Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments

Gennady Gorin[a,1], **John J. Vastola**[b,1], **Meichen Fang**[c], **and Lior Pachter**[c,d,2]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; [b]Department of Neurobiology, Harvard Medical School, Boston, MA 02115; [c]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125; [d]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125

The question of how cell-to-cell differences in transcription rate affect RNA count distributions is fundamental for understanding biological processes underlying transcription. We argue that answering this question requires quantitative models that are both interpretable (describing concrete biophysical phenomena) and tractable (amenable to mathematical analysis). This enables the identification of experiments which best discriminate between competing hypotheses. As a proof of principle, we introduce a simple but flexible class of models involving a stochastic transcription rate coupled to a discrete stochastic RNA transcription and splicing process, and compare and contrast two biologically plausible hypotheses about transcription rate variation. One assumes variation is due to DNA experiencing mechanical strain, while the other assumes it is due to regulator number fluctuations. Although biophysically distinct, these models are mathematically similar, and we show they are hard to distinguish without comparing whole predicted probability distributions. Our work illustrates the importance of theory-guided data collection, and introduces a general framework for constructing and solving mathematically nontrivial continuous–discrete stochastic models.

Chemical master equation | Markov modeling | mRNA transcription | Stochastic differential equations | Extrinsic noise

Single-cell RNA counts fluctuate due to a combination of dynamic processes in living cells, such as DNA supercoiling, gene regulation, and RNA processing; however, it is unclear how much we can learn about these processes' kinetics and relative importance from counts alone. By generating enormous amounts of single-cell data, modern transcriptomics has the potential to shed light on such fundamental aspects of transcription on a genome-wide scale. However, the field's standard data-driven and phenomenological analyses are descriptive: even though they can summarize data, they do not make specific claims about the mechanisms that generated it. To make mechanistic sense of measurements of gene expression and submolecular features in thousands of single cells at a time (1–4), we seek a framework for systematically distinguishing different plausible hypotheses about transcription.

In principle, models of transcription that are both interpretable and tractable would allow us to be more hypothesis-driven. Interpretability means fitting model parameters conveys clear biological information about the kinetics of microscopic phenomena. Tractability means a thorough mathematical analysis of model behavior is possible. These properties enable a 'rational' design of transcriptomic experiments (Figure 1a), analogous to ideas about rational drug design (5–9) and the optimal design of single-cell experiments (10–13), since one can mathematically determine the kind of experiment that best distinguishes two such models.

The common *post-hoc* approach of fitting negative binomial-like distributions to RNA count data (14–18) is mathematically tractable, but not biologically interpretable. On the other hand, detailed mathematical models of transcription (19–27) are certainly interpretable, but tend not to be tractable: complexity makes a thorough analysis challenging, and identifiability

Significance Statement

The interpretation of transcriptomic observations requires detailed models of biophysical noise that can be compared and fit to experimental data. Models of *intrinsic* noise, describing stochasticity in molecular reactions, and *extrinsic* noise, describing cell-to-cell variation, are particularly common. However, integrating and solving them is challenging, and previous results are largely limited to summary statistics. We examine two mechanistically grounded stochastic models of transcriptional variation and demonstrate that (1) well-known regimes naturally emerge in limiting cases, and (2) the choice of noise model significantly affects the RNA distributions, but not the lower moments, offering a route to model identification and inference. This approach provides a simple and biophysically interpretable means to construct and unify models of transcriptional variation.

J.J.V. and G.G. conceived of the work, derived the mathematical results, and drafted the manuscript. G.G, M.F., and J.J.V. worked on simulating the models and numerically implementing their analytic solutions. L.P. supervised the work. All authors reviewed and edited the manuscript.

No competing interests.

[1] G.G. and J.J.V. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: lpachter@caltech.edu

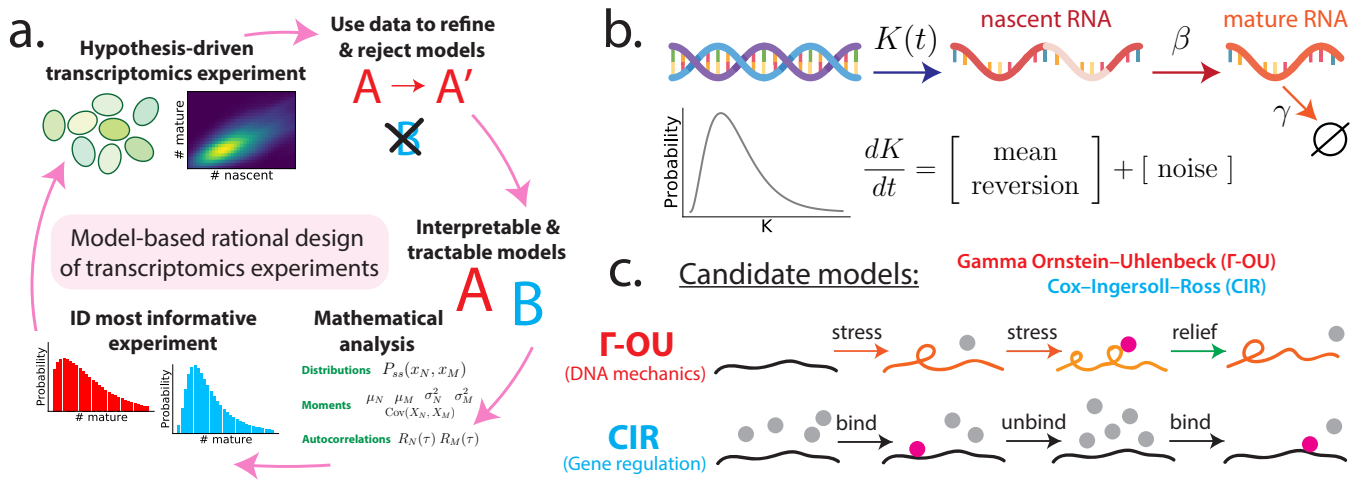

Fig. 1. Framework for the rational design of transcriptomics experiments. **a.** Model-based closed loop paradigm. A researcher begins by representing two or more competing hypotheses as interpretable and tractable mathematical models (middle right of circle). Next, they perform a detailed mathematical analysis of each model, computing quantities (e.g., RNA count distributions and moments) that can help distinguish one hypothesis from another. Using the results of that analysis as input, they identify the experiment that best distinguishes the two models. Finally, they perform this experiment on some population of cells, use the resulting data to refine and/or reject models, and repeat the process with an updated ensemble of models. **b.** Interpretable and tractable modeling framework for transcription rate variation. We consider stochastic models of transcription involving (i) nascent/unspliced RNA, (ii) mature/spliced RNA, and (iii) a stochastic and time-varying transcription rate $K(t)$. The transcription rate is assumed to evolve in time according to a simple, one-dimensional SDE that includes a mean-reversion term (which tends to push $K(t)$ towards its mean value) and a noise term (which causes $K(t)$ to randomly fluctuate). Here, we have specifically chosen dynamics for which the long-time probability distribution of $K(t)$ is a gamma distribution (gray curve), because this assumption yields empirically plausible negative binomial-like RNA distributions. However, the framework does not require this in general. **c.** Two plausible models studied in this paper. The Γ-OU model describes DNA mechanics, whereas the CIR model describes regulation by a high-copy number regulator.

issues mean that it can be difficult or impossible to use the data one has to distinguish competing hypotheses.

