Uncovering the effect of RNA polymerase steric interactions on gene expression noise: Analytical distributions of nascent and mature RNA numbers

Juraj Szavits-Nossan ●* and Ramon Grima ●†

School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JH, United Kingdom

The telegraph model is the standard model of stochastic gene expression, which can be solved exactly to obtain the distribution of mature RNA numbers per cell. A modification of this model also leads to an analytical distribution of nascent RNA numbers. These solutions are routinely used for the analysis of single-cell data, including the inference of transcriptional parameters. However, these models neglect important mechanistic features of transcription elongation, such as the stochastic movement of RNA polymerases and their steric (excluded-volume) interactions. Here we construct a model of gene expression describing promoter switching between inactive and active states, binding of RNA polymerases in the active state, their stochastic movement including steric interactions along the gene, and their unbinding leading to a mature transcript that subsequently decays. We derive the steady-state distributions of the nascent and mature RNA numbers in two important limiting cases: constitutive expression and slow promoter switching. We show that RNA fluctuations are suppressed by steric interactions between RNA polymerases, and that this suppression can in some instances even lead to sub-Poissonian fluctuations; these effects are most pronounced for nascent RNA and less prominent for mature RNA, since the latter is not a direct sensor of transcription. We find a relationship between the parameters of our microscopic mechanistic model and those of the standard models that ensures excellent consistency in their prediction of the first and second RNA number moments over vast regions of parameter space, encompassing slow, intermediate, and rapid promoter switching, provided the RNA number distributions are Poissonian or super-Poissonian. Furthermore, we identify the limitations of inference from mature RNA data, specifically showing that it cannot differentiate between highly distinct RNA polymerase traffic patterns on a gene.

I. INTRODUCTION

The widespread availability of RNA data in single cells has spurred a large amount of theoretical work on gene expression over the past two decades [1–9]. Snapshot measurements of the transcript numbers over a population of cells reveal large cell-to-cell variability in bacteria [10], yeast [11], and mammalian cells [12]. This variability can be more directly appreciated by following transcription dynamics in a single cell, which demonstrates that transcription does not occur continuously, but rather at random times and in bursts [13]. A main focus of mathematical models of gene expression has been to understand the origin of this noise, and also how cells tolerate, control, and possibly exploit it from the perspective of biological function [14–17].

Stochastic models of gene expression are predominantly based on the (random) telegraph model [18], which describes switching between two promoter states (active and inactive),

synthesis of a mature transcript from the active state and its subsequent degradation. By assuming that the dynamics are Markovian, one can write a time-evolution equation for the joint probability distribution of the promoter state and the number of RNA molecules, which can be solved exactly in steady state and also in time [19]. It has become common to fit the steady-state solution of this model to distributions of the number of RNA per cell obtained from single molecule fluorescence *in situ* hybridization (smFISH) [11,20] or single-cell RNA sequencing (scRNA-seq) experiments [21,22]. Provided the RNA decay rate is estimated experimentally, this fitting leads to estimates of the transcriptional parameters (the rate of switching to the active state, the rate of switching to the off state and the synthesis rate) for any gene of interest.

A criticism of this fitting procedure is that mature RNA numbers are not a direct sensor of transcription, as they are affected by other processes downstream of transcription, such as splicing and nuclear export. To overcome this criticism, models were developed to predict the distributions of nascent RNA, i.e., RNA that is attached to transcribing RNA polymerases (RNAPs) moving along a gene during transcriptional elongation [23,24]. These models are non-Markovian because nascent RNA does not decay via a first-order reaction (as mature RNA), but rather its removal is assumed to occur after a fixed deterministic time equal to the total time of elongation and termination. Fitting the steady-state solution to single-cell nascent RNA data also leads to estimation of transcriptional

---

*juraj.szavits.nossan@ed.ac.uk

†ramon.grima@ed.ac.uk

parameters, which may differ significantly from those estimated using mature mRNA data [25]. This makes a strong case for developing more accurate mathematical models of nascent RNA fluctuations.

A disadvantage of current stochastic models of nascent RNA dynamics with explicit analytical solutions [23,24,26–29] is that they implicitly assume RNAPs move deterministically along the gene, hence they do not interact with each other. This is clearly an over-simplification, since the frequency of such interactions should at least be significant for highly transcribed genes [30]. A significant number of computational studies have been undertaken to model the fine-scale dynamics of transcriptional elongation, such as volume-excluded (steric) interactions between RNAPs, ubiquitous pausing and backtracking, interaction of RNAPs with nucleosomes, and the mechanochemical cycle of RNAP movement [30–46]. However, none of these studies have analytically derived the steady-state single-cell distributions of nascent and mature RNA numbers.

In this paper, we take a first step toward achieving this goal by deriving the nascent and mature RNA number distributions for a stochastic model of gene expression that explicitly includes volume-excluded interactions between RNAPs. Specifically, the movement of RNAPs in this model is the same as that of the well known totally asymmetric simple exclusion process (TASEP) with uniform hopping rates [47,48]. While our model does not have all the detailed fine-scale biological detail of some of the aforementioned simulation-based studies, it provides a minimalist description of RNAP traffic on a gene that makes the transcript number distributions analytically tractable.

The paper is structured as follows. The constitutive model with RNAP volume exclusion is studied in Sec. II. The telegraph model with RNAP volume exclusion is studied in Sec. III. The results from the two models are summarized and discussed in Sec. IV.

II. CONSTITUTIVE MODEL WITH RNAP VOLUME EXCLUSION

We first consider a model of constitutive gene expression, i.e., a gene that is constantly expressed and is not subject to regulation. The gene body is coarse-grained into L segments of length $\ell \approx 35$ nucleotides, which is the footprint size of RNAP [49–51]. Hence, each segment $i = 1, \ldots, L$ is either empty or is occupied by an RNAP. Transcription starts by the binding of an RNAP to a promoter, which is followed by a sequence of steps, including promoter opening, promoter escape, promoter-proximal pausing, and pause release, after which the elongation of the nascent transcript starts. We lump all these initiation phase processes together into a single-step reaction with rate α. We note that the elongation phase can only start, i.e., RNAP can only be released into the first segment, provided this segment is empty. Subsequently, the RNAP moves forward stochastically one segment at a time, provided the neighboring segment in front of the RNAP is empty. We denote by ω the hopping rate at which the RNAP moves from a segment to the next one. This rate has units of inverse time, and it is related to the measured elongation rate in the absence of volume exclusion (in nucleotides per second)
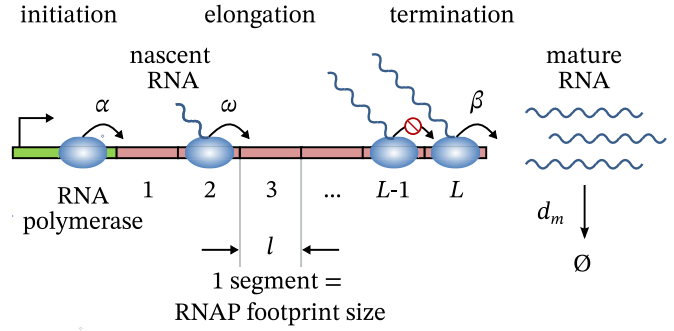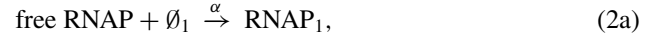

FIG. 1. Illustration of the constitutive model of transcription with RNAP volume exclusion. The gene is divided into L segments, each having a size equal to the footprint of a single RNAP. Transcription initiation is modelled by a single-step reaction that occurs at rate α, provided the first segment is empty. RNAPs move along the gene with rate ω, provided the segment in front is empty. As elongation progresses, the nascent RNA tail attached to the RNAP grows. Termination with rate β leads to the unbinding of the RNAP-nascent RNA complex (transcription elongation complex) from the gene and its dissociation into the free RNAP and the free RNA (mature RNA) that is subsequently degraded with rate d_M.

by the formula

$$\omega = \frac{\text{elongation rate [nt/s]}}{\text{RNAP footprint size [nt]}}. \quad (1)$$

Transcription termination occurs at the rate β from the last segment, after which the RNAP is removed from the lattice and a mature RNA is produced. The mature RNA degrades stochastically with rate d_M. The model is schematically presented in Fig. 1. The full model can be summarized as

$$\text{free RNAP} + \emptyset_1 \xrightarrow{\alpha} \text{RNAP}_1, \quad (2a)$$

$$\text{RNAP}_i + \emptyset_{i+1} \xrightarrow{\omega} \emptyset_i + \text{RNAP}_{i+1}, \quad i \in [1, L-1] \quad (2b)$$

$$\text{RNAP}_L \xrightarrow{\beta} \emptyset_L + \text{free RNAP} + \text{RNA}, \quad (2c)$$

$$\text{RNA} \xrightarrow{d_M} \emptyset, \quad (2d)$$

where free RNAP denotes an RNAP that is not actively engaged in transcription, RNAP_i denotes an RNAP positioned at the segment i, $\emptyset_i$ denotes that the segment i is empty, and RNA denotes the mature RNA. Since the number of free RNAPs is large, we can approximate the second-order reaction in Eq. (2a) by a quasi first-order reaction $\emptyset_1 \rightarrow \text{RNAP}_1$, where the rate α is proportional to the number of free RNAPs.

We note that a subset of the model describing the RNAP dynamics [Eqs. (2a)–(2c)] has been studied in various contexts, some of which are nonbiological, and is known as the totally asymmetric simple exclusion process (TASEP) [52]. In the TASEP, there is no explicit tracking of mature RNA. Hence, to clearly distinguish our model from the TASEP and from other models that we consider later on, we refer to the model defined by Eqs. (2a)–(2d) as the constitutive model with RNAP volume exclusion (vCM for short). We note that since our model describes movement on the coarse length scale of an RNAP, not at the single nucleotide level, it does not have an explicit description of processes occurring over few base pairs such as backtracking or pausing; to some extent,

these processes can be captured by a suitable renormalization of the hopping rate (see Sec. IV for further discussion).

To track RNAPs along the gene, we introduce a variable τ_i such that $\tau_i = 0$ if the segment i is empty, and $\tau_i = 1$ if the segment i is occupied by an RNAP. The number of RNAPs on the gene, denoted by n, is given by

$$n = \sum_{i=1}^{L} \tau_i. \tag{3}$$

We denote the number of mature RNA by m. We refer to $C = \{\tau_1, \ldots, \tau_L\}$ as a configuration of RNAPs along the gene. The joint probability to find RNAPs in a configuration C and m copies of mature RNA at time t is denoted by $P(C, m, t)$. We are interested in the steady state, in which case we drop the time dependence and consider only the joint probability $P(C, m)$. The marginal distribution of RNAPs along the gene is obtained by summing $P(C, m)$ over all m, $P(C) = \sum_{m=0}^{\infty} P(C, m)$. We define a local density of RNAPs at a segment i as

$$\rho_i = \langle \tau_i \rangle = \sum_C P(C)\delta_{\tau_i, 1}, \tag{4}$$

where $\delta_{i,j}$ is the Kronecker delta. The transcription rate at which new mature RNA is produced is defined as the local density ρ_L, multiplied by the termination rate β, $k_{syn} = \rho_L \beta$. In the steady state, the transcription rate is equal to the current of RNAPs along the gene,

$$k_{syn} = J = \alpha(1 - \langle \tau_1 \rangle) = \omega\langle \tau_i(1 - \tau_{i+1}) \rangle = \beta\rho_L. \tag{5}$$

The probability distributions of the number of RNAPs on the gene and the number of mature RNA are defined as, respectively,

$$P_N(n) = \sum_C P(C)\delta_{\sum_i \tau_i, n}, \tag{6a}$$
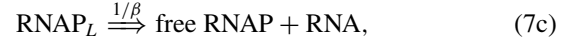
$$P_M(m) = \sum_C P(C, m). \tag{6b}$$

We note that the distribution of the number of RNAPs on the gene is the same as the distribution of the number of nascent RNA since to each gene-bound RNAP, a nascent RNA tail is attached. Analytical results are known for $P(C)$, J and ρ_i from the exact solution of the TASEP [48], and are summarized in Sec. II B. Results for the distributions $P_N(n)$ and $P_M(m)$ are new, and are derived in Secs. II C and II D, respectively.

Before we present the results for $P_N(n)$ and $P_M(m)$, we consider a much simpler model in which elongation and termination processes are assumed to be deterministic, and in which there are no RNAP volume-exclusion effects. This model is exactly solvable and will serve as a useful benchmark for understanding the effect of RNAP collisions on the distributions of nascent and mature RNA numbers.

A. Delay constitutive model

This model approximates elongation and termination by a series of delay reactions (indicated by double arrows) that take a fixed amount of time to finish. The model is effectively defined by the reaction scheme

$$\text{free RNAP} \xrightarrow{\alpha} \text{RNAP}_1, \tag{7a}$$

$$\text{RNAP}_i \xRightarrow{1/\omega} \text{RNAP}_{i+1} \text{ for } i \in [1, L-1], \tag{7b}$$

$$\text{RNAP}_L \xRightarrow{1/\beta} \text{free RNAP} + \text{RNA}, \tag{7c}$$

$$\text{RNA} \xrightarrow{d_M} \emptyset. \tag{7d}$$

Here, $1/\omega$ is the fixed time it takes an RNAP to move across one segment, and $1/\beta$ is the fixed time it takes an RNAP to terminate from the last segment. As before, we assume that the number of free RNAPs is large so that it can be absorbed in the initiation rate α. We refer to this model as the delay constitutive model.

The total time of elongation and termination is fixed and equal to

$$T_{el} = \frac{L-1}{\omega} + \frac{1}{\beta}. \tag{8}$$

Since elongation and termination are deterministic, the transcription rate k_{syn} equals the rate of initiation α,

$$k_{syn} = \alpha. \tag{9}$$

The local density of RNAPs on the gene can be computed from that fact that the number of initiation events in a given time interval t is a Poisson random variable with parameter αt. The time an RNAP stays in the segment i is equal to $1/\omega$ for $i = 1, \ldots, L-1$ and $1/\beta$ for $i = L$, hence the local density ρ_i is equal to

$$\rho_i = \begin{cases} \frac{\alpha}{\omega} & i = 1, \ldots, L-1, \\ \frac{\alpha}{\beta}, & i = L. \end{cases} \tag{10}$$

Since the model ignores excluded volume interactions between RNAPs, the local density ρ_i becomes larger than 1 if $\alpha > \omega$ or $\alpha > \beta$, which is not physical. This model is therefore justified only for $\alpha < \omega$ and $\alpha < \beta$.

The number of RNAPs on the gene n is equal to the number of initiations in the time interval T_{el}, hence

$$P_N(n) = \frac{(\alpha T_{el})^n}{n!} e^{-\alpha T_{el}}. \tag{11}$$

The mean and the variance of this distribution are

$$\mu_N = \alpha T_{el}, \quad \sigma_N^2 = \alpha T_{el}, \tag{12}$$

and the Fano factor FF_N (the variance divided by the mean) is equal to 1. Combining Eqs. (9) and (12) gives

$$\mu_N = k_{syn} T_{el}. \tag{13}$$

In queuing theory, this relationship is known as Little's law [53,54], which states that the long-time average number of customers in a queue is equal to the arrival rate k_{syn} multiplied by the long-time average of the time spent in the queue, T_{el}.

Due to the deterministic nature of elongation and termination, the mature RNA effectively follows a birth-death process with the birth rate equal to α and the death rate equal to d_M,

$$\emptyset \xrightarrow{\alpha} \text{RNA} \xrightarrow{d_M} \emptyset. \tag{14}$$

TABLE I. Different regimes of the steady-state TASEP.

Regime	Parameter range	Local RNAP density ρ_i	Transcription rate J
Initiation-limited regime (IL)	$\alpha < \omega/2$ and $\beta > \alpha$	α/ω	$\alpha(1 - \alpha/\omega)$
Termination-limited regime (TL)	$\beta < \omega/2$ and $\alpha > \beta$	$1 - \beta/\omega$	$\beta(1 - \beta/\omega)$
Coexistence line (IL/TL)	$\alpha = \beta < \omega/2$	Linearly increasing	$\alpha(1 - \alpha/\omega)$
Elongation-limited regime (EL)	$\alpha, \beta > \omega/2$	$1/2$	$\omega/4$

The steady-state probability distribution $P_M(m)$ is the Poisson distribution with the parameter α/d_M,

$$P_M(m) = \frac{(\alpha/d_M)^m}{m!} e^{-\alpha/d_M}. \tag{15}$$

The mean and the variance of this distribution are

$$\mu_M = \frac{\alpha}{d_M}, \quad \sigma_M^2 = \frac{\alpha}{d_M}, \tag{16}$$

and the Fano factor FF_M is equal to 1.

