Incorporating age and delay into models for biophysical systems

Wasiur R. KhudaBukhsh [1], **Hye-Won Kang** [2], **Eben Kenah** [3], **and Grzegorz A. Rempała** [4]

[1] Mathematical Biosciences Institute and the College of Public Health, The Ohio State University, 1735 Neil Avenue, Columbus OH 43210, USA
[2] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore MD 21250, USA
[3] Division of Biostatistics, College of Public Health, The Ohio State University, 1841 Neil Avenue, Columbus OH 43210, USA
[4] Mathematical Biosciences Institute and the College of Public Health, The Ohio State University, 1735 Neil Avenue, Columbus OH 43210, USA

E-mail: `khudabukhsh.2@osu.edu`

Abstract. In many biological systems, chemical reactions or changes in a physical state are assumed to occur instantaneously. For describing the dynamics of those systems, Markov models that require exponentially distributed inter-event times have been used widely. However, some biophysical processes such as gene transcription and translation are known to have a significant gap between the initiation and the completion of the processes, which renders the usual assumption of exponential distribution untenable. In this paper, we consider relaxing this assumption by incorporating age-dependent random time delays (distributed according to a given probability distribution) into the system dynamics. We do so by constructing a measure-valued Markov process on a more abstract state space, which allows us to keep track of the "ages" of molecules participating in a chemical reaction.

We study the large-volume limit of such age-structured systems. We show that, when appropriately scaled, the stochastic system can be approximated by a system of Partial Differential Equations (PDEs) in the large-volume limit, as opposed to Ordinary Differential Equations (ODEs) in the classical theory. We show how the limiting PDE system can be used for the purpose of further model reductions and for devising efficient simulation algorithms. In order to describe the ideas, we use a simple transcription process as a running example. We, however, note that the methods developed in this paper apply to a wide class of biophysical systems.

Keywords: stochastic transcription; translation; random time delays; multiscale analysis; survival dynamical systems; age-dependent processes; non-Markovian systems.

1. Introduction

We consider biophysical systems described by a set of chemical reactions. The chemically identical molecular entities in the system are called (chemical) species. A chemical reaction refers to the event of creation, annihilation, or conversion of a number of molecules of one or more species. Here, we assume the system is well mixed spatially in that a randomly chosen molecule of a species has an equal chance to chemically interact with any other molecule of any species in the system. A Continuous Time Markov Chain (CTMC) is a natural choice to model the species copy numbers of such systems.

When modeling Chemical Reaction Networks (CRNs) stochastically using CTMCs, one assumes that every reaction occurs instantaneously after an exponentially distributed amount of time. Whenever a reaction takes place, we update the system state. A random time-change representation of the Poisson process is often used to write the trajectory equations and to analyze the system dynamics [1, 2, 3, 4]. The sample paths of the CTMC are simulated exactly using the Doob–Gillespie's Stochastic Simulation Algorithm (SSA) [5, 6, 7] or the next reaction method by Gibson and Bruck [8].

1.1. Delays are inherent and a useful model reduction tool

It has been reported that some biological processes do not take place instantaneously. In other words, there is a time lag between the initiation and the completion of the process. Time delays are observed inherently in many biological systems, such as gene transcription [9, 10, 11] and translation [12], cell cycle in cancer treatment [13], intracellular viral dynamics [14, 15], control of infectious diseases [16], population growth [17, 18], RNA and protein folding [19, 20], and enzyme catalyzed reactions [21, 22]. Sometimes time delays are introduced purposefully as a useful means to reduce model complexity and compensate for the lack of experimental observation in both deterministic and stochastic models of biological processes.

Intermediate, ancillary processes or unobserved reactions can be replaced by time delays. For example, production of hes1 mRNA from hes1 gene has been modeled using delay differential equations where detailed mRNA synthesis and processing steps are replaced by a time delayed reaction [23]. While modeling the mammalian circadian clock, intermediate protein dynamics can be simplified as transcriptional feedback loops with time delayed variables in delay differential equations [24]. In enzyme catalyzed reactions with multiple intermediates, the production of the final product can be expressed as a distributed delay equation, which is a useful tool when measurements on multiple intermediates in the experiment are not available [25].

Introduction of time delays as a model reduction technique has also been applied in discrete stochastic models for CRNs. For instance, model complexity of unimolecular reaction networks is reduced by generating delay distributions with key model features that are derived by computing first passage times of target species [26]. In [27], the production of yellow fluorescent protein has been described using a time-delayed birth and death process where a randomly distributed time delay was generated to simplify a sequence of steps in gene activation.

1.2. Our contribution

In most previous works in this area, the focus was on investigating stochastic models for CRNs with constant or random time delays. In those models, the probability that a reaction occurs within the next short interval of time is commonly described by a propensity (also known as intensity) function of the reaction. The waiting time for non-delayed reactions is exponentially distributed [28]. In practice, the occurrence of some reactions is not only determined by the molecular counts of the reactants but also affected by the age distributions or lifetimes of the reactant molecules. For example, mRNA decay rates vary depending on the age of each mRNA. Moreover, the age of the mRNAs determines polysome size distributions and protein synthesis rates in translation ([29, 30], Chapters 3 and 5 in [31]). It was also reported that an mRNA tail length distribution depends on the average age of mRNA population and that the tail-length distribution plays a significant role in deadenylation and decay dynamics of mRNA populations [32, 33].

When time delays are used to aggregate out ancillary or unobserved processes and reduce model complexity, it makes more sense that the length of time delay depends on the age of each reactant molecule (*e.g.*, mRNA, protein, and enzymes). Therefore, it is worthwhile to consider an individual-based age-structured stochastic model for CRNs.

In this work, we develop a way to describe CRNs with random time delays and non-delayed reaction rates incorporating the *age* of each reactant and making use of *hazard functions* in survival analysis [34, 35]. See Appendix C for some preliminaries on relevant mathematical and statistical concepts. In our approach, the hazard functions are set as constant, time-dependent, or age-dependent functions generalizing the notion of reaction rate constants in propensity functions. Our model keeps track of the age of each reactant molecule and provides a new way to express time delays in non-Markovian models. Moreover, the method also allows us to describe discrete stochastic CRNs with constant or random time delays without age dependence, as considered in previous works. We study the large-volume limit of the proposed non-Markov CRN and provide a mean-field PDE limit for the age densities by virtue of the Law of Large Numbers (LLN), as opposed to an ODE limit in the classical theory. The PDE limit is based on existing results in the literature [36, 37] and follows from the standard limit theory for measure-valued Markov processes. However, novel usage of the PDE limit can provide further approximations and pave the way for efficient simulation algorithms. For the sake of illustration, we show how the PDE limit can be used to approximate Mean First Passage Times (MFPTs) in the context of CRNs. As another by-product of the LLN, we show how an efficient (fast) hybrid simulation algorithm can be devised when a subset of the CRN is abundantly available, giving a flavor of multiscale approximation. Finally, as simple applications of our approach, we briefly discuss a prokaryotic auto-regulation and the Quasi-Steady State Approximation (QSSA) in the context of the Michaelis–Menten enzyme kinetic reactions. Numerical examples have been provided wherever deemed necessary. For the sake of ready usage of our methods, the Julia scripts used in the numerical examples have been made available via a GitHub repository [38].

The following notational conventions are adhered to throughout the paper. We use $\mathbb{1}_{\{A\}}(x)$ to denote the indicator (or characteristic) function of a set A, *i.e.*, $\mathbb{1}_{\{A\}}(x) = 1$ if and only if $x \in A$. Given a suitable space E, let $D([0,\infty), E)$ (or $D([0,T], E)$) denote the space of E-valued càdlàg functions defined on $[0,\infty)$ (or $[0,T]$, for some $T > 0$). The set of Borel subsets of a set A will be denoted by $\mathcal{B}(A)$. The set of natural numbers are denoted by $\mathbb{N}$. The set of real numbers is denoted by $\mathbb{R}$. Other notations will be introduced as and when needed.

2. The simplest model with a delay

Let us consider a simple CRN with two chemical species A and B. First, we shall describe the standard Markovian approach and then introduce an age structure to allow non-exponential holding times. The following network describes the production and the degradation of A along with a conversion from A to B

$$\begin{aligned} &\emptyset \longrightarrow^{b} A \longrightarrow^{\tau} B, \\ &A \longrightarrow^{d} \emptyset. \end{aligned} \tag{2.1}$$

where b, τ, and d, depending on whether we are in the Markovian or non-Markovian setup, will be either reaction rate constants or hazard functions for the production of A, the conversion from A to B, and the degradation of A, respectively.

An example similar to the CRN in Equation (2.1) was investigated in some previous works with time delays [39, 40]. It is worth noting that the simplistic CRN described in Equation (2.1) can be thought of as a model reduction of a more complex CRN. For instance, a series of conversion type reactions

$$A \longrightarrow^{k_1} A_1 \longrightarrow^{k_2} A_2 \longrightarrow^{k_3} \cdots \longrightarrow^{k_n} B \tag{2.2}$$

can be described by a single conversion reaction $A \longrightarrow^{\tau} B$ with an appropriate hazard function τ. For the sake of illustration, let us assume we are in the Markovian setup so that $k_1, k_2, \ldots, k_n$ are positive constants. We can interpret the CRN in Equation (2.2) as follows: One molecule of A gets transformed into a molecule of A_1 after an exponentially distributed (with rate k_1) amount of time. Then, the molecule of A_1 gets transformed into a molecule of A_2 after an exponentially distributed (with rate k_2 this time) amount of time. This process goes on until the molecule finally gets transformed into a molecule of B from a molecule of A_{n-1}. Therefore, from the perspective of a single A molecule, the amount of time required for the molecule to finally get transformed into a molecule of B is the sum of those exponentially distributed amounts of times (with rates $k_1, k_2, \ldots, k_n$). Under independence, the probability distribution of the total amount of time required for a single A molecule to get transformed into a B molecule can be described by a convolution of the individual exponential distributions. Denoting the corresponding hazard function by τ, one can describe the CRN in Equation (2.2) by a single conversion reaction $A \longrightarrow^{\tau} B$. Similarly, a series of birth-death-conversion type reactions

$$\begin{aligned}&\emptyset \longrightarrow^{b} A \longrightarrow^{k_1} A_1 \longrightarrow^{k_2} A_2 \longrightarrow^{k_3} \cdots \longrightarrow^{k_n} B,\\ &A \longrightarrow^{d} \emptyset, A_1 \longrightarrow^{d_1} \emptyset, A_2 \longrightarrow^{d_2} \emptyset, \cdots, A_n \longrightarrow^{d_n} \emptyset\end{aligned}$$

can be approximated by a single birth type reaction $\emptyset \longrightarrow^{\tau} B$ with an appropriate hazard function τ. Therefore, even a simplistic model such as the CRN in Equation (2.1) covers a nontrivial class of CRNs and builds the foundation for studying more complex CRNs.

