Title

Urn models for stochastic gene expression yield intuitive insights into the probability distributions of single-cell mRNA and protein counts

Authors

Choudhary, Krishna
Narang, Atul

Publication Date

2020-08-09

DOI

10.1088/1478-3975/aba50f

Data Availability

The data associated with this publication are in the supplemental files.

Peer reviewed

Physical Biology

OFFICIAL JOURNAL OF
Sociedad de Biofísicos Latino Americanos

ACCEPTED MANUSCRIPT

Urn models for stochastic gene expression yield intuitive insights into the probability distributions of single-cell mRNA and protein counts

To cite this article before publication: Krishna Choudhary *et al* 2020 *Phys. Biol.* in press https://doi.org/10.1088/1478-3975/aba50f

Urn models for stochastic gene expression yield intuitive insights into the probability distributions of single-cell mRNA and protein counts

Krishna Choudhary[1,*], Atul Narang[2,#]

1 Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA
2 Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology, Delhi, India

* krishna.choudhary@gladstone.ucsf.edu, kchoudhary@ucdavis.edu
# anarang@dbeb.iitd.ac.in

Abstract

Fitting the probability mass functions from analytical solutions of stochastic models of gene expression to the single-cell count distributions of mRNA and protein molecules can yield valuable insights into mechanisms underlying gene expression. Solutions of chemical master equations are available for various kinetic schemes but, even for the basic ON-OFF genetic switch, they take complex forms with generating functions given as hypergeometric functions. Interpretation of gene expression dynamics in terms of bursts is not consistent with the complete range of parameters for these functions. Physical insights into the probability mass functions are essential to ensure proper interpretations but are lacking for models considering genetic switches. To fill this gap, we develop urn models for stochastic gene expression. We sample RNA polymerases or ribosomes from a master urn, which represents the cytosol, and assign them to recipient urns of two or more colors, which represent time intervals in which no switching occurs. Colors of the recipient urns represent sub-systems of the promoter states, and the assignments to urns of a specific color represent gene expression. We use elementary principles of discrete probability theory to solve a range of kinetic models without feedback, including the Peccoud-Ycart model, the Shahrezaei-Swain model, and models with an arbitrary number of promoter states. In the last case, we obtain a novel result for the protein distribution. For activated genes, we show that transcriptional lapses, which are events of gene inactivation for short time intervals separated by long active intervals, quantify the transcriptional dynamics better than bursts. We show that the intuition gained from our urn models may also be useful in understanding existing solutions for models with feedback. We contrast our models with urn models for related distributions, discuss a generalization of the Delaporte distribution for single-cell data analysis, and highlight the limitations of our models.

Introduction

Gene expression occurs in multiple steps [1]. The biochemical mechanisms of its steps are of great interest [2–4]. In particular, a majority of studies have focused on switching of promoter states, transcription, and translation [5]. Genes might be expressed at a uniform rate or transition between two or more states with different rates of expression [6]. In the latter case, the transitions might be mediated by gene-specific mechanisms such as interactions of the promoters with specific transcription factors or gene-independent mechanisms such as DNA supercoiling [7–13]. When genes are in transcriptionally active states, mRNA molecules might be produced, which might be translated further into proteins. Experimental data for the distribution of mRNA/protein molecules in single cells could be harnessed for model selection out of a candidate set of mechanistic models [14–16]. In this direction, numerous stochastic models of gene expression have been developed to study a range of kinetic schemes [5,17–21]. Their analytical solutions for the probability distributions of molecular counts have been obtained in many cases [16,22–43], and comparisons with single-cell RNA-seq and single-molecule imaging data have facilitated inferences in mechanistic studies [6, 11, 23, 44–50].

An elementary model of constitutive gene expression uniformly allows transcription at all times [36, 37]. This is identical to the classical *birth-and-death* process, which results in the Poisson distribution for mRNA molecules at stationary state [51]. A physically intuitive method to derive the mRNA distribution utilizes an *urn model*, whereby the kinetic scheme of mRNA production (which, say, occurs with rate constant v_0) and degradation (say, with rate constant d_0) is mapped to an urn scheme. To this end, one considers an urn with balls of two colors —black and white. Let the proportion of black balls in the urn be π. As time progresses, we sample balls one-at-a-time from the urn, i.e., we perform Bernoulli trials [51]. The outcome of each trial, a black or white ball, corresponds to an outcome of transcription or no transcription in physical terms, respectively. Each trial consists of drawing a ball, recording its color, replacing the ball in the urn, and mixing the urn to prepare for the next trial. The probability of m black balls in, say, n_{trials} trials is $\binom{n_{\text{trials}}}{m}\pi^m\left(1-\pi\right)^{n_{\text{trials}}-m}$, which is a binomial distribution [51]. Let us say that we draw balls without taking any break and the time duration per trial, Δt is infinitesimal. The kinetic scheme is mapped to the urn scheme by defining $\pi = v_0\Delta t$, i.e., the proportion of black balls in the urn is the same as the probability of transcription in Δt time, which is $v_0\Delta t$. Finally, the probability of observing m copies of mRNA molecules is obtained as the probability of drawing m black balls in infinitely many trials during the mean lifetime of mRNAs, d_0^{-1}, i.e., requiring that $n_{\text{trials}} = {d_0^{-1}}/{\Delta t}$ is very large. Poisson distribution with the parameter ${v_0}/{d_0}$ is the special case of the thus obtained binomial distribution when $\Delta t \to 0$, i.e., when $n_{\text{trials}} \to \infty$ and $\pi \to 0$ such that $\pi n_{\text{trials}} = {v_0}/{d_0}$ [51]. The steps of transcription and translation are mechanistically similar and hence, the urn scheme for the Poisson process applies to both. The stationary state count of proteins in a cell is given by a sum of random variables denoting the number of translations per mRNA molecule that is produced in the time needed to reach stationarity. This results in the negative binomial distribution for the count if the noise in transcriptions can be ignored (i.e., $d_0 \gg d_1$) [16, 36] and the Neyman type A distribution if not (i.e., $d_1 \gg d_0$) [38].

The models for constitutive expression have been extended to include switching of promoter states (henceforth, called a genetic switch regardless of the switching mechanism, which may be mediated by transcription factors, or by other factors that may or may not be actively regulated). Peccoud and Ycart studied a gene whose promoter switches between active and inactive states [32]. Shahrezaei and Swain extended the Peccoud-Ycart model by accounting for translation and solved it assuming $d_0 \gg d_1$, where d_1 is the rate constant for protein degradation [16]. Numerous generalizations and extensions of these models exist and many have been solved analytically, e.g., the leaky two-state model where the promoter switches between two states with different levels of activity [22–24], multi-state models that consider a promoter with more than two states [27, 31, 34, 35, 52], models with auto-regulation [25, 26, 30, 33], etc. All of these models result in probability generating functions that are related to the Kemp families of distributions, which have been derived using various urn models of contagion and population heterogeneity, and as compound or mixture distributions [53–55]. Notably, in each of their applications, the urn model has a distinct design, which is systematically developed to capture the physical characteristics of the natural system under study. Their distinctive features provide an intuitive mapping to the mechanisms behind their respective systems. These urn models have proven fundamental to studies of their intended systems, have immense pedagogical value and have been called a "standard expression" in statistical language [56–58]. For the system of a genetic switch, while an approach of solving chemical master equations can provide analytical solutions for probability distributions, physical insight into the solutions can be greatly facilitated by the application of urn models. Yet, to the best of our knowledge, urn schemes with well-defined mapping to models considering genetic switches are still lacking.

In this article, we develop an urn model approach to address this gap. We demonstrate its utility by applying it to diverse kinetic schemes with genetic switches and deriving stationary state probability distributions of mRNA and protein counts. Central to our approach are two principles. First, while transcriptions and translations are affected by promoter state transitions, arrivals of RNA polymerases or ribosomes occur with fixed rate constants independent of the transitions. In other words, the promoter state determines whether these arrivals result in gene expression, but none of the promoter states exclude polymerases or ribosomes from arriving (Fig. 1a). Second, while there is heterogeneity in promoter activity over long time intervals, i.e., a promoter switches between active and inactive states, in short intervals, the activity is homogeneous. Hence, we map each kinetic model that we consider to an urn model with a master

