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

Gennady Gorin[1], John J. Vastola[2], Meichen Fang[3], and Lior Pachter[3,4]

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125
[2]Department of Neurobiology, Harvard Medical School, Boston, MA, 02115
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, 91125
[4]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, 91125
[*]Address correspondence to Lior Pachter (lpachter@caltech.edu)

September 6, 2021

Abstract

To what extent do cell-to-cell differences in transcription rate affect RNA copy number distributions, and what can this variation tell us about biological processes underlying transcription? We argue that successfully answering such questions requires quantitative models that are both interpretable (describing concrete biophysical phenomena) and tractable (amenable to mathematical analysis); in particular, such models enable 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 (governed by a stochastic differential equation) coupled to a discrete stochastic RNA transcription and splicing process, and compare and contrast two biologically plausible hypotheses about observed transcription rate variation. One hypothesis assumes transcription rate variation is due to DNA experiencing mechanical strain and relaxation, while the other assumes that variation is due to fluctuations in the number of an abundant regulator. Through a thorough mathematical analysis, we show that these two models are challenging to distinguish: properties like first- and second-order moments, autocorrelations, and several limiting distributions are shared. However, our analysis also points to the experiments which best discriminate between them. Our work illustrates the importance of theory-guided data collection in general, and multimodal single-molecule data in particular for distinguishing between competing hypotheses. We use this theoretical case study to introduce and motivate a general framework for constructing and solving such nontrivial continuous–discrete models.

Contents

1 Introduction **3**

2 Why might transcription rates vary across cell populations? **5**
2.1 The constitutive model of transcription . . . 5
2.2 Transcription rate variation accounts for empirically observed variance . . . 6
2.3 High-level overview of modeling approach . . . 7
2.4 Gamma Ornstein–Uhlenbeck production rate model . . . 10
2.5 Cox–Ingersoll–Ross production rate model . . . 11

3 Behavior of each transcription model is interpretable **14**
3.1 Generic behavior of each model . . . 14
3.2 Behavior across representative parameter regimes . . . 15
3.3 Simulation framework for numerically solving models . . . 15

4 Behavior of each transcription model is tractable **19**
4.1 Summary of mathematical calculations and approach . . . 19
4.2 Notation . . . 21
4.3 Steady-state probability distribution solutions . . . 22
4.3.1 Gamma Ornstein–Uhlenbeck model . . . 22
4.3.2 Cox–Ingersoll–Ross model . . . 23
4.4 Distribution properties . . . 24
4.4.1 Moments . . . 24
4.4.2 Autocorrelation functions . . . 25
4.5 Limiting cases . . . 26
4.5.1 Fast mean-reversion limit ($\kappa \to \infty$, $a \to \infty$, a/κ fixed) . . . 26
4.5.2 Slow mean-reversion limit ($\kappa \to 0$, $a \to 0$, a/κ fixed) . . . 27
4.5.3 Low gain limit ($\theta \to 0$, $\kappa \to 0$, θ/κ fixed) . . . 27
4.5.4 High gain limit ($\theta \to \infty$, $\kappa \to \infty$, θ/κ fixed) . . . 27

5 Discriminating between noise models **30**

6 Discussion **32**

7 Code Availability **34**

8 Acknowledgments **34**

References **34**

1 Introduction

A wealth of available single-cell RNA count data, including measurements of gene expression and submolecular features in thousands of single cells at a time [1–4], has opened the possibility of learning much about the molecular biology of the cell. However, despite this abundance of data, there is still much we do not know about the biology underlying key processes. For example, the numbers of RNA molecules inside single cells fluctuate due to a combination of cell-to-cell differences (extrinsic noise) and randomness associated with the timing of events like transcription and degradation (intrinsic noise) [5–9]; how much does stochastic variation tell us about the fine details of transcription, such as the influence of DNA supercoiling and molecular regulators?

The current largely data-driven approach to the analysis of single-cell transcription information makes these questions difficult to answer. Typically, only *post-hoc* analyses are performed by fitting a negative binomial, or similar distribution, to observed RNA count distributions [10–14]. But such approaches merely summarize the data, and indicate little about the underlying biology. They are mathematically *tractable*, but unfortunately not biologically *interpretable*. On the other hand, there is a rich tradition of constructing detailed mathematical models of transcription [15–23]. These models, while certainly interpretable, tend not to be tractable: high levels of detail make a thorough mathematical analysis challenging, and identifiability issues mean that it can be difficult or impossible to use the data one has to distinguish among many competing hypotheses.

In principle, models of transcription that are both interpretable and tractable would allow us to go beyond the current data-driven approach, and to be more hypothesis-driven. To understand why, suppose one represents several competing hypotheses using models of this type. Because the models are interpretable, fitting their parameters to data conveys clear biological information, e.g. about the dynamics of DNA supercoiling and relaxation. Because the models are tractable, one can perform a thorough mathematical analysis of their behavior, and use this analysis to identify the experiment that best distinguishes between them. This enables a 'rational' design of transcriptomic experiments (Figure 1), analogous to ideas about rational drug design [24–28] and the optimal design of single-cell experiments [29–31].

In this paper, we will not actually *implement* the entire paradigm depicted in Figure 1; indeed, most of the loop falls under the well-established umbrella of Bayesian updating. Instead, we will develop one possible mathematical and computational foundation for the mathematical analysis, which turns out to be challenging in its own right. We propose a class of transcription models that is fairly simple, yet flexible enough to account for a range of biological phenomena. This class of models, which we will show satisfies both interpretability and tractability, assumes that a stochastic and time-varying transcription rate drives a discrete stochastic RNA transcription and splicing process. We focus on two specific examples of models from this class: one which assumes transcription rate variation is due to random changes in the mechanical state of DNA (the gamma Ornstein–Uhlenbeck model), and one which assumes variation is due to random changes in the number of an abundant regulator (the Cox–Ingersoll–Ross model).

After thoroughly analyzing these two models using a variety of numerical and mathematical tools, we find that they are extremely difficult to tell apart: standard summaries such as the first and second-order moments, autocorrelations, and several limiting distributions are identical between the models. This suggests that a naïve approach may fail to discern between them. Nonetheless, they *can* be distinguished, and our analysis suggests the best way to do so. In finding that these models yield different predictions, we answer 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 long-time distribution of those fluctuations, can qualitatively affect model predictions.

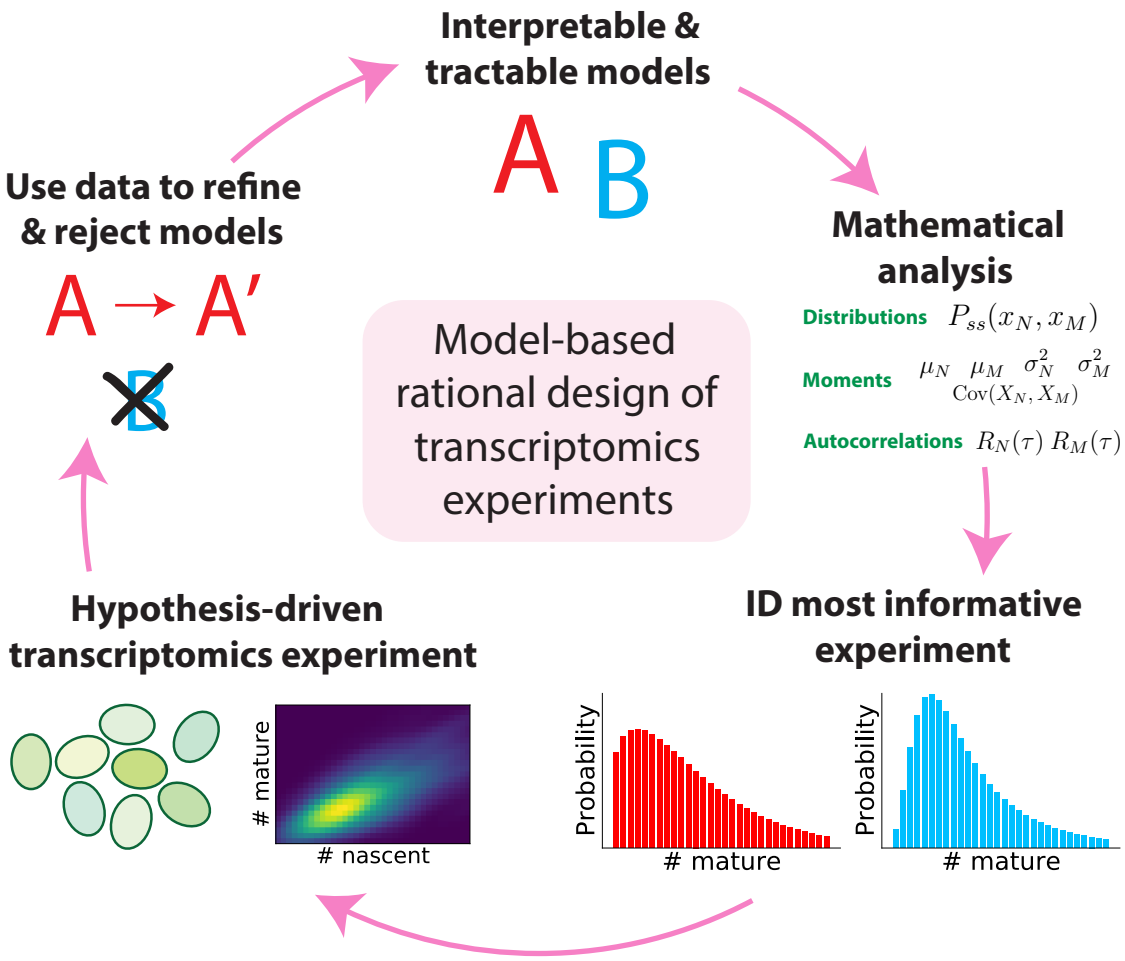

Figure 1: Paradigm for model-based rational design of transcriptomics experiments. One begins by representing two or more competing hypotheses as interpretable and tractable mathematical models (top of circle). Next, one performs a detailed mathematical analysis of each model, computing quantities (e.g. RNA count distributions and moments) that can help distinguish one hypothesis from another. Using the results of that analysis as input, one identifies the experiment that best distinguishes the two models. Finally, one performs this experiment on some population of cells, uses the resulting data to refine the models, and repeats the process with an updated ensemble.

The paper is organized as follows. In Section 2, we motivate the class of models we are considering and the two biologically plausible examples we focus on. In Section 3, we study each model's predictions and show that their behavior is interpretable (predicted distributions clearly convey information about fine details of transcription). In Section 4, we perform a thorough mathematical analysis of each model, and show that their behavior is tractable (closed-form mathematical formulas are available for important model properties). In Section 5, we consider how best to distinguish between them. Finally, in Section 6 we consider some implications of our work for hypothesis-driven transcriptomics, and discuss the broader utility of our novel numerical and mathematical methods.

2 Why might transcription rates vary across cell populations?

In this section, we motivate the class of transcription models we are considering, along with the two specific examples (the gamma Ornstein–Uhlenbeck and Cox–Ingersoll–Ross models) we study in detail throughout this paper. The key question motivating these models is: why might different cells in a population have different transcription rates, and how do these differences affect the RNA count distributions we observe?

To motivate and formalize this question, we first review the constitutive model of transcription (which assumes a constant transcription rate that is identical for all cells in a population) and its empirical shortcomings. Next, we review how this model can be extended to fit observed distributions if cell-to-cell differences in transcription rate are incorporated. This suggests models with transcription rate variation *can* fit the data, but does not identify the biological source of that variation. To address this issue, we present a high-level overview of our modeling approach, which involves instantiating specific biological hypotheses about transcription rate variation using tractable mathematical models. Finally, we derive and explain the intuition behind the gamma Ornstein–Uhlenbeck and Cox–Ingersoll–Ross models.

2.1 The constitutive model of transcription

The most well-studied quantitative model of transcription is the so-called *constitutive* model, which assumes that RNA is produced at a constant rate [32]. More precisely, the constitutive model of nascent and mature RNA dynamics is defined by the chemical reaction list

$$\begin{aligned} \varnothing &\xrightarrow{K} \mathcal{N} \\ \mathcal{N} &\xrightarrow{\beta} \mathcal{M} \\ \mathcal{M} &\xrightarrow{\gamma} \varnothing \end{aligned}$$

where $\mathcal{N}$ denotes nascent RNA, $\mathcal{M}$ denotes mature RNA, K is the transcription/production rate, β is the splicing rate, and γ is the mature RNA degradation rate. Note that the transcription rate K is constant, and does not depend on time or any kind of regulation.

To appropriately model changes in RNA counts as discrete and stochastic, one must use the theoretical framework associated with the chemical master equation (CME) [33–39]. The CME for the constitutive model, which fully specifies its behavior, takes the following form:

$$\begin{aligned} \frac{\partial P(x_N, x_M, t)}{\partial t} = K\left[P(x_N - 1, x_M, t) - P(x_N, x_M, t)\right] \\ + \beta\left[(x_N + 1)P(x_N + 1, x_M - 1, t) - x_N P(x_N, x_M, t)\right] \\ + \gamma\left[(x_M + 1)P(x_N, x_M + 1, t) - x_M P(x_N, x_M, t)\right] \end{aligned} \tag{1}$$

where $x_N \in \mathbb{N}_0$ denotes the number of nascent RNA, $x_M \in \mathbb{N}_0$ denotes the number of mature RNA, and $P(x_N, x_M, t)$ is the probability that the system is in state (x_N, x_M) at time t.

