# Molecular-memory-driven phenotypic switching in a genetic toggle switch without cooperative binding

Baohua Qiu,[1] Tianshou Zhou,[1,2] and Jiajun Zhang[1,2,*]

[1] *School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China*
[2] *Key Laboratory of Computational Mathematics, Guangdong Province, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China*

(Received 9 October 2019; accepted 17 January 2020; published 14 February 2020)

A genetic toggle switch would involve multistep reaction processes (e.g., complex promoter activation), creating memories between individual reaction events. Revealing the effect of this molecular memory is important for understanding intracellular processes such as cellular decision making. We propose a generalized genetic toggle switch model and use a generalized chemical master equation theory to account for the memory effect. Interestingly, we find that molecular memory can induce bimodality in this memory system although the corresponding memoryless counterpart is not bimodal. This finding implies a plausible alternative mechanism for phenotypic switching that is driven by molecular memory rather than by ultrasensitivity or cooperative binding as shown in previous studies. We also find that unbalanced memories arising from the processes by which mutually inhibiting transcription factors are produced can give rise to asymmetric bimodality without changing the positions of two peaks in the bimodal protein distribution. Given the prevalence of molecular memory in gene regulation, our findings would provide insights into cell fate decisions in growth and development.

DOI: 10.1103/PhysRevE.101.022409

## I. INTRODUCTION

Phenotypic switching, which often relies on bimodal distributions of gene expression levels, has been observed in bacteria [1], yeast [2], and cancer cells [3] with single-cell experimental methods. It has been argued that stochastic transition in gene activity is a cause of phenotypic diversity in a population of genetically identical cells [1,4] and is critical for the cell population survivial in a fluctuating environment [5]. Closely related to phenotypic switching, bimodality (i.e., two peak modes of a distribution) is usually associated with different physiological states of a living system, such as different stem-cell fates, different disease states, and cancer subtypes [1,3,6,7]. Therefore, revealing the mechanism of phenotypic switching or bimodality is of significance.

Several mechanisms underlying bimodality have been identified to date. One common belief is that bimodality is directly related to deterministic bistability, i.e., two deterministic stable steady states of a system in the absence of noise [8]. Common examples of bimodality include lactoseoperon of *Escherichia coli* [9,10], $\lambda$ phage lysis or lysogeny circuit [11], competence development for genetic transformation in Bacillus subtilis [12]. These systems have been studied both experimentally and theoretically as well as in the context of synthetic biology, and it has been shown that bistability and switching are two important properties of gene regulatory networks [8,13,14]. The second mechanism is noise-induced bimodality, i.e., the noise can induce a bimodal response that does not occur in the deterministic case [15–19]. For instance, using a synthetic system in budding yeast, To *et al.* found that positive feedback involving a promoter with multiple transcription factor (TF) binding sites can induce a steady-state bimodal response without cooperative bindings of the TFs [16]. The third mechanism is noise filter-induced bimodality, which results from the fact that a nonlinear noise filter characterized by a Hill function can transform a unimodal distribution into a bimodal one of transcription rates in a cell population [20]. For example, Ochab-Marcinek *et al.* found that a unimodal distribution of TFs over a cell population can generate a bimodal steady-state output without cooperative binding of the TFs [20].

We note that the modeling of the reaction systems corresponding to the above mechanisms is based on the Markovian assumption, that is, the stochastic motion of the reactants is assumed to be uninfluenced by previous states and only by the current state. This memoryless property implies that the reaction kinetics can be described as a Poisson process, which is characterized by an exponential distribution of the interevent time between consecutive reaction events [21–23]. However, gene expression is a complex stochastic process and in particular, epigenetic regulation, gene activation, and transcription would involve numerous chemical events: from repressors falling off DNA to RNA polymerase elongating nascent transcripts, and to protein translation. These subprocesses can lead to nonexponential time intervals between transcription windows or during protein synthesis even if every single elementary reaction involved is rate limiting [4,24,25]. In other words, a multistep reaction process can create molecular memory between single reaction events, and this case cannot be described as a Markovian process. As a matter of fact, such molecular memory can significantly affect reaction kinetics and gene expression levels. For example, a nonexponential waiting time distribution between the arrival

[*] zhjiajun@mail.sysu.edu.cn

of mRNA bursts can amplify gene expression noise [26,27], and the memory resulting from time delay can cause a system of stochastic gene expression to oscillate, although its deterministic counterpart does not oscillate under the same condition of parameters [28]. In addition, molecular memory can increase the average residence time near a stable state of a bistable gene system [29]. In spite of these facts on memory effects, many questions remain unsolved, e.g., how molecular memory contributes to bimodality and whether it can give rise to the bimodality in a biochemical network without ultrasensitivity or cooperative interactions? In order to address these questions, we will introduce a generalized toggle switch model (gTSM) without cooperative bindings, where two TFs mutually repress each other [18].

For Markovian reaction systems or networks such as genetic toggle switch models, there have been many methods of modeling and analysis. On the modeling side, there have been different approaches to model biochemical systems where molecular memory would implicitly exist [25,28,30–32]. The most straightforward way is to model these systems with both the total biochemical reactions and the known parameter values if available. But it is difficult to identify all rate limiting elementary reactions and measure/estimate parameter values. Even if all model details are given, the computational complexity would also hamper the revealing of the memory effect [25–28]. An alternative way is to reduce dimensionality while keeping the system tractable without losing its important dynamical properties. For example, an explicit nonexponential waiting time or delay is used to model a multistep reaction process and this is formally correct in the limit of many steps. Some successful examples include queuing models [25,30,33–35], delay models [36–38], and continuous time random walk (CTRW) models [39–42]. On the analysis side, there have been rather few theories for non-Markovian processes so far. This is because reliable mathematical tools and machineries in the traditional Markovian theory cannot be directly translated into those for non-Markovian reaction systems.

In this paper, we adopt a CTRW framework to model and analyze the non-Markovian gTSM. First, we introduce general (exponential or nonexponential) waiting-time distributions for the reactions involved in the reaction system. Then, we establish a stationary generalized master equation for this system, which explicitly captures the effect of internal or external variability in the waiting times between reaction events on the probabilistic behavior of the system [32]. This framework lays a solid theoretical basis and provides an effective mathematical tool for studying various genetic toggle switch models with molecular memory [24,25,32,42]. In addition, based on the framework, we find interesting phenomena due to the effect of molecular memory, e.g., molecular memory can induce phenotypic switching.

The rest of this paper is organized as follows. In Sec. II, we first explain how a complex protein synthesis process creates molecular memory, and introduce this memory to the gTSM without the cooperative binding of TFs. Then, we develop a stationary generalized master equation and a stationary reaction rate equation to study the dynamics of the gTSM, where the former actually converts a non-Markovian problem into a Markovian one whereas the later can help us predict the system's the macroscopic behavior [or average behavior] underlying the memory effect. In Sec. III, we present main results on how molecular memory rather than cooperative binding drives phenotypic switching. In Sec. IV, we conclude this paper by discussing our results and their applicability.

## II. BACKGROUND AND MODELS

### A. The complex process of gene expression can create molecular memory

Generally, each birth or death of a macromolecule (e.g., protein) could involve several intermediate reaction steps, thus creating a memory between individual reaction events. For example, the process of gene promoter activation can create narrowly distributed gestation periods between transcription windows, and this implies a multistep process, which stems from the fact that the chromatin template accumulates over time until the gene promoter becomes active [25]. Some studies have shown that the arrival time of messenger RNA (mRNA) or protein bursts is nonexponentially distributed [25,27]. In addition, transcription delay may lead to the delayed response time of a gene system [28,29]. These examples indicate that molecular memory exists extensively in gene expression.

Molecular memory can be characterized by nonexponential waiting-time distribution such as $\gamma$ or Weibull distribution [25]. Here we take protein synthesis as an example to explain the biophysical foundation of $\gamma$ waiting-time distribution. As mentioned above, the synthesis of a protein involves a multistep process where each single-step process is assumed as an elementary reaction. Note that although the waiting time for each single-step reaction follows an exponential distribution, the waiting time for the entire multistep reaction process in general does not follow an exponential distribution but follows a $\gamma$ distribution [see Fig. 1(a)]. If this multistep process is regarded as the composition of $\boldsymbol{n}$ identical subprocesses where the waiting time for each single-step reaction obeys an exponential distribution characterized by parameter $\mu$, then the waiting time distribution [denoted by $\psi(t)$] for the entire reaction process is the $\boldsymbol{n}$-fold convolution of the exponential distribution. Mathematically, this can be described as [24,43]

$$\psi(t) = \underbrace{[\mu e^{-\mu t} * \mu e^{-\mu t} * \ldots]}_{n \text{ times}} = \frac{t^{n-1} e^{-\mu t} \mu^n}{(n-1)!}, \qquad (1)$$

where $*$ denotes convolution and $\boldsymbol{n}$ is a shape parameter. This distribution is also known as the Erlang distribution (a special case of $\gamma$ distribution, i.e., $\boldsymbol{n}$ is an integer). Note that we only consider unbranched multistep processes in this example. If a branched multistep process (i.e., in the multiple process analyzed above, some single reaction itself is a multistep reaction process) is introduced, the waiting-time distribution $\psi(t)$ in Eq. (1) is different and could become more complex. Moreover, the branched process may change the shape of the waiting-time distribution for the entire process. Nevertheless, the resulting distribution can be well approximated by a $\gamma$ distribution with the same mean and variance [34]. In addition, a large number of experimental measurements on waiting times offer another possibility for the versatility of $\gamma$ distributions

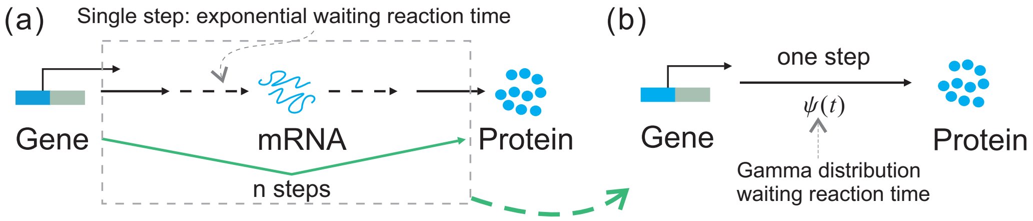

FIG. 1. Schematic diagram that maps a multistep reaction process into a single-step reaction process. (a) A multistep process for protein synthesis, where the waiting time for each single-step reaction is exponentially distributed. (b) One-step reaction for protein synthesis, where the reaction waiting time is assumed to follow a $\gamma$ distribution characterizing molecular memory that is introduced to simplify the modeling of the multistep reaction process.

