What can we learn when fitting a complex gene expression model to a simple telegraph model?

Feng Jiao[1], Jing Li[1], Ting Liu[1], Yifeng Zhu[1], Wenhao Che[1], Leonidas Bleris[2], Chen Jia[3,*]

[1] Guangzhou Center for Applied Mathematics, Guangzhou University, Guangzhou, Guangzhou, 510006, China

[2] Center for Systems Biology, The University of Texas at Dallas, Richardson, Texas, U.S.A

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

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

Abstract

In experiments, the distributions of mRNA or protein numbers in single cells are often fitted to the random telegraph model which includes synthesis and degradation of mRNA or protein, and switching of the gene between active and inactive states. While commonly used, this model does not describe how fluctuations are influenced by crucial biological mechanisms such as feedback regulation, non-exponential gene inactivation durations, and multiple gene activation pathways. Here we investigate the dynamical properties of four complex gene expression models by fitting their steady-state mRNA or protein number distributions to the simple telegraph model. We show that despite the underlying complex biological mechanisms, the telegraph model with three effective rate parameters can accurately capture the steady-state gene product distributions, as well as the conditional distributions in the active gene state, of the complex models. Some effective rate parameters are reliable and can reflect realistic dynamic behaviors of the complex models, while others may deviate significantly from their real values in the complex models. The effective parameters can be also applied to characterize the capability for a complex model to exhibit multimodality. Using additional information such as single-cell data at multiple time points, we provide an effective method of distinguishing the complex models from the telegraph model. Furthermore, using measurements under varying experimental conditions, we show that fitting the mRNA or protein number distributions to the telegraph model may even reveal the underlying gene regulation mechanisms of the complex models. The effectiveness of these methods is confirmed by analysis of single-cell data for *E. coli* and mammalian cells.

Introduction

Recent experiments have revealed a large cell-to-cell variation in the numbers of mRNA and protein molecules in isogenic populations due to stochasticity in gene expression and the low copy numbers of many components, including DNA and important regulatory molecules [1–3]. Live-cell imaging approaches allow a direct visualization of stochastic bursts of gene expression in living cells [4]. However these experiments are challenging and thus more commonly one measures the mRNA or protein expression in individual cells using flow cytometry, single-molecule fluorescence *in situ* hybridization (smFISH) [4], and single-cell RNA sequencing (scRNA-seq) [5]. Together with mathematical models, the large amounts of single-cell gene expression data have been used to understand stochastic gene regulations in various biological problems, ranging from genetic engineering [6, 7] to cell fate decision [8, 9] and therapeutic targets of disease [10, 11].

The experimental distributions of mRNA and protein numbers are often fitted to the predictions of mathematical models [12–18]. The most common and well-studied model of this type is the two-state telegraph model [19, 20], which is composed of four effective reactions (Fig. 1(a))

$$G \xrightarrow{\bar{\lambda}} G^*, \quad G^* \xrightarrow{\bar{\gamma}} G, \quad G^* \xrightarrow{\bar{\rho}} G^* + P, \quad P \xrightarrow{d} \varnothing, \tag{1}$$

where the first two reactions describe the switching of the gene between an active state G^* and an inactive state G, the third reaction describes synthesis of the gene product P, which can be either mRNA [12] or protein [20] when the gene is active, and the fourth reaction describes degradation of the gene product. The chemical master equations describing the telegraph model can be exactly solved in steady-state, as well as in time [20, 21]. Extensions of this model to include more than two gene states have also been considered [22–24]. With tremendous efforts of quantitative and qualitative analysis, the telegraph model has been successfully applied to understand stochasticity in gene expression by (i) clarifying the biological reasons of different distribution patterns of mRNA or protein numbers [25], (ii) performing a fast and reliable inference of all rate parameters [17], and (iii) unravelling gene regulation mechanisms in response to environmental changes in various cell types [26].

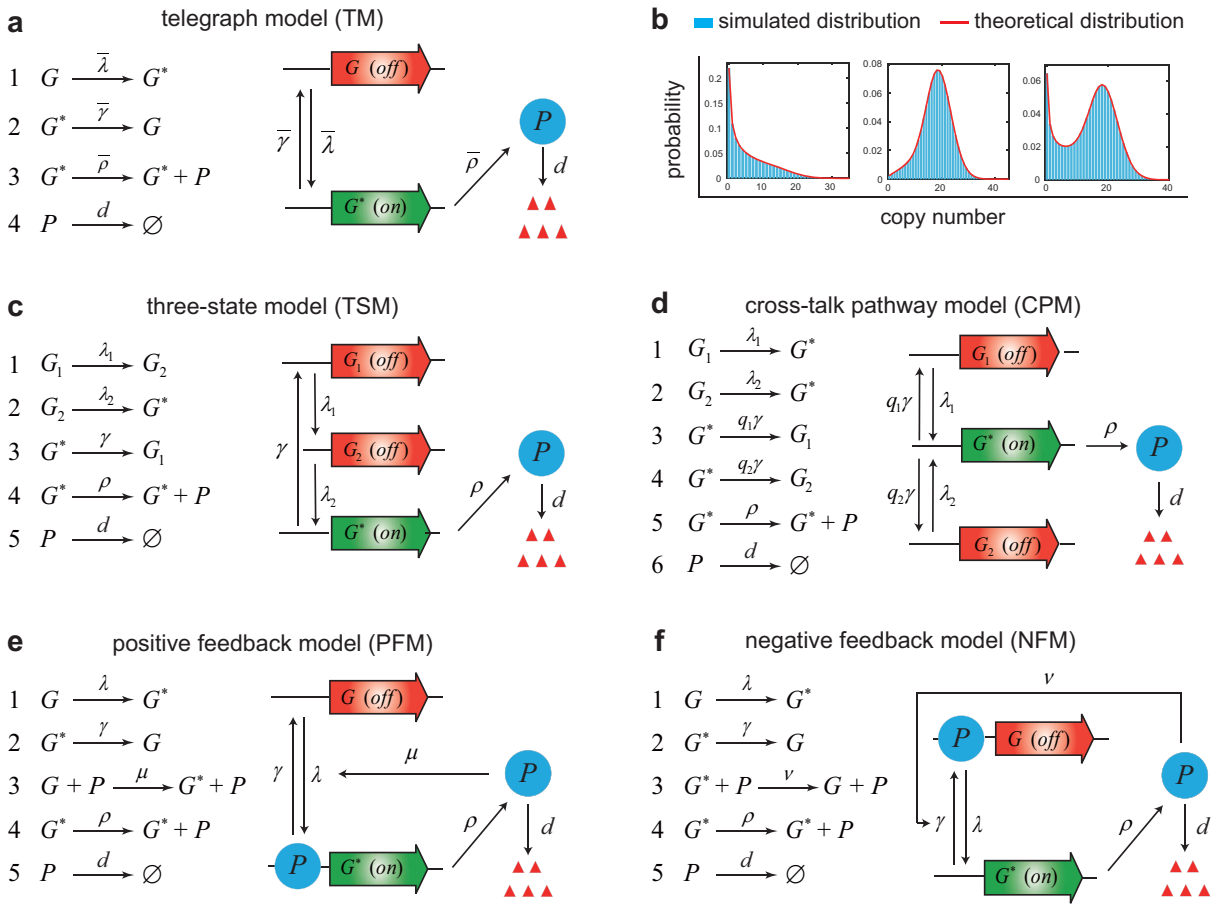


Figure 1: **The simple telegraph model and four relatively complex gene expression models.** **(a)** In the random telegraph model (TM), the gene switches between an inactive (off) and an active (on) state with rates $\bar{\lambda}$ and $\bar{\gamma}$. The gene product (mRNA or protein, denoted by P) is synthesized with rate $\bar{\rho}$ when the gene is active, and is degraded with rate d. **(b)** The telegraph model can generate three different patterns of the mRNA or protein number distributions: a unimodal distribution with a zero peak (left panel), a unimodal distribution with a nonzero peak (middle panel), and a bimodal distribution with both a zero and a nonzero peak (right panel). Computation of the theoretical distribution for the telegraph model can be used to perform parameter inference based on real gene expression data. **(c)** In the three-state model (TSM), the gene exhibits a "refractory" behavior: after leaving the active state with rate γ, the gene has to progress through two sequential inactive states with rates λ_1 and λ_2 before becoming active again. **(d)** In the cross-talk pathway model (CPM), the gene can be activated via two signaling pathways with rates λ_1 and λ_2. The competition between the two pathways is modelled by equipping them with two selection probabilities q_1 and $q_2 = 1 - q_1$. **(e)** In the positive feedback model (PFM), the protein produced from the gene activates its own expression with feedback strength μ. **(f)** In the negative feedback model (NFM), the protein produced from the gene inhibits its own expression with feedback strength ν.

In experiments, there are three commonly observed patterns for the distributions of mRNA and protein numbers: a unimodal distribution with a zero peak, a unimodal distribution with a nonzero peak, and a bimodal distribution with both a zero and a nonzero peak (Fig. 1(b)) [25]. Actually, the telegraph model can only produce

the above three distribution patterns. Among the three patterns, the bimodal distribution attracts massive attention since it separates isogenic cells into two distinct phenotypes [27, 28]. Bimodality of mRNA and protein number distributions has been used to describe the bet-hedging strategy in microorganisms [29, 30], and to quantify cell fate decisions such as the differentiation of embryonic stem cells [31] and the activation of HIV latency [11]. For the telegraph model, it has been shown that the occurrence of bimodality requires relatively slow rates of gene state switching — a bimodal distribution can only occur when both the gene activation and inactivation rates are smaller than the degradation rate [32, 33].

In previous studies, the mRNA or protein number distributions for various genes (and even for the whole genome) are often fitted to the telegraph model [12–18], by which one can obtain estimates of the rates of the underlying gene expression processes. One of the most prevalent methods of parameter inference is the maximum likelihood method which minimizes the log-likelihood function [16, 17]

$$\log L(\theta) = \sum_n N(n) \log(P_n(\theta)), \tag{2}$$

where θ is the parameter set, $N(n)$ denotes the number of cells with n copies of mRNA or protein, and $P_n(\theta)$ denotes the mRNA or protein number distribution with parameter set θ. The degradation rate d in the telegraph model is often determined by measuring the half-life of mRNA or protein [34]. To achieve fast and accurate estimation of the gene activation rate $\bar{\lambda}$, gene inactivation rate $\bar{\gamma}$, and synthetic rate $\bar{\rho}$, one of the important steps is to select their initial values $\bar{\lambda}_0$, $\bar{\gamma}_0$, and $\bar{\rho}_0$ for optimization. A common choice is to set $\bar{\rho}_0$ to be the maximum number of mRNA or protein molecules among single cells [16]. Once $\bar{\rho}_0$ is determined, the initial values of the other two parameters, $\bar{\lambda}_0$ and $\bar{\gamma}_0$, can be determined by matching the mean and variance of gene product fluctuations [16, 18].

Accurate inference of model parameters can further reveal various gene regulation mechanisms in response to varying environmental conditions. By fitting the gene expression data to the telegraph model, one can understand how all rate parameters change in response to varying conditions [12, 15]. Previous studies have revealed rich gene regulation mechanisms under different induction conditions or promoter architectures. For instance, the up-regulation of gene expression levels can be achieved by increasing the gene activation rate $\bar{\lambda}$ for the zinc-induced yeast ZRT1 gene [6], decreasing the gene inactivation rate $\bar{\gamma}$ for over 20 *Escherichia coli* (*E. coli*) promoters under different growth conditions [7, 13], increasing the synthesis rate $\bar{\rho}$ for serum-induced mammalian *ctgf* gene [35], or a combined effect of both the burst frequency $\bar{\lambda}$ and burst size $\bar{\rho}/\bar{\gamma}$ in prokaryotic and eukaryotic cells [26, 36].