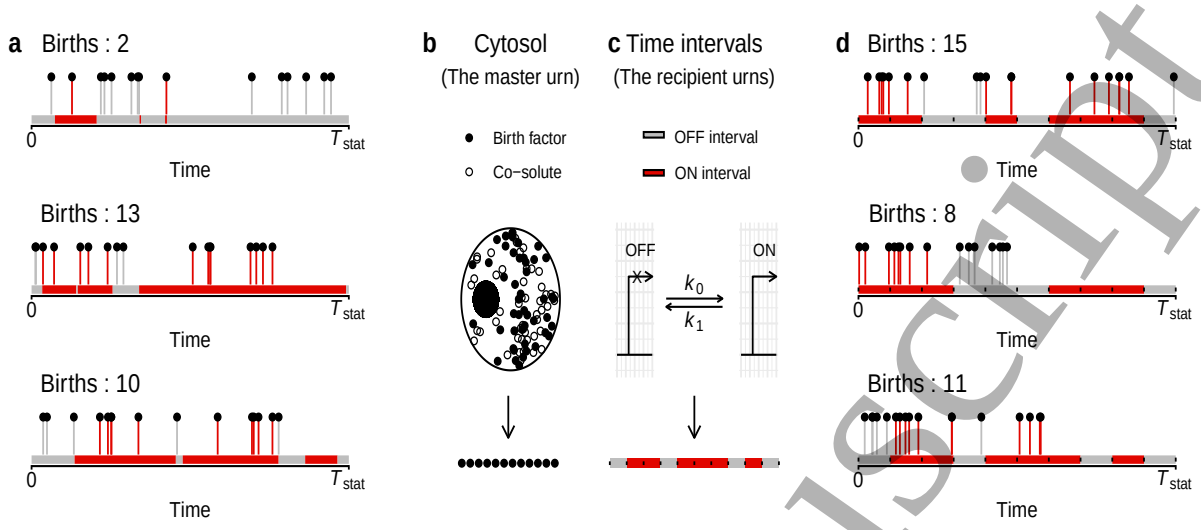

Figure 1. The urn scheme for a system with active and inactive states. **(a)** Three sample trajectories of a gene expression system. The black balls at the heads of spikes are birth factors, which represent RNA polymerases for models of mRNA counts and ribosomes for models of protein counts. The colored bar along the x-axis gives the state of the promoter at different time points, with the red color indicating the active state, grey color indicating the inactive state and T_{stat} denoting the time scale for stationarity. Births are observed if birth factors arrive when the promoter is active. The illustrative examples here consider birth factor arrivals as a Poisson process with 20 expected arrivals in $T_{\text{stat}} = 20$ s, $k_0 = 0.3$ s^{-1}, and $k_1 = 0.2$ s^{-1}. **(b)** The master urn represents the cytosol of a cell and contains black and white balls, which represent birth factors and solutes other than the birth factors, respectively. Each trial consists of a random sampling of black balls from this urn. **(c)** The recipient urns of red and grey colors represent time intervals with the promoter in the active (ON) or inactive (OFF) states, respectively. Each trial consists of a random ordering of these urns as shown at the bottom. The boundary of each recipient urn is marked with a black dot. **(d)** A trial concludes with a random assignment of the black balls sampled from the master urn to the recipient urns. Assignments to red recipient urns are counted as births. In keeping with the trajectories in **a**, these trials consider birth factor arrivals as a Poisson process with 20 expected arrivals in $T_{\text{stat}} = 20$ s, six red urns and four grey urns.

urn (cytosol) and a set of recipient urns of two or more colors (time intervals; see Fig. 1b,c). Each trial in our urn scheme consists of two steps. By virtue of the first principle, the first step of sampling balls (RNA polymerases or ribosomes) from the master urn is done independently of considerations for promoter state transitions. By utilizing the second principle, we devise recipient urns such that each of them represents a time interval with homogeneous promoter activity. In the second step, we assign the balls sampled from the master urn to the recipient urns. We show that the probability distributions of counts of the mRNA and protein molecules from a broad range of models without feedback are identical to that of the balls in the recipient urns of a specific color, say red, which represents the active time intervals (Fig. 1d). If the sampling distribution of balls from the master urn is a Poisson distribution and there are recipient urns of two colors, our urn scheme yields the solution of Peccoud and Ycart. If the sampling distribution is negative binomial instead, the urn scheme yields the solution of Shahrezaei and Swain. If there are urns of more than two colors, it yields the solutions for models with multiple promoter states. Our approach yields intuitive solutions for the probability distributions in all cases, and physical interpretations of the parameters of the solutions of the Peccoud-Ycart and Shahrezaei-Swain models. Using a simplified version of the model of Kumar *et al.* [33] as an example, we illustrate that the intuition gained from our analysis of models without feedback may also be useful in understanding the existing solutions of models with feedback. Additionally, we validate the urn schemes by proving that they yield the same probability distributions as the chemical master equations, and by comparing their simulations with simulations of the corresponding kinetic schemes.

Our urn model yields the probability distribution of protein counts for a model with an arbitrary promoter architecture and one active state. This distribution is also a member of the Kemp families of distributions, and has a ${}_{p+1}F_p$ generalized hypergeometric function as its generating function for a model with p promoter states. Approximation of one of our solutions leads to the interpretation of active transcription dynamics in terms of transcriptional *lapses*, which we define as short-lived events of transcriptional inactivation separated by relatively long active intervals. We find that transcriptional lapses are a more accurate description of expression dynamics than transcriptional bursts if the promoter spends more time in the active state than in the inactive state (henceforth called *activated* expression regardless of whether there are any factors regulating the activation or not). Additionally, we discuss a generalization of the Delaporte distribution to fit single-cell data in cases where a priori knowledge about the activation status of genes is lacking, and highlight the current limitations of our models.

Results

The urn scheme

Sampling from the master urn

The master urn represents the cytosol and contains balls of two colors — black and white (Fig. 1b). Each black ball represents a *birth factor*, which we define to be the RNA polymerase for mRNAs and the ribosome for proteins. The white balls represent solutes other than the birth factor. From this urn, we sample black balls over one mean lifetime of the mRNA or protein depending on the time scale, T_{stat} for the intended solution to reach stationarity. The sampling process is defined by the kinetic process under consideration. Most models of transcription implicitly assume that the rate constant for RNA polymerases to collide with the promoter site does not vary with time but whether a colliding polymerase successfully binds the promoter and transcribes the gene depends on the promoter state. Hence, the number of arrivals follows the Poisson distribution, say with mean μ per mRNA lifetime and the urn scheme for Poisson process applies for sampling balls that represent RNA polymerases. We denote the probability distribution of the count, m_1 of polymerase arrivals in one mRNA lifetime as $\text{Pois}\left(\bullet_{m_1}|\mu\right)$, which also denotes the sampling distribution of black balls from the master urn for mRNA distributions (see Supplementary Section 1.1 for a detailed derivation of the Poisson distribution using the urn model). On the other hand, most models of translation assume that an mRNA molecule is degraded much faster than its protein counterparts, and that in its negligibly short lifetime, it binds a geometrically distributed number of ribosomes resulting in a *burst* of proteins. In our urn scheme, we draw a geometrically distributed number of ribosomes for each potential event of transcription. Hence, the cumulative count, m_2 of ribosome arrivals on mRNAs over one protein lifetime follows the negative binomial distribution with parameters α and β, which represent the number of potential transcriptions in one protein lifetime and the mean number of ribosomes that bind to each mRNA, respectively (the interpretation of α is given in more detail later). We denote this $\text{NB}\left(\bullet_{m_2}|\alpha,\beta\right)$ (see Supplementary Section 1.2 for a detailed derivation of the negative binomial distribution using the urn model).

Assignment to the recipient urns

Each recipient urn represents a time interval with a fixed rate constant for gene expression. Each urn has one of two or more colors, with the number of colors dependent on the number of promoter states (Fig. 1c). We defer the mathematical exposition of our urn scheme to a later section. Here, it suffices that at stationary state, the length of time interval captured by an urn is the same for all the urns and represents the characteristic time scale of promoter state transitions. Furthermore, the total time captured by the set of recipient urns equals T_{stat}. For a genetic switch with two states, let grey and red urns represent the inactive and active states, respectively. Also, say n_{grey} and n_{red} represent the numbers of grey and red urns, respectively. At stationary state, n_{red} and n_{grey} are fixed and their permutations represent samples of state-transition trajectories. The interaction of the time points of arrivals of birth factors with the promoter state transitions determines the outcome of interest, which is the number of births (Fig. 1d). In the urn model parlance, we capture this by assigning the black balls from the master urn to the recipient urns. If a

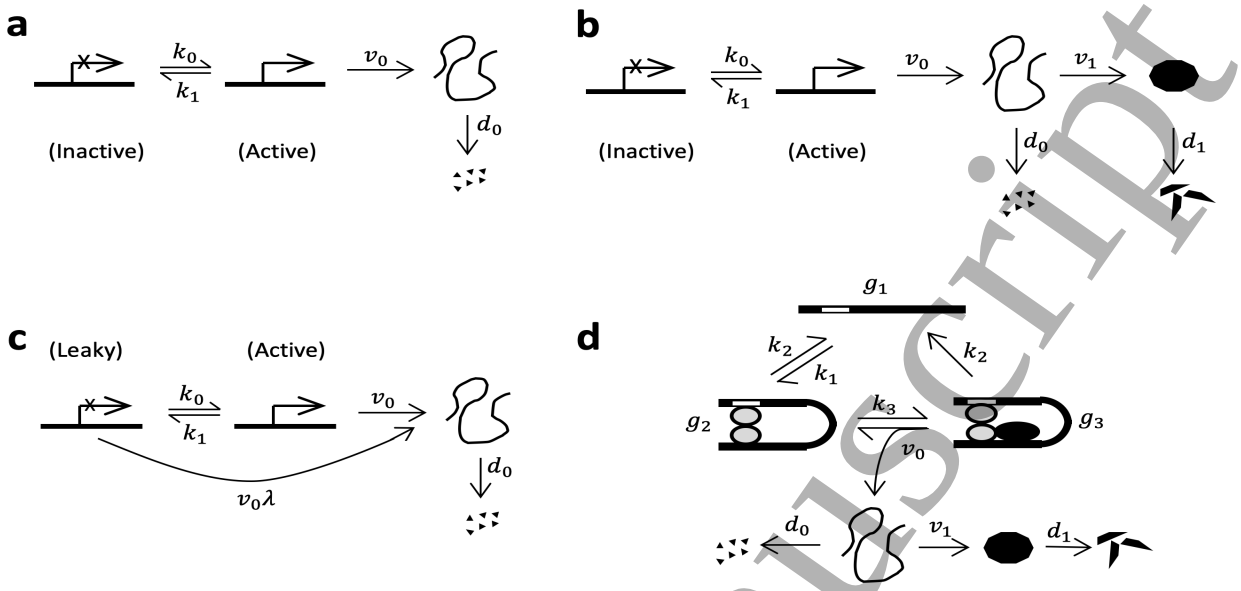

Figure 2. Kinetic schemes for gene expression with switching of promoter states. **(a)** The Peccoud-Ycart model considers a promoter that switches between inactive and active states with rate constants k_0 and k_1. The active state transcribes with rate constant v_0 and mRNAs are degraded with rate constant d_0. **(b)** The Shahrezaei-Swain model builds on the Peccoud-Ycart model by accounting for translations of mRNAs with rate constant v_1 and degradation of proteins with rate constant d_1. **(c)** The leaky two-state model generalizes the Peccoud-Ycart model by replacing the inactive state with a leaky state that transcribes with rate constant $v_0\lambda$. **(d)** Cao *et al.* consider a promoter that transitions between three states — g_1, g_2 and g_3. In state g_1, neither the transcription factor nor the RNA polymerase are bound to the promoter. In state g_2, the transcription factor is bound to the promoter and in state g_3, both the polymerase and the transcription factor are bound. State g_3 releases the polymerase with rate constant v_0, which results in transcription and translations as well as a transition to g_2. In addition, the model allows reversible transitions between g_1 and g_2 with rate constants k_1 and k_2, transition from g_2 to g_3 with rate constant k_3 and transition from g_3 to g_1 with rate constant k_2.