B. Summary of known results for the TASEP in the steady state

The steady-state probability $P(C)$ is known exactly and can be written in the following matrix-product form [48],

$$P(\tau_i, \dots, \tau_L) = \frac{1}{Z_L} \langle W | \prod_{i=1}^{L} [\tau_i D + (1 - \tau_i)E] | V \rangle, \tag{17}$$

where D and E are infinite-dimensional matrices that satisfy

$$DE = D + E, \tag{18}$$

$\langle W |$ and $| V \rangle$ are infinite-dimensional vectors that satisfy

$$\langle W | E = \frac{\omega}{\alpha} \langle W |, \quad D | V \rangle = \frac{\omega}{\beta} | V \rangle, \tag{19}$$

and Z_L is the normalization,

$$Z_L = \langle W | (D + E)^L | V \rangle. \tag{20}$$

We note that the matrices D and E are not uniquely defined [48]; however, the relations above are sufficient for all practical calculations.

An explicit formula for Z_L in terms of α, β, and ω is

$$Z_L = \sum_{p=0}^{L} B_{L,p} \sum_{q=0}^{p} \left(\frac{\omega}{\alpha}\right)^q \left(\frac{\omega}{\beta}\right)^{p-q}, \tag{21}$$

where

$$B_{k,p} = \begin{cases} \frac{p}{2k-p}\binom{2k-p}{k}, & p = 1, \dots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

The steady-state RNAP current J, which equals the transcription rate k_{syn}, reads

$$k_{\text{syn}} = J = \frac{\beta \langle W | (D + E)^{L-1} D | V \rangle}{Z_L} = \frac{\omega Z_{L-1}}{Z_L}, \tag{23}$$

and the local RNAP density ρ_i is given by

$$\rho_i = \frac{1}{Z_L} \langle W | (D + E)^{i-1} D (D + E)^{L-i} | V \rangle$$
$$= \sum_{p=1}^{L-i} B_{p,1} \frac{Z_{L-p}}{Z_L} + \frac{Z_{i-1}}{Z_L} \sum_{p=1}^{L-i} \frac{B_{L-i,p}}{\beta^{p+1}}. \tag{24}$$

In the limit $L \to \infty$, J simplifies to

$$J = \begin{cases} \alpha\left(1 - \frac{\alpha}{\omega}\right), & \alpha < \frac{\omega}{2}, \ \beta > \alpha, \\ \beta\left(1 - \frac{\beta}{\omega}\right), & \beta < \frac{\omega}{2}, \ \alpha > \beta, \\ \frac{\omega}{4}, & \alpha, \beta > \frac{\omega}{2}. \end{cases} \tag{25}$$

Away from the boundaries, the local density ρ_i is approximately constant and equal to

$$\rho = \begin{cases} \frac{\alpha}{\omega}, & \alpha < \frac{\omega}{2}, \ \beta > \alpha, \\ 1 - \frac{\beta}{\omega}, & \beta < \frac{\omega}{2}, \ \alpha > \beta, \\ \frac{1}{2}, & \alpha, \beta > \frac{\omega}{2}. \end{cases} \tag{26}$$

The exception is for $\alpha = \beta < \omega/2$, for which the local density increases linearly from α/ω at the left boundary to $1 - \beta/\omega$ at the right boundary.

Different regimes of the TASEP depending on the initiation and termination rates with respect to the hopping rate ω are summarized in Table I. In the low-density or initiation-limited (IL) regime ($\alpha < \omega/2$ and $\beta > \alpha$), $\rho = \alpha/\omega$ and the transcription rate is controlled by the initiation rate, $k_{\text{syn}} = \alpha(1 - \alpha/\omega)$. In this regime, transcription is rate-limited by initiation, which is usually stated as the most likely scenario under physiological conditions. In the high-density or termination-limited (TL) regime ($\beta < \omega/2$ and $\alpha > \beta$), $\rho = 1 - \beta/\omega$, and the transcription rate is controlled by the termination rate, $k_{\text{syn}} = \beta(1 - \beta/\omega)$. In this regime, there is a long queue of RNAPs spanning from the termination site toward the beginning of the gene. We are unaware of such scenario being observed *in vivo*, though. On the coexistence (IL/TL) line between these two phases ($\alpha = \beta < \omega/2$), the local density increases linearly along the gene and the transcription rate is equal to $k_{\text{syn}} = \alpha(1 - \alpha/\omega)$. This is a very specific regime that occurs only when the initiation rate is equal to the termination rate. Finally, in the maximum-current or elongation-limited (EL) regime ($\alpha > \omega/2$ and $\beta > \omega/2$), $\rho = 1/2$, and the transcription rate depends only on the hopping rate ω, $k_{\text{syn}} = \omega/4$. In this regime, transcription dynamics is fully controlled by the elongation rate ω. Any further increase in the initiation and termination rates has no effect on the dynamics.

In comparison to these results, the delay constitutive model predicts $\rho = \alpha/\omega$ and $k_{\text{syn}} = \alpha$, which makes sense only if $\alpha < \omega$.

C. Probability distribution of the number of nascent RNA

Our strategy is to compute the probability generating function G_N defined as

$$G_N(z) = \sum_{n=0}^{L} P_N(n) z^n. \tag{27}$$

Using Eqs. (6a) and (17), it is straightforward to show that

$$G_N(z) = \frac{1}{Z_L} \langle W | (zD + E)^L | V \rangle. \qquad (28)$$

The calculation of $G_N(z)$ is presented in Appendix A, and the final result is

$$G_N(z) = \frac{1}{Z_L} \sum_{k=1}^{L} \frac{k}{L} u_{L-k,L}(z) v_k(z), \qquad (29)$$

where $u_{p,n}(z)$ and $v_k(z)$ are given by

$$u_{p,n}(z) = \sum_{m=1}^{n} \binom{n}{m} \binom{p+m-1}{p} (z-1)^{n-m}, \qquad (30a)$$

$$v_0(z) = 1, \qquad (30b)$$

$$v_k(z) = \sum_{i=0}^{k} \left(\frac{\alpha z - \alpha + 1}{\alpha z} \right)^i \frac{1}{\beta^{k-i}}$$
$$- \frac{z-1}{z} \sum_{i=0}^{k-1} \left(\frac{\alpha z - \alpha + 1}{\alpha z} \right)^i \frac{1}{\beta^{k-1-i}}, \ k \geqslant 1. \quad (30c)$$

From here, we get $P_N(n)$ by expanding $G_N(z)$ in z and collecting the terms containing z^n,

$$P_N(n) = \frac{1}{n!} \frac{d^n}{dz^n} G_N(z) \Big|_{z=0}. \qquad (31)$$

The mean μ_N is equal to the spatial average of the local density ρ_i given by Eq. (24),

$$\mu_N = \frac{1}{L} \sum_{i=1}^{L} \rho_i, \qquad (32)$$

and the variance σ_N^2 can be computed from

$$\sigma_N^2 = \frac{d^2 G_N(z)}{dz^2} \Big|_{z=1} + \mu_N (1 - \mu_N). \qquad (33)$$

The Fano factor of the nascent RNA number is given by

$$\mathrm{FF}_N = 1 - \mu_N + \frac{1}{\mu_N} \frac{d^2 G_N(z)}{dz^2} \Big|_{z=1}. \qquad (34)$$

We were not able to find a simple expression for the variance and the Fano factor FF_N for arbitrary α/ω and β/ω. For the special case $\alpha/\omega = \beta/\omega = 1$, $\sigma_N^2 = L(L+2)/(8L+4)$ was computed previously in Ref. [55]. In the limit in which $L \to \infty$, this result yields $\mathrm{FF}_N = 1/4$. We have checked numerically for $L = 100$ that $\mathrm{FF}_N \approx 1/4$ for other values of α and β in the elongation-limited regime.

While the distribution $P_N(n)$ does not seem to be related to any known distribution, we know from the exact solution of the TASEP that for $\alpha + \beta = \omega$, $P(\tau_1, \ldots, \tau_L)$ simplifies to a product of Bernoulli distributions [56],

$$P(\tau_1, \ldots, \tau_L) = \prod_{i=1}^{L} [\tau_i \rho + (1 - \tau_i)(1 - \rho)], \qquad (35)$$

where $\rho = \alpha/\omega = 1 - \beta/\omega$. For this distribution, the local density ρ_i is equal to ρ at any segment, i.e., the density

of RNAP is uniform along the gene body. By inserting this distribution into Eq. (6a), we get a binomial distribution

$$P_N(n) = \binom{L}{n} \rho^n (1 - \rho)^{L-n}. \qquad (36)$$

The mean and the variance of this binomial distribution are $L\rho$ and $L\rho(1 - \rho)$, respectively, and the Fano factor is given by

$$\mathrm{FF}_N = 1 - \rho. \qquad (37)$$

This Fano factor is always less than 1, and becomes vanishingly small in the limit $\rho \to 1$. The skewness of the binomial distribution in Eq. (36) is given by $(1 - 2\rho)/\sqrt{\rho(1-\rho)L}$, hence the distribution is right-skewed for $\rho < 1/2$, symmetrical for $\rho = 1/2$, and left-skewed for $\rho > 1/2$. In the limit in which $L \to \infty$ and $\rho \to 0$ such that $\lambda \equiv L\rho$ is fixed, the binomial distribution becomes the Poisson distribution with the rate parameter λ. This limit corresponds to the delay constitutive model.

To check how well the binomial distribution [Eq. (36) with ρ given by Eq. (26)] approximates the exact one [Eq. (31)] across the parameter space, i.e., without imposing the condition $\alpha + \beta = \omega$, we computed the Hellinger distance (which varies between 0 and 1) between these two distributions for 2500 combinations of α/ω and β/ω, equally spaced between 0 and 1 [Fig. 2(a)]. As expected, the Hellinger distance is small in the initiation-limited and termination-limited regimes away from the regime boundaries [Figs. 2(b) and 2(c)], and is exactly zero on the special line $\alpha + \beta = \omega$. The largest difference between the distributions is observed at the coexistence line between the initiation-limited and termination-limited regimes ($\alpha = \beta < \omega/2$) [Fig. 2(d)], where the local density has a linear profile [48]. Significant differences are also observed at the boundaries between the initiation-limited and the elongation-limited regimes, between the termination-limited and the elongation-limited regimes, and in the entire elongation-limited regime [Fig. 2(e)]. These discrepancies are due to the algebraic decay of the local density toward its value of $1/2$ in the middle of the gene [48]. In the elongation-limited regime, the Fano factor is approximately equal to $1/4$, whereas the binomial distribution predicts the value of $1/2$. Simulations in Fig. 2 also confirm the theoretical transition from the left-skewed to the right-skewed nascent RNA distributions as ρ crosses the threshold of $1/2$ [Figs. 2(b) and 2(c)]—similar transitions have been reported in studies of crowding-induced phenomena in other chemical reaction systems [57]. The simulations in Fig. 2 and in the rest of the paper were performed using the Gillespie algorithm [58].

From Eq. (12), it follows that the mean number of nascent RNA predicted by the delay constitutive model is $\alpha T_{\mathrm{el}} \approx L\alpha/\omega$. Comparing this result to that from the constitutive model with RNAP exclusion, ρL, we conclude that the two match only in the initiation-limited regime for small α, such that $J \approx \alpha$ and $\rho = \alpha/\omega$. It is, however, possible to extend the constitutive delay model such that its mean nascent RNA number matches that of the constitutive model with RNAP beyond this regime. This can be achieved by replacing the initiation rate α in the delay model with an effective initiation rate given by $\alpha_{\mathrm{eff}} = J$, where J is the transcription rate in the
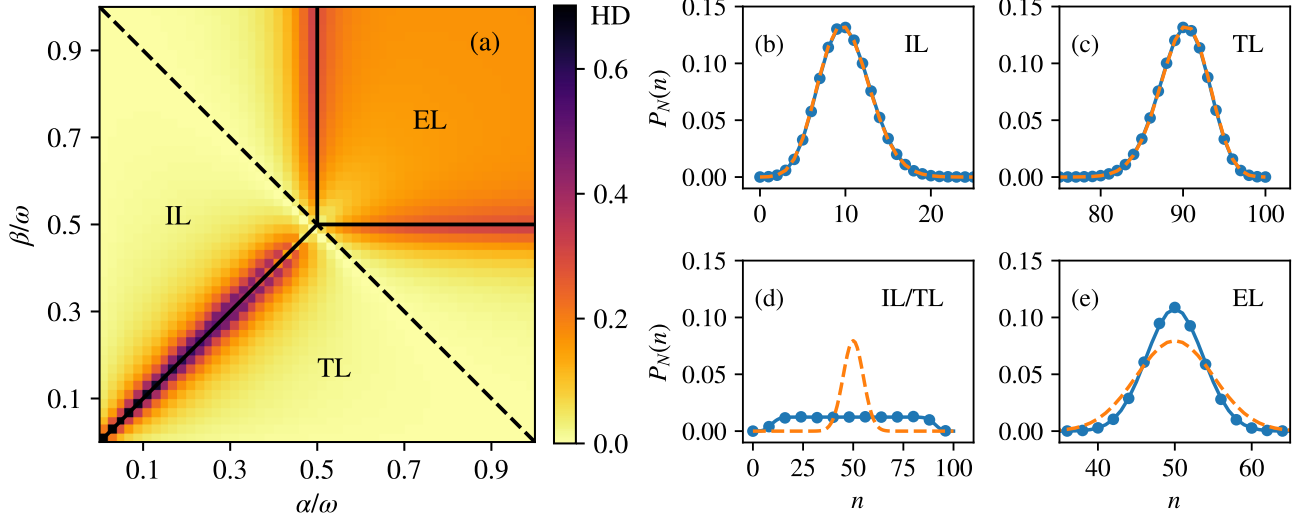
FIG. 2. Accuracy of the binomial approximation for the steady-state nascent RNA number distribution in the constitutive model with RNAP volume exclusion. (a) Heat map of the Hellinger distance (HD) between the exact probability distribution $P_N(n)$ in Eq. (31) and the binomial distribution in Eq. (36). Solid black lines are regime boundaries and the dashed black line is $\alpha + \beta = \omega$, in which case the Hellinger distance is exactly zero. (b–e) Compare the nascent RNA number distribution obtained using stochastic simulations (blue points), the exact distribution given by Eq. (31) (solid blue lines) and the binomial distribution approximation given by Eq. (36) (dashed orange lines). The number of segments is $L = 100$. (b) Initiation-limited regime (IL, $\alpha/\omega = 0.1$, $\beta/\omega = 0.7$). (c) Termination-limited regime (TL, $\alpha/\omega = 0.7$, $\beta/\omega = 0.1$). (d) Coexistence line (IL/TL, $\alpha/\omega = \beta/\omega = 0.1$). (e) Elongation-limited regime (EL, $\alpha/\omega = 0.7$, $\beta/\omega = 0.7$).

constitutive model with RNAP volume exclusion [Eq. (23)]. Similarly, we define an effective hopping rate at segment i as $\omega_{\text{eff},i} = J/\rho_i$, where ρ_i is the local RNAP density in the constitutive model with RNAP volume exclusion [Eq. (24)]. Finally, to satisfy Little's law, we define the effective elongation and termination time $T_{\text{el,eff}} = \sum_{i=1}^{L} 1/\omega_{\text{eff},i}$. We refer to the constitutive delay model with these parameters as the effective delay constitutive model (edCM for short). Assuming $\rho_i \approx \rho$, which is true for large L and outside the coexistence line where ρ_i is linearly increasing, we get $\omega_{\text{eff},i} \approx \omega(1 - \rho)$ and $T_{\text{el,eff}} \approx L/\omega_{\text{eff}}$. We note that the effective hopping rate ω_{eff} is the "true" rate of nascent RNA elongation taking into account slowing down of RNAP due to other RNAPs on the gene, whereas ω is the "bare" hopping rate in the absence of other RNAPs.

D. Probability distribution of the number of mature RNA

We cannot compute $P_M(m)$ directly from Eq. (6b), because the joint distribution $P(C, m)$ is unknown. Instead, we find an approximate expression for $P_M(m)$ by replacing the process of RNA production and degradation with the following queuing process,

$$\emptyset \xrightarrow{f_{\text{ter}}(t)} M \xrightarrow{d_M} \emptyset. \tag{38}$$

Here, $f_{\text{ter}}(t)$ is the probability density function (pdf) of the waiting time between two successive termination events, and d_M is the degradation rate of mature RNA. This process is known as a $G/M/\infty$ queue in Kendall's notation [59], where G stands for general interarrival distribution [mature RNAs arrive at time intervals distributed according to $f_{\text{ter}}(t)$], M stands for Markovian service process (service times are exponentially distributed with rate d_M), and the number of servers

is infinite (the RNA degradation machinery is assumed to be abundant).

We note that the queuing process described by Eq. (38) is not an exact representation of the original model. In the queuing process, the waiting times between successive termination events are mutually independent, which is not true in the original model. Therefore, we will refer to Eq. (38) as the *renewal approximation* of the original process, because the production of mature RNA in this approximation constitutes a renewal process (a generalization of the Poisson process to an arbitrary interarrival time distribution).