However, the conventional telegraph model is limited in their predictive power because it lacks a description of some biological mechanisms that are known to have a profound impact on the distributions of mRNA and protein numbers in single cells, such as feedback regulation, non-exponential gene inactivation durations, and multiple gene activation pathways (Fig. 1(c)-(f)). The telegraph model can only be used to study genes that are unregulated, and it fails for regulated genes. One of the most common gene network motifs is an autoregulatory feedback loop whereby protein expressed from a gene activates or represses its own expression (Fig. 1(e),(f)) [37–40]. It has been estimated that 40% of all transcription factors self-regulate in *E. coli* [41] with most of them participating in negative autoregulation [42]. An effective method of inferring the sign of autoregulation has been proposed based on gene expression measurements under different feedback strengths [43].

Except feedback regulation, another important mechanism that regulates gene expression is non-exponential gene inactivation periods. In the two-state telegraph model, the time spent in the active or inactive gene state has an exponential distribution. The exponential active period is generally a reasonable assumption [44]. However, recent studies in mammalian and bacterial cells have shown that the inactive periods for many genes may have a non-exponential peaked distribution [45–48]. This suggests that the gene dynamics in the inactive period may contain two rate-limiting steps and exhibit a "refractory" behavior: after leaving the active state, the promoter has to progress through two inactive states before becoming active again (Fig. 1(c)). This refractory behavior is probably due to the fact that the activation of the promoter is a complex multi-step biochemical process due to chromatin remodeling and the binding and release of transcription factors [22, 49]. A Bayesian method has been also proposed to estimate all rate parameters of this refractory model based on time-course gene expression data [50].

Another possible mechanism that regulates gene expression is the existence of multiple signalling pathways during gene activation [44, 51]. In the telegraph model, there is only one gene activation pathway. Recent modelling studies [52, 53] have shown that the competition between two gene activation pathways (Fig. 1(d)) can well capture the rapid overshooting behavior of transcription levels observed in mouse fibroblasts under the induction of tumor necrosis factor [54]. In contrast, such rapid overshooting dynamics cannot be explained by the single gene activation pathway with one or more rate-limiting steps as it either generates monotonic transcription dynamics or triggers a long lag to reach the peak of the transcription level. Moreover, the competition between two gene activation pathways can also capture the dynamic mRNA expression data observed in yeast *HSP12* gene under NaCl osmotic stress which exhibit unimodal distributions with a zero peak for short and long times while exhibit bimodal distributions for intermediate times [55, 56]. Such dynamic transitions among different distribution patterns are rarely observed in the telegraph model and other gene expression models [57, 58]. The multiple gene activation pathways are probably due to the exclusive binding of multiple transcription factors to the promoter.

Integrating the above mechanisms (feedback regulation, non-exponential gene inactivation periods, and multiple gene activation pathways) into the telegraph model can generate more complex models of stochastic gene expression (Fig. 1(c)-(f)). An essential problem is, compared to the telegraph model, there is still a lack of efficient methods of theoretical analysis and parameter inference for these complex models due to the increased complexity of model structures and increased number of model parameters. Furthermore, it is also difficult to distinguish these complex models from the simple telegraph model since they often exhibit similar distribution shapes. This raises the questions of (i) whether some rate parameters for a complex model can be accurately inferred, (ii) whether we can distinguish a complex model from the telegraph model, and (iii) whether we can infer the gene regulation mechanism of a complex model by using gene expression data under different induction conditions.

In this paper, we will provide insights to these questions. Our strategy is not to investigate the complex models themselves; rather, we examine these models by fitting their steady-state distributions of mRNA or protein numbers to the telegraph model and then obtain estimates of the "effective" rate parameters. In fact, the idea of fitting a complex model to the telegraph model has been previously carried out in [17], where the authors realized that the three-state model shown in Fig. 1(c) may be more accurate in mammalian cells but they still fitted the mRNA number distributions for thousands of genes to the telegraph model since, as explained in [17], "the resulting steady-state distribution for the extended (three-state) model is very close to the two-state model and to distinguish between these similar models, additional information such as multiple time measurements within the same cell is needed." In general, the estimated values of the rate parameters (the mRNA or protein synthesis rate and the gene activation and inactivation rates) in the "artificial" telegraph model may deviate largely from the real values of these parameters in the complex models. However, we find that the effective rate parameters are still sometimes reliable and can also reveal important dynamical properties of the complex models such as the ability of a complex model to produce bimodality. Furthermore, using additional information such as gene expression data at multiple time points or measurements under varying experimental conditions, we provide an effective method of distinguishing the complex models from the telegraph model, and we also show that fitting the mRNA or protein number distributions to the telegraph model may even reveal the underlying gene regulation mechanisms of the complex models. The effectiveness of these methods are confirmed by analysis of published data for *E. coli* and mammalian cells.

Results

Four complex gene expression models revisited

Here we recall four relatively complex models of stochastic gene expression including the three-state model, cross-talk pathway model, positive feedback model, and negative feedback model (see Fig. 1(c)-(f) for illustration and the detailed reaction schemes). All of them have been extensively studied in the literature and are established by integrating a particular genetic regulation mechanism into the telegraph model. Like the telegraph model, all the

four complex models are composed of gene state switching, synthesis of the gene product, and degradation of the gene product. The gene product for the former two models can be either mRNA [12] or protein [20], while for the latter two models, the gene product must be proteins since feedback regulation is realized by binding of proteins to the promoter.

In the two-state telegraph model, the durations in the active and inactive gene states both have an exponential distribution. However, recent studies showed that while the exponential active period seems to be universal [44], the inactive periods in mammalian cells may have a non-exponential distribution with a nonzero peak [45–48]. This suggests that the process of gene activation may contain two rate-limiting steps and in sum, the gene would undergo a three-state switching process (Fig. 1(c)) [45]. The dynamics of the three-state model (also called the refractory model [45]) is controlled by two consecutive gene activation steps with rates λ_1 and λ_2, gene inactivation rate γ, synthesis rate ρ of mRNA or protein, and degradation rate d. For convenience, we set $d = 1$ in what follows. This is not an arbitrary choice but stems from the fact that the time in the chemical master equations can always be non-dimensionalized using d. Specifically, the time given below should be understood to be non-dimensional and equal to the real time multiplied by d, while all other parameters λ_i, γ, and ρ given below should be understood to also be non-dimensional and equal to the real values of these parameters divided by d. Let P_n denote the steady-state probability of having n copies of the gene product in a single cell. Then the corresponding generating function $F(z) = \sum_{n=0}^{\infty} P_n z^n$ of the gene product distribution for the three-state model is given by [22]

$$F(z) = {}_2F_2(\lambda_1, \lambda_2; \alpha_1, \beta_1; \rho(z-1)), \tag{3}$$

where $\alpha_1 + \beta_1 = \lambda_1 + \lambda_2 + \gamma$, $\alpha_1\beta_1 = \lambda_1\lambda_2 + \lambda_1\gamma + \lambda_2\gamma$, and ${}_2F_2$ denotes the generalized hypergeometric function. Once the generating function $F(z)$ is known, the gene product distribution P_n can be recovered by taking the derivative of $F(z)$ at $z = 0$:

$$P_n = \frac{1}{n!}\frac{d^n F(z)}{dz^n}\bigg|_{z=0}.$$

The cross-talk pathway model describes competitive binding of two transcriptional factors to the promoter: one actives the gene via a weak signalling pathway with rate λ_1, and the other actives the gene via a strong signalling pathway with a larger rate $\lambda_2 > \lambda_1$ (Fig. 1(d)) [59, 60]. The competition between the two pathways is modelled by equipping them with two selection probabilities q_1 and q_2 satisfying $q_1 + q_2 = 1$. In other words, the gene is activated via the weak pathway with probability q_1 and is activated via the strong pathway with probability q_2. The mRNA or protein is synthesized with rate ρ and is degraded with rate $d = 1$. This model has been successfully used to explain the rich transcription dynamics observed in mouse fibroblasts and yeast under different induction conditions [53, 56]. The generating function $F(z)$ of the gene product distribution for the cross-talk pathway model is given by [61]

$$F(z) = {}_2F_2(\lambda_1, \lambda_2; \alpha_2, \beta_2; \rho(z-1)), \tag{4}$$

where $\alpha_2 + \beta_2 = \lambda_1 + \lambda_2 + \gamma$ and $\alpha_2\beta_2 = \lambda_1\lambda_2 + (q_2\lambda_1 + q_1\lambda_2)\gamma$.

One of the most common gene network motifs is an autoregulatory feedback loop whereby protein expressed from a gene activates or represses its own expression [42]. It has been estimated that 40% of all transcription factors self-regulate in *E. coli* [41]. The positive feedback model describes an autoregulatory gene circuit whereby protein expressed from a gene activates its own expression (Fig. 1(e)). It has the same reaction scheme as the telegraph model except that the protein activates the gene with rate constant μ, which characterizes the strength of positive feedback. Note that when $\mu = 0$, it reduces to the telegraph model. The generating function $F(z)$ of the gene product distribution for the positive feedback model is given by [62]

$$F(z) = \frac{{}_1F_1(\alpha_3; \beta_3; \rho(z - z_3))}{{}_1F_1(\alpha_3; \beta_3; \rho(1 - z_3))}, \tag{5}$$

where $\alpha_3 = \lambda/(\mu+1)$, $\beta_3 = (\lambda+\gamma)/(\mu+1)$, $z_3 = 1/(\mu+1)$, and ${}_1F_1$ denotes the confluent hypergeometric function.

Among the 40% transcription factors that regulate their own expression in *E. coli*, most of them participate in negative autoregulation [42]. The negative feedback model describes an autoregulatory gene circuit whereby protein expressed from a gene represses its own expression (Fig. 1(f)). It has the same reaction scheme as the telegraph model except that the protein represses the gene with rate constant ν, which characterizes the strength of negative feedback. The generating function $F(z)$ of the gene product distribution for the negative feedback model is given by [37]

$$F(z) = \frac{{}_1F_1(\alpha_4; \beta_4; \rho(z - z_4))}{{}_1F_1(\alpha_4; \beta_4; \rho(1 - z_4))}, \tag{6}$$

where $\alpha_4 = (\lambda + \gamma)/(\nu + 1) + \rho\nu/(\nu + 1)^2$, $\beta_4 = \rho/(\nu + 1)$, and $z_4 = 1/(\nu + 1)$.

Note that there are three parameters in the classical telegraph model (assuming that $d = 1$), four parameters in the three-state, positive feedback, and negative feedback models, and five parameters in the cross-talk pathway model. In recent experiments, it has been shown that the distributions of protein numbers in *E. coli* measured using single-molecule fluorescence microscopy often have a unimodal distribution [3] and the distributions of mRNA numbers measured using single-cell RNA sequencing often have a negative binomial or a zero-inflated negative binomial distribution [63–65]. Given these (relatively simple) mRNA and protein number distributions, it is almost impossible to accurately infer all the four or five parameters of a complex model. A solution of this is to fit the experimental distributions of mRNA or protein numbers to the predictions of a simple telegraph model, by which one can obtain estimates of the three "effective" rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ [12–17]. However, it is not clear whether a simple telegraph model can always capture the distribution of a complex model, and it is also not clear whether these effective rate parameters estimated can reflect the properties of the transcriptional and translational processes behind a complex model.

The telegraph model can accurately capture the distributions of complex models

We first examine whether the gene product distribution of a complex model can be well approximated by that of the telegraph model. To this end, for each of the four complex models, we generate synthetic data of mRNA or protein numbers for $N = 10^4$ cells using the stochastic simulation algorithm (SSA), and then fit the steady-state simulated distributions to the telegraph model using the maximum-likelihood method [12–17] (Fig. 2(a)). When performing simulations, the gene activation rate of the weak signalling pathway for the cross-talk pathway model is fixed to be $\lambda_1 = 0.2$ so that each of the four complex models has four independent parameters. To proceed, we proportionally select five different values for each of the four parameters, which cover large swathes of parameter space and give $5^4 = 625$ different parameter sets for each model (see Methods for details).

Like the telegraph model, each complex model can generate unimodal or bimodal distributions of gene product numbers (Fig. 2(a)). For each of the 625 parameter sets, the synthetic data obtained using the SSA are then fitted to the telegraph model. Interestingly, we find that the simulated distributions for all complex models and all parameter sets can be well approximated by the predictions of the telegraph model with Hellinger distance (HD) between the two distributions always less than 0.08 (Fig. 2(a),(b)). This shows that the distributions of complex models can generally be well captured by the telegraph model with effective rate parameters $\bar{\rho}$, $\bar{\lambda}$, and $\bar{\gamma}$. In what follows, the telegraph model equipped with the effective rate parameters is called the *effective telegraph model* of a complex model.