[43]. Therefore, we will focus on multistep processes with $\gamma$ waiting-times distribution in the following analysis.

### B. A generalized toggle switch model

A bistable genetic toggle switch is often used to explain cell fate differentiation and decision making, e.g., lactoseoperon of *E. coli* [9,10], λ phage lysis or lysogeny circuit [11], and competence development for genetic transformation in Bacillus subtilis [12]. The model of a synthetic genetic toggle switch constructed in *E. coli* [44] provides a theoretical basis for analyzing bistability and state switching. This bistable system consists of two genes, *lacI* and *tetR*, which mutually repress each other via promoter binding [44]. The protein product of one gene (*lacI* or *tetR*) first binds to the promoter of the other gene as a TF and then represses its output. Subsequently, this protein activates its lineage-determining downstream targets in each differentiating cell. Indeed, some protein molecules themselves can also form a polymer complex to produce protein homomultimers, which in turn bind to a promoter site and inhibit the gene expression, thus carrying out gene regulation. To better explain the complex dynamics of mutually repressing genes *lacI* and *tetR*, we construct a synthetic genetic toggle switch model, which is schematically depicted in Fig. 2(a).

The genetic model shown in Fig. 2(a) assumes that two genes $A$ and $B$ mutually repress each other via promoter binding (possibly via protein homomultimer binding). Gene activation and transcription would involve numerous chemical reaction events, e.g., those from repressors falling off DNA to RNA polymerase elongating nascent transcripts, and even to protein translation, as well as protein polymerization, and other regulated components involved in gene expression such as repressors, TFs and mediators. All these subprocesses can affect the final expression level of a gene. Each process including the protein synthesis and the production of protein homomultimers that bind the promoter site can be viewed as a multistep process and described by a sequence of chemical reactions. If all biological details including reactions and the corresponding reaction rate parameter values are available, the involved reaction processes can be described as Markovian ones, implying that all the waiting-time distributions between inter-reaction events are exponential. But unfortunately, it is difficult or even impossible to access the complete information for this Markovian modeling due to some unobservable variables. Nevertheless, we can alternatively integrate a sequence of reactions with exponential waiting times into a single reaction with non-exponential waiting time or with molecular memory. This treatment in generally does not lose the essential property of reaction kinetics. Therefore, we introduce a stochastic model of the genetic toggle switch with molecular memory, which consists of the following four reactions [referring to

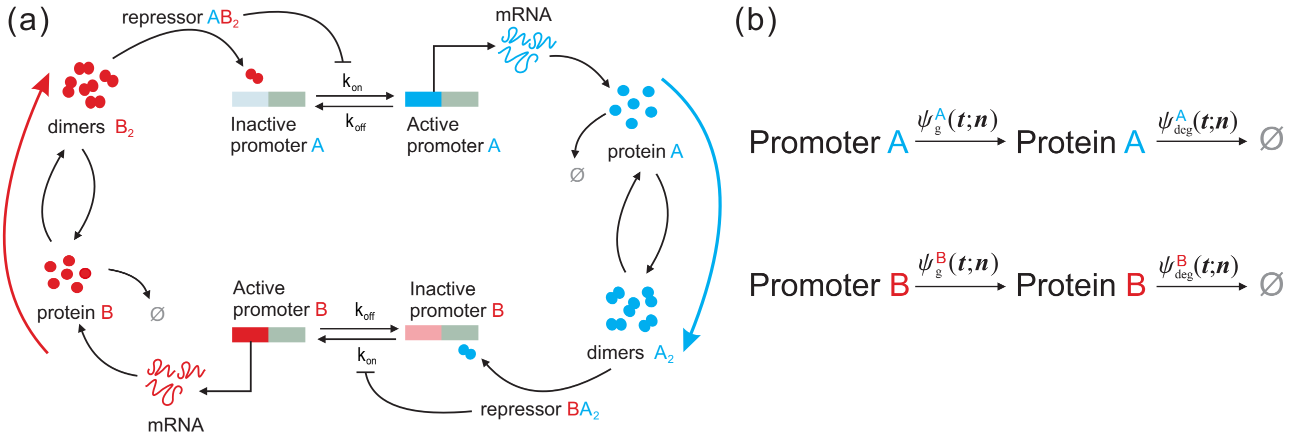

FIG. 2. (a) Schematic of a genetic toggle switch model, where gene $A(B)$ is transcribed into mRNA and further translated into protein $A(B)$. Dimer $A_2$, formed by protein polymerization, binds to promoter $B$ and inhibits the gene $B$ activity, thus repressing the transcription of this gene. Similarly, dimer $B_2$ binds to promoter A and represses the expression of gene $A$. (b) A generalized genetic toggle switch model corresponding to (a), which consists of four chemical reactions involving protein generation and degradation for which the waiting-time distributions are characterized by $\gamma$ distribution $\psi_g^X(t;\boldsymbol{n})$ and exponential distribution $\psi_{\text{deg}}^X(t;\boldsymbol{n})$ $(X = A, B)$, respectively.

Fig. 2(b)]:

$$\begin{aligned}
&\text{Gene-}A \xrightarrow{\psi_g^A(t;\boldsymbol{n})} \text{Gene-}A + \text{Protein-}A,\\
&\text{Protein-}A \xrightarrow{\psi_{\text{deg}}^A(t;\boldsymbol{n})} \varnothing,\\
&\text{Gene-}B \xrightarrow{\psi_g^B(t;\boldsymbol{n})} \text{Gene-}B + \text{Protein-}B,\\
&\text{Protein-}B \xrightarrow{\psi_{\text{deg}}^B(t;\boldsymbol{n})} \varnothing,
\end{aligned} \tag{2}$$

where $\psi_g^X(t;\boldsymbol{n})$ is the probability density function of the reaction waiting time for the synthesis of protein $A$ or $B$, and $\psi_{\text{deg}}^X(t;\boldsymbol{n})$ is that for the degradation of protein $A$ or $B$. Here $n{=}(n_A, n_B)^T$ represents the system's state vector at time $t$ with $n_A, n_B$ being the molecule numbers of reactive proteins $A$ and $B$ at time $t$, and T represents transpose. In the following, we assume that $\psi_g^X(t;\boldsymbol{n})$ obeys a $\gamma$ distribution with parameter $k_X$ and rate function $\lambda_g^X(\boldsymbol{n})$ (an inverse scale parameter), which is a nonexponential waiting-time distribution if $k_X \neq 1$. In addition, we assume that $\psi_{\text{deg}}^X(t;\boldsymbol{n})$ follows an exponential distribution. Thus, the four reaction waiting-time distributions are given by

$$\begin{aligned}
\psi_g^A(t;\boldsymbol{n}) &= \frac{(\lambda_g^A(\boldsymbol{n}))^{k_A}}{\Gamma(k_A)} t^{k_A-1} e^{-\lambda_g^A(\boldsymbol{n})t},\\
\psi_g^B(t;\boldsymbol{n}) &= \frac{(\lambda_g^B(\boldsymbol{n}))^{k_B}}{\Gamma(k_B)} t^{k_B-1} e^{-\lambda_g^B(\boldsymbol{n})t},
\end{aligned} \tag{3a}$$

$$\begin{aligned}
\psi_{\text{deg}}^A(t;\boldsymbol{n}) &= \lambda_{\text{deg}}^A n_A e^{-\lambda_{\text{deg}}^A n_A t},\\
\psi_{\text{deg}}^B(t;\boldsymbol{n}) &= \lambda_{\text{deg}}^B n_B e^{-\lambda_{\text{deg}}^B n_B t},
\end{aligned} \tag{3b}$$

where $\Gamma(k_A), \Gamma(k_B)$ are $\gamma$ functions with $k_A > 0, k_B > 0$. The rate function $\lambda_g^X(\boldsymbol{n})$ $(X = A, B)$ is described by a Hill function that explains the mutual repression characteristic of proteins $A$ and $B$, that is, $\lambda_g^A(\boldsymbol{n}) = g_A/(1 + r_A n_B^H)$, $\lambda_g^B(\boldsymbol{n}) = g_B/(1 + r_B n_A^H)$, where $g_A, g_B$ represent the maximal production rates of proteins $A$ and $B$ respectively, $\lambda_{\text{deg}}^A, \lambda_{\text{deg}}^B$ are the degradation rates of proteins $A$ and $B$, respectively, $r_A, r_B$ are repression strengths representing the ratio of the promoter binding rate to the dissociation rate of protein $A$ $(B)$, and $H$ is a Hill coefficient. Note that $H = 1$ implies the noncooperative binding of a single protein, that is, a protein binding to any site does not affect the others even if several binding sites exist, while $H > 1$ implies the cooperative binding of two or more of the same proteins, that is, the first protein molecule bound can facilitate the binding of the second one [45]. We term the gene model shown in Fig. 2(b) as a gTSM. Note that shape parameters $k_A$,$k_B$ can determine not only the shapes of the waiting-time distributions but also the relationship between the gTSM and the conventional TSM, e.g., the former reduces to the latter if two $\gamma$ distributions are exponential ones (i.e., if $k_A = 1$,$k_B = 1$). We emphasize that the gTSM with $H = 1$ is nothing but a genetic toggle switch model without cooperative binding.