The advantage of this approximation is that the steady-state distribution of the number of customers (the number of mature RNA in our case) in a $G/M/\infty$ queue can be computed analytically for any $f(t)$ whose mean interarrival time is finite [60]. Here we write the final result for $P_M(m)$ and refer the reader to the original paper for the derivation. We denote by $\tilde{f}(s)$ the Laplace transform of $f(t)$, and by μ the mean interarrival time,

$$\tilde{f}(s) = \int_0^\infty dt\, f(t)e^{-st}, \quad \mu = \int_0^\infty dt\, t f(t) < \infty. \tag{39}$$

Next, we define a coefficient C_i as

$$C_0 = 1, \quad C_i = \prod_{j=1}^{i} \frac{\tilde{f}(jd_M)}{1 - \tilde{f}(jd_M)}, \quad i = 1, 2, 3 \ldots. \tag{40}$$

The steady-state distribution of the mature RNA number m is then given by

$$P_M(0) = 1 - \sum_{k=1}^{\infty} (-1)^{k-1} \frac{C_{k-1}}{\mu d_M k}, \tag{41a}$$

$$P_M(m) = \sum_{k=m}^{\infty} (-1)^{k-m} \binom{k}{m} \frac{C_{k-1}}{\mu d_M k}, \quad m \geqslant 1. \tag{41b}$$

Hence, to compute $P_M(m)$ we need to compute the pdf $f_{\text{ter}}(t)$ of the waiting time between two successive termination events.

To this end, we denote by $\mathcal{G}$ the set of all configurations $C = \{\tau_1, \dots, \tau_L\}$ in which the last segment is occupied by an RNAP, $\mathcal{G} = \{C \mid \tau_L(C) = 1\}$. For a configuration $C \in \mathcal{G}$, we define the gap size $g(C)$ as the number of consecutive empty segments in front of the last segment. For example, $g(C) = 0$ if the closest RNAP trailing behind is at the segment $L-1$, and $g(C) = L-1$ if there are no trailing RNAPs. Next, we denote by $P_{\text{gap}}(k, L)$ the probability that a configuration $C \in \mathcal{G}$ has a gap of size k,

$$P_{\text{gap}}(k; L) = \frac{\sum_{C \in A} P(C)\delta_{g(C),k}}{\sum_{C \in A} P(C)}, \qquad (42)$$

where $\delta_{i,j}$ is the Kronecker delta. Note that the denominator is equal to the local density ρ_L.

Let us now assume that an RNAP terminated transcription at time $t = 0$, and that the gap size right before that was equal to k. We denote by $r_k(t)$ the pdf of the waiting time t until the next termination event. The pdf $f_{\text{ter}}(t)$ is then equal to $P_{\text{gap}}(k; L)$ multiplied by $r_k(t)$ and summed over all $k = 0, \dots, L-1$,

$$f_{\text{ter}}(t) = \sum_{k=0}^{L-1} P_{\text{gap}}(k; L) r_k(t). \qquad (43)$$

For $k < L - 1$, the time between two successive termination events is equal to the time it takes an RNAP at the segment $L - k - 1$ to move $k + 1$ segments and terminate from the last segment. Hence, the pdf $r_k(t)$ is a convolution of the Erlang distribution with shape $k + 1$ and rate ω, and the exponential distribution with rate β, which gives

$$r_k(t) = \beta e^{-\beta t}\left(\frac{\omega}{\omega - \beta}\right)^{k+1} \frac{\gamma(k+1, (\omega - \beta)t)}{\Gamma(k+1)}, \qquad (44)$$

where $\gamma(n, x)$ is a lower incomplete Gamma function. For $k = L - 1$, the time between two successive termination events is equal to the time it takes a free RNAP to initiate transcription, move $L - 1$ segments, and terminate from the last segment. Hence, the pdf $r_{L-1}(t)$ is given by a convolution of $r_{L-2}(t)$ and the exponential distribution with the rate parameter α, which gives

$$r_{L-1}(t) = \frac{\alpha\beta}{\alpha - \beta} e^{-\beta t}\left(\frac{\omega}{\omega - \beta}\right)^{L-1} \frac{\gamma(L-1, (\omega - \beta)t)}{\Gamma(L-1)}$$
$$- \frac{\alpha\beta}{\alpha - \beta} e^{-\alpha t}\left(\frac{\omega}{\omega - \alpha}\right)^{L-1} \frac{\gamma(L-1, (\omega - \alpha)t)}{\Gamma(L-1)}. \qquad (45)$$

The final expression for $f_{\text{ter}}(t)$ is complicated, so we consider the asymptotic limit of $L \to \infty$, in which $f_{\text{ter}}(t)$ becomes

$$f_{\text{ter,as}}(t) = \sum_{k=0}^{\infty} P_{\text{gap}}(k) r_k(t), \qquad (46)$$

where $P_{\text{gap}}(k) = \lim_{L \to \infty} P_{\text{gap}}(k; L)$. The gap-size distribution $P_{\text{gap}}(k, L)$ and the limiting distribution $P_{\text{gap}}(k)$ were previously derived in Ref. [61]. The limiting distribution $P_{\text{gap}}(k)$

reads

$$P_{\text{gap}}(k) = \frac{\alpha}{\beta}\left(1 - \frac{\alpha}{\omega}\right)^{k+1} \frac{\beta - \alpha}{\omega - 2\alpha} + \frac{\omega - \alpha}{\beta}\left(\frac{\alpha}{\omega}\right)^{k+1}$$
$$\times \left(1 - \frac{\beta - \alpha}{\omega - 2\alpha}\right), \ \alpha < \frac{\omega}{2}, \ \beta > \alpha, \qquad (47a)$$

$$P_{\text{gap}}(k) = \left(1 - \frac{\beta}{\omega}\right)\left(\frac{\beta}{\omega}\right)^k, \ \beta < \frac{\omega}{2}, \ \alpha \geqslant \beta, \qquad (47b)$$

$$P_{\text{gap}}(k) = \frac{\omega}{\beta 2^{k+1}}\left(\frac{1-k}{2} + \frac{\beta k}{\omega}\right), \ \alpha, \beta > \frac{\omega}{2}. \qquad (47c)$$

This limiting distribution is valid as long as the typical gap size is much less than L and L is large, in which case extending L to infinity does not have much impact on the gap-size distribution. If we now insert Eqs. (44) and (47) into Eq. (46), then we get a remarkably simple result for $f_{\text{ter,as}}(t)$,

$$f_{\text{ter,as}}(t) = \frac{\omega\nu(1-\nu)}{1 - 2\nu}(e^{-\omega\nu t} - e^{-\omega(1-\nu)t}), \qquad (48)$$

where ν is given by

$$\nu = \begin{cases} \frac{\alpha}{\omega}, & \alpha < \frac{\omega}{2}, \ \beta > \alpha, \\ 1 - \frac{\beta}{\omega}, & \beta < \frac{\omega}{2}, \ \alpha \geqslant \beta, \\ \frac{1}{2}, & \alpha, \beta > \frac{\omega}{2}. \end{cases} \qquad (49)$$

There is a subtle difference between ν in Eq. (49) and ρ in Eq. (26): $\nu = \rho$ everywhere except at the coexistence line $\alpha = \beta < 1/2$ for which $\nu = 1 - \beta/\omega$, whereas ρ is not properly defined on this line (the local density increases linearly from α/ω at the left boundary to $1 - \alpha/\beta$ at the right boundary). The expression similar to the one in Eq. (48) was previously derived for the discrete-time TASEP with parallel [32,62] and random-sequential [63] hopping.

Equation (48) describes a hypoexponential pdf of the sum of two exponentially distributed random variables with rates $\omega\nu$ and $\omega(1 - \nu)$. As expected, the mean waiting time μ_{ter} is equal to $1/J$, where J is the transcription rate given by Eq. (25). Note that $f_{\text{ter,as}}(t)$ is invariant to the exchange of $\nu \leftrightarrow 1 - \nu$. In the elongation-limited regime, $f_{\text{ter,as}}(t)$ becomes the gamma distribution with the shape parameter 2 and the rate parameter $\omega/2$.

We are now ready to compute the probability distribution of the number of mature RNA in the renewal approximation of our model. The Laplace transform of $f_{\text{ter,as}}(t)$ in Eq. (48) is given by

$$\tilde{f}_{\text{ter,as}}(s) = \frac{\omega^2\nu(1-\nu)}{(s + \omega\nu)[s + \omega(1-\nu)]}. \qquad (50)$$

Inserting this expression into Eq. (41) yields the following expression for the mature RNA distribution in the constitutive model,

$$P_M(m) = \frac{(ab)^m}{m!(a+b)_m} \, {}_0F_1(a + b + m, -ab), \qquad (51)$$

where $(x)_n = \Gamma(x + n)/\Gamma(x)$ is the Pochhammer symbol, ${}_0F_1$ is the confluent hypergeometric limit function, and a and b are given by

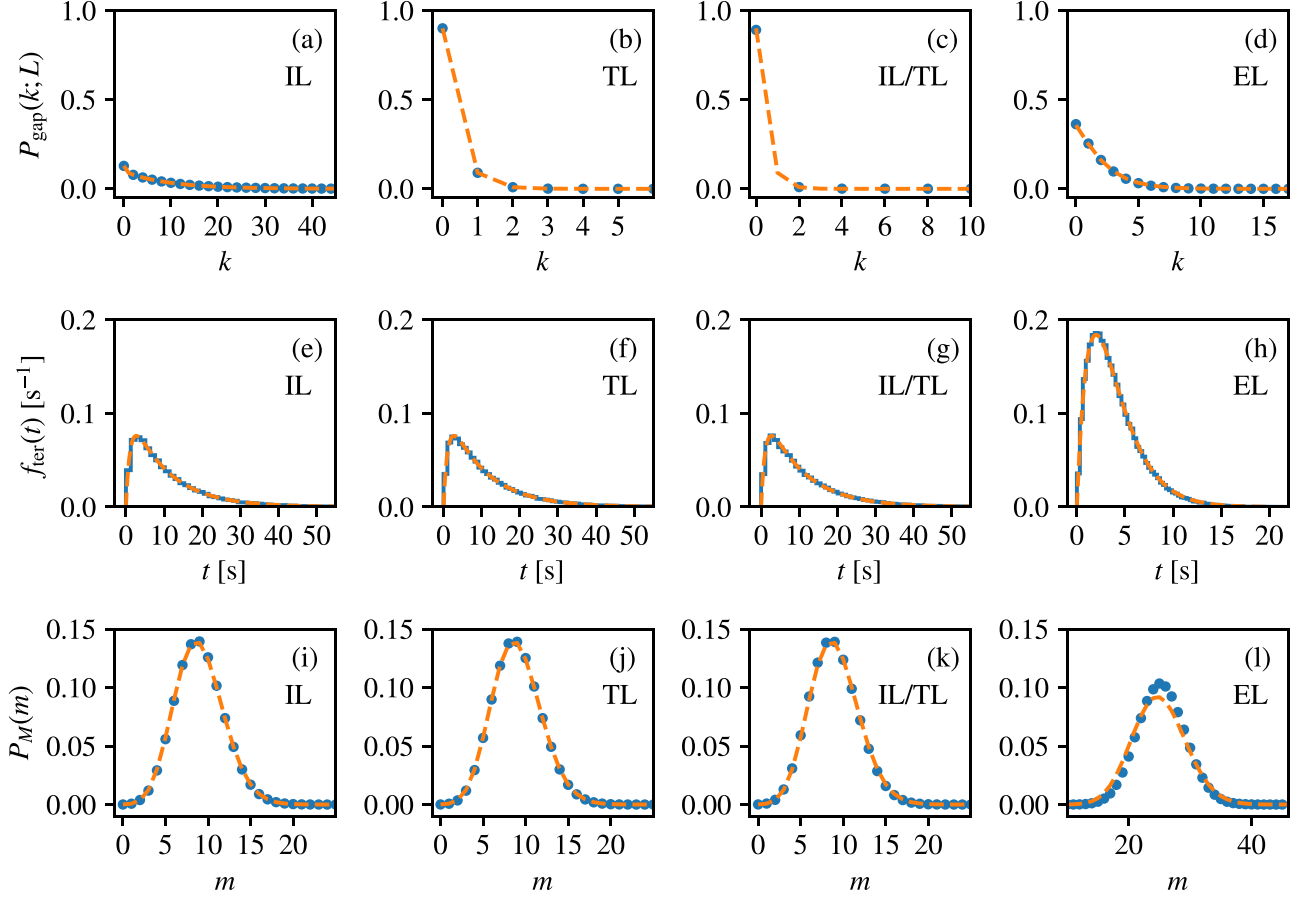$$a = \frac{\omega\nu}{d_M}, \quad b = \frac{\omega(1-\nu)}{d_M}. \qquad (52)$$

FIG. 3. Accuracy of the theoretical results for the gap-size distribution, the waiting-time distribution between successive termination events and the mature RNA number distribution in the constitutive model with RNAP volume exclusion. [(a)–(d)] Compare the gap-size distribution $P_{gap}(k; L)$ computed using stochastic simulations (blue points) and $P_{gap}(k)$ computed from the asymptotic theory in Eq. (47) (dashed orange line). [(e)–(h)] Compare the pdf $f_{ter}(t)$ of the waiting time between two successive termination events computed using stochastic simulations (blue solid line) and $f_{ter,as}(t)$ computed from the asymptotic theory in Eq. (48) (dashed orange line). [(i)–(l)] compare the mature RNA number distribution $P_M(m)$ computed using stochastic simulations (blue points) and the one computed from the renewal approximation in Eq. (51) (dashed orange line). Note that the renewal approximation theory is also based on the asymptotic theory for the waiting time between two successive termination events. The parameters are: $\alpha = 0.1$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, and $d_M = 0.01$ s^{-1} for the initiation-limited regime (IL, first column); $\alpha = 0.7$ s^{-1}, $\beta = 0.1$ s^{-1}, $\omega = 1.0$ s^{-1}, and $d_M = 0.01$ s^{-1} for the termination-limited regime (TL, second column); $\alpha = 0.1$ s^{-1}, $\beta = 0.1$ s^{-1}, $\omega = 1.0$ s^{-1}, and $d_M = 0.01$ s^{-1} for the coexistence line (IL/TL, third column); and $\alpha = 0.7$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, and $d_M = 0.01$ s^{-1} for the elongation-limited regime (EL, fourth column). The system size is $L = 100$.

The waiting-time distribution matching procedure that we used to obtain an expression for the approximate mature RNA distribution is similar in principle to the model reduction technique described in Ref. [28]. The probability generating function for $P_M(m)$ reads

$$G_M(z) = \sum_{m=0}^{\infty} P_M(m) z^m = {}_0F_1[a + b, ab(z - 1)]. \quad (53)$$

The mean and the variance of the RNA number m are given by

$$\mu_M = \frac{J}{d_M}, \quad \sigma_M^2 = \mu_M \left(1 - \frac{J}{\omega + d_M} \right). \quad (54)$$

The Fano factor of the mature RNA number in the renewal approximation is equal to

$$FF_M = 1 - \frac{J}{\omega + d_M} = 1 - \frac{\omega \rho (1 - \rho)}{\omega + d_M} \leqslant 1. \quad (55)$$

The Fano factor is always less than 1, which means that the RNA number distribution is sub-Poissonian. When ω is much larger than d_M, we get a simple expression for FF_M that depends only on ρ, $FF_M = 1 - \rho(1 - \rho)$. We note that $FF_M = 1$ in the (effective) delay constitutive model. Hence, the deviation of the mature RNA distribution from the Poisson distribution is directly related to the level of RNAP traffic on the gene, as measured by the RNAP density ρ. By comparison of Eqs. (37) and (55), it is also clear that the Fano factor of the nascent RNA number is always less than that of the mature RNA number distribution. This indicates that RNAP volume exclusion effects become less apparent for RNA involved in processes downstream of transcription.