While the distributions of complex models can be well fitted by the effective telegraph model, it is not clear whether the conditional gene product distributions of complex models in the inactive and active gene states can also be captured by the effective telegraph model. Specifically, let $P_{i,n}$ denote the steady-state probability of observing n copies of the gene product when the gene is in state i, with $i = 0, 1$ corresponding to the inactive and active states, respectively. Then the conditional gene product distribution in gene state i can be calculated as

$$P_{n|i} = \frac{P_{i,n}}{\sum_{n=0}^{\infty} P_{i,n}}.$$

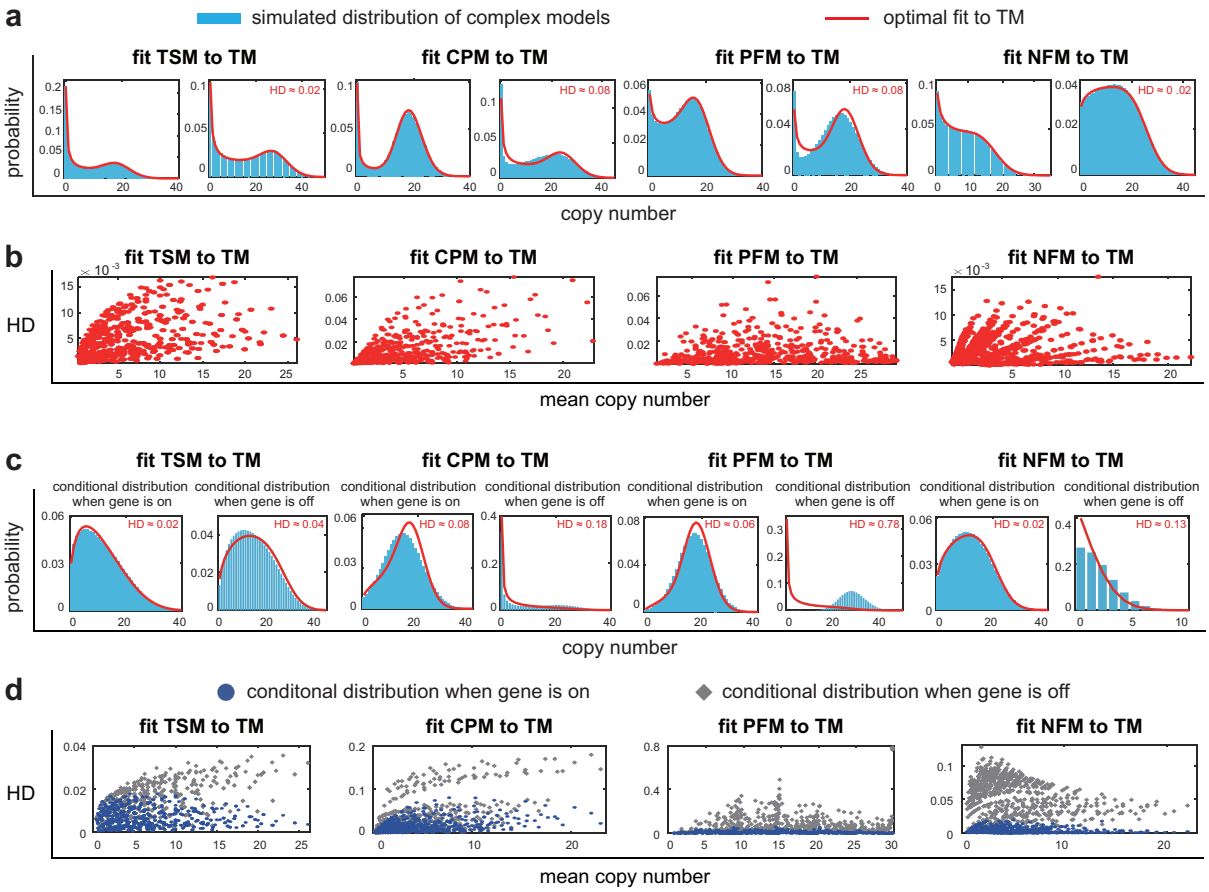


Figure 2: **Fitting the gene product distributions of complex models to the telegraph model.** For each complex model, synthetic data of gene product numbers are generated under 625 parameter sets using the SSA. **(a)** In steady state, all the simulated distributions (blue bars) are well captured by the distributions of the telegraph model (red curve). For each complex model, the left panel shows a typical gene product distribution and the right panel shows the distribution with worse telegraph model approximation, i.e. maximum Hellinger distance (HD). **(b)** For each complex model, the HD between the simulated distribution and its telegraph model approximation is shown as a function of the mean expression level for the 625 parameter sets. The HD is less than 0.08 for all complex models. **(c)** In steady state, the telegraph model not only captures the total gene product distribution of a complex model, but also captures the conditional distribution in the active gene state. In contrast, for all complex models except the three-state model, the conditional distribution in the inactive gene state in general fails to be captured by the telegraph model. For each complex model, the left (right) panel shows the conditional distribution when the gene is on (off) with worse telegraph model approximation, i.e. maximum HD. **(d)** For each complex model, the HD is shown as a function of the mean expression level for the 625 parameter sets. The blue circles (grey diamonds) show the HD between the conditional distribution when the gene is on (off) and its telegraph model approximation. The maximum HD for blue circles is only 0.08 for all complex models, while the maximum HD for grey diamonds can be as large as 0.78.

The expressions of the conditional distributions for the four complex models can also be computed in closed form and can be found in [22, 61, 62]. For each complex model, we generate the simulated conditional distributions in the inactive and active gene states ($P_{n|0}$ and $P_{n|1}$) using the SSA under all 625 parameter sets. Similarly to the total gene product distribution (P_n), the conditional distribution in the active gene state ($P_{n|1}$) for each complex model can always be well approximated by the predictions of the effective telegraph model with HD always less than 0.08 (see Fig. 2(c) and the blue dots in Fig. 2(d)). The situation is different for the inactive gene state. We find that for each complex model except the three-state model, the conditional distribution in the inactive gene state ($P_{n|0}$) fails to be captured by the effective telegraph model, manifested by significantly larger HD values. The worst approximation occurs for the positive feedback model where the HD can be as large as 0.78 (see Fig. 2(c) and the grey dots in Fig. 2(d)).

Linking effective rate parameters to realistic gene expression processes

Recall that in the telegraph model, $\bar{\rho}$ represents the synthesis rate of the gene product, while $\bar{\lambda}$ and $\bar{\gamma}$ represent the frequencies of the gene being activated and inactivated, respectively (Fig. 1(a)). In other words, $1/\bar{\lambda}$ and $1/\bar{\gamma}$ represent the mean inactive and active durations, respectively. However, it is not clear whether the effective rate parameters $\bar{\rho}$, $\bar{\lambda}$, and $\bar{\gamma}$ of the four complex models can reflect the same dynamic properties.

Note that the synthesis rate of the gene product is ρ for each complex model (Fig. 1(c)-(f)). However, the mean holding times in the active and inactive states for the four complex models have completely different analytical expressions. For the three-state model, since gene inactivation consists of only one exponential step and gene activation consists of two exponential steps, the mean active and inactive durations can be easily calculated as

$$\langle T_{\text{on}} \rangle = \frac{1}{\gamma}, \quad \langle T_{\text{off}} \rangle = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}.$$

For the cross-talk pathway model, since there is only one pathway for gene inactivation and two pathways for gene activation, the mean active and inactive durations can be easily calculated as

$$\langle T_{\text{on}} \rangle = \frac{1}{\gamma}, \quad \langle T_{\text{off}} \rangle = \frac{q_1}{\lambda_1} + \frac{q_2}{\lambda_2}.$$

The computations of the mean holding times for positive and negative feedback models are more complicated. Note that for each of the two models, the mean protein number can be represented as [62]

$$\langle n \rangle = \rho \times \frac{\langle T_{\text{on}} \rangle}{\langle T_{\text{on}} \rangle + \langle T_{\text{off}} \rangle}, \tag{7}$$

where the second term on the right-hand side represents the probability of the gene being active. The mean protein number $\langle n \rangle = F'(1)$ can be easily recovered by taking the derivatives of the generating function given in Eqs. (5) and (6) at $z = 1$. Then the mean holding times can be computed analytically using Eq. (7). Specifically, for the positive feedback model, the mean active and inactive durations are given by (see Methods for the detailed proof)

$$\langle T_{\text{on}} \rangle = \frac{1}{\gamma}, \quad \langle T_{\text{off}} \rangle = \frac{\beta_3 \cdot {}_1F_1(\alpha_3; \beta_3; \rho(1-z_3)) - \alpha_3 \cdot {}_1F_1(\alpha_3+1; \beta_3+1; \rho(1-z_3))}{\gamma\alpha_3 \cdot {}_1F_1(\alpha_3+1; \beta_3+1; \rho(1-z_3))},$$

while for the negative feedback model, we have (see Methods for the detailed proof)

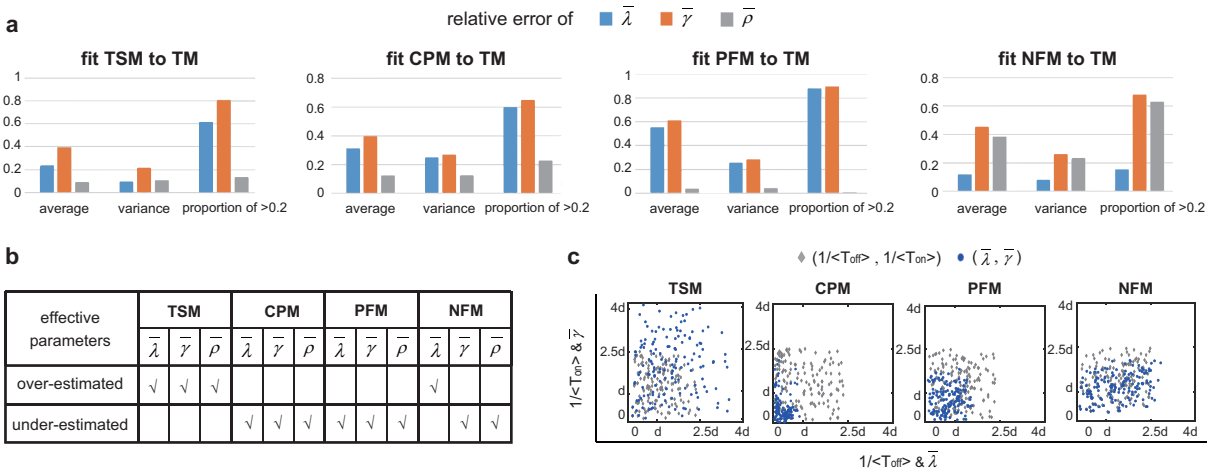
$$\langle T_{\text{on}} \rangle = \frac{\alpha_4 \cdot {}_1F_1(\alpha_4+1; \beta_4+1; \rho(1-z_4))}{\lambda\beta_4 \cdot {}_1F_1(\alpha_4; \beta_4; \rho(1-z_4)) - \lambda\alpha_4 \cdot {}_1F_1(\alpha_4+1; \beta_4+1; \rho(1-z_4))}, \quad \langle T_{\text{off}} \rangle = \frac{1}{\lambda}.$$

Next we examine whether the three effective parameters of a complex model can reflect the realistic dynamic properties of transcription and translation. To this end, we consider the relative error between $\bar{\rho}$ and ρ, the relative error between $\bar{\lambda}$ and $1/\langle T_{\text{off}} \rangle$, and the relative error between $\bar{\gamma}$ and $1/\langle T_{\text{on}} \rangle$, i.e.

$$\text{RE}_{\bar{\rho}} = \frac{\bar{\rho} - \rho}{\rho}, \quad \text{RE}_{\bar{\lambda}} = \frac{\bar{\lambda} - 1/\langle T_{\text{off}} \rangle}{1/\langle T_{\text{off}} \rangle}, \quad \text{RE}_{\bar{\gamma}} = \frac{\bar{\gamma} - 1/\langle T_{\text{on}} \rangle}{1/\langle T_{\text{on}} \rangle}.$$

Moreover, we say that an effective parameter is *over-estimated* (*under-estimated*) if the corresponding relative error is greater (less) than zero. For each complex model, we calculate the relative errors of the three effective parameters under all 625 parameter sets which are chosen to be the same as in Fig. 2(a). Fig. 3(a) illustrates the sample mean and sample standard deviation of $|\text{RE}_{\bar{\rho}}|$, $|\text{RE}_{\bar{\lambda}}|$, and $|\text{RE}_{\bar{\gamma}}|$ for all parameter sets (also see Supplementary Fig. S1 for the empirical distributions of $|\text{RE}_{\bar{\rho}}|$, $|\text{RE}_{\bar{\lambda}}|$, and $|\text{RE}_{\bar{\gamma}}|$). In addition, Fig. 3(a) also illustrates the empirical proportions of $|\text{RE}_{\bar{\rho}}|$, $|\text{RE}_{\bar{\lambda}}|$, and $|\text{RE}_{\bar{\gamma}}|$ being greater than 0.2. If the relative error of an effective parameter has an absolute value less than 0.2, then we believe that it can reflect the realistic dynamic property of the corresponding complex model.