It is this equation which must be solved in order to understand the model's behavior. Using various mathematical techniques [40–42], or numerical techniques such as variants of Gillespie's algorithm [43, 44], one finds particularly simple behavior in the long-time limit relevant to single-cell experiments: the long-time/steady-state distribution $P_{ss}(x_N, x_M)$ is Poisson, i.e.

$$P_{ss}(x_N, x_M) := \lim_{t\to\infty} P(x_N, x_M, t) = \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!} \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!} \ .$$

In other words, since Poisson distributions are unimodal and have little variance, the number of nascent RNA in a single cell is typically around K/β, and the number of mature RNA in a single cell is typically around K/γ.

The constitutive model has many shortcomings, and clearly fails to capture a great deal of important biology. Although it has been shown to fit certain single cell data (for example, from some yeast housekeeping genes [32,45]) surprisingly well, its most serious problem for our purposes is that it fails to account for the fact that most observed eukaryotic RNA count distributions are 'overdispersed'—i.e., they have a higher variance than Poisson distributions with the same mean.

2.2 Transcription rate variation accounts for empirically observed variance

One interesting way to account for 'overdispersion' in RNA count distributions is to assume that different cells in a population have distinct transcription rates. Depending on the postulated transcription rate distribution, one can obtain various RNA count distributions, including the negative binomial distribution often used to fit count data. In particular, consider the constitutive model again, but assume that K is distributed according to

$$f(K) = \frac{K^{\alpha-1}\, e^{-K/\theta}}{\theta^{\alpha}\, \Gamma(\alpha)},$$

i.e., assume K is gamma-distributed with shape parameter α and scale parameter θ. Then, in the long-time limit, RNA counts will be distributed according to

$$P_{ss}(x_N, x_M) = \int_0^{\infty} dK\, \frac{K^{\alpha-1}\, e^{-K/\theta}}{\theta^{\alpha}\, \Gamma(\alpha)}\, \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!}\, \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!}.$$

The marginal distributions of this joint distribution will each be negative binomial, e.g.

$$P_{ss}(x_N) := \sum_{x_M=0}^{\infty} P_{ss}(x_N, x_M) = \binom{x_N + \alpha - 1}{x_N} \left(\frac{\beta}{\theta+\beta}\right)^{\alpha} \left(\frac{\theta}{\theta+\beta}\right)^{x_N}.$$

In this way, we can slightly extend the constitutive model so that it actually fits observed single-cell data. But this approach—which is essentially equivalent to the *post-hoc* fitting of negative binomial distributions—is not physically 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?

2.3 High-level overview of modeling approach

We would like models which are biologically detailed enough to fit the data, but which are simple enough to allow a thorough mathematical study of their behavior. The model presented in the previous section is certainly mathematically tractable, but is biologically opaque; meanwhile, many existing biophysical models of transcription incorporate plenty of relevant biology, but are extremely challenging to solve and fit parameters to.

We propose a class of transcriptional models that balance interpretability and tractability. Although various biological details underlying transcription may be quite complicated, we assume they can be captured by an effective transcription rate $K(t)$ which is stochastic and varies with time. This transcription rate is assumed to randomly fluctuate around some 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 a stochastic differential equation (SDE)

$$\begin{aligned} \dot{K} &= [\text{ mean reversion }] + [\text{ noise }] \\ &= A - BK + [\text{ noise }] \end{aligned} \tag{2}$$

for some coefficients A and B, where [noise] denotes a model-dependent term that introduces stochastic variation. The mean-reversion term pushes $K(t)$ towards its mean value, while the noise term causes random fluctuations. The transcription rate $K(t)$ is coupled to the dynamics of RNA production and splicing as in the constitutive model, so that the reaction list

$$\begin{aligned} \varnothing &\xrightarrow{K(t)} \mathcal{N} \\ \mathcal{N} &\xrightarrow{\beta} \mathcal{M} \\ \mathcal{M} &\xrightarrow{\gamma} \varnothing \end{aligned}$$

describes the dynamics of the nascent ($\mathcal{N}$) and mature ($\mathcal{M}$) RNA. Just as before, we can write down a master equation like Eq. 1 that characterizes the dynamics of the overall system (whose precise form depends on how one chooses the terms in Eq. 2). This modeling framework is summarized schematically in Figure 2a.

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, and to generalize the naïve negative binomial model from the previous section, we will specifically consider candidate models for which the long-time 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 (Section 4.5) that the popular bursting model of RNA production [1,46–48] 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. In the following two sections, we present the biophysical motivation for each. See Table 1 for a guide to the parameters of both models, and Figure 2b for a cartoon illustration of their biological interpretations. In Figure 3, we qualitatively compare the transcriptional models mentioned so far.

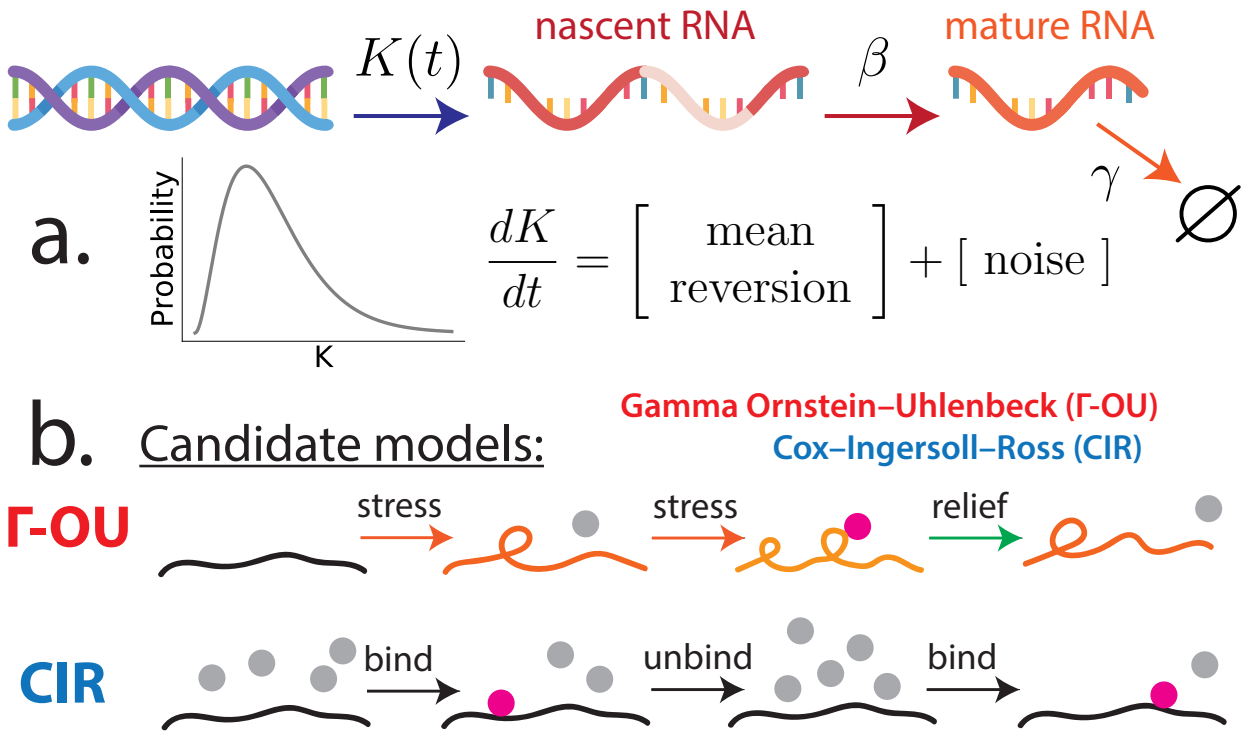

Figure 2: The modeling framework we introduce in this paper. (a) We consider stochastic models of transcription involving (i) nascent/unspliced RNA, (ii) mature/spliced RNA, and (iii) a stochastic and time-varying transcription rate $K(t)$. The transcription rate is assumed to evolve in time according 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. (b) The two candidate models we focus on. The gamma Ornstein–Uhlenbeck (Γ-OU) model views transcription as becoming increasingly inefficient over time due to mechanical stress, which can create topological obstructions like DNA loops. This stress is randomly relieved by topoisomerases. The Cox–Ingersoll–Ross (CIR) model assumes that transcription depends on some abundant regulator molecule, and that fluctuations in the transcription rate are due to fluctuations in the number of this molecule. As the number of regulator molecules changes, how frequently they tend to bind to DNA changes, causing them to modulate transcription more or less.

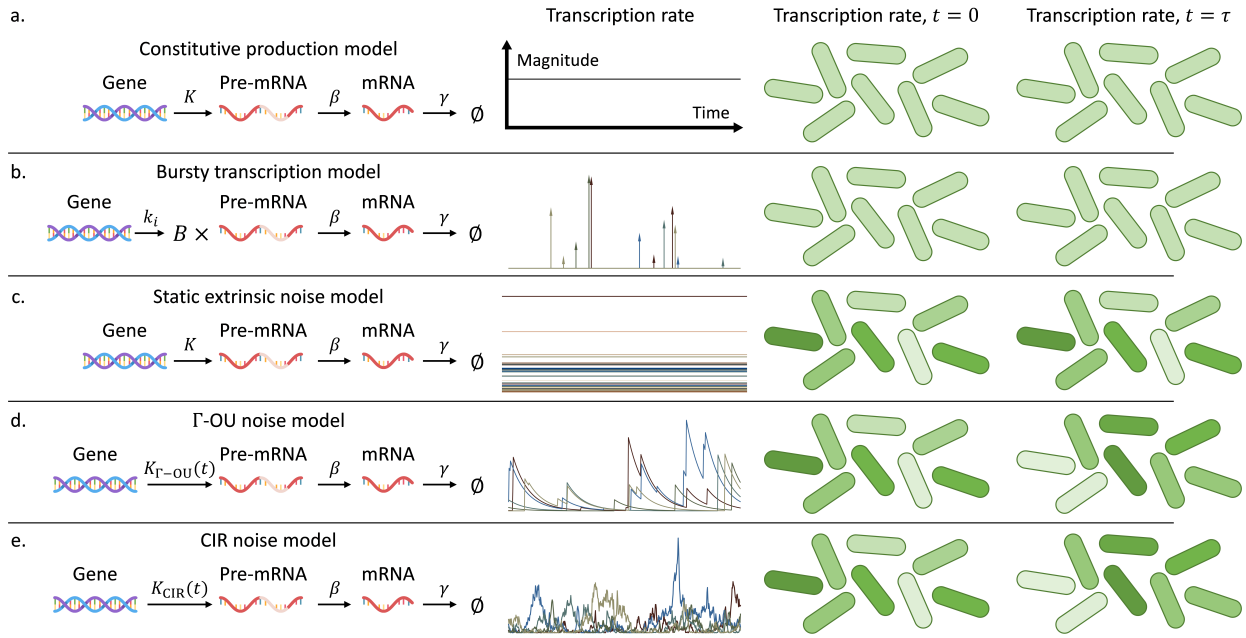

Figure 3: An overview of the qualitative features of common transcription rate models. The constitutive production model has a constant and identical transcription rate for all cells. The intrinsic noise model exhibits bursts of production, which instantaneously generate one or more RNA molecules. The extrinsic noise model has time-invariant cell-specific transcription rates. The Γ-OU and CIR processes considered in the present study vary between cells and over time, but with qualitatively different trajectory dynamics.
(a) Schema of the constitutive production model (K: transcription rate; β: pre-RNA splicing rate; γ: RNA degradation rate. Uniform shade of green indicates identical parameter values across all cells). (b) Schema of the bursty noise model (k_i: burst frequency; B: burst size drawn from a geometric distribution. Transcription rate variation corresponds to instantaneous transcription events with Dirac delta intensity. Uniform shade of green indicates identical burst statistics across all cells). (c) Schema of the static extrinsic noise model (K: transcription rate. Different shades of green indicate different, but time-independent, values of K across cells). (d) Schema of the Γ-OU noise model ($K_{\Gamma\text{-OU}}(t)$: time-dependent transcription rate. Different shades of green indicate different values of K across cells and throughout time). (e) Schema of the CIR noise model (parameters analogous to Γ-OU).

Symbol	Parameter	Γ-OU interpretation	CIR interpretation
κ	Mean reversion rate	DNA frustration rate	Regulator degradation rate
a	Transcription-increasing process rate	Topoisomerase arrival frequency	Regulator production rate
$\alpha := a/\kappa$	Shape parameter of gamma distribution	Mean relaxation level	Mean regulator abundance
θ	Scale parameter of gamma distribution	Relaxation gain	Regulator gain
β	Splicing rate		
γ	Mature RNA degradation rate		

Table 1: Model variables and parameters. The Γ-OU and CIR models both have five independent parameters (three transcription-related parameters, a splicing rate, and a degradation rate). The parameters in each model mathematically correspond to one another, but have different biological interpretations.