2.1. Standard Markov approach

The standard way to model the CRN in Equation (2.1) is to use a CTMC to describe the counts of molecules of the species A and B over time. In such a model, whenever each reaction fires, the consumption and the production of molecules are instantaneous. Let $\tilde{X}_A, \tilde{X}_B$ denote the stochastic processes counting the copy numbers of the species A and B respectively. Here, the quantities b, τ, and d are reaction rate constants. The propensity functions corresponding to the three chemical reactions are defined as

$$\lambda_b(t) = b, \qquad \lambda_\tau(t) = \tau \times x_A(t), \qquad \lambda_d(t) = d \times x_A(t),$$

where $x_i(t)$ denotes the number of molecules of the chemical species i at time t, for $i = A, B$. Define T_k to be the waiting time until the next reaction of type birth ($k = b$), conversion ($k = \tau$), and death ($k = d$). Then, T_k is exponentially distributed with rate $\lambda_k(t)$ for $k = b, \tau, d$. The probability of each reaction's occurrence is expressed in terms of the corresponding propensity function as follow:

$$\mathsf{P}\left(t \le T_k < t + \Delta t \mid \tilde{X}_A(t) = x_A, \tilde{X}_B(t) = x_B\right) \approx \lambda_k(t)\Delta t + o(\Delta t)$$

for $k = b, \tau, d$ when Δt is small enough. Then, the trajectory equations can be written in a straightforward fashion following the random time changed representation of Poisson processes as

$$\begin{aligned}\tilde{X}_A(t) &= \tilde{X}_A(0) + R_1(bt) - R_2\left(\int_0^t \tau \tilde{X}_A(s)\,\mathrm{d}s\right) - R_3\left(\int_0^t d\,\tilde{X}_A(s)\,\mathrm{d}s\right),\\ \tilde{X}_B(t) &= R_2\left(\int_0^t \tau \tilde{X}_A(s)\,\mathrm{d}s\right),\end{aligned}$$

where R_1, R_2, and R_3 are unit rate Poisson processes [2]. We assume we do not have any B molecules in the system initially, *i.e.*, $\tilde{X}_b(0) = 0$. Now, if we scale the stochastic processes by a scaling parameter n, *e.g.*, volume of the system, it follows directly from the LLN for Poisson processes [41, 42] that the scaled process $(n^{-1}\tilde{X}_A, n^{-1}\tilde{X}_B)$ can be approximated by the solution to the following system of ODEs:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}x_A(t) &= -(\tau + d)x_A(t),\\ \frac{\mathrm{d}}{\mathrm{d}t}x_B(t) &= \tau x_A(t).\end{aligned}$$

Notice that the birth rate b vanishes in the limit because we did not assume any scaling of b with respect to n. In general, one would assume that the overall birth rate scales linearly with n so that it is sustained in the limit.

2.2. Age-structured model

Now, let us introduce *age* and *delay* into the CRN described by Equation (2.1). We assume that the production rate of B and the degradation rate of A depend on the age of the reactant molecule of species A. We use "age" as an umbrella term to refer to the amount of elapsed time since a specific event. Thus, "age" could mean different things depending on the application area. The most straightforward way is the biological or the physical age, which we take as the time duration since the molecule was born or created. In systems where a certain reaction can fire only when a gene is activated, one could define age as the time duration since activation of the gene. In some cases, it may be desirable to define delays in terms of time duration since the initiation of a reaction. The notion of age is sufficiently general to account for those cases as well. For example, a reaction $A \to B$ in which the delay is defined purely in terms of time difference between initiation and completion of the reaction, can be replaced by the reaction system $A \to F \to B$ where F is a fictitious species. The physical age of this fictitious species F is precisely the time since the initiation of the reaction $A \to B$. Now, putting an appropriate hazard function on the reaction $F \to B$, we can introduce a random or a deterministic delay in the reaction $A \to B$. Therefore, for the CRN in Equation (2.1), it seems sufficient to define the age to be the physical age of the molecules of A.

When we have an age-structured model, the counts (copy numbers of the species A and B) are inherently non-Markovian unless the holding times are exponentially distributed. However, if we keep track of the ages of the molecules in addition to the counts, we can get a Markov system, albeit on a more abstract state space. A neat way to do so is to use measure-valued processes that keep track of the age distribution of the molecules over time. Moreover, the measure-valued processes are also Markovian, which allows us to make use of the already existing limit theory for Banach space-valued Markov processes. This approach to age-structured modeling in biology is not new. Our work builds on the existing literature [36, 37, 43, 44]. In the next section, we describe how the measure-valued processes can be utilized in the context of the CRN in Equation (2.1).

2.3. The measure-valued process and the limiting system

Let us denote by $N_A(t)$ and $N_B(t)$ the numbers of molecules of the chemical species A and B at time t. Then, individual molecules of A are labelled $1, 2, \cdots, N_A(t)$. We denote the age of the i-th molecule of the species A by $a_i(t)$ for $i = 1, 2, \cdots, N_A(t)$. Similarly, we denote by $b_j(t)$ the age of the j-th molecule of the species B at time t. Now, we define a measure-valued process $X_t = \left(X_t^A, X_t^B\right)$ where X_t^A and X_t^B describe the age distributions of chemical species A and B at time t. To be more precise, we define

$$X_t^A := \sum_{i=1}^{N_A(t)} \delta_{a_i(t)}, \quad X_t^B := \sum_{i=1}^{N_B(t)} \delta_{b_i(t)}, \tag{2.3}$$

where δ_x is the Dirac measure, a function that takes value 1 if the argument to the function (a measurable set) contains x and zero otherwise. The components X_t^A and X_t^B of X_t are finite point measures with atoms placed on the individual ages of the molecules. For example, $X_t^A\left((0.5, 11.25]\right) = \sum_{i=1}^{N_A(t)} \delta_{a_i(t)}\left((0.5, 11.25]\right)$ gives us the count of species A molecules with ages in the set $(0.5, 11.25]$ at time t. In general, $X_t^A(F)$ gives us the count of species A molecules whose ages lie in the set F at time t.

For any point measure $\mu = \sum_{i=1}^{n} \delta_{x_i}$ and a measurable function f, we denote the integration of the function f with respect to the measure μ by

$$\langle \mu, f \rangle := \int f \, d\mu = \sum_{i=1}^{n} f(x_i).$$

If $\mu := (\mu_1, \mu_2, \ldots, \mu_L)$, for some positive integer L, is a vector of point measures and f is a measurable function, we use the notation $\langle\langle \mu, f \rangle\rangle$ to denote

$$\langle\langle \mu, f \rangle\rangle := \sum_{i=1}^{L} \langle \mu_i, f \rangle.$$

Therefore, we have $N_A(t) = \langle X_t^A, 1 \rangle = X_t^A(\mathbb{R}_+)$ and $N_B(t) = \langle X_t^B, 1 \rangle = X_t^B(\mathbb{R}_+)$ where 1 stands for the identity function. The set of non-negative real numbers is denoted by $\mathbb{R}_+$. The total population size is given by

$$N(t) := \langle\langle X_t, 1 \rangle\rangle = N_A(t) + N_B(t).$$

The process X_t is a Markov process on the space $D([0,T], \mathcal{M}_P(\mathbb{R}_+) \times \mathcal{M}_P(\mathbb{R}_+))$ where $T > 0$ is a finite time horizon and $\mathcal{M}_P(\mathbb{R}_+)$ is the space of finite, point measures on $\mathbb{R}_+$.

In order to simplify notations, we introduce maps $\sigma_i : \mathcal{M}_P(\mathbb{R}_+) \to \mathbb{R}_+$, for $i = 1, 2, 3, \ldots$, the purpose of which is to extract the i-th atom (the age of the i-th molecule) from a point measure following some partial order (*e.g.*, "greater or equal to" relation). Therefore, $\sigma_i(X_t^A)$ gives us the age of the i-th molecule of the species A at time t. We can now write down the trajectory equations:

$$\begin{aligned} X_t^A = {} & \sum_{k=1}^{N_A(0)} \delta_{t+\sigma_k(X_0^A)} + \int_0^t \int_0^\infty \delta_{t-s} \, 1_{\{\theta \le b\}} \, Q_1(ds, d\theta) \\ & - \int_0^t \int_{\mathbb{N}} \int_0^\infty \delta_{t-s+\sigma_i(X_{s-}^A)} \, 1_{\{i \le N_A(s-)\}} \, 1_{\{\theta \le \tau(\sigma_i(X_{s-}^A))\}} \, Q_2(ds, di, d\theta) \\ & - \int_0^t \int_{\mathbb{N}} \int_0^\infty \delta_{t-s+\sigma_i(X_{s-}^A)} \, 1_{\{i \le N_A(s-)\}} \, 1_{\{\theta \le d(\sigma_i(X_{s-}^A))\}} \, Q_3(ds, di, d\theta), \end{aligned} \tag{2.4}$$

$$X_t^B = \int_0^t \int_{\mathbb{N}} \int_0^\infty \delta_{t-s} \, 1_{\{i \le N_A(s-)\}} \, 1_{\{\theta \le \tau(\sigma_i(X_{s-}^A))\}} \, Q_2(ds, di, d\theta), \tag{2.5}$$

where Q_1, Q_2, Q_3 are independent Poisson Point Measures (PPMs) with intensity measures $\mathrm{d}s \times \mathrm{d}\theta$, $\mathrm{d}s \times \mathrm{d}i \times \mathrm{d}\theta$, and $\mathrm{d}s \times \mathrm{d}i \times \mathrm{d}\theta$ respectively, where $\mathrm{d}i$ is a counting measure on $\mathbb{N}$, and $\mathrm{d}s$ and $\mathrm{d}\theta$ are Lebesgue measures on $\mathbb{R}_+$. Provided the global jump rates are upper bounded by a finite quantity and the initial population size does not explode $\sup_n \mathsf{E}\left[n^{-1}N_A(0)\right] < \infty$, the trajectory equations admit a unique pathwise solution (X_t^A, X_t^B) (see [37, Theorem 2.5] for a similar derivation).

