RESEARCH ARTICLE | MARCH 13 2013

Reduction of chemical reaction networks through delay distributions

Manuel Barrio; André Leier; Tatiana T. Marquez-Lago

J. Chem. Phys. 138, 104114 (2013)
https://doi.org/10.1063/1.4793982

Reduction of chemical reaction networks through delay distributions

Manuel Barrio,[1,a),b)] André Leier,[2,a),b)] and Tatiana T. Marquez-Lago[3,b)]
[1]*Departamento de Informática, Universidad de Valladolid, Valladolid, Spain*
[2]*Okinawa Institute of Science and Technology, Okinawa, Japan*
[3]*Integrative Systems Biology Unit, Okinawa Institute of Science and Technology, Okinawa, Japan*

(Received 29 October 2012; accepted 17 February 2013; published online 13 March 2013)

Accurate modelling and simulation of dynamic cellular events require two main ingredients: an adequate description of key chemical reactions and simulation of such chemical events in reasonable time spans. Quite logically, posing the right model is a crucial step for any endeavour in Computational Biology. However, more often than not, it is the associated computational costs which actually limit our capabilities of representing complex cellular behaviour. In this paper, we propose a methodology aimed at representing chains of chemical reactions by much simpler, reduced models. The abridgement is achieved by generation of model-specific delay distribution functions, consecutively fed to a delay stochastic simulation algorithm. We show how such delay distributions can be analytically described whenever the system is solely composed of consecutive first-order reactions, with or without additional "backward" bypass reactions, yielding an exact reduction. For models including other types of monomolecular reactions (constitutive synthesis, degradation, or "forward" bypass reactions), we discuss why one must adopt a numerical approach for its accurate stochastic representation, and propose two alternatives for this. In these cases, the accuracy depends on the respective numerical sample size. Our model reduction methodology yields significantly lower computational costs while retaining accuracy. Quite naturally, computational costs increase alongside network size and separation of time scales. Thus, we expect our model reduction methodologies to significantly decrease computational costs in these instances. We anticipate the use of delays in model reduction will greatly alleviate some of the current restrictions in simulating large sets of chemical reactions, largely applicable in pharmaceutical and biological research. [http://dx.doi.org/10.1063/1.4793982]

I. INTRODUCTION

One of the fundamental goals of Systems Biology is to understand complex interactions between components of biological systems. At the cellular level, such interactions give rise to specific biological functions such as gene expression, molecular transport, and cell signal transduction, and are typically represented by chemical reaction networks.

However, even when biological functions have been studied in scrutinizing detail and can be reliably represented by sets of chemical reactions, one may not be able to accurately simulate such networks, nor explore alternatives to wild-type scenarios. The reason for this is the associated computational costs, limiting the time spans in which phenomena can be simulated. Even when solely considering deterministic scenarios, the network of interactions can be very large, making the simulation of a system potentially unfeasible.

In fact, the large number of interactions is not the only bottleneck limiting computational efficiency. Chemical reactions are discrete stochastic events, and should be treated as such. They are deemed stochastic as it is impossible to say – with absolute certainty – the specific type of reaction that will happen during a prescribed time interval, or when or where such event is to occur. Hence, an accurate description of chemical kinetics can often solely happen in a probabilistic sense, prescribing rates of change between different states of the system.

If the state space is enumerated, one can define a linear ordinary differential equation (ODE) for the time evolution of the probability associated with each state of the system (as opposed to each molecular species in the system), and the set of all such ODEs compose the so-called chemical master equation (CME). Obtaining direct analytic solutions of the CME is possible, but limited to simplified scenarios with reduced applicability (cf. Sec. IV). In reality, the CME is sometimes studied through finite state projections (FSP),[1] and more generally by simulating exact trajectories with a stochastic simulation algorithm (SSA).[2,3] Naturally, either method can be prohibitively expensive, especially so when there are large numbers of molecules (and by consequence, distinct states of the system). Additionally, computational costs of SSAs are very large whenever the system has widely varying rate constants or increasing molecular populations. In such cases, simulating biological systems in adequate time spans may even be unfeasible. Thus, there is a great need to reduce networks of chemical reactions.

Well-known examples of model reductions are the Michaelis-Menten model for enzyme-substrate reactions,[4] the chemical lumping of reaction sequences,[5] and examples of one-reaction abridgment.[6]

[a)]Authors to whom correspondence should be addressed. Electronic addresses: mbarrio@infor.uva.es and andre.leier@oist.jp.
[b)]M. Barrio, A. Leier, and T. T. Marquez-Lago contributed equally to this work.