2.4 Gamma Ornstein–Uhlenbeck production rate model

Although the wide use of mass action-type models of transcription obscures the mechanical details of the process, biomechanics can have important consequences for transcriptional dynamics. Each nascent RNA produced by an RNA polymerase induces a small amount of mechanical stress in DNA, making transcription slightly more difficult. This mechanical stress builds up with each transcription event; if the stress is sufficiently high (i.e., if the DNA is excessively supercoiled), transcription is mechanically frustrated, and more nascent RNA cannot be produced until this stress is relieved. Topoisomerases arrive to relieve stress, creating a dynamic balance between transcription-mediated frustration and topoisomerase-mediated recovery. This model has been explored by Sevier, Kessler, and Levine [19,21], and shown to recapitulate gene bursting. However, this detailed mechanical model requires the description of submolecular features and feedback between regulatory and transcriptional events—features which make it difficult to work with in practice.

We can simplify this model while retaining crucial qualitative aspects. Let the transcription rate be proportional to the level of DNA relaxation, i.e. $K(t) = \theta \cdot \mathrm{rel}$, where θ is a scaling factor/'gain'. We assume that DNA relaxation continuously decreases (as transcription happens roughly continuously), and that topoisomerases randomly arrive to increase relaxation. In other words, we will describe the dynamics of relaxation via the SDE

$$\dot{\mathrm{rel}} = -\kappa\ f(\mathrm{rel}) + [\ \text{noise}\ ],$$

where f is some functional dependence on the current level of relaxation, κ encodes its time scale, and [noise] denotes the random topoisomerase-induced increases in relaxation.

We choose the following plausible model for these phenomena. The functional dependence is simply given by $f(\mathrm{rel}) = \mathrm{rel}$, corresponding to linear frustration. In a small amount of time Δt,

a number $n \sim Poisson(a\Delta t)$ of topoisomerases arrive to relieve stress, and (having chosen the units of 'rel' appropriately) the ith topoisomerase increases relaxation by an amount $r_i \sim Exp(1)$. Thus, topoisomerase arrival is a Poisson process, and the stress relief of individual topoisomerases is described by Poisson shot noise.

Mathematically, this means the noise comes from a specific kind of Lévy process: a compound Poisson process with arrival frequency a and exponentially distributed jumps with expectation 1. We can write

$$\dot{\text{rel}} = -\kappa \text{ rel} + \epsilon(t),$$

where $\epsilon(t)$ denotes the infinitesimal Lévy process. Now the production rate $K(t) = \theta \cdot \text{rel}$ satisfies the SDE

$$\dot{K}(t) = -\kappa K(t) + \epsilon(t),$$

where the exponentially distributed random variables r_i that appear in the Lévy process now have $r_i \sim Exp(1/\theta)$, i.e. the expected jump size is θ.

This is the gamma Ornstein–Uhlenbeck (Γ-OU) model of transcription [49]. To summarize, 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 rate κ, and the stochastic relaxation by topoisomerases that arrive at rate a. The scaling between the relaxation rate and the transcription rate is set by the gain 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, among other things [50–53]. 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.

2.5 Cox–Ingersoll–Ross production rate model

The rate of RNA production often depends on the concentration of regulatory molecules that do not get consumed by transcription, such as RNA polymerases, inducers, and activators. When there are more of such molecules available, we expect more transcription to occur; when there are fewer, we expect less transcription. Exactly how many of these molecules there are at any given time depends on how frequently *they* are produced and degraded.

We can codify this intuition in the following crude model. Let $\mathcal{T}$ denote our RNA *t*ranscript, and $\mathcal{R}$ label a *r*egulator that enables its transcription. Consider the following reaction list:

$$\varnothing \xrightarrow{a} \mathcal{R}$$
$$\mathcal{R} \xrightarrow{\kappa} \varnothing$$
$$\mathcal{R} \xrightarrow{\theta} \mathcal{R} + \mathcal{T}$$

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.

Let $r(t)$ denote the number of $\mathcal{R}$ molecules. If the number of regulator molecules is very large, we can accurately approximate $r(t)$ as a continuous stochastic process. The continuous process which best approximates the true discrete dynamics of $r(t)$ is described by the chemical Langevin equation (CLE) [35, 54], which says

$$\dot{r} = a - \kappa r + \sqrt{a + \kappa r}\ \xi(t),$$

where $\xi(t)$ is a Gaussian white noise term. A troublesome feature of this approximation is that the domain of $r(t)$ is $(-a/\kappa, \infty)$, i.e., it includes negative regulator concentrations; we can remedy this by making an additional approximation. If $r(t)$ spends most of its time around its mean value, $a \sim \kappa r$, we can write

$$\dot{r} \approx a - \kappa r + \sqrt{2\kappa r}\ \xi(t),$$

so that dynamics are now most naturally defined on $(0, \infty)$. The effective transcription rate $K(t) := \theta r(t)$ then satisfies the SDE

$$\dot{K} = a\theta - \kappa K + \sqrt{2\kappa\theta K}\ \xi(t). \tag{3}$$

This is the Cox–Ingersoll–Ross (CIR) model of transcription [49] we will study in this paper.

Although the CIR model is popular as a description of interest rates in quantitative finance [55–57], 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 [58–61]. Interestingly, the Γ-OU model can arise from an analogous analysis with Poisson shot noise synthesis of $\mathcal{R}$ [62].

If we are modeling the effect of RNA polymerase dynamics on transcription, the above derivation is fairly satisfying. However, if we are modeling the effect of an inducer or activator, the above derivation does not offer a mechanistic understanding of the gain parameter θ. If we wish to make the model slightly more biologically interpretable in such cases, and to keep Eq. 3 as a reasonable description of the transcription rate dynamics, we can instead consider the following modified underlying model.

Suppose the gene that produces $\mathcal{T}$ has two states, denoted by $\mathcal{G}_{off}$ and $\mathcal{G}_{on}$. Consider the reaction list:

$$\varnothing \xrightarrow{a} \mathcal{R}$$
$$\mathcal{R} \xrightarrow{\kappa} \varnothing$$
$$\mathcal{G}_{off} + \mathcal{R} \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} \mathcal{G}_{on}$$
$$\mathcal{G}_{on} \xrightarrow{k_{ini}} \mathcal{G}_{on} + \mathcal{T}$$

where k_{on} is the 'on' rate, k_{off} is the 'off' rate, and k_{ini} is the transcription initiation rate. If the binding of the regulator molecule $\mathcal{R}$ to the promoter is sufficiently fast and weak, this reaction list is well-described by the previous one with an effective gain $\theta = k_{ini}(k_{on}/k_{off})$.

To see why, let $g_{off}(t)$ denote the fraction of time the gene spends in the 'off' state, and $g_{on}(t)$ the fraction of time the gene spends in the 'on' state. If the binding/unbinding dynamics of $\mathcal{R}$ occur on a sufficiently fast time scale, then we can use a quasi-steady-state argument to treat the kinetics of binding and unbinding as roughly at equilibrium, so that

$$k_{on} r g_{off} = k_{off} g_{on}.$$

Combining this with the constraint that $g_{on} + g_{off} = 1$, we find

$$g_{off} = \frac{1}{1 + \frac{k_{on} r}{k_{off}}}$$
$$g_{on} = \frac{\frac{k_{on} r}{k_{off}}}{1 + \frac{k_{on} r}{k_{off}}}.$$

If the binding of $\mathcal{R}$ to the promoter is sufficiently weak (that is, if $k_{on}r \ll k_{off}$ for typical values of r), we can approximate the above expressions by first-order Taylor expansions

$$g_{off} \approx 1 - \frac{k_{on}r}{k_{off}}$$
$$g_{on} \approx \frac{k_{on}r}{k_{off}}.$$

Given these expressions, our effective transcription rate $K(t)$ is

$$\begin{aligned} K(t) &= 0 \cdot g_{off}(t) + k_{ini} \cdot g_{on}(t) \\ &\approx \left(\frac{k_{ini}k_{on}}{k_{off}}\right) r(t). \end{aligned}$$

For ease of reference, the SDEs describing the Γ-OU and CIR models are summarized in Table 2.

Model	Transcription rate SDE	Noise term	Associated biology
Γ-OU	$\dot{K}(t) = -\kappa K(t) + \epsilon(t)$	$\epsilon(t)$: Dirac delta process with arrival frequency a and exponentially distributed weights with expectation θ	Mechanical frustration and recovery
CIR	$\dot{K} = a\theta - \kappa K + \sqrt{2\kappa\theta K}\ \xi(t)$	$\xi(t)$: Gaussian white noise	Regulator dynamics

Table 2: The SDEs characterizing the transcription rate dynamics of the Γ-OU and CIR models.

3 Behavior of each transcription model is interpretable

With our modeling framework now motivated, we proceed to study the behavior and predictions of the Γ-OU and CIR models in the remainder of the paper. In what follows, we will pay particular attention to whether and when each model can be discriminated from the 'null' models (the constitutive/Poisson and negative binomial models), as well as from each other.

In this section, we study the qualitative predictions of the Γ-OU and CIR models. We show that the *behavior* of each model is interpretable, in the sense that predicted distribution shapes can be fairly informative regarding underlying parameter values. This allows one to make claims like: "According to this data, the time scale of DNA frustration is somewhat longer than the time scale of splicing." It also indicates that the biological details of transcription rate variation can indeed qualitatively affect RNA count distributions, allowing them to be distinguished from naïve Poisson and negative binomial models.

First, we describe the generic behavior of each model. Then we more exhaustively describe model behavior across a range of representative parameter sets. Finally, we discuss our framework for simulating such discrete-continuous hybrid systems.

3.1 Generic behavior of each model

Generically, the distribution shapes predicted by the Γ-OU and CIR models seem to interpolate between the Poisson and negative binomial extremes. To see why this makes biological sense, consider the effect of changing κ, the parameter that controls the extrinsic noise time scale. When κ is very fast, the transcription rate very quickly reverts to its mean value whenever it is perturbed, so the transcription rate is effectively constant: in this limit, both models predict Poisson distributions. When κ is very slow, the transcription rates of individual cells appear 'frozen' in time from the point of view of the time scales of RNA dynamics: in this case, both models predict negative binomial distributions for the nascent and mature marginals. As a consequence of this interpolation, variances (of both kinds of RNA, in both models) increase as κ becomes slower and decrease as κ becomes faster.

One can run through a similar heuristic argument to describe the effect of changing the process gain θ, the parameter that controls the scale of transcription rate fluctuations. If θ is very small, fluctuations in underlying biological factors (the DNA relaxation state or the number of regulator molecules) are damped, and do not significantly affect the transcription rate; hence, it is roughly constant, leading both models to predict Poisson distributions. If θ is very large, fluctuations are amplified, and predicted distributions become increasingly overdispersed.

Because the CIR model predicts somewhat heavier-tailed distributions than the Γ-OU model, and because increasing θ adds weight to distribution tails, the two models can significantly differ in this limit. This difference in tail region probabilities stems from fact that the noise term in the CIR model is state-dependent (i.e., a function of K), while the noise term in the Γ-OU model is not. In biological terms, the number of topoisomerases which randomly arrive to relieve mechanical stress does not depend on the current relaxation state of the DNA; on the other hand, birth-death fluctuations in the number of regulator molecules tend to be greater when there are more regulator molecules present.

3.2 Behavior across representative parameter regimes

To get a more detailed picture of model behavior, we visualized model predictions—including autocorrelation functions (Figure 4) and full long-time RNA count distributions (Figures 5 and 6)—for six representative parameter sets. Table 3 reports the parameters used to define the regimes of interest.

Parameter set	κ	a	θ	T_{ss}	T_R
High gain	10	0.1	150	20	10
Slow reversion	0.12	0.01	15	200	50
Low gain	8.33×10^{-4}	0.1	0.05	60	10
Fast reversion	100	100	14.93	7.143	10
Intermediate 1	0.6765	2.3	0.7692	7.391	10
Intermediate 2	1.25	4.25	1.493	7.143	10

Table 3: Representative parameter sets used to explore model predictions. Four parameter sets lie in limiting regimes, while two lie in intermediate regimes. In all cases, $\beta = 1.2$ and $\gamma = 0.7$ were used. Simulations tracked 10^4 cells until an 'equilibration' time T_{ss}, and then continued until a time $T_{ss} + T_R$ to compute autocorrelation functions.

Autocorrelation functions quantify how a stochastic system approaches equilibrium. In our case, they answer the question: "How correlated are nascent/mature RNA counts right now with nascent/mature RNA counts some time τ in the future?" In principle, because they depend on model details, experimental measurements of autocorrelation functions from live-cell data can be used to discriminate between competing models. But the autocorrelation functions of the Γ-OU and CIR models exactly match (Figure 4), eliminating this as a discrimination method.

Γ-OU distribution shape predictions are shown in Figure 5, and CIR distribution shape predictions are shown in Figure 6. Overall, the plots are consistent with the intuition developed in the previous section: both joint and marginal distributions appear to interpolate between Poisson-like and overdispersed. In spite of the similarities of each model's predictions, tail predictions significantly differ in the high gain regime (where θ and κ are both very large), as previously discussed. This difference is somewhat larger for the mature count distribution than for the nascent count distribution (compare the third row, second column of Figures 5 and 6).

3.3 Simulation framework for numerically solving models

Numerically extracting model predictions for models like these requires developing efficient methods for simulating hybrid discrete-continuous stochastic systems. Both computationally and conceptually, this is somewhat challenging. Because our models involve no feedback, we split this problem into two parts: first, simulate the continuous stochastic dynamics of the transcription rate $K(t)$; then, simulate the discrete stochastic dynamics of the nascent and mature RNA using a variant of Gillespie's direct method [43].