Under some assumptions on the hazard functions and the initial age distribution of the A molecules, the scaled process $n^{-1}X_t$ converges to a deterministic, continuous function $x_t := (x_t^A, x_t^B)$ whose components x_t^A and x_t^B are themselves measure-valued functions satisfying

$$\begin{aligned}\langle x_t^A, f_t\rangle &= \langle x_0^A, f_0\rangle \\ &\quad + \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) - f_s(a)(\tau(a) + d(a))\right) x_s^A(\,\mathrm{d}a)\,\mathrm{d}s \\ \langle x_t^B, f_t\rangle &= \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) + f_s(0)\tau(a)\right) x_s^A(\,\mathrm{d}a)\,\mathrm{d}s,\end{aligned}$$

for a sufficiently large class of test functions $f : (a, s) \to f_s(a)$. The convergence of the scaled stochastic process $n^{-1}X_t$ to the deterministic function x_t can be proved using techniques similar to those in [36, 45, 37, 43, 44]. However, for the sake of completeness, a brief, intuitive argument is presented in Appendix D.

Since the measure-valued function x_t^B is determined entirely by x_t^A, it suffices to study x_t^A. The densities $y_A(t, a)$ of the measure x_t^A, when they exist, are an important quantity describing the distribution of age of the species A molecules in the large-volume mean-field limit. The density function y_A should satisfy

$$(\partial_t + \partial_s)\, y_A(t, s) = -\left(\tau(s) + d(s)\right) y_A(t, s), \tag{2.6}$$

with the initial and the boundary conditions

$$y_A(0, s) = f_A(s), \quad y_A(t, 0) = 0,$$

where $f_A(s)$ specifies the age distribution of A molecules at time $t = 0$. To be more precise, it is the density of the limiting measure x_0^A, which we assume exists, with respect to the Lebesgue measure. Notice that the birth rate b vanishes in the limit, as in case of CTMC model, because we did not assume any scaling of the birth rate with respect to n.

Let y_B denote the limiting proportion of B molecules in the system. Then, y_B can be described entirely in terms of the density y_A as a solution to the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t} y_B(t) = \int_0^\infty \tau(s) y_A(t, s)\,\mathrm{d}s, \tag{2.7}$$

with the initial condition $y_B(0) = 0$. Luckily, the limiting system Equation (2.6) can be solved explicitly using standard analysis techniques:

$$y_A(t, s) = f_A(s - t) S_\tau(s) S_d(s) / \left(S_\tau(s - t) S_d(s - t)\right),$$

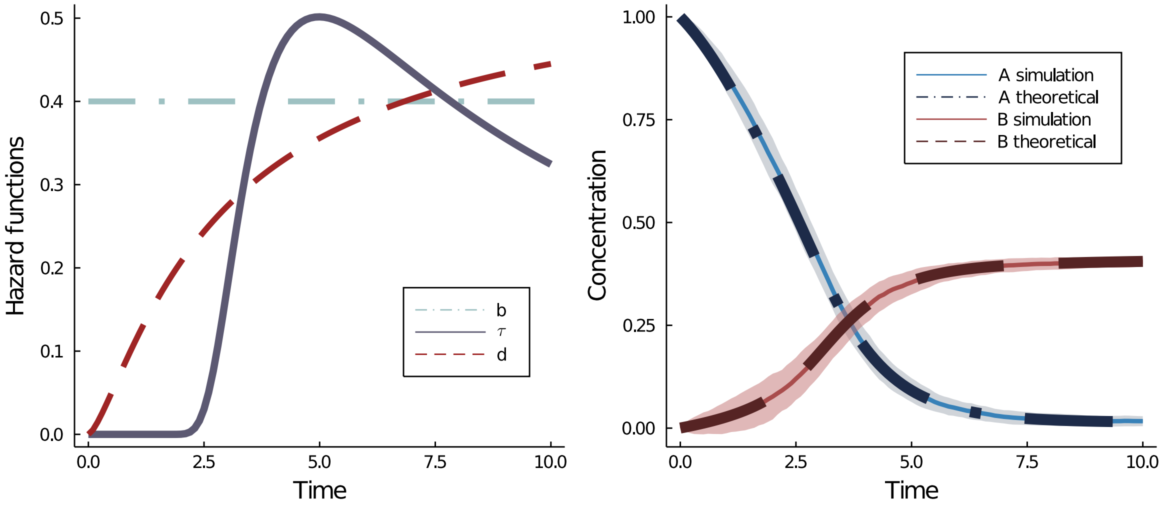

Figure 1. (Left) The shapes of the three hazard functions in the CRN described by Equation (2.1). Here, $b = 0.4$. The hazard functions τ and d correspond to a Generalized Extreme Value distribution with parameters $(1.25/0.30, 1.250, 0.30)$ and a Gamma distribution with parameters $(2.5, 1.75)$ respectively. Here, the conversion reaction has been explicitly made a delayed one. (Right) Comparison of the theoretical limiting trajectory and the simulated trajectories of concentrations of A and B molecules. The mean of the simulated trajectories is shown in solid lines, while the theoretical mean curve (given by the PDE limit) is shown in dashed lines. The width of the ribbons indicate 1 standard derivation fluctuation around the mean. Here, $n = 100$, *i.e.*, the initial number of A molecules is 100. It is evident that the theoretical limit provides a fairly accurate approximation to the scaled processes.

where S_τ and S_d are the survival functions of the probability distributions characterized by the hazard functions τ and d respectively. Therefore, the limiting concentration of B molecules can be described by

$$y_B(t) = \int_0^t \int_0^\infty \tau(v) y_A(u, v) \, \mathrm{d}v \, \mathrm{d}u.$$

In Figure 1, we numerically show the agreement between the theoretical limits in Equations (2.6) and (2.7) and the stochastic simulation. More specifically, we compare $\int_0^\infty y_A(t, s) \, \mathrm{d}s$ with stochastic simulations of $\langle n^{-1} X_t^A, 1\rangle$ and $y_B(t)$, with $\langle n^{-1} X_t^B, 1\rangle$. As it can be verified, the approximation error vanishes in the limit. Because X_t is a Markov process, the simulation of the stochastic CRN in Equation (2.1) can be carried out by adapting the Doob–Gillespie's SSA, which involves simulating two quantities at each step: 1) simulating the next reaction time; and 2) determining the reaction type. Note that, for the CRN in Equation (2.1), there are $(2N_A(t) + 1)$ different reactions possible at time t, even though there are only three types of reactions. The next reaction time can be simulated by drawing an exponential random variable with rate equal to the total hazard (the sum of the hazards corresponding to those $(2N_A(t) + 1)$ possible reactions). The total hazard is given by $b + \langle X_t^A, \tau\rangle + \langle X_t^A, d\rangle$. The type of reaction is then decided by drawing a categorical random variable whose probability masses are the ratios of the individual hazards and the total hazard. This discrete event simulation algorithm is

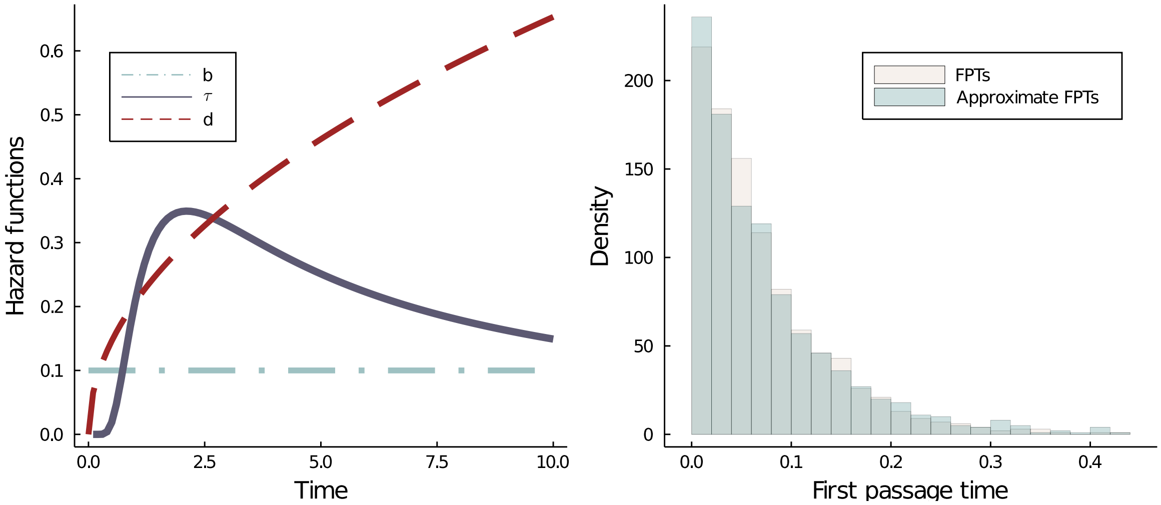

Figure 2. (Left) The shapes of the three hazard functions in the CRN described by Equation (2.1). Here, $b = 0.1$. The hazard functions τ and d characterize an Inverse Gamma distribution with parameters $(1.75, 4.25)$ and a Weibull distribution with parameters $(1.5, 3.75)$ respectively. (Right) The density of approximate First Passage Times (FPTs) match that of the true FPTs. Here, $n = 100$.

a straightforward adaptation of Doob–Gillespie's SSA for CTMCs. However, it must be noted that the simulation of a non-Markovian CRN is computationally more expensive than the CTMCs. For the sake of completeness, a pseudocode describing the above procedure is given in Algorithm 2.1. An implementation in the Julia programming language [46] is also made available in [38].