For the three-state model, the sample mean of $|\text{RE}_{\bar{\lambda}}|$ and $|\text{RE}_{\bar{\gamma}}|$ are large, while $|\text{RE}_{\bar{\rho}}|$ has a relatively small sample mean; in particular, there are only 13.6% of parameter sets such that $|\text{RE}_{\bar{\rho}}| > 0.2$ (Fig. 3(a)). This suggests



effective parameters	TSM			CPM			PFM			NFM		
	$\bar{\lambda}$	$\bar{\gamma}$	$\bar{\rho}$	$\bar{\lambda}$	$\bar{\gamma}$	$\bar{\rho}$	$\bar{\lambda}$	$\bar{\gamma}$	$\bar{\rho}$	$\bar{\lambda}$	$\bar{\gamma}$	$\bar{\rho}$
over-estimated	√	√	√							√		
under-estimated				√	√	√	√	√	√		√	√

Figure 3: **Linking effective rate parameters to their real values in complex models.** **(a)** For each complex model, the absolute values of relative errors of the three effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ are computed under 625 parameter sets, along with their sample means, sample variances, and the sample frequencies of relative errors being greater than 0.2. The effective parameter $\bar{\rho}$ is closed to the synthesis rate ρ for the three-state, cross-talk pathway, and positive feedback models, while the effective parameter $\bar{\lambda}$ is closed to the gene activation frequency of the negative feedback model. **(b)** Accuracy of the three effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ for each complex model. For the three-state model, all effective parameters are over-estimated; for the cross-talk pathway and positive feedback models, all effective parameters are under-estimated; for the negative feedback model, $\bar{\lambda}$ is over-estimated, while $\bar{\gamma}$ and $\bar{\rho}$ are under-estimated. **(c)** For each complex model, 150 parameter sets are randomly generated such that $1/\langle T_{\text{off}}\rangle$ and $1/\langle T_{\text{on}}\rangle$ are between 0 and $2.5d$ (grey diamonds). For the three-state model, the scatter plot of $(\bar{\lambda}, \bar{\gamma})$ escapes from the potential bimodal region, i.e. $\bar{\lambda}, \bar{\gamma} < d$; for the cross-talk pathway and positive feedback models, the scatter plot of $(\bar{\lambda}, \bar{\gamma})$ moves towards the potential bimodal region; for the negative feedback model, the scatter plot of $(\bar{\lambda}, \bar{\gamma})$ neither escapes from nor moves towards the potential bimodal region.

that in most cases, the estimated value of the effective rate $\bar{\rho}$ is very close to the real synthesis rate ρ. Similar phenomenon is also observed for the cross-talk pathway and positive feedback models (Fig. 3(a)). In particular, for the positive feedback model, almost all values of $|\text{RE}_{\bar{\rho}}|$ are less than 0.2, suggesting that fitting the protein number distribution of the positive feedback model to the telegraph model can always provide a reliable estimation of the synthesis rate ρ. The situation is different for the negative feedback model, where both $|\text{RE}_{\bar{\gamma}}|$ and $|\text{RE}_{\bar{\rho}}|$ have a relatively large sample mean, while $|\text{RE}_{\bar{\lambda}}|$ has a relatively small sample mean (Fig. 3(a)). This suggests that the estimated value of the effective parameter $\bar{\lambda}$ can reflect the realistic gene activation frequency of a negative feedback model.

Interestingly, our simulations also reveal that the relative errors of the three effective parameters follow some consistent principles: (i) for the three-state model, all effective parameters are over-estimated; (ii) for the cross-talk pathway and positive feedback models, all effective parameters are under-estimated; (iii) for the negative feedback model, $\bar{\lambda}$ is over-estimated, while $\bar{\gamma}$ and $\bar{\rho}$ are under-estimated. Here the principles are consistent in the sense that it is impossible that an effective parameter is over-estimated for a certain parameter set, while it is under-estimated for another parameter set. These rules of over-estimation and under-estimation are summarized in Fig. 3(b).

These rules can be further used to characterize the ability for a complex model to exhibit bimodality. For each complex model, we randomly select 150 parameter sets such that the values of $1/\langle T_{\text{off}}\rangle$ and $1/\langle T_{\text{on}}\rangle$ are between 0 and $2.5d$ (see Methods for details), and for each parameter set, we fit the synthetic data obtained from the SSA to the telegraph model. The values of the real gene switching frequencies $1/\langle T_{\text{off}}\rangle$ and $1/\langle T_{\text{on}}\rangle$ are illustrated by the grey dots in Fig. 3(c). We then superpose the values of the effective rate parameters $\bar{\lambda}$ and $\bar{\gamma}$ for the 150 parameter sets (shown by the blue dots) onto the same figure.

For the telegraph model, it has been proved that a bimodal gene product distribution can only occur when both gene switching rates are smaller than the degradation rate, i.e. $\bar{\lambda} < d$ and $\bar{\gamma} < d$ [32]. The consistent principles summarized in Fig. 3(b) show that $\bar{\lambda}$ and $\bar{\gamma}$ are under-estimated for the cross-taking pathway and positive feedback models. This is also shown in Fig. 3(c), where the scatter plots of $(\bar{\lambda}, \bar{\gamma})$ for the two models move towards the

potential bimodal region, i.e. $\bar{\lambda} < d$ and $\bar{\gamma} < d$. Hence compared to the telegraph model, the cross-taking pathway and positive feedback models are more likely to produce a bimodal gene product distributions. In contrast, both $\bar{\lambda}$ and $\bar{\gamma}$ are over-estimated for the three-state model (Fig. 3(b)) and thus the scatter plot of $(\bar{\lambda}, \bar{\gamma})$ tends to escape from the potential bimodal region (Fig. 3(c)). This shows that the three-state model is less likely to exhibit bimodality, possible due to the fact that a multi-step gene activation process reduces gene expression noise [24, 66]. For the negative feedback model, $\bar{\lambda}$ is over-estimated and $\bar{\gamma}$ is under-estimated (Fig. 3(b)). Hence the scatter plot of $(\bar{\lambda}, \bar{\gamma})$ neither moves towards nor escape from the bimodal region (Fig. 3(c)). This indicates that negative autoregulation has weak influence on bimodality.

Identification of complex models using snapshot at multiple time points

Given the experimental distributions of mRNA or protein numbers in steady state, it is impossible to distinguish a complex model from the telegraph model since the latter can accurately capture the steady-state distribution of the former. To distinguish between them, additional information such as the snapshot data at multiple discrete time points within the same cell population measured using e.g., live-cell imaging, flow cytometry, smFISH, and scRNA-seq, is needed [55]. Recent studies have proposed various statistical methods, such as the maximum likelihood method [55, 67] and the Bayesian method [50, 68], to search optimal kinetic parameters based on single-cell data at multiple time points. However, no matter which method is used, the first and most important step is to determine which model (the telegraph model or more complex models) is the most competitive to describe the snapshot data [55].

We next examine the dynamic distinctions between a complex model and its effective telegraph model. Here we assume that initially there is no mRNA or protein molecules in the cell and the gene is off. This mimics the situation where the gene has been silenced by some repressor over a period of time such that all gene product molecules have been removed via degradation. At time $t = 0$, the repressor is removed and we investigate how gene expression recovers. Note that a complex model and its effective telegraph models share very close gene product distributions in steady state. However, they may exhibit completely different dynamic behaviors since their time-dependent distributions are generally different. For each complex model, we compute the time-dependent mean $M(t)$ and variance $\sigma^2(t)$ of gene product fluctuations using the finite-state projection algorithm (FSP) [67] under all 625 parameter sets which are chosen to be the same as in Fig. 2. The time-dependent mean and variance for the effective telegraph model are denoted by $\bar{M}(t)$ and $\bar{\sigma}^2(t)$, respectively.

Interestingly, for the three-state and positive feedback models, we find that the mean curve $M(t)$, as a function of time t, is always below its counterpart $\bar{M}(t)$ for the effective telegraph model for all 625 parameter sets. In contrast, the mean curve $M(t)$ for the cross-talk pathway and negative feedback models is always above the mean curve $\bar{M}(t)$ for the effective telegraph model (Fig. 4(a)). This is probably because compared to the telegraph model, the three-state and positive feedback models have slower relaxation speed to the steady state, while the cross-talk pathway and negative feedback models relax to the steady state faster [69]. In particular, the latter two models may even perform overshooting behavior [70] where the maximum value of the mean curve $M(t)$ exceeds its steady-state value (shown by the blue dashed curve in Fig. 4(a)) (it has long been known that a negative feedback system is capable of producing overshooting behavior [71]). Similar dynamic features are also observed for the time-dependent second moment $\langle n^2 \rangle(t)$; however, common indicators of gene expression noise, such as the coefficient variation squared $\sigma^2(t)/M^2(t)$ and the Fano factor $\sigma^2(t)/M(t)$, present less easily distinguishable dynamic differences between a complex model and its effective telegraph model (Supplementary Fig. S2).