ball is assigned to a red urn, a birth is observed. The outcome of interest is the number of balls assigned to the red urns collectively. Since the times of arrivals of birth factors are independent of the trajectory of promoter state transitions, the process of assigning balls to the recipient urns allows equal likelihood of assignment to all urns. Let us say that in an experiment, we sample $m+i$ black balls from the master urn. The probability that a random assignment to the recipient urns results in exactly m balls assigned to the red urns is given by a negative hypergeometric distribution. To see this, note that the number of ways to divide $m+i$ balls in $n_{\text{red}}+n_{\text{grey}}$ urns is $\binom{n_{\text{red}}+n_{\text{grey}}+m+i-1}{m+i}$, which is the same as the number of ways to permute $m+i$ identical balls and $n_{\text{red}}+n_{\text{grey}}-1$ identical dividers [51]. Of these, $\binom{n_{\text{red}}+m-1}{m}\binom{n_{\text{grey}}+i-1}{i}$ are such that exactly m balls are assigned to the red urns. Hence, the probability of the outcome m is the ratio of this quantity to the total number of ways to assign $m+i$ balls. We denote the resulting negative hypergeometric distribution as $\text{NH}(\bullet_m \mapsto n_{\text{red}} | \bullet_{m+i} \mapsto \{n_{\text{red}}, n_{\text{grey}}\})$, where '$\mapsto$' represents the process of assignment (see Supplementary Section 2).

Application to the Peccoud-Ycart model

The Peccoud-Ycart model for probability distribution of mRNA counts considers a promoter that can exist in active and inactive states (Fig. 2a). Using the urn model parlance here, the black balls represent RNA polymerases. To observe an outcome of m_1 transcriptions, the sample drawn from the master urn must contain m_1 or more black balls, say m_1+i_1 with $i_1 \geq 0$. The probability of this event is $\text{Pois}(\bullet_{m_1+i_1}|\mu)$. Given m_1+i_1 balls, n_{red} red recipient urns and n_{grey} grey recipient urns, the probability that exactly m_1

balls are assigned to the red urns is $\mathrm{NH}\left(\bullet_{m_1} \mapsto n_{\mathrm{red}} | \bullet_{m_1+i_1} \mapsto \{n_{\mathrm{red}}, n_{\mathrm{grey}}\}\right)$. The joint probability of m_1 balls assigned to the red urns and i_1 to the grey urns is given by the product of the said Poisson and negative hypergeometric distributions. A marginal of the joint probability yields the probability of m_1 transcriptions,

$$P\left(\bullet_{m_1} \mapsto n_{\mathrm{red}} | \mu, n_{\mathrm{red}}, n_{\mathrm{grey}}\right) = \sum_{i_1=0}^{\infty} \mathrm{Pois}\left(\bullet_{m_1+i_1} | \mu\right) \times \mathrm{NH}\left(\bullet_{m_1} \mapsto n_{\mathrm{red}} | \bullet_{m_1+i_1} \mapsto \{n_{\mathrm{red}}, n_{\mathrm{grey}}\}\right). \quad (1)$$

Its generating function is given in terms of the Kummer's hypergeometric function of the first kind, ${}_1F_1\left[\begin{smallmatrix} n_{\mathrm{red}} \\ n_{\mathrm{red}}+n_{\mathrm{grey}} \end{smallmatrix}; \mu(z-1)\right]$, which is formally identical to the solution of Peccoud and Ycart (see Supplementary Section 3.1 for the proof). We defer the mapping of the kinetic parameters to parameters of the urn model to a later section.

Application to the Shahrezaei-Swain model

Similarly to the Peccoud-Ycart model, the Shahrezaei-Swain model considers transcriptionally inactive and active states but also accounts for translation to solve for the probability distribution of protein counts (Fig. 2b). Hence, for urn modeling in this case, we let the black balls represent ribosomes. To observe m_2 translations, the sample drawn from the master urn must contain $m_2 + i_2$ black balls with $i_2 \geq 0$, which happens with probability $\mathrm{NB}\left(\bullet_{m_2+i_2} | \alpha, \beta\right)$. Once again, we consider n_{red} red recipient urns and n_{grey} grey recipient urns, but in this case they represent *translationally* active or inactive intervals, respectively. Note that a subset of recipient urns that represented transcriptionally active time intervals in the case of Peccoud-Ycart model might not receive any polymerase arrivals, which renders them translationally inactive. Hence, n_{red} and n_{grey} have a different mapping to the kinetic parameters than in case of the Peccoud-Ycart model, as we show later. Regardless, given n_{red} and n_{grey}, the probability of m_2 translations follows from similar arguments as before,

$$P\left(\bullet_{m_2} \mapsto n_{\mathrm{red}} | \alpha, \beta, n_{\mathrm{red}}, n_{\mathrm{grey}}\right) = \sum_{i_2=0}^{\infty} \mathrm{NB}\left(\bullet_{m_2+i_2} | \alpha, \beta\right) \times \mathrm{NH}\left(\bullet_{m_2} \mapsto n_{\mathrm{red}} | \bullet_{m_2+i_2} \mapsto \{n_{\mathrm{red}}, n_{\mathrm{grey}}\}\right). \quad (2)$$

The generating function for this distribution is given in terms of the Gaussian hypergeometric function, ${}_2F_1\left[\begin{smallmatrix} \alpha,\ n_{\mathrm{red}} \\ n_{\mathrm{red}}+n_{\mathrm{grey}} \end{smallmatrix}; \beta(z-1)\right]$, which is formally identical to the solution of Shahrezaei and Swain (see Supplementary Section 3.2 for the proof).

Application to the leaky two-state model

The leaky two-state model for mRNAs considers two states with the rate constants of transcription differing by a constant factor, say λ (Fig. 2c). Let the expected numbers of arrivals of RNA polymerase in one mRNA lifetime be $\mu\lambda$ and μ in the leaky and fully active states, respectively, with $0 < \lambda < 1$. Since we consider a model for probability distribution of mRNA counts, the black balls drawn from the master urn represent RNA polymerases. In this case, we consider two parallel experiments in our urn scheme, which represent the contributions of a constitutive component with the expected value of $\mu\lambda$, and a regulated component with the expected value of $\mu(1-\lambda)$, i.e., the component of gene expression under control of the genetic switch. In the first experiment, we draw a sample of $m_{1,a}$ black balls from the master urn with the probability $\mathrm{Pois}\left(\bullet_{m_{1,a}} | \mu\lambda\right)$, which represents the leakage that is unaffected by promoter state transitions. Hence, with respect to this sample, all recipient urns are red and all of it is counted towards transcriptions. In the second experiment, we draw a sample of $m_{1,b} + i_1$ balls from the master urn with the probability $\mathrm{Pois}\left(\bullet_{m_{1,b}+i_1} | \mu - \mu\lambda\right)$, which represents the regulated component of polymerase arrivals. This experiment proceeds as described earlier for the Peccoud-Ycart model and allows us to derive the probability of $m_{1,b}$ transcriptions from the regulated component by substituting m_1 with $m_{1,b}$ and μ with $\mu(1-\lambda)$ in Eq. 1. The overall outcome of interest is $m_1 = m_{1,a} + m_{1,b}$. Its probability is given by the convolution rule and has a generating function that is the product of the generating functions for $\mathrm{Pois}\left(\bullet_{m_{1,a}} | \mu\lambda\right)$ and that for the regulated component, i.e., $e^{\mu\lambda(z-1)} {}_1F_1\left[\begin{smallmatrix} n_{\mathrm{red}} \\ n_{\mathrm{red}}+n_{\mathrm{grey}} \end{smallmatrix}; \mu(1-\lambda)(z-1)\right]$, which is consistent with the solution of

Cao and Grima [22]. Note that previously, we have shown in the context of a model for the *lac* operon of *E. coli* that this generating function can be viewed as a convolution of contributions from a leaky sub-system of Lac repressor-bound states and transitions to the repressor-free state [24]. A recent manuscript also utilizes an identical concept [59]. In this section, we have shown that the same concept is easily accommodated in the framework of our urn model. Further, the distribution of proteins from a leaky two-state model can be derived similarly.