### C. A mathematical model

According to the chemical CTRW theory [42] and the stationary generalized chemical master equation (sgCME) theory [32], we can firstly derive the sgCME for the gTSM (see Appendix A for details). If the stationary joint probability distribution function of the protein molecule number exists and is denoted by $p(\boldsymbol{n}) = p(n_A, n_B)$, the sgCME takes the form

$$\begin{aligned}
&\big(\mathbb{E}_A^{-1} - 1\big)\big(K_g^A(n_A, n_B)p(n_A, n_B)\big)\\
&\quad + (\mathbb{E}_A - 1)\big(K_{\text{deg}}^A(n_A, n_B)p(n_A, n_B)\big)\\
&\quad + \big(\mathbb{E}_B^{-1} - 1\big)\big(K_g^B(n_A, n_B)p(n_A, n_B)\big)\\
&\quad + (\mathbb{E}_B - 1)\big(K_{\text{deg}}^B(n_A, n_B)p(n_A, n_B)\big) = 0,
\end{aligned} \tag{4}$$

where $\mathbb{E}$ is a step operator, defined by its effect on an arbitrary function $f(n_A, n_B)$ as $\mathbb{E}_A^{-1} f(n_A, n_B) = f(n_A - 1, n_B)$, $\mathbb{E}_A f(n_A, n_B) = f(n_A + 1, n_B)$, and functions $K_g^X(n_A, n_B)$ and $K_{\text{deg}}^X(n_A, n_B)$ represent the effective transition propensity functions for the generation and degradation reaction of proteins $X$ $(X = A, B)$, respectively, which depend only on state $(n_A, n_B)$ and is independent of the prior history. Importantly, we can show that $K_g^X(n_A, n_B)$ and $K_{\text{deg}}^X(n_A, n_B)$ are explicitly expressed by the given waiting-time distributions, that is,

$$\begin{aligned}
K_i^X(\boldsymbol{n}) &= \frac{\int_0^{+\infty} \psi_i^X(t;\boldsymbol{n}) \prod_{\{Y,j\}\neq\{X,i\}} \big[1 - \Psi_j^Y(t;\boldsymbol{n})\big]dt}{\int_0^{+\infty} \prod_{\mathrm{Y}\in G, j\in R} \big[1 - \Psi_j^Y(t;\boldsymbol{n})\big]dt},\\
X &\in G = \{A, B\}, i \in R = \{g, \text{deg}\},
\end{aligned} \tag{5}$$

where $\Psi_i^X(t;\boldsymbol{n})$ is the cumulative distribution function corresponding to the waiting-time distribution $\psi_i^X(t;\boldsymbol{n})$, i.e., $\Psi_i^X(t;\boldsymbol{n}) = \int_0^t \psi_i^X(\tau';\boldsymbol{n})d\tau'$. By substituting Eq. (3) into Eq. (5), we obtain

$$K_g^A(\boldsymbol{n}) = \frac{\lambda_g^A(\boldsymbol{n}) \sum_{\beta=0}^{k_B-1} \binom{\beta + k_A - 1}{\beta} (\omega_B(\boldsymbol{n}))^\beta (\omega_A(\boldsymbol{n}))^{k_A-1}}{\sum_{\alpha=0}^{k_A-1} \sum_{\beta=0}^{k_B-1} \binom{\alpha+\beta}{\alpha} (\omega_A(\boldsymbol{n}))^\alpha (\omega_B(\boldsymbol{n}))^\beta}, \tag{6a}$$

$$K_g^B(\boldsymbol{n}) = \frac{\lambda_g^B(\boldsymbol{n}) \sum_{\alpha=0}^{k_A-1} \binom{\alpha + k_B - 1}{\alpha} (\omega_A(\boldsymbol{n}))^\alpha (\omega_B(\boldsymbol{n}))^{k_B-1}}{\sum_{\alpha=0}^{k_A-1} \sum_{\beta=0}^{k_B-1} \binom{\alpha+\beta}{\alpha} (\omega_A(\boldsymbol{n}))^\alpha (\omega_B(\boldsymbol{n}))^\beta}, \tag{6b}$$

where $\omega_A(\boldsymbol{n}) = \lambda_g^A(\boldsymbol{n})/\Sigma(\boldsymbol{n})$, $\omega_B(\boldsymbol{n}) = \lambda_g^B(\boldsymbol{n})/\Sigma(\boldsymbol{n})$ with $\Sigma(\boldsymbol{n}) = \lambda_g^A(\boldsymbol{n}) + \lambda_g^B(\boldsymbol{n}) + \lambda_{\text{deg}}^A n_A + \lambda_{\text{deg}}^B n_B$. Note that if $\psi_{\text{deg}}^X(t;\boldsymbol{n})$ follows an exponential distribution, the effective reaction propensity functions of proteins $A$ and $B$ are $K_{\text{deg}}^A(\boldsymbol{n}) = \lambda_{\text{deg}}^A n_A$ and $K_{\text{deg}}^B(\boldsymbol{n}) = \lambda_{\text{deg}}^B n_B$, respectively. And the corresponding sgCME reduces to the common stationary CME if all reactions are exponential distributions. Interestingly, the sgCME [i.e., Eq. (4)] successfully converts the original non-Markovian problem into a Markovian one where the effective reaction propensity functions [i.e., Eq. (6)] explicitly encode non-Markovian effects. Thus, analysis and numerical methods for solving Eq. (4) are routine. Moreover, one can obtain analytical solutions in some special cases [32]. In most cases, however, we need to resort to numerical methods. Here we propose a stationary generalized finite state projection algorithm, which can be used to calculate an approximate joint probability distribution of proteins $A$ and

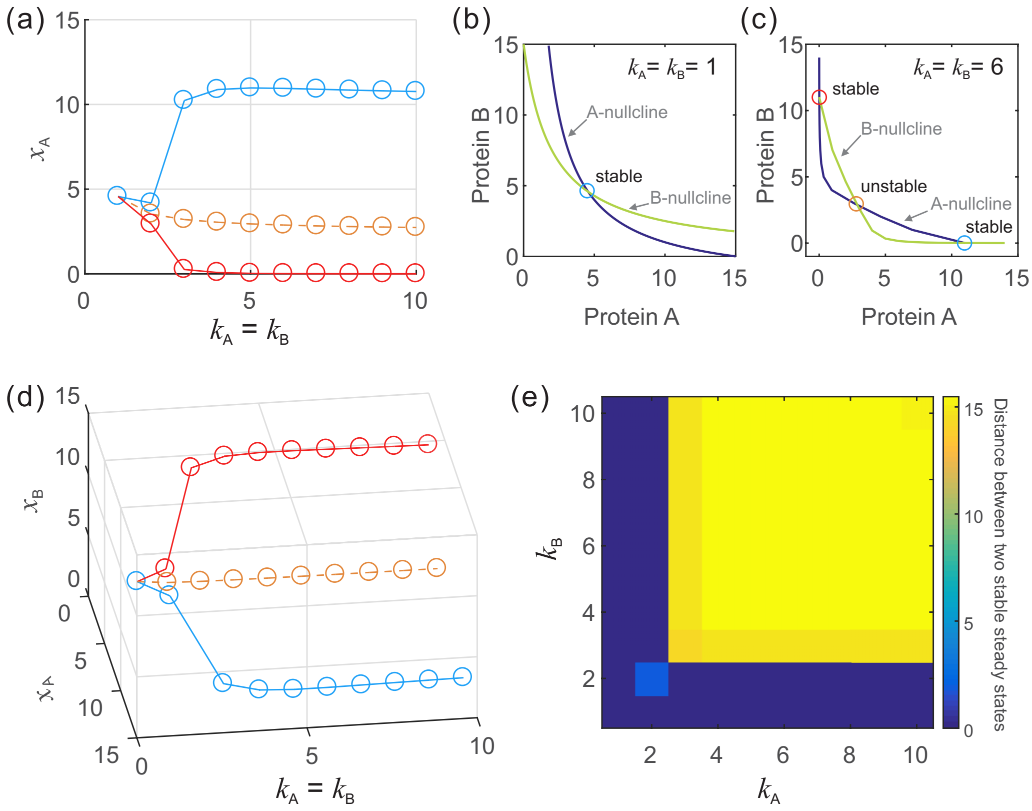

FIG. 3. It is shown how bistability is generated in the sgTSM. (a) A bifurcation diagram, where protein concentration $x_A$ is taken as a function of $k_A(= k_B)$. (b),(c) Nullclines for protein concentrations $x_A$ and $x_B$ under different parameter conditions. In the state space, there is only one stable steady state in (b), and there are two stable steady states and one unstable steady state in (c). (d) A three-dimensional bifurcation diagram corresponding to (a), where $x_A$ and $x_B$ are taken as functions of $k_A(= k_B)$. (e) A heat map for the Euclidean distance between two stable steady states as a function of $k_A$ and $k_B$. In (a)–(e), the other parameter values are set as $H = 1$, $r_A = r_B = 0.5$, $\lambda_{\text{deg}}^A = \lambda_{\text{deg}}^B = 1$, $g_A = 15k_A$, $g_B = 15k_B$. Note that $H = 1$ correspond to a genetic toggle switch without cooperative binding.