The above results provide an effective method of distinguishing a complex model from the simple telegraph model. Given single-cell data at multiple time points, we can first fit the steady-state distribution to the telegraph model and estimate the effective rate parameters $\bar{\lambda}/d, \bar{\gamma}/d$ and $\bar{\rho}/d$, where the degradation rate d can be determined by measuring the half-life of the gene product [1, 12, 13]. We then compare the experimental mean curve $M(t)$ obtained from the snapshot data and the mean curve $\bar{M}(t)$ of the effective telegraph model. If the former is always

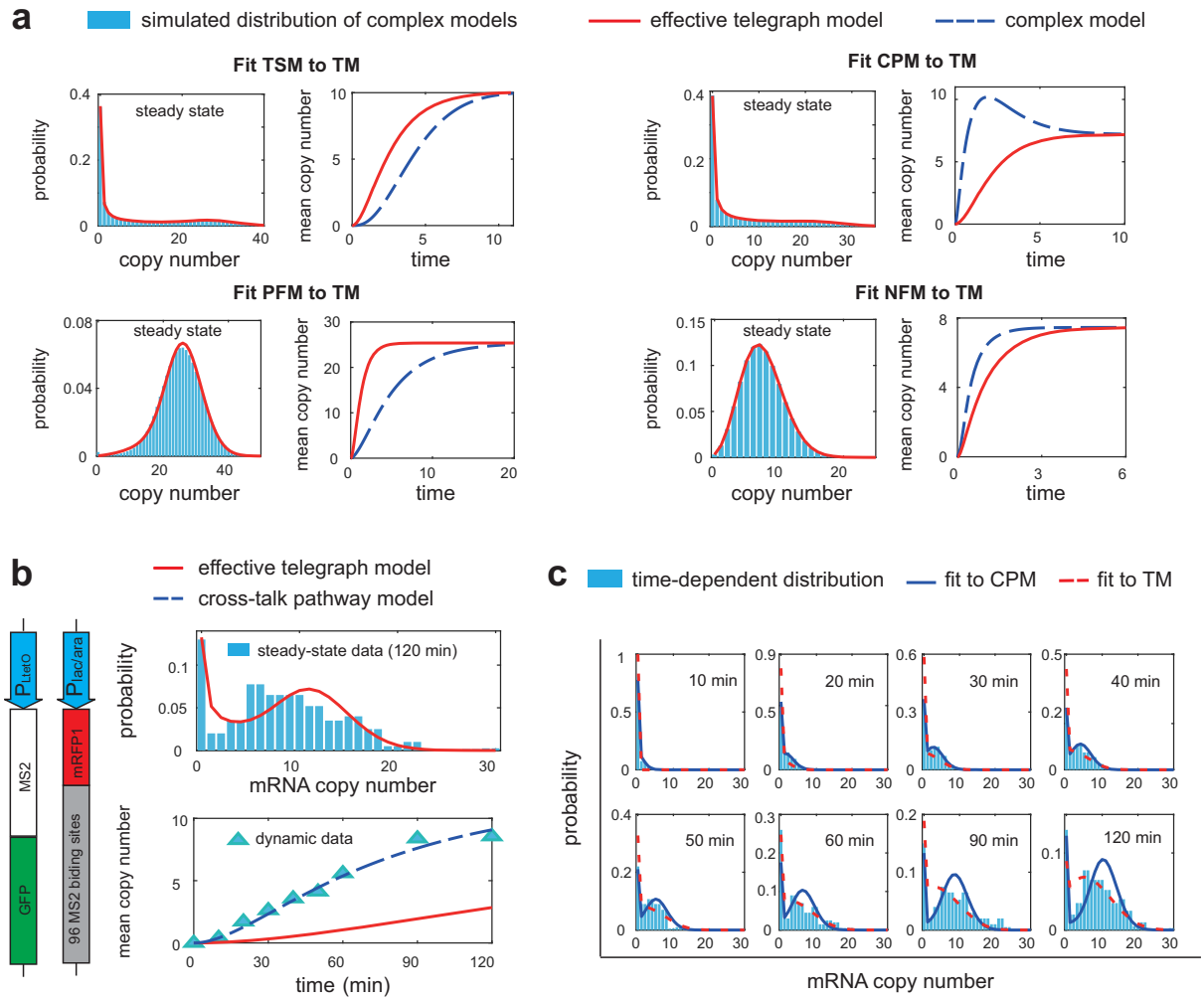


Figure 4: **Determining the most competitive model to describe single-cell data at multiple time points.** **(a)** Under the initial condition that the gene is off and there is no gene product molecules in the cell, the three-state and positive feedback models have smaller time-dependent mean curve compared to the effective telegraph model, while the cross-talk pathway and negative feedback models have larger time-dependent mean curve. **(b)** In *E. coli* cells, the mRNA of interest, under the control of an inducible promoter $\mathrm{P_{lac/ara}}$, was consisted of the coding region for a red fluorescent protein mRFP1, followed by a tandem array of 96 MS2 binding sites [1]. The green fluorescent protein (GFP), independently produced by the promoter $\mathrm{P_{LtetO}}$, tagged the target mRNA by binding to the MS2 sites. The number of target mRNA molecules in an individual cell was computed using the fluorescence intensities of GFP at nine time points: 0, 10, 20, 30, 40, 50, 60, 90, 120 min. The steady-state mRNA number distribution (at 120 min) was fitted to the telegraph model with the measured degradation rate $d = 0.014$ min^{-1} [1] and the three effective rate parameters are estimated to be $(\bar{\lambda}/d, \bar{\gamma}/d, \bar{\rho}/d) = (0.36, 0.18, 12.5)$. The time-dependent mean expression levels predicted by the effective telegraph model (red curve in the lower-right panel) are much less than the real mean expression levels (blue triangles), suggesting that the cross-talk pathway model is a potential candidate to describe the transcription data at discrete time points. **(c)** The cross-talk pathway model (blue curves) provides a much better fit to the time-dependent mRNA number distributions than the telegraph model (red dashed curves). The estimated parameters for the cross-talk pathway model are given by $q_1 = 0.13$, $q_2 = 1 - q_1$, and $(\lambda_1, \lambda_2, \gamma, \rho) = (0.0013, 0.055, 0.0042, 0.21)$ min^{-1}. The snapshot transcription data used are from [1].

above the latter, then we have good reasons to believe that the cross-talk pathway or the negative feedback model is more competitive to describe the snapshot data. In contrast, if the former is always below the latter, then we may select the three-state or the positive feedback model to explain the snapshot data.

Validation of theoretical results using snapshot data at multiple time points

To validate our theory, we apply it to a data set of mRNA expression at multiple discrete time points measured in living *E. coli* cells [1]. In this experiment, anhydrotetracycline was first added to a growing culture, which

induced the production of green fluorescent protein (GFP) to tag the mRNA of interest. The mRNA target, under the control of an inducible promoter $P_{\mathrm{lac/ara}}$, is consisted of the coding region for a red fluorescent protein mRFP1, followed by a tandem array of 96 MS2 binding sites. Activation of the promoter was induced using high levels of arabinose and isopropylthio-β-D-galactoside (IPTG). The GFP can then bind to the MS2 binding sites and the synthesized transcripts were tagged by MS2-GFP fusion protein. Samples were imaged using fluorescence microscopy at nine different time points from 0-120 min, and the number of mRNA molecules in individual cells was computed according to green fluorescence intensities.

All cells contained no green foci at time $t = 0$, suggesting that initially there are no mRNA molecules. Under the induction of arabinose and IPTG, the mean number of transcripts increases monotonically and approaches the steady state at 120 min (Fig. 4(b)). The mRNA expression in steady state exhibits an appearent bimodal distribution which has a zero and a nonzero peak. We then use a generalized moment-matching method [18] to fit the steady-state distribution to the telegraph model since this method performs better in capturing bimodality than the conventional maximum likelihood method (Fig. 4(b)). The three effective rate parameters are then estimated to be $\bar{\lambda}/d = 0.36$, $\bar{\gamma}/d = 0.18$, and $\bar{\rho}/d = 12.5$. Furthermore, the mRNA tagged by MS2-GFP fusion protein is very stable and its degradation rate is measured to be $d = 0.014$ min^{-1} [1]. In Fig. 4(b), we compare the experimental mean curve $M(t)$ and the mean curve $\bar{M}(t)$ of the effective telegraph model computed using the effective parameters. Clearly, $M(t)$ is always above $\bar{M}(t)$ for all time points. Our theory then suggests that the cross-talk pathway and negative feedback models are more reasonable than the telegraph model. Since the mRFP1 protein does not bind to its own promoter to form an autoregulatory feedback loop, we have good reasons to believe that the cross-talk pathway model may be the most competitive to describe the data set.

We next estimate all the rate parameters of the cross-talk pathway model based on the time-dependent distributions of mRNA numbers. An accurate parameter inference can be achieved by minimizing the following log-likelihood function summed over all nine time points [67]:

$$\log L(\theta) = \sum_{l=1}^{9} \sum_{n} N(t_l, n) \log(P_n(t_l, \theta)),$$

where θ is the parameters set of the cross-talk pathway model, $N(t_l, n)$ denotes the number of cells with n transcripts at time $t = t_l$, and $P_n(t_l, \theta)$ denotes the theoretical mRNA number distribution at time $t = t_l$. Here the theoretical distribution is computed using FSP assuming that initially there are no mRNA molecules and the gene is off. To handle the positive constraint on rate parameters, we rewrite $\theta = e^{\tilde{\theta}}$ and set $\tilde{\theta}$ to be the optimization variables [68]. One exception is the selection probability $q_2 \in (0, 1)$ of the strong signalling pathway, for which we rewrite $q_2 = 1 - e^{-|\tilde{q}_2|}$ and set $\tilde{q}_2$ to be the optimization variable. The initial parameters for optimization are chosen to be $q_1 = q_2 = 0.5, \lambda_1 = \lambda_2 = \bar{\lambda}, \gamma = \bar{\gamma}$ and $\rho = \bar{\rho}$, where $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ are the effective rate parameters of the telegraph model.

Fig. 4(c) illustrates the experimental mRNA number distributions at all measured time points and the optimal fit of these distributions to the cross-talk pathway model (blue curves) and the telegraph model (red dash curves). It can be seen that the former indeed behaves much better than the latter. First, the total HD value for the cross-talk pathway model summed over all nine time points is 1.03, which is less than a much higher HD value of 1.67 for the telegraph model. Second, the bimodal distributions after 30 min can be very well captured by the cross-talk pathway model but fail to be captured by the telegraph model. Our results imply that the activation of the $P_{\mathrm{lac/ara}}$ promoter is likely to be realized via the competition between a weak and strong signalling pathways. This is supported by the biological fact that activation of the promoter is regulated by both the repressor LacI and the activator AraC, which compete to bind to the promoter [1].

Inference of gene regulation mechanisms using parameter-varying data

Experimentally, gene expression data are often measured under different experimental conditions. One of the most common experimental strategies is to modulate the value of only one parameter and keep the values of other

parameters invariant, e.g. to modulate the feedback strength in a genetic feedback loop while keep the protein synthesis and degradation rates, as well as the gene switching rates the same [72]. Given gene expression data under different experimental conditions, a natural question is whether we can infer the underlying gene regulation mechanisms, e.g. whether there is a feedback loop, multiple gene states, or multiple gene activation signalling pathways.

To answer this, for each of the four complex models, we generate synthetic data of mRNA or protein numbers using the SSA under 40 different experimental conditions, where we tune the value of only one parameter and fix the values of other parameters. The parameter that is modulated will be referred to as the *tuning parameter* in what follows. When performing simulations, we fix $\lambda_1 = 0.2$ for the cross-talk pathway model so that each complex model has four tuning parameters. We choose 10 different values for each tuning parameter and thus there are a total of $4 \times 10 = 40$ experimental conditions for each complex model. For each experimental condition, we then fit the synthetic data to the telegraph model and obtain the effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$.

Note that tuning a single parameter will lead to a change in the mean expression level and will also give rise to changes in the effective rate parameters. For each complex model and each tuning parameter, we illustrate $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ as a function of the mean expression level (Fig. 5(a)). For convenience, the value of each effective parameter is normalized to unity at the lowest mean expression level. Interestingly, we find that when modulating a single parameter, the changes in the effective parameters follow some consistent principles. The rules for the three-state model are simple: variations in the gene activation rate λ_1 (or λ_2), gene inactivation rate γ, and synthesis rate ρ result in changes in their counterparts $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ in the effective telegraph model, respectively (Fig. 5(a)). For the cross-talk pathway model, variations in γ and ρ result in changes in their counterparts $\bar{\gamma}$ and $\bar{\rho}$, respectively, while tuning either the selection probability q_2 or the gene activation rate λ_2 of the strong signaling pathway gives rise to simultaneous variations in both $\bar{\lambda}$ and $\bar{\gamma}$ (Fig. 5(a)). Furthermore, we can distinguish between the regulations of q_2 and λ_2 since the increase of q_2 leads to increasing $\bar{\lambda}$ and decreasing $\bar{\gamma}$, while the increase of λ_2 leads to non-monotonic $\bar{\lambda}$ and decreasing $\bar{\gamma}$.

