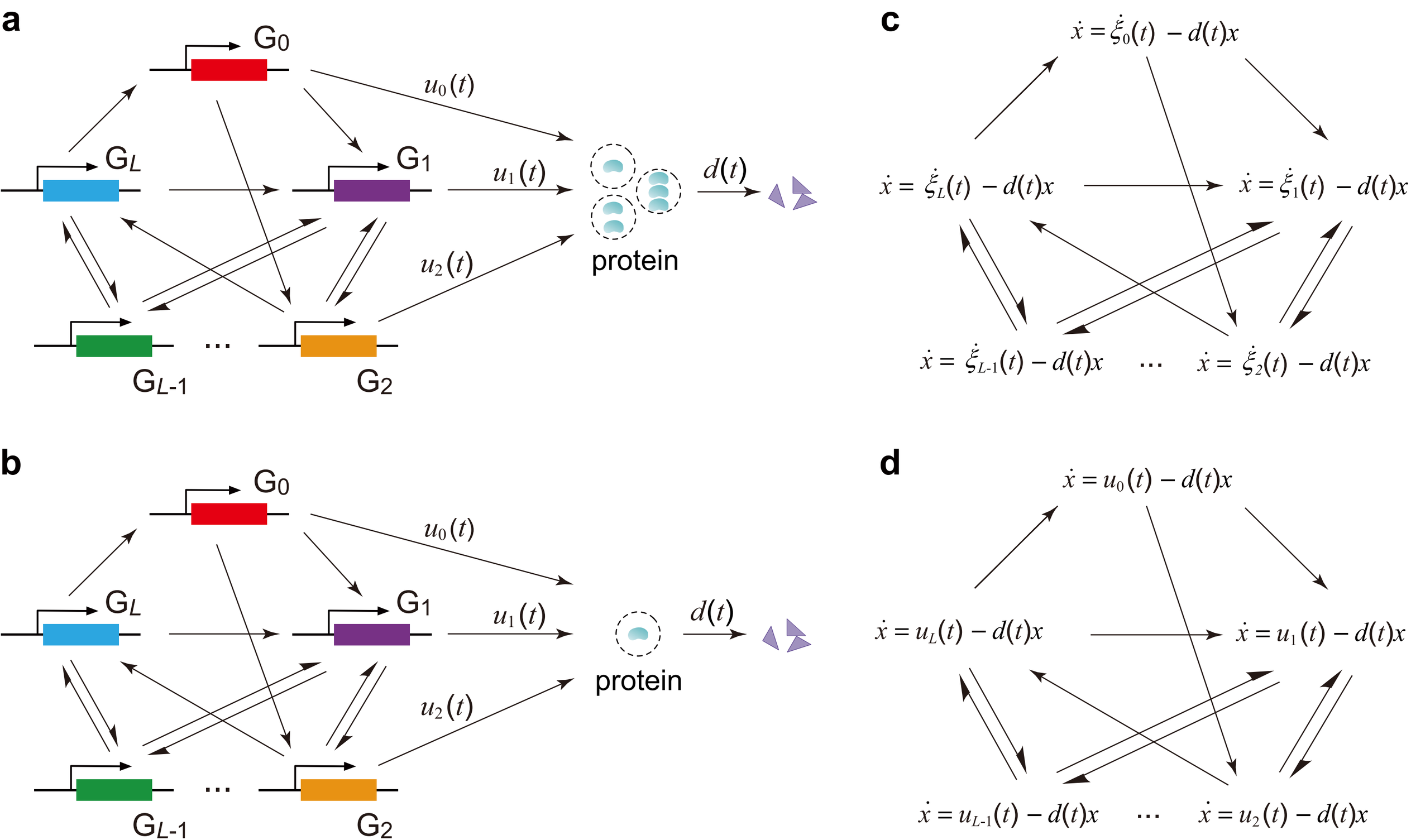
[83] Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).

[84] Zhang, Z., Deng, Q., Wang, Z., Chen, Y. & Zhou, T. Exact results for queuing models of stochastic transcription with memory and crosstalk. *Physical Review E* **103**, 062414 (2021).