$B$ by solving a truncated version of stochastic process (see Appendix B for details).

Recall that for a given TSM, the equations governing the dynamics of the first-order raw moments of the state variables have been derived from the common CME [14]. These equations can help us to roughly analyze the macroscopic behavior of the TSM before we further perform stochastic analysis. Similarly, the reaction rate equations for the gTSM can also be used to analyze the corresponding macroscopic behavior. Specifically, the stationary generalized reaction rate equation (sgRRE) is given by

$$\boldsymbol{S}\boldsymbol{K}(x_A, x_B) = 0, \tag{7}$$

where $x_A, x_B$ represent the concentrations of proteins $A$ and $B$, $\boldsymbol{S} = \left[\begin{smallmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{smallmatrix}\right]$ is a stoichiometric matrix, and $\boldsymbol{K}(x_A, x_B) = (K_g^A(x_A, x_B), K_{\text{deg}}^A(x_A, x_B), K_g^B(x_A, x_B), K_{\text{deg}}^B(x_A, x_B))^T$ is a four-dimensional vector. Note that Eq. (7) is equivalent to

$$K_g^A(x_A, x_B) - K_{\text{deg}}^A(x_A, x_B) = 0$$
$$K_g^B(x_A, x_B) - K_{\text{deg}}^B(x_A, x_B) = 0, \tag{8}$$

where the effective transition propensity function $\boldsymbol{K}_i^X(x_A, x_B)$ is calculated according to Eq. (5).

## III. RESULTS

### A. Molecular memory can induce bistability

We start by numerically investigating how molecular memory affects macroscopic behaviors of proteins $A$ and $B$ in the gTSM. Numerical results for the steady state of Eq. (8) are demonstrated in Fig. 3. In numerical simulations, we set $k_A/\lambda_{\text{deg}}^A$ and $k_B/\lambda_{\text{deg}}^B$ as constants so that the average waiting times remain unchanged, where $k_A$,$k_B$ are the corresponding shape parameters of $\gamma$ distributions $\psi_g^A(t;\boldsymbol{n})$ and $\psi_g^B(t;\boldsymbol{n})$ in Eq. (3a) respectively, while $\lambda_{\text{deg}}^A, \lambda_{\text{deg}}^B$ are the degradation rates of proteins $A$ and $B$ in Eq. (3b), respectively. Note that if $k_A = k_B = 1$, the system described by Eq. (2) is a common TSM, which corresponds to a Markovian process since both $\psi_g^X(t;\boldsymbol{n})$ and $\psi_{\text{deg}}^X(t;\boldsymbol{n})$, $X = (A, B)$, are exponential distributions. However, if $k_A \neq 1$ or/and $k_B \neq 1$, the TSM corresponds to a non-Markovian process with waiting-time distribution $\psi_g^A(t;\boldsymbol{n})$ or $\psi_g^B(t;\boldsymbol{n})$ being nonexponential.

From Figs. 3(a) and 3(d), we observe that there is only one stable steady state when $k_A = k_B = 1$ that corresponds to the common TSM without molecular memory, whereas there are two stable steady states and one unstable steady state when $k_A = k_B > 1$ that corresponds to a non-Markovian process with molecular memory. Figures 3(b) and 3(c) also verify

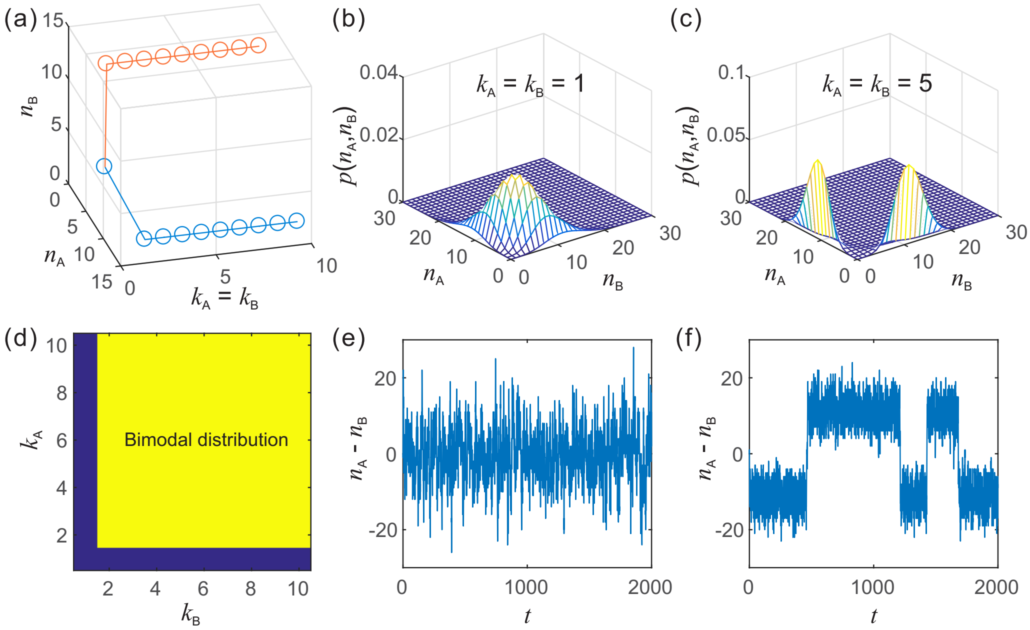

FIG. 4. (a) Global peak modes of distribution, i.e., the most probable protein numbers as a function of $k_A(= k_B)$. (b),(c) Joint probability distributions of proteins $A$ and $B$, obtained by a numerical algorithm given in Appendix A, where parameter values are set as $k_A = k_B = 1$ (b), $k_A = k_B = 5$ (c), respectively. (d) Classification of distributions by $k_A$ and $k_B$, where the yellow and blue regions correspond to bimodal and unimodal distributions, respectively. (e),(f) Time series of the difference between the levels of proteins $A$ and $B$ corresponding to (b) and (c), respectively, obtained by stochastic simulation using a modified Gillespie algorithm [32]. (b) and (e) correspond to exponential waiting times, where parameter values are set as $k_A = k_B = 1$, $g_A = g_B = 15$; (c) and (f) correspond to nonexponential waiting times, where parameter values are set as $k_A = k_B = 5$, $g_A = g_B = 75$. In (a)–(f), the other parameter values are set as $H = 1$, $\lambda_{\text{deg}}^A = \lambda_{\text{deg}}^B = 1$, $r_A = r_B = 0.5$, $g_A = 15k_A$, $g_B = 15k_B$.

these results due to the intersection of two nullclines under parameter conditions. This suggests that bistability exists in the sgRRE after molecular memory is introduced to the genetic switch model without cooperative binding. In order to obtain the region of parameter space for bistability, we numerically compute the Euclidean distance between two stable steady states (defined as a bistability index) in a large region of parameter pair $k_A$ and $k_B$. The results are shown as a heatmap in Fig. 3(e). From this subfigure, we observe that the distance between two stable steady states first has an increasing trend with increasing $k_A$ and $k_B$, and then remain nearly a constant with the further increase of these two parameters. We also observe that there is a very sharp boundary for $k_A = k_B \geqslant 3$. This is possibly because parameters $k_A$ and $k_B$ are set as integers, but if they are set as real numbers, this boundary will not be so sharp.

It should be noted that if $k_A = k_B = 1$ and $H = 1$, the gTSM can reduce to the common TSM without cooperative binding, mainly since $H = 1$ indicates the noncooperative binding of TFs. On the other hand, it has been theoretically proven that the cooperative binding is a necessary condition for the generation of bistability. The absence of the cooperative binding only yields a single stable steady state and implies that the two TFs coexist in our case [46]. Figure 3 indicates that after molecular memory is introduced into the TSM, the bistability can occur. This is a very interesting result, and implies that molecular memory can induce bistability in the gTSM.