In this paper, we propose a class of interpretable and tractable transcription models that is fairly simple, yet flexible enough to account for a range of biological phenomena. It assumes that a stochastic and time-varying transcription rate drives a discrete stochastic RNA transcription and splicing process. This model class incorporates both intrinsic noise (randomness associated with the timing of events like transcription and degradation) and extrinsic noise (due to cell-to-cell differences) (28–32) in a principled way, with the latter due to transcription rate variation. We focus on two specific examples of models from this class, which assume variation is due to (i) random changes in the mechanical state of DNA, or (ii) random changes in the number of an abundant regulator.

We find that these models, although mathematically similar, yield different predictions, answering our initial question in the affirmative: the fine details of transcription *can*, at least some of the time, be inferred from transcription rate variation. This is because the details of *how* the transcription rate fluctuates (i.e., its dynamics), rather than just the steady-state distribution of those fluctuations, can qualitatively affect model predictions. We also find that a naïve approach to distinguishing between them fails, and that comparing whole distributions far outperforms other approaches. While we will not actually *implement* the entire closed loop paradigm depicted in Figure 1a, our work constructs one possible mathematical and computational foundation for it.

Results

Transcription rate variation accounts for empirically observed variance. If we would like to understand and fit available transcriptomic data—especially *multimodal* data sets that report the numbers of both nascent and mature transcripts inside single cells (33, 34)—what kind of models of transcription should we consider? Given that single cell RNA counts are often low, we would like our models to be able to account for the production, processing, and degradation of individual RNA molecules. From experiments in living cells, these processes are known to be random (35). Crucially, the molecule counts are low enough that the variation in molecule numbers should be explicitly described by a stochastic model.

The theoretical framework associated with the chemical master equation (CME) (36–42) can be used to define discrete and stochastic models of cellular processes. The *constitutive* model of transcription, which assumes RNA is produced at a constant rate, is one particularly simple and well-studied example. It can be defined via the chemical reactions

$$\varnothing \xrightarrow{K} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing \qquad [1]$$

where $\mathcal{N}$ denotes nascent RNA, $\mathcal{M}$ denotes mature RNA, K is the transcription rate, β is the splicing rate, and γ is the degradation rate. It predicts (43) that the long-time probability $P_{ss}^{con}(x_N, x_M)$ of observing $x_N \in \mathbb{N}_0$ nascent RNA and $x_M \in \mathbb{N}_0$ mature RNA in a single cell is Poisson, so that

$$P_{ss}^{con}(x_N, x_M) = \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!} \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!}. \qquad [2]$$

While mathematically tractable, a model like this is too simple to fit existing data. Most observed eukaryotic RNA count distributions are 'overdispersed': they have a higher variance than Poisson distributions with the same mean (44).

One way to account for overdispersion is to assume that different cells in a population have different transcription rates, but that each individual cell otherwise follows the constitutive model. For various choices of transcription rate distribution, one can obtain results that look much closer to eukaryotic transcriptomic data. For example, one reasonable choice (which has been explored by other authors (45)) is to assume that the transcription rate K is gamma-distributed with shape parameter α and scale parameter θ, i.e. $K \sim \Gamma(\alpha, \theta)$. The long-time/steady-state probability of observing x_N nascent and x_M mature RNA would then be described by the 'mixture'

$$P_{ss}^{mix}(x_N, x_M) = \int_0^\infty dK \; \frac{K^{\alpha-1} \; e^{-K/\theta}}{\theta^\alpha \; \Gamma(\alpha)} \; P_{ss}^{con}(x_N, x_M). \qquad [3]$$

The marginal distributions of this joint distribution will be negative binomial rather than Poisson, allowing us to actually fit observed single-cell data. But this approach—which is equivalent to the *post-hoc* fitting of negative binomial distributions—is not biophysically interpretable. What is the biological meaning of the parameters α and θ? And why do different cells have different transcription rates? Is it really reasonable to assume, as we have here, that these rates are 'frozen', and remain as they are for all time in a given cell?

Interpretable and tractable modeling framework for transcription rate variation. We propose a class of transcriptional models that balance interpretability and tractability, and generalize the mixture model. Although various biological details underlying transcription may be complicated, we assume they can be captured by an effective transcription rate $K(t)$ which is stochastic and varies with time. This transcription rate randomly fluctuates about its mean value, with the precise nature of its fluctuations dependent upon the fine biophysical details of transcription. Mathematically, we assume that $K(t)$ is a continuous stochastic process described by an (Itô-interpreted) stochastic differential equation (SDE)

$$\begin{aligned} \dot{K}(t) &= [\text{ mean reversion }] + [\text{ noise }] \\ &= A - BK(t) + [\text{ noise }] \end{aligned} \qquad [4]$$

for some coefficients A and B, where [noise] denotes a model-dependent term that introduces stochastic variation. The transcription rate $K(t)$ is coupled to RNA dynamics as in the constitutive model:

$$\varnothing \xrightarrow{K(t)} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing. \qquad [5]$$

This reaction list defines a master equation model that couples discrete stochastic RNA dynamics to the continuous stochastic process $K(t)$ (Figure 1b). Although this model class is not completely realistic (for example, there is no feedback), it is fairly flexible, and can recapitulate empirically plausible negative binomial-like RNA count distributions. To guarantee this, we will specifically consider candidate models for which the steady-state distribution of $K(t)$ is a gamma distribution.

Other kinds of transcriptional models can also be viewed as special cases of this model class. The constitutive model is a degenerate case that arises from the limit of no noise and fast mean-reversion. We will see later that the popular bursting model of RNA production, which describes intermittent production of multiple nascent transcripts at a time (1, 46–49) is also a degenerate case. For the rest of this paper, we examine two specific cases of this model class more closely: the gamma Ornstein–Uhlenbeck (Γ-OU) model and Cox–Ingersoll–Ross (CIR) model, which are depicted in Figure 1c. In particular, we will motivate the underlying biophysics, solve the models, outline major similarities and differences, and discuss how and when they can be distinguished given transcriptomic data.

A. Gamma Ornstein–Uhlenbeck production rate model. Transcription rate variation may emerge due to mechanical changes in DNA that make producing RNA more or less kinetically favorable. Each nascent RNA produced by an RNA polymerase induces a small amount of mechanical stress/supercoiling in DNA, which builds over time and can mechanically frustrate transcription unless it is relieved. Because topoisomerases arrive to relieve stress, there is a dynamic balance between transcription-mediated stress and topoisomerase-mediated recovery, models of which can recapitulate gene overdispersion and bursting (23, 25).

We can simplify the detailed mechanistic model of Sevier, Kessler, and Levine while retaining crucial qualitative aspects. Assuming that the transcription rate $K(t)$ is proportional to how mechanically relaxed DNA is, that relaxation continuously decreases due to transcription-associated events, and that topoisomerases randomly arrive to increase relaxation, we find (see Section S3.2.1) that $K(t)$ can be modeled by the SDE

$$\dot{K}(t) = -\kappa K(t) + \epsilon(t), \qquad [6]$$

where $\epsilon(t)$ is an infinitesimal Lévy process (a compound Poisson process with arrival frequency a and exponentially distributed jumps with expected size θ) capturing random topoisomerase arrival. This is the gamma Ornstein–Uhlenbeck (Γ-OU) model of transcription (50). It naturally emerges from a biomechanical model with two opposing effects: the continuous mechanical frustration of DNA undergoing transcription, which is a first-order process with relaxation rate κ, and the stochastic relaxation by topoisomerases that arrive at rate a. The scaling between the relaxation state and the transcription rate is set by a gain parameter θ.

The Γ-OU model is perhaps better known in finance applications, where it has been used to model the stochastic volatility of the prices of stocks and options (51–54). Its utility as a financial model is largely due to its ability to capture asset behavior that deviates from that of commonly used Gaussian Ornstein–Uhlenbeck models, such as skewness and frequent price jumps.