In particular, reduction methods exploiting separation of time scales have been developed. For instance, Mastny *et al.*[7] applied singular perturbation theory to remove highly reactive intermediates (quasi-steady state approximation (QSSA) species) in low numbers from the CME. Their method was successfully applied to reaction networks where all species occur only in small numbers but QSSA species are zero most of the time, and to those where non-QSSA species occur in large numbers while QSSA species populations are small. Recently, Thomas *et al.*[8] proposed a reduced linear Langevin equation, the so-called slow-scale linear noise approximation (ssLNA), describing the fluctuations in the slowly varying species only. The ssLNA follows rigorously from the LNA using the projection operator technique. It accurately describes stochastic dynamics of monostable biochemical networks, including bimolecular reactions, in conditions characterized by small intrinsic noise and time-scale separation – namely, those required for the QSSA as well.

However, there are different issues limiting such reductions. For instance, new reaction rates need to be derived using information from the original model. Moreover, the existence of a reduced model is not guaranteed, depending not only on the original network "topology" (a general ailment of reduction methods), but also on the value of the reaction rates and compliance with time-scale separation conditions (see Ref. 6 for a discussion on this issue). Furthermore, as has been previously discussed,[9] adopting a QSSA may mask important discrete stochastic effects (such as those observed in tight regulation scenarios), while a linear noise approximation can yield completely erroneous results due to closure of moments of the CME. Thus, a method that is independent of time-scale separation conditions would be preferable.

Model reduction can also be achieved through the use of time delays: chemical events are not instantaneous, and a succession of them can be explicitly defined as a "time lapse." Such description is particularly useful when we know the final effect of a complex process and we can estimate the time it takes to be completed. The key idea here is to replace chains of processes by equivalent delayed reactions that transform reactants into products after a predefined time delay. This has already been implemented in delay stochastic simulation algorithms.[10,11] However, one may ask: can such descriptions be accurate with the consideration of a constant delay or, as one may expect, are time delays actually random variables obeying specific probability distributions? Moreover, can arbitrary biochemical networks be replaced by delayed reactions, assuming one can find an appropriate rate constant and delay distribution?

In this work, we answer both questions, by introducing a new method of model reduction using delays in stochastic chemical kinetics. We study different systems of reversible reactions with the additional restriction that the chain of reactions must be finished by an irreversible reaction. The sequence of reversible reactions would work as a chemical "block" whereas the last irreversible reaction would be used as a "connector" between neighbouring blocks.[5,6] Such combination of blocks and connections allows us to study chemical systems with different degrees of complexity, under the premise that each block can be lumped into a single reaction with a rate constant and a delay distribution. Abridged systems can then be part of much larger systems with additional reactions outside these blocks, reducing computational costs of large systems simulations.

Our methodology provides an exact reduction in scenarios solely composed of unimolecular and/or backward bypass reactions, as the delay distributions can be derived analytically. For all other monomolecular reactions (constitutive creation, degradation, or forward bypass reactions), our methodology's accuracy can be tailored at will, as the delay distributions can be derived numerically, either in terms of first-passage time (SSA) runs or matrix exponentials for sampled time points. In these cases, the accuracy depends on the number of SSA simulations obtained for the first-passage distribution, or the number of time points at which the matrix exponential is calculated, respectively.

To the best of our knowledge, this is the first work on accurate model reduction through delays, and we show the salient effects of all presented abridged models are probabilistically equivalent to those of the corresponding complete models. Also, a major advantage of our methodology over other abridgment methods is that it does not rely on time-scale separation conditions. Hence, it is more universally applicable.

In what follows, we will show how to reduce chains of consecutive chemical reactions and systems including "backward" and "forward" bypass reactions, degradation of involved molecular species, and constitutive creation of intermediate species. We will highlight what types of systems can be reduced exactly by analytical and/or numerical means, and will illustrate how to deal with large chemical systems. In these cases, one may still obtain exact solutions or, alternatively, a good approximation through Arnoldi estimates, at much lower computational cost. Furthermore, we will apply our methodologies to a stochastic model of eukaryotic mRNA turnover[12] and discuss under which conditions even Michaelis-Menten reactions can be lumped with high accuracy. Finally, we discuss the current limitations of our approach with respect to the chemical reaction network topology, and possible extensions of our methodology.

II. RESULTS

Let us first assume a chain of $n-1$ consecutive reversible reactions finished by an irreversible reaction,

$$S_1 \underset{c_{1_r}}{\overset{c_1}{\rightleftarrows}} S_2 \underset{c_{2_r}}{\overset{c_2}{\rightleftarrows}} S_3 \underset{c_{3_r}}{\overset{c_3}{\rightleftarrows}} \dots S_{n-1} \underset{c_{n-1_r}}{\overset{c_{n-1}}{\rightleftarrows}} S_n \overset{c_n}{\rightarrow} S_{n+1} \qquad (1)$$

with non-zero $c_1, c_{1_r}, c_2 \dots .c_{n-1}$ reaction rates. The goal is to replace this system with a single reaction,

$$S_1 \overset{\tau^*}{\rightarrow} S_{n+1}, \qquad (2)$$

where $\tau*$ is a delay distribution. We want this "lumping" to be exact in the sense that S_{n+1} in (2) follows the same distribution as in (1). The abridged model will not include the intermediate species ($S_2 \dots S_n$) and, consequently, will not provide any information on their dynamics. Moreover, the dynamics of S_1 in (2) is not equivalent to that in (1), since there is a

modified representation of the reactions consuming and producing S_1 molecules.