[85] Bratsun, D., Volfson, D., Tsimring, L. S. & Hasty, J. Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. USA* **102**, 14593–14598 (2005).

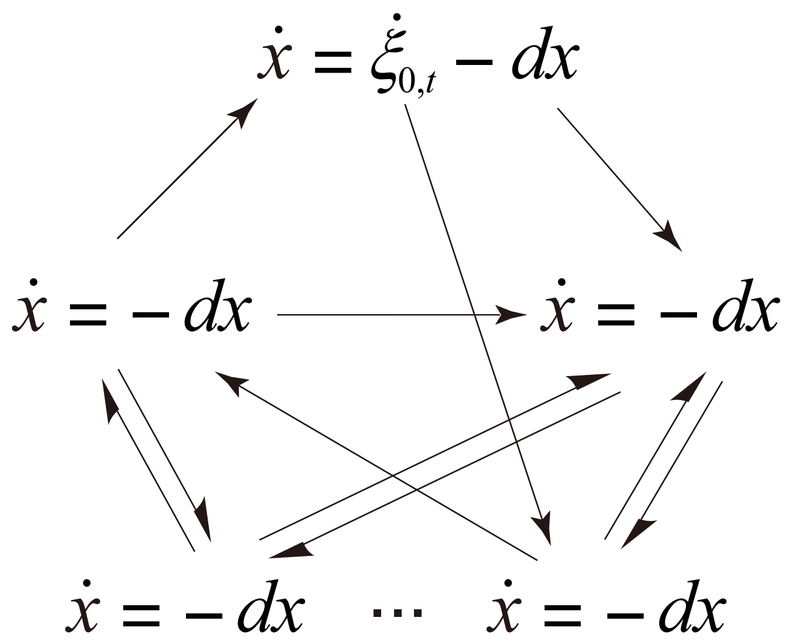
[86] Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

[87] Evans, L. C. *Partial differential equations*, vol. 19 (American Mathematical Soc., 2010).