The first step is similar in spirit for both models: use specific technical knowledge to exactly compute $K(t)$, without directly simulating the corresponding SDE. In the case of the Γ-OU model, $K(t)$ is a weighted sum of exponentially distributed random variables [63]. In the case of the CIR model, $K(t)$ can be simulated by drawing from a noncentral chi-squared distribution (see

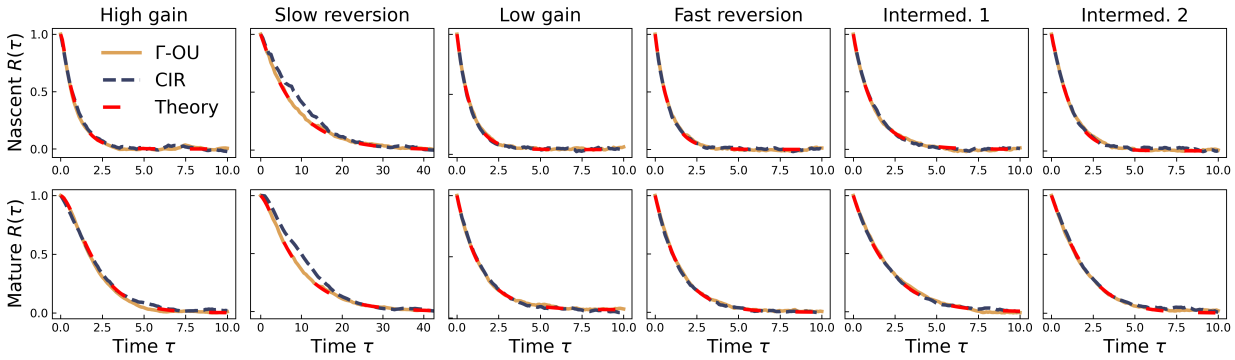

Figure 4: Comparison of theoretical and simulated autocorrelation functions. First row: autocorrelation of $\mathcal{N}$ counts at equilibrium. Second row: autocorrelation of $\mathcal{M}$ counts at equilibrium.

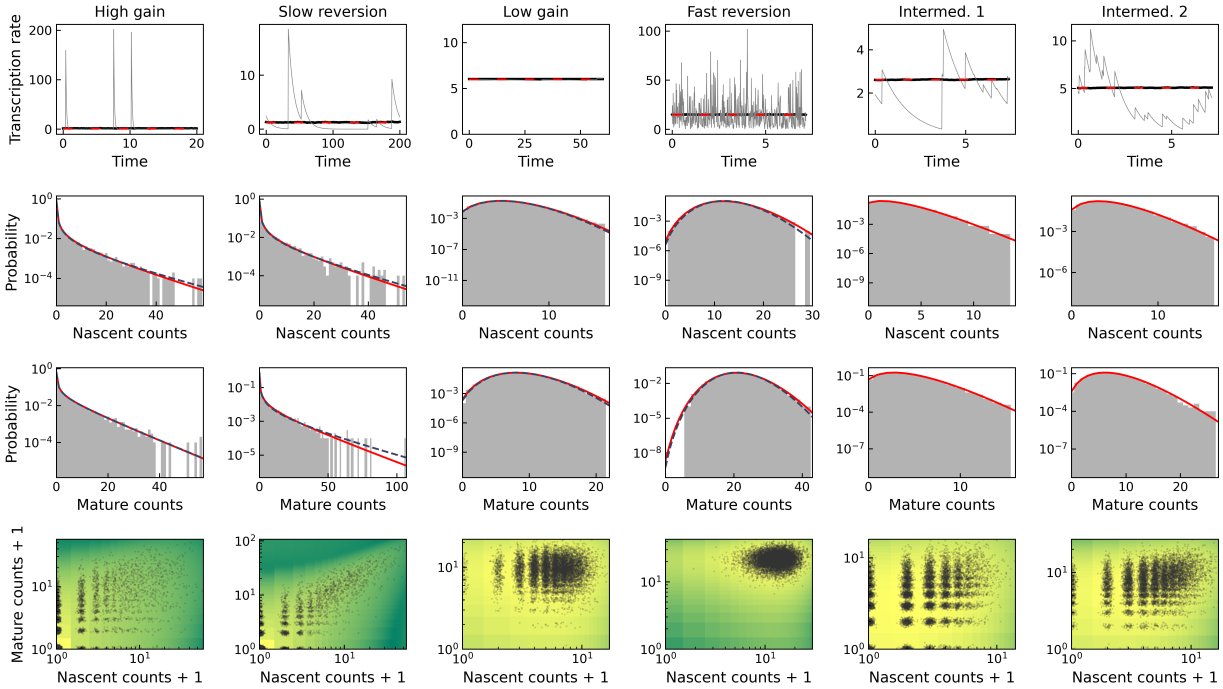

Figure 5: Γ-OU simulation results in six regimes, compared to steady-state solutions obtained by numerical integration and exact closed-form solutions to limiting cases. The simulations closely match the analytical results.

Top row: transcription rate time series (black line: mean of all simulation; grey line: single simulation, red dashed line: expected stationary mean). Second row: nascent RNA stationary distributions (grey histogram: observed distribution; red line: expected analytical distribution; dashed blue line: limiting regime solution). Third row: mature RNA stationary distributions. Bottom row: empirical joint distribution (color: log analytical joint probability mass function (PMF); black points: cells; normal jitter with $\sigma = 0.05$ added).

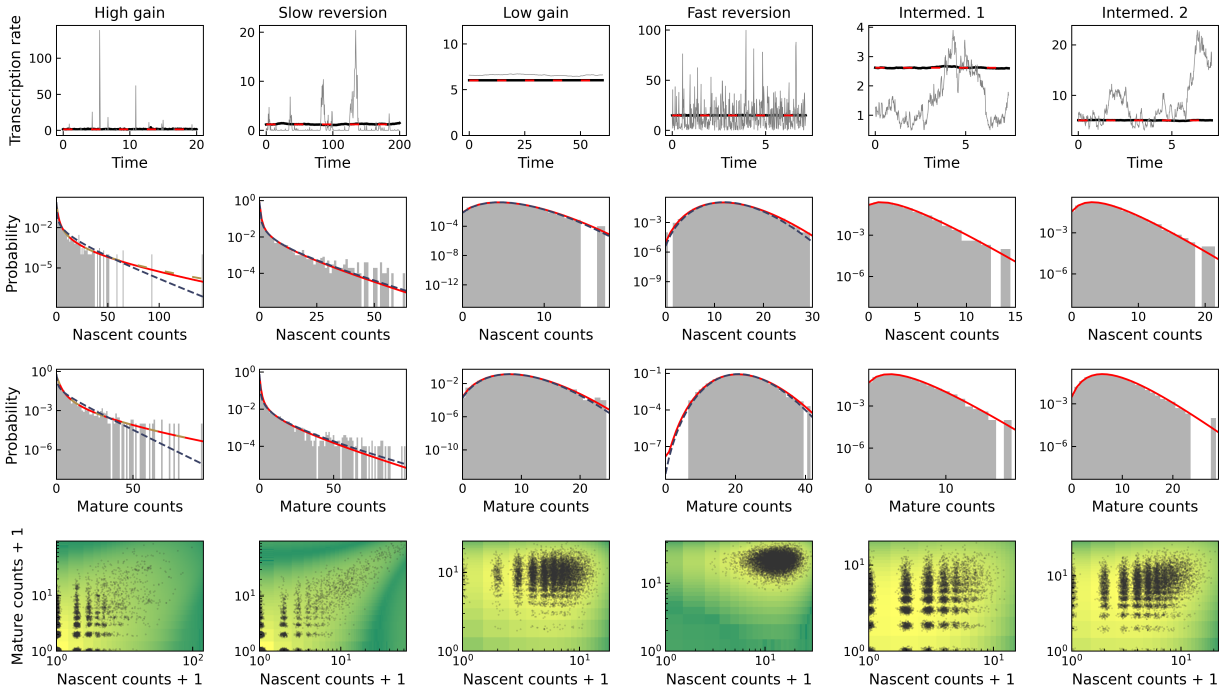

Figure 6: CIR simulation results in six regimes, compared to steady-state solutions obtained by numerical integration and exact closed-form solutions to limiting cases. The simulations closely match the analytical results; the intrinsic regime converges to distributions distinct from the corresponding Γ-OU limit.
All parameters and conventions as in Figure 5. Gold dashed line: OU-IG solution.

Section S2). The second step differs somewhat for each model. Gillespie's direct method requires knowing $\int K(t')dt'$, which can be explicitly calculated using special functions in the case of the Γ-OU model [63]. Because it cannot be computed exactly for the CIR model, $\int K(t')dt'$ was approximated by evaluating $K(t)$ (in the previous step) at many closely spaced time points [64], and using the trapezoidal rule [43, 65, 66].

The full details of the Γ-OU simulation procedure were described previously in [63], while the full details of the CIR simulation procedure are described in Section S2. One advantage of this simulation framework is its flexibility: it can be readily extended to treat more complicated variants of the systems we have considered, even if those generalizations are no longer mathematically tractable. For example, it is straightforward to incorporate (i) more realistic splicing kinetics (involving multiple introns that are removed in an arbitrary order), (ii) proteins, and (iii) gene-protein feedback.

4 Behavior of each transcription model is tractable

While the qualitative analysis of the previous section was insightful, we can make our intuitions about these models more concrete by translating them into mathematical statements. For example: what exactly do we *mean* when we say that the Γ-OU and CIR models interpolate between Poisson and negative binomial extremes? What formula does the interpolating, and under what conditions does one recover the limiting cases?

A thorough mathematical study of these models can serve at least two important purposes. First, theoretical results can validate our simulations. Second, making intuitions about these models concrete makes them *actionable*: we can use mathematical formulas describing model properties to more efficiently distinguish between competing models, and to more precisely understand how underlying biological features affect model predictions.

By design, the Γ-OU and CIR transcription models are mathematically tractable, which allows us to derive exact expressions for many useful model predictions by hand. In this section, we present a detailed mathematical analysis of the behavior of each model, all of which supports the heuristic picture described in the previous section. Among our theoretical results are exact formulas for the steady-state distributions, moments, autocorrelation functions, and limiting distributions.

4.1 Summary of mathematical calculations and approach

The models in our model class keep track of three things: nascent transcripts $\mathcal{N}$ (whose number we denote by $X_N \in \mathbb{N}_0$), mature transcripts $\mathcal{M}$ (whose number we denote by $X_M \in \mathbb{N}_0$), and the transcription rate $K(t) = K_t \in (0, \infty)$. Each of these evolves in time according to a stochastic process.

In some sense, obtaining a full mathematical understanding of the stochastic dynamics of these systems reduces to exactly computing the probability density $P(x_N, x_M, K, t)$, which quantifies the probability that the system is in the state $(X_N = x_n, X_M = x_M, K_t = K)$ at some time t (given some initial condition $P_0(x_N, x_M, K, 0)$).

In the case of the constitutive model, the equation characterizing the time evolution of the probability density was Eq. 1; for these more complex models, the time evolution of $P(x_N, x_M, K, t)$ is also completely characterized by a master equation (see Section S3). For the Γ-OU model, it is

$$\begin{aligned}\frac{\partial P(x_N, x_M, K, t)}{\partial t} = \; & K\left[P(x_N - 1, x_M, K, t) - P(x_N, x_M, K, t)\right] \\ & + \beta\left[(x_N + 1)P(x_N + 1, x_M - 1, K, t) - x_N P(x_N, x_M, K, t)\right] \\ & + \gamma\left[(x_M + 1)P(x_N, x_M + 1, K, t) - x_M P(x_N, x_M, K, t)\right] \\ & - \frac{\partial}{\partial K}\left[(-\kappa K)\, P(x_N, x_M, K, t)\right] + a\sum_{n=1}^{\infty}(-\theta)^n \frac{\partial^n}{\partial K^n}\left[P(x_N, x_M, K, t)\right].\end{aligned}$$

For the CIR model, this equation is

$$\begin{aligned}\frac{\partial P(x_N, x_M, K, t)}{\partial t} = \; & K\left[P(x_N - 1, x_M, K, t) - P(x_N, x_M, K, t)\right] \\ & + \beta\left[(x_N + 1)P(x_N + 1, x_M - 1, K, t) - x_N P(x_N, x_M, K, t)\right] \\ & + \gamma\left[(x_M + 1)P(x_N, x_M + 1, K, t) - x_M P(x_N, x_M, K, t)\right] \\ & - \frac{\partial}{\partial K}\left[(a\theta - \kappa K)\, P(x_N, x_M, K, t)\right] + \kappa\theta \frac{\partial^2}{\partial K^2}\left[K P(x_N, x_M, K, t)\right].\end{aligned}$$

Our task is essentially to solve these two equations—or, at least, to understand their behavior well enough to extract experimentally relevant properties and summaries. We restrict our analysis to long-time/steady-state probability distributions, which describe 'natural' equilibria independent of the system's initial condition:

$$P_{ss}(x_N, x_M, K) := \lim_{t\to\infty} P(x_N, x_M, K, t).$$

As suggested by Figure 4, the transcriptional systems approach these equilibria exponentially fast. Because the transcription rate is usually not measurable, we are also primarily interested in distributions marginalized over K, i.e.

$$P_{ss}(x_N, x_M) := \int_0^\infty dK\ P_{ss}(x_N, x_M, K).$$

Aside from the steady-state distributions marginalized over K, we are also interested in steady-state first- and second-order moments (e.g. means and variances), which offer a partial look at how transcription rate details can affect the scale and dispersion of count distributions, and autocorrelation functions, which quantify these systems' approach to equilibrium.