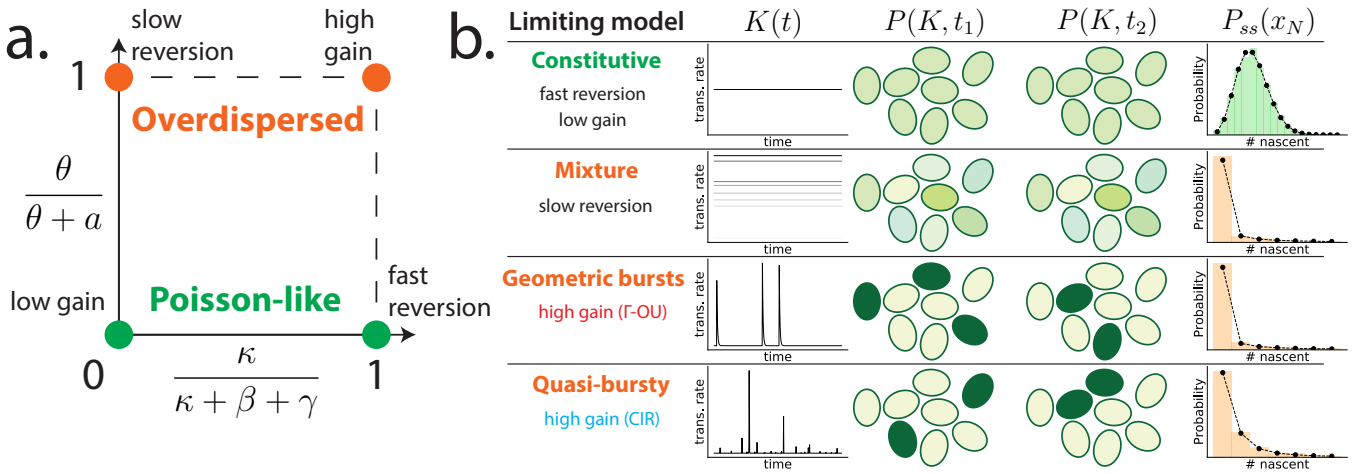

Fig. 2. Summary of the qualitative behavior of the Γ-OU and CIR models. **a.** Qualitative behavior can be visualized in a two-dimensional parameter space, with $\kappa/(\kappa+\beta+\gamma)$ on one axis and the gain ratio $\theta/(\theta+a)$ on the other. The four limits discussed in the text correspond to the four corners of this space. When $a \gg \theta$, we obtain Poisson-like behavior (green). When $a \ll \theta$, we obtain overdispersed distributions (orange). **b.** Dynamics of limiting models. Both the Γ-OU and CIR models reduce to the constitutive model in the fast reversion and low gain limits, where the transcription rate $K(t)$ is effectively constant in time and identical for all cells in the population. Both reduce to the mixture model in the slow reversion limit, so that $K(t)$ is inhomogeneous across the population but constant in time for individual cells. In the high gain limit, the Γ-OU and CIR models yield different heavy-tailed distributions, with the CIR limiting model appearing to be novel. In both cases, $K(t)$ exhibits sporadic large fluctuations within single cells.

B. Cox–Ingersoll–Ross production rate model. Alternatively, transcription rate variation may be due to fluctuations in the concentration of some regulator molecule, which affects RNA transcription without getting consumed by it (e.g., RNA polymerases, inducers, and activators). The following reaction list ($\mathcal{N}$: nascent RNA, $\mathcal{R}$: regulator) crudely models this idea:

$$\begin{aligned} \varnothing &\xrightarrow{a} \mathcal{R} \xrightarrow{\kappa} \varnothing \\ \mathcal{R} &\xrightarrow{\theta} \mathcal{R} + \mathcal{N}, \end{aligned} \qquad [7]$$

where a is the $\mathcal{R}$ production rate, κ is the $\mathcal{R}$ degradation rate, and θ is the 'gain' relating the number of regulator molecules to the rate of transcription. If the number of regulator molecules $r(t)$ is very large, we can accurately approximate regulator dynamics as a continuous stochastic process using the framework associated with the chemical Langevin equation (38, 55). The effective transcription rate $K(t) := \theta\, r(t)$ satisfies the SDE

$$\dot{K}(t) = a\theta - \kappa K(t) + \sqrt{2\kappa\theta K(t)}\; \xi(t) \qquad [8]$$

where $\xi(t)$ is a Gaussian white noise term (see Section S3.3.1). This is the Cox–Ingersoll–Ross (CIR) model of transcription (50).

Because the CIR model remains agnostic about the precise mechanism by which $K(t)$ depends on regulator concentration, it can be used to represent various more biophysically detailed hypotheses. Consider the dynamics of a regulator $\mathcal{R}$ that activates a promoter $\mathcal{G}$ by binding to it:

$$\begin{aligned} \varnothing &\xrightarrow{a} \mathcal{R} \xrightarrow{\kappa} \varnothing \\ \mathcal{G}_{off} + \mathcal{R} &\underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} \mathcal{G}_{on} \\ \mathcal{G}_{on} &\xrightarrow{k_{ini}} \mathcal{G}_{on} + \mathcal{N}, \end{aligned} \qquad [9]$$

where k_{on} and k_{off} describe the binding/unbinding kinetics, and k_{ini} is the promoter initiation rate. If the binding of $\mathcal{R}$ to the promoter is sufficiently rapid and weak, the effective transcriptional dynamics are described by the CIR model with $\theta = k_{ini}k_{on}/k_{off}$ (see Section S3.3.1).

Although the CIR model is most familiar as a description of interest rates in quantitative finance (56–58), it has been previously used to describe biochemical input variation based on the CLE, albeit with less discussion of the theoretical basis and limits of applicability (59–62).

The models are interpretable and unify known results. Qualitatively, the distribution shapes predicted by the Γ-OU and CIR models interpolate between Poisson and negative binomial-like extremes, with behavior controlled mostly by two of the transcription noise parameters: the mean-reversion rate κ and the gain parameter θ (Figure 2a). Remarkably, where one is in this landscape of qualitative behavior is independent of the mean transcription rate $\langle K \rangle = a\theta/\kappa$, since a can vary to accommodate any changes in κ or θ. It is also independent of the steady-state distribution of transcription rates, which is the

same (i.e., $\Gamma(a/\kappa, \theta)$) in all cases. We find that the details of how the transcription rate fluctuates in time strongly impact the shape of RNA count distributions, a fact which may have previously gone underappreciated.

When κ is very fast, the transcription rate very quickly reverts to its mean value whenever it is perturbed, so it is effectively constant, and we recover the constitutive model. When κ is very slow, the transcription rates of individual cells appear 'frozen' on the time scales of RNA dynamics, and we recover the mixture model discussed earlier. When θ is very small, fluctuations in underlying biological factors (DNA relaxation state or regulator concentration) are significantly damped, so $K(t)$ is also effectively constant in this case.

Interestingly, while the two models agree in the aforementioned limits, their predictions markedly differ in the large θ limit, where fluctuations are amplified and predicted count distributions become increasingly overdispersed. The Γ-OU model predicts that nascent RNA is produced in geometrically distributed bursts in this limit, recapitulating the conventional model of bursty gene expression (35, 49). However, the CIR model predicts a novel family of count distributions with heavier tails than their Γ-OU counterparts. The difference is shown in Figure S4. This deviation is a consequence of state-dependent noise: while the number of topoisomerases which arrive to relieve stress does not depend on the current relaxation state of the DNA, birth-death fluctuations in the number of regulators tend to be greater when there are more regulator molecules present. We illustrate the four limiting regimes of interest in Figure 2b, present their precise quantitative forms in Section S2.5, and derive them in Section S5.