In Fig. 3, we compare the predictions of our asymptotic theory for $P_{gap}(k)$, $f_{ter,as}(t)$, and $P_M(m)$ with the results of stochastic simulations. The results were obtained for four sets of parameters α, β, and ω representing the initiation-limited regime (IL), the termination-limited regime (TL), the
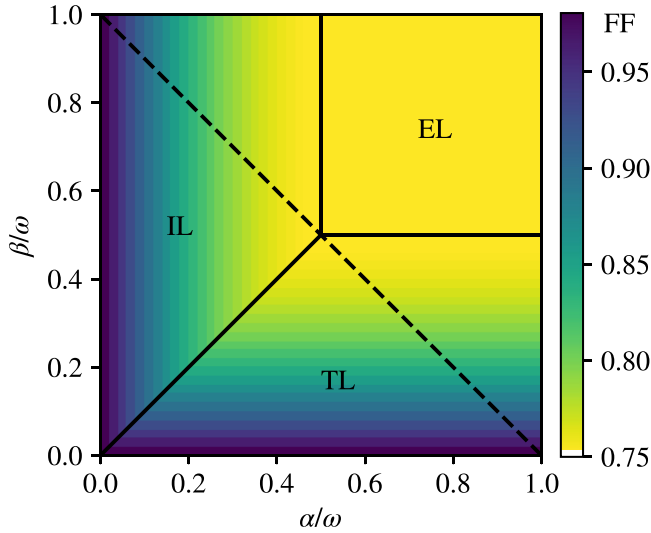
FIG. 4. Heat map of the Fano factor (FF) of the mature RNA distribution for the constitutive model with RNAP volume exclusion, computed from the renewal approximation given by Eq. (55). The Fano factor is bounded between 3/4 and 1. The model parameters are: $L = 100$, $\omega = 1$ s^{-1}, and $d_M = 0.01$ s^{-1} ($\omega \gg d_M$ is common for many genes).
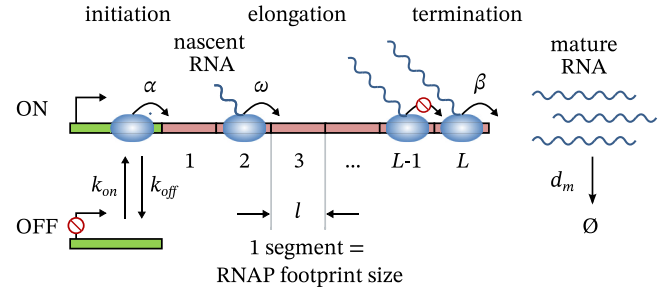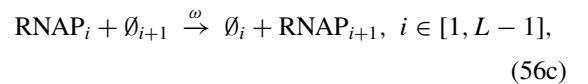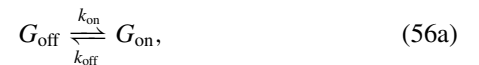

FIG. 5. Illustration of the telegraph model of transcription with RNAP volume exclusion. The gene is divided into L segments, whereby one segment equals in size to one RNAP footprint size (≈ 35 bp). The gene switches between two states of activity and inactivity with rates k_{on} and k_{off}. Transcription initiation occurs from the active state at rate α, provided the first segment is empty. RNAPs move along the gene at rate ω segments per unit time, provided the segment in front is empty. The rates of termination and mature RNA degradation are β and d_M, respectively.

coexistence line (IL/TL), and the elongation-limited regime (EL), respectively. We find an excellent agreement between $P_{\text{gap}}(k, L)$ obtained using stochastic simulations and $P_{\text{gap}}(k)$ predicted by Eq. (47) (the top row). Consequently, the pdf $f_{\text{ter,as}}(t)$ computed from Eq. (48) and the one obtained using stochastic simulations are practically indistinguishable (the middle row). The analytical and simulated results for the mature RNA distribution $P_M(m)$ agree in all the regimes, however a small but noticeable disagreement is observed in the elongation-limited regime (the bottom row). Since $f_{\text{ter}}(t)$ is well approximated by $f_{\text{ter,as}}(t)$ in the elongation-limited regime [Fig. 3(h)], we conclude that this disagreement must have originated from the renewal approximation (the assumption that the waiting times between successive termination events are uncorrelated).

We emphasize that the good agreement of theory and simulations for $L = 100$ implies that the asymptotic theory for the waiting-time distributions between successive termination events and the renewal theory for mature RNA distributions provide accurate results for genes of length larger than approximately 3500 bp (since each segment is the length of an RNAP footprint). This value is much smaller than the typical gene length in humans (the median value is 26.4 kb for protein-coding and 11.2 kb for noncoding genes, respectively [64]), but larger than the typical gene length in *Saccharomyces cerevisiae* (the average gene length is 1.4 kb). For such short genes, the asymptotic result in Eq. (48) may not be applicable if the mean gap size is of the order of the system size L, which occurs if the initiation rate is sufficiently small. However, in that case the RNAP volume exclusion can be ignored, and the results of the delay constitutive model can be used instead.

In Fig. 4, we show the Fano factor FF$_M$ computed from Eq. (55) across the whole phase diagram of the constitutive model. The smallest value of 3/4 is achieved in the

elongation-limited regime in which $\rho = 1/2$, whereas the largest value 1 is achieved in the limit of small transcription rate J, which is either when the initiation rate α or the termination rate β are much smaller than the hopping rate ω.

III. TELEGRAPH MODEL WITH RNAP VOLUME EXCLUSION

Next, we consider an extension of the constitutive model of gene expression that allows for promoter switching (transitions between two states of activity and inactivity). A cartoon illustrating the new model is shown in Fig. 5. We denote by k_{on} the rate at which the gene switches to the active state, and by k_{off} the rate at which the gene switches to the inactive state. Initiation occurs at the rate α if the gene is in the active state, and the first segment is empty. As in the original telegraph model of gene expression [18], the gene remains in the active state immediately after the initiation. The elongation, termination, and RNA degradation proceed as in the constitutive model. The model can be summarized by the following reactions:

$$G_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} G_{\text{on}}, \tag{56a}$$

$$G_{\text{on}} + \text{free RNAP} + \emptyset_1 \overset{\alpha}{\rightarrow} G_{\text{on}} + \text{RNAP}_1, \tag{56b}$$

$$\text{RNAP}_i + \emptyset_{i+1} \overset{\omega}{\rightarrow} \emptyset_i + \text{RNAP}_{i+1}, \ i \in [1, L-1], \tag{56c}$$

$$\text{RNAP}_L \overset{\beta}{\rightarrow} \emptyset_L + \text{free RNAP} + \text{RNA}, \tag{56d}$$

$$\text{RNA} \overset{d_M}{\rightarrow} \emptyset, \tag{56e}$$

Similar to the constitutive model, we can approximate the second-order reaction in Eq. (56b) by a quasi-first-order reaction $G_{\text{on}} \rightarrow G_{\text{on}} + \text{RNAP}_1$, where the rate α is proportional to the number of free RNAPs. Henceforth, we refer to this model as the telegraph model with RNAP volume exclusion.

We introduce a random variable σ describing promoter activity, whereby $\sigma = 0$ when the promoter is inactive and $\sigma = 1$ when the promoter is active. As before, C denotes the

configuration of RNAPs on the gene, n the number of RNAPs that are actively engaged in transcription, and m the number of mature RNAs. The probability to find the gene in state σ and configuration C, along with m copies of mature RNA, is denoted by $P(\sigma, C, m)$. We are interested in computing the distributions $P_N(n)$ and $P_M(m)$ defined by
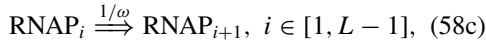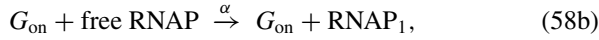
$$P_N(n) = \sum_C \sum_{\sigma=0,1} P(\sigma, C) \delta_{\sum_i \tau_i, n}, \tag{57a}$$

$$P_M(m) = \sum_C \sum_{\sigma=0,1} P(\sigma, C, m). \tag{57b}$$

Before we present results for the telegraph model with RNAP volume exclusion, we consider a simpler benchmark model in which RNAPs move deterministically on the gene, which can be solved in full for any choice of model parameters.

A. Delay telegraph model

In this model, elongation and termination take a fixed amount of time to finish, and the excluded-volume interactions between RNAPs are ignored. The reactions for this model are

$$G_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} G_{\text{on}}, \tag{58a}$$

$$G_{\text{on}} + \text{free RNAP} \xrightarrow{\alpha} G_{\text{on}} + \text{RNAP}_1, \tag{58b}$$

$$\text{RNAP}_i \xRightarrow{1/\omega} \text{RNAP}_{i+1}, \ i \in [1, L-1], \tag{58c}$$

$$\text{RNAP}_L \xRightarrow{1/\beta} \text{free RNAP} + \text{RNA}, \tag{58d}$$

$$\text{RNA} \xrightarrow{d_M} \emptyset. \tag{58e}$$

As before, double arrows denote a delay reaction that takes a fixed amount of time to finish. We refer to this model as the delay telegraph model.

This model can be solved using renewal theory, which generalizes the Poisson process to an arbitrary distribution of interarrival times [65]. To this end, we denote by $f_{\text{in}}(t)$ the probability density function of the waiting time between two successive initiation events, which has been computed in Ref. [37]. The mean waiting time between two successive initiation events is given by

$$\mu_{\text{in}} = \frac{k_{\text{off}} + k_{\text{on}}}{\alpha k_{\text{on}}}, \tag{59}$$

and the transcription rate k_{syn} is equal to $1/\mu_{\text{in}}$,

$$k_{\text{syn}} = \frac{1}{\mu_{\text{in}}} = \frac{\alpha k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}}. \tag{60}$$

The number of RNAPs that reside at segment i at time t is equal to the number of initiation events that occurred between $t - i/\omega$ and $t - (i-1)/\omega$. According to renewal theory, the mean number of initiation events in an interval Δt in the steady state is equal to $\Delta t/\mu_{\text{in}}$ [65], from which it follows that the local density $\rho_i = 1/(\mu_{\text{in}}\omega)$ for $i = 1, \dots, L-1$ and $\rho_L = 1/(\mu_{\text{in}}\beta)$. Using the expression for μ_{in} in Eq. (59), we

get that

$$\rho_i = \begin{cases} \frac{\alpha}{\omega} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}, & i = 1, \dots, L-1, \\ \frac{\alpha}{\beta} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}, & i = L. \end{cases} \tag{61}$$

The probability distribution of the total number of nascent RNA can be computed by noting that the elongation and termination steps can be grouped into one delayed reaction of duration T_{el} given by Eq. (8). The distribution $P_N(n)$ for this process has been derived in Refs. [24,29]. Alternatively, $P_N(n)$ can be computed from the Taylor expansion of the probability generating function of the nascent RNA number given by

$$G_N(u) = \frac{e^{-\kappa(u)T}}{2(k_{\text{on}}+k_{\text{off}})\delta(u)} \{(k_{\text{on}}+k_{\text{off}})\delta(u) \\ \times [e^{\delta(u)T_{\text{el}}}+1] + [(k_{\text{on}}+k_{\text{off}})^2 + \alpha u(k_{\text{on}}-k_{\text{off}})] \\ \times [e^{\delta(u)T_{\text{el}}}-1]\}, \tag{62}$$

where $u = z - 1$, $\kappa(u) = [k_{\text{on}} + k_{\text{off}} - k_{\text{syn}}u + \delta(u)]/2$, and $\delta(u) = \sqrt{(k_{\text{on}} + k_{\text{off}} - \alpha u)^2 + 4k_{\text{on}}\alpha u}$. The mean and the variance of the number of RNAPs are equal to

$$\mu_N = \frac{T_{\text{el}}}{\mu_{\text{in}}} = k_{\text{syn}}T_{\text{el}} = \frac{\alpha k_{\text{on}}T_{\text{el}}}{k_{\text{on}}+k_{\text{off}}}, \tag{63a}$$

$$\sigma_N^2 = \mu_N \left\{ 1 + \frac{2\alpha k_{\text{off}}}{T_{\text{el}}(k_{\text{on}}+k_{\text{off}})^3} [e^{-(k_{\text{on}}+k_{\text{off}})T_{\text{el}}} \\ -1 + (k_{\text{on}}+k_{\text{off}})T_{\text{el}}] \right\}. \tag{63b}$$

The Fano factor of the nascent RNA number is

$$\text{FF}_N = 1 + \frac{2\alpha k_{\text{off}}}{T_{\text{el}}(k_{\text{on}}+k_{\text{off}})^3} [e^{-(k_{\text{on}}+k_{\text{off}})T_{\text{el}}} - 1 \\ + (k_{\text{on}}+k_{\text{off}})T_{\text{el}}]. \tag{64}$$

Note that because $\exp(x) \geqslant 1 + x$ for any real x, it follows that $\text{FF}_N \geqslant 1$.

Since elongation and termination are deterministic, the waiting time between two mature RNA production events is the same as the waiting time between two RNAP binding events, hence the probability distribution of the number of mature RNA is that of the telegraph model [18,66],

$$P_M(m) = \frac{(\alpha/d_M)^m e^{-\alpha/d_M}}{m!} \frac{(k_{\text{on}}/d_M)_m}{(k_{\text{on}}/d_M + k_{\text{off}}/d_M)_m} \\ \times M(k_{\text{off}}/d_M, k_{\text{on}}/d_M + k_{\text{off}}/d_M + m, \alpha/d_M), \tag{65}$$

where $(x)_n = \Gamma(x+n)/\Gamma(x)$ is the Pochhammer symbol and M is Kummer's (confluent hypergeometric) function. In particular, the mean and the variance of this distribution are

$$\mu_M = \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{\alpha}{d_M}, \tag{66a}$$

$$\sigma_M^2 = \mu_M \left[1 + \frac{\alpha k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+d_M)} \right], \tag{66b}$$

and the Fano factor FF_M is given by

$$FF_M = 1 + \frac{\alpha k_{off}}{(k_{on} + k_{off})(k_{on} + k_{off} + d_M)} \geqslant 1. \quad (67)$$

B. Telegraph model with RNAP volume exclusion in the slow-switching regime

The steady-state probability distribution $P(C)$ is known for the constitutive model with RNAP volume exclusion, but not for the telegraph model with RNAP volume exclusion. To make progress, we therefore need to apply some approximation method. Inspired by the success of timescale separation methods to simplify stochastic models of gene regulatory networks (where RNAP dynamics is not explicitly taken into account) [67,68], here we focus on the regime of slow switching.

In the slow-switching regime, we require that the nascent RNA in each gene state reaches the steady state of the TASEP corresponding to that gene state in the absence of switching, *before* the gene switches its state again. We discuss these conditions and the constraints they impose on the parameters of the telegraph model with RNAP volume exclusion in detail in Appendix B. In general, we find that transcription is in the slow-switching regime if the mean time the gene spends in the on and off states is much larger than the total time of elongation and termination of a single RNAP.

To put these conditions in the context of eukaryotic transcription, we analyzed a large dataset of the on and off rates that were inferred from single-cell RNA sequencing data in mouse fibroblasts [22]. We performed the analysis for three values of the elongation rate (0.6, 1.8, and 2.4 kb/min), which were reported in Ref. [69]. Of 3236 genes from the dataset, the slow-switching conditions were met in 40% of the genes at 0.6 kb/min, 62% of the genes at 1.8 kb/min, and 66% of the genes at 2.4 kb/min. Details of this analysis are presented in Appendix B.

The advantage of the slow-switching regime is that the calculation of the RNAP distribution $P(C)$ is simplified, since the conditional probability $P(C|\sigma)$ can be approximated by

$$P(C|\sigma = 0) \approx \begin{cases} 1, & \tau_1 = \cdots = \tau_L = 0, \\ 0, & \text{otherwise}, \end{cases} \quad (68a)$$

$$P(C|\sigma = 1) \approx P(C)\big|_{vCM}, \quad (68b)$$

where $P(C)|_{vCM}$ is the RNAP distribution in the constitutive model with RNAP exclusion [Eq. (17)]. In simple terms, Eqs. (68a) and (68b) are the steady states of the TASEP in which the initiation rates has been set to zero and α, respectively. The probability distribution $P(C)$ can be now computed from

$$P(C) = \sum_{\sigma=0,1} P(\sigma)P(C|\sigma), \quad (69)$$

where $P(\sigma)$ is the probability to find the gene in the state $\sigma = 0, 1$,

$$P(\sigma) = \frac{k_{on}\sigma + k_{off}(1 - \sigma)}{k_{on} + k_{off}}. \quad (70)$$

1. Local RNAP density, transcription rate and the nascent RNA number distribution

Using Eq. (69) yields the following expressions for the local density ρ_i, the transcription rate k_{syn} and the nascent RNA number distribution $P_N(n)$, respectively,

$$\rho_i = \frac{k_{on}}{k_{on} + k_{off}} \rho_i\big|_{vCM}, \quad i = 1, \ldots, L, \quad (71a)$$

$$k_{syn} = \frac{k_{on}}{k_{on} + k_{off}} J, \quad (71b)$$

$$P_N(n) = \frac{k_{on}}{k_{on} + k_{off}} P_N(n)\big|_{vCM} + \frac{k_{off}}{k_{on} + k_{off}} \delta_{n,0}, \quad (71c)$$

where $\rho_i|_{vCM}$ is the local RNAP density [Eq. (24)], J is the transcription rate [Eq. (23)] and $P_N(n)|_{vCM}$ is the nascent RNA number distribution [Eq. (31)] predicted by the constitutive model with RNAP volume exclusion. The expressions for the mean and the variance of n are

$$\mu_N = \frac{k_{on}}{k_{on} + k_{off}} \mu_N\big|_{vCM}, \quad (72a)$$

$$\sigma_N^2 = \frac{k_{on}}{k_{on} + k_{off}} \sigma_N^2\big|_{vCM} + \frac{k_{on}k_{off}}{(k_{on} + k_{off})^2} \mu_N\big|_{vCM}, \quad (72b)$$

where $\mu_N|_{vCM}$ and $\sigma_M^2|_{vCM}$ are given by Eqs. (32) and (33), respectively. The Fano factor FF_N is thus given by

$$FF_N = FF_N\big|_{vCM} + \frac{k_{off}}{k_{on} + k_{off}} \mu_N\big|_{vCM}, \quad (73)$$

where $FF_N|_{vCM}$ is given by Eq. (34). Note that since $FF_N|_{vCM} < 1$, the nascent RNA number distribution can be (depending on the model parameters) either sub-Poissonian ($FF_N < 1$) or super-Poissonian ($FF_N > 1$) distribution. The transition from the sub-Poissonian to the super-Poissonian behavior occurs at the ratio $k_{off}/k_{on} \simeq 1/L$ in the initiation-limited and termination-limited regimes, and at the ratio $k_{off}/k_{on} \simeq 3/(2L)$ in the elongation-limited regime.