To compute these quantities, we use a variety of tricks from theoretical physics and the mathematics of stochastic processes. For example, we solve the Γ-OU model by identifying a mathematical correspondence between it and the well-known bursting model of transcription; we solve the CIR model by computing $P(x_N, x_M, K, t)$ using a state-space path integral representation [54, 67].

A central idea in all of our calculations is to consider the discrete Fourier transform of the probability density, the so-called *probability generating function* (PGF), instead of the probability density itself. In general, it is defined as

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

with $g_N, g_M \in \mathbb{C}$ both on the complex unit circle and $h \in \mathbb{R}$. As with the probability density, it is helpful to consider variants marginalized over the transcription rate and/or with the $t \to \infty$ limit taken. We are mainly interested in $\psi_{ss}(g_N, g_M)$, the PGF of $P_{ss}(x_N, x_M)$.

The generating function ψ satisfies a somewhat simpler equation than $P(x_N, x_M, K, t)$, and can be exploited to compute moments and autocorrelation functions. Moreover, there is no loss of information in considering the generating function, because one can straightforwardly recover the probability density from it via an inverse Fourier transform:

$$P(x_N, x_M, K, t) = \int_{-\infty}^{\infty} \frac{dh}{2\pi} \oint \frac{dg_N dg_M}{(2\pi i)^2}\ \frac{1}{g_N^{x_N+1} g_M^{x_M+1}} e^{-ihK} \psi(g_N, g_M, h, t).$$

Numerically, this step can be performed extremely efficiently using the inverse fast Fourier transform [46, 68]. For technical reasons, we will also consider the so-called *factorial-cumulant generating function* ϕ, defined via

$$\phi(u_N, u_M, h, t) := \log \psi(g_N, g_M, h, t)$$

whose RNA-related arguments are written as $u_N := g_N - 1$ and $u_M := g_M - 1$. The steady-state version of this marginalized over transcription rate, i.e. $\phi_{ss}(u_N, u_M) := \log \psi_{ss}(g_N, g_M)$, is what we will use to report our answers for the steady-state distributions of the Γ-OU and CIR models.

In the rest of the section, we report our mathematical results. For reference purposes, we discuss our notation in Section 4.2. In Section 4.3, we outline our solution methodology in more detail. Finally, in Section 4.4, we report moments, autocorrelation functions, and limiting distributions, which are mostly identical for the Γ-OU and CIR models.

4.2 Notation

A guide to important notation is presented in Table 4. Below, we describe our notation for common probability distributions.

Symbol	Meaning
$K(t), K_t$	Stochastic and time-varying transcription rate $\in (0, \infty)$
$\langle K \rangle$	Mean transcription rate at steady state, $\langle K \rangle = (a\theta)/\kappa = \alpha/\kappa$
$X_N \in \mathbb{N}_0$	Nascent RNA copy number
$X_M \in \mathbb{N}_0$	Mature RNA copy number
$P(x_N, x_M, K, t)$	Density of state $(x_N, x_M, K) \in \mathbb{N}_0 \times \mathbb{N}_0 \times (0, \infty)$ at time t
$P_{ss}(x_N, x_M, K)$	Steady-state density of state $(x_N, x_M, K) \in \mathbb{N}_0 \times \mathbb{N}_0 \times (0, \infty)$
$P_{ss}(x_N, x_M)$	Steady-state probability of observing (x_N, x_M) RNA counts
$\psi(g_N, g_M, h, t)$	Generating function of $P(x_N, x_M, K, t)$ (see Equation 4.1)
$\psi_{ss}(g_N, g_M, h)$	Steady-state generating function of $P_{ss}(x_N, x_M, K)$
$\psi_{ss}(g_N, g_M)$	Steady-state generating function of $P_{ss}(x_N, x_M)$
$\phi(u_N, u_M, h, t)$	Factorial-cumulant generating function $\log \psi(u_N + 1, u_M + 1, h, t)$
$\phi_{ss}(u_N, u_M, h)$	Factorial-cumulant generating function $(t \to \infty)$, $\log \psi_{ss}(u_N + 1, u_M + 1, h)$
$\phi_{ss}(u_N, u_M)$	Factorial-cumulant generating function $(t \to \infty,\ h = 0)$, $\log \psi_{ss}(u_N + 1, u_M + 1)$
μ_N	Mean nascent RNA count at steady state
μ_M	Mean mature RNA count at steady state
σ_N^2	Variance of nascent RNA count at steady state
σ_M^2	Variance of mature RNA count at steady state
$\mathrm{Cov}(X_N, X_M)$	Covariance of nascent and mature RNA counts at steady state
$\mathrm{Cov}(X_N, K)$	Covariance of nascent RNA count and transcription rate at steady state
$\mathrm{Cov}(X_M, K)$	Covariance of mature RNA count and transcription rate at steady state
$R_N(\tau)$	Autocorrelation of $\mathcal{N}$ (normalized by its variance) at lag time τ
$R_M(\tau)$	Autocorrelation of $\mathcal{M}$ (normalized by its variance) at lag time τ

Table 4: Probability distributions, generating functions, and moments of interest.

- The Poisson distribution is defined as follows: if $X \sim Poisson(\lambda)$, $P(X = k; \lambda) = \frac{1}{k!}\lambda^k e^{-\lambda}$, where $k \in \mathbb{N}_0$ and $\lambda > 0$.
- The geometric distribution is defined as follows: if $X \sim Geom(p)$, $P(X = k; p) = (1-p)^k p$, where $k \in \mathbb{N}_0$ and $p \in (0, 1]$. The geometric distribution is well-known to arise in the short-burst limit of the two-state transcription model [69].
- The negative binomial distribution is defined as follows: if $X \sim NegBin(r, p)$, $P(X = k; r, p) = \frac{\Gamma(r+k)}{k!\Gamma(r)}(1-p)^r p^k$, where $k \in \mathbb{N}_0$, $p \in [0, 1]$, and $r > 0$. We note that MATLAB and the NumPy library take the opposite convention, with a $\tilde{p}$ parameter defined as $1 - p$.

- The exponential distribution is defined in two alternative ways: if $X \sim Exp(\eta)$, $f(x;\eta) = \eta e^{-\eta x}$, where $x, \eta > 0$. This is the rate parametrization. Conversely, MATLAB and the NumPy library take the opposite scale parametrization, with parameter $\theta = \eta^{-1}$.

- The gamma distribution is defined in two alternative ways: if $X \sim Gamma(\alpha, \eta)$, $f(x;\alpha,\eta) = \frac{\eta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\eta x}$, where $x, \alpha, \eta > 0$. This is the shape/rate parametrization. Conversely, the MATLAB and the NumPy library take the opposite shape/scale parametrization with parameter $\theta = \eta^{-1}$. In the literature, the rate η is usually given the variable name 'β'; however, we use the current convention to preclude confusion with the splicing rate parameter. $Exp(\eta)$ is equivalent to $Gamma(1, \eta)$.

- The normal distribution $N(\mu, \sigma^2)$ has probability density $f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\big(-\frac{1}{2\sigma^2}(x-\mu)^2\big)$.

- The continuous uniform distribution $U(a, b)$, used in simulation, has density $f(x;a,b) = (b-a)^{-1}$ on $[a, b]$ and 0 elsewhere.

- The inverse Gaussian distribution $IG(A, B)$ arises in the high gain limit of the CIR model and has probability density $f(x;A,B) = \frac{A}{\sqrt{2\pi}} e^{AB} x^{-3/2} \exp\big(-\frac{1}{2}(A^2 x^{-1} + B^2 x)\big)$, where $x, A, B > 0$.

4.3 Steady-state probability distribution solutions

In this section, we sketch our approach to solving the Γ-OU and CIR models. A complete treatment of each problem can be found in the appendix (Section S3.2 for Γ-OU and Section S3.3 for CIR).

4.3.1 Gamma Ornstein–Uhlenbeck model

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; it is fully outlined in Section S3.2.

In brief, we may consider a bursting gene coupled to a 3-step birth-death process, characterized by the path graph $\varnothing \xrightarrow{k} B \times \mathcal{T}_0 \xrightarrow{\beta_0} \mathcal{T}_1 \xrightarrow{\beta} \mathcal{T}_2 \xrightarrow{\gamma} \varnothing$, where $B \sim Geom(b)$. The stochastic process describing the Poisson intensity of $\mathcal{T}_0$ is precisely the Γ-OU process [70]. This implies that the joint distribution of species $\{\mathcal{T}_1, \mathcal{T}_2\}$ in the burst-driven process is identical to the joint distribution of $\{\mathcal{N}, \mathcal{M}\}$ in the system driven by Γ-OU transcription, under the substitutions $b \leftarrow \theta/\kappa$, $k \leftarrow a$, and $\beta_0 \leftarrow \kappa$. Indeed, this relation holds with no loss of generality for an arbitrary number of species.

The generating function of SDE-driven system can be immediately computed using the solution of the bursty system:

$$\phi_{ss}(u_N, u_M) = a \int_0^\infty \frac{\frac{\theta}{\kappa} U_0(s; 0, u_N, u_M)}{1 - \frac{\theta}{\kappa} U_0(s; 0, u_N, u_M)} ds, \tag{4}$$

where $U_0(s; 0, u_N, u_M)$ is the exponential sum solution of the following ODE system:

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

This system of linear first-order ODEs can be straightforwardly solved to find that

$$U_0 = A_0 e^{-\kappa s} + A_1 e^{-\beta s} + A_2 e^{-\gamma s} \tag{5}$$

with

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

Alternatively, these coefficients can be computed numerically, which may be useful for treating generalizations of this system involving a more realistic splicing topology. An algorithm for this is provided in [48].

4.3.2 Cox–Ingersoll–Ross model

The CIR production rate model can be solved using a state space path integral representation of the master equation, as described in Section S3.3 and [67]. This involves writing the most general probability of a transition between two states as a sum over all possible transition paths. Simplifying the resulting expression involves evaluating many integrals and sums, the end result of which allows us to write $\phi_{ss}(u_N, u_M) = \log \psi_{ss}(u_N, u_M)$ as

$$\phi_{ss}(u_N, u_M) = a\theta \int_0^\infty U(s; u_N, u_M) ds$$

where $U(s; u_N, u_M)$ is the solution to

$$\frac{dU}{ds} = -\kappa U + \kappa\theta U^2 + \left[\left(u_N - \frac{\beta}{\beta - \gamma} u_M \right) e^{-\beta s} + \frac{\beta}{\beta - \gamma} u_M e^{-\gamma s} \right]$$

with initial condition $U(s = 0) = 0$. While the above ODE has an exact solution [67], it is somewhat cumbersome. Fortunately, it is more computationally efficient to solve the ODE numerically.

4.4 Distribution properties

Several salient observables of the two transcription rate models are identical. These include low order moments, autocorrelation functions, and certain limiting cases of the steady-state probability distribution. In this subsection, we report these matching features.

4.4.1 Moments

The full derivation of our moment results is provided in Sections S3.2 and S3.3. In brief, the moments can be computed by taking partial derivatives of the PGF and evaluating them at $g_N = g_M = 1$ (or equivalently but more usefully, taking derivatives of ϕ_{ss} and evaluating them at $u_N = u_M = 0$). For example:

$$\mu_N = \sum_{x_N=0}^{\infty} \sum_{x_M=0}^{\infty} x_N P_{ss}(x_N, x_M) = \left. \frac{\partial \psi_{ss}(g_N, g_M)}{\partial g_N} \right|_{g_N=g_M=1}.$$

Our results are reported in Table 5, along with side-by-side comparisons to moment results for the Poisson/constitutive model (Section 2.1), the negative binomial model (Section 2.2), and the bursty model of RNA production. The Γ-OU and CIR models, whose first and second order moments all match, apparently generalize the moment results from those more naïve models.

Moment	Γ-OU and CIR	Poisson model	NB model	Bursty model
μ_N	$\langle K \rangle / \beta$	$\langle K \rangle / \beta$	$\langle K \rangle / \beta$	$\langle K \rangle / \beta$
μ_M	$\langle K \rangle / \gamma$	$\langle K \rangle / \gamma$	$\langle K \rangle / \gamma$	$\langle K \rangle / \gamma$
$\sigma_N^2 - \mu_N$	$(\mu_N \theta)/(\kappa + \beta)$	0	$(\mu_N \theta)/\beta$	$(\mu_N \theta)/\kappa$
$\sigma_M^2 - \mu_M$	$\frac{\mu_M \beta \theta (\kappa+\beta+\gamma)}{(\kappa+\beta)(\kappa+\gamma)(\beta+\gamma)}$	0	$(\mu_M \theta)/\gamma$	$\frac{\mu_M \theta}{\kappa} \frac{\beta}{\beta+\gamma}$
$\mathrm{Cov}(X_N, X_M)$	$\frac{\langle K \rangle \theta (\kappa+\beta+\gamma)}{(\kappa+\beta)(\kappa+\gamma)(\beta+\gamma)}$	0	$\frac{\langle K \rangle \theta}{\beta \gamma}$	$\frac{\langle K \rangle \theta}{\kappa(\beta+\gamma)}$

Table 5: Moments of the Γ-OU and CIR models (Γ-OU and CIR), the constitutive/Poisson model (Poisson model), and the negative binomial model from Section 2.2 (NB model). Note that the Γ-OU and CIR results match the Poisson results in the $\kappa \to \infty$ limit and $\theta \to 0$ limit; they match the negative binomial results in the $\kappa \to 0$ limit.