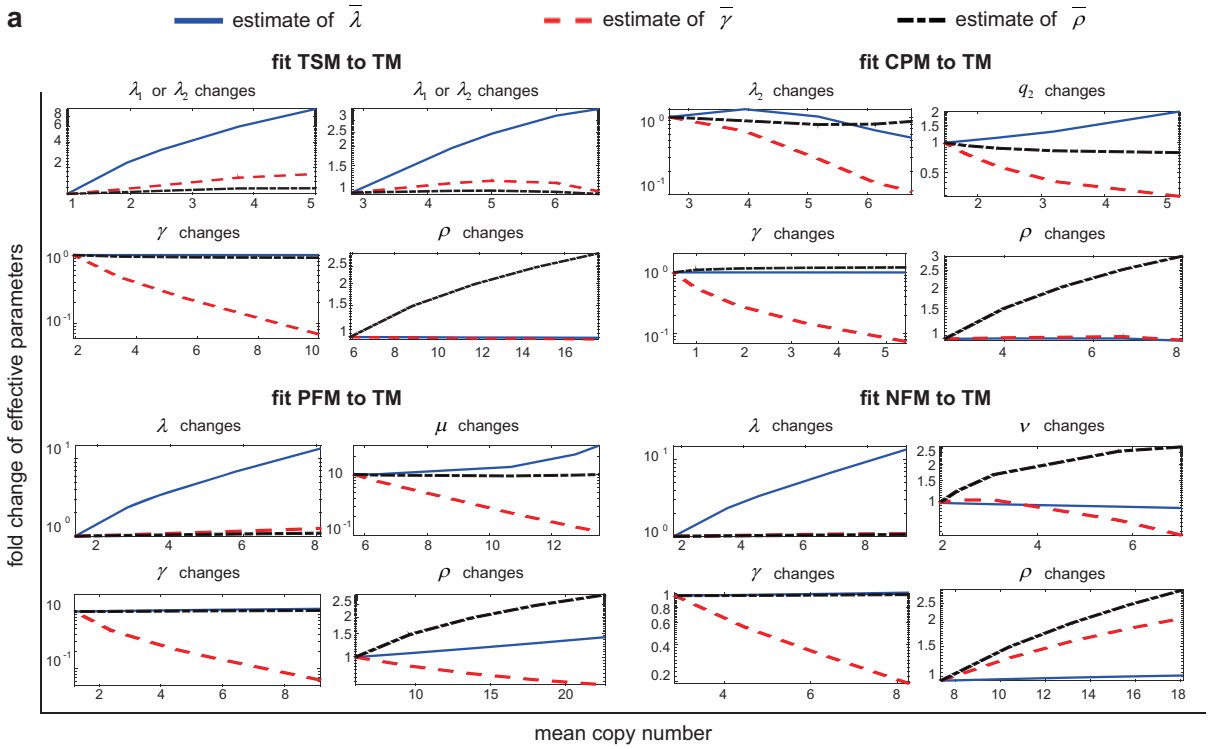
For the positive and negative feedback models, variations in the gene switching rates λ and γ result in changes in their counterparts $\bar{\lambda}$ and $\bar{\gamma}$ in the effective telegraph model, respectively (Fig. 5(a)). The situation is different when modulating the feedback strengths μ and ν, as well as the synthesis rate ρ. From Fig. 5(a), it is clear that the increase of the positive feedback strength μ leads to increasing $\bar{\lambda}$ and decreasing $\bar{\gamma}$; the decrease of the negative feedback strength ν gives rise to increasing $\bar{\rho}$ and decreasing $\bar{\gamma}$; the increase of the synthetic rate ρ results in simultaneous increase of both $\bar{\rho}$ and $\bar{\gamma}$ for the negative feedback model and results in the variations of all effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ for the positive feedback model.

We emphasize that the above rules are actually independent of the choice of model parameters. To see this, for each complex model and each tuning parameter, we repeat the above procedures under $5^3 = 125$ parameter sets, where we choose five different values for each of the three fixed parameters (see Methods for details). All 125 parameter sets give rise to the same principles as in Fig. 5(a) (also see Supplementary Tables S1-S4). For clarity, we summarize them in Fig. 5(b). These consistent rules not only reveal the influence of the tuning parameter on the effective rate parameters, but also provide a potentially useful way of inferring the underlying gene regulation mechanism by using gene expression data under different induction conditions. For example, if we observe increasing λ (ρ) and decreasing γ as a function of the mean expression level in a gene network in response to varying experimental conditions, then we have good reasons to conjecture that there is a positive (negative) feedback loop within the network.

Validation of theoretical results using synthetic gene networks

To validate our theory, we apply it to a synthetic gene regulatory network (orthogonal property of a synthetic network can minimize extrinsic noise) stably integrated in human kidney cells, as illustrated in Fig. 6(a) [72]. In this network, a bidirectional promoter is designed to control the expression of two fluorescent proteins: zsGreen and

a



b

		parameters varied in TSM				parameters varied in CPM				parameters varied in PFM				parameters varied in NFM			
		λ_1	λ_2	γ	ρ	λ_2	q_2	γ	ρ	λ	μ	γ	ρ	λ	ν	γ	ρ
effective parameters varied in TM	$\bar{\lambda}$	√	√							√				√			
	$\bar{\gamma}$			√				√				√				√	
	$\bar{\rho}$				√				√								
	$\bar{\lambda}$ & $\bar{\gamma}$					√	√				√						
	$\bar{\gamma}$ & $\bar{\rho}$														√		√
	$\bar{\lambda}$ & $\bar{\gamma}$ & $\bar{\rho}$												√				

Figure 5: **Variation pattern of effective rate parameters under different induction conditions in each of the four complex models. (a)** Tuning a single parameter of a complex model can generate a series of distributions of gene product numbers, along with different mean expression levels. Fitting these distributions by the telegraph model leads to a series of effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$. Plotting $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ as a function of the corresponding mean expression level reveals how the effective rate parameters vary when the single parameter of a complex model is tuned. **(b)** Effective rate parameters changed when tuning a single parameter of each complex model. For example, for the positive feedback model, the effective parameter $\bar{\lambda}$ changes when tuning the parameter λ, while all effective parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ change when tuning the the parameter ρ.

dsRed. The activity of the promoter can be activated in the presence of Doxycycline (Dox). The green fluorescent protein, zsGreen, is fused upstream from the transcriptional repressor LacI. The LacI protein binds to its own gene and inhibits its own transcription, forming a negative autoregulatory feedback loop. The negative feedback strength can be tuned by induction of IPTG. As a control architecture, the red fluorescent protein, dsRed, is not regulated by induction of IPTG, forming a network with no feedback. The steady-state fluorescence intensities of zsGreen and dsRed are measured under ten different IPTG concentrations (0, 0.2, 0.4, 0.8, 1.6, 3.1, 6.2, 12.5, 25, 50 μM) and two different Dox concentrations (low and high) using flow cytometry.

Note that in this experiment, it is the fluorescence intensities of the two proteins that are measured, rather than their copy numbers. Hence it is crucial to determine the proportionality constant between fluorescence intensities and copy numbers. In other words, we need to convert the fluorescence intensity x into the copy number $n = [x/\beta]$, where β is the fluorescence intensity per protein copy and $[a]$ denotes the integer part of a. For zsGreen, since negative feedback is weak when IPTG concentration is high, the value of β is chosen such that the mean number of

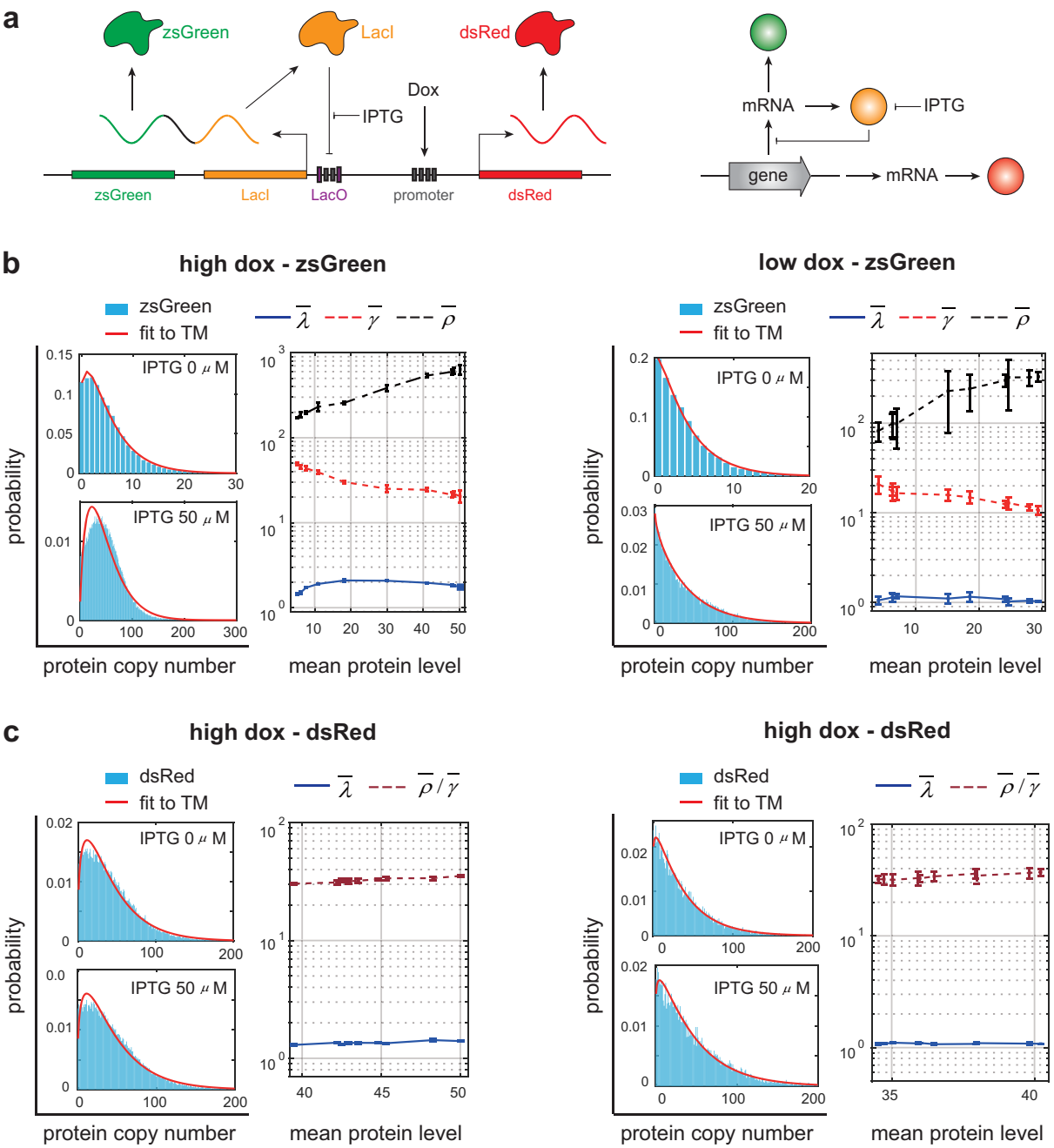


Figure 6: **Unravelling the regulation mechanism in a synthetic gene network integrated in human kidney cells. (a)** The bidirectional promoter transcribes the zsGreen-LacI and dsRed transcripts. The gene network includes two architectures: a negative-feedback network and a network with no feedback. The zsGreen-LacI transcripts are inhibited by LacI, forming a network with negative autoregulation. The dsRed transcripts are not regulated, forming a network with no feedback. The activity of the promoter can be activated in the presence of Doxycycline (Dox), and the negative feedback strength can be tuned by induction of isopropylthio-β-D-galactoside (IPTG). **(b)** Under both high and low Dox levels, fitting the distributions of zsGreen levels under different IPTG concentrations to the telegraph model leads to increasing $\bar{\rho}$, decreasing $\bar{\gamma}$, and almost invariant $\bar{\lambda}$ against the mean expression level. Such variation pattern of the three effective parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ exactly coincides with that in the negative feedback model when the feedback strength ν is tuned. **(c)** Under both high and low Dox levels, fitting the distributions of dsRed levels under different IPTG concentrations to the telegraph model leads to almost invariant values of $\bar{\lambda}$ and $\bar{\rho}/\bar{\gamma}$ against the mean expression level. The error bars in (b) and (c) show the standard deviation of three repeated experiments.

zsGreen molecules $\langle n \rangle = 50$ at the highest IPTG concentration (50 μM) and at high Dox concentration, which is compatible with the typical number of LacI repressor in the *lac* operon [73]. For dsRed, since its expression is not regulated by IPTG induction, the value of β is chosen so that the mean number of dsRed molecules $\langle n \rangle = 50$ at

high Dox concentration (see Supplementary Table S5 for the specific values of β for the two proteins).