### B. Molecular memory can induce bimodal distribution

In the above subsection, we have analyzed the deterministic counterpart of the gTSM, focusing on how bistability is generated by molecular memory. Here, we investigate the stochastic counterpart of the gSTM, focusing on how molecular memory induces bimodal protein distributions. To analyze memory effect on switching states due to transitions from a steady state to another, we calculate the joint probability distribution $p(n_A, n_B; t)$ with different parameter values of $k_A$ and $k_B$, and simulate switching dynamics. Figure 4(a) demonstrates the (global) peak modes of the stationary distribution $p(n_A, n_B)$, which correspond to the most probable number of proteins $n_B$ as a function of $k_A(= k_B)$. We observe that the distribution has only one peak if $k_A = k_B = 1$ that corresponds to the common TSM without molecular memory, but the distribution is bimodal otherwise. Note that $k_A = k_B \geqslant 2$ corresponds to non-Markovian processes with molecular memory, which are assumed to be characterized by nonexponential waiting-time distributions $\psi_g^A(t; \boldsymbol{n})$ and $\psi_g^B(t; \boldsymbol{n})$ given in Eq. (3a). These results indicate that molecular memory can induce bimodal distribution in the stochastic framework, and the distribution $p(n_A, n_B)$ is bimodal if $k_A = k_B \geqslant 2$. Further, if $k_A \neq k_B$, the distribution $p(n_A, n_B)$ in the gTSM also has the bimodal characteristic, seen in Fig. 4(d), which is similar to the case of $k_A = k_B$. From Fig. 4(b), we observe that the distribution $p(n_A, n_B)$ is unimodal for a symmetric parameter pair of $k_A = k_B = 1$, since the distribution has only one local maximum at the molecule number $(n_A, n_B) \approx (10, 10)$ of

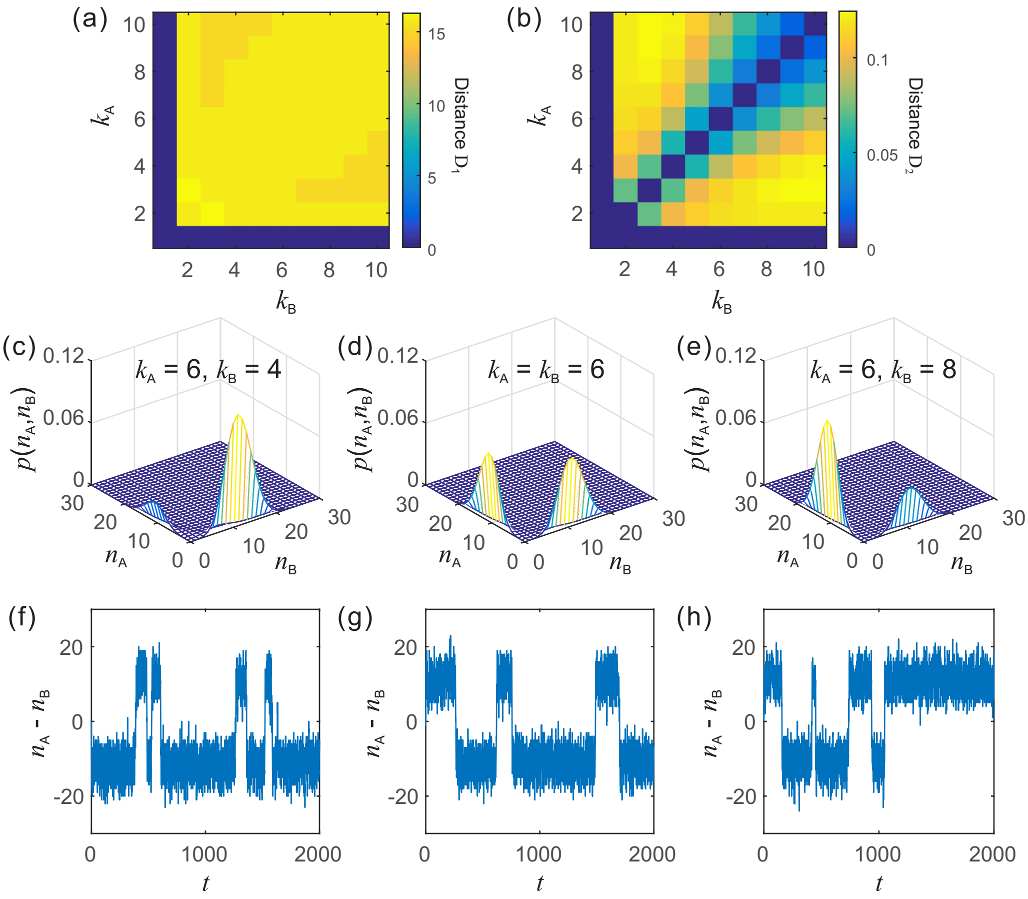

FIG. 5. (a) The distance between two probability peak points as a function of parameter pair $k_A$ and $k_B$. (b) The probability difference between double probability peak points as a function of parameter pair $k_A$ and $k_B$. (c)–(e) The stationary joint probability distributions of proteins $A$ and $B$, obtained by a numerical method (see Appendix). (f)–(h) Time series of the difference $n_A - n_B$ between the levels of proteins $A$ and $B$ corresponding to (c)–(e), obtained by stochastic simulation. In (a)–(h), all parameters are set as $H = 1$, $r_A = r_B = 0.5$, $\lambda_{\text{deg}}^A = \lambda_{\text{deg}}^B = 1$, $g_A = 15k_A$, $g_B = 15k_B$, where $k_A = 6$, $k_B = 4$ in (c),(f), $k_A = k_B = 6$ in (d),(g), and $k_A = 6, k_B = 8$ in (e),(h), respectively.

proteins $A$ and $B$ in this case. However, the distribution $p(n_A, n_B)$ exhibits two peaks $(n_A, n_B) \approx (10, 0)$ and $(n_A, n_B) \approx (0, 10)$ if $k_A = k_B = 5$ in Fig. 4(c). These two peak states correspond to the dominant numbers of proteins $A$ and $B$, respectively. In addition, Fig. 4(d) shows the regions in the $k_A - k_B$ plane, where switching behavior occurs. Clearly, there is a bimodal distribution region where $k_A \geqslant 2$ and $k_B \geqslant 2$ are simultaneously satisfied. In Fig. 4, parameters $k_A$ and $k_B$ are set as integers, but if they are nonintegers, we also obtain qualitatively similar results.

We now turn to the analysis of switching dynamics. Figures 4(e) and 4(f) show two representative time series of the difference $n_A - n_B$ for the gTSM. We observe that there are no obvious transition behaviors but there are only noise-induced fluctuations in this difference if $k_A = k_B = 1$ [Fig. 4(e)]. We can clearly see two switching states from Fig. 4(f), where one protein is strongly repressed compared to the other one. These states correspond to the probability maxima in Fig. 4(c). Thus, we conclude that bimodality can arise in stochastic cases without cooperativity, due to the effect of molecular memory.

### C. Unbalanced memory can induce asymmetric bimodality

In the above subsection, we have shown that molecular memory can lead to the bimodal joint distribution of proteins $A$ and $B$. Here, we further quantify bimodality by the distance between double peak points of the probability [Fig. 5(a)] and the probability difference between these peak points [Fig. 5(b)]. The distance index is defined as a simple Euclidean distance $D_1 = \sqrt{\left(n_A^{(1)} - n_A^{(2)}\right)^2 + \left(n_B^{(1)} - n_B^{(2)}\right)^2}$ (called the bimodality index) where $(n_A^{(1)}, n_B^{(1)})$ and $(n_A^{(2)}, n_B^{(2)})$ are two peak points, respectively. A higher value of $D_1$ indicates a larger gap between two gene phenotypic clusters. Figure 5(a) shows the bimodality index $D_1$ as a function of $k_A$ and $k_B$. We observe that $D_1$ is larger in some parameter regions of $k_A > 1$ and $k_B > 1$, which implies that unbalanced memories can affect $D_1$. The probability difference index is defined as a distance $D_2 = |P(n_A^{(1)}, n_B^{(1)}) - P(n_A^{(2)}, n_B^{(2)})|$ where $P(n_A^{(1)}, n_B^{(1)})$ and $P(n_A^{(2)}, n_B^{(2)})$ are the stationary joint probabilities of $n_A$ and $n_B$ at two peak points respectively. A higher value of $D_2$ indicates a larger imbalance between the number of two populations. Figure 5(b) shows the bimodality index $D_2$

as a function of $k_A$ and $k_B$. We find that under the condition of $k_A > 1$ and $k_B > 1$, the bigger the difference between $k_A$ and $k_B$, the bigger $D_2$.

In addition, we performed numerical calculations and stochastic simulations, obtaining the stationary probability distributions [Figs. 5(c)–5(e)] and tracking the protein status over time [Figs. 5(f)–5(h)] at different values of $k_A$ and $k_B$. All these results demonstrate that unbalanced memories can induce asymmetric bimodal distributions in the gTSM without cooperative bindings.

## IV. DISCUSSION

As a representative bistable example, the conversional genetic toggle switch model without molecular memory has been extensively studied with the aim to answer how bistability is generated and how intrinsic variability induces stochastic switching between stable states. On the other hand, molecular memory extensively exists in biomolecular interacting systems, but its effect remains elusive and even unexplored. Here we have studied a generalized genetic toggle switch model with molecular memory but without cooperative binding, using a generalized chemical master equation theory [32]. We have shown that molecular memory can give rise to nontrivial results. Specifically, molecular memory can induce bistability in the deterministic system of the genetic toggle switch without cooperative binding, whereas molecular memory cannot only induce the bimodal protein distribution but also adjust the symmetry of the two peaks of this probability in a stochastic framework. Our study indicates that molecular memory, which exists extensively in gene regulation, is an important factor impacting gene expression.

The above results could have important biological implications. First, introducing molecular memory into biological regulatory systems such as genetic toggle switch models would well explain cell fate differentiation and fate decision. Second, molecular memory can significantly contribute to genetic variability as shown in this paper, and is therefore an unneglectable source of noise in gene expression. Third, molecular memory can induce phenotypic switching in the gSTM without cooperative binding as shown above. Molecular memory can even adjust which part of the cell population survives in this environment since we have shown that the unbalanced memory can induce asymmetric bimodality of the protein distribution. Therefore, molecular memory can be taken as an effective strategy for a population of genetically identical cells that survive in a noisy environment.