These results allow us to revisit well-known noise source decompositions from the stochastic gene expression literature [6, 7, 9], which describe how intrinsic and extrinsic noise sources contribute to overall cell-to-cell variation in RNA copy numbers. For models in our model class, intrinsic noise is due to randomness associated with the timing of transcription, splicing, and degradation; meanwhile, extrinsic noise is due to variation in the transcription rate $K(t)$.

Define the squared coefficient of variation, $\eta^2 := \sigma^2/\mu^2$. For the nascent species, we have

$$\begin{aligned}
\eta_N^2 &= \frac{1}{\mu_N}\left[1 + \frac{\theta}{\kappa+\beta}\right] = \varsigma_{N,int}^2 + \varsigma_{N,ext}^2 \\
\eta_{N,int}^2 &= \frac{1}{\mu_N} \\
\eta_{N,ext}^2 &= \frac{1}{\mu_N}\frac{\theta}{\kappa+\beta} = \frac{\beta}{\langle K\rangle}\frac{\theta}{\kappa+\beta} = \frac{\theta}{\langle K\rangle}\frac{1/\kappa}{1/\kappa+1/\beta},
\end{aligned}$$

which exactly matches previously derived results [9], as long as the average environmental signal is appropriately normalized by its scale θ to provide a nondimensional η_N^2. For the mature species, we have

$$\begin{aligned}
\eta_M^2 &= \frac{1}{\mu_M}\left[1 + \frac{\beta\theta(\kappa+\beta+\gamma)}{(\kappa+\beta)(\kappa+\gamma)(\beta+\gamma)}\right] = \varsigma_{M,int}^2 + \varsigma_{M,ext}^2 \\
\eta_{M,int}^2 &= \frac{1}{\mu_M} \\
\eta_{M,ext}^2 &= \frac{1}{\mu_M}\left[\frac{\beta\theta(\kappa+\beta+\gamma)}{(\kappa+\beta)(\kappa+\gamma)(\beta+\gamma)}\right] = \frac{\theta}{\langle K\rangle}\left[\frac{\beta\gamma(\kappa+\beta+\gamma)}{(\kappa+\beta)(\kappa+\gamma)(\beta+\gamma)}\right].
\end{aligned}$$

Our results are more complicated, but consistent with previous ideas about the separability of intrinsic and extrinsic noise sources. As one might expect, the contribution of extrinsic noise to each coefficient of variation vanishes when transcription rate fluctuations become negligible (for example, when the rate of mean-reversion κ is very fast, or when the gain θ is very small).

Interestingly enough, in spite of substantial differences in the details of the model, the combinatorial form of the extrinsic noise result matches the form of a result previously derived for a two-state model of transcription [7], with $\ell \leftarrow \kappa$ aggregating the gene locus timescales and $n \leftarrow \langle K\rangle/\theta$ describing the gene copy number or promoter strength. However, in the current model, the extrinsic noise contribution is *positive* rather than negative, because there is no constraint on the promoter strength.

4.4.2 Autocorrelation functions

We define the normalized autocorrelation function of a stationary process X_t with mean μ and variance σ^2 as follows:

$$R(\tau) := \lim_{t\to\infty}\frac{1}{\sigma^2}\mathbb{E}[(X_t-\mu)(X_{t+\tau}-\mu)]$$

where the expectation here is taken over all possible stochastic trajectories. We can also define autocorrelation functions in terms of transition probabilities (see Section S4.3 for the details, and for a full derivation of autocorrelation results). To actually compute autocorrelation functions, we defined a special generating function relating the system's behavior at times t and $t+\tau$ and took partial derivatives.

The autocorrelation of the nascent species takes the following functional form:

$$\begin{aligned}
R_N(\tau) &= e^{-\beta\tau} + \frac{\mathrm{Cov}(x_N, K)}{\sigma_N^2}\frac{(e^{-\kappa\tau}-e^{-\beta\tau})}{\beta-\kappa} \\
&= e^{-\beta\tau} + \frac{\theta\beta}{\beta+\kappa+\theta}\frac{(e^{-\kappa\tau}-e^{-\beta\tau})}{\beta-\kappa}.
\end{aligned}$$

The autocorrelation function of the mature species, found using the same method, is:

$$
\begin{aligned}
R_M(\tau) &= e^{-\gamma\tau} + \beta\,\frac{\mathrm{Cov}(X_N, X_M)}{\sigma_M^2}\,\frac{\left[e^{-\beta\tau} - e^{-\gamma\tau}\right]}{\gamma-\beta} \\
&\quad + \beta\,\frac{\mathrm{Cov}(X_M, K)}{\sigma_M^2}\left[\frac{e^{-\beta\tau}}{(\beta-\gamma)(\beta-\kappa)} + \frac{e^{-\gamma\tau}}{(\gamma-\beta)(\gamma-\kappa)} + \frac{e^{-\kappa\tau}}{(\kappa-\beta)(\kappa-\gamma)}\right] \\
&= e^{-\gamma\tau} + \frac{\theta\gamma(\beta+\gamma+\kappa)}{\theta\beta(\beta+\gamma+\kappa) + (\beta+\kappa)(\gamma+\kappa)(\beta+\gamma)}\,\beta\frac{\left[e^{-\beta\tau} - e^{-\gamma\tau}\right]}{\gamma-\beta} \\
&\quad + \frac{\theta\beta\gamma(\beta+\gamma)}{\theta\beta(\beta+\gamma+\kappa) + (\beta+\kappa)(\gamma+\kappa)(\beta+\gamma)}\,\beta \\
&\quad \times \left[\frac{e^{-\beta\tau}}{(\beta-\gamma)(\beta-\kappa)} + \frac{e^{-\gamma\tau}}{(\gamma-\beta)(\gamma-\kappa)} + \frac{e^{-\kappa\tau}}{(\kappa-\beta)(\kappa-\gamma)}\right].
\end{aligned}
$$

In the limit of very fast κ, each term but the first vanishes in both $R_N(\tau)$ and $R_M(\tau)$, so that we recover the autocorrelation functions of the constitutive model. In the limit of very fast splicing ($\beta \to \infty$), $R_M(\tau)$ matches the $R_N(\tau)$ result with γ in place of β—i.e., the system effectively behaves as if there is only one kind of RNA species.

4.5 Limiting cases

We have already discussed several interesting limits, and in this subsection catalog them more systematically. These limits are derived in Section S5, summarized in Table 6, and depicted schematically in Figure 7.

Limit	Parameter conditions	Held fixed	Nascent	Mature
Fast reversion	$\kappa \to \infty,\ a \to \infty$	$\alpha := a/\kappa$	Poisson	Poisson
Slow reversion	$\kappa \to 0,\ a \to 0$	$\alpha := a/\kappa$	NB	NB
Low gain	$\theta \to 0,\ \kappa \to 0$	$b := \theta/\kappa$	Poisson	Poisson
High gain	$\theta \to \infty,\ \kappa \to \infty$	$b := \theta/\kappa$	NB (Γ-OU) see text (CIR)	see text (Γ-OU) see text (CIR)

Table 6: Four interesting limiting regimes for both models. In each regime, two parameters are both taken to either infinity or zero, with their ratio held fixed to avoid degeneration. For the fast and slow reversion limits, the shape parameter $\alpha := a/\kappa$ is held fixed. For the low and high gain limits, the burst size $b := \theta/\kappa$ is held fixed. The last two columns ('Nascent' and 'Mature') indicate the steady-state marginal distributions of nascent and mature counts in each limit. NB: negative binomial. Limiting distributions match for the Γ-OU and CIR models except in the high gain limit.

4.5.1 Fast mean-reversion limit ($\kappa \to \infty$, $a \to \infty$, a/κ fixed)

In the fast mean-reversion limit ($\kappa \to \infty$ and $a \to \infty$ with $\alpha := \alpha/\kappa$ held constant), the transcription rate dynamics are much more rapid than mRNA processing, so we expect the effect of the trajectory shape to vanish. By computing the exact solution, we find that the effect is even more severe and everything but the *location* of the transcription rate distribution ceases to matter: the limit is

equivalent to a constitutive-like mean-field treatment. Quantitatively, we recover uncorrelated bivariate Poisson distributions for both models.

$$\phi_{ss}(u_N, u_M) = \frac{\langle K \rangle}{\beta} u_N + \frac{\langle K \rangle}{\gamma} u_M$$
$$P_{ss}(x_N, x_M) = \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!} \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!}.$$

4.5.2 Slow mean-reversion limit ($\kappa \to 0$, $a \to 0$, a/κ fixed)

In the slow mean-reversion limit ($\kappa \to 0$ and $a \to 0$ with $\alpha := \alpha/\kappa$ held constant), κ is so small that the transcription rates of each cell in a population do not change much on experimental time scales; for this reason, we expect the system to behave as if each cell's transcription rate is 'frozen' in time, with the distribution of these transcription rates corresponding to the long-time distribution of $K(t)$ (i.e., a gamma distribution). This suggests we should recover the Poisson-gamma mixture model from Section 2.2, which turns out to be true for both models:

$$\phi_{ss}(u_N, u_M) = -\alpha \log\left[1 - \theta\left(\frac{u_N}{\beta} + \frac{u_M}{\gamma}\right)\right]$$
$$P_{ss}(x_N, x_M) = \int_0^\infty dK \, \frac{K^{\alpha-1} \, e^{-K/\theta}}{\theta^\alpha \, \Gamma(\alpha)} \, \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!} \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!}$$
$$P_{ss}(x_N) = \binom{x_N + \alpha - 1}{x_N} \left(\frac{\beta}{\theta + \beta}\right)^\alpha \left(\frac{\theta}{\theta + \beta}\right)^{x_N}$$
$$P_{ss}(x_M) = \binom{x_M + \alpha - 1}{x_M} \left(\frac{\gamma}{\theta + \gamma}\right)^\alpha \left(\frac{\theta}{\theta + \gamma}\right)^{x_M}.$$

We remind the reader that the marginal distributions $P_{ss}(x_N)$ and $P_{ss}(x_M)$ are both negative binomial for this mixture model.

4.5.3 Low gain limit ($\theta \to 0$, $\kappa \to 0$, θ/κ fixed)

In the low gain limit ($\theta \to 0$ and $\kappa \to 0$ with $b := \theta/\kappa$ held constant), the gain θ is so small that fluctuations in the underlying biology (the DNA's relaxation state in the case of Γ-OU, and the concentration of regulator molecules in the case of CIR) hardly impact the transcription rate $K(t)$, leaving it effectively constant; as in the fast mean-reversion limit, we expect to recover constitutive model-like behavior. Once again, we indeed obtain Poisson distributions in this limit for both models:

$$\phi_{ss}(u_N, u_M) = \frac{\langle K \rangle}{\beta} u_N + \frac{\langle K \rangle}{\gamma} u_M$$
$$P_{ss}(x_N, x_M) = \frac{\left(\frac{K}{\beta}\right)^{x_N} e^{-K/\beta}}{x_N!} \frac{\left(\frac{K}{\gamma}\right)^{x_M} e^{-K/\gamma}}{x_M!}.$$

4.5.4 High gain limit ($\theta \to \infty$, $\kappa \to \infty$, θ/κ fixed)

The high gain limit ($\theta \to \infty$ and $\kappa \to \infty$ with $b := \theta/\kappa$ held constant), in which the gain θ is so high that fluctuations in the underlying biology become greatly amplified, is somewhat more interesting

and subtle than the others. This is the only limiting regime in which the predictions of the Γ-OU and CIR models markedly differ, and the only regime in which the mathematics associated with taking the limit becomes substantially more demanding. One obvious reason for this is that the transcription rate dynamics become somewhat singular: for example, the steady-state transcription rate variance $\sigma_K^2 = \langle K \rangle \theta$ (see Section S4.2) becomes infinite.

In this limit, the Γ-OU model precisely recapitulates the well-known bursting model of transcription, which has RNA produced in geometrically-distributed bursts. According to earlier work, the solution to this system is [46]:

$$\phi_{ss}(u_N, u_M) = a \int_0^\infty \frac{b\left[\left(u_N - u_M \frac{\beta}{\beta-\gamma}\right)e^{-\beta s} + \frac{\beta}{\beta-\gamma} u_M e^{-\gamma s}\right]}{1 - b\left[\left(u_N - u_M \frac{\beta}{\beta-\gamma}\right)e^{-\beta s} + \frac{\beta}{\beta-\gamma} u_M e^{-\gamma s}\right]} ds.$$

While there is no nice way to simplify the mature marginal $P_{ss}(x_M)$, the nascent marginal $P_{ss}(x_N)$ is negative binomial:

$$\begin{aligned}
\phi_{ss}(u_N) &= -\frac{a}{\beta} \log\left[1 - b u_N\right] \\
P_{ss}(x_N) &= \binom{x_N + a/\beta - 1}{x_N} \left(\frac{\kappa}{\theta + \kappa}\right)^\alpha \left(\frac{\theta}{\theta + \kappa}\right)^{x_N}.
\end{aligned}$$

Note that this is a different negative binomial distribution than the one that arises in the slow mean-reversion limit. Interestingly, though, this distribution is identical to that one except that β and κ are swapped.