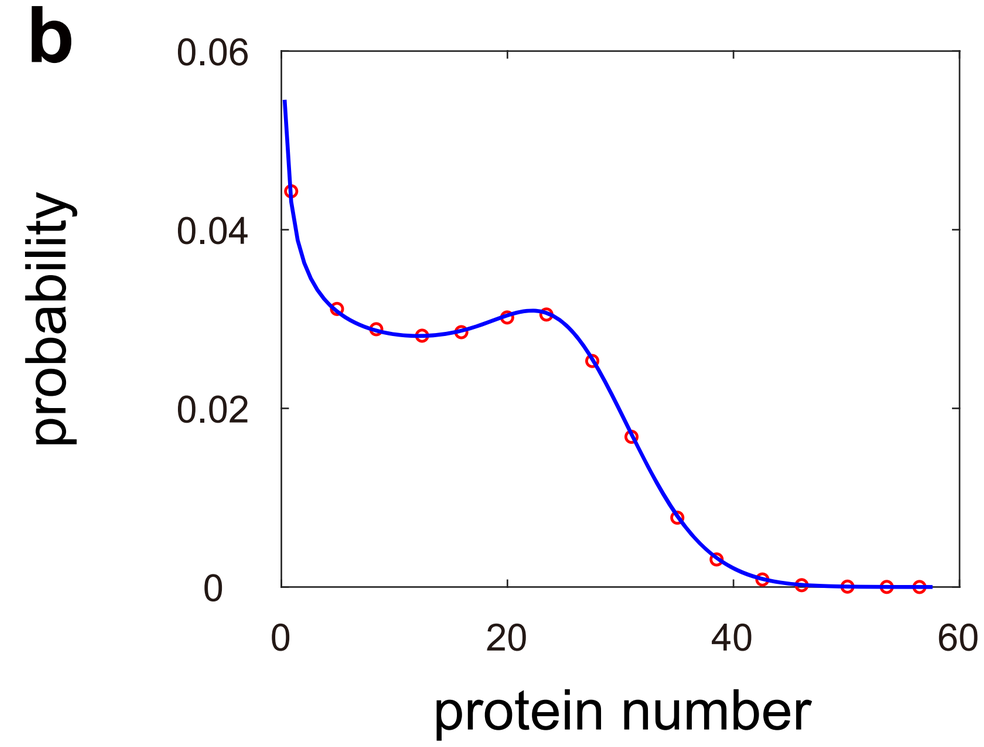

a
G0
GL
G1
GL-1
G2
u0(t)
u1(t)
u2(t)
d(t)
protein
b
G0
GL
G1
GL-1
G2
u0(t)
u1(t)
u2(t)
d(t)
protein
c
ẋ = ξ̇0(t) − d(t)x
ẋ = ξ̇L(t) − d(t)x
ẋ = ξ̇1(t) − d(t)x
ẋ = ξ̇L-1(t) − d(t)x
ẋ = ξ̇2(t) − d(t)x
d
ẋ = u0(t) − d(t)x
ẋ = uL(t) − d(t)x
ẋ = u1(t) − d(t)x
ẋ = uL-1(t) − d(t)x
ẋ = u2(t) − d(t)x