Application to models with multiple states

Analytical solutions are available for models with more than two but a fixed number of promoter states [27,34] as well as those with an arbitrary number of states [31]. Next, we consider the model by Cao *et al.*, where the promoter exists in three states, say g_1, g_2 and g_3 with g_3 being active (Fig. 2d). For the mRNA counts, the sampling distribution of $m_1 + i_1 + j_1$ balls (RNA polymerases) from the master urn is $\text{Pois}\left(\bullet_{m_1+i_1+j_1}|\mu\right)$, where $i_1, j_1 \geq 0$. We account for the additional promoter state by adding another layer of recipient urns. First, we divide the balls between urns that correspond to the sub-system of state g_1, and the sub-system of g_2 and g_3 collectively (grouping the promoter states into the sub-system of state g_2, and the sub-system of g_1 and g_3 collectively is also allowed). Let there be n_{blue} blue and n_{ppl} purple urns for these sub-systems, respectively. Then, the probability of assigning $m_1 + i_1$ balls to the purple urns is $\text{NH}\left(\bullet_{m_1+i_1} \mapsto n_{\text{ppl}}|\bullet_{m_1+i_1+j_1} \mapsto \{n_{\text{ppl}}, n_{\text{blue}}\}\right)$. The balls assigned to the purple urns are further re-assigned to another layer of red and grey recipient urns — n_{red} and n_{grey} in number, respectively. Let the grey and red urns represent the states g_2 and g_3, respectively (or the states g_1 and g_3, respectively, if purple urns represent the sub-system of g_1 and g_3). The outcome of interest, assignment of m_1 balls to the red recipient urns has the probability $\text{NH}\left(\bullet_{m_1} \mapsto n_{\text{red}}|\bullet_{m_1+i_1} \mapsto \{n_{\text{red}}, n_{\text{grey}}\}\right)$. Hence, the probability of m_1 transcriptions is

$$\begin{aligned} P\left(\bullet_{m_1} \mapsto n_{\text{red}}|\mu, n_{\text{red}}, n_{\text{grey}}, n_{\text{ppl}}, n_{\text{blue}}\right) &= \sum_{i_1=0}^{\infty}\sum_{j_1=0}^{\infty} \text{Pois}\left(\bullet_{m_1+i_1+j_1}|\mu\right) \times \\ &\quad \text{NH}\left(\bullet_{m_1+i_1} \mapsto n_{\text{ppl}}|\bullet_{m_1+i_1+j_1} \mapsto \{n_{\text{ppl}}, n_{\text{blue}}\}\right) \times \\ &\quad \text{NH}\left(\bullet_{m_1} \mapsto n_{\text{red}}|\bullet_{m_1+i_1} \mapsto \{n_{\text{red}}, n_{\text{grey}}\}\right), \end{aligned} \tag{3}$$

which has a generalized hypergeometric function, ${}_2F_2\left[\begin{matrix}n_{\text{red}}, & n_{\text{ppl}} \\ n_{\text{red}}+n_{\text{grey}}, & n_{\text{ppl}}+n_{\text{blue}}\end{matrix}; \mu(z-1)\right]$ as its generating function (see Supplementary Section 3.3 for the proof). This solution can be extended to the protein distribution for a model with two inactive and one active promoter states by replacing the Poisson distribution with the negative binomial distribution. This yields the solution by Cao et al. [27] (see Supplementary Section 3.4 for the proof). It can be extended to a model of an arbitrary number of promoter states with all but one inactive by adding additional layers of recipient urns. Then, if sampling from the master urn follows the Poisson distribution, we get the solution by Zhou and Liu for the probability distribution of mRNA counts [31] (see Supplementary Section 3.5 for the proof). Once again, if sampling from the master urn follows the negative binomial distribution, we get the solution for protein counts (see Supplementary Section 4). Note that Qiu et al. [35] provide a procedure to iteratively obtain binomial moments of various orders for a model with arbitrary promoter architecture, which could be used to obtain analytical values for probabilities of the resulting protein counts. However, they do not provide an explicit expression for the probability distribution. To the best of our knowledge, for a model considering an arbitrary number, say p, of promoter states with all but one inactive, our derivation of the probability distribution of protein counts and its generating function as a generalized hypergeometric function with $p + 1$ numerator and p denominator parameters is a novel result. This distribution is also a member of the Kemp families of distributions (see Supplementary Section 4 for the proof).

Relationship between the kinetic and urn model parameters

We have shown that our urn model yields probability distributions that are formally identical to those from kinetic models. In this section, we obtain the relationship between parameters of the kinetic and urn models. To this end, say, we follow a cell in real time starting at time $t = 0$ when the gene system under consideration

is at stationary state (e.g., Fig. 1a shows trajectories of three cells). We define that event E_m occurs when m births are observed in a time interval given by the time scale, T_{stat} for reaching stationarity (e.g., the top panel in Fig. 1a illustrates the event E_2). Let M be a random variable representing the number of arrivals of birth factors in T_{stat} time (the number of black balls in any trajectory shown in Fig. 1a), $\{T_1, T_2, ..., T_M\}$ be the random variables representing the time points of arrivals (x-coordinates of the spikes in Fig. 1a), and R_{ON} be a random variable representing the set of time points when the promoter is active (set of all the time points in red colored segments along the x-axis in any trajectory shown in Fig. 1a). Then, E_m occurs when $\sum_{\ell=1}^{M} 1_{T_\ell \in R_{\text{ON}}} = m$, where $1_{T_\ell \in R_{\text{ON}}}$ is an indicator variable that equals 1 if $T_\ell \in R_{\text{ON}}$ and 0 otherwise. For this to happen, there must be at least m arrivals of the birth factors in total, i.e., $M = m + i$ such that $i \geq 0$. Given that this condition is met, there must be exactly m arrivals during the active time intervals, i.e., $\sum_{\ell=1}^{m+i} 1_{T_\ell \in R_{\text{ON}}} = m$. Essentially, we find that the event of m births can be decomposed into two simpler events, whose probabilities can be derived separately. Next, let us define R_{red} as the counterpart of R_{ON} in the urn space (Fig. 1d). In other words, R_{red} is the subset of time points from the set $[0, T_{\text{stat}}]$ that fall in red urns for a sample permutation of the recipient urns. Say, $E_{\text{urn},m}$ represents the event $\sum_{\ell=1}^{M} 1_{T_\ell \in R_{\text{red}}} = m$. Then, we must choose n_{red} and n_{grey} such that $P(E_m) = P(E_{\text{urn},m})$.

For the Peccoud-Ycart model, $T_{\text{stat}} = d_0^{-1}$ and RNA polymerases function as the birth factors. As we mentioned, the sampling of polymerases from the master urn can be modeled as a Poisson process, which yields $\mu = {}^{v_0}/_{d_0}$. Next, we solve for n_{red} and n_{grey}. Since the T_ℓ's are independent of each other, $P(T_{\ell_1} \in R_{\text{ON}})$ is independent of $P(T_{\ell_2} \in R_{\text{ON}})$ for $\ell_1 \neq \ell_2$. Hence, $P(E_m) = P(E_{\text{urn},m})$ if $P(T_\ell \in R_{\text{ON}}) = P(T_\ell \in R_{\text{red}})$ for all ℓ. Let the rate constants for promoter state transitions be k_0 and k_1 (Fig. 2a). At stationary state, $P(T_\ell \in R_{\text{ON}}) = {}^{k_0}/_{k_0+k_1}$. To derive $P(T_\ell \in R_{\text{red}})$, let w_{red} and w_{grey} represent the time duration captured by each of the red and grey urns, respectively, and $\min(w_{\text{red}}, w_{\text{grey}})$ be the minimum of the two. Then, for T_ℓ close to the boundaries of the set $[0, T_{\text{stat}}]$, i.e. for T_ℓ less than $\min(w_{\text{red}}, w_{\text{grey}})$ away from the boundaries, $P(T_\ell \in R_{\text{red}}) = {}^{n_{\text{red}}}/_{n_{\text{red}}+n_{\text{grey}}}$. This is because all of the duration from $t = 0$ to $\min(w_{\text{red}}, w_{\text{grey}})$ is contained within a single urn, which can either be red or grey. On the other hand, for an arbitrary choice of T_ℓ, $P(T_\ell \in R_{\text{red}}) = {}^{n_{\text{red}} w_{\text{red}}}/_{(n_{\text{red}} w_{\text{red}} + n_{\text{grey}} w_{\text{grey}})}$. In other words, the probability that a randomly chosen time point falls in a red urn is given by the fraction of time in the interval $[0, T_{\text{stat}}]$ that is covered by the red urns. At stationary state, whether T_ℓ falls in R_{red} should be the same for all T_ℓ, which requires $w_{\text{red}} = w_{\text{grey}}$. Let us replace w_{red}, w_{grey} with w. Now, if we were to arbitrarily pick a polymerase arrival and shift the corresponding T_ℓ to the left or right by a fixed amount (say, by grabbing one of the spikes in Fig. 1d), whether the shifted spike still falls in an urn of the same color depends on w. In simple words, while there is temporal heterogeneity in transcriptional activity over long periods, in short time windows around any event of polymerase arrival, transcriptional activity is homogeneous. Given a suitable choice of w, these time windows can be modeled as if they were composed of urns of the same color. w is determined by the transient time scale for switching of transcriptional activity, which is $w = (k_0 + k_1)^{-1}$ (see the section on reversible unimolecular reactions in McQuarrie [60] for derivation of the transient time scale). Finally, for $P(T_\ell \in R_{\text{ON}}) = P(T_\ell \in R_{\text{red}})$, ${}^{n_{\text{red}}}/_{n_{\text{red}}+n_{\text{grey}}} = {}^{k_0}/_{k_0+k_1}$. Since, $w(n_{\text{red}} + n_{\text{grey}}) = d_0^{-1}$, we obtain $n_{\text{red}} = {}^{k_0}/_{d_0}$ and $n_{\text{grey}} = {}^{k_1}/_{d_0}$. Essentially, we find that n_{red} and n_{grey} are equal to the rate constants for the promoter to switch to the active and inactive states, respectively, scaled with respect to the mRNA degradation rate constant. By replacing these parameter values in Eq. 1, we retrieve the analytical solution of Peccoud and Ycart.