Another lens through which to view qualitative behavior is the coefficient of variation ($\eta^2 := \sigma^2/\mu^2$), which quantifies the amount of 'noise' in a system. We find (see Section S2.4.2), consistent with previous results (29, 30, 32), that the total noise can be written as a sum of 'intrinsic' (due to the stochasticity inherent in chemical reactions) and 'extrinsic' (due to transcription rate variation) contributions. For both models,

$$\eta_N^2 = \frac{1}{\mu_N} + \frac{\theta}{\langle K \rangle} \frac{1/\kappa}{1/\kappa + 1/\beta}$$
$$\eta_M^2 = \frac{1}{\mu_M} + \frac{\theta}{\langle K \rangle} \frac{1/\kappa}{1/\kappa + 1/\beta} \frac{1/\kappa}{1/\kappa + 1/\gamma} \frac{1/\kappa + 1/(\beta+\gamma)}{1/\kappa}$$

where η_N^2 and η_M^2 quantify the amount of noise in nascent and mature RNA counts, and μ_N and μ_M denote the average number of nascent and mature RNA. In the 'overdispersed' regimes, where θ is large or κ is small, the extrinsic noise contributions become significant. These results are exact.

The models are analytically tractable. Using a suite of novel theoretical approaches—including path integral methods, generating function computations, a correspondence between the Poisson representation of the CME and SDEs, and tools from the mathematics of stochastic processes—we were able to exactly solve the Γ-OU and CIR models. This includes computing all steady-state probability distributions $P_{ss}(x_N, x_M)$, first order moments, second order moments, and autocorrelation functions.

A central idea in all of our calculations is to consider transforms of the probability distribution—variants of the generating function—instead of the distribution itself. Once a generating function is available, the PMF can be computed by computationally inexpensive Fourier inversion. The joint generating function $\psi(g_N, g_M, h, t)$ is defined as

$$\psi := \sum_{x_N=0}^{\infty} \sum_{x_M=0}^{\infty} \int_0^{\infty} dK \; g_N^{x_N} g_M^{x_M} e^{ihK} \; P(x_N, x_M, K, t) \qquad [10]$$

with $g_N, g_M \in \mathbb{C}$ both on the complex unit circle, $h \in \mathbb{R}$, and $P(x_N, x_M, K, t)$ encoding the probability density over counts *and* transcription rates. As these rates are not usually observable, and the previous body of work treats stationary distributions, we are most interested in $\psi_{ss}(g_N, g_M)$, the probability-generating function (PGF) of $P_{ss}(x_N, x_M)$. We find it most convenient to report our results in terms of $\phi_{ss}(u_N, u_M) := \log \psi_{ss}(g_N, g_M)$, the log of the PGF with an argument shift $u_N := g_N - 1$ and $u_M := g_M - 1$.

The solution of the Γ-OU model is

$$\phi_{ss}(u_N, u_M) = \langle K \rangle \int_0^{\infty} \frac{U_0(s; u_N, u_M)}{1 - \frac{\theta}{\kappa} U_0(s; u_N, u_M)} ds \qquad [11]$$

where $U_0(s; u_N, u_M)$ is obtained by solving the characteristic ODEs obtained from the generating function (63):

$$\begin{aligned} \frac{dU_2}{ds} &= -\gamma \; U_2, & U_2(0) &= u_M, \\ \frac{dU_1}{ds} &= \beta \; (U_2 - U_1), & U_1(0) &= u_N, \\ \frac{dU_0}{ds} &= \kappa \; (U_1 - U_0), & U_0(0) &= 0. \end{aligned} \qquad [12]$$

This system of linear first-order ODEs can be solved analytically (48), and the generating function can be obtained by quadrature. The solution to the CIR model is

$$\phi_{ss}(u_N, u_M) = \langle K \rangle \int_0^{\infty} U_0(s; u_N, u_M) ds \qquad [13]$$

where $U_0(s; u_N, u_M)$ is obtained from analogous ODEs:

$$\begin{aligned} \frac{dU_2}{ds} &= -\gamma\, U_2, & U_2(0) &= u_M, \\ \frac{dU_1}{ds} &= \beta\, (U_2 - U_1), & U_1(0) &= u_N, \\ \frac{dU_0}{ds} &= \kappa(U_1 - U_0) + \theta\, U_0^2, & U_0(0) &= 0. \end{aligned} \qquad [14]$$

While the above ODEs have an exact solution (64), it is cumbersome, and preferable to evaluate numerically. We derive these solutions in Section S3, and validate them against stochastic simulations in Section S6.

Summary statistics cannot distinguish between the models. The tractability of these two models allows us to analytically compute common (steady-state) summary statistics. Despite the models' distinct biological origins, their means (μ_N and μ_M), variances (σ_N^2 and σ_M^2), covariances, and autocorrelation functions ($R_N(\tau)$ and $R_M(\tau)$) match exactly (Table 1; see Section S4). This means that such summary statistics cannot be used as the basis for model discrimination. More fundamentally, it implies that experimental technologies that only report averages—such as RNA sequencing without single-cell resolution—cannot possibly distinguish between noise models.

Table 1. Molecular distribution moments

Moment	Value
$\langle K \rangle$	$\frac{a\theta}{\kappa}$
μ_N	$\langle K \rangle / \beta$
μ_M	$\langle K \rangle / \gamma$
$\sigma_N^2 - \mu_N$	$\frac{\mu_N \theta}{\kappa+\beta}$
$\sigma_M^2 - \mu_M$	$\frac{\mu_M \theta}{\kappa+\gamma} \cdot \frac{\beta}{\kappa+\beta} \cdot \frac{\kappa+\beta+\gamma}{\beta+\gamma}$
$\mathsf{Cov}(X_N, K)$	$\frac{\langle K \rangle \theta}{\kappa+\beta}$
$\mathsf{Cov}(X_M, K)$	$\frac{\langle K \rangle \theta}{\kappa+\gamma} \cdot \frac{\beta}{\kappa+\beta}$
$\mathsf{Cov}(X_N, X_M)$	$\frac{\langle K \rangle \theta}{(\kappa+\beta)(\kappa+\gamma)} \cdot \frac{\kappa+\beta+\gamma}{\beta+\gamma}$
$R_N(\tau)$	$e^{-\beta\tau} + \frac{\mathsf{Cov}(X_N, K)}{\sigma_N^2} \frac{\left[e^{-\kappa\tau} - e^{-\beta\tau}\right]}{\beta - \kappa}$
$R_M(\tau)$	$e^{-\gamma\tau} + \beta \frac{\mathsf{Cov}(X_N, X_M)}{\sigma_M^2} \frac{\left[e^{-\beta\tau} - e^{-\gamma\tau}\right]}{\gamma - \beta} + \beta \frac{\mathsf{Cov}(X_M, K)}{\sigma_M^2}$ $\times \left[\frac{e^{-\beta\tau}}{(\beta-\gamma)(\beta-\kappa)} + \frac{e^{-\gamma\tau}}{(\gamma-\beta)(\gamma-\kappa)} + \frac{e^{-\kappa\tau}}{(\kappa-\beta)(\kappa-\gamma)}\right]$

Models can be distinguished using multimodal count data. The lower moments cannot distinguish between these two models even in principle. But even if this were not the case, the shortcomings of using them to perform model discrimination are becoming increasingly clear (65). Does the situation improve if we compare whole count distributions? Our analytic solutions (Eq. 11 and Eq. 13) show that they are in principle discriminable.