a

multi-state bursty model

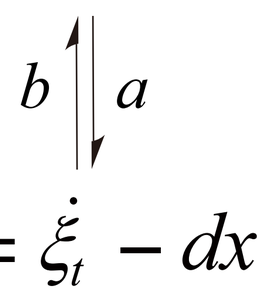


b

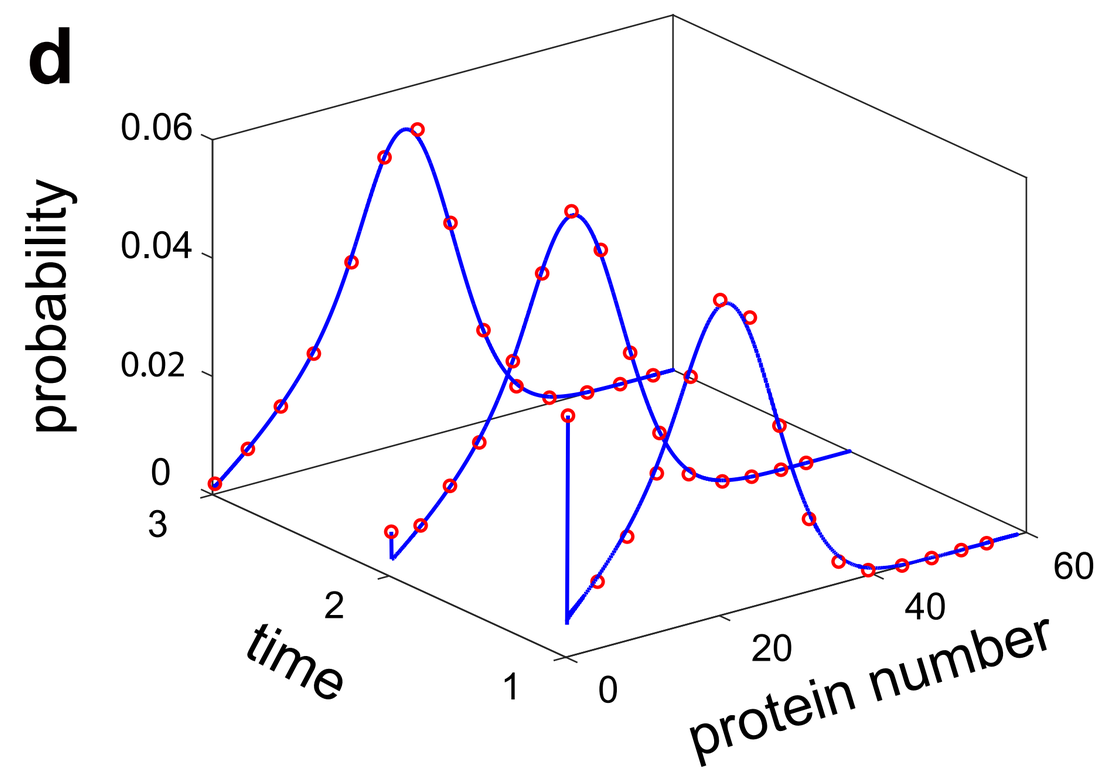


c

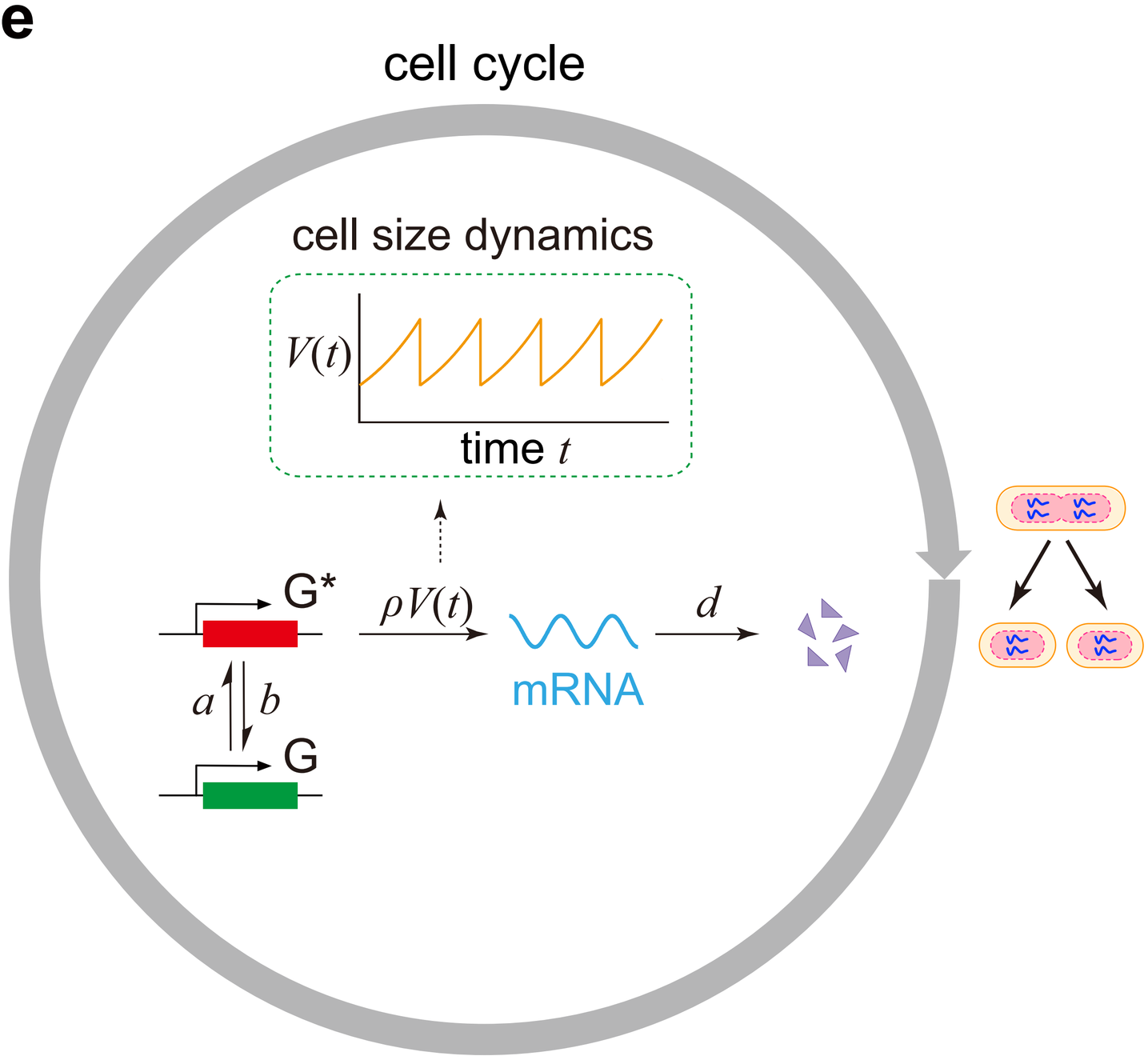
two-state bursty model



d

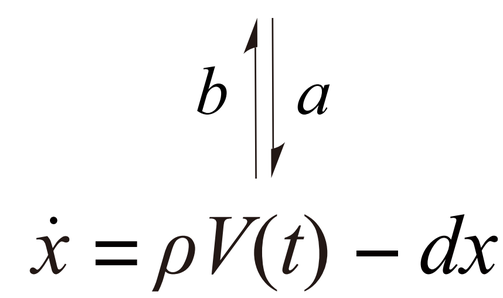


e