Shahrezaei and Swain solved a kinetic model for protein production assuming $d_0 \gg d_1$, which implies $T_{\text{stat}} = d_1^{-1}$. In this case, ribosomes function as the birth factors and for the system to be active for protein production, the gene must be transcriptionally active and RNA polymerase arrivals must occur. In other words, if we start our observation with the promoter in the transcriptionally inactive state, the waiting time for transition to the translationally active state is greater than k_0^{-1}. In the remaining part of this section, we first derive the rate constant for translational activity, and then apply the approach described above to obtain n_{red} and n_{grey} for the Shahrezaei-Swain model. To this end, let us pick a time point randomly as $t = 0$. Next, we define $p_0(t)$ as the probability that the gene is transcriptionally inactive at time t and no mRNA has been produced in time $[0, t]$. Similarly, $p_1(t)$ is the probability that the gene is transcriptionally

active at time t but no mRNA has been produced in time $[0, t]$. Then,

$$\frac{dp_0(t)}{dt} = k_1 p_1(t) - k_0 p_0(t), \quad (4)$$

$$\frac{d[p_0(t) + p_1(t)]}{dt} = -v_0 p_1(t). \quad (5)$$

We are interested in the time scale at which $p(t) = p_0(t) + p_1(t)$ approaches 0, i.e., the time scale at which the marginal probability of no mRNA production decays to 0. To this end, Eqs. 4-5 can be combined to get a second order differential equation,

$$\frac{d^2 p(t)}{dt^2} + (v_0 + k_0 + k_1)\frac{dp(t)}{dt} + v_0 k_0 p(t) = 0, \quad (6)$$

which admits solutions of the form $e^{-t/\tilde{t}}$, where $\tilde{t} > 0$ is a characteristic time scale of the system. Substituting $e^{-t/\tilde{t}}$ in Eq. 6 yields a quadratic equation, with the roots

$$\tilde{t}^{-1} \equiv \frac{1}{2}\left(v_0 + k_0 + k_1 \pm \sqrt{(v_0 + k_0 + k_1)^2 - 4v_0 k_0}\right). \quad (7)$$

Inverse of the larger of the roots, i.e. the slow time scale (say, $\tilde{t}_s$) gives the waiting time for mRNA production and the inverse of the smaller one yields the fast time scale (say, $\tilde{t}_f$). We interpret $\tilde{t}_f^{-1}$ as the rate constant for occurrence of any event, i.e., promoter state transition or polymerase arrival. For example, if we know that a polymerase arrival occurs at any time point t, it is likely that there will be no arrival in the time window $(t, t + \tilde{t}_f)$ because the polymerase arrival occurs as fast as the fast time scale of the system allows. Now, we can derive the relationship between the kinetic and urn model parameters in terms of these time scales. First, we sample ribosomes from the master urn, such that for each $\tilde{t}_f$ time window in T_{stat}, we draw a geometrically distributed number of ribosomes. The geometric distribution has the mean $\beta = {}^{v_1}/_{d_0}$ and we draw $\alpha = {}^{T_{\text{stat}}}/_{\tilde{t}_f}$ geometrically distributed samples. Hence, the probability distribution for a sample of $m_2 + i_2$ ribosomes is given by the negative binomial distribution with the said values for α and β. Next, we distribute these in the red and grey recipient urns. Similarly to arrivals of polymerases, arrivals of ribosomes are independent of each other. We can use the same method as described for Peccoud-Ycart model to derive n_{red} and n_{grey}. The difference is that here, red urns represent translationally active time windows. Hence, their number is given by $n_{\text{red}} = {}^{\tilde{t}_s^{-1}}/_{d_1}$ and $n_{\text{grey}} = {}^{(k_0+k_1-\tilde{t}_s^{-1})}/_{d_1}$, which are the rate constants for the system to switch to the translationally active and inactive states, respectively, scaled with respect to the protein degradation rate constant. By using these parameter values in Eq. 2, we retrieve the solution of Shahrezaei and Swain.

For the mRNA distribution from the leaky two-state model, the μ, n_{red} and n_{grey} parameters have the same interpretations and relationship to the kinetic parameters as in the Peccoud-Ycart model. For the mRNA distribution from the model of Cao *et al.* in Eq. 3, $\mu = {}^{v_0}/_{d_0}$ also has the same interpretation as in the Peccoud-Ycart model. Further, the above results suggest that the n_{red}, n_{grey}, n_{ppl}, and n_{blue} parameters correspond to the rate constants for switching between sub-systems of the promoter states scaled with respect to the mRNA degradation rate constant. However, deriving the expressions for these parameters in terms of the kinetic parameters based on the intuitive arguments presented above is challenging. The relationships between the urn and kinetic model parameters can be obtained by comparing Eq. 3 with the solution available in Cao *et al.* [27].

Solving the urn model using the inclusion-exclusion principle

In the previous sections, we have written the probability of *exactly* m_1 assignments to the red urns out of a sample of $m_1 + i_1$ balls from the master urn directly as a negative hypergeometric distribution. The same probability could be written from an alternative perspective, where we start with the probabilities for at least i_1 assignments to the grey urns, at least $i_1 + 1$ assignments to the grey urns, ..., at least $i_1 + m_1 - 1$ assignments to the grey urns and all $i_1 + m_1$ assignments to the grey urns. Then, we combine these as an

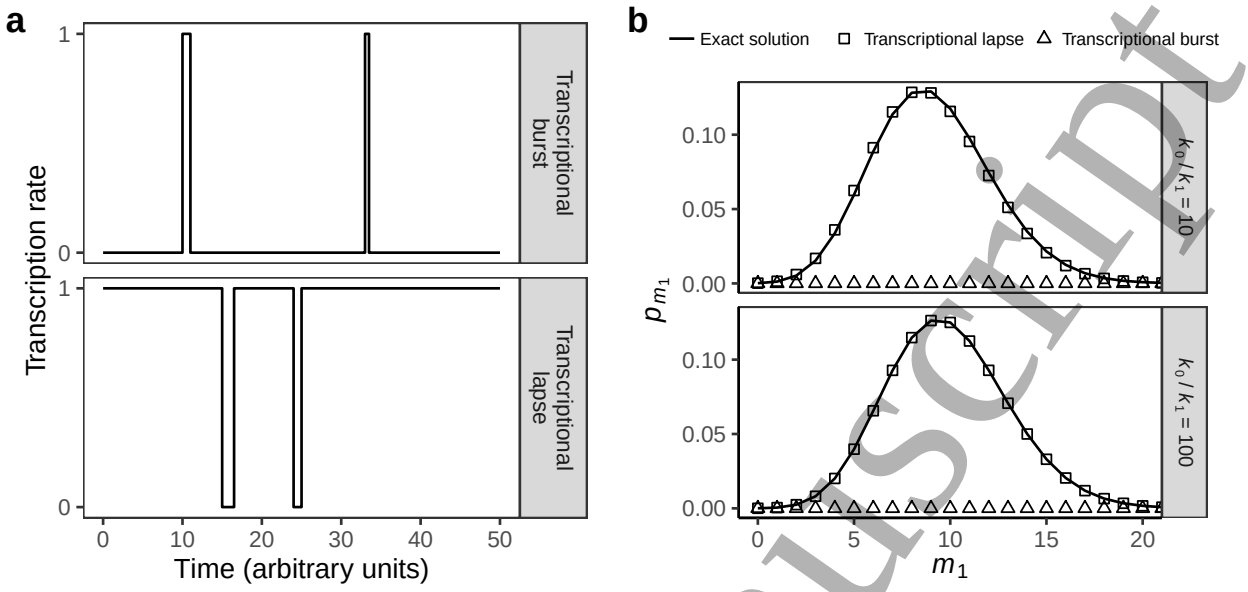

Figure 3. The dynamics of transcription for activated genes involve transcriptional lapses. (a) The transcriptional dynamics of repressed genes are characterized by short-lived periods of activity (bursts) and that of activated genes by short-lived periods of inactivity (lapses). **(b)** Stationary state probability of m_1 mRNAs obtained from exact solution of the Peccoud-Ycart model (solid line) and its limiting forms for transcriptional lapses (rectangles) and bursts (triangles). Both the panels show the probabilities in the activated case, when k_0 is 10 times k_1 (top) or 100 times k_1 (bottom) with $k_1 = 0.01$ s^{-1}, $v_0 = 0.05$ s^{-1} and $d_0 = 0.005$ s^{-1}. The peaks from the transcriptional burst model appear at $m_1 \gg 10$ for both the panels.

alternating sum using the principle of inclusion-exclusion to write the probability of exactly m_1 assignments to the red urns (see Supplementary Section 5). For the Peccoud-Ycart model, this approach yields an expression for $P\left(\bullet_{m_1} \mapsto n_{\rm red} | \mu, n_{\rm red}, n_{\rm grey}\right)$ that is the $m_1^{\rm th}$ coefficient in the Maclaurin series expansion of Kummer transformation of the Peccoud-Ycart solution, $e^{\mu(z-1)}{}_1F_1\left[\begin{smallmatrix} n_{\rm grey} \\ n_{\rm red}+n_{\rm grey}\end{smallmatrix}; -\mu(z-1)\right]$ with $\mu = {}^{v_0}/_{d_0}$, $n_{\rm red} = {}^{k_0}/_{d_0}$ and $n_{\rm grey} = {}^{k_1}/_{d_0}$. This way of representing the Peccoud-Ycart solution leads to an interpretation of the transcriptional dynamics in terms of transcriptional *lapses* as we show next, and contrast with the dynamics of transcriptional bursts [61].

For a repressed gene with ${}^{k_1}/_{d_0} \gg 1$, the hypergeometric function, ${}_1F_1\left[\begin{smallmatrix} n_{\rm red} \\ n_{\rm red}+n_{\rm grey}\end{smallmatrix}; \mu(z-1)\right]$ reduces to the generating function for the negative binomial distribution, $\left[1 - {}^{\mu(z-1)}/_{n_{\rm grey}}\right]^{-n_{\rm red}}$ [62]. Here, ${}^{\mu}/_{n_{\rm grey}}$ is interpreted as the mean number of mRNAs produced when the promoter switches to the active state (burst size), $n_{\rm red}$ as the frequency of switching to the active state (burst frequency), and the hypergeometric function is the generating function for the Peccoud-Ycart solution. In Supplementary Section 6.1, we use a perturbation theoretic approach to show that this is valid for $k_1 \gg k_0$. Using this limiting form of the hypergeometric function, when $k_0 \gg k_1$, i.e., for an activated gene, $e^{\mu(z-1)}{}_1F_1\left[\begin{smallmatrix} n_{\rm grey} \\ n_{\rm red}+n_{\rm grey}\end{smallmatrix}; -\mu(z-1)\right]$ reduces to $e^{\mu(z-1)}\left[1 + {}^{\mu(z-1)}/_{n_{\rm red}}\right]^{-n_{\rm grey}}$ (see Supplementary Section 6.2 for proof). Hence, in both the activated and repressed scenarios, the limiting form of the Peccoud-Ycart solution has a term of the kind $\left[1 - \beta\,(z-1)\right]^{-\alpha}$, which is the generating function for a negative binomial distribution if $\alpha, \beta > 0$. However, $\beta \equiv {}^{\mu}/_{n_{\rm grey}} > 0$ in the repressed limit but $\beta \equiv -{}^{\mu}/_{n_{\rm red}} < 0$ in the activated limit. Hence, this term is not consistent with the interpretation of transcriptional dynamics in the activated case as being bursty and $\left[1 + {}^{\mu(z-1)}/_{n_{\rm red}}\right]^{-n_{\rm grey}}$ is not a probability distribution. We interpret the dynamics in the activated case as one that is composed of transcriptional lapses, whereby the gene is expressed with the rate constant v_0 for majority of the time and for short-lived intervals, there is a lapse in transcriptional activity brought about by switching of the promoter states (Fig. 3a). To the best of our knowledge, this is a novel interpretation of activated expression dynamics. In Fig. 3b, we show that for an activated gene, the approximation in terms of transcriptional

lapses is in significantly better agreement with the exact solution than that in terms of transcriptional bursts. This approximation might be useful in fitting single-cell data for activated genes.