To establish that the Γ-OU and CIR models can indeed be distinguished using whole count distributions, we first generated noise-free synthetic data from the CIR model for many different parameter sets, and computed Bayes factors for each one (the probability of the data given the CIR model, divided by the probability of the data given the Γ-OU model), assuming we know the model parameters. We found that, if we plot log Bayes factors (averaged over many synthetic data sets) in the same space that depicts the models' qualitative regimes (see Figure 2a), the models are strongly distinguishable for large swaths of parameter space if one has on the order of 1000 data points (Figure 3a), but not fewer. They are easier to distinguish if the gain ratio $\theta/(\theta + a)$ is larger, and if one uses both nascent and mature count data.

How much does one gain from using multimodal data instead of nascent counts only? Given a data set of 1000 cells, distinguishability can improve by about an order of magnitude (Figure 3b) on average. To verify that the Γ-OU and CIR models are not distinguishable for trivial reasons (for example: one becomes Poisson-like, while the other does not), we checked whether each model is separately discriminable from the constitutive and mixture models (Figure 3c). We found extremely high Bayes factors when comparing against Poisson distributions, meaning these two models should almost never be confused for Poisson distributions. We also found that both models are strongly discriminable from the static mixture model.

How different are the predictions of the Γ-OU and CIR models in the most optimistic case? We used a gradient descent search to identify a parameter set that maximized the KL divergence between the Γ-OU and CIR joint count distributions, and found that even when predictions are maximally divergent, the distributions are still visually alike (Figure 3d). This suggests that, while probabilistic inference using whole distributions may succeed at performing model discrimination, many naïve approaches may fail.

Even if one can distinguish between the two noise models, can one precisely infer biophysically interpretable noise model parameters like κ and θ, or are the two models 'sloppy' with respect to them? Using the Python package PyMC3, and a

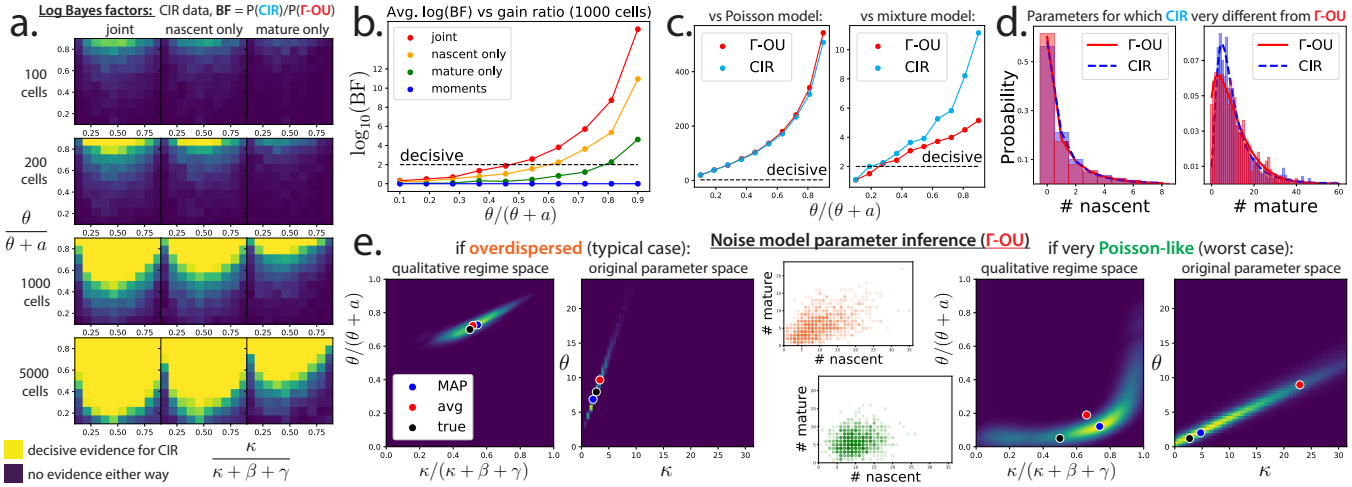

Fig. 3. Model distinguishability and parameter inference. **a.** Log Bayes factors show models are distinguishable in most of parameter space. Given $[\langle K\rangle = 5, \beta = 1, \gamma = 1.5]$, 100 noise-free synthetic data sets were sampled for each point on a 10×10 grid covering the qualitative parameter space. For each parameter set, $\log_{10}(P(\text{data}|\text{CIR})/P(\text{data}|\Gamma\text{-OU}))$ was computed. Plotted are the average log Bayes factors capped at 2 (a common threshold for decisive evidence in favor of one model). This procedure was done for data sets containing 100, 200, 1000, and 5000 cells, assuming nascent data was used to compute the Bayes factors, mature data was used, or both were used. As expected, distinguishability (i.e., higher Bayes factors) increases as the number of cells increases, and as the gain ratio $\theta/(\theta + a)$ increases. **b.** Models are often strongly distinguishable. A slice of the 1000 cell row of the previous plot (without the cap at 2) for a moderate value of $\kappa/(\kappa + \beta + \gamma)$. Using both nascent and mature data is better than using either individually, usually by at least an order of magnitude. **c.** It is easy to distinguish the Γ-OU and CIR models from trivial models. Same axes as in b. Discriminability of Γ-OU and CIR models versus Poisson and mixture models for a somewhat small value of κ ($\kappa/(\kappa + \beta + \gamma) = 0.1$), where discriminability is expected to be difficult. **d.** Nascent and mature marginal distributions for the Γ-OU (red) and CIR (blue) models for a maximally divergent parameter set ($[\kappa = 0.6044, a = 0.2428, \theta = 5.568, \beta = 2.442, \gamma = 0.212]$). Histograms show synthetic data (1000 cells), while the smooth lines show the exact results. **e.** Bayesian inference of noise model parameters. We sampled the posterior distribution of the parameters of the Γ-OU model (assuming it is known that $\beta = 1, \gamma = 1.7$, and $\langle K\rangle = 10$), given a synthetic data set of 1000 cells. Posteriors are presented in both the qualitative regimes space, and in terms of the original parameters. For very Poisson-like data, posteriors are broad in both spaces, because κ is no longer identifiable. MAP: mode of posterior, avg: average of posterior, true: true parameter values.

differential-evolution-based Markov chain Monte Carlo approach, we sampled the posterior distribution of parameters (Figure 3e). To simplify this computation, we assumed that β, γ, and $\langle K\rangle$ were known, because they can be accurately and robustly inferred from empirical means (one can also imagine performing separate experiments to determine them first).

We find that, in the typical scenario (overdispersed data), the posterior is fairly tight in both the qualitative regimes space and in the original parameter space. Both κ and θ are fairly identifiable, allowing us to be optimistic that it is possible to infer biophysical parameters related to transcription rate variation from single-cell data. In the pessimistic scenario (Poisson-like data), model predictions appear to be sloppy with respect to κ. This is because, if the gain ratio $\theta/(\theta + a)$ is small, both the Γ-OU and CIR models predict Poisson-like distributions independent of the value of κ.

Discussion

We have introduced a class of interpretable and tractable models of transcription, and characterized the properties of two biologically plausible members of that class. Our results foreground several considerations for experimental design and modeling in modern transcriptomics.

Interpretable stochastic models encode mechanistic insights, and motivate the collection of data necessary to distinguish between mechanisms. A variety of stochastic differential equations can describe a variety of biophysical phenomena. Through the methods explored in the current study, they can be coupled to models of downstream processing and used to generate testable hypotheses about RNA distributions. Therefore, our SDE–CME framework can guide experiments to parametrize and distinguish between biologically distinct models of transcription.