We then fit the copy number distributions of zsGreen to the telegraph model at all IPTG and Dox concentrations and obtain the effective rate parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ (Supplementary Table S6). Note that the increase of IPTG concentration will lead to the increase of the mean expression level of zsGreen. In Fig. 6(b), we illustrate $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ against the mean expression level as IPTG concentration varies. It is clear that at both low and high Dox concentrations, the increase of IPTG concentration results in increasing $\bar{\rho}$ and decreasing $\bar{\gamma}$, while the value of the $\bar{\lambda}$ is almost unaffected by IPTG induction. This is in perfect agreement with the consistent rule shown in Fig. 5(a),(b) for the negative feedback model as the negative feedback strength ν changes.

Similarly, we also repeat the above procedures for dsRed. Interestingly, we find that fitting the copy number distributions of dsRed to the telegraph model will lead to extremely large values of $\bar{\gamma}$ and $\bar{\rho}$, suggesting the copy number of dsRed has a negative binomial distribution (the gene product distribution of the telegraph model reduces to a negative binomial when gene expression is sufficiently bursty, i.e. $\bar{\gamma} \gg \bar{\lambda}$ and $\bar{\rho}/\bar{\gamma}$ is finite [74]). In this case, only the burst frequency $\bar{\lambda}$ and the burst size $\bar{\rho}/\bar{\gamma}$ can be accurately inferred [18]. Since dsRed expression is unregulated by IPTG induction, in both the high and low Dox cases, the values of $\bar{\lambda}$ and $\bar{\rho}/\bar{\gamma}$ are almost invariant when plotted against the mean expression level as IPTG concentration varies (Fig. 6(c)).

Conclusions and discussion

A central question in molecular biology is to understand various genetic regulation mechanisms and how they modulate the production of mRNA and protein at the single-cell level [25, 44]. The classical two-state telegraph model has been extensively used to explain massive single-cell gene expression data so that one can estimate the underlying kinetic parameters and unravel the gene regulation mechanisms in response to varying environmental changes [15, 17, 26]. However, the telegraph model is limited in its predictive power since it lacks a description of some biological mechanisms that are known to have a profound impact on the mRNA and protein number distributions in single cells. In the presence of complex biological mechanisms, fitting the gene expression data to the simple telegraph model [17] may lead to inaccurate parameter inference and even incorrect predictions of the underlying gene regulation mechanisms.

In the present paper, we investigate the dynamical properties of four relatively complex gene expression models, including the three-state model, cross-talk pathway model, positive feedback model, and negative feedback models. Compared with the telegraph model, these models describe how fluctuations are influence by complex biological mechanisms such as non-exponential gene inactivation durations, multiple gene activation pathways, and feedback regulation. Despite the potential risks, we found that fitting the steady-state mRNA or protein number distributions of these complex models to the simple telegraph model may still provide a large amount of useful information.

First, we showed that the gene product number distributions, as well as the conditional distributions in the active gene state, of the four complex models can all be well fitted by the telegraph model. The telegraph model with optimal fit was referred to as the effective telegraph model and the associated three optimal parameters were referred to as the effective rate parameters. We found that while most effective parameters may deviate significantly from their real values in the complex models, there are still some parameters that can be reliably estimated with very small relative relative. For the three-state, cross-talk pathway, and positive feedback models, the effective synthesis rate is very closed to the its real value, while the effective gene activation and inactivation rates deviate largely from their real values. At the first glance, only the gene activation mechanisms in the three complex models differ from that in the telegraph model. However, our results showed that fitting the steady-state distributions of the complex models to the telegraph model may lead to an unreliable estimation of both the gene activation and inactivation rates [7], but does not significantly influence the synthesis rate. For the negative feedback model, we showed that the effective synthesis and gene inactivation rates are unreliable, while the effective gene activation rate is very closed to its real value.

The effective rate parameters also provide a natural and convenient way of characterizing the capability for a complex model to exhibit bimodal gene product distributions. This characterization is based on a mathematical result [32, 33] which shows that the telegraph model can generate a bimodal distribution only when its gene activation and inactivation rates are both smaller than the degradation rate. For the three-state model, the effective gene switching rates are both over-estimated compared to their real values, which make bimodal distributions difficult to occur. In contrast, for the cross-talk pathway and positive feedback models, the effective gene switching rates are both under-estimated, and thus these two models are more likely to exhibit bimodality. For the negative feedback model, one of the effective gene switching rates is over-estimated while the other is under-estimated, which exerts a weak influence on bimodality.

Furthermore, we showed that the effective rate parameters can be used to distinguish a complex model from the telegraph model by using additional single-cell data at multiple time points. The good fit of complex models to the telegraph model in steady state indicates that it is impossible to distinguish between a complex model and the telegraph model by only using the steady-state gene expression data. Previous studies showed that a non-monotonic dynamic feature of gene expression mean can rule out the telegraph model since the telegraph model can only display a monotonic time-dependent curve of gene expression mean [52, 53]. However, this idea is limited when the gene expression mean displays a monotonic dynamics. To solve this, we compared the time-dependent mean curves of a complex model and its effective telegraph model, where the effective rate parameters were estimated by using the steady-state data. We showed that if the mean curve for the effective telegraph model is always below the real mean curve, then the three-state or positive feedback model is more component to describe the data compared to the telegraph model. In contrast, if the former is always above the latter, we may select the cross-talk pathway or negative feedback model to explain the data. As a validation of our theory, we apply our method to the snapshot mRNA expression data of the $P_{\mathrm{lac/ara}}$ promoter at multiple time points measured in *E. coli* cells [1]. We showed that among the four complex models, the cross-talk pathway model is the most competitive and thus we predict that the activation of $P_{\mathrm{lac/ara}}$ is very likely to be regulated by the competition between two signalling pathways.

Finally, we showed that the effective rate parameters can be used to unravel the gene regulation mechanisms of a complex model in response to varying environmental conditions. For a series of gene product distributions generated by tuning a single parameter of a complex model, fitting those distributions to the telegraph model gives a certain variation pattern of the three effective parameters. We found that the variation pattern of the three effective parameters is independent of the choice of parameters of the complex model. Hence it is possible to determine the underlying gene regulatory mechanism in response to environmental changes by identifying the variation pattern of effective parameters. We test our method by using a synthetic autoregulatory gene circuit in human kidney cells where the feedback loop is designed to suppress gene expression in response to the induction of varying IPTG concentrations [72]. Fitting the data under all induction conditions reveals a certain variation pattern of effective parameters, which is in perfect agreement with that of the negative feedback model by tuning the negative feedback strength. In contrast, fitting the data for an unregulated system to the telegraph model results in almost constant effective parameters as IPTG concentration varies.

In summary, the simple telegraph model should be used with caution when there are complex mechanisms behind the underling gene expression system. However, fitting the steady-state distribution of a complex gene expression model to the telegraph model can still reveal rich information such as the estimation of some kinetic parameters of the complex model, the capability for the complex model to produce bimodality, model selection, and gene regulation mechanisms in response to environmental changes. Future work is required to test our methods by adding more detailed biological mechanisms into the telegraph model, including the bursty production of mRNA or protein [75, 76], cell cycle events such as cell growth and division [77–79], cell-volume dependence [80–82], as well as complex gene regulatory networks [83, 84]. In addition, we anticipate that our theory can be enriched by fitting the time-dependent distributions (rather than only the steady-state distributions) of complex models to the simple telegraph model.

Methods

Initial values of parameter inference for the telegraph model

We use the "fminsearch" function in MATLAB to minimize the log-likelihood function Eq. (2) so that we can obtain estimates of the three parameters $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ of the telegraph model. One of the most important steps for optimization is to select the initial values $\bar{\lambda}_0$, $\bar{\gamma}_0$, and $\bar{\rho}_0$ of the three parameters. The conventional moment estimation method determines $\bar{\lambda}_0$, $\bar{\gamma}_0$, and $\bar{\rho}_0$ as functions of the first three moments of gene expression data across a cell population by solving three algebraic equations [14, 20]. Although this method is straightforward, it may produce negative values of $\bar{\lambda}_0$, $\bar{\gamma}_0$, and $\bar{\rho}_0$ [14]. To avoid the possible negative initial values, we modify the moment estimation method as follows. If $\bar{\rho}_0$ is known, then the two initial values $\bar{\lambda}_0$ and $\bar{\gamma}_0$ can be determined by [16, 18]

$$\bar{\lambda}_0 = \frac{M(\bar{\rho}_0 + 1 - \eta M - M)}{\bar{\rho}_0(\eta M - 1)}, \quad \bar{\gamma}_0 = \frac{\bar{\rho}_0 - M}{M}\bar{\lambda}_0, \tag{8}$$

where M and η are the mean and noise (coefficient of variation squared) of gene product numbers, respectively. In previous papers [16], $\bar{\rho}_0$ is commonly chosen as the maximum number of gene product molecules within the cell population. However, this choice may still give rise to negative values of $\bar{\lambda}_0$ and $\bar{\gamma}_0$ in Eq. (8) when the maximum number of molecules $\bar{\rho}_0 < \eta M + M - 1$. If this happens, then we simply reset $\bar{\rho}_0 = \eta M + M$. To test this method, we use the telegraph model to generate synthetic data of gene products numbers for $N = 10^4$ cells using the SSA under 10^3 randomly selected parameter sets within the regions $\bar{\lambda}, \bar{\gamma} \in [0.1, 4]$ and $\bar{\rho} \in [10, 30]$. The synthetic data are then fitted to the telegraph model using the maximum-likelihood method with the initial values $\bar{\lambda}_0$, $\bar{\gamma}_0$, and $\bar{\rho}_0$ being chosen as above. For all 10^3 parameter sets, the estimated values of $\bar{\lambda}$, $\bar{\gamma}$, and $\bar{\rho}$ are almost as the same as their real values. This shows that our parameter inference method is highly reliable.

Calculation of the mean active and inactive periods for feedback models

In [62], it has been shown that no matter there is autoregulation or not, the mean of gene product numbers can always be represented by the mean active and inactive periods as

$$\langle n \rangle = \rho \times \frac{\langle T_{\text{on}} \rangle}{\langle T_{\text{on}} \rangle + \langle T_{\text{off}} \rangle},$$

where $\langle T_{\text{on}} \rangle$ and $\langle T_{\text{off}} \rangle$ are the mean active and inactive periods of the gene. This formula is intuitive for unregulated genes. However, for regulated genes, it is highly non-trivial and was proved in [62] using the analytical distributions of gene product numbers. For the positive feedback model, the gene expression mean $\langle n \rangle$ has the form of [62]

$$\langle n \rangle = \rho \times \frac{\alpha_3 \cdot {}_1F_1(\alpha_3 + 1; \beta_3 + 1; \rho(1 - z_3))}{\beta_3 \cdot {}_1F_1(\alpha_3; \beta_3; \rho(1 - z_3))}.$$

Then substituting $\langle T_{\text{on}} \rangle = 1/\gamma$ into Eq. (7) gives the expression of $\langle T_{\text{off}} \rangle$. Similarly, for the negative feedback model, the gene expression mean is given by [62]

$$\langle n \rangle = \rho \times \frac{\alpha_4 \cdot {}_1F_1(\alpha_4 + 1; \beta_4 + 1; \rho(1 - z_4))}{\beta_4 \cdot {}_1F_1(\alpha_4; \beta_4; \rho(1 - z_4))}.$$

This formula, together with $\langle T_{\text{off}} \rangle = 1/\lambda$, gives the expression of $\langle T_{\text{on}} \rangle$.