Interpretation of existing solutions for models with feedback

In the previous sections, we analyzed models that did not consider feedback loops. Hence, our assumption that the numbers of red and grey recipient urns is fixed and independent of the number of black balls sampled from the master urn was valid. In the presence of feedback mediated by say, the protein products of the gene under consideration, the rate of switching between promoter states depends on the number of proteins in the cell. In the urn model parlance, the numbers of recipient urns of different colors and the number of black balls assigned to the red urns are interdependent. As a consequence, the steps of sampling from the master urn and assignment to the recipient urns cannot be performed independently. Due to this limitation, our urn model needs to be developed further to account for dependence between the two steps and thereby, to account for feedback, which will be subject of future work. Despite this, it is noteworthy that solutions of models with feedback are also given in terms of hypergeometric functions [25, 26, 30, 33, 63]. This is expected because the models with feedback admit Peccoud-Ycart model for mRNAs and Shahrezaei-Swain model for proteins as their special cases. Hence, solutions in presence of feedback reduce to the solutions of the Peccoud-Ycart and Shahrezaei-Swain models when the parameters related to the feedback mechanism are ignored. Note that if a feedback model considers two active states, upon ignoring feedback, its solution reduces to the solution of leaky two-state model instead. For example, if the binding rate constant for proteins to the promoter is zero in the model of Kumar *et al.* [33], we retrieve the solution to the leaky two-state model. Importantly, this means that we can write series expansions of the solutions of models with feedback in a way that some of the terms can be interpreted as the negative hypergeometric and Poisson/negative binomial probability terms that appear in solutions of the Peccoud-Ycart and Shahrezaei-Swain models. In addition, there may be other terms, which originate due to feedback. We call these feedback-dependent "correction" terms. We find that these terms might also be physically interpretable. Here, we discuss the solution to a simplified version of the model of Kumar *et al.* as an illustrative example (see Supplementary Section 7 for details). This model generalizes the Shahrezaei-Swain model by allowing the protein product of a gene to inactivate its promoter by binding to it, which is in addition to a basal rate of inactivation independent of the protein product (see Supplementary Figure 1). Using the solution in Kumar *et al.*, we expand the probability of m_2 proteins as follows,

$$\sum_{k_2=0}^{\infty}\left[\sum_{i_2=0}^{\infty}\mathrm{NB}\left(\bullet_{m_2+i_2+k_2}|\alpha,\beta\right)\times\mathrm{NH}\left(\bullet_{m_2+k_2}\mapsto n_{\mathrm{red}}|\bullet_{m_2+i_2+k_2}\mapsto\{n_{\mathrm{red}},n_{\mathrm{grey}}\}\right)\right]\times\binom{m_2+k_2}{m_2}\frac{\rho^{k_2}}{C},\quad(8)$$

where C is a normalization constant, and $\rho = {}^{r}/_{(d_1+r)}$ such that r is the rate constant for binding to the promoter and d_1 is the rate constant for protein degradation (see Supplementary Section 7). If $d_1 \gg r$, the proteins are likely to degrade before binding to the promoter. In that sense, ρ represents the feedback efficiency, i.e., higher degradation rate leads to smaller ρ. Some of the terms in Eq. 8 can now be interpreted in terms of our urn model. Say, we sample $m_2 + i_2 + k_2$ black balls (ribosomes) from the master urn, which occurs with the probability $\mathrm{NB}\left(\bullet_{m_2+i_2+k_2}|\alpha,\beta\right)$. Assign the balls to the red and grey urns (translationally active and inactive intervals, respectively), such that m_2+k_2 are assigned to the red urns, which happens with the probability $\mathrm{NH}\left(\bullet_{m_2+k_2}\mapsto n_{\mathrm{red}}|\bullet_{m_2+i_2+k_2}\mapsto\{n_{\mathrm{red}},n_{\mathrm{grey}}\}\right)$. Summing the product of the said negative binomial and negative hypergeometric probability functions over all values of i_2 gives the probability of $m_2 + k_2$ assignments to the red urns (terms inside square brackets in Eq. 8). To interpret the remaining terms $\binom{m_2+k_2}{m_2}\frac{\rho^{k_2}}{C}$, note that the effect of negative feedback is to switch the state of the promoter from active to inactive for some of the time intervals when the promoter would otherwise be active. Hence, if negative feedback were added to the Shahrezaei-Swain model, we would find that some of the $m_2 + k_2$ ribosomes that are assigned to the red urns may not result in protein production. The likelihood of such a failure to produce protein is proportional to the feedback efficiency, ρ. If exactly k_2 ribosomes out of the $m_2 + k_2$ assigned to the red urns fail, we would observe m_2 proteins produced in the duration of the time scale for stationarity. Summing over all possible values of k_2 and normalizing yields the probability function for m_2 counts in Eq. 8. Note that since we do not have an urn model that directly accounts for feedback, the correction terms,

which are collectively $\binom{m_2+k_2}{m_2}\frac{\rho^{k_2}}{C}$, do not constitute a probability distribution, although our interpretation is reminiscent of the binomial distribution $\binom{m_2+k_2}{m_2}\rho^{k_2}(1-\rho)^{m_2}$. Further development of our urn model may provide better interpretations of probability expressions resulting from models with feedback.

Validation by comparisons with stochastic simulations of the kinetic models

Besides proofs in Supplementary Section 3, we validated our urn models by comparisons with simulations of the kinetic models. We simulated the reaction schemes in Fig. 2 using the optimized direct method implementation of the Gillespie's stochastic simulation algorithm [64]. We performed 10^5 realizations of each model and saved the count distributions of mRNAs and proteins at the end of 36000 s in simulation time. We saved distributions from simulating each model for three choices of parameters (see Supplementary Tables 1-4 for the parameter values). To compare with, we simulated the corresponding urn schemes of sampling black balls from the master urn and assigning them to recipient urns. For example, for the Peccoud-Ycart model, we sampled a random number from a Poisson distribution, which represented the number of black balls from the master urn, and randomly assigned the balls to recipient urns of red and grey colors as described earlier (see links for additional data containing the R scripts). We performed 10^5 realizations of the urn scheme in each case. We computed the urn model parameters for simulations from the kinetic parameters using the relationships in Supplementary Section 8 and rounded the fractional values for numbers of recipient urns to their nearest integer. Our test cases included instances when the nearest integers were 0, in which case we used the value of 1 instead to ensure at least one recipient urn of each color. In each case, we saved the count distributions of balls in red urns.

We compared the count distributions resulting from simulations of the urn and kinetic models (Fig. 4). In majority of the cases, we observed that the two were in good agreement. Figs. 4a-d show the comparisons for the Peccoud-Ycart model, the Shahrezaei-Swain model, the leaky two-state model, and the model of Cao *et al.* as shown in Figs. 2a-d, respectively. We found that for fractional values of n_{grey} close to 0, the count distributions from the urn and the kinetic models were substantially different (see the plots labelled PY-3, SS-3, and L2S-3 in Figs. 4a-c, respectively). Due to the restriction that we can simulate the urn model only if n_{grey} is a positive integer, we must replace fractional n_{grey} with its nearest integer. However, the nearest integer of the exact value of n_{grey} in the PY-3, SS-3 and L2S-3 cases is 0. Setting n_{grey} to 0 would render the gene without promoter switching. Hence, we set n_{grey} to 1. A criterion for bimodality of count distributions in these kinetic models is that the sojourn time of the promoter in both active and inactive states (e.g., $1/k_0$ and $1/k_1$ for the Peccoud-Ycart model) be larger than or same order of magnitude as the time scale for stationarity (e.g., $1/d_0$ for the Peccoud-Ycart model) [24]. In other words, one of the criteria is that n_{red} and n_{grey} be less than 1. However, in urn model simulations, the minimum value that n_{red} and n_{grey} can take is 1. Hence, we did not observe bimodal distributions in urn model simulations. Collectively, from these simulations, we find that the count distributions from simulations of the urn schemes agree with those from the corresponding kinetic schemes.