Conversely, the dramatic effect of dynamic contributions suggests that simple noise models need to be questioned. The slow-reversion regime assumed by the mixture model, which presupposes that the evolution of parameters is substantially slower than RNA dynamics, is attractive but potentially implausible. The parameter set we use to illustrate the slow-reversion regime (see Table S5) has a noise time scale κ^{-1} an order of magnitude longer than degradation, yet still produces distributions that noticeably deviate from the mixture model in their tail regions (Figures S2 and S3). The lifetime of a human mRNA is on the order of tens of hours (66). Therefore, using a mixture model is formally equivalent to postulating a driving process with an autocorrelation time of *weeks*. In practice, if the noise time scale is assumed to be on the order of hours to tens of hours, it is useful to explore non-stationary effects, especially if the analysis focuses on tail effects (45). Our SDE–CME tools facilitate this exploration.

The collection and representation of multimodal data are particularly fruitful directions for experimental design. Even if individual marginals are too similar to use for statistics, joint distributions may be able to distinguish between mechanisms. Aggregating distinct molecular species as a single observable (i.e., modeling the variable $X = X_N + X_M$) neglects biologically important (67–69) regulatory processes of splicing and export buffering. Further, as demonstrated in Figure 3d, marginal

distributions may be insufficiently distinct to identify one of two competing model hypotheses, *even with perfect knowledge of the stationary distribution, autocorrelation, and chemical parameters*. The bioinformatic barriers to generating full gene-specific splicing graphs based on uncharacterized and infrequent intermediate isoforms are formidable. However, the analytical solutions easily accommodate such data, by solving slightly more complicated versions of the ODEs in Equations 12 and 14 (48). Therefore, the deliberate collection of multimodal data is a natural direction for the rational and model-guided planning of high-throughput sequencing experiments.

The identical analytical results for the models' lower moments underscore the need to consider full distributions of molecular species. Although moment-based estimates are useful for qualitative comparisons, and computationally efficient for large bioinformatics datasets, they are insufficient for resolving distinctions even between relatively simple models (65).

In studying the Γ-OU and CIR models, we found and validated several distinct asymptotic regimes. Both models recapitulate the constitutive and mixture models in the slow-driving limit (κ very small). However, in the limit of bursty production (κ large and θ large), they produce qualitatively different behaviors: the Γ-OU model yields geometric bursts of transcription, whereas the CIR model yields inverse Gaussian driving (see Section S5.3) with an infinite number of bursts in each finite time interval. We explicitly solved the inverse Gaussian-driven system and computed the generating function, filling an apparent lacuna (70) in the quantitative finance literature. Discrepancies between the models motivate the quantitative investigation of the effects of jump drivers on the molecule distributions, as even this preliminary study shows that they produce drastically different tail behaviors. Further, we identified a fast-mean-reversion, mean-field regime with rapid fluctuations (κ very fast), which yields effectively constitutive behavior.

The mathematical methods bear further mention, as they can be substantially generalized. The solution for the Γ-OU model given in Section S3.2 exploits an isomorphism between the CME and the underlying driving SDE (48). However, this relation is not practical to apply to broader classes of models. As shown in Section S3.3, the path integral method can recapitulate the solution, with robust performance under wider classes of driving processes (64). More generally, stochastic path integral and physics-inspired methods have recently proven useful for obtaining analytical solutions to relatively complicated stochastic models (71–74). As discussed in Section S3.1, we take this opportunity to explore the diversity of solution methods and emphasize useful unifying themes.

Interestingly, certain superficially different models of regulation can be described by the same models. We have motivated the SDEs by endogenous mechanisms, localized to a single cell. However, these models can also describe *exogenous* variability, such as the transport of regulators into and out of a cell. For example, the mean-reversion term in Equation 4 can model passive equilibration with an extracellular medium, while the noise term can model active transport into the cell. The form of the noise coarsely encodes the physics of the transport: if a regulator is introduced in bursts (e.g., by vesicle transport), the regulator's concentration can be described by a Γ-OU process, whereas if it is introduced by a constant-rate transporter, its concentration can be described by a CIR process. This interpretation is intriguing in light of extensively characterized gene co-expression patterns observed in cultured cells (75–77). Inspired by these results, we propose that the toolbox of SDE-CME models can achieve a mechanistic, yet tractable, treatment of co-expression, modeling the concentration of a multi-gene regulator by a continuous stochastic process.

The availability of numerical solvers suggests natural directions for future study. So far, we have treated the case of transcription rates with time-independent parameters at steady state. However, if the parameters vary with time, it is straightforward to adapt the numerical routines to produce full time-dependent distributions for even more general drivers with a combination of stochastic and deterministic effects. This extension provides a route to explicitly modeling the non-stationary behavior of systems with relatively rapid driver time scales, such as differentiation pathways and the cell cycle. Conversely, the stochastic simulations designed for this study can be easily adapted to describe systems with complex phenomena, such as protein synthesis, reversible binding, and diffusion, which are intractable by analytical approaches in all but the simplest cases.

As we have shown, fine details of transcription—including DNA mechanics and gene regulation—may have signatures in single-cell data, and a model-based, hypothesis-driven paradigm may help identify them. Just as microscopes permit biologists to see beyond their eyes when inspecting a plate of cells, so too can mathematical tools allow them to extract finer insight from the same transcriptomic data.

Materials and Methods

A complete list of major technical results is presented in Section S2. The Γ-OU and CIR models are fully motivated and solved in Section S3. Moments and autocorrelations are derived in Section S4. Limiting cases are derived in Section S5. Simulation details and validation of our exact results are presented in Section S6.

Brief summaries of certain aspects of this work covered more fully in the supplement, and important miscellaneous information, are provided below.

Notation. A complete guide to our mathematical notation is presented in Section S2.2. The molecular species of interest are nascent transcripts $\mathcal{N}$ and mature transcripts $\mathcal{M}$. Their respective counts are denoted by random variables X_N and X_M. The gene locus produces $\mathcal{N}$ with a time-dependent rate $K(t) = K_t$, described by a stochastic process. Therefore, the probability density of the system is given by $P(X_N = x_N, X_M = x_M, K_t \in [K, K + dK], t)$, i.e., the density associated with finding the system in a state with x_N molecules of $\mathcal{N}$, x_M molecules of $\mathcal{M}$, and a transcription rate of K at time t. Having introduced this rather formal notation, we use a shorthand that elides the random variables.

Master equations. The Γ-OU and CIR models are mathematically defined via master equations, which describe how probability flows between different possible states. In particular,

$$\frac{dP(x_N, x_M, K, t)}{dt} = \text{CME} + \text{FPE}, \tag{15}$$

$$\text{FPE}_{\Gamma\text{-OU}} = -\frac{\partial}{\partial K}[(a\theta - \kappa K)P] + a\sum_{n=2}^{\infty}(-\theta)^n \frac{\partial^n P}{\partial K^n}, \tag{16}$$

$$\text{FPE}_{\text{CIR}} = -\frac{\partial}{\partial K}[(a\theta - \kappa K)P] + \kappa\theta\frac{\partial^2 (KP)}{\partial K^2}. \tag{17}$$

The CME term is identical for both models, and encodes transcription, splicing, and degradation reactions as in the constitutive model (63) (see Section S2). However, the Fokker-Planck equation (FPE) terms, which encode transcription rate variation, are different.

Numerically computing generating functions. To numerically obtain distributions, we first compute the generating function (Equation 10) by numerically solving ODEs and integrating the results (i.e., using Equations 11 and 12 or Equations 13 and 14). Then we take an inverse fast Fourier transform (46, 78). The ODEs must be evaluated for a sufficiently fine grid of g_N and g_M on the complex unit sphere.