It is instructive to compare these results to those of the delay telegraph model analyzed in Sec. III A. Using the effective parameters discussed in Sec. II C, we find that the two models agree on the local RNAP density, transcription rate and mean nascent RNA number, but differ on the Fano factor of the nascent RNA number. To see this, we compute the Fano factor for the delay constitutive model in the slow-switching regime by expanding the right-hand side of Eq. (64) in $(k_{on} + k_{off})T_{el,eff}$, which gives up to the leading order

$$FF_N\big|_{edTM} = 1 + \frac{k_{off}}{k_{on} + k_{off}} \mu_N\big|_{vCM}, \quad (74)$$

where "edTM" stands for the effective delay telegraph model. Comparing Eqs. (73) and (74), we conclude that the Fano factor in the telegraph model with RNAP volume exclusion is always smaller than the Fano factor in the effective delay telegraph model.

In Fig. 6, we compare $P_N(n)$ obtained using stochastic simulations with the predictions of our slow-switching theory and the effective delay telegraph model. The simulations of the reaction scheme in Eq. (56) were performed in the slow-switching regime for three sets of parameters corresponding to the initiation-limited [IL, Fig. 6(a)], termination-limited [TL,
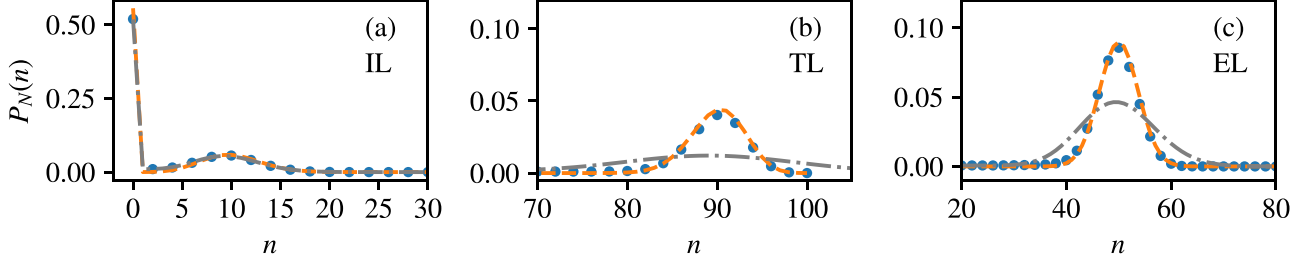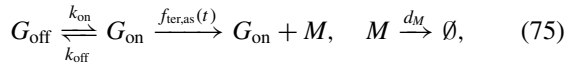
FIG. 6. Accuracy of the theoretical prediction of the nascent RNA number distribution $P_N(n)$ in the telegraph model with RNAP volume exclusion in the slow-switching regime. The blue points are from stochastic simulations of the reaction scheme in Eq. (56), the dashed orange lines are the slow-switching prediction computed from Eq. (71c) and the dot-dashed gray lines are the prediction from the effective delay telegraph model. Note that the effective delay telegraph model has the same mean nascent RNA number as the telegraph model with RNAP volume exclusion (see main text for details). The slow-switching theory provides accurate results for all levels of RNAP traffic, whereas the effective delay telegraph model is only accurate in the initiation-limited regime. The parameters are: (a) $\alpha = 0.1$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 8 \times 10^{-4}$ s^{-1}, $k_{\text{off}} = 0.001$ s^{-1}, and $d_M = 0.01$ s^{-1} (the initiation-limited regime, IL); (b) $\alpha = 0.7$ s^{-1}, $\beta = 0.1$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 10^{-4}$ s^{-1}, $k_{\text{off}} = 2 \times 10^{-4}$ s^{-1}, and $d_M = 0.01$ s^{-1} (the termination-limited regime, TL); and (c) $\alpha = 0.7$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 5 \times 10^{-4}$ s^{-1}, $k_{\text{off}} = 10^{-4}$ s^{-1}, and $d_M = 0.01$ s^{-1} (the elongation-limited regime, EL). The system size is $L = 100$ for all plots.

Fig. 6(b)], and elongation-limited regimes [Fig. 6(c)]. Our slow-switching theory agrees well with the simulations for all three levels of RNAP traffic. In contrast, the effective delay telegraph model correctly predicts the nascent RNA number distribution in the initiation-limited regime, but fails to do so in the termination-limited and elongation-limited regimes, which can be explained as follows. In the slow-switching regime, the nascent RNA number distribution predicted by the effective delay telegraph model is a mixture distribution of the delta function at $n = 0$ (corresponding to the off state), and the Poisson distribution with the rate parameter ρL (corresponding to the on state). However, the nascent RNA number distribution in the on state, $P_N(n)|_{\text{vCM}}$, is close to the Poisson distribution only in the initiation-limited regime, and it is markedly sub-Poissonian in the termination-limited and elongation-limited regimes.

2. Mature RNA number distribution

To compute the probability distribution of mature RNA number, we replace the TASEP with the following stochastic process for the turnover of mature RNA,

$$G_{\text{off}} \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} G_{\text{on}} \xrightarrow{f_{\text{ter,as}}(t)} G_{\text{on}} + M, \quad M \xrightarrow{d_M} \emptyset, \quad (75)$$

where $f_{\text{ter,as}}(t)$ is given by Eq. (48). The assumption of slow switching is crucial here, otherwise the pdf of the waiting time between two successive termination events conditioned on not leaving the on state is not necessarily described by $f_{\text{ter,as}}(t)$, which was derived for the TASEP in the steady state. Since we are interested in the probability distribution of the mature RNA number m irrespective of the gene state, we can rewrite this process as the following queuing process,

$$\emptyset \xrightarrow{h(t)} M \xrightarrow{d_M} \emptyset, \quad (76)$$

where $h(t)$ is the waiting-time distribution between two successive mature RNA production events in the process defined by Eq. (75). The steady-state distribution of the mature RNA number m for the queuing process (76) is given by Eq. (41) in which $\tilde{f}(s)$ is replaced by $\tilde{h}(s)$, the Laplace transform of

$h(t)$. The calculation of $\tilde{h}(s)$ is presented in Appendix C, and the final result is given by Eq. (C16). Inserting Eq. (C16) into Eq. (41) yields

$$P_M(0) = 1 - \frac{b_1 b_2}{\mu_h K a_1 a_2}$$
$$\times \left[1 - {}_2F_2 \left(\frac{a_1}{d_M}, \frac{a_2}{d_M}; \frac{b_1}{d_M}, \frac{b_2}{d_M}; -\frac{K}{d_M} \right) \right], \quad (77a)$$

$$P_M(m) = \frac{b_1 b_2 \left(\frac{K}{d_M} \right)^m \left(\frac{a_1}{d_M} \right)_m \left(\frac{a_2}{d_M} \right)_m}{\mu_h K a_1 a_2 (m!) \left(\frac{b_1}{d_M} \right)_m \left(\frac{b_2}{d_M} \right)_m}$$
$$\times {}_2F_2 \left(\frac{a_1}{d_M} + m, \frac{a_2}{d_M} + m; \right.$$
$$\left. \frac{b_1}{d_M} + m, \frac{b_2}{d_M} + m; -\frac{K}{d_M} \right), \quad m \geqslant 1. \quad (77b)$$

In the expression above, μ_h is the first moment of $h(t)$,

$$\mu_h = \frac{(k_{\text{off}} + k_{\text{on}})(k_{\text{off}} + \omega)}{k_{\text{on}} J \omega}. \quad (78)$$

Note that $_2F_2$ is a generalized hypergeometric function, and the constants K, a_1, a_2, b_1 and b_2 are defined through the factorization of the polynomials in $\tilde{h}/[1 - \tilde{h}(s)]$ such that

$$\frac{\tilde{h}(s)}{1 - \tilde{h}(s)} = \frac{K(s + a_1)(s + a_2)}{s(s + b_1)(s + b_2)}. \quad (79)$$

The mean and the variance of the mature RNA number m are given by

$$\mu_M = \frac{1}{\mu_h d_M} = \frac{k_{\text{on}}}{(k_{\text{on}} + k_{\text{off}})} \frac{J\omega}{d_M(\omega + k_{\text{off}})}, \quad (80a)$$

$$\sigma_M^2 = \frac{1}{\mu_h d_M} \left[1 + \frac{\tilde{h}(d_M)}{1 - \tilde{h}(d_M)} - \frac{1}{\mu_h d_M} \right], \quad (80b)$$

from which we get the following expression for the Fano factor FF_M

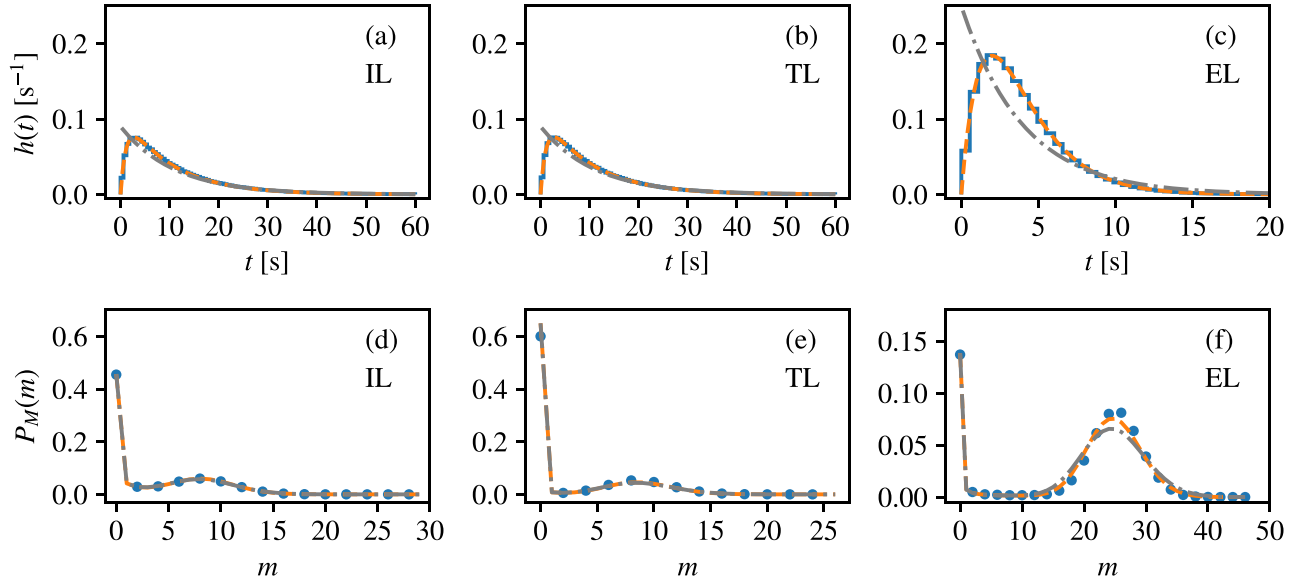$$\text{FF}_M = 1 + \frac{\tilde{h}(d_M)}{1 - \tilde{h}(d_M)} - \frac{1}{\mu_h d_M}. \quad (81)$$

FIG. 7. Accuracy of the theoretical predictions of the statistics of the time between termination events and of mature RNA numbers in the slow-switching regime of the telegraph model with RNAP volume exclusion. [(a)–(c)] Compare the pdf $h(t)$ of the waiting-time distribution between two successive termination events computed using stochastic simulations (solid blue lines), the slow-switching prediction obtained from $\tilde{h}(s)$ in Eq. (C16) (dashed orange lines), and the prediction of the effective delay telegraph model (dash-dotted gray lines). [(d)–(f)] Compare the mature RNA number distribution $P_M(m)$ computed using stochastic simulations (blue points), the slow-switching prediction computed from Eq. (77) (dashed orange line), and the prediction of the effective delay telegraph model (dash-dotted gray line). Note that the effective delay telegraph model has the same mean nascent and mature RNA numbers as the telegraph model with RNAP volume exclusion (see main text for details). The parameters are: (a),(d) $\alpha = 0.1$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 8 \times 10^{-4}$ s^{-1}, $k_{\text{off}} = 0.001$ s^{-1}, and $d_M = 0.01$ s^{-1} (the initiation-limited regime, IL); (b),(e) $\alpha = 0.7$ s^{-1}, $\beta = 0.1$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 10^{-4}$ s^{-1}, $k_{\text{off}} = 2 \times 10^{-4}$ s^{-1}, and $d_M = 0.01$ s^{-1} (the termination-limited regime, TL); and (c),(f) $\alpha = 0.7$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{\text{on}} = 5 \times 10^{-4}$ s^{-1}, $k_{\text{off}} = 10^{-4}$ s^{-1}, and $d_M = 0.01$ s^{-1} (the elongation-limited regime, EL). The system size is $L = 100$ for all plots.

Interestingly, the mean mature RNA number in Eq. (80a) differs from the one predicted by the effective delay telegraph model [Eq. (66a)]—the former is smaller by a factor of $\omega/(\omega + k_{\text{off}})$, which is close to 1 only when $k_{\text{off}} \ll \omega$. Given that the elongation rate is typically in the range of 10–100 nt/s (corresponding to the hopping rate ω in the range of 0.3–3.0 segments/s), this condition requires that the average time the gene spends in the on state is much larger than a few seconds, which is in line with values reported in the literature that measure in minutes. Following this argument, we set $\omega = k_{\text{off}}/x$ and expand the Fano factor of the mature RNA number given by Eq. (81) in x, keeping the first two terms. The calculation, not shown here, reveals that the leading term is precisely the Fano factor predicted by the delay telegraph model [Eq. (67)], whereas the leading correction is always negative. Hence, in the biologically relevant regime of $k_{\text{off}} \ll \omega$, the Fano factor of the mature RNA number predicted by the telegraph model with RNAP volume exclusion is smaller than the one predicted by the delay telegraph model.

In Fig. 7, we compare $h(t)$ and $P_M(m)$ obtained using stochastic simulations with the predictions of our slow-switching theory and the effective delay telegraph model. The simulations of the reaction scheme in Eq. (56) were performed in the slow-switching regime for the same three sets of parameters as in Fig. 6 corresponding to the initiation-limited, termination-limited and elongation-limited regimes. Overall, our slow-switching theory agrees well with the simulations for all three levels of RNAP traffic. (A small discrepancy between the slow-switching theory and the simulations is found in the elongation-limited regime, which is due to the renewal approximation that ignores correlations between successive termination events.)

The mature RNA number distribution predicted by the effective delay telegraph model matches closely that of the telegraph model with RNAP volume exclusion, especially in the initiation-limited and termination-limited regimes [Figs. 7(d) and 7(e)], whereas a small but visible discrepancy between the distributions is found in the elongation-limited regime [Fig. 7(f)]. While this agreement may seem surprising, we note that the results were obtained in the biologically relevant regime of $k_{\text{off}} \ll \omega$ in which the mean and the variance of the mature RNA number, as well as the first three moments of the waiting-time distribution, agree between the two models (provided the delay telegraph model is used with the effective parameters discussed in Sec. II C).