Discussion

Comparison with existing models for the Kemp families of distributions

All problems concerning probabilities could be addressed using urn models [57]. However, it is a non-trivial task to develop physically meaningful urn schemes for any given system, even if solutions for the probability distributions of interest were known. This is because many different models can give rise to identical distributions [56]. Indeed, this is true for distributions resulting from models considering genetic switches, which are related to the Kemp families of distributions that are well-known in the urn modeling literature [65]. A general feature of some of the models that lead to such distributions is that in a series of trials, multiple instances of the same outcome tend to occur in close succession. This could result due to the presence of contagion or population heterogeneity in the stochastic process under consideration [54]. We consider a known model for each and discuss how they differ from models considering genetic switching. The Pólya-Eggenberger urn model studies the spread of contagious diseases [56]. In this model, one considers a finite urn with a

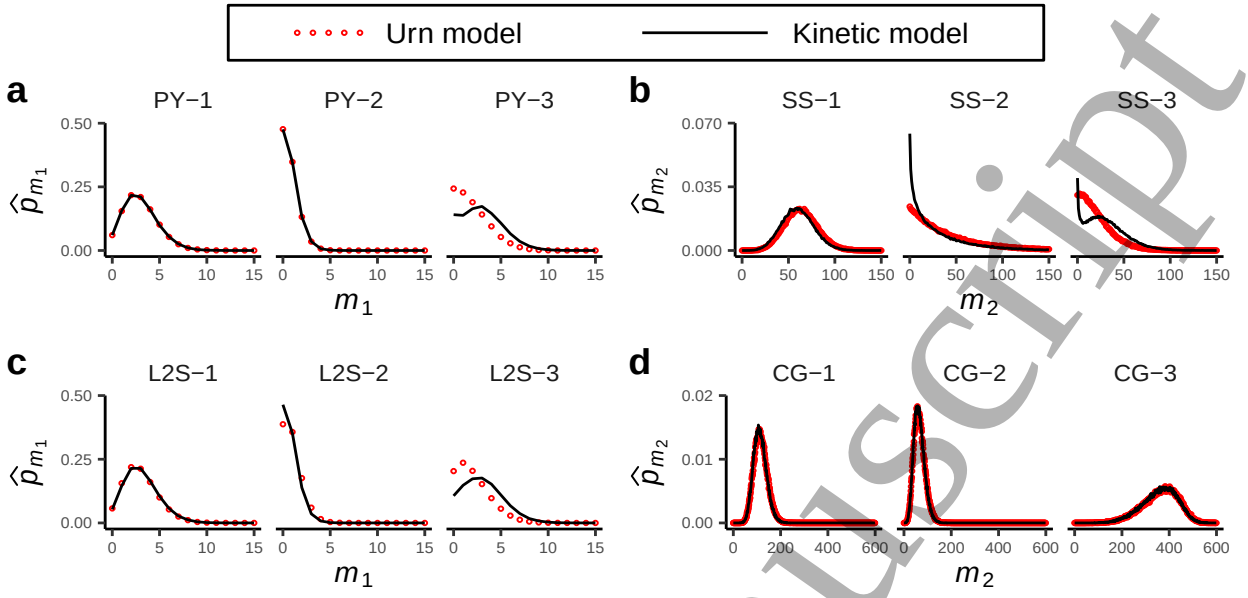

Figure 4. Count distributions from simulations of the urn models agree with those from simulating the corresponding kinetic models. Each plot is showing the observed frequencies ($\hat{p}$) of counts of mRNAs (m_1) or proteins (m_2) in distributions from simulations of the urn (red circles) and kinetic models (black solid line). The panels **a**,**b**,**c**, and **d** correspond to results for the Peccoud-Ycart model (PY), the Shahrezaei-Swain model (SS), the leaky two-state model (L2S), and the model of Cao *et al.* (CG), respectively. For the parameter values, see Supplementary Tables 1-4.

pre-specified number of white and black balls, and draws balls from the urn randomly and one at a time. The drawn ball is replaced along with additional balls of the same color, which increases the probability of drawing this color in the next trial. This process simulates the spread of pathogens once they appear in a population. However, in the models without feedback, such contagion does not exist because the concentrations of birth factors do not change with time. Hence, the process of replacing a ball with additional balls of the same color is not physically meaningful. Alternatively, Gurland considered a compound Poisson process, which is a Poisson process with a rate parameter that is also a random variable due to heterogeneity of the population in consideration [54]. For example, consider a constitutively expressed gene in a population of cells such that the rate of RNA polymerase arrival at the promoter varies from cell to cell. If the rate parameter depended on a random variable with Beta distribution, the probability distribution of the corresponding mRNA counts would be identical to the solution of the Peccoud-Ycart model. It is noteworthy that this interpretation has been recently utilized to fit single-cell RNA-seq data [45, 46]. However, the Peccoud-Ycart model, like most mechanistic models of gene expression, is solved for a homogeneous population with a fixed rate of polymerase arrivals. Hence, the interpretation as a compound Poisson process is not consistent with its kinetic scheme.

Implications for analysis of single-cell data

Our study provides helpful insights for single-cell data analysis. First, we find that the urn model parameters are related to the time scales of switching between sub-systems of the promoter states. Particularly, in the case of the Shahrezaei-Swain model, the numerator parameters of the Gaussian hypergeometric generating function are related to the fast time scale and the waiting time for transcriptions. To the best of our knowledge, these parameters have not been ascribed physical interpretations previously. Our interpretations would facilitate mechanistic understanding from analysis of single-cell data for proteins if the data were fit using the Shahrezaei-Swain solution. Second, the field has recently witnessed a rapid development of single-cell technologies and parallel advances in analytical solutions of models incorporating mechanistic details of gene expression [27, 66, 67]. As such, solutions of the detailed models would likely be utilized for

analyzing single-cell data in the future. We find that solutions of disparate mechanistic models (e.g., the models of Cao et al. and Karmakar [27, 34]) may result in identical distributions if they involve the same number of timescales for switching between the sub-systems of promoter states. Hence, if a probability distribution yields good fits to data on mRNA or protein counts, it indicates that the number of time scales represented in the distribution is adequate to describe the gene system. Additional data must be collected to distinguish between the mechanistic models leading to these time scales as well as to assess whether the probability distribution might have resulted from the presence of features such as contagion or population heterogeneity in the system. Further insights in this direction could be gained from existing statistical literature to distinguish between such features [68]. Finally, we note that in some cases, the researchers utilize the negative binomial distribution to fit their data [11]. This could be motivated by the challenges of computing hypergeometric functions for the complete range of parameter values [69]. Furthermore, the parameters of negative binomial distribution have well-understood interpretations as the burst size and frequency of expression. We find that at least for activated genes, the transcriptional dynamics are better described in terms of transcriptional lapses. In the general case, it might be worthwhile to consider a distribution with the generating function $e^{\mu(z-1)}\left[1-\beta\left(z-1\right)\right]^{-\alpha}$ to fit the data. If all the parameters, α, β and μ were non-negative, this is a Delaporte distribution, which could be interpreted in terms of a leaky two-state model with transcriptional bursts (see Supplementary Section 6) [65]. If β were allowed to be negative, the parameters could be interpreted in terms of a leaky two-state model with transcriptional lapses.

Limitations

Limitations of our approach are worth considering. First, as the mechanistic models grow in complexity, identifying the relationship between the urn model and kinetic parameters becomes a challenging task. We found the probability distribution for protein counts from a model with arbitrary promoter architecture but did not derive the corresponding relationships between the urn and kinetic parameters. It will be interesting to address this gap in future studies. Despite this limitation, our approach complements the chemical master equation approach by providing an intuitive alternative to derive the probability distributions. Second, our model assumes that birth factors arrive independently of each other and that there are no feedback loops. If these assumptions do not hold, the birth factors cannot be assigned to the recipient urns independently of each other and the numbers of recipient urns of different colors are not independent of the counts of births. Nevertheless, the solution for a model with feedback could be viewed as correction terms combined with the solution for its simplified version that ignores feedback (see Results and Supplementary Section 7). Further, we note that, like most models that have been solved analytically, our model implicitly assumes that the reactants and catalysts needed for gene expression such as tRNAs, amino acids, ribonucleotides, etc. are available in sufficient quantity [5]. In the future, generalizations of our model might be able to relax these assumptions. Third, our model makes physical sense only if the parameters such as n_{red} and n_{grey} are integer-valued. This might not be true for arbitrary values of the kinetic parameters. In this case, the factorial functions appearing in our expressions must be replaced by gamma functions as done for the urn model of the negative binomial distribution [16]. This results in probability distributions identical to the solutions of chemical master equations, as evidenced by our proofs in Supplementary Section 3. Furthermore, the replacement of factorial functions with gamma functions is justifiable. Since our models consider a natural system, the resulting probability distributions should be continuous with respect to the parameters. Hence, the function that replaces the factorial function for positive and non-integer parameter values should interpolate the factorials of natural numbers. Of the many functions that can interpolate the factorial function, that the gamma function is the correct replacement is consistent with the well-known fact that they commonly appear in the context of systems involving integrals with exponentially decaying functions [70]. In our case, the models of gene expression that we study rely on the Markovian assumption of a memoryless process, whereby the time to switch a promoter state is exponentially distributed (for example, see the solution of Eq. 6). This would lead to exponential terms within integrals of the chemical master equations, and thereby the gamma function in expressions of probability distributions. If the Markovian assumption is violated, extension to non-integer parameters using the gamma function may not apply with the urn model in its current form. Fourth, we provide physical interpretations of parameters of the probability distributions, but we have not addressed the challenges with fitting them to data. This can be an issue due to the presence

of hypergeometric functions in expressions and deserves attention in the future studies.

Conclusion

We developed an urn model and applied it to study a broad range of stochastic models of gene expression with promoter switching. Our urn model generalizes the classical birth-death model by considering the regulation of births. The classical model makes no distinction between arrivals of birth factors and births, as all arrivals cause births. However, in the presence of regulation, as in the case of genetic switches, arrivals of birth factors do not result in births if they occur when genes are inactive. Hence, we reinterpret the classical scheme of births as a scheme to sample birth factors from a master urn, which represents the cytosol of a cell. Next, we note that despite temporal heterogeneity in expression activity of genes over long time intervals, there is homogeneity in short intervals. We use this concept to devise recipient urns of two or more colors, which are discretized time intervals with the promoter existing in a single sub-system of its states for the duration of each urn. Then, we assign the birth factors to the recipient urns and count the assignments to urns of a specific color as births. Given physically intuitive choices of sampling distribution from the master urn and the numbers of recipient urns for each color, our model yields probability distributions that are identical to solutions of a range of kinetic models. We describe the physical principles that lead to our urn scheme and provide kinetic interpretations for the urn model parameters. Finally, we discuss our approach in the broader context of urn models and single-cell data analysis as well as highlight its limitations. We conclude by noting that the solutions from chemical master equations are obtained in terms of generating functions and physical intuition into origins of the expressions for probability distributions have thus far been limited for models with genetic switches. Our model facilitates direct interpretation of the probability expressions, which underscores its significance for pedagogical purposes and to interpret results from data fitting. The physical insights developed in this work will facilitate the adoption of the analytical solutions in single-cell studies.