It is worth pointing out that the mathematically tractable framework developed in this paper for a generalized genetic toggle switch with molecular memory characterized by nonexponential waiting-time distributions lays a solid theoretical foundation for analyzing stochastic dynamics of the underlying system. The similar framework can also be extended to other biochemical processes including other genetic toggle switches [19,47,48] such as exclusive switch, genetic switch with bound repressor degradation or protein-protein interactions. As such, one can expect discovery of new biological knowledge. In addition, the lifetimes of stable steady states and the optimal transition paths connecting these stable steady states are also important quantities underlying the memory effect [49,50]. The corresponding mechanism studies are ongoing in our laboratory, based on the mathematical framework developed here.

Finally, we point out that there are some other approaches or models that are used to model molecular memory, such as queuing model [34,35] and delay model [36–38]. A detailed comparison between these models is outside the scope of this article and will be therefore discussed elsewhere.

## ACKNOWLEDGMENTS

This work was supported by Grants No. 11931019, No. 11475273, No. 11775314, and No. 11631005 from the Natural Science Foundation of P. R. China and No. 201707010117 from the Science and Technology Program of Guangzhou, P. R. China.

## APPENDIX A: DERIVING THE sgCME FROM THE gTSM

In this Appendix, we derive a stationary CME from the gTSM based on the CTRW theory [32,42]. In contrast to the conventional TSM that models a Markovian process, the gTSM models a non-Markovian process (since complex gene expression involves a multistep process, creating a molecular memory between chemical reaction events).

In order to reduce complexity, we characterize a genetic toggle switch by the waiting time of chemical reactions, where the waiting time for protein synthesis is assumed to obey a $\gamma$ distribution while protein decay time is assumed to follow an exponential distribution. Thus, a simple genetic toggle switch model with molecular memory can be described by four chemical reactions given by Eq. (2) in the main text. The stoichiometric numbers of the state change for four reactions are $\boldsymbol{v}_g^A{=}(1,0)^T$ $\boldsymbol{v}_{\text{deg}}^A{=}(-1,0)^T$, $\boldsymbol{v}_g^B{=}(0,1)^T$, $\boldsymbol{v}_{\text{deg}}^B{=}(0,-1)^T$ respectively. Let $\boldsymbol{v}{=}(\boldsymbol{v}_g^A, \boldsymbol{v}_{\text{deg}}^A, \boldsymbol{v}_g^B, \boldsymbol{v}_{\text{deg}}^B)_{2\times4}$ represent a $2\times4$ stoichiometric matrix, where the $i$th column element is the state change for the $i$th reaction. Let $\psi_g^X(t;\boldsymbol{n})$ be the probability density function of the reaction waiting time for proteins $A$ and $B$ produced, and $\psi_{\text{deg}}^X(t;\boldsymbol{n})$ be protein degradations,$X = A, B$, which are given by Eq. (3) in the main text.

Next, we discuss the probability that the system is at state $\boldsymbol{n}$ at time $t$, denoted by $p(\boldsymbol{n};t)$. For convenience, we denote by symbol $\Omega \equiv \{\{X,i\}|X \in G = \{A,B\}, i \in R = \{g, \text{deg}\}\}$ all chemical reactions in the gTSM, i.e., $\Omega$ consists of four chemical reactions. Then, the stoichiometric matrix of the reaction system is $\boldsymbol{v} = (\boldsymbol{v}_i^X)_{2\times4}$, $\{X,i\} \in \Omega$. The corresponding cumulative distribution functions $\Psi_i^X(t;\boldsymbol{n})$ for four reaction waiting-time distributions $\psi_i^X(t;\boldsymbol{n})$, $\{X,i\} \in \Omega$, are, respectively, expressed by

$$\Psi_g^A(t;\boldsymbol{n}) = 1 - \sum_{\alpha=0}^{k_A-1} \frac{\left(\lambda_g^A(\boldsymbol{n})\right)^\alpha}{\alpha!} t^\alpha e^{-\lambda_g^A(\boldsymbol{n})t},$$

$$\Psi_g^B(t;\boldsymbol{n}) = 1 - \sum_{\beta=0}^{k_B-1} \frac{\left(\lambda_g^B(\boldsymbol{n})\right)^\beta}{\beta!} t^\beta e^{-\lambda_g^B(\boldsymbol{n})t},$$

$$\Psi_{\text{deg}}^A(t;\boldsymbol{n}) = 1 - e^{-\lambda_{\text{deg}}^A n_A t},$$

$$\Psi_{\text{deg}}^B(t;\boldsymbol{n}) = 1 - e^{-\lambda_{\text{deg}}^B n_B t}. \tag{A1}$$

To determine the probability that the system is at state $\boldsymbol{n}$ at time $t$, we first assume that $\phi_i^X(t;\boldsymbol{n})$ represents the joint probability density function of one reaction $\{X,i\}$ in the system with $\{X,i\}\in\Omega$, defined by

$$\phi_i^X(t;\boldsymbol{n}) = \psi_i^X(t;\boldsymbol{n}) \prod_{\{Y,j\}\neq\{X,i\}} \left[1-\Psi_j^Y(t;\boldsymbol{n})\right], \{X,i\}\in\Omega, \tag{A2}$$

and the corresponding cumulative distribution function of $\phi_i^X(t;\boldsymbol{n})$ is calculated by $\Phi_i^X(t;\boldsymbol{n}) = \int_0^t \phi_i^X(t';\boldsymbol{n})dt'$, where $\phi_i^X(t;\boldsymbol{n})$ satisfies the probabilistic conservative condition $\sum_{\{X,i\}\in\Omega}\int_0^\infty \phi_i^X(t';\boldsymbol{n})dt' = 1$.

Furthermore, we define $R_k(t;\boldsymbol{n})$ as the probability density function of its waiting time for which the system reaches state $\boldsymbol{n}$ after undergoing $k$ reaction steps. Based on the renewal theory [24,33], the probability density function of next reaction step can be given by

$$R_{k+1}(t;\boldsymbol{n}) = \int_0^t \sum_{\{X,i\}\in\Omega} R_k\left(t';\boldsymbol{n}-\boldsymbol{v}_i^X\right)\phi_i^X\left(t-t';\boldsymbol{n}-\boldsymbol{v}_i^X\right)dt', \tag{A3}$$

where $R_0(t;\boldsymbol{n}) = p(\boldsymbol{n};t)\delta(t)$ with $\delta(t)$ being the Dirac $\delta$ function. Note that, if system is at state $\boldsymbol{n}$ at time $t$ after going through arbitrarily many reaction steps, the corresponding probability density function of its waiting time, denoted by $R(t;\boldsymbol{n})$, is given by

$$R(t;\boldsymbol{n}) = \sum_{k=0}^{\infty} R_k(t;\boldsymbol{n}). \tag{A4}$$

Thus, the probability $p(\boldsymbol{n};t)$ can be expressed by two parts: the one is the probability that the system arrives at state $\boldsymbol{n}$ at earlier time $t'$, and the other is the probability that the system has not a reaction that occurred within the remaining time $t-t'$, over the time interval $[0,t]$. That is,

$$p(\boldsymbol{n};t) = \int_0^t R(t';\boldsymbol{n})\left[1-\sum_{\{X,i\}\in\Omega}\Phi_i^X(t-t';\boldsymbol{n})\right]dt', \tag{A5}$$

where the equality $1-\sum_{\{X,i\}\in\Omega}\Phi_i^X(t-t';\boldsymbol{n}) = \prod_{\{Y,j\}\in\Omega}[1-\Psi_j^Y(t;\boldsymbol{n})]$ holds.

Based on the CTRW theory, we introduce a memory function $M_i^X(t;\boldsymbol{n})$ for every reaction $\{X,i\}\in\Omega$, which is defined by the Laplace transform as [24,32,42]

$$\tilde{M}_i^X(s;\boldsymbol{n}) = \frac{s\tilde{\phi}_i^X(s;\boldsymbol{n})}{1-\sum_{\{X,i\}\in\Omega}\tilde{\phi}_i^X(s;\boldsymbol{n})}, \tag{A6}$$

where $\tilde{\phi}_i^X(s;\boldsymbol{n})$ is Laplace transform of $\phi_i^X(s;\boldsymbol{n})$. Therefore, combining Eq. (A3)–(A6) yields

$$\begin{aligned} s\tilde{p}(\boldsymbol{n};s) = p(\boldsymbol{n};0) &+ \sum_{\{X,i\}\in\Omega}\tilde{M}_i^X\left(s;\boldsymbol{n}-\boldsymbol{v}_i^X\right)\tilde{p}\left(\boldsymbol{n}-\boldsymbol{v}_i^X;s\right) \\ &- \sum_{\{X,i\}\in\Omega}\tilde{M}_i^X(s;\boldsymbol{n})\tilde{p}(\boldsymbol{n};s), \end{aligned} \tag{A7}$$

which is a gCME in the Laplace domain. Interestingly, if we take the inverse Laplace transform in Eq. (A7), we obtain a gCME in the time domain

$$\begin{aligned} \frac{\partial p(\boldsymbol{n};t)}{\partial t} &= \int_0^t \left(\mathbb{E}_A^{-1}-1\right)\left(M_g^A(t-t';\boldsymbol{n})p(\boldsymbol{n};t')\right)dt' \\ &+ \int_0^t \left(\mathbb{E}_A-1\right)\left(M_{\text{deg}}^A(t-t';\boldsymbol{n})p(\boldsymbol{n};t')\right)dt' \\ &+ \int_0^t \left(\mathbb{E}_B^{-1}-1\right)\left(M_g^B(t-t';\boldsymbol{n})p(\boldsymbol{n};t')\right)dt' \\ &+ \int_0^t \left(\mathbb{E}_B-1\right)\left(M_{\text{deg}}^B(t-t';\boldsymbol{n})p(\boldsymbol{n};t')\right)dt'. \end{aligned} \tag{A8}$$