In Section 1, we mentioned that introduction of delay into a CRN could also serve the purpose of model reduction. Indeed, the LLN limit $y := (y_A, y_B)$ provides a model reduction of the original non-Markovian CRN in Equation (2.1). In the following, we discuss two other examples of usefulness of the LLN limit in the form of a PDE system. The first one approximates MFPTs, while the second one describes a faster simulation algorithm.

2.4. Mean First Passage Times

Mean First Passage Times are important quantities in the study of stochastic processes and dynamical systems. In the context of CRNs, they could arise in several ways [47, 48]. For instance, natural questions that could arise for the CRN in Equation (2.6) are how long it takes to deplete all molecules of species A or to produce the first molecule of B. One of the benefits of the LLN limit is that it can be used to approximate FPTs when the scaling parameter n is sufficiently large. The following illustrates this point.

Suppose we are interested in the time required to produce the first molecule of B. Following the exact simulation Algorithm 2.1 adapted from Doob–Gillespie's SSA, the total hazard for the production of a B molecule is $\langle X_0^A, \tau \rangle$. In the large-volume limit, we can approximate this hazard by $\int_0^\infty n\tau(s) y_A(0,s)\,\mathrm{d}s$. Therefore, for a sufficiently large

Algorithm 2.1 Pseudocode for the exact simulation of the CRN in Equation (2.1).

```
Require: $n$, $X_0$, $K$    ▷ $K$: Maximum number of iterations
Ensure: $(t_1, X_{t_1}), (t_2, X_{t_2}), \ldots$    ▷ Timings of the reactions along with the measures

Set $t = 0$
for $i = 1, 2, \ldots, K$ do
                                        ▷ Compute the next reaction time
    Calculate $\Lambda = \left(b + \langle X^A_{t_{i-1}}, \tau\rangle + \langle X^A_{t_{i-1}}, d\rangle\right)^{-1}$    ▷ $\Lambda^{-1}$: Total hazard
    if $0 < \Lambda < \infty$ then
        Draw an exponential random variable $T$ with mean $\Lambda$, i.e., $T \sim \textsc{Exponential}(\Lambda)$
        Set $t_i = t_{i-1} + T$    ▷ Advance time to the next reaction time

                                        ▷ Determine the reaction type
        Define $\pi_1 = \Lambda b$    ▷ Probability for the birth reaction
        Define $\pi_j = \Lambda\tau(\sigma_{j-1}(X^A_{t_{i-1}}))$ for $j = 2, 3, \ldots, (N_A(t_{i-1}) + 1)$    ▷ Probabilities for the transformation reaction
        Define $\pi_j = \Lambda d(\sigma_{j-N_A(t_{i-1})-1}(X^A_{t_{i-1}}))$ for $j = (N_A(t_{i-1}) + 2), (N_A(t_{i-1}) + 3), \ldots, (2N_A(t_{i-1}) + 1)$    ▷ Probabilities for the death reaction
        Set $\pi := (\pi_1, \pi_2, \ldots, \pi_{2N_A(t_{i-1})+1})$
        Draw a categorical random variable $L$ with probability distribution $\pi$
        if $L = 1$ then    ▷ Birth reaction
            $X^A_{t_i} = \delta_0 + \sum_{k=1}^{N_A(t_{i-1})} \delta_{\sigma_k(X^A_{t_{i-1}})+T}$    ▷ Advance ages of all $A$ molecules by $T$ and add an atom $\{0\}$
            $X^B_{t_i} = \sum_{k=1}^{N_B(t_{i-1})} \delta_{\sigma_k(X^B_{t_{i-1}})+T}$    ▷ Advance ages of all $B$ molecules by $T$
        else if $L \leq (N_A(t_{i-1}) + 1)$ then    ▷ Transformation reaction
            $X^A_{t_i} = \sum_{k=1}^{N_A(t_{i-1})} \delta_{\sigma_k(X^A_{t_{i-1}})+T} - \delta_{\sigma_{L-1}(X^A_{t_{i-1}})+T}$    ▷ Remove the atom $\{\sigma_{L-1}(X^A_{t_{i-1}})\}$ from the measure $X^A_{t_{i-1}}$ and advance ages of all other $A$ molecules by $T$
            $X^B_{t_i} = \delta_0 + \sum_{k=1}^{N_B(t_{i-1})} \delta_{\sigma_k(X^B_{t_{i-1}})+T}$    ▷ Advance ages of all $B$ molecules by $T$ and add an atom $\{0\}$
        else    ▷ Death reaction
            $X^A_{t_i} = \sum_{k=1}^{N_A(t_{i-1})} \delta_{\sigma_k(X^A_{t_{i-1}})+T} - \delta_{\sigma_{L-N_A(t_{i-1})-1}(X^A_{t_{i-1}})+T}$    ▷ Remove the atom $\{\sigma_{L-N_A(t_{i-1})-1}(X^A_{t_{i-1}})\}$ from the measure $X^A_{t_{i-1}}$ and advance ages of all other $A$ molecules by $T$
            $X^B_{t_i} = \sum_{k=1}^{N_B(t_{i-1})} \delta_{\sigma_k(X^B_{t_{i-1}})+T}$    ▷ Advance ages of all $B$ molecules by $T$
        end if
    else
        Stop and break loop
    end if
    Set $i = i + 1$.
end for
```

n, the MFPT can be approximated by

$$m = \left(\int_0^\infty n\tau(s) y_A(0, s)\,\mathrm{d}s\right)^{-1}, \tag{2.8}$$

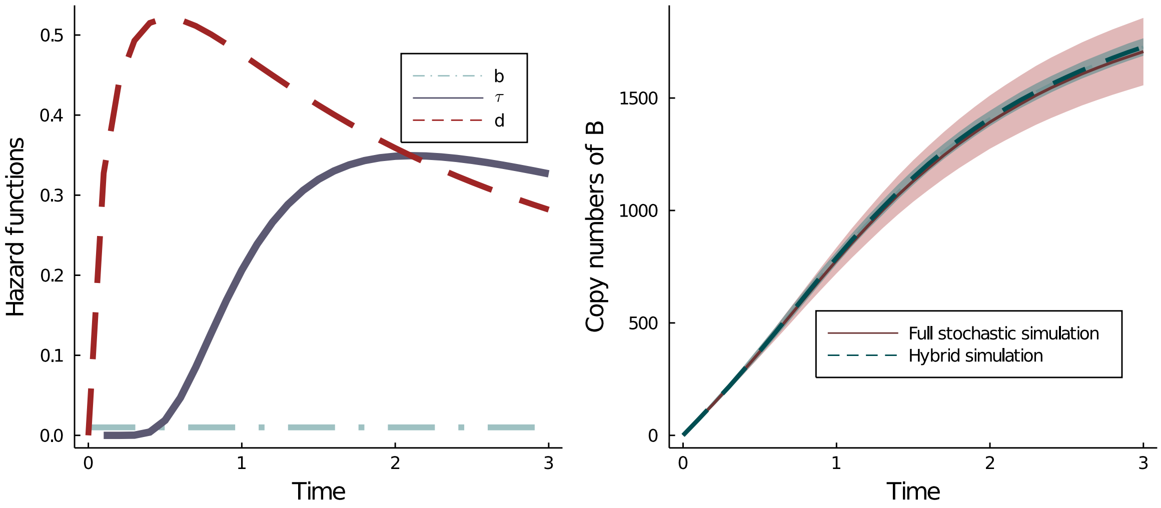

Figure 3. An example of the hybrid simulation approach. (Left) The shapes of the three hazard functions the CRN described by Equation (2.1). (Right) Comparison of the hybrid simulation algorithm (Algorithm 2.2) with the full stochastic simulation algorithm (Algorithm 2.1). Here, the birth rate $b = 0.01$. The distributions characterized by τ and d are inverse gamma distribution with parameters $(1.75, 4.25)$ and a Beta prime distribution with parameters $(1.75, 1.25)$. The value of n in this example is 5000. The full stochastic simulation took 305.714216 seconds, while the hybrid simulation took only 62.093832 seconds on a 2.3 GHz 18-Core Intel Xeon W machine.

which, of course, vanishes in the limit of $n \to \infty$. Moreover, the FPTs can be approximated by a random variable following an exponential distribution with mean m, whenever n is sufficiently large. It follows that we can use a simple likelihood function (based on the exponential distribution) for the purpose of statistical inference of the underlying parameters, provided we have observations on the FPTs. This method, called dynamic survival analysis, of estimating parameters based on timings rather than counts was recently explored in the context of an epidemiology in [34]. Dynamic survival analysis of general CRNs will be discussed elsewhere.

In Figure 2, we show the accuracy of this approximation when $n = 100$. The approximation appears to be reasonably accurate. More importantly, this suggests we might be able to devise an efficient simulation algorithm using such approximate results. We explore this idea next.

2.5. Fast hybrid simulation

Consider a situation when the species A is abundantly available at the beginning of the reaction. Naturally, we expect the PDE approximation to the age density of the species A to be quite accurate, even though there will be considerable stochastic fluctuations in the copy numbers of B, at least initially. However, if we approximate the age density of A by the limiting PDE, we can also approximate the initial growth of the B molecules by a Poisson process whose time-varying intensity is driven by the PDE. We use this idea to devise a hybrid simulation algorithm, which is, again, essentially an adaptation of the Doob–Gillespie's SSA in the sense that it only draws next reaction times from an exponential distribution whose mean depends on the solution to the PDE. For the sake of completeness, a pseudocode describing the idea is provided in Algorithm 2.2.