Acknowledgements

We are grateful to two anonymous reviewers for their constructive criticism that helped improve the manuscript.

Supporting Information

Supplementary information for this article is available online. The scripts used for simulations and generating the figures are available here.

Declaration of interest

The authors declare that they have no competing interests.

Author contributions

KC and AN conceived the project. KC designed the research, developed the method, performed the simulations, generated the figures and wrote the manuscript. AN provided critical feedback and helped shape the research, analysis and manuscript. KC did the proofs other than the perturbation theoretic proofs, which were done by AN. Both the authors read and approved the manuscript.

References

1. George Orphanides and Danny Reinberg. A unified theory of gene expression. *Cell*, 108(4):439–451, 2002.

2. Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

3. Avigdor Eldar and Michael B Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167, 2010.

4. Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451, 2005.

5. Johan Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175, 2005.

6. Roy D Dar, Brandon S Razooky, Abhyudai Singh, Thomas V Trimeloni, James M McCollum, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, 2012.

7. Caroline R Bartman, Nicole Hamagami, Cheryl A Keller, Belinda Giardine, Ross C Hardison, Gerd A Blobel, and Arjun Raj. Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Molecular cell*, 73(3):519–532, 2019.

8. Shasha Chong, Chongyi Chen, Hao Ge, and X Sunney Xie. Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326, 2014.

9. Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004.

10. Keisuke Fujita, Mitsuhiro Iwaki, and Toshio Yanagida. Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. *Nature communications*, 7:13788, 2016.

11. Lok H. So, Anandamohan Ghosh, Chenghang Zong, Leonardo A. Sepulveda, Ronen Segev, and Ido Golding. General properies of the transcriptional time-series in *E. coli*. *Nature Genetics*, 43(6):554–560, 2011.

12. Daniel L Jones, Robert C Brewster, and Rob Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, 2014.

13. Tineke L Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R Larson. Transcription dynamics in living cells. *Annual review of biophysics*, 45:25–47, 2016.

14. Brian Munsky, Gregor Neuert, and Alexander Van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

15. KA Geiler-Samerotte, CR Bauer, S Li, N Ziv, David Gresham, and ML Siegal. The details in the distributions: why and how to study phenotypic variability. *Current opinion in biotechnology*, 24(4):752–759, 2013.

16. Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–61, 2008.

17. Alvaro Sanchez, Hernan G. Garcia, Daniel Jones, Rob Phillips, and Jané Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Computational Biology*, 7(3):e1001100, 2011.

18. JM Vilar and L Saiz. Suppression and enhancement of transcriptional noise by DNA looping. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 89(6):062703–062703, 2014.

19. Tyler M Earnest, Elijah Roberts, Michael Assaf, Karin Dahmen, and Zaida Luthey-Schulten. DNA looping increases the range of bistability in a stochastic model of the *lac* genetic switch. *Physical biology*, 10(2):026002, 2013.

20. Thomas B Kepler and Timothy C Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–3136, 2001.

21. Masaki Sasai and Peter G Wolynes. Stochastic gene expression as a many-body problem. *Proceedings of the National Academy of Sciences*, 100(5):2374–2379, 2003.

22. Zhixing Cao and Ramon Grima. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nature communications*, 9(1):3305, 2018.

23. Krishna Choudhary, Stefan Oehler, and Atul Narang. Protein distributions from a stochastic model of the *lac* operon of *E. coli* with DNA looping: Analytical solution and comparison with experiments. *PLoS One*, 9(7):1–14, 2014.

24. Krishna Choudhary and Atul Narang. Analytical expressions and physics for single-cell mRNA distributions of the *lac* operon of *E. coli*. *Biophysical journal*, 117(3):572–586, 2019.

25. Ramon Grima, Deena R Schmidt, and Timothy J Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *The Journal of Chemical Physics*, 137(3):035104, 2012.

26. Yves Vandecan and Ralf Blossey. Self-regulatory gene: an exact solution for the gene gate model. *Physical Review E*, 87(4):042705, 2013.

27. Zhixing Cao, Tatiana Filatova, Diego A Oyarzún, and Ramon Grima. Multi-scale bursting in stochastic gene expression. *bioRxiv*, page 717199, 2019.

28. Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 2020.

29. Hodjat Pendar, Thierry Platini, and Rahul V Kulkarni. Exact protein distributions for stochastic models of gene expression using partitioning of poisson processes. *Physical Review E*, 87(4):042720, 2013.

30. José EM Hornos, Daniel Schultz, Guilherme CP Innocentini, JAMW Wang, Aleksandra M Walczak, José N Onuchic, and Peter G Wolynes. Self-regulating gene: an exact solution. *Physical Review E*, 72(5):051907, 2005.

31. Tianshou Zhou and Tuoqi Liu. Quantitative analysis of gene expression systems. *Quantitative Biology*, 3(4):168–181, 2015.

32. Jean Peccoud and Bernard Ycart. Markovian modeling of gene product synthesis. *Theoretical Population Biology*, 48:222–234, 1995.

33. Niraj Kumar, Thierry Platini, and Rahul V Kulkarni. Exact distributions for stochastic gene expression models with bursting and feedback. *Physical review letters*, 113(26):268105, 2014.

34. Rajesh Karmakar. Conversion of graded to binary response in an activator-repressor system. *Physical Review E*, 81(2):021905, 2010.

35. Huahai Qiu, Bengong Zhang, and Tianshou Zhou. Influence of complex promoter structure on gene expression. *Journal of Systems Science and Complexity*, pages 1–15, 2018.

36. Otto G Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of Theoretical Biology*, 71(4):587–603, 1978.

37. D. R. Rigney and W. C. Schieve. Stochastic model of linear, continuous protein synthesis in bacterial populations. *Journal of Theoretical Biology*, 69(4):761–766, 1977.

38. Pavol Bokes, John R King, Andrew TA Wood, and Matthew Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of Mathematical Biology*, 64(5):829–854, 2012.

39. Jiajun Zhang and Tianshou Zhou. Markovian approaches to modeling intracellular reaction processes with molecular memory. *Proceedings of the National Academy of Sciences*, 116(47):23542–23550, 2019.

40. Zihao Wang, Zhenquan Zhang, and Tianshou Zhou. Exact distributions for stochastic models of gene expression with arbitrary regulation. *Science China Mathematics*, 63(3):485–500, 2020.

41. Tuoqi Liu, Jiajun Zhang, and Tianshou Zhou. Effect of interaction between chromatin loops on cell-to-cell variability in gene expression. *PLoS computational biology*, 12(5):e1004917, 2016.

42. Tianshou Zhou and Jiajun Zhang. Analytical results for a multistate gene model. *SIAM Journal on Applied Mathematics*, 72(3):789–818, 2012.

43. Nir Friedman, Long Cai, and X Sunney Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical review letters*, 97(16):168302, 2006.

44. Daphne Ezer, Victoria Moignard, Berthold Göttgens, and Boris Adryan. Determining physical mechanisms of gene expression regulation from single cell gene expression data. *PLoS computational biology*, 12(8):e1005072, 2016.

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

46. Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome biology*, 14(1):R7, 2013.

47. Long Cai, Nir Friedman, and X Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358, 2006.

48. Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular cell*, 58(1):147–156, 2015.

49. Paul J Choi, Long Cai, Kirsten Frieda, and X Sunney Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science (New York, N.Y.)*, 322(5900):442–6, 2008.

50. Jiajun Zhang, Qing Nie, and Tianshou Zhou. Revealing dynamic mechanisms of cell fate decisions from single-cell transcriptomic data. *Frontiers in Genetics*, 10(1280), 2019.

51. William Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, Inc., New York, London, Sydney.

52. Guilherme da Costa Pereira Innocentini, Michael Forger, Alexandre Ferreira Ramos, Ovidiu Radulescu, and José Eduardo Martinho Hornos. Multimodality and flexibility of stochastic gene expression. *Bulletin of mathematical biology*, 75(12):2600–2630, 2013.

53. Adrienne W Kemp and CD Kemp. A family of discrete distributions defined via their factorial moments. *Communications in Statistics-Theory and Methods*, 3(12):1187–1196, 1974.

54. John Gurland. A generalized class of contagious distributions. *Biometrics*, 14(2):229–249, 1958.

55. Ram C Tripathi and John Gurland. Some aspects of the Kemp families of distributions. *Communications in statistics-theory and methods*, 8(9):855–869, 1979.

56. Norman Lloyd Johnson and Samuel Kotz. *Urn models and their application; an approach to modern discrete probability theory*. New York, NY (USA) Wiley, 1977.

57. George Pólya. *Mathematics and plausible reasoning: Patterns of Plausible Inference*, volume 2. Princeton University Press, 1954.

58. Hans Freudenthal. Models in applied probability. In *The concept and the role of the model in mathematics and natural and social sciences*, pages 78–88. Springer, 1961.

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

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

61. Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

62. Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, 2006.

63. Paolo Visco, Rosalind J Allen, and Martin R Evans. Exact solution of a model DNA-inversion genetic switch with orientational control. *Physical review letters*, 101(11):118104, 2008.

64. Kevin R Sanft, Sheng Wu, Min Roh, Jin Fu, Rone Kwei Lim, and Linda R Petzold. StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics (Oxford, England)*, 27(17):2457–8, 2011.

65. Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

66. Johan Elf and Irmeli Barkefors. Single-molecule kinetics in living cells. *Annual review of biochemistry*, 88:635–659, 2019.

67. Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research*, 5, 2016.

68. Stanley Wasserman. Distinguishing between stochastic models of heterogeneity and contagion. *Journal of Mathematical Psychology*, 27(2):201–215, 1983.

69. John W Pearson, Sheehan Olver, and Mason A Porter. Numerical methods for the computation of the confluent and Gauss hypergeometric functions. *Numerical Algorithms*, 74(3):821–866, 2017.

70. Philip J Davis. Leonhard Euler's integral: A historical profile of the Gamma function. *The American Mathematical Monthly*, 66(10):849–869, 1959.