Next, we use Eq. (A7) in the Laplace domain to derive a sgCME for the sgTSM. For this, we apply the final value theorem to derive a practical equation from Eq. (A7). After introducing two limit functions $p(\boldsymbol{n}) = \lim_{t\to\infty}p(\boldsymbol{n};t)$ and $\boldsymbol{K}_i^X(\boldsymbol{n}) = \lim_{s\to 0}\tilde{M}_i^X(s;\boldsymbol{n})$, and taking the limit $s\to 0$ of Eq. (A7) multiplied $s$ on both sides, we get

$$\begin{aligned} &\left(\mathbb{E}_A^{-1}-1\right)\left(K_g^A(\boldsymbol{n})p(\boldsymbol{n})\right) + \left(\mathbb{E}_A-1\right)\left(K_{\text{deg}}^A(\boldsymbol{n})p(\boldsymbol{n})\right) \\ &\quad + \left(\mathbb{E}_B^{-1}-1\right)\left(K_g^B(\boldsymbol{n})p(\boldsymbol{n})\right) + \left(\mathbb{E}_B-1\right)\left(K_{\text{deg}}^B(\boldsymbol{n})p(\boldsymbol{n})\right) = 0, \end{aligned} \tag{A9}$$

where $p(\boldsymbol{n})$ is a stationary probability in the gTSM system and $\boldsymbol{K}_i^X(\boldsymbol{n})$ is the mean reaction propensity function of the reaction $\{X,i\}\in\Omega$ [The expression of $\boldsymbol{K}_i^X(\boldsymbol{n})$ will be deduced latter]. Notably, this function $\boldsymbol{K}_i^X(\boldsymbol{n})$ is memoryless since it does not depend on the prior history of the reaction process. For convenience, Eq. (A9) is called the stationary CME (sgCME) for the genetic toggle switch with molecular memory.

Substituting the Laplace transforms of $\phi_i^X(t;\boldsymbol{n})$ into the Eq. (A6) yields

$$\tilde{M}_i^X(s;\boldsymbol{n}) = \frac{\int_0^{+\infty} e^{-st}\psi_i^X(t;\boldsymbol{n})\prod_{\{Y,j\}\neq\{X,i\}}\left[1-\Psi_j^Y(t;\boldsymbol{n})\right]dt}{\int_0^{+\infty} e^{-st}\prod_{\{Y,j\}\in\Omega}\left[1-\Psi_j^Y(t;\boldsymbol{n})\right]dt}. \tag{A10}$$

After taking the limit $s\to 0$ of $\tilde{M}_i^X(s;\boldsymbol{n})$, the expression of $\boldsymbol{K}_i^X(\boldsymbol{n})$ is given by

$$K_i^X(\boldsymbol{n}) = \frac{\int_0^{+\infty}\psi_i^X(t;\boldsymbol{n})\prod_{\{Y,j\}\neq\{X,i\}}\left[1-\Psi_j^Y(t;\boldsymbol{n})\right]dt}{\int_0^{+\infty}\prod_{\{Y,j\}\in\Omega}\left[1-\Psi_j^Y(t;\boldsymbol{n})\right]dt}, \quad \{X,i\}\in\Omega. \tag{A11}$$

Thus, the mean reaction propensity function $\boldsymbol{K}_i^X(\boldsymbol{n})$ of four chemical reactions in the sgCME can be calculated by the Eq. (A11).

Substituting the expression $\psi_g^X(t;\boldsymbol{n})$ [Eq. (3a) in the main text] into Eq. (A11), and combining their cumulative distributions, we find that the explicit expression of $\boldsymbol{K}_g^X(\boldsymbol{n})$ is given by

$$K_g^A(\boldsymbol{n}) = \frac{\frac{(\lambda_g^A(\boldsymbol{n}))^{k_A}}{(k_A-1)!}\sum_{\beta=0}^{k_B-1}\frac{(\lambda_g^B(\boldsymbol{n}))^{\beta}}{\beta!}\frac{(\beta+k_A-1)!}{(\lambda_g^A(\boldsymbol{n})+\lambda_g^B(\boldsymbol{n})+\lambda_{\deg}^A n_A+\lambda_{\deg}^B n_B)^{\beta+k_A}}}{\sum_{\alpha=0}^{k_A-1}\frac{(\lambda_g^A(\boldsymbol{n}))^{\alpha}}{\alpha!}\sum_{\beta=0}^{k_B-1}\frac{(\lambda_g^B(\boldsymbol{n}))^{\beta}}{\beta!}\frac{(\alpha+\beta)!}{(\lambda_g^A(\boldsymbol{n})+\lambda_g^B(\boldsymbol{n})+\lambda_{\deg}^A n_A+\lambda_{\deg}^B n_B)^{\alpha+\beta+1}}}, \tag{A12a}$$

$$K_g^B(\boldsymbol{n}) = \frac{\frac{(\lambda_g^B(\boldsymbol{n}))^{k_B}}{(k_B-1)!}\sum_{\alpha=0}^{k_A-1}\frac{(\lambda_g^A(\boldsymbol{n}))^{\alpha}}{\alpha!}\frac{(\alpha+k_B-1)!}{(\lambda_g^A(\boldsymbol{n})+\lambda_g^B(\boldsymbol{n})+\lambda_{\deg}^A n_A+\lambda_{\deg}^B n_B)^{\alpha+k_B}}}{\sum_{\alpha=0}^{k_A-1}\frac{(\lambda_g^A(\boldsymbol{n}))^{\alpha}}{\alpha!}\sum_{\beta=0}^{k_B-1}\frac{(\lambda_g^B(\boldsymbol{n}))^{\beta}}{\beta!}\frac{(\alpha+\beta)!}{(\lambda_g^A(\boldsymbol{n})+\lambda_g^B(\boldsymbol{n})+\lambda_{\deg}^A n_A+\lambda_{\deg}^B n_B)^{\alpha+\beta+1}}}. \tag{A12b}$$

That is, we have

$$\begin{aligned} K_g^A(\boldsymbol{n}) &= \frac{\lambda_g^A(\boldsymbol{n})\sum_{\beta=0}^{k_B-1}\binom{\beta+k_A-1}{\beta}(\omega_B(\boldsymbol{n}))^{\beta}(\omega_A(\boldsymbol{n}))^{k_A-1}}{\sum_{\alpha=0}^{k_A-1}\sum_{\beta=0}^{k_B-1}\binom{\alpha+\beta}{\alpha}(\omega_A(\boldsymbol{n}))^{\alpha}(\omega_B(\boldsymbol{n}))^{\beta}}, \\ K_g^B(\boldsymbol{n}) &= \frac{\lambda_g^B(\boldsymbol{n})\sum_{\alpha=0}^{k_A-1}\binom{\alpha+k_B-1}{\alpha}(\omega_A(\boldsymbol{n}))^{\alpha}(\omega_B(\boldsymbol{n}))^{k_B-1}}{\sum_{\alpha=0}^{k_A-1}\sum_{\beta=0}^{k_B-1}\binom{\alpha+\beta}{\alpha}(\omega_A(\boldsymbol{n}))^{\alpha}(\omega_B(\boldsymbol{n}))^{\beta}}, \end{aligned} \tag{A13}$$

where $\omega_A(\boldsymbol{n}) = \lambda_g^A(\boldsymbol{n})/\Sigma(\boldsymbol{n})$, $\omega_B(\boldsymbol{n}) = \lambda_g^B(\boldsymbol{n})/\Sigma(\boldsymbol{n})$, $\Sigma(\boldsymbol{n}) = \lambda_g^A(\boldsymbol{n}) + \lambda_g^B(\boldsymbol{n}) + \lambda_{\deg}^A n_A + \lambda_{\deg}^B n_B$, $\lambda_g^A(\boldsymbol{n}) = g_A/(1 + r_A n_B^H)$, $\lambda_g^B(\boldsymbol{n}) = g_B/(1 + r_B n_A^H)$. Similarly, since $\psi_{\deg}^X(t;\boldsymbol{n})$ follows an exponential distribution [Eq. (3b) in the main text], its mean reaction propensity function can be reduced to

$$K_{\deg}^A(\boldsymbol{n}) = \lambda_{\deg}^A n_A, \quad K_{\deg}^B(\boldsymbol{n}) = \lambda_{\deg}^B n_B. \tag{A14}$$

Equations (A9) and (A13)–(A14) altogether define the sgCME for the gTSM with molecular memory.

## APPENDIX B: A NUMERICAL ALGORITHM FOR COMPUTING PROBABILITY DISTRIBUTION

Motivated by the final state projection method [51], we derive a stationary generalized finite state projection, which is capable of computing the stationary probability distribution in the gTSM.