Algorithm 2.2 Pseudocode for the hybrid simulation algorithm

```
Require: $n$, $y_A$, $K$                        ▷ $K$: Maximum number of reactions
Ensure: $t_1, t_2, \ldots$                        ▷ Timings of creation of $B$ molecules

Set $t_0 = 0$
for $i = 1, 2, \ldots, K$ do
    Calculate $\Lambda = \left(\int_0^\infty n\tau(s) y_A(t_{i-1}, s)\,\mathrm{d}s\right)^{-1}$.
    if $0 < \Lambda < \infty$ then
        Draw an exponential random variable $T$ with mean $\Lambda$, i.e., $T \sim$ Exponential($\Lambda$)
        Set $t_i = t_{i-1} + T$
    else
        Stop and break loop
    end if
    Set $i = i + 1$.
end for
```

In Figure 3, we show the accuracy of the hybrid simulation algorithm. Expectedly, the hybrid simulation is considerably faster than the full stochastic simulation of the CRN in Equation (2.1). A more elaborate comparison of performance is shown in Figure 4. However, it is worth noting that the hybrid simulation algorithm, by design, will underestimate the variance in the counting process for the species B. Therefore, one should use the hybrid simulation when it suffices to get the mean trajectory accurately. Alternatively, one can borrow ideas to estimate the variance correctly in other simulation algorithms [49, 50, 51]. Similar ideas to expedite simulations have been proposed previously. For instance, Ganguly et al. [52] propose a jump-diffusion approximation to the stochastic CRNs and provide error analysis while others [28, 53] introduce hybrid simulation methods using a piecewise deterministic Markov process.

3. Michaelis–Menten enzyme-kinetic reactions

Michaelis–Menten enzyme-catalyzed chemical reactions form an important class of CRNs particularly because of their vast applications in the industry [55, 56]. Several descriptions of this class of reactions are available in the literature. For the sake of simplicity, in what follows we shall adopt the simplest form of the Michaelis–Menten enzyme-catalyzed reactions. In this form, the CRN comprises a reversible binding of a molecule of a substrate (S) and a molecule of an enzyme (E) into a molecule of a substrate-enzyme complex, and an irreversible conversion of a molecule of the complex into a molecule of a product (P) leaving the molecule of the enzyme free. That is, the

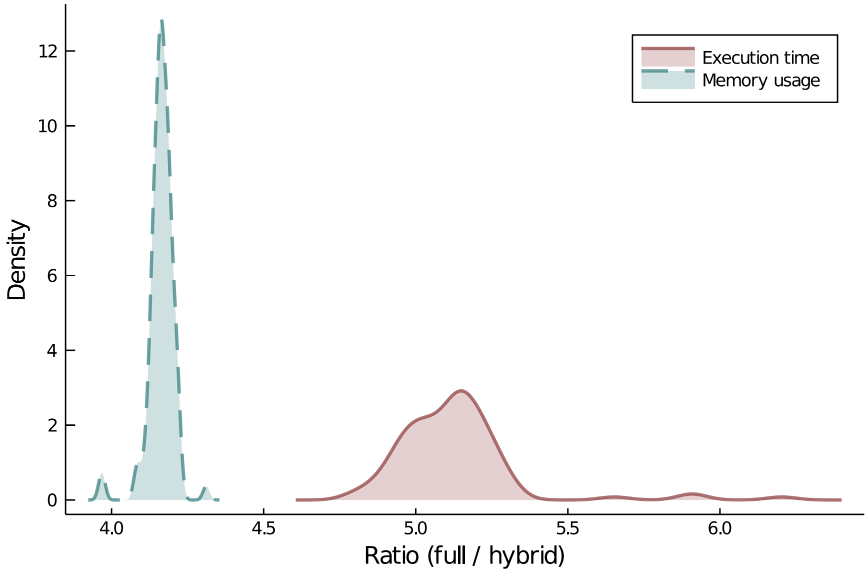

Figure 4. Efficiency of the hybrid simulation algorithm. The figure shows the empirical density of the ratios of execution times and memory usage of the full stochastic simulation and those of the hybrid simulation algorithm described in Algorithm 2.2. It is evident that the hybrid simulation algorithm is at least five times faster in terms of execution times and at least four times more efficient in terms of memory usage. The simulation set-up is the same as Figure 3. The performance evaluation of the hybrid simulation is done using the *BenchmarkTools.jl* package [54] in Julia language [46]

system consists of the following reactions:

$$\begin{aligned} &E + S \longrightarrow^{k_1} C, \\ &C \qquad \longrightarrow^{k_{-1}} E + S, \\ &C \qquad \longrightarrow^{k_2} P + E. \end{aligned} \tag{3.9}$$

In traditional models of enzyme kinetics, the quantities k_1, k_{-1}, and k_2 are reaction rate constants. When modeled stochastically using a CTMC, the mean-field limit of the scaled concentrations is described by the following set of ODEs (see [57] for more details):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}[E] &= -k_1[E][S] + (k_{-1} + k_2)[C], \\ \frac{\mathrm{d}}{\mathrm{d}t}[S] &= -k_1[E][S] + k_{-1}[C], \\ \frac{\mathrm{d}}{\mathrm{d}t}[C] &= k_1[E][S] - (k_{-1} + k_2)[C], \\ \frac{\mathrm{d}}{\mathrm{d}t}[P] &= k_2[C]. \end{aligned} \tag{3.10}$$

The $[\cdot]$ notation is used to denote the concentrations. The ODE system in Equation (3.10) has been studied extensively in the literature. We will adopt our measure-valued representation to incorporate potential age structure in the Michaelis–Menten CRN.

3.1. Enzyme kinetics with age structure

We assume the binding reaction depends on the age of the participating molecule of the enzyme. That is, only k_1 is age-dependent; k_{-1} and k_2 are constants. For the species E, S, C, and P, define the measure-valued stochastic processes

$$X_t^E := \sum_{i=1}^{N_E(t)} \delta_{e_i(t)}, \quad X_t^S := \sum_{i=1}^{N_S(t)} \delta_{s_i(t)}, \quad X_t^C := \sum_{i=1}^{N_C(t)} \delta_{c_i(t)}, \quad X_t^P := \sum_{i=1}^{N_P(t)} \delta_{p_i(t)},$$

where N_E, N_S, N_C, N_P denote the counts of molecules of E, S, C, and P respectively. Similarly, e_i, s_i, c_i, p_i denote the age of the i-th molecule of E, S, C, and P respectively. The process $X := (X^E, X^S, X^C, X^P)$ is a Markov process on the space $D([0,T], \mathcal{M}_P(\mathbb{R}_+)^4)$. Please note that we need to scale the hazard function k_1 corresponding to the bimolecular reaction by n^{-1} following the stochastic law of mass actions [1].

As before, we are interested in the large-volume limit of the scaled process $n^{-1}X_t$. The scaled stochastic process $n^{-1}X_t$ converges to a deterministic function $x_t := (x_t^E, x_t^S, x_t^C, x_t^P)$ whose components $x_t^E, x_t^S, x_t^C, x_t^P$ are finite measures on $\mathbb{R}_+$ by virtue of the LLN.

Let y_E denote the density of the measure x_t^E with respect to the Lebesgue measure. Also, let y_S, y_C, y_P denote the concentrations of the S, C, and P molecules. Then, we get the following limiting system:

$$\begin{aligned}
(\partial_t + \partial_s)\, y_E(t,s) &= -\, k_1(s) y_E(t,s) y_S(t), \\
\frac{\mathrm{d}}{\mathrm{d}t} y_S(t) &= -\, y_S(t) \int_0^\infty k_1(s) y_E(t,s)\,\mathrm{d}s + k_{-1} y_C(t), \\
\frac{\mathrm{d}}{\mathrm{d}t} y_C(t) &= y_S(t) \int_0^\infty k_1(s) y_E(t,s)\,\mathrm{d}s - (k_{-1} + k_2) y_C(t), \\
\frac{\mathrm{d}}{\mathrm{d}t} y_P(t) &= k_2 y_C(t),
\end{aligned} \tag{3.11}$$

with the boundary condition

$$y_E(t,0) = (k_{-1} + k_2) y_C(t)$$

and the initial condition $y_E(0,s) = f_E(s)$ such that $\int_0^\infty f_E(s)\,\mathrm{d}s = [E_0]$. Appropriate initial conditions for S, C, and P are also assumed. This limiting system can now be used to study the effects of delay in the binding reaction. One interesting approximation that has been widely applied in the context of Michaelis–Menten enzyme kinetic reactions is what is known as a Quasi-Steady State Approximation [58]. There are many forms of QSSAs, namely, standard QSSA (sQSSA), total QSSA (tQSSA), and reversible QSSA (rQSSA). Detailed analysis of any of the QSSAs is beyond the scope of the present work. For the purpose of illustration, we informally describe an analogue of the sQSSA here.

3.2. The standard QSSA

The QSSAs are a multiscale approximation of the Michaelis–Menten enzyme-kinetic reactions. The basic assumption behind the standard QSSA is that the substrate-enzyme complex C reaches its steady-state much quicker than the other species. In the deterministic set-up, the approximation is achieved by setting $\frac{\mathrm{d}}{\mathrm{d}t} y_C(t) = 0$ in Equation (3.11), which allows one to work with a smaller system of ODEs. Several conditions for the validity of the sQSSA have been proposed in the literature. See [57] for a detailed discussion.