We note that besides the effective delay telegraph model here considered, there are other ways of finding a reduced model for the mature RNA distributions which might lead to equally good fits to the telegraph model with RNAP volume exclusion. For example, following the approach in Ref. [70], one could find parameters of the conventional telegraph model (promoter switching and synthesis rate) such that its predictions for the fraction of time that the gene is active, the mean mature RNA number and the variance of the mature RNA fluctuations agree with those of the reaction scheme in Eq. (56).
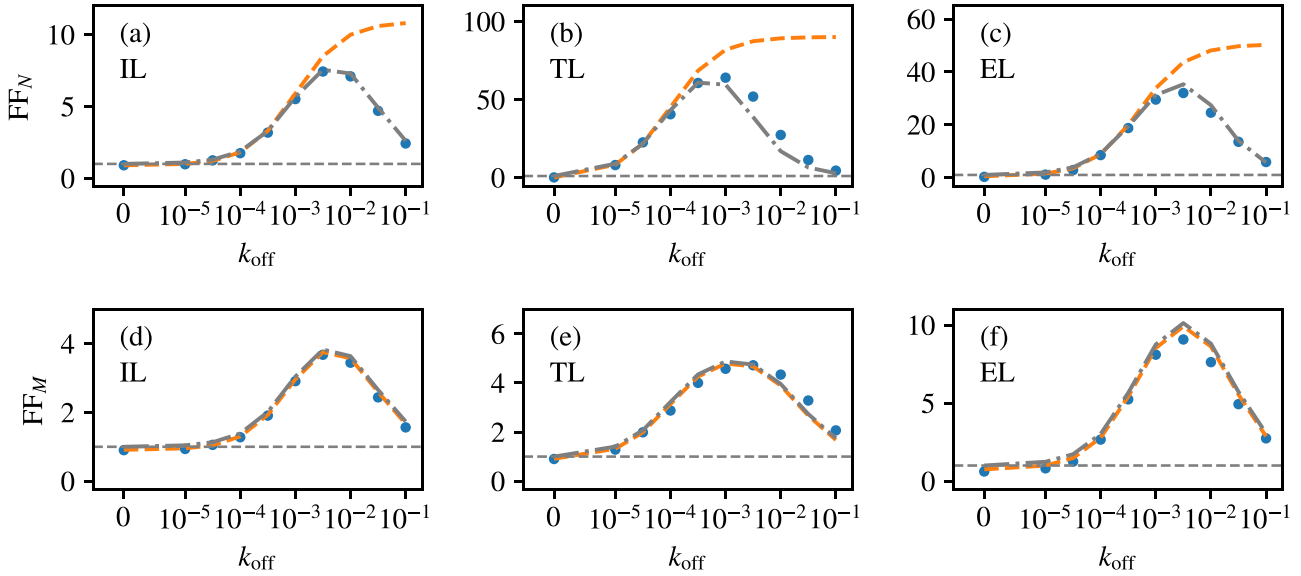
FIG. 8. Accuracy of the theoretical predictions for the Fano factor of nascent and mature RNA numbers in the telegraph model with RNA volume exclusion as a function of the switching off rate. [(a)–(c)] Compare the Fano factor FF_N of the nascent RNA number distribution computed using stochastic simulations (blue points), the slow-switching prediction obtained from Eq. (73) (dashed orange lines) and the prediction of the effective delay telegraph model (dash-dotted gray lines). [(d)–(f)] Compare the Fano factor FF_M of the mature RNA number distribution computed using stochastic simulations (blue points), the slow-switching prediction computed from Eq. (81) (dashed orange line), and the prediction of the effective delay telegraph model (dash-dotted gray line). The dashed gray line is at the value of 1, marking a transition from the sub-Poissonian to the super-Poissonian behavior. The parameters are: (a),(d) $\alpha = 0.1$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{on} = 10^{-3}$ s^{-1}, and $d_M = 0.02$ s^{-1} (the initiation-limited regime, IL); (b),(e) $\alpha = 0.3$ s^{-1}, $\beta = 0.1$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{on} = 10^{-4}$ s^{-1}, and $d_M = 0.02$ s^{-1} (the termination-limited regime, TL); and (c),(f) $\alpha = 0.7$ s^{-1}, $\beta = 0.7$ s^{-1}, $\omega = 1.0$ s^{-1}, $k_{on} = 5 \times 10^{-4}$ s^{-1}, and $d_M = 0.02$ s^{-1} (the elongation-limited regime, EL). The system size is $L = 100$ for all plots.

C. Telegraph model with RNAP volume exclusion outside the slow-switching regime

We consider here what happens when we relax the slow-switching condition in Eq. (B8), but keep the condition in Eq. (B1) satisfied. Relaxing the condition in Eq. (B8) means that the RNAP distribution in the on state is no longer given by the steady-state distribution of the TASEP [Eq. (17)], whereas keeping the condition in Eq. (B1) satisfied means that the gene in the off state has no active RNAPs. The motivation to study this regime comes from mammalian genes which stay inactive for hours, but are active only for few minutes, which may not be enough for the RNAP dynamics to reach the steady state.

In Fig. 8, we show the Fano factor of the nascent and mature RNA numbers obtained using stochastic simulations for various values of the off rate k_{off}, including values outside the slow-switching regime. These results are compared to the predictions of the slow-switching theory and the predictions of the effective delay telegraph model. As can be seen from Figs. 8(a)–8(c) depicting the Fano factor FF_N dependence on k_{off}, the slow-switching theory eventually fails as k_{off} is increased, whereas the effective delay telegraph model prediction provides a good fit for all values of k_{off}, except for a small but visible disagreement in the termination-limited regime [Fig. 8(b)]. In contrast, the Fano factor of the mature RNA number is well accounted for by both the slow-switching theory and the effective delay telegraph model.

We emphasize that an implicit reason for the agreement of the effective delay telegraph model and the telegraph model with RNAP volume exclusion in Fig. 8 is that for all the

values of k_{off} considered (except the value of $k_{off} = 0$), the RNA fluctuations in the latter model are super-Poissonian. If the parameters were such that the RNA fluctuations are sub-Poissonian (by choosing k_{off} to be much less than k_{on}), then this model matching would be impossible.

The agreement between the models can be understood by looking at the time evolution of the TASEP, starting from an empty lattice (the initial state at the time of switching to the on state). We note that this problem was previously studied in the context of mRNA translation [71] and stochastic resetting [72]. We begin with the initiation-limited regime in which $\alpha < \omega/2$ and $\beta > \alpha$ [Figs. 8(a) and 8(d)]. In this case, the RNAP density is a shock wave of density α/ω traveling at the speed of $\omega(1 - \alpha/\omega)$. The nascent RNA production rate is given by $\alpha(1 - \alpha/\omega)$, and the total time it takes the RNAP density wave to travel across the gene is equal to $L/[\omega(1 - \alpha/\omega)]$. We note that the RNAP density of α/ω is compatible with the termination rate $\beta > \alpha$. Hence, once the RNAP density wave reaches the end of the gene, the RNAP density does not change, and the rate of mature RNA production is given by $\alpha(1 - \alpha/\omega)$. This explains why the effective delay telegraph model provides a good fit for all values of k_{off}.

We next consider the termination-limited regime in which $\beta < \omega/2$ and $\alpha > \beta$ [Figs. 8(b) and 8(e)]. For $\alpha < \omega/2$, the RNAP density wave is a shock wave of density α/ω as in the initiation-limited regime. Once the RNAP density wave reaches the end of the gene, the RNAPs begin to accumulate, creating a second shock wave of density $1 - \beta/\omega$ that travels backwards toward the promoter. This happens because the termination rate $\beta < \alpha$ is not compatible with the density of

α/ω. Because the RNAP density at the end of the gene is now equal to $1 - \beta/\omega$, the rate of mature RNA production is given by $\beta(1 - \beta/\omega)$. Once the second shock wave reaches the start of the gene, provided the gene stays in the on state long enough, the rate of nascent RNA production changes from $\alpha(1 - \alpha/\omega)$ to $\beta(1 - \beta/\omega)$. Otherwise, the RNAP density becomes nonuniform along the gene, which explains why the Fano factor of the nascent RNA number predicted by the effective delay telegraph model does not quite match the simulations for large values of k_{off} [Fig. 8(b)].

Finally, we consider the elongation-limited regime in which $\alpha > \omega/2$ and $\beta > \omega/2$ [Figs. 8(b) and 8(e)]. In this case, the RNAP density is not a shock wave, i.e., the initiation rate of α has no influence on the RNAP density. Instead, the RNAP density is a rarefaction wave that decays linearly along the gene, and the rate of nascent RNA production is given by $\omega/4$. Once the tip of the RNAP density wave reaches the end of the gene, which happens after time L/ω, the RNAP density begins to increase toward its steady-state value of $1/2$. If the gene stays long enough for this to happen, then the total time of elongation and termination is equal to $2L/\omega$, and the rate of mature RNA production is given by $\omega/4$. We would therefore expect the effective delay telegraph model not to provide a good fit for large values of k_{off}. However, that is not what we see in Figs. 8(b) and 8(e), i.e., the effective delay telegraph model fits the data obtained using simulations even for large values of k_{off} (up to 10% value of the elongation rate ω). Presently, we do not have a good explanation for this agreement.

IV. SUMMARY AND DISCUSSION

The main question we asked in this study is how the excluded volume interactions between RNAPs affect the distributions of nascent and mature RNA numbers. To answer this question, we developed a stochastic model of gene expression that accounts for: (i) transcription initiation, (ii) transcription elongation by RNAPs moving along the gene, (iii) transcription termination and nascent RNA processing into a mature RNA, and (iv) degradation of mature RNA. The movement of RNAPs was modelled by the totally asymmetric simple exclusion process (TASEP) in which each lattice site corresponded to ≈ 35 bp (the footprint of RNAP) and RNAPs could only move forward with nonzero probability if the next segment is free. We considered two initiation mechanisms, one in which the gene is always active (the constitutive model with RNAP volume exclusion, see Sec. II), and the other in which the gene switches between two states of activity and inactivity (the telegraph model with RNAP volume exclusion, Sec. III). To determine the importance of RNAP volume exclusion, we compared these models with another set of models in which RNAPs move deterministically and do not interact with each other (the delay constitutive and telegraph models).

A. Constitutive model with RNAP volume exclusion

For the constitutive model with RNAP volume exclusion, we obtained an exact expression for the distribution of nascent RNA number in the steady state. The shape of this distribution is strongly determined by the RNAP density ρ, which equals the average probability of a gene segment being occupied by an RNAP. The distribution is very close to a binomial distribution in the initiation-limited ($\rho < 1/2$) and termination-limited ($\rho > 1/2$) regimes, but differs substantially from the binomial distribution at the coexistence line between the initiation-limited and termination-limited regimes, and in the elongation-limited regime ($\rho = 1/2$). A direct consequence of the excluded volume interactions between RNAPs is that the nascent RNA number distribution is sub-Poissonian. Consequently, the Fano factor of the nascent RNA number ranges between 0 when $\rho \to 1$ and 1 when $\rho \to 0$. In contrast, the delay constitutive model predicts a Poisson distribution of the nascent RNA number whose Fano factor equals exactly 1. We note that the Poisson distribution is a special limit of the binomial distribution when the RNAP density is small and the gene length is large. Hence, the two models are expected to agree when the RNAP density is low. A similar result that the nascent RNA number distribution becomes sub-Poissonian due to excluded-volume interaction has been obtained before using stochastic simulations of the TASEP in which RNAPs occupy more than one lattice site [44].

The distribution of the mature RNA number was computed in the renewal approximation, in which the correlations between waiting times between successive termination events are ignored. This approximation allowed us to map the production and degradation of mature RNA to a queuing process. The distribution of mature RNA number was computed analytically using known results from queuing theory. This distribution is also sub-Poissonian, whereby its Fano factor ranges from $3/4$ in the elongation-limited regime to 1 when $\rho \to 0$ or $\rho \to 1$. In contrast, the delay constitutive model predicts a Poisson distribution whose Fano factor equals exactly 1. Our results also show that the Fano factor of the nascent RNA number is less than that of the mature RNA number, implying that a signature of RNAP volume exclusion effects (the sub-Poissonian nature of fluctuations) is gradually erased as RNA progresses through its lifecycle.

It is important here to note that experimentally observed sub-Poissonian fluctuations in nascent and nuclear transcript numbers are generally not solely due to RNAP volume exclusion effects; they can also be caused by multiple rate-limiting steps in initiation [28,73,74], which we have not considered in this paper. In particular, the same waiting-time distribution between two mature RNA production events that was obtained in the constitutive model with RNAP volume exclusion [the hypoexponential distribution in Eq. (48)], can also be obtained in a constitutive model with no RNAP volume exclusion, but with a two-step initiation mechanism of the type: $G_0 \to G_1 \to G_0 + \text{RNA}, \text{RNA} \to \emptyset$ (G_i are the promoter states).

Furthermore, we note that the mature RNA distribution is invariant to $\rho \leftrightarrow 1 - \rho$, which originates from the waiting-time distribution between two successive termination events possessing the same symmetry. This implies a fundamental limitation to the accurate estimation of the RNAP density from experimental measurements of mature RNA number alone, i.e., without measuring the nascent RNA number.

B. Telegraph model with RNAP volume exclusion

Similarly, for the telegraph model with RNAP volume exclusion, we obtained analytical expressions for the nascent and mature RNA distributions in the slow-switching regime, in which case the TASEP has enough time to reach the steady state it would relax to in the absence of switching, before the next switching event occurs. In this regime, the nascent RNA number distribution simplified to a mixture distribution of the delta function at zero and the nascent RNA number distribution derived for the constitutive model. To compare this distribution to the one predicted by the delay telegraph model, we adjusted the initiation rate and the elongation time of the delay telegraph model to match those predicted by the TASEP, which we called the effective delay telegraph model. This made sure that the two nascent RNA number distributions had the same mean. Nevertheless, significant deviations between the two distributions were observed in the termination-limited and elongation-limited regimes, suggesting that caution is needed when inferring parameters by fitting nascent RNA data to the delay telegraph model [25,75,76]. In contrast, the two models agreed on the mature RNA number distribution both in the initiation-limited and termination-limited regimes, whereas a small discrepancy between the distribution was found in the elongation-limited regime. We also found that the slow promoter switching causes the Fano factor of the nascent RNA number to be larger than in the constitutive model, potentially changing the nature of the fluctuations from sub-Poissonian to super-Poissonian. This transition occurs when $k_{\text{off}}/k_{\text{on}} \simeq 1/L$, i.e., when the gene spends most of the time in the on state. As k_{off} increases and becomes much larger than k_{on}, a condition that is often associated with bursty gene expression [6,12], the effects of RNAP volume exclusion on the nascent RNA fluctuations are expected to diminish.

Outside the slow-switching regime, limited results were obtained, as the telegraph model with RNAP volume exclusion becomes difficult to solve analytically. Motivated by mammalian genes which have short periods of activity followed by long periods of inactivity, we focused on the regime in which the nascent RNA reaches the steady state in the off state (an empty lattice), but not necessarily in the on state. We computed the Fano factor of the nascent and mature RNA numbers using stochastic simulations, and compared it to the predictions of the effective delay telegraph model. We showed that the effective delay telegraph model provides a good match for the telegraph model with RNAP volume exclusion well beyond the slow-switching regime.

C. Model-based inference of transcription kinetics from single-cell mature RNA data

Although nascent RNA is a direct reflection of the transcription process, most of the gene expression data comes from measuring the mature RNA. The question is then, what can the mature RNA data tell us about transcription kinetics and in particular about the RNAP density (assuming mature RNA fluctuations are not dominated by post-transcriptional noise)?

For the constitutive expression and for the slow promoter switching, the mature RNA number distribution is determined by the waiting-time distribution between two successive termination events, and the degradation rate d_m. The waiting-time distribution is parametrized by $\omega\rho$ and $\omega(1-\rho)$, and is invariant under the exchange of $\rho \leftrightarrow 1-\rho$, where ρ is the RNAP density in the on state. Hence, it is not possible to distinguish from the mature RNA data alone whether the transcription is in the initiation-limited or termination-limited regime.

The same is true outside the slow-switching regime. We remind the reader that in this case, the effective delay telegraph model provides a reasonably good fit to the mature RNA number distribution. For this model, the effective transcription rate in the on state is equal to J, the RNAP current of the TASEP [Eq. (25)]. Since J is invariant under the exchange of $\alpha \leftrightarrow \beta$, inferring the value of J from the mature RNA data is not enough to determine whether the transcription is in the initiation-limited or termination-limited regime. There is a fundamental loss of information due to the downstream processing of nascent RNA, which makes it difficult to infer transcriptional dynamics from the mature RNA data.

If one is interested in the estimation of transcriptional parameters, then one can use the nascent RNA number distributions derived in this paper to construct the likelihood function. The maximization of this function, given the measured single-cell nascent RNA data, leads to the desired estimates. This approach is presently limited to constitutive expression and the slow-switching regime, for which we have derived analytical expressions for the nascent RNA number distributions. For genes with a substantial RNAP traffic, this approach may produce more accurate estimates than models of nascent RNA dynamics that ignore steric interactions. Examples of such genes are the ribosomal genes. In *Escherichia coli*, one RNA polymerase was observed every 85 bp on rRNA operons [77], which is equivalent to the RNAP density of 41%. In yeast, between 30 and 70 RNA polymerases were observed on the 35S rRNA gene [46], which given the gene length of 6858 bp amounts to the RNAP density in the range of 15–36%. Other genes which might have a large amount of RNAP traffic might be those whose transcription rate scales with the cell volume, particularly when the volume is large [78].

D. Possible extensions to include more details of transcription initiation and elongation

In general, there might be more than two promoter states [79], for example, to describe off duration times that are nonexponential, which has been measured in some mammalian cells [12]. As well, extra promoter states could capture multiple rate-limiting steps in initiation [29,74]. The promoter-proximal pausing of RNAP II in metazoans, which occurs within 100 bp from the transcription start site, can also be considered as one of the promoter states since the paused RNAP blocks another RNAP from being recruited to the promoter [28,80]. The methods developed in this paper can be extended in these directions, provided the RNAP distribution is allowed to reach the steady state in each of the

promoter states much sooner than the next change of state occurs.