Analytically solving the Γ-OU and CIR models. The Γ-OU model can be analytically solved using previous results for the n-step birth-death process coupled to a bursting gene. This approach exploits the fact that the source species of such a system has a Poisson intensity described by the Γ-OU process, and is fully outlined in Section S3.2. We set up a system with a bursting gene coupled to a 3-step birth-death process, characterized by the path graph $\varnothing \xrightarrow{a} B \times \mathcal{T}_0 \xrightarrow{\kappa} \mathcal{N} \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} \varnothing$, where $B \sim Geom$ with mean θ/κ.

The stochastic process describing the Poisson intensity of $\mathcal{T}_0$ is precisely the Γ-OU process (79). This implies that the joint distribution of the downstream species coincides with the system driven by Γ-OU transcription. The generating function of SDE-driven system can be computed using the solution of the bursty system, reported in Equation 11, where $U_0(s; u_N, u_M) = A_0 e^{-\kappa s} + A_1 e^{-\beta s} + A_2 e^{-\gamma s}$ can be computed by solving Equation 12:

$$\begin{aligned} A_2 &= u_M \frac{\beta}{\beta - \gamma}\frac{\kappa}{\kappa - \gamma}, \\ A_1 &= \frac{\kappa}{\kappa - \beta}\left(u_N - u_M \frac{\beta}{\beta - \gamma}\right), \\ A_0 &= -A_1 - A_2. \end{aligned}$$

The CIR model is solved using a state space path integral representation of $P(x_N, x_M, K, t)$ which combines the path integral representation of the CME from (55) with a more conventional continuous state space path integral. The Γ-OU model can also be solved using this method, along with a plethora of other discrete-continuous hybrid models.

Analytically computing moments and autocorrelation functions. The master equation satisfied by $P(x_N, x_M, K, t)$ can be recast as a partial differential equation (PDE) satisfied by $\phi(u_N, u_M, s, t)$ (see Section S3):

$$\begin{aligned} \frac{\partial \phi}{\partial t} = u_N \frac{\partial \phi}{\partial s} + \beta(u_M - u_N)\frac{\partial \phi}{\partial u_N} - \gamma u_M \frac{\partial \phi}{\partial u_M} \\ + a\theta s - \kappa s \frac{\partial \phi}{\partial s} + f(s), \end{aligned} \tag{18}$$

$$f_{\Gamma\text{-OU}}(s) = a\sum_{n=2}^{\infty}\theta^n s^n, \tag{19}$$

$$f_{\text{CIR}}(s) = s^2 \kappa\theta \frac{\partial \phi}{\partial s}. \tag{20}$$

By taking certain partial derivatives of the above PDEs, we can recover ODEs satisfied by moments and autocorrelation functions. These can then be straightforwardly solved to compute them.

Simulation. Stochastic simulations can verify our analytical results and enable further facile extensions to SDE-driven systems that are otherwise analytically intractable. Because our models involve no feedback, we split this problem into two parts: first, we simulate the continuous stochastic dynamics of the transcription rate $K(t)$, and then we simulate the discrete stochastic dynamics of the nascent and mature RNA using a variant of Gillespie's direct method (80). This approach requires evaluating reaction waiting times for time-varying transcription rates. For the Γ-OU model, we computed these times exactly *via* the Lambert W function. For the CIR model, we used a trapezoidal approximation of the integral.

To ensure that all regimes of interest are verified, we chose six parameter sets to test: four of these lie in the extreme limits shown in Figure 2, and two lie in intermediate regimes. We performed 10^4 simulations for each parameter set, with $\beta = 1.2$ and $\gamma = 0.7$. The trajectories were equilibrated until a putative steady-state time T_{ss}. Afterward, the simulations were left to run until $T_{ss} + T_R$ to enable the computation of autocorrelations. The parameters as well as values of T_{ss} and T_R are reported in Table S5. The implementation details and simulation results are given in Section S6.

Parameter inference. The Python package PyMC3 was used to sample the parameter posteriors. Because we did not have gradients of our likelihood functions available, we used the non-gradient-based DEMetropolisZ as our Markov chain Monte Carlo sampler.

Availability. The simulated data, algorithms, and Python notebooks used to generate the figures are available at https://github.com/pachterlab/GVFP_2021.

ACKNOWLEDGMENTS. The DNA, pre-mRNA, and mature mRNA used in Figure 1 are derivatives of the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. G.G. acknowledges the help of Victor Rohde in exploration of the stochastic process literature. G.G., M.F. and L.P. were partially funded by NIH U19MH114830. J.J.V. was supported by NSF Grant # DMS 1562078.