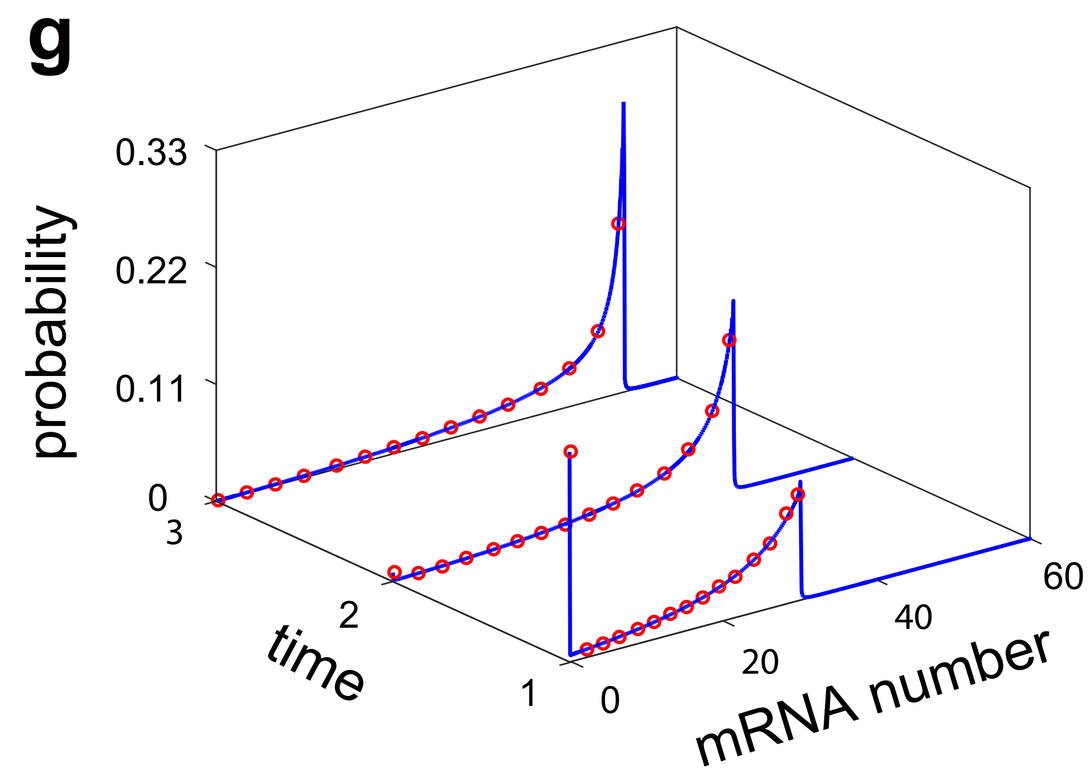


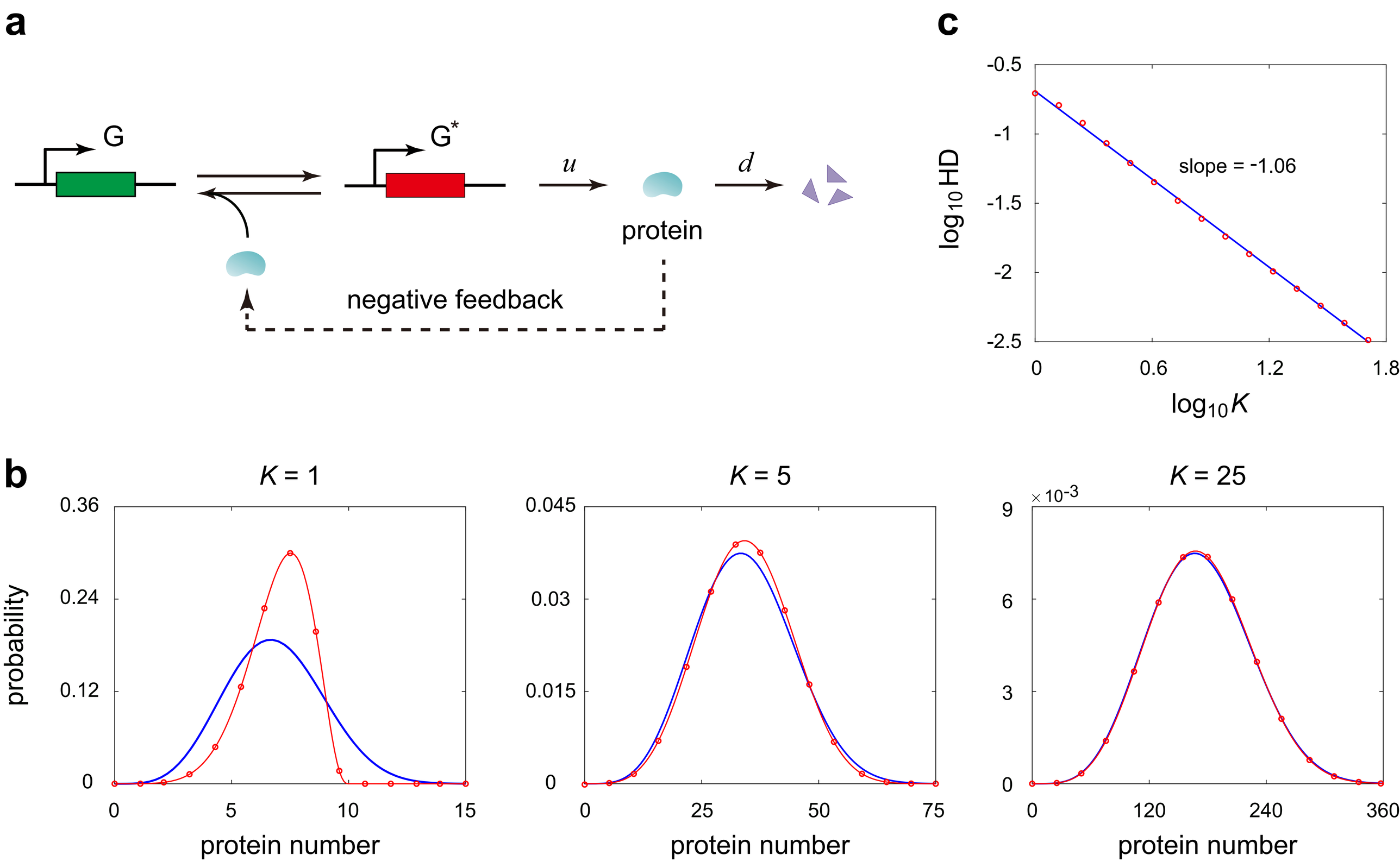
f

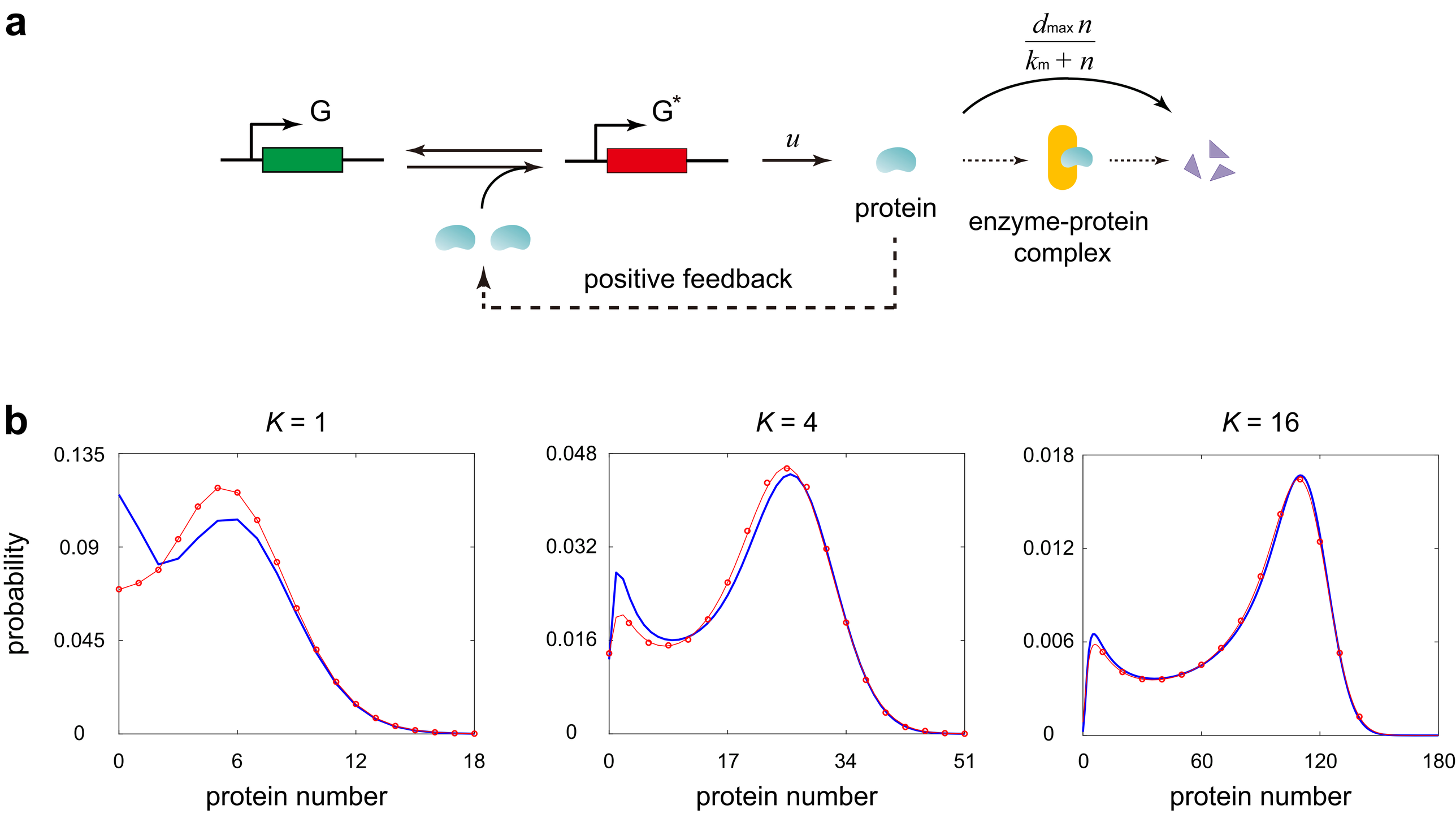
two-state coupled model

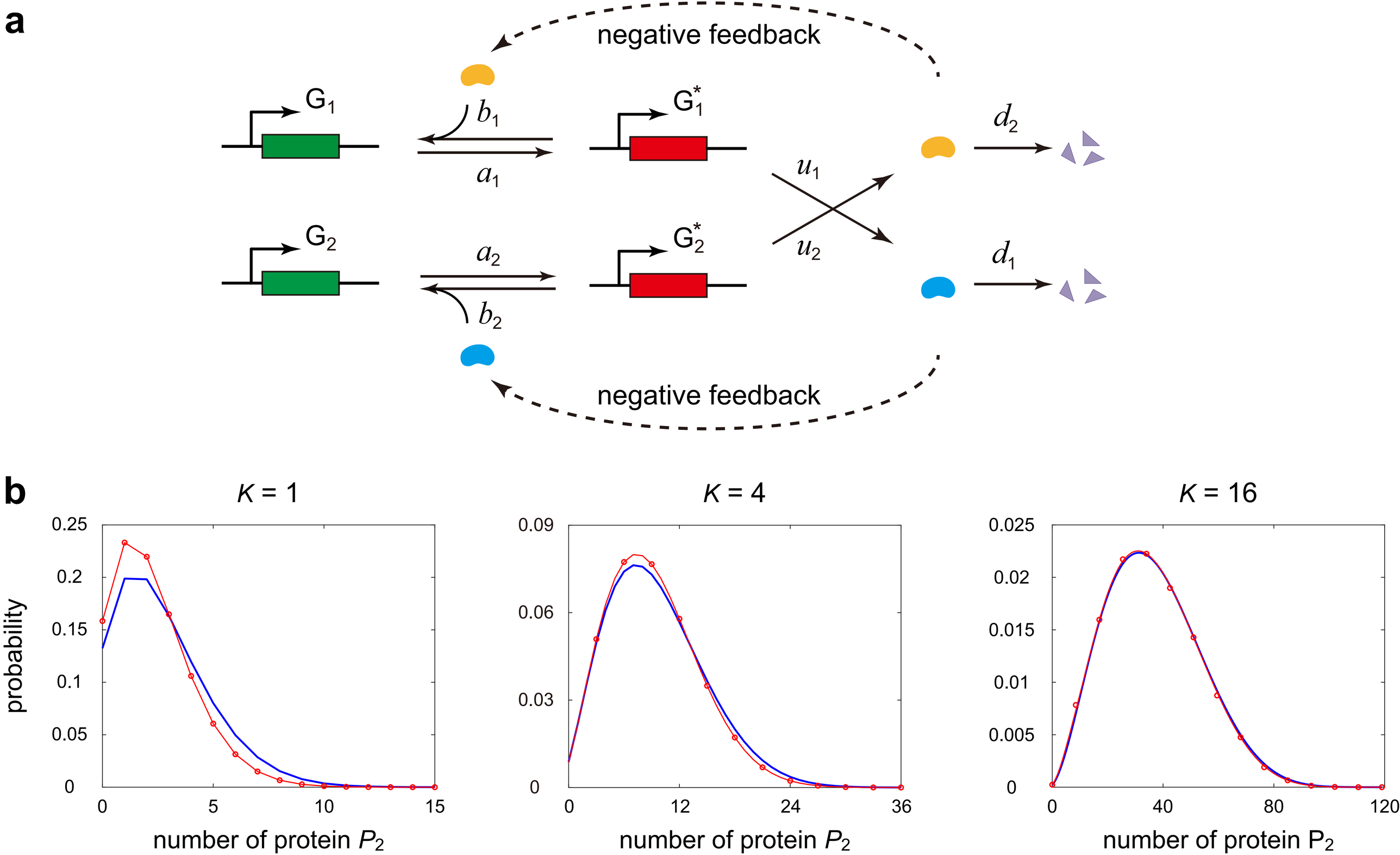


g




a
G
G*
u
d
protein
negative feedback
c
slope = -1.06
log10HD
log10K
-0.5
-1
-1.5
-2
-2.5
0
0.6
1.2
1.8
b
K = 1
K = 5
K = 25
probability
protein number
protein number
protein number
0.36
0.24
0.12
0
0.045
0.03
0.015
0
× 10-3
9
6
3
0
0
5
10
15
0
25
50
75
0
120
240
360


a
G
G*
u
protein
$\frac{d_{max}\, n}{k_m + n}$
enzyme-protein complex
positive feedback
b
K = 1
K = 4
K = 16
probability
0.135
0.09
0.045
0
0
6
12
18
protein number
0.048
0.032
0.016
0
0
17
34
51
protein number
0.018
0.012
0.006
0
0
60
120
180
protein number


a
negative feedback
G1
b1
a1
G1*
d2
u1
u2
G2
a2
b2
G2*
d1
negative feedback
b
K = 1
K = 4
K = 16
probability
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
number of protein P2
0.09
0.06
0.03
0
0
12
24
36
number of protein P2
0.025
0.02
0.015
0.01
0.005
0
0
40
80
120
number of protein P2