Since our model describes the movement of RNAPs at the resolution of ≈ 35 bp (the footprint length of the RNAP), it does not have an explicit description of the processes occurring at the single nucleotide level, such as ubiquitous pausing and backtracking. These processes can be included implicitly in our model by setting ω to $1/(\ell\tau)$, where τ is the mean residence time the RNAP spends at each nucleotide and ℓ is the footprint length of the RNAP. For example, in a model with ubiquitous pausing in which an RNAP enters the paused state at rate k_p, returns to the active state at rate k_a, and moves to the next nucleotide in the active state at rate ϵ [30], the mean residence time τ is equal to $(k_p + k_a)/(k_a\epsilon)$. Backtracking is usually modelled as a biased random walk, in which case τ is equal to the mean first-passage time it takes the RNAP to return to the original position it backtracked from [33]. We note that setting ω to $1/(\ell\tau)$ in these examples will work only if the RNAP pausing does not cause a substantial queuing of the RNAPs. Otherwise, the mean residence time τ is not only determined by the pausing process, but also by the time the RNAP spends waiting in a queue. Finally, we note that our model does not account for long-range cooperation between RNAPs during transcription elongation [81,82]. This cooperation results in a reduced elongation time when multiple RNAPs transcribe the gene at the same time. Several theoretical studies have been put forward to explain this cooperation via DNA supercoiling [42,83–85]. However, the RNAP density at which these effects become evident is poorly understood. For example, a several-fold increase in RNAP density on *lacZ* operon showed no effect on the elongation rate [82]. Future work will focus on resolving how the aforementioned mechanistic details affect the nascent and mature RNA number distributions.

In conclusion, we have developed a theory of stochastic gene expression that elucidates the link between fluctuations in RNA numbers and the strength of mechanical interactions between RNAPs. The specialty of the model is its analytic tractability, which enables us to make several predictions of physical and biological relevance. The theory (i) uncovers a fundamental limitation in the deduction of RNA polymerase traffic patterns on a gene from the steady-state distribution of mature RNA numbers; (ii) proves that steric interactions result in a suppression of fluctuations in the RNA numbers, potentially even leading to sub-Poissonian fluctuations; (iii) provides a novel interpretation of standard models of gene expression by expressing their parameters in terms of the kinetic parameters controlling the microscopic processes of transcription. Our work shows that there is a wealth of information about transcriptional processes hidden in the nascent RNA number fluctuations. This information can be extracted using statistical-mechanical techniques previously developed for the TASEP, giving us in turn an unprecedented view of stochastic gene expression.

APPENDIX A: THE PROBABILITY GENERATING FUNCTION $G_N(n)$ IN THE CONSTITUTIVE MODEL WITH RNAP VOLUME EXCLUSION

Starting from Eq. (28), we introduce

$$Y_L(z) = \langle W|(zD + E)^L|V\rangle, \tag{A1}$$

and denote by $H(w, z)$ the generating function for $Y_L(z)$,

$$H(w, z) = \sum_{L=0}^{\infty} Y_L(z)w^L$$
$$= \langle W|[1 - w(zD + E)]^{-1}|V\rangle. \tag{A2}$$

Next, we note that since $DE = D + E$, we can write

$$1 - w(zD + E) = (1 - aD)(1 - bE), \tag{A3}$$

where a and b satisfy

$$wz = a(1 - b), \quad w = b(1 - a). \tag{A4}$$

Hence, the inverse of $1 - w(zD + E)$ is given by

$$[1 - w(zD + E)]^{-1} = (1 - bE)^{-1}(1 - aD)^{-1}. \tag{A5}$$

Using Eq. (A4), we express b in terms of a and z, which gives $b = a/(z - za + a)$. Altogether,

$$H(a, z) = \frac{1 - \left(\frac{z-1}{z}\right)a}{\left[1 - \left(\frac{\alpha z - \alpha + 1}{\alpha z}\right)a\right]\left(1 - \frac{1}{\beta}a\right)}, \tag{A6}$$

where a is a function of w and z defined implicitly by the equation

$$w = \frac{a(1 - a - z)}{z(1 - a)}. \tag{A7}$$

To find $Y_L(z)$, we look for the coefficient of w^L in the Taylor expansion of $H(w, z)$ around $w = 0$,

$$Y_L(z) = [w^L]H(w, z) = \frac{1}{L!}\frac{\partial^L H}{\partial w^L}\bigg|_{w=0}, \tag{A8}$$

where $[w^L]$ is an operator which extracts the coefficient of w^L in the Taylor series of a function of w. According to the Lagrange inversion formula, if we can find a function $\phi(a, z)$ such that $w = a/\phi(a, z)$ whereby $\phi(0, z) \neq 0$, then

$$[w^L]H(w, z) = \frac{1}{L}[a^{L-1}]\left(\frac{\partial H(a, z)}{\partial a}[\phi(a, z)]^L\right). \tag{A9}$$

The advantage of this formula is that the right-hand side is much easier to calculate than the left-hand side. According to Eq. (A7),

$$w = \frac{a}{\phi(a, z)} \text{ and } \phi(a, z) = z - 1 + \frac{1}{1 - a}. \tag{A10}$$

By expanding the right-hand side of Eq. (A9) and collecting the terms of the same order in a yields $G_N(z)$ given by Eqs. (29).

APPENDIX B: SLOW-SWITCHING CONDITIONS AND THEIR PREVALENCE IN MAMMALIAN GENES

We consider the TASEP in the absence of switching and with the initiation rate set to α. We assume that the TASEP

is initially in the steady state, i.e., the RNAP distribution is given by Eq. (17). At time $t = 0$, we set the initiation rate to zero, and observe the relaxation of the TASEP to the new steady state in which the lattice is empty. We denote by T_{off} the mean time it takes the TASEP to reach that state. For the approximation in Eq. (68a) to hold, we require that

$$T_{\text{off}} \ll \frac{1}{k_{\text{on}}}, \tag{B1}$$

where $1/k_{\text{on}}$ is the mean time that the gene spends in the off state. To compute T_{off}, we denote by k the position of the leftmost RNAP on the lattice at $t = 0$, and by $P_1(k)$ the probability distribution of k. The mean time T_{off} is then given by

$$T_{\text{off}} = \sum_{k=1}^{L} P_1(k) \sum_{i=k}^{L} \frac{\rho_i}{J}, \tag{B2}$$

where $\sum_{i=k}^{L} \rho_i / J$ is the mean time it takes the leftmost RNAP at the segment k to leave the gene. The probability distribution $P_1(k)$ can be written as

$$P_1(k) = \frac{\langle W | E^{k-1} DC^{L-k} | V \rangle}{Z_L}, \quad k = 1, \dots, L. \tag{B3}$$

For $k = L$, we get

$$P_1(L) = \frac{1}{Z_L} \left(\frac{\omega}{\alpha} \right)^{L-1} \frac{\omega}{\beta}, \tag{B4}$$

where Z_L is given by Eq. (21). For $k = 1, \dots, L-1$, we use the following identity [48]:

$$DC^{L-k} = \sum_{i=0}^{L-k-1} B_{i+1,1} C^{L-k-i} + \sum_{i=2}^{L-k+1} B_{L-k,i-1} D^i, \tag{B5}$$

where $B_{k,p}$ is defined in Eq. (22). This gives

$$P_1(k) = \left(\frac{\omega}{\alpha} \right)^{k-1} \sum_{i=0}^{L-k-1} B_{i+1,1} \frac{Z_{L-k-i}}{Z_L}$$
$$+ \left(\frac{\omega}{\alpha} \right)^{k-1} \sum_{i=2}^{L-k+1} B_{L-k,i-1} \frac{1}{Z_L} \left(\frac{\omega}{\beta} \right)^{i}. \tag{B6}$$

If we assume that $\rho_i \approx \rho$, then $P_1(k) = (1-\rho)^{k-1} \rho$, in which case the expression for T_{off} simplifies to

$$T_{\text{off}} = \frac{L}{\omega(1-\rho)} - \frac{1}{\omega\rho} [1 - (1-\rho)^L]. \tag{B7}$$

We now consider the TASEP in the absence of switching and with the initiation rate set to zero. We assume that the TASEP is initially in the steady state, i.e., the lattice is empty. At time $t = 0$, we set the initiation rate to α, and observe the relaxation of the TASEP to the new steady state in which the RNAP probability distribution is given by Eq. (17). We denote by T_{on} the mean time it takes the TASEP to reach that state. For the approximation in Eq. (68b) to hold, we require that

$$T_{\text{on}} \ll \frac{1}{k_{\text{off}}}. \tag{B8}$$

The time-evolution of the TASEP from an empty lattice has been studied in the context of mRNA translation [71,86], and

more recently in the context of stochastic resetting [72]. In the initiation-limited regime, $T_{\text{on}} = T_{\text{el}}$, where $T_{\text{el}} \approx L/[\omega(1 - \rho)]$ is the mean time of elongation and termination of a single RNAP. For the expression for T_{on} in the termination-limited regime, we refer to Ref. [72]. Unfortunately, no analytical expression for T_{on} is available on the coexistence line and in the elongation-limited regime. However, we know that the large-time relaxation of the TASEP is determined by the eigenvectors of the transition matrix with the largest real parts of the corresponding eigenvalues (note that the real parts of all eigenvalues are nonpositive) [87]. On the coexistence line, the spectral gap Δ (the difference between the zero eigenvalue corresponding to the steady state and the next largest eigenvalue) approaches zero as L^{-2} for large L, signaling a diffusive relaxation time $\Delta^{-1} \propto L^2$, whereas in the elongation-limited regime the spectral gap approaches zero as $L^{-3/2}$ for large L [87]. Hence, we expect T_{on} to scale with L^2 on the coexistence line, and with $L^{3/2}$ in the elongation-limited regime.

The conditions (B1) and (B8) were tested on a large dataset of the on and off rates inferred from single-cell RNA sequencing data in mouse fibroblasts [22]. Gene lengths for each gene in the dataset were obtained from the GRCm39 genome assembly for *Mus musculus* published by Ensembl (release 106, Nov. 2022) using `gffutils` package in Python. The data was analyzed for three values of the elongation rate (0.6, 1.8, and 2.4 kb/min) reported for this type of cells in Ref. [69]. These elongation rates were measured by observing the depletion of RNAPs after stopping new transcription initiation, therefore we assumed that these values represent the effective elongation rates $\omega_{\text{eff}} = \omega(1 - \rho)$, rather than the bare ones (ω). The conditions (B1) and (B8) were tested assuming $T_{\text{on}} \simeq T_{\text{el}}$ and $T_{\text{off}} \simeq T_{\text{el}}$, where $T_{\text{el}} \approx L/[\omega(1 - \rho)]$. A gene was considered to be in the slow-switching regime if both $k_{\text{on}}/T_{\text{el}} < 0.2$ and $k_{\text{off}}/T_{\text{el}} < 0.2$ conditions were satisfied.

Of 3236 genes, 78% of the genes at 0.6 kb/min, 96% of the genes at 1.8 kb/min, and 98% of the genes at 2.4 kb/min satisfied the condition (B1), meaning that for most genes the off state should last long enough for all RNAPs to leave the gene before the gene switches back to the on state. This also means that most of the genes will have only one transcriptional burst of RNAPs active on the gene. In contrast, the conditions (B1) and (B8) were simultaneously satisfied in 40% of the genes at 0.6 kb/min, 62% of the genes at 1.8 kb/min, and 66% of the genes at 2.4 kb/min.

APPENDIX C: WAITING-TIME DISTRIBUTION $h(t)$ BETWEEN TWO SUCCESSIVE TERMINATION EVENTS IN THE TELEGRAPH MODEL WITH RNAP VOLUME EXCLUSION

For the process described in Eq. (75), we define the probability density functions $f_{0 \to 1}(t)$, $f_{1 \to 0}(t)$, and $f_{1 \to 2}(t)$ as follows: $f_{0 \to 1}(t)$ is the probability density function of the waiting time in the off state, after which the gene switches to the on state, $f_{1 \to 0}(t)$ is the probability density function of the waiting time in the on state after which the gene switches to the off state, and $f_{1 \to 2}(t)$ is the probability density function of the waiting time in the on state, after which the gene produces a nascent RNA (and remains in the on state). We denote

their respective Laplace transforms by $\tilde{f}_{0\to1}(s)$, $\tilde{f}_{1\to0}(s)$, and $\tilde{f}_{1\to2}(s)$.

Using these definitions, the Laplace transform of $h(t)$, denoted by $\tilde{h}(s)$, can be written as

$$
\tilde{h}(s) = \sum_{n=0}^{\infty} [\tilde{f}_{1\to0}(s)\tilde{f}_{0\to1}(s)]^n \tilde{f}_{1\to2}(s)
$$

$$
= \frac{\tilde{f}_{1\to2}(s)}{1 - \tilde{f}_{1\to0}(s)\tilde{f}_{0\to1}(s)}. \tag{C1}
$$

The expression in the square brackets accounts for the progression through the cycle $G_{\text{on}} \to G_{\text{off}} \to G_{\text{on}}$ during which no nascent RNA is produced. The integer n denotes the number of such cycles before a nascent RNA is produced, and the summation goes over all $n \geqslant 0$.

The remaining problem is to compute $\tilde{f}_{0\to1}(s)$, $\tilde{f}_{1\to0}(s)$, and $\tilde{f}_{1\to2}(s)$. Since the waiting time in the off state is exponentially distributed, and there is only one reaction that leads from the off state, $\tilde{f}_{0\to1}(s)$ is given by

$$
\tilde{f}_{0\to1}(s) = \int_0^\infty k_{\text{on}} e^{-(s+k_{\text{on}})t} = \frac{k_{\text{on}}}{s + k_{\text{on}}}. \tag{C2}
$$

From the definitions of $\tilde{f}_{1\to0}(s)$ and $\tilde{f}_{1\to2}(s)$, it follows that

$$
f_{1\to0}(t) = P(1 \to 0)p(t), \tag{C3}
$$

$$
f_{1\to2}(t) = P(1 \to 2)p(t), \tag{C4}
$$

where $p(t)$ is the probability density function of the waiting time in the on state until the gene either switches to the off state or produces a nascent RNA, $P(1 \to 0)$ is the probability that the gene switches to off state, and $P(1 \to 2)$ is the probability that the gene produces a nascent RNA in the on state. We note that

$$
\int_0^\infty p(t)dt = 1, \quad P(1 \to 0) + P(1 \to 2) = 1. \tag{C5}
$$

To find $p(t)$, we denote by t_1 a random variable whose pdf is given by $f_1(t_1)$ and by t_2 a random variable whose pdf is given by $f_2(t)$, where $f_1(t_1)$ and $f_2(t_2)$ are given by

$$
f_1(t_1) = k_{\text{off}} e^{-k_{\text{off}}t}, \quad f_2(t_2) = f_{\text{ter,as}}(t_2), \tag{C6}
$$

and $f_{\text{ter,as}}(t)$ is given by Eq. (48). The waiting time t in the on state before either switching to the off state or producing a nascent RNA is the minimum of t_1 and t_2, $t = \min\{t_1, t_2\}$. Hence,

$$
P(t > x) = e^{-k_{\text{off}}} P_2(t_2 > x). \tag{C7}
$$

From here, we get $p(x)$ by taking the derivative of $P(t \leqslant x) = 1 - P(t > x)$,

$$
p(x) = e^{-k_{\text{off}}x}[k_{\text{off}} P_2(t_2 > x) + f_2(x)]. \tag{C8}
$$

The probability $P(1 \to 0)$ can be found from the condition that $t_1 < t_2$,

$$
P(1 \to 0) = P(t_1 < t_2) = \int_0^\infty dt_1 f_1(t_1) \int_{t_1}^\infty dt_2 f_2(t_2)
$$

$$
= k_{\text{off}} \int_0^\infty dt_1 e^{-k_{\text{off}}t_1} P_2(t_2 > t_1). \tag{C9}
$$

Similarly, $P(1 \to 2)$ follows from the condition that $t_2 < t_1$,

$$
P(1 \to 2) = P(t_2 < t_1) = \int_0^\infty dt_2 f_2(t_2) \int_{t_2}^\infty dt_1 f_1(t_1)
$$

$$
= \int_0^\infty dt_2 e^{-k_{\text{off}}t_2} f_2(t_2). \tag{C10}
$$

It is convenient to introduce $\lambda_1 = \omega v$ and $\lambda_2 = \omega(1 - v)$, in which case $f_{\text{ter,as}}(t)$ can be written as

$$
f_{\text{ter,as}}(t) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}(e^{-\lambda_2 t} - e^{-\lambda_1 t}). \tag{C11}
$$

Inserting this expression into the expressions for $p(t)$, $P(1 \to 0)$, and $P(1 \to 2)$ yields

$$
p(t) = \frac{\lambda_2(k_{\text{off}} + \lambda_1)}{\lambda_2 - \lambda_1} e^{-(k_{\text{off}}+\lambda_1)t}
$$
$$
- \frac{\lambda_1(k_{\text{off}} + \lambda_2)}{\lambda_2 - \lambda_1} e^{-(k_{\text{off}}+\lambda_2)t}, \tag{C12}
$$

$$
P(1 \to 0) = \frac{k_{\text{off}}(k_{\text{off}} + \lambda_1 + \lambda_2)}{(k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)}, \tag{C13}
$$

$$
P(1 \to 2) = \frac{\lambda_1 \lambda_2}{(k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)}. \tag{C14}
$$

From Eq. (C12), it follows that the Laplace transform of $p(t)$ is given by

$$
\tilde{p}(s) = \frac{k_{\text{off}}s + (k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)}{(k_{\text{off}} + \lambda_1 + s)(k_{\text{off}} + \lambda_2 + s)}. \tag{C15}
$$

Inserting $\tilde{p}(s)$ into $\tilde{f}_{1\to0}(s) = P(1 \to 0)\tilde{p}(s)$ and $\tilde{f}_{1\to2}(s) = P(1 \to 2)\tilde{p}(s)$, and then in the expression for $\tilde{h}(s)$ yields

$$
\tilde{h}(s) = \frac{A(s)}{B(s)}, \tag{C16}
$$

where $A(s)$ and $B(s)$ are given by

$$
A(s) = \lambda_1 \lambda_2(s + k_{\text{on}})[k_{\text{off}}s + (k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)], \tag{C17a}
$$

$$
\begin{aligned}
B(s) &= (k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)(s + k_{\text{off}} + \lambda_1)(s + k_{\text{off}} + \lambda_2) \\
&\quad \times (s + k_{\text{on}}) - k_{\text{on}}k_{\text{off}}(k_{\text{off}} + \lambda_1 + \lambda_2) \\
&\quad \times [k_{\text{off}}s + (k_{\text{off}} + \lambda_1)(k_{\text{off}} + \lambda_2)].
\end{aligned} \tag{C17b}
$$

This expression can be inverted to get $h(t)$ by finding the roots of $B(s)$ and performing the partial fraction decomposition.