Following the deterministic approach in our case, we set $\frac{\mathrm{d}}{\mathrm{d}t} y_C(t) = 0$ in Equation (3.11) to get a reduced PDE system that is analogous to the sQSSA. To be more precise, $\frac{\mathrm{d}}{\mathrm{d}t} y_C(t) = 0$ yields

$$y_C(t) = \frac{y_S(t) \int_0^\infty k_1(s) y_E(t,s)\,\mathrm{d}s}{k_{-1} + k_2},$$

which further yields an approximate system

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} y_S(t) &= -\frac{k_2}{k_{-1}+k_2} y_S(t) \int_0^\infty k_1(s) y_E(t,s)\,\mathrm{d}s, \\ \frac{\mathrm{d}}{\mathrm{d}t} y_P(t) &= \frac{k_2}{k_{-1}+k_2} y_S(t) \int_0^\infty k_1(s) y_E(t,s)\,\mathrm{d}s. \end{aligned} \tag{3.12}$$

Recall that y_E solves $(\partial_t + \partial_s)\, y_E(t,s) = -k_1(s) y_E(t,s) y_S(t)$ with boundary condition $y_E(t,0) = (k_{-1}+k_2) y_C(t)$ and initial condition $y_E(0,s) = f_E(s)$. As a consequence, y_E is determined by y_S and y_C, and can be partially solved in terms of y_S and y_C. Therefore, the reduced system of ODEs in Equation (3.12) is indeed autonomous and therefore, serves as an sQSSA of the CRN in Equation (3.9).

In the stochastic set-up, the QSSAs are obtained by means of the probabilistic multiscaling techniques developed in [3, 4]. The stochastic and the deterministic QSSAs mostly agree with each other with some notable differences. Please see [57] for examples of discrepancies as well as more details on the methods. Here, for paucity of space, we do not consider the stochastic QSSAs or possible discrepancies between stochastic and deterministic methods in the present age-structured models.

4. Prokaryotic auto-regulation

As another example, we consider a simple genetic network with feedback. We apply our approach using an age-dependent measure-valued process to build a model for a simple prokaryotic auto-regulation with a time delay. We modify an auto-regulation mechanism in the prokaryote gene network in [59] (Section 1.5.7). We simplify the example by approximating transcription and translation as a one-step process with a time delay and replacing repression of the gene by a protein dimer to repression by a single protein instead. For other related examples for the gene transcription and translation, see Section 2.1.1 in [60] and [61, 62, 63, 64].

Consider a genetic network with a gene (G), a protein (P), and a gene-protein complex (C). The gene activates production of protein following a hazard function b_P and the protein degrades following a hazard function d_P. The protein can reversibly bind with the gene to form a complex with binding hazard b_C and unbinding hazard d_C.

Since the gene-protein complex cannot participate in the production of protein, this is auto-regulation of the gene by its complex. Schematically, the reactions are as follows:

$$\begin{aligned} G \quad &\longrightarrow^{b_P} P+G, \\ P+G &\longrightarrow^{b_C} C, \\ C \quad &\longrightarrow^{d_C} P+G, \\ P \quad &\longrightarrow^{d_P} \emptyset. \end{aligned} \tag{4.13}$$

In (4.13), we assume that the age of the gene is important. Therefore, the hazard functions b_P and b_C are assumed to be age-dependent whereas d_C and d_P are assumed to be constants. Note that after unbinding of the gene-protein complex, the age of the gene is reset to zero. On the other hand, the age of the gene is not affected by the protein production.

Denote by $N_G(t)$, $N_P(t)$, and $N_C(t)$ the total molecular counts of the gene, the protein, and the gene-protein complex at time t, respectively. For the species G, P, and C, define the measure-valued processes

$$X_t^G := \sum_{i=1}^{N_G(t)} \delta_{g_i(t)}, \quad X_t^P := \sum_{i=1}^{N_P(t)} \delta_{p_i(t)}, \quad X_t^C := \sum_{i=1}^{N_C(t)} \delta_{c_i(t)},$$

where we denote the age of the i-th molecule of the species G, P, and C by g_i, p_i, and c_i respectively. As in the case of the Michaelis–Menten enzyme kinetic reaction, we scale the hazard function b_C corresponding to the bimolecular reaction by n^{-1} following the stochastic law of mass actions [1].

The LLN limit of the scaled process $n^{-1}X_t := (n^{-1}X_t^G, n^{-1}X_t^P, n^{-1}X_t^C)$ can be derived following by now familiar arguments of the previous examples. As one would expect, the scaled process $n^{-1}X_t$ converges to a deterministic function $x_t := (x_t^G, x_t^P, x_t^C)$ whose components are finite measures on $\mathbb{R}_+$. Since we assume only the age of the gene is important, we consider the limiting age density y_G of the gene, which we obtain as the density, when it exists, of the measure x_t^G with respect to the Lebesgue measure. Similarly, define the limiting concentrations of the product y_P and the complex y_C. The limiting system is then described by

$$\begin{aligned} (\partial_t + \partial_s)\, y_G(t,s) &= -b_C(s)\, y_G(t,s) y_P(t), \\ \frac{\mathrm{d}}{\mathrm{d}t} y_P(t) &= \int_0^\infty b_P(s) y_G(t,s)\,\mathrm{d}s - y_P(t) \int_0^\infty b_C(s) y_G(t,s)\,\mathrm{d}s \\ &\quad + d_C\, y_C(t) - d_P\, y_P(t), \\ \frac{\mathrm{d}}{\mathrm{d}t} y_C(t) &= y_P(t) \int_0^\infty b_C(s) y_G(t,s)\,\mathrm{d}s - d_C\, y_C(t), \end{aligned} \tag{4.14}$$

with the boundary condition

$$y_G(t,0) = d_C\, y_C(t)$$

and the initial condition $y_G(0,s) = f_G(s)$, which specifies the initial ages of the gene. Note that the hazard function for unbinding of the gene-protein complex appears in the boundary condition since we assumed that the age of the gene is reset to zero when the complex breaks into the gene and the protein. Also, recall that $b_P(s)$ encodes a time delay in transcription and translation. For example, we may set $b_P(s) = r1_{[\tau,\infty)}(s)$, which asserts that protein is produced only when the age of the gene is greater than τ with a hazard function r.

5. Discussion

Many biological processes with time delays, including CRNs, cannot be directly modeled using CTMCs due to non-exponentially distributed inter-event times of the processes. The simulation and analysis of systems with an age structure and time delays become challenging since the system dynamics are affected by the inherent randomness (stochasticity) as well as time delays. One way to simulate such stochastic systems with age structure and time delays is to modify simulation algorithms for CTMC models where the next reaction time and type are determined based on molecule counts of reactants. Bratsun et al. [39], Barrio et al. [65] and Cai [66] constructed modified SSAs, while Anderson [67] introduced a modified next reaction method to simulate discrete stochastic chemical reaction networks with delays. Notably, all of those works assume that the time lags in the delayed reactions are constant. Furthermore, in [68], Caravagna and Hillston described a non-Markovian stochastic process algebra, called Bio-PEPAd, to incorporate deterministic delays and perform formal analysis. Mura et al. [69] described how general holding time distributions can be incorporated in the programming language BlenX and studied the effect of the choice of the reaction time distributions. A stochastic simulation algorithm for non-Markovian biochemical reactions based on constraint programming is presented in [70].

CRNs with an age structure and random time delays provide a more realistic description of stochastic biophysical or chemical systems compared to the ones with fixed time delays. Unfortunately, the literature on stochastic systems with random time delays remains sparse. In a previous work by Koyama (Chapter 4 in [40]), the author investigated a stochastic kinetic network with a random time delay where a delayed reaction can be interrupted by another reaction and can fail to complete. In another work by Marquez-Lago et al. [71], the authors utilized probability distributed time delays to incorporate spatial effects such as diffusion or translocation of molecules in temporal stochastic models. In a recent work by Choi et al. [27], the authors described protein production in transcription and translation as a birth and death process with a random time delay.

In this paper, we developed a new way to incorporate an age structure and time delays in CRNs using age-dependent processes. We availed ourselves of previous theoretical works [36, 37, 43, 44] designed to study age-dependent population dynamics. We applied those stochastic models in the context of CRNs to account for the non-Markovian property due to the time delays. The use of age-dependent hazard functions not only enables us to model age-dependent time delays or reaction rates but also covers the modeling of constant and random time delays in the existing literature. We illustrated our method using simple biophysical systems in gene regulation and enzyme kinetics, but it will easily apply to general CRNs.

One potential disadvantage of the age-dependent processes is that simulation can be prohibitive since the age of each individual molecule of the chemical species of interest needs to be tracked over the entire simulation time. Therefore, we derived a large-volume limit of the age-dependent process for CRNs in the form of PDEs using the analytic methods in [36, 37, 43, 44] and used the PDE limit to construct a hybrid simulation algorithm, which, in our example, turned out to be five times faster than the full stochastic simulation. Moreover, we approximated a Mean First Passage Time efficiently utilizing the theoretical limit.

In this work, we emphasized how age-structured processes and their large-volume limits can be applied to model CRNs, in particular, biophysical or chemical systems with time delays. Many previous findings for general CRNs under Markovian assumption can be reinvestigated and extended to non-Markovian settings using age-structured processes. It would be interesting to see how the long time behavior of stochastic CRNs is affected by incorporating age structure. For example, it would be interesting to study stationary distributions of autocatalytic CRNs with switching behavior [72], to identify a class of CRNs maintaining product-form Poisson distributions for all times [73] and to find when CRNs show nonexplosive behavior [74]. Another interesting direction will be to study stability of CRNs [75] and to estimate transition times between different attractors in CRNs [76].

For the sake of simplicity, we have assumed in this paper that the molecular entities of all chemical species are abundant at the same order of magnitude so as to obtain the large-volume limit under the classical scaling. A natural extension of this work is to consider general CRNs with a wide range of molecular abundances and reaction rates where we can apply multiscale approximations to reduce model complexity [1, 3, 77]. We leave such investigation to future work. In this paper, we briefly described how an analogue of QSSA can be derived in the Michaelis-Menten enzyme-kinetic reactions. As shown in the related previous work [57, 58], both deterministic and stochastic QSSAs can be revisited with an extension of our approach to multiscale approximations in enzyme kinetics under non-Markovian setting. Another promising application of our approach seems to be in parameter inference and survival analysis of general CRNs with age structure. Given the current interests in pandemic modeling, such CRNs could lead to interesting examples in population dynamics and epidemiology. We hope to be able to pursue such work in the near future.