1. N Battich, T Stoeger, L Pelkmans, Control of Transcript Variability in Single Mammalian Cells. *Cell* **163**, 1596–1610 (2015).
2. J Cao, et al., The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
3. C Xia, J Fan, G Emanuel, J Hao, X Zhuang, Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci.* p. 201912459 (2019).
4. SY Anvar, et al., Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).
5. D Rognan, Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **152**, 38–52 (2007).
6. S Mandal, M Moudgil, SK Mandal, Rational drug design. *Eur. J. Pharmacol.* **625**, 90–100 (2009) New Vistas in Anti-Cancer Therapy.
7. GR Bowman, ER Bolin, KM Hart, BC Maguire, S Marqusee, Discovery of multiple hidden allosteric sites by combining markov state models and experiments (2015).
8. BJ Williams-Noonan, E Yuriev, DK Chalmers, Free energy methods in drug design: Prospects of "alchemical perturbation" in medicinal chemistry. *J. Medicinal Chem.* **61**, 638–649 (2018) PMID: 28745501.
9. D Liu, Y Mao, X Gu, Y Zhou, D Long, Unveiling the "invisible" druggable conformations of gdp-bound inactive ras. *Proc. Natl. Acad. Sci.* **118** (2021).
10. Z Fox, G Neuert, B Munsky, Finite state projection based bounds to compare chemical master equation models using single-cell data. *The J. Chem. Phys.* **145**, 074101 (2016).
11. ZR Fox, B Munsky, The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments. *PLOS Comput. Biol.* **15**, e1006365 (2019).
12. ZR Fox, G Neuert, B Munsky, Optimal design of single-cell experiments within temporally fluctuating environments. *Complexity* **2020**, 8536365 (2020).
13. D Silk, PDW Kirk, CP Barnes, T Toni, MPH Stumpf, Model Selection in Systems Biology Depends on Experimental Design. *PLoS Comput. Biol.* **10**, e1003650 (2014).
14. MD Robinson, GK Smyth, Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).
15. MD Robinson, DJ McCarthy, GK Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
16. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **15**, 550 (2014).
17. R Rostom, V Svensson, SA Teichmann, G Kar, Computational approaches for interpreting scRNA-seq data. *FEBS Lett.* **591**, 2213–2225 (2017).
18. MD Luecken, FJ Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
19. LF Liu, JC Wang, Supercoiling of the dna template during transcription. *Proc. Natl. Acad. Sci.* **84**, 7024–7027 (1987).
20. J Peccoud, B Ycard, Markovian Modeling of Gene Product Synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
21. S Iyer-Biswas, F Hayot, C Jayaprakash, Stochasticity of gene products from transcriptional pulsing. *Phys. Rev. E* **79**, 031911 (2009).
22. L Huang, Z Yuan, P Liu, T Zhou, Effects of promoter leakage on dynamics of gene expression. *BMC Syst. Biol.* **9** (2015).
23. SA Sevier, DA Kessler, H Levine, Mechanical bounds to transcriptional noise. *Proc. Natl. Acad. Sci.* **113**, 13983–13988 (2016).
24. SA Sevier, H Levine, Mechanical properties of transcription. *Phys. Rev. Lett.* **118**, 268101 (2017).
25. SA Sevier, H Levine, Properties of gene expression and chromatin structure with mechanically regulated elongation. *Nucleic Acids Res.* **46**, 5924–5934 (2018).
26. Z Cao, T Filatova, DA Oyarzún, R Grima, A stochastic model of gene expression with polymerase recruitment and pause release (2020).
27. Z Cao, R Grima, Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proc. Natl. Acad. Sci.* **117**, 4682–4692 (2020).
28. M Thattai, A van Oudenaarden, Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci.* **98**, 8614–8619 (2001).
29. MB Elowitz, AJ Levine, ED Siggia, PS Swain, Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).
30. PS Swain, MB Elowitz, ED Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci.* **99**, 12795–12800 (2002).
31. A Raj, A van Oudenaarden, Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* **135**, 216–226 (2008).
32. A Hilfinger, J Paulsson, Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proc. Natl. Acad. Sci.* **108**, 12167–12172 (2011).
33. G La Manno, et al., RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
34. S Shah, et al., Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell* **174**, 363–376.e16 (2018).
35. I Golding, J Paulsson, SM Zawilski, EC Cox, Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025–1036 (2005).
36. DA McQuarrie, Stochastic approach to chemical kinetics. *J. Appl. Probab.* **4**, 413–478 (1967).
37. DT Gillespie, A rigorous derivation of the chemical master equation. *Phys. A: Stat. Mech. its Appl.* **188**, 404–425 (1992).
38. DT Gillespie, The chemical Langevin equation. *The J. Chem. Phys.* **113**, 297–306 (2000).
39. DT Gillespie, Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).
40. DT Gillespie, A Hellander, LR Petzold, Perspective: Stochastic algorithms for chemical kinetics. *The J. Chem. Phys.* **138**, 170901 (2013).
41. Z Fox, B Munsky, Stochasticity or Noise in Biochemical Reactions, (arXiv: 1708.09264), Preprint (2017).
42. B Munsky, WS Hlavacek, LS Tsimring, eds., *Quantitative Biology: Theory, Computational Methods, and Models.* (The MIT Press), (2018).
43. T Jahnke, W Huisinga, Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.* **54**, 1–26 (2006).
44. A Sanchez, I Golding, Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science* **342**, 1188–1193 (2013).
45. L Ham, RD Brackston, MPH Stumpf, Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Phys. Rev. Lett.* **124**, 108101 (2020).
46. A Singh, P Bokes, Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophys. J.* **103**, 1087–1096 (2012).
47. G Gorin, L Pachter, Special function methods for bursty models of transcription. *Phys. Rev. E* **102**, 022409 (2020).
48. G Gorin, L Pachter, Analytical solutions of the chemical master equation with bursty production and isomerization reactions, (bioRxiv: 2021.03.24.436847), Preprint (2021).
49. RD Dar, et al., Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci.* **109**, 17454–17459 (2012).
50. R Cont, P Tankov, *Financial Modeling with Jump Processes*, Financial Mathematics. (Chapman & Hall), (2004).
51. OE Barndorff-Nielsen, N Shephard, Modelling by lévy processess for financial econometrics in *Lévy Processes: Theory and Applications*, eds. OE Barndorff-Nielsen, SI Resnick, T Mikosch. (Birkhäuser Boston, Boston, MA), pp. 283–318 (2001).
52. OE Barndorff-Nielsen, N Shephard, Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **63**, 167–241 (2001).
53. Y Qu, A Dassios, H Zhao, Exact simulation of gamma-driven Ornstein–Uhlenbeck processes with finite and infinite activity jumps. *J. Oper. Res. Soc.* pp. 1–14 (2019).
54. G Bernis, R Brignone, S Scotti, C Sgarra, A gamma ornstein–uhlenbeck model driven by a hawkes process. *Math. Financial Econ.* (2021).
55. JJ Vastola, WR Holmes, Chemical Langevin equation: A path-integral view of Gillespie's derivation. *Phys. Rev. E* **101**, 032417 (2020).
56. JC Cox, JE Ingersoll, SA Ross, A theory of the term structure of interest rates in *Theory of Valuation*, eds. S Bhattacharya, GM Constantinides. (World Scientific Publishing Company), pp. 129–164 (2005).
57. SJ Brown, PH Dybvig, The empirical implications of the cox, ingersoll, ross theory of the term structure of interest rates. *The J. Finance* **41**, 617–630 (1986).
58. J Hull, A White, Pricing Interest-Rate-Derivative Securities. *The Rev. Financial Stud.* **3**, 573–592 (2015).
59. B Hu, DA Kessler, WJ Rappel, H Levine, How input fluctuations reshape the dynamics of a biological switching system. *Phys. review. E, Stat. nonlinear, soft matter physics* **86**, 061910 (2012).
60. C Zechner, H Koeppl, Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. *PLoS Comput. Biol.* **10**, e1003942 (2014).
61. W Saelens, R Cannoodt, H Todorov, Y Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
62. A Pratapa, AP Jalihal, JN Law, A Bharadwaj, TM Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
63. PJ Gans, Open First-Order Stochastic Processes. *The J. Chem. Phys.* **33**, 691–694 (1960).
64. JJ Vastola, The information transmission problem in RNA splicing networks, and a path integral framework for exactly solving coupled discrete and continuous stochastic dynamics (2021).
65. B Munsky, G Li, ZR Fox, DP Shepherd, G Neuert, Distribution shapes govern the discovery of predictive models for gene regulation. *Proc. Natl. Acad. Sci.* **115**, 7533–7538 (2018).
66. R Milo, R Phillips, *Cell Biology by the Numbers.* (Garland Science), (2015).
67. Q Wang, T Zhou, Alternative-splicing-mediated gene expression. *Phys. Rev. E* **89**, 012713 (2014).
68. T Alpert, L Herzel, KM Neugebauer, Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip. Rev. RNA* **8**, e1401 (2017).
69. M Schmid, TH Jensen, Controlling nuclear RNA levels. *Nat. Rev. Genet.* **19**, 518–529 (2018).
70. OE Barndorff-Nielsen, N Shephard, Integrated OU Processes and Non-Gaussian OU-based Stochastic Volatility Models. *Scand. J. Stat.* **30**, 277–295 (2003).
71. JJ Vastola, The chemical birth-death process with additive noise, (arXiv: 1910.09117), Preprint (2019).
72. JJ Vastola, The chemical birth-death process with Gillespie noise, (arXiv: 1910.10807), Preprint (2019).
73. JJ Vastola, Solving the chemical master equation for monomolecular reaction systems analytically: a Doi-Peliti path integral view. *arXiv:1911.00978 [q-bio]* (2019) arXiv: 1911.00978.
74. JJ Vastola, G Gorin, L Pachter, WR Holmes, Analytic solution of chemical master equations involving gene switching. I: Representation theory and diagrammatic approach to exact solution, (arXiv: 2103.10992), Preprint (2021).
75. B Munsky, G Neuert, A van Oudenaarden, Using Gene Expression Noise to Understand Gene Regulation. *Science* **336**, 183–187 (2012).
76. SJ Gandhi, D Zenklusen, T Lionnet, RH Singer, Transcription of functionally related constitutive genes is not coordinated. *Nat. structural & molecular biology* **18**, 27–34 (2011).
77. J Stewart-Ornstein, JS Weissman, H El-Samad, Cellular Noise Regulons Underlie Fluctuations in Saccharomyces cerevisiae. *Mol. Cell* **45**, 483–493 (2012).
78. P Bokes, JR King, ATA Wood, M Loose, Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *J. Math. Biol.* **64**, 829–854 (2012).

79. L Amrhein, K Harsha, C Fuchs, A mechanistic model for the negative binomial distribution of single-cell mRNA counts, (bioRxiv: 657619), Preprint (2019).
80. DT Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).