The behavior of the CIR model in this limit is considerably more complicated, and the corresponding count distribution does not seem to belong to any well-characterized parametric family. Still, it is clear that the behavior of the CIR model diverges from that of the Γ-OU model in this regime (see Section S5.2.4 for the details). Our result is that

$$\phi_{ss}(u_N, u_M) = \frac{a}{2} \int_0^\infty 1 - \sqrt{1 - 4b\left[\left(u_N - \frac{\beta}{\beta-\gamma} u_M\right) e^{-\beta s} + \frac{\beta}{\beta-\gamma} u_M e^{-\gamma s}\right]} ds.$$

Once again, while there does not appear to be a simple expression for $\phi_{ss}(u_M)$, we can write

$$\begin{aligned}
\phi_{ss}(u_N) &= \frac{a}{2} \int_0^\infty 1 - \sqrt{1 - 4 b u_N e^{-\beta s}} ds \\
&= \frac{a}{\beta}\left(1 - \sqrt{1 - 4 b u_N}\right) + \frac{a}{\beta} \log\left(\frac{1 + \sqrt{1 - 4 b u_N}}{2}\right)
\end{aligned}$$

for the factorial-cumulant generating function of the nascent marginal. Generally, this expression appears to represent a heavy-tailed and overdispersed nascent count distribution, with the burst size b controlling the dispersion. In the small-burst limit ($b \to 0$), we have that $\phi_{ss}(u_N) \to (ab/\beta) u_N$, i.e. this complicated-looking expression approaches the Poisson result. The first term of $\phi_{ss}(u_N)$ appears to be related to the moment-generating function of an inverse Gaussian distribution.

While the Γ-OU model reduces to describing bursty RNA production in this limit, the CIR model appears to yield a novel 'quasi-bursty' regime not previously studied in biology. RNA

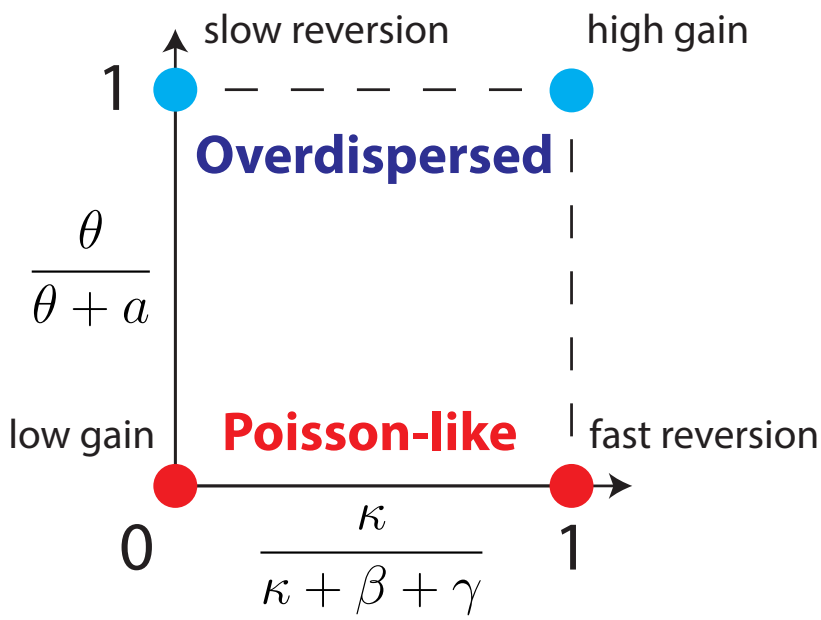

Figure 7: A schematic summary of the limiting regimes of the Γ-OU and CIR models. When $a \gg \theta$, we obtain Poisson-like behavior. When $a \ll \theta$, we obtain overdispersed distributions. Red: Poisson-like regimes; blue: overdispersed regimes.

production here can be viewed as being driven by an *inverse Gaussian* (IG) subordinator with an infinite number of jumps in every time interval (see Section S5.3).

Interestingly, a continuous stochastic process that arises in this limit, which could be labeled the OU-IG process in accordance with standard stochastic processes terminology, has been previously studied in finance [71, 72]. The analytic result for $\phi_{ss}(u_N)$ appears not to have been previously reported in this context, so we have unintentionally filled a lacuna in the financial literature on non-Gaussian OU processes.

5 Discriminating between noise models

Having ascertained the relationship between the static and dynamic models, we are naturally led to consider the relationship between the two classes of dynamic models. The overarching goal of developing stochastic models of biophysics is to identify physical laws that govern living cells. Therefore, we focus on the problem of distinguishing the Γ-OU and CIR models based solely on transcript distributions.

We treat the simplest version of this problem, which we find to be the most illustrative: can we distinguish between the two models, if we know everything *but* the model identity? We must be able to: the models' generating functions have substantially different functional forms. Further, we already know the probability mass functions (PMFs) are distinct based on the first column of Figures 5 and 6, which show different distributions in the high gain limit.

To sum up, even under identical $[\kappa, a, \theta, \beta, \gamma]$, the two models produce distinct behaviors. As shown in Section 4.4, the lower stationary moments and the autocorrelations do not reflect these differences, as they take identical functional forms. Therefore, the most promising path appears to be the comparison of entire joint PMFs.

In light of this, we performed exploratory model analysis to identify *where* in parameter space the two models are easiest to distinguish. We selected the parameter set labeled Intermediate 1 in Table 3 and modulated its parameters across three orders of magnitude. At each parameter set, we computed joint transcript PMFs for both models. From these, we calculated Kolmogorov-Smirnov (KS) distances and Kullback-Leibler (KL) divergences, and visualized the nascent marginals at the highest KL divergence.

These measures serve distinct, yet complementary purposes. As the maximum difference between distribution functions, the KS distance is easily interpretable, symmetric, and robust to tail effects; however, it does not extend to multivariate distributions and does not directly inform the model hypothesis testing problem. On the other hand, the KL divergence provides a criterion for model selection (e.g., via likelihood ratio testing) and does not depend on the system dimensionality; however, it is asymmetric, and potentially underestimates tail-region deviations.

Remarkably, the distributions are very similar, with extremely small KS and KL measures and virtually identical marginals (Figure 8). The sole exception is the very high-noise/high-gain regime (low η or high θ), which demonstrates the highest divergence and produces distributions that can be visually distinguished.

Motivated by this result, we used a gradient descent search to identify points in the parameter space where the KL divergence between Γ-OU and CIR models is maximized. One such distribution, corresponding to $[\kappa, a, \theta, \beta, \gamma] = [0.6044, 0.2428, 5.568, 2.442, 0.212]$, is shown in Figure 9. Validation using simulated data (10^4 cells) confirms that the discordance between the models is a real feature, rather than a numerical artifact (for clarity, we omit the simulated joint histograms). The noise magnitude θ is rather high, again suggesting that the models are best distinguished in high-gain regimes. However, by far the most interesting qualitative feature is the striking difference between the nascent and mature marginals: the nascent distributions are nearly identical, whereas the mature distributions are easily distinguishable. This behavior matches the higher KS distances between mature marginals under SDE parameter modulation in Figure 8, and provides a counterpoint to the heuristic that nascent transcripts are more reflective of transcriptional signal than mature transcripts [73].

In sum, the lower moments cannot be used to distinguish between these mechanistic models. However, full distributions *can*, and the identifiability improves with more data modalities. This

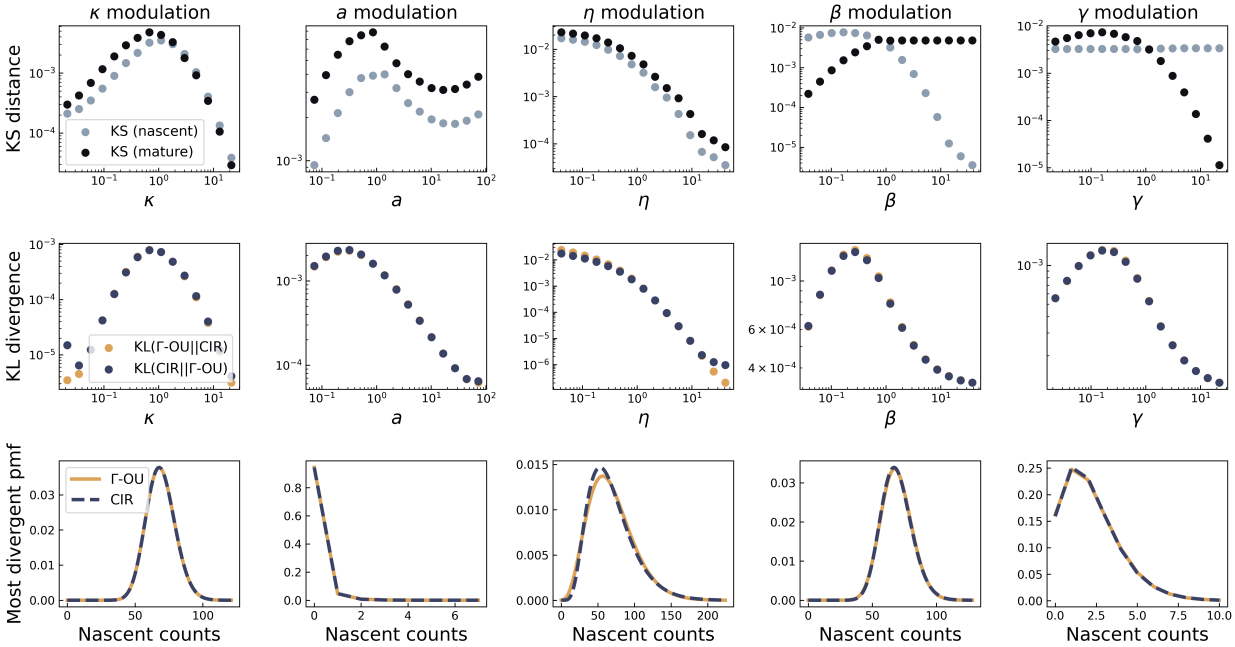

Figure 8: Exploratory analysis of differences between Γ-OU and CIR noise across the model parameter space, modulating one parameter at a time. Top row: Kolmogorov-Smirnov distance between marginal count distributions. Middle row: Kullback-Leibler divergence between joint distributions. Bottom row: Comparison of Γ-OU and CIR-driven nascent RNA distributions at the parameter set achieving the highest KL divergence.

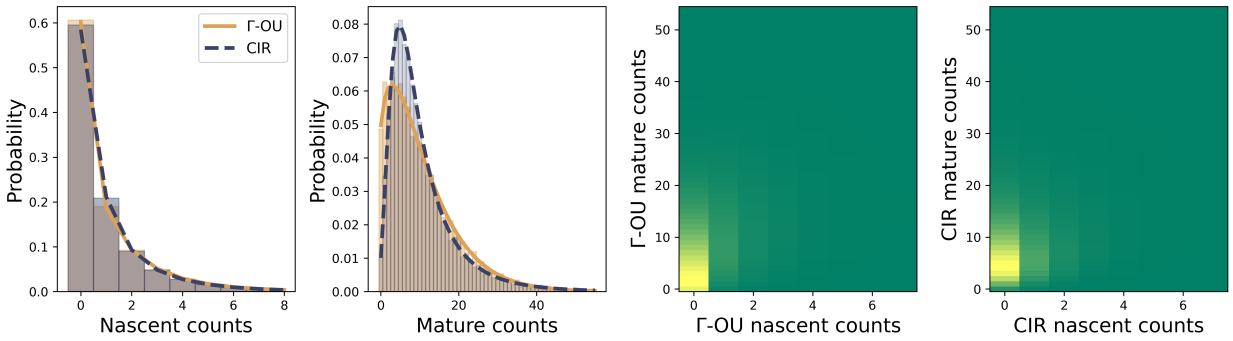

Figure 9: Comparison of observables resulting from Γ-OU and CIR models at a locally maximally divergent parameter set.

finding strongly motivates the use of *joint* models that incorporate both mechanistic hypotheses about transcription and a description of the downstream processes.

6 Discussion

We introduced a class of interpretable and tractable models of transcription, and studied the properties of two biologically plausible members of that class in great detail. 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 biological observables of interest. Therefore, our SDE–CME framework can guide experiments to parameterize 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 extrinsic regime, which presupposes that the evolution of parameters is substantially slower than CME dynamics, is attractive, but potentially implausible. The extrinsic regime we use for illustration has a noise time scale an order of magnitude longer than degradation, yet still produces the tail region deviations shown in Figures 5 and 6. The lifetime of a human mRNA is on the order of tens of hours [74]. Therefore, using an extrinsic noise model is formally equivalent to postulating a driving process with an autocorrelation time of *weeks*, which is implausible. The theoretical and computational tools discussed in the current study allow the analysis of potential noise models and the robustness of predictions under a range of time scale scenarios converging to the limiting case of interest. 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 [75].