We conclude our discussion by briefly mentioning a class of CRNs modeled using Poisson processes with time-varying intensities. While retaining the Markov property, time-varying intensities provide a flexible way to aggregate out unobserved processes and to account for heterogeneity in the system such as cell-to-cell variability, changes in the volume or temperature of a cell affecting reaction rates [78, 67, 79]. However, the crucial difference between those models and ours is that time-varying intensities alone cannot induce a dependence structure of time delays on the initiation times of reactions whereas introduction of an age structure can. This is because time-varying intensities are a property of the system, whereas the age is a property of the individual molecule. Therefore, making the intensities depend explicitly on the individual ages of the molecules, as we do in this paper, provides a richer class of models.

A. Table of symbols

Symbol	Meaning
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of reals
$\mathbb{R}_+$	The set of non-negative reals
$1_{\{A\}}(x)$	Indicator (characteristic) function of the set A
δ_x	Dirac delta function at x
$\mathcal{B}(A)$	The Borel σ-field of subsets of a set A
$\mathcal{M}_P(E)$	The space of finite point measures on the set E
$D([0,T],E)$	The space of E-valued càdlàg functions defined on $[0,T]$
$\langle \mu, f \rangle$	The integral $\int f \, \mathrm{d}\mu$

B. Acronyms

CDF	Cumulative Distribution Function
CRN	Chemical Reaction Network
CTMC	Continuous Time Markov Chain
FPT	First Passage Time
LLN	Law of Large Numbers
MFPT	Mean First Passage Time
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PDF	Probability Density Function
PPM	Poisson Point Measure
QSSA	Quasi-Steady State Approximation
rQSSA	reversible QSSA
sQSSA	standard QSSA
SSA	Stochastic Simulation Algorithm
tQSSA	total QSSA

C. Preliminaries

For the sake of completeness, we briefly describe some statistical and mathematical preliminaries here. Consider a continuous random variable U taking nonnegative values with Cumulative Distribution Function (CDF) G_U and Probability Density Function (PDF) g_U. The *survival function* S_U of the random variable U is defined as

$$S_U(t) := \mathsf{P}\left(U > t\right) = 1 - G_U(t). \tag{C.15}$$

The *hazard function* h_U of the random variable U is defined as

$$h_U(t) := \frac{g_U(t)}{S_U(t)}. \tag{C.16}$$

Hazard and survival functions are extensively used in survival analysis to model *time to event* data, *e.g.*, time to death, time to hospitalization, time to default, time to failure etc. Intuitively, the hazard function describes the probability of failure in an infinitesimally small time period $(t, t + \Delta t)$ given survival till time t. With little application of calculus, one can see that

$$h_U(t) = \lim_{h\to 0} \frac{\mathsf{P}\left(t < U < t + h \mid U > t\right)}{h} = -\frac{\mathrm{d}}{\mathrm{d}t} \log S_U(t),$$

which yields another useful relationship between the hazard function and the survival function:

$$S_U(t) = \exp\left(-\int_0^t h_U(u)\,\mathrm{d}u\right) = \exp\left(-\Lambda_U(t)\right),$$

where $\Lambda_U(t) := \int_0^t h_U(u)\,\mathrm{d}u$ is called the *cumulative hazard function.* Hazard and survival functions cannot always be obtained in closed form. Probability distributions for which we can obtain them in closed form include Weibull, exponential, log-logistic distributions. The case of exponential distribution is unique in that it is the only probability distribution for which the hazard function is constant. However, a constant hazard is unrealistic in models for many biophysical systems.

D. Brief derivation of the PDE limit

In this section, we provide a brief, intuitive derivation of the PDE limit mentioned in Section 2.3. The line of argument follows the standard tightness-uniqueness route for abstract Markov processes and has been used in several prior works [36, 45, 37, 43, 44]. A rigorous proof of convergence for a general class of non-Markovian CRNs will be discussed elsewhere.

Consider the CRN in Equation (2.1) with the measure-valued process X_t as defined in Section 2.3. The components X_t^A, X_t^B satisfy the trajectory equations given in Equations (2.4) and (2.5). In order to study moments and martingale properties of

X_t^A and X_t^B, it is worthwhile to check that

$$\begin{aligned}
\langle X_t^A, f_t\rangle &= \sum_{k=1}^{N_A(0)} f_t(t+\sigma_k(X_0^A)) + \int_0^t \int_0^\infty f_t(t-s)\, 1_{\{\theta\le b\}}\, Q_1(ds, d\theta)\\
&\quad - \int_0^t \int_{\mathbb{N}} \int_0^\infty f_t(t-s+\sigma_i(X_{s-}^A))\, 1_{\{i\le N_A(s-)\}}\, 1_{\{\theta\le \tau(\sigma_i(X_{s-}^A))\}}\, Q_2(ds, di, d\theta)\\
&\quad - \int_0^t \int_{\mathbb{N}} \int_0^\infty f_t(t-s+\sigma_i(X_{s-}^A))\, 1_{\{i\le N_A(s-)\}}\, 1_{\{\theta\le d(\sigma_i(X_{s-}^A))\}}\, Q_3(ds, di, d\theta),\\
\langle X_t^B, f_t\rangle &= \int_0^t \int_{\mathbb{N}} \int_0^\infty f_t(t-s)\, 1_{\{i\le N_A(s-)\}}\, 1_{\{\theta\le \tau(\sigma_i(X_{s-}^A))\}}\, Q_2(ds, di, d\theta),
\end{aligned}$$

for a sufficiently large class of test functions $f : (a, s) \to f_s(a)$.

As in the case of standard Markov model in Section 2.1, we are now interested in the large-volume limit ($n \to \infty$) of the scaled stochastic process $n^{-1}X_t$. By virtue of the LLN, if we assume i) the hazard functions are continuous, ii) the global jump rates are bounded above by a finite quantity, iii) a finite second moment condition on the initial population size $\sup_n \mathsf{E}\,[n^{-2}N_A(0)^2] < \infty$, and iv) the initial age distribution does not explode, we have that the scaled process $n^{-1}X_t$ converges to a deterministic function $x_t := (x_t^A, x_t^B)$ whose components x_t^A and x_t^B are themselves measure-valued functions. This can be formally justified by verifying that the sequence of processes $n^{-1}X_t$ is tight and then, showing that the limit points (along subsequences) are unique. We can identify the limit points by studying certain martingale processes associated with the scaled processes $n^{-1}X_t$. Outline of the argument is provided below.

D.1. Martingale property and tightness-uniqueness

First, under the above mentioned assumptions, we can show that the components of the scaled process $n^{-1}X_t$ do not explode (similar derivation in [37, Proposition 2.7]). Now, note that the trajectory equations for the processes X_t^A and X_t^B given in Equations (2.4) and (2.5) are driven by PPMs. Since we have

$$f_t(a+t-s) = f_s(a) + \int_s^t \left(\frac{\partial}{\partial u} f_u(a+u-s) + \frac{\partial}{\partial a} f_u(a+u-s)\right) \mathrm{d}u,$$

and using the compensated PPMs of the PPMs Q_1, Q_2, Q_3, we can show the processes

$$\begin{aligned}
M_t^{A,f} &= \langle n^{-1}X_t^A, f_t\rangle - \langle n^{-1}X_0^A, f_0\rangle\\
&\quad - \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) - f_s(a)(\tau(a)+d(a))\right) n^{-1}X_s^A(\,\mathrm{d}a)\,\mathrm{d}s\\
M_t^{B,f} &= \langle n^{-1}X_t^B, f_t\rangle\\
&\quad - \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) + f_s(0)\tau(a)\right) n^{-1}X_s^A(\,\mathrm{d}a)\,\mathrm{d}s
\end{aligned}$$

are zero mean, square integrable, càdlàg martingale processes with predictable quadratic variations of the order n^{-1}. Since we expect the predictable quadratic variations to vanish in the limit of $n \to \infty$, the scaled process $n^{-1}X_t$ converges to a deterministic, continuous function x_t. The tightness of the process $n^{-1}X_t$ can be established by verifying a criterion due to Roelly [80] in the vague topology and the Aldous–Rebolledo criteria [81]. See [36] or [37, Proposition 3.1] for similar calculations. Furthermore, thanks to the martingale representations above, we expect the limit x_t to satisfy

$$\begin{aligned}\langle x_t^A, f_t\rangle &= \langle x_0^A, f_0\rangle \\ &\quad + \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) - f_s(a)(\tau(a) + d(a))\right) x_s^A(\,\mathrm{d}a)\,\mathrm{d}s \\ \langle x_t^B, f_t\rangle &= \int_0^t \int_0^\infty \left(\frac{\partial}{\partial a} f_s(a) + \frac{\partial}{\partial s} f_s(a) + f_s(0)\tau(a)\right) x_s^A(\,\mathrm{d}a)\,\mathrm{d}s.\end{aligned}$$

The uniqueness of the solutions can be shown by first establishing that the solutions remain bounded on finite time intervals (recall the global jump rates are assumed bounded) and then invoking Grönwall's lemma to show the distance between two possible solutions must vanish proving the desired uniqueness.

E. Software

The numerical results in this paper are obtained by the Julia programming language [46]. The Julia scripts (compatible with version 1.4.1) used in this paper have been made available publicly at a dedicated GitHub repository [38].

Funding

WKB was supported by the National Institute of Allergy and Infectious Diseases (NIAID) grant R01 AI116770, the National Science Foundation (NSF) grant DMS-2027001 and the Ohio State University's President's Postdoctoral Scholars Program (PPSP). EK was supported by NIAID grant R01 AI116770 and the NSF grants DMS-2027001 and DMS-1853587. GAR was supported by the NSF grants DMS-2027001 and DMS-1853587. HWK was supported in part by the NSF under the grant DMS-1620403. The project was initiated when HWK was visiting the Mathematical Biosciences Institute (MBI) at the Ohio State University in Summer 2019. The authors acknowledge the hospitality and the support of MBI. The content of this manuscript is solely the responsibility of the authors and does not represent the official views of NSF, NIGMS, NIAID, or NIH.