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

[2] N. Friedman, L. Cai, and X. S. Xie, Linking Stochastic Dynamics to Population Distribution: An analytical Framework of Gene Expression, Phys. Rev. Lett. **97**, 168302 (2006).

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

[4] N. Kumar, T. Platini, and R. V. Kulkarni, Exact Distributions for Stochastic Gene Expression Models with Bursting and Feedback, Phys. Rev. Lett. **113**, 268105 (2014).

[5] Y. R. Lim, J.-H. Kim, S. J. Park, G.-S. Yang, S. Song, S.-K. Chang, N. K. Lee, and J. Sung, Quantitative Understanding of Probabilistic Behavior of Living Cells Operated by Vibrant Intracellular Networks, Phys. Rev. X **5**, 031014 (2015).

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

[7] L. Ham, R. D. Brackston, and M. P. H. Stumpf, Extrinsic Noise and Heavy-Tailed Laws in Gene Expression, Phys. Rev. Lett. **124**, 108101 (2020).

[8] C. Jia and R. Grima, Frequency Domain Analysis of Fluctuations of mRNA and Protein Copy Numbers Within a Cell Lineage: Theory and Experimental Validation, Phys. Rev. X **11**, 021032 (2021).

[9] A. Gupta and M. Khammash, Frequency spectra and the color of cellular noise, Nat. Commun. **13**, 4305 (2022).

[10] L.-H. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding, General properties of transcriptional time series in escherichia coli, Nat. Genet. **43**, 554 (2011).

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

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

[13] B. T. Donovan, A. Huynh, D. A. Ball, H. P. Patel, M. G. Poirier, D. R. Larson, M. L. Ferguson, and T. L. Lenstra, Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting, EMBO J. **38**, e100809 (2019).

[14] C. V. Rao, D. M. Wolf, and A. P. Arkin, Control, exploitation and tolerance of intracellular noise, Nature (London) **420**, 231 (2002).

[15] A. Singh and J. P. Hespanha, Optimal feedback strength for noise suppression in autoregulatory gene networks, Biophys. J. **96**, 4013 (2009).

[16] G. Balázsi, A. Van Oudenaarden, and J. J. Collins, Cellular decision making and biological noise: From microbes to mammals, Cell **144**, 910 (2011).

[17] R. A. Kellogg and S. Tay, Noise facilitates transcriptional control under dynamic inputs, Cell **160**, 381 (2015).

[18] J. Peccoud and B. Ycart, Markovian modeling of gene-product synthesis, Theoretical Population Biology **48**, 222 (1995).

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

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

[21] J. K. Kim and J. C. Marioni, Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data, Genome Biol. **14**, R7 (2013).

[22] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg, Genomic encoding of transcriptional burst kinetics, Nature (London) **565**, 251 (2019).

[23] S. Choubey, J. Kondev, and A. Sanchez, Deciphering transcriptional dynamics *in vivo* by counting nascent RNA molecules, PLoS Comput. Biol. **11**, e1004345 (2015).

[24] H. Xu, S. O. Skinner, A. M. Sokac, and I. Golding, Stochastic Kinetics of Nascent RNA, Phys. Rev. Lett. **117**, 128101 (2016).

[25] X. Fu, H. P. Patel, S. Coppola, L. Xu, Z. Cao, T. L. Lenstra, and R. Grima, Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions, Elife **11**, e82493 (2022).

[26] S. Choubey, Nascent RNA kinetics: Transient and steady-state behavior of models of transcription, Phys. Rev. E **97**, 022402 (2018).

[27] Q. Jiang, X. Fu, S. Yan, R. Li, W. Du, Z. Cao, F. Qian, and R. Grima, Neural network aided approximation and parameter inference of non-Markovian models of gene expression, Nat. Commun. **12**, 2618 (2021).

[28] S. Braichenko, J. Holehouse, and R. Grima, Distinguishing between models of mammalian gene expression: Telegraph-like models versus mechanistic models, J. R. Soc. Interface **18**, 20210510 (2021).

[29] J. Szavits-Nossan and R. Grima, Steady-state distributions of nascent RNA for general initiation mechanisms, Phys. Rev. Res. **5**, 013064 (2023).

[30] S. Klumpp, Pausing and backtracking in transcription under dense traffic conditions, J. Stat. Phys. **142**, 1252 (2011).

[31] T. Tripathi and D. Chowdhury, Interacting RNA polymerase motors on a DNA track: Effects of traffic congestion and intrinsic noise on RNA synthesis, Phys. Rev. E **77**, 011921 (2008).

[32] T. Tripathi and D. Chowdhury, Transcriptional bursts: A unified model of machines and mechanisms, Europhys. Lett. **84**, 68004 (2008).

[33] M. Voliotis, N. Cohen, C. Molina-París, and T. B. Liverpool, Fluctuations, pauses, and backtracking in DNA transcription, Biophys. J. **94**, 334 (2008).

[34] S. Klumpp and T. Hwa, Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination, Proc. Natl. Acad. Sci. USA **105**, 18159 (2008).

[35] S. Kim and C. Jacobs-Wagner, Effects of mRNA degradation and site-specific transcriptional pausing on protein expression noise, Biophys. J. **114**, 1718 (2018).

[36] T. Tripathi, G. M. Schütz, and D. Chowdhury, RNA polymerase motors: Dwell time distribution, velocity and dynamical phases, J. Stat. Mech.: Theory Exp. (2009) P08018.

[37] M. Dobrzyński and F. J. Bruggeman, Elongation dynamics shape bursty transcription and translation, Proc. Natl. Acad. Sci. USA **106**, 2583 (2009).

[38] A. S. Ribeiro, A. Häkkinen, S. Healy, and O. Yli-Harja, Dynamical effects of transcriptional pause-prone sites, Comput. Biol. Chem. **34**, 143 (2010).

[39] T. Rajala, A. Häkkinen, S. Healy, O. Yli-Harja, and A. S. Ribeiro, Effects of transcriptional pausing on gene expression dynamics, PLoS Comput. Biol. **6**, e1000704 (2010).

[40] D. Chowdhury, Stochastic mechano-chemical kinetics of molecular motors: A multidisciplinary enterprise from a physicist's perspective, Phys. Rep. **529**, 1 (2013).

[41] M. Sahoo and S. Klumpp, Backtracking dynamics of RNA polymerase: Pausing and error correction, J. Phys.: Condens. Matter **25**, 374104 (2013).

[42] T. Heberling, L. Davis, J. Gedeon, C. Morgan, and T. Gedeon, A mechanistic model for cooperative behavior of co-transcribing RNA polymerases, PLoS Comput. Biol. **12**, e1005069 (2016).

[43] J. Cholewa-Waclaw, R. Shah, S. Webb, K. Chhatbar, B. Ramsahoye, O. Pusch, M. Yu, P. Greulich, B. Waclaw, and A. P. Bird, Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic, Proc. Natl. Acad. Sci. USA **116**, 14995 (2019).

[44] M. Z. Ali, S. Choubey, D. Das, and R. C. Brewster, Probing mechanisms of transcription elongation through cell-to-cell variability of RNA polymerase, Biophys. J. **118**, 1769 (2020).

[45] J. Szavits-Nossan and B. Waclaw, Current-density relation in the exclusion process with dynamic obstacles, Phys. Rev. E **102**, 042117 (2020).

[46] T. W. Turowski, E. Petfalski, B. D. Goddard, S. L. French, A. Helwak, and D. Tollervey, Nascent transcript folding plays a major role in determining RNA polymerase elongation rates, Mol. Cell **79**, 488 (2020).

[47] G. Schütz and E. Domany, Phase transitions in an exactly soluble one-dimensional exclusion process, J. Stat. Phys. **72**, 277 (1993).

[48] B. Derrida, M. R. Evans, V. Hakim, and V. Pasquier, Exact solution of a 1D asymmetric exclusion model using a matrix formulation, J. Phys. A: Math. Gen. **26**, 1493 (1993).

[49] P. Tongaonkar, S. L. French, M. L. Oakes, L. Vu, D. A. Schneider, A. L. Beyer, and M. Nomura, Histones are required for transcription of yeast rRNA genes by RNA polymerase I, Proc. Natl. Acad. Sci. USA **102**, 10129 (2005).

[50] D. Schneider, S. French, Y. Osheim, A. Bailey, L. Vu, J. Dodd, J. Yates, A. Beyer, and M. Nomura, RNA polymerase II elongation factors Spt4p and Spt5p play roles in transcription elongation by RNA polymerase I and rRNA processing, Proc. Natl. Acad. Sci. USA **103**, 12707 (2006).

[51] J. A. Claypool, S. L. French, K. Johzuka, K. Eliason, L. Vu, J. A. Dodd, A. L. Beyer, and M. Nomura, Tor pathway regulates Rrn3p-dependent recruitment of yeast RNA polymerase I to the promoter but does not participate in alteration of the number of active genes, Mol. Biol. Cell **15**, 946 (2004).

[52] C. T. MacDonald, J. H. Gibbs, and A. C. Pipkin, Kinetics of biopolymerization on nucleic acid templates, Biopolymers **6**, 1 (1968).

[53] J. D. C. Little, A proof for the queuing formula, Oper. Res. **9**, 383 (1961).

[54] V. Elgart, T. Jia, and R. V. Kulkarni, Applications of Little's law to stochastic models of gene expression, Phys. Rev. E **82**, 021901 (2010).

[55] B. Derrida and M. R. Evans, Exact correlation functions in an asymmetric exclusion model with open boundaries, J. Phys. I France **3**, 311 (1993).

[56] B. Derrida, E. Domany, and D. Mukamel, An exact solution of a one-dimensional asymmetric exclusion model with open boundaries, J. Stat. Phys. **69**, 667 (1992).

[57] C. Cianci, S. Smith, and R. Grima, Molecular finite-size effects in stochastic models of equilibrium chemical systems, J. Chem. Phys. **144**, 084101 (2016).

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

[59] D. G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, Ann. Math. Stat. **24**, 338 (1953).

[60] L. Takács, On a coincidence problem concerning telephone traffic, Acta Math. Acad. Sci. Hung. **9**, 45 (1958).

[61] M. Krbálek and P. Hrabák, Interparticle gap distribution and spectral rigidity of the totally asymmetric simple exclusion process with open boundaries, J. Phys. A: Math. Theor. **44**, 175203 (2011).

[62] K. Ghosh, A. Majumdar, and D. Chowdhury, Distribution of time-headways in a particle-hopping model of vehicular traffic, Phys. Rev. E **58**, 4012 (1998).

[63] P. Hrabák, Time-headway distribution for random-sequential-update TASEP with periodic and open boundaries, J. Traff. Transport. Eng. **7**, 30 (2020).

[64] A. Piovesan, M. Caracausi, F. Antonaros, M. C. Pelleri, and L. Vitale, GeneBase 1.1: A tool to summarize data from NCBI Gene datasets and its application to an update of human gene statistics, Database **2016**, baw153 (2016).

[65] D. Cox, *Renewal Theory* (Methuen, London, 1967).

[66] T. Filatova, N. Popovic, and R. Grima, Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation, elongation, pausing, and termination, Bull. Math. Biol. **83**, 3 (2021).

[67] P. Thomas, N. Popović, and R. Grima, Phenotypic switching in gene regulatory networks, Proc. Natl. Acad. Sci. USA **111**, 6994 (2014).

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

[69] I. Jonkers, H. Kwak, and J. T. Lis, Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons, eLife **3**, e02407 (2014).

[70] J. Szavits-Nossan and R. Grima, Mean-field theory accurately captures the variation of copy number distributions across the mRNA lifecycle, Phys. Rev. E **105**, 014410 (2022).

[71] A. Nagar, A. Valleriani, and R. Lipowsky, Translation by ribosomes with mRNA degradation: Exclusion processes on aging tracks, J. Stat. Phys. **145**, 1385 (2011).

[72] S. Karthika and A. Nagar, Totally asymmetric simple exclusion process with resetting, J. Phys. A: Math. Theor. **53**, 115003 (2020).

[73] J. Lloyd-Price, S. Startceva, V. Kandavalli, J. G. Chandraseelan, N. Goncalves, S. M. Oliveira, A. Häkkinen, and A. S. Ribeiro, Dissecting the stochastic transcription initiation process in live *Escherichia coli*, DNA Res. **23**, 203 (2016).

[74] D. E. Weidemann, J. Holehouse, A. Singh, R. Grima, and S. Hauf, The minimal intrinsic stochasticity of constitutively expressed eukaryotic genes is sub-Poissonian, Sci. Adv. **9**, eadh5138 (2023).

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

[76] C. Fritzsch, S. Baumgärtner, M. Kuban, D. Steinshorn, G. Reid, and S. Legewie, Estrogen-dependent control and cell-to-cell variability of transcriptional bursting, Mol. Syst. Biol. **14**, e7678 (2018).

[77] S. L. French and O. L. Miller, Transcription mapping of the *Escherichia coli* chromosome by electron microscopy, J. Bacteriol. **171**, 4207 (1989).

[78] O. Padovan-Merhar, G. Nair, A. Biaesch, A. Mayer, S. Scarfone, S. Foley, A. Wu, L. Churchman, A. Singh, and A. Raj, Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms, Mol. Cell **58**, 339 (2015).

[79] T. Zhou and J. Zhang, Analytical results for a multistate gene model, SIAM J. Appl. Math. **72**, 789 (2012).

[80] W. Shao and J. Zeitlinger, Paused RNA polymerase II inhibits new transcriptional initiation, Nat. Genet. **49**, 1045 (2017).

[81] V. Epshtein and E. Nudler, Cooperation between RNA polymerase molecules in transcription elongation, Science **300**, 801 (2003).

[82] S. Kim, B. Beltran, I. Irnov, and C. Jacobs-Wagner, Long-distance cooperative and antagonistic RNA polymerase dynamics via DNA supercoiling, Cell **179**, 106 (2019).

[83] P. Chatterjee, N. Goldenfeld, and S. Kim, DNA Supercoiling Drives a Transition Between Collective Modes of Gene Synthesis, Phys. Rev. Lett. **127**, 218101 (2021).

[84] S. Tripathi, S. Brahmachari, J. N. Onuchic, and H. Levine, DNA supercoiling-mediated collective behavior of co-transcribing RNA polymerases, Nucleic Acids Res. **50**, 1269 (2022).

[85] A. Klindziuk and A. B. Kolomeisky, Long-range supercoiling-mediated RNA polymerase cooperation in transcription, J. Phys. Chem. B **125**, 4692 (2021).

[86] J. Szavits-Nossan and M. R. Evans, Dynamics of ribosomes in mRNA translation under steady- and nonsteady-state conditions, Phys. Rev. E **101**, 062404 (2020).

[87] J. de Gier and F. H. L. Essler, Bethe Ansatz Solution of the Asymmetric Exclusion Process with Open Boundaries, Phys. Rev. Lett. **95**, 240601 (2005).