The collection and representation of multimodal data are particularly fruitful directions for experimental design. As demonstrated previously [76], joint mRNA distributions can imply very different mechanistic foundations, even if the marginals are identical. Aggregating distinct molecular species as a single observable (i.e., modeling a single observable $X = X_N + X_M$) violates the Markovian assumption and neglects biologically important [77–79] regulatory processes of splicing

and export buffering. Strictly considering well-characterized and annotated mature isoforms, as is custom in single-cell RNA sequencing [80], gives partial information about the transcription kinetics, but also includes contributions from downstream processes, complicating the analysis and giving rise to non-identifiability issues. Further, as demonstrated in Figure 9, nascent mRNA 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, as described elsewhere in the context of bursty models [48], the analytical solutions easily accommodate such data. Therefore, the deliberate collection of multimodal data is a natural idea for the rational and model-guided planning of high-throughput sequencing experiments.

Further, 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 [81].

In studying the Γ-OU and CIR models, we found and validated several distinct asymptotic regimes. Both models recapitulate extrinsic noise and constitutive production in the slow-driving limit. However, in the limit of bursty production, they produce qualitatively different behaviors: the Γ-OU model yields geometric bursts of transcription, whereas the CIR model yields inverse Gaussian driving 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 [72] 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 characterized a fast-mean-reversion, mean-field regime with rapid fluctuations, which yields constitutive-like uncorrelated Poisson copy number distributions.

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 [67]. More generally, stochastic path integral and physics-inspired methods have recently proven useful for obtaining analytical solutions to relatively complicated stochastic models [41,42,82,83]. As discussed in Section S3.1, we take this opportunity to explore the diversity of solution methods and emphasize useful unifying themes.

The availability of solvers for such general classes of driving functions suggest 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 quadrature 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.

7 Code Availability

Python code that can be used to reproduce the figures is available at `https://github.com/pachterlab/GVFP_2021`.

8 Acknowledgments

The DNA, pre-mRNA, and mature mRNA images used in Figure 3 are derivatives of the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. Color palettes used in all figures are partially derived from dutchmasters by EdwinTh, used under the MIT license. 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.

References

[1] Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610, December 2015.

[2] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, February 2019.

[3] Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, page 201912459, September 2019.

[4] Seyed Yahya Anvar, Guy Allard, Elizabeth Tseng, Gloria M. Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H. Yin, Hans E. Johansson, Yavuz Ariyurek, Johan T. den Dunnen, Stephen W. Turner, and Peter A. C. 't Hoen. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biology*, 19(1):46, December 2018.

[5] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, July 2001.

[6] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, 2002.

[7] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, October 2002.

[8] Arjun Raj and Alexander van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226, October 2008.

[9] A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–12172, July 2011.

[10] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 09 2007.

[11] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[12] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, Dec 2014.

[13] Raghd Rostom, Valentine Svensson, Sarah A. Teichmann, and Gozde Kar. Computational approaches for interpreting scRNA-seq data. *FEBS Letters*, 591(15):2213–2225, 2017.

[14] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019.

[15] L F Liu and J C Wang. Supercoiling of the dna template during transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027, 1987.

[16] Jean Peccoud and Bernard Ycard. Markovian Modeling of Gene Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[17] Srividya Iyer-Biswas, F. Hayot, and C. Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79(3):031911, March 2009.

[18] Lifang Huang, Zhanjiang Yuan, Peijiang Liu, and Tianshou Zhou. Effects of promoter leakage on dynamics of gene expression. *BMC Systems Biology*, 9, March 2015.

[19] Stuart A. Sevier, David A. Kessler, and Herbert Levine. Mechanical bounds to transcriptional noise. *Proceedings of the National Academy of Sciences*, 113(49):13983–13988, December 2016.

[20] Stuart A. Sevier and Herbert Levine. Mechanical properties of transcription. *Phys. Rev. Lett.*, 118:268101, Jun 2017.

[21] Stuart A Sevier and Herbert Levine. Properties of gene expression and chromatin structure with mechanically regulated elongation. *Nucleic Acids Research*, 46(12):5924–5934, July 2018.

[22] Z. Cao, T. Filatova, D. A. Oyarzún, and R. Grima. A stochastic model of gene expression with polymerase recruitment and pause release. *bioRxiv*, 2020.

[23] Zhixing Cao and Ramon Grima. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*, 117(9):4682–4692, March 2020.

[24] D Rognan. Chemogenomic approaches to rational drug design. *British Journal of Pharmacology*, 152(1):38–52, 2007.

[25] Soma Mandal, Mee'nal Moudgil, and Sanat K. Mandal. Rational drug design. *European Journal of Pharmacology*, 625(1):90–100, 2009. New Vistas in Anti-Cancer Therapy.

[26] Gregory R. Bowman, Eric R. Bolin, Kathryn M. Hart, Brendan C. Maguire, and Susan Marqusee. Discovery of multiple hidden allosteric sites by combining markov state models and experiments. *Proceedings of the National Academy of Sciences*, 2015.

[27] Billy J. Williams-Noonan, Elizabeth Yuriev, and David K. Chalmers. Free energy methods in drug design: Prospects of "alchemical perturbation" in medicinal chemistry. *Journal of Medicinal Chemistry*, 61(3):638–649, 2018. PMID: 28745501.

[28] Dan Liu, Yunyun Mao, Xue Gu, Yang Zhou, and Dong Long. Unveiling the "invisible" druggable conformations of gdp-bound inactive ras. *Proceedings of the National Academy of Sciences*, 118(11), 2021.

[29] Zachary Fox, Gregor Neuert, and Brian Munsky. Finite state projection based bounds to compare chemical master equation models using single-cell data. *The Journal of Chemical Physics*, 145(7):074101, August 2016.

[30] Zachary R Fox and Brian Munsky. The finite state projection based Fisher information matrix approach to estimate information and optimize single-cell experiments. *PLOS Computational Biology*, 15(1):e1006365, January 2019.

[31] Zachary R. Fox, Gregor Neuert, and Brian Munsky. Optimal design of single-cell experiments within temporally fluctuating environments. *Complexity*, 2020:8536365, Jun 2020.

[32] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336(6078):183–187, 2012.

[33] Donald A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[34] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.

[35] Daniel T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[36] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007.

[37] Daniel T. Gillespie, Andreas Hellander, and Linda R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013.

[38] Zachary Fox and Brian Munsky. Stochasticity or Noise in Biochemical Reactions. Preprint, arXiv: 1708.09264, Aug 2017.

[39] Brian Munsky, William S. Hlavacek, and Lev S. Tsimring, editors. *Quantitative Biology: Theory, Computational Methods, and Models*. The MIT Press, 2018.

[40] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, December 2006.

[41] John J. Vastola. Solving the chemical master equation for monomolecular reaction systems analytically: a Doi-Peliti path integral view. *arXiv:1911.00978 [q-bio]*, November 2019. arXiv: 1911.00978.

[42] John J. Vastola, Gennady Gorin, Lior Pachter, and William R. Holmes. Analytic solution of chemical master equations involving gene switching. I: Representation theory and diagrammatic approach to exact solution. Preprint, arXiv: 2103.10992, March 2021. arXiv: 2103.10992.

[43] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.

[44] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977.

[45] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271, 2008.

[46] Abhyudai Singh and Pavol Bokes. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. *Biophysical Journal*, 103(5):1087–1096, September 2012.

[47] Gennady Gorin and Lior Pachter. Special function methods for bursty models of transcription. *Physical Review E*, 102(2):022409, August 2020.

[48] Gennady Gorin and Lior Pachter. Analytical solutions of the chemical master equation with bursty production and isomerization reactions. Preprint, bioRxiv: 2021.03.24.436847, June 2021.

[49] Rama Cont and Peter Tankov. *Financial Modeling with Jump Processes*. Financial Mathematics. Chapman & Hall, 2004.

[50] Ole E. Barndorff-Nielsen and Neil Shephard. Modelling by lévy processess for financial econometrics. In Ole E. Barndorff-Nielsen, Sidney I. Resnick, and Thomas Mikosch, editors, *Lévy Processes: Theory and Applications*, pages 283–318. Birkhäuser Boston, Boston, MA, 2001.

[51] Ole E. Barndorff-Nielsen and Neil Shephard. Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):167–241, 2001.

[52] Yan Qu, Angelos Dassios, and Hongbiao Zhao. Exact simulation of gamma-driven Ornstein–Uhlenbeck processes with finite and infinite activity jumps. *Journal of the Operational Research Society*, pages 1–14, December 2019.

[53] Guillaume Bernis, Riccardo Brignone, Simone Scotti, and Carlo Sgarra. A gamma ornstein–uhlenbeck model driven by a hawkes process. *Mathematics and Financial Economics*, Mar 2021.

[54] John J. Vastola and William R. Holmes. Chemical Langevin equation: A path-integral view of Gillespie's derivation. *Phys. Rev. E*, 101:032417, Mar 2020.

[55] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A theory of the term structure of interest rates. In Sudipto Bhattacharya and George Michael Constantinides, editors, *Theory of Valuation*, pages 129–164. World Scientific Publishing Company, 2005.

[56] Stephen J. Brown and Philip H. Dybvig. The empirical implications of the cox, ingersoll, ross theory of the term structure of interest rates. *The Journal of Finance*, 41(3):617–630, 1986.

[57] John Hull and Alan White. Pricing Interest-Rate-Derivative Securities. *The Review of Financial Studies*, 3(4):573–592, 04 2015.

[58] Bo Hu, David A. Kessler, Wouter-Jan Rappel, and Herbert Levine. How input fluctuations reshape the dynamics of a biological switching system. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 86(6 Pt 1):061910, December 2012.

[59] Christoph Zechner and Heinz Koeppl. Uncoupled Analysis of Stochastic Reaction Networks in Fluctuating Environments. *PLoS Computational Biology*, 10(12):e1003942, December 2014.

[60] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 37(5):547–554, May 2019.

[61] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020.

[62] Nir Friedman, Long Cai, and X. Sunney Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):168302, October 2006.

[63] Gennady Gorin and Lior Pachter. Direct simulation of a stochastically driven multi-step birth-death process. Preprint, bioRxiv: 2021.01.20.427480, January 2021.

[64] Anqi Shao. *A fast and exact simulation for CIR process*. PhD thesis, University of Florida Gainesville, FL, 2012.

[65] A. Prados, J. J. Brey, and B. Sánchez-Rey. A dynamical monte carlo algorithm for master equations with time-dependent transition rates. *Journal of Statistical Physics*, 89(3-4):709–734, November 1997.

[66] M. M. Tai. A Mathematical Model for the Determination of Total Area Under Glucose Tolerance and Other Metabolic Curves. *Diabetes Care*, 17(2):152–154, February 1994.

[67] John J. Vastola. The information transmission problem in RNA splicing networks, and a path integral framework for exactly solving coupled discrete and continuous stochastic dynamics. *in preparation*, 2021.

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

[69] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, December 2005.

[70] Lisa Amrhein, Kumar Harsha, and Christiane Fuchs. A mechanistic model for the negative binomial distribution of single-cell mRNA counts. Preprint, bioRxiv: 657619, June 2019.

[71] O. E. Barndorff-Nielsen and J. Schmiegel. A Stochastic Differential Equation Framework for the Timewise Dynamics of Turbulent Velocities. *Theory of Probability & Its Applications*, 52(3):372–388, January 2008.

[72] Ole E. Barndorff-Nielsen and Neil Shephard. Integrated OU Processes and Non-Gaussian OU-based Stochastic Volatility Models. *Scandinavian Journal of Statistics*, 30(2):277–295, June 2003.

[73] Heng Xu, Samuel O. Skinner, Anna Marie Sokac, and Ido Golding. Stochastic Kinetics of Nascent RNA. *Physical Review Letters*, 117(12):128101, 2016.

[74] Ron Milo and Rob Phillips. *Cell Biology by the Numbers*. Garland Science, July 2015.

[75] Lucy Ham, Rowan D. Brackston, and Michael P.H. Stumpf. Extrinsic Noise and Heavy-Tailed Laws in Gene Expression. *Physical Review Letters*, 124(10):108101, March 2020.

[76] Gennady Gorin and Lior Pachter. Intrinsic and extrinsic noise are distinguishable in a synthesis – export – degradation model of mRNA production. Preprint, bioRxiv: 10.1101/2020.09.25.312868, September 2020.

[77] Qianliang Wang and Tianshou Zhou. Alternative-splicing-mediated gene expression. *Physical Review E*, 89(1):012713, January 2014.

[78] Tara Alpert, Lydia Herzel, and Karla M. Neugebauer. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdisciplinary Reviews: RNA*, 8(2):e1401, March 2017.

[79] Manfred Schmid and Torben Heick Jensen. Controlling nuclear RNA levels. *Nature Reviews Genetics*, 19(8):518–529, August 2018.

[80] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, April 2017.

[81] Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences*, 115(29):7533–7538, 2018.

[82] John J. Vastola. The chemical birth-death process with additive noise. Preprint, arXiv: 1910.09117, October 2019.

[83] John J. Vastola. The chemical birth-death process with Gillespie noise. Preprint, arXiv: 1910.10807, October 2019.