References

[1] Ball K, Kurtz T G, Popovic L and Rempala G 2006 *The Annals of Applied Probability* **16** 1925–1961

[2] Anderson D F and Kurtz T G 2011 Continuous Time Markov Chain Models for Chemical Reaction Networks *Design and Analysis of Biomolecular Circuits* ed Koeppl H, Setti G, di Bernardo M and Densmore D (Springer) pp 3–42

[3] Kang H W and Kurtz T G 2013 *The Annals of Applied Probability* **23** 529–583

[4] Kang H W 2012 *BMC systems biology* **6** 143

[5] Gillespie D T 1976 *Journal of computational physics* **22** 403–434

[6] Gillespie D T 1977 *The journal of physical chemistry* **81** 2340–2361

[7] Gillespie D T 2007 *Annu. Rev. Phys. Chem.* **58** 35–55

[8] Gibson M A and Bruck J 2000 *The journal of physical chemistry A* **104** 1876–1889

[9] Palangat M, Meier T I, Keene R G and Landick R 1998 *Molecular cell* **1** 1033–1042

[10] Hoyle N P and Ish-Horowicz D 2013 *Proceedings of the National Academy of Sciences* **110** E4316–E4324

[11] Swinburne I A and Silver P A 2008 *Developmental cell* **14** 324–330

[12] Baron J W and Galla T 2019 *Journal of the Royal Society Interface* **16** 20190436

[13] Zhou Y, Gwadry F G, Reinhold W C, Miller L D, Smith L H, Scherf U, Liu E T, Kohn K W, Pommier Y and Weinstein J N 2002 *Cancer Research* **62** 1688–1695

[14] Herz A, Bonhoeffer S, Anderson R M, May R M and Nowak M A 1996 *Proceedings of the National Academy of Sciences* **93** 7247–7251

[15] Bai F, Huff K E and Allen L J 2019 *Journal of biological dynamics* **13** 47–73

[16] Fraser C, Riley S, Anderson R M and Ferguson N M 2004 *Proceedings of the National Academy of Sciences* **101** 6146–6151

[17] Kuang Y 1993 *Delay differential equations: with applications in population dynamics* (Academic press)

[18] Giang D V, Lenbury Y and Seidman T I 2005 *Journal of mathematical analysis and applications* **305** 631–643

[19] van den Berg B, Wain R, Dobson C M and Ellis R J 2000 *The EMBO journal* **19** 3870–3875

[20] van Meerten D, Girard G and van Duin J 2001 *Rna* **7** 483–494

[21] Easterby J S 1981 *Biochemical Journal* **199** 155–161

[22] Kekenes-Huskey P M, Eun C and McCammon J 2015 *The Journal of chemical physics* **143** 09B601_1

[23] Monk N A 2003 *Current Biology* **13** 1409–1413

[24] Korenčič A, Bordyugov G, Košir R, Rozman D, Goličnik M and Herzel H 2012 *PloS one* **7**

[25] Hinch R and Schnell S 2004 *Journal of mathematical chemistry* **35** 253–264

[26] Barrio M, Leier A and Marquez-Lago T T 2013 *The Journal of chemical physics* **138** 104114

[27] Choi B, Cheng Y Y, Cinar S, Ott W, Bennett M R, Josić K and Kim J K 2020 *Bioinformatics* **36** 586–593

[28] Hepp B, Gupta A and Khammash M 2015 *The Journal of chemical physics* **142** 034118

[29] Valleriani A, Ignatova Z, Nagar A and Lipowsky R 2010 *EPL (Europhysics Letters)* **89** 58003

[30] Deneke C, Lipowsky R and Valleriani A 2013 *PloS one* **8**

[31] Deneke C 2012 *Ph. D. Thesis*

[32] Prieto S, Bernard J and Scheffler I E 2000 *Journal of Biological Chemistry* **275** 14155–14166

[33] Eisen T J, Eichhorn S W, Subtelny A O, Lin K S, McGeary S E, Gupta S and Bartel D P 2020 *Molecular Cell*

[34] KhudaBukhsh W R, Choi B, Kenah E and Rempała G A 2020 *Interface Focus* **10** 20190048

[35] Calderazzo S, Brancaccio M and Finkenstädt B 2019 *Bioinformatics* **35** 1380–1387

[36] Fournier N and Méléard S 2004 *The Annals of Applied Probability* **14** 1880–1919 ISSN 1050-5164

[37] Tran V C 2008 *ESAIM. Probability and Statistics* **12** 345–386 ISSN 1292-8100

[38] KhudaBukhsh W R, Kang H W, Kenah E and Rempała G 2020 DelayModel: A Julia implementation of age-structured stochastic chemical reaction networks GitHub repository URL `https://github.com/wasiur/DelayModel`

[39] Bratsun D, Volfson D, Tsimring L S and Hasty J 2005 *Proceedings of the National Academy of Sciences* **102** 14593–14598

[40] Koyama M 2013 *Ph. D. Thesis*

[41] Kurtz T G 1970 *Journal of applied Probability* **7** 49–58

[42] Kurtz T G 1978 *Stochastic Processes and their Applications* **6** 223–240

[43] Ferrière R and Tran V C 2009 Stochastic and deterministic models for age-structured populations

with genetically variable traits *CANUM 2008* (*ESAIM Proc.* vol 27) (EDP Sci., Les Ulis) pp 289–310

[44] Méléard S and Tran V C 2012 *Stochastic Processes and their Applications* **122** 250–276

[45] Champagnat N, Ferrière R and Méléard S 2008 Individual-Based Probabilistic Models of Adaptive Evolution and Various Scaling Approximations *Seminar on Stochastic Analysis, Random Fields and Applications V* ed Dalang R C, Russo F and Dozzi M (Birkhäuser Basel) pp 75–113

[46] Bezanson J, Edelman A, Karpinski S and Shah V B 2017 *SIAM Review* **59** 65–98

[47] MacNamara S, Burrage K and Sidje R B 2008 *Multiscale Modeling & Simulation* **6**(4)

[48] Kim J, Dark J, Eciso G and Sindi S 2020 *arXiv preprint* URL `https://arxiv.org/pdf/2005.05503.pdf`

[49] Franz B, Flegg M B, Chapman S J and Erban R 2013 *SIAM Journal on Applied Mathematics* **73** 1224–1247

[50] Harrison J U and Yates C A 2016 *Journal of The Royal Society Interface* **13** 20160335

[51] Kang H W and Erban R 2019 *Bulletin of mathematical biology* **81** 3185–3213

[52] Ganguly A, Altintan D and Koeppl H 2015 *Multiscale Modeling & Simulation* **13**(4)

[53] Gupta A and Khammash M 2019 *Bulletin of Mathematical Biology* **81** 3121–3158

[54] Chen J and Revels J 2016 *arXiv e-prints* arXiv:1608.04295 (*Preprint* `1608.04295`)

[55] Cornish-Bowden A 2012 *Fundamentals of Enzyme Kinetics* 4th ed (Wiley)

[56] Segel I H 1975 *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems* (Wiley)

[57] Kang H W, KhudaBukhsh W R, Koeppl H and Rempała G 2019 *Bulletin of Mathematical Biology* **81** 13031336

[58] Eilertsen J and Schnell S 2020 *Mathematical Biosciences* 108339

[59] Wilkinson D J 2012 *Stochastic modelling for systems biology* 2nd ed (CRC press)

[60] Anderson D F and Kurtz T G 2015 *Stochastic analysis of biochemical systems* (Springer)

[61] Rempala G A, Ramos K S and Kalbfleisch T 2006 *Journal of theoretical biology* **242** 101–116

[62] Kim J K, Rempala G A and Kang H W 2017 *Multiscale Modeling & Simulation* **15** 1376–1403

[63] Cappelletti D, Majumder A P and Wiuf C 2019 *arXiv preprint arXiv:1912.00401*

[64] Cao Z and Grima R 2019 *Journal of The Royal Society Interface* **16** 20180967

[65] Barrio M, Burrage K, Leier A and Tian T 2006 *PLoS computational biology* **2**

[66] Cai X 2007 *The Journal of chemical physics* **126** 124108

[67] Anderson D F 2007 *The Journal of chemical physics* **127** 214107

[68] Caravagna G and Hillston J 2012 *Theoretical Computer Science* **419** 26 – 49

[69] Mura I, Prandi D, Priami C and Romanel A 2009 *Electronic Notes in Theoretical Computer Science* **253** 83 – 98 proceedings of Seventh Workshop on Quantitative Aspects of Programming Languages (QAPL 2009)

[70] Chiarugi D, Falaschi M, Hermith D, Olarte C and Torella L 2015 *BMC Systems Biology* **9**

[71] Marquez-Lago T T, Leier A and Burrage K 2010 *BMC systems biology* **4** 19

[72] Bibbona E, Kim J and Wiuf C 2020 *arXiv* arXiv–2001

[73] Anderson D F, Schnoerr D and Yuan C 2020 *Journal of Mathematical Biology* 1–33

[74] Anderson D F, Cappelletti D, Koyama M and Kurtz T G 2018 *Bulletin of mathematical biology* **80** 2561–2579

[75] Agazzi A and Mattingly J C 2018 *arXiv preprint arXiv:1810.06547*

[76] Agazzi A, Dembo A and Eckmann J P 2018 *Annals of Applied Probability* **28** 1821–1855

[77] Kang H W, Kurtz T G and Popovic L 2014 *The Annals of Applied Probability* **24** 721–759

[78] Lu T, Volfson D, Tsimring L and Hasty J 2004 *Systems Biology* **1** 121–128

[79] Chevalier M W and El-Samad H 2014 *The Journal of Chemical Physics* **141** 214108

[80] RoellyCoppoletta S 1986 *Stochastics* **17** 43–65

[81] Joffe A and Metivier M 1986 *Advances in Applied Probability* **18** 20–65