First, we set $0 \leqslant i \leqslant m_A$, $0 \leqslant j \leqslant m_B$. For convenience, if we denote $i \equiv n_A$, $j \equiv n_B$, then $P_{ij} \equiv p(n_A, n_B)$. In addition, we denote

$$\begin{aligned} \tilde{\mathbf{P}}_{ij} &= \left(\mathbf{P}_{i,0}, \mathbf{P}_{i,1}, \cdots, \mathbf{P}_{i,m_B}\right)^{\mathrm{T}} \\ &= \left(\left(P_{0,0}, P_{1,0}, \cdots, P_{m_A,0}\right)^{\mathrm{T}}, \left(P_{0,1}, P_{1,1}, \cdots, P_{m_A,1}\right)^{\mathrm{T}}, \cdots, \left(P_{0,m_B}, P_{1,m_B}, \cdots, P_{m_A,m_B}\right)^{\mathrm{T}}\right)^{\mathrm{T}}, \end{aligned} \tag{B1}$$

where the conservative condition $\sum_{j=0}^{m_B}\sum_{i=0}^{m_A} P_{ij} = 1$ holds. Then, the sgCME, Eq. (A9), can become

$$\begin{aligned} &K_g^A(i-1,j)P_{i-1,j} + K_{\deg}^A(i+1,j)P_{i+1,j} + K_g^B(i,j-1)P_{i,j-1} + K_{\deg}^B(i,j+1)P_{i,j+1} \\ &\quad - \left[K_g^A(i,j) + K_g^B(i,j) + K_{\deg}^A(i,j) + K_{\deg}^B(i,j)\right]P_{ij} = 0. \end{aligned} \tag{B2}$$

Thereby, the matrix form of this truncated equation is given by

$$\mathbb{M}\vec{\mathbf{P}}_{ij} = 0 \tag{B3}$$

where

$$\mathbb{M} = \begin{pmatrix} D^0 & U_{\deg}^1 & & & & \\ L_g^0 & D^1 & U_{\deg}^2 & & & \\ & L_g^1 & D^2 & U_{\deg}^3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & L_g^{m_B-2} & D^{m_B-1} & U_{\deg}^{m_B} \\ & & & & L_g^{m_B-1} & D^{m_B} \end{pmatrix} \tag{B4}$$

with

$$D^j = A_g^j + A_{\deg}^j - L_g^j - U_{\deg}^j, 0 \leqslant j \leqslant m_B, \tag{B5}$$

which satisfies the following matrix:

$$L_g^j = \mathrm{diag}\big(K_g^B(0,j), K_g^B(1,j), K_g^B(2,j), \cdots, K_g^B(m_A,j)\big), \tag{B6}$$

$$U_{\mathrm{deg}}^j = \mathrm{diag}\big(K_{\mathrm{deg}}^B(0,j), K_{\mathrm{deg}}^B(1,j), K_{\mathrm{deg}}^B(2,j), \cdots, K_{\mathrm{deg}}^B(m_A,j)\big), \tag{B7}$$

$$A_g^j = \begin{pmatrix} -K_g^A(0,j) & & & & \\ K_g^A(0,j) & -K_g^A(1,j) & & & \\ & K_g^A(1,j) & -K_g^A(2,j) & & \\ & & \ddots & \ddots & \\ & & & K_g^A(m_A-1,j) & -K_g^A(m_A,j) \end{pmatrix}, \tag{B8}$$

$$A_{\mathrm{deg}}^j = \begin{pmatrix} -K_{\mathrm{deg}}^A(0,j) & K_{\mathrm{deg}}^A(1,j) & & & \\ & -K_{\mathrm{deg}}^A(1,j) & K_{\mathrm{deg}}^A(2,j) & & \\ & & -K_{\mathrm{deg}}^A(2,j) & \ddots & \\ & & & \ddots & K_{\mathrm{deg}}^A(m_A,j) \\ & & & & -K_{\mathrm{deg}}^A(m_A,j) \end{pmatrix}. \tag{B9}$$

By solving the algebraic equation Eq. (B3) with conservative condition $\sum_{j=0}^{m_B}\sum_{i=0}^{m_A} P_{ij} = 1$, we can obtain a numerical $\tilde{\mathbf{P}}_{ij}$.

[1] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie, Science **322**, 442 (2008).

[2] M. Acar, A. Becskei, and A. van Oudenaarden, Nature (London) **435**, 228 (2005).

[3] P. B. Gupta, C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, and E. S. Lander, Cell **146**, 633 (2011).

[4] N. Folguera-Blasco, R. Pérez-Carrasco, E. Cuyàs and J. A. Menendez, and T. Alarcón, PLoS Comput. Biol. **15**, e1006592 (2019).

[5] E. Kussell and S. Leibler, Science **309**, 2075 (2005).

[6] M. Wu, R. Q. Su, X. H. Li, T. Ellis, Y. C. Lai, and X. Wang, Proc. Natl. Acad. Sci. USA **110**, 10610 (2013).

[7] T. M. Norman, N. D. Lord, J. Paulsso, and R. Losic, Annu. Rev. Microbiol. **69**, 381 (2015).

[8] Y. Xu, Y. Li, H. Zhang, X. Li, and J. Kurths, Sci. Rep. **6**, 31505 (2016).

[9] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden, Nature (London) **427**, 737 (2004).

[10] M. Santillán, M. C. Mackey, and E. S. Zeron, Biophys. J. **92**, 3830 (2007).

[11] T. Tian and K. Burrage, J. Theor. Biol. **227**, 229 (2004).

[12] W. K. Smits, C. C. Eschevins, K. A. Susanna, S. Bron, O. P. Kuipers, and L. W. Hamoen, Mol. Microbiol. **56**, 604 (2005).

[13] J. Jaruszewicz-Błonska and T. Lipniacki, BMC Syst. Biol. **11**, 117 (2017).

[14] P. Wang, J. Lu, and X. Yu, IEEE/ACM Trans. Comput. Biol. Bioinform. **12**, 579 (2015).

[15] P. Bokes and A. Singh, in *2019 18th European Control Conference (ECC)* (IEEE, 2019), pp. 698-703.

[16] T. L. To and N. Maheshri, Science **327**, 1142 (2010).

[17] K. H. Klm and H. M. Sauro, BMC Biol. **10**, 89 (2012).

[18] A. Lipshtat, A. Loinger, N. Q. Balaban, and O. Biham, Phys. Rev. Lett. **96**, 188101 (2006).

[19] T. Biancalani and M. Assaf, Phys. Rev. Lett. **115**, 208101 (2015).

[20] A. Ochab-Marcinek and M. Tabaka, Proc. Natl. Acad. Sci. USA **107**, 22096 (2010).

[21] P. B. Warren and P. R. ten Wolde, J. Phys. Chem. B **109**, 6812 (2005).

[22] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences* (Springer, New York, 2009).

[23] E. Pardoux, *Markov Processes and Applications: Algorithms, Networks, Genome and Finance* (Wiley, New York, 2008).

[24] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 2007).

[25] J. M. Pedraza and J. Paulsson, Science **319**, 339 (2008).

[26] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Science **332**, 472 (2011).

[27] C. V. Harper, B. Finkenstädt, D. J. Woodcock, S. Friedrichsen, S. Semprini, L. Ashall, D. G. Spiller, J. J. Mullins, D. A. Rand, J. R. E. Davis, and M. R. H. White, PLoS Biol. **9**, e1000607 (2011).

[28] R. Schlicht and G. Winkler, J. Math. Biol. **57**, 613 (2008).

[29] C. Gupta, J. M. López, W. Ott, K. Josić, and M. R. Bennett, Phys. Rev. Lett. **111**, 058104 (2013).

[30] T. Jia and R. V. Kulkarni. Phys. Rev. Lett. **106**, 058102 (2011).

[31] J. Klafter and R. Silbey, Phys. Rev. Lett. **44**, 55 (1980).

[32] J. J. Zhang and T. S. Zhou, Proc. Natl. Acad. Sci. USA **116**, 23542 (2019).

[33] N. Kumar, A. Singh, and R. V. Kulkarni, PLoS Comput. Biol. **11**, e1004292 (2015).

[34] A. Schwabe, K. N. Rybakova, and F. J. Bruggeman, Biophys. J. **103**, 1152 (2012).

[35] S. J. Park, S. Song, G. S. Yang, P. M. Kim, S. Yoon, J. H. Kim, and J. Sung, Nat. Commun. **9**, 297 (2018).

[36] M. Barrio, A. Leier, and T. T. Marquez-Lago, J. Chem. Phys. **138**, 104114 (2013).

[37] A. Leier and T. T. Marquez-Lago, Proc. R. Soc. A. **471**, 20150049 (2015).

[38] A. Grönlund, P. Grönlund, and J. Elf, Nat. Commun. **2**, 419 (2011).

[39] E. W. Montroll and G. H. Weiss, J. Math. Phys. **6**, 167 (1965).

[40] V. M. Kenkre, E. W. Montroll, and M. F. Shlesinger, J. Stat. Phys. **9**, 45 (1973).

[41] U. Landman, E. W. Montroll, and M. F. Shlesinger, Proc. Natl. Acad. Sci. USA **74**, 430 (1977).

[42] T. Aquino and M. Dentz, Phys. Rev. Lett. **119**, 230601 (2017).

[43] K. Thurley, L. F. Wu, and S. J. Altschuler, Cell Syst. **6**, 355 (2018).

[44] T. S. Gardner, C. R. Cantor, and J. J. Collins, Nature (London) **403**, 339 (2000).

[45] M. I. Stefan and N. Le Novère, PLoS Comput. Biol. **9**, e1003106 (2013).

[46] J. Cherry and F. R. Adler, J. Theo. Biol. **203**, 117 (2000).

[47] M. Strasser, F. J. Theis, and C. Marr, Biophys. J. **102**, 19 (2012).

[48] B. Barzel and O. Biham, Phys. Rev. E. **78**, 041919 (2008).

[49] Y. Xu, Y. Zhu, J. Shen, and J. Su, Physica A **416**, 461 (2014).

[50] H. Chen, P. Thill, and J. Cao, J. Chem. Phys. **144**, 175104 (2016).

[51] B. Munsky and M. Khammash, J. Chem. Phys. **124**, 044104 (2006).