Selection of parameter sets for complex models

In Fig. 2, we generate synthetic data of gene products numbers under 625 different parameter sets with the degradation rate $d \equiv 1$ for each of the four complex models. For the three-state model, we set $\rho = 10, 15, 20, 25, 30$ and $\lambda_1, \lambda_2, \gamma = 0.3, 0.7, 1, 2, 4$, which gives $5^4 = 625$ combinations of the four parameters. Note that λ_1 and λ_2 are symmetric in the analytical distribution given in Eq. (3), and thus there are actually $(5 + 4 + 3 + 2 + 1) \times 5^2 = 375$

different parameter sets for this model. For the cross-talk pathway model, the gene activation rate for the weak signaling pathway is fixed to be $\lambda_1 = 0.2$. The other four parameters are chosen to be $\rho = 10, 15, 20, 25, 30$ $\lambda_2, \gamma = 0.5, 1, 2, 4, 8$, and $q_1 = 0.1, 0.3, 0.5, 0.7, 0.9$. Similarly, for the positive feedback and negative feedback models, we set $\rho = 10, 15, 20, 25, 30$, $\lambda, \gamma = 0.3, 0.7, 1, 2, 4$, and $\mu, \nu = 0.05, 0.1, 0.5, 1, 1.5$.

In Fig. 3(c), we use the "rand" function in MATLAB to randomly generate 150 parameter sets such that the values of $1/\langle T_{\text{off}}\rangle$ and $1/\langle T_{\text{on}}\rangle$ are between 0 and $2.5d$ with the degradation rate $d \equiv 1$ for each of the four complex models. The synthesis rate ρ is randomly selected so that $\rho \in [10, 30]$ for each complex model. For the three-state, cross-talk pathway, and positive feedback models, $1/\langle T_{\text{on}}\rangle = \gamma$ is randomly selected so that $\gamma \in [0.1, 2.5]$. For the negative feedback model, $1/\langle T_{\text{off}}\rangle = \lambda$ is randomly selected so that $\lambda \in [0.1, 2.5]$. Moreover, for the three-state model, both λ_1 and λ_2 are selected so that $\lambda_1, \lambda_2 \in [0.1, 5]$; this ensures that

$$1/\langle T_{\text{off}}\rangle = 1/(1/\lambda_1 + 1/\lambda_2) \in [0.05, 2.5].$$

For the cross-talk pathway model, q_1, λ_1, and λ_2 are randomly selected so that $q_1 \in [0.1, 0.9], \lambda_1 \in [0.05, 0.5]$, and $\lambda_2 \in [1, 8]$; this ensures that

$$1/\langle T_{\text{off}}\rangle = 1/(q_1/\lambda_1 + q_2/\lambda_2) \in [0.05, 3.2],$$

and we randomly select 150 parameter sets that satisfy $1/\langle T_{\text{off}}\rangle \leq 2.5$. The analytical formula of $\langle T_{\text{off}}\rangle$ for the positive feedback model and the analytical formula of $\langle T_{\text{on}}\rangle$ for the negative feedback model are too complicated to directly calculate their upper and lower bounds. To overcome this, we restrict $\lambda \in [0.1, 1.5]$ and $\mu \in [0.01, 0.15]$ for the positive feedback model and randomly select 150 parameter sets so that $1/\langle T_{\text{off}}\rangle \leq 2.5$. Similarly, we restrict $\gamma \in [0.1, 1.5]$ and $\nu \in [0.01, 0.15]$ for the negative feedback model and randomly select 150 parameter sets so that that $1/\langle T_{\text{on}}\rangle \leq 2.5$.

In Fig. 5, we tune a single parameter while fix the other parameters for each of the four complex models. The values for each parameter are selected to be the same as in Fig. 2. Hence each complex model has four parameters to be tuned and each tuning parameter is equipped with five different values. For each complex model, we tune a single parameter among five different values, and hence there are total $5^3 = 125$ combinations of the other three parameters. To observe a significant effect of the tuning parameter on the gene expression level, we only consider those combinations of the other three parameters for which the mean expression level changes by at least two folds.

Author contributions

C. J. conceived the original idea. F. J. and C. J. performed the theoretical derivations. F. J., J. L., T. L., Y. Z., and W. C. perform the numerical simulations and data analysis. F. J., L. B., and C. J. interpreted the theoretical results. F. J. and C. J. jointly wrote the manuscript.

Acknowledgements

F. J. acknowledges support from National Natural Science Foundation of China with grant No. 12271118. C. J. acknowledges support from National Natural Science Foundation of China with NSAF grant No. U2230402, grant No. 12131005, and grant No. 12271020.

References

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

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

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

[4] Lenstra, T. L., Rodriguez, J., Chen, H. & Larson, D. R. Transcription dynamics in living cells. *Annual review of biophysics* **45**, 25–47 (2016).

[5] Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).

[6] Carey, L. B., Van Dijk, D., Sloot, P. M., Kaandorp, J. A. & Segal, E. Promoter sequence determines the relationship between expression level and noise. *PLoS Biol.* **11**, e1001528 (2013).

[7] Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346**, 1533–1536 (2014).

[8] Zong, C., So, L.-h., Sepúlveda, L. A., Skinner, S. O. & Golding, I. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. *Mol. Syst. Biol.* **6**, 440 (2010).

[9] Jiao, F. & Tang, M. Quantification of transcription noises impact on cell fate commitment with digital resolutions. *Bioinformatics* **38**, 3062–3069 (2022).

[10] Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

[11] Dar, R. D., Hosmane, N. N., Arkin, M. R., Siliciano, R. F. & Weinberger, L. S. Screening for noise in gene expression identifies drug synergies. *Science* **344**, 1392–1396 (2014).

[12] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).

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

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

[15] Dey, S. S., Foley, J. E., Limsirichai, P., Schaffer, D. V. & Arkin, A. P. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Mol. Syst. Biol.* **11**, 806 (2015).

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

[17] Larsson, A. J. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

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

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

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

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

[22] Zhou, T. & Zhang, J. Analytical results for a multistate gene model. *SIAM J. Appl. Math.* **72**, 789–818 (2012).

[23] Chen, J. & Jiao, F. A novel approach for calculating exact forms of mRNA distribution in single-cell measurements. *Mathematics* **10**, 27 (2021).

[24] Jia, C. & Li, Y. Analytical time-dependent distributions for gene expression models with complex promoter switching mechanisms. *bioRxiv* (2022).

[25] Munsky, B., Neuert, G. & Van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).

[26] Sanchez, A. & Golding, I. Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**, 1188–1193 (2013).

[27] Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

[28] Kærn, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).

[29] Jia, C., Qian, M., Kang, Y. & Jiang, D. Modeling stochastic phenotype switching and bet-hedging in bacteria: stochastic nonlinear dynamics and critical state identification. *Quant. Biol.* **2**, 110–125 (2014).

[30] Ackermann, M. A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol.* **13**, 497–508 (2015).

[31] Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS*

Biol. **7**, e1000149 (2009).

[32] Jiao, F., Sun, Q., Tang, M., Yu, J. & Zheng, B. Distribution modes and their corresponding parameter regions in stochastic gene transcription. *SIAM J. Appl. Math.* **75**, 2396–2420 (2015).

[33] Jia, C., Zhang, M. Q. & Qian, H. Analytic theory of stochastic oscillations in single-cell gene expression. *arXiv preprint arXiv:1909.09769* (2019).

[34] Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337 (2011).

[35] Molina, N. *et al.* Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Natl. Acad. Sci. USA* **110**, 20563–20568 (2013).

[36] Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. USA* **109**, 17454–17459 (2012).

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

[38] Grima, R., Schmidt, D. & Newman, T. Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).

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

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

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

[42] Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).

[43] Jia, C., Xie, P., Chen, M. & Zhang, M. Q. Stochastic fluctuations can reveal the feedback signs of gene regulatory networks at the single-molecule level. *Sci. Rep.* **7**, 1–9 (2017).

[44] Larson, D. R. What do expression dynamics tell us about the mechanism of transcription? *Current opinion in genetics & development* **21**, 591–599 (2011).

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

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

[47] Oliveira, S. M. *et al.* Temperature-dependent model of multi-step transcription initiation in Escherichia coli based on live single-cell measurements. *PLoS Comput. Biol.* **12**, e1005174 (2016).

[48] Zimmer, C., Häkkinen, A. & Ribeiro, A. S. Estimation of kinetic parameters of transcription from temporal single-RNA measurements. *Math. Biosci.* **271**, 146–153 (2016).

[49] Wang, Y., Liu, F., Li, J. & Wang, W. Reconciling the concurrent fast and slow cycling of proteins on gene promoters. *J. R. Soc. Interface* **11**, 20140253 (2014).

[50] Zoller, B., Nicolas, D., Molina, N. & Naef, F. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Mol. Syst. Biol.* **11**, 823 (2015).

[51] De Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12**, 833–845 (2011).

[52] Jiao, F. & Zhu, C. Regulation of gene activation by competitive cross talking pathways. *Biophys. J.* **119**, 1204–1214 (2020).

[53] Chen, L., Lin, G. & Jiao, F. Using average transcription level to understand the regulation of stochastic gene activation. *R. Soc. Open Sci.* **9**, 211757 (2022).

[54] Hao, S. & Baltimore, D. The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat. Immunol.* **10**, 281–288 (2009).

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

[56] Sun, Q., Cai, Z. & Zhu, C. A novel dynamical regulation of mRNA distribution by cross-talking pathways. *Mathematics* **10**, 1515 (2022).

[57] Jiao, F., Ren, J. & Yu, J. Analytical formula and dynamic profile of mRNA distribution. *Discrete and Continuous Dynamical Systems-B* **25**, 241–257 (2019).

[58] Jia, C. & Grima, R. Dynamical phase diagram of an auto-regulating gene in fast switching conditions. *J. Chem. Phys.* **152**, 174110 (2020).

[59] Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20**, 1726–1738 (2001).

[60] Nixon, C. C. *et al.* Systemic HIV and SIV latency reversal via non-canonical NF-κB signalling in vivo. *Nature* **578**, 160–165 (2020).

[61] Jiao, F., Sun, Q., Lin, G. & Yu, J. Distribution profiles in gene transcription activated by the cross-talking pathway. *Discrete & Continuous Dynamical Systems-B* **24**, 2799 (2019).

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

[63] Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).

[64] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).

[65] Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM J. Appl. Math.* **80**, 1336–1355 (2020).

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

[67] Munsky, B., Fox, Z. & Neuert, G. Integrating single-molecule experiments and discrete stochastic models to understand heterogeneous gene transcription dynamics. *Methods* **85**, 12–21 (2015).

[68] Cao, Z. & Grima, R. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *J. R. Soc. Interface* **16**, 20180967 (2019).

[69] Jia, C., Qian, H., Chen, M. & Zhang, M. Q. Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks. *J. Chem. Phys.* **148**, 095102 (2018).

[70] Jia, C., Qian, M. & Jiang, D. Overshoot in biological systems modelled by Markov chains: a non-equilibrium dynamic phenomenon. *IET Syst. Biol.* **8**, 138–145 (2014).

[71] Ma, W., Trusina, A., El-Samad, H., Lim, W. A. & Tang, C. Defining network topologies that can achieve biochemical adaptation. *Cell* **138**, 760–773 (2009).

[72] Shimoga, V., White, J. T., Li, Y., Sontag, E. & Bleris, L. Synthetic mammalian transgene negative autoregulation. *Mol. Syst. Biol.* **9**, 670 (2013).

[73] Bintu, L. *et al.* Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* **15**, 125–135 (2005).

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

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

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

[77] Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 12795–12800 (2002).

[78] Bertaux, F., Marguerat, S. & Shahrezaei, V. Division rate, cell size and proteome allocation: impact on gene expression noise and implications for the dynamics of genetic circuits. *R. Soc. Open Sci.* **5**, 172234 (2018).

[79] Jia, C. & Grima, R. Frequency domain analysis of fluctuations of mRNA and protein copy numbers within a cell lineage: theory and experimental validation. *Phys. Rev. X* **11**, 021032 (2021).

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

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

[82] Jia, C. & Grima, R. Coupling gene expression dynamics to cell size dynamics and cell cycle events: Exact and approximate solutions of the extended telegraph model. *Iscience* **26**, 105746 (2023).

[83] Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proc. Natl. Acad. Sci. USA* **111**, 6994–6999 (2014).

[84] Hilfinger, A., Norman, T. M. & Paulsson, J. Exploiting natural fluctuations to identify kinetic mechanisms in sparsely characterized systems. *Cell Syst.* **2**, 251–259 (2016).