An alternative reaction scheme is

$$S_0 \xrightarrow{c_0} S_1 \underset{c_{1_r}}{\overset{c_1}{\rightleftarrows}} S_2 \underset{c_{2_r}}{\overset{c_2}{\rightleftarrows}} S_3 \underset{c_{3_r}}{\overset{c_3}{\rightleftarrows}} \dots S_{n-1} \underset{c_{n-1_r}}{\overset{c_{n-1}}{\rightleftarrows}} S_n \xrightarrow{c_n} S_{n+1} \qquad (1')$$

with an additional non-zero reaction rate c_0. If we can replace system (1) with the abridged model (2), we can obviously replace (1′) with the single reaction

$$S_0 \xrightarrow{c_0,\tau^*} S_{n+1}. \qquad (2')$$

Since reaction scheme (1′) starts with an irreversible reaction, the dynamics of S_0 in (1′) and (2′) are now identical.

A. Consecutive reversible reactions

How long will it take one S_1 molecule to become a S_{n+1} molecule? We can formally model it through a *random walker* and the *first-passage* problem.[6,13] Suppose a random walker, initially at "position" S_1, can "jump" from one species to another. The transition probabilities are determined by the coefficients c_j, c_{j_r} so that c_j is the probability per unit time that, being at S_j, the walker jumps into S_{j+1}, while c_{j_r} is the equivalent from S_{j+1} to S_j.

The first-passage time corresponds to the total time of the overall transformation or, in other words, to the system delay. We will follow the approach of the *absorbing boundary*[13] due to the last irreversible reaction, trapping the walker in S_{n+1} once it has arrived. Let T_{n+1} be the time for a walker that started in S_1 at $t = 0$, to first reach S_{n+1}. The latter state is absorbing and, therefore, the cumulative distribution function (CDF) of T_{n+1} can be defined in terms of the more familiar *occupation* probability $p(S_{n+1}, t|S_1, 0.) = Prob\{T_{n+1} \leq t\}$, where $p(S_{n+1}, t|S_1, 0.)$ is the probability that the walker is at S_{n+1} at time t, having started in S_1 at $t = 0$.

From the master equation, we know that the probability density function (PDF) of the first-passage time to the *state* S_{n+1} is related to the occupation probability of S_n by

$$f_{T_{n+1}} = \frac{d}{dt} p\left(S_{n+1}, t\,|S_1, 0\right) = c_n\ p\left(S_n, t\,|S_1, 0\right). \qquad (3)$$

In order to obtain $p(S_n, t|S_1, 0.)$ (in shorter notation $p_n(t)$), we have to solve the n-dimensional differential equation

$$\frac{d}{dt} p\left(t\right) = A p\left(t\right),$$

where $p(t) = [p(S_1, t|S_1, 0)\ p(S_2, t|S_1, 0) \dots p(S_n, t|S_1, 0)]^T$ is the occupancy probability vector of the n sites, the walker can be at prior to reaching S_{n+1}, and A is the $n \times n$ rate (or transition) matrix of the system (cf. Sec. III A for a description of A for system (1)). Note that each species is considered to be an occupancy site for the walker that is known to be initially at site 1 (at time $t = 0$), so $p(0) = [1\,0 \dots 0]^T$. It is well known that the solution of this differential equation is the matrix exponential mapping the initial probability to the probability at time t, $p(t) = e^{At}p(0)$.

Here, we choose the Laplace transform as a solution method for the matrix exponential and obtain

$$e^{At} = \mathcal{L}^{-1}\{(sI - A)^{-1}\},$$

where $(sI - A)^{-1}$ is the so-called *resolvent* of A. It can be shown that A is an invertible matrix, and its eigenvalues are real and negative (cf. Secs. III A and III B). This is a crucial point for the resolution of the matrix exponential. Moreover, it can be shown (cf. Secs. III C and III D) that the probability distribution of the first-passage time for our system (1) is

$$f_{T_{n+1}} = e_1(t) \star e_2(t) \star \dots \star e_n(t),$$

where $e_k(t) = \overline{\lambda_k} e^{-\overline{\lambda_k} t}$ and $\overline{\lambda_k} = -\lambda_k$, the positive values of the eigenvalues of A. In other words, $f_{T_{n+1}}$ is the convolution of exponential distributions with parameters $\overline{\lambda_k}$. By sampling from this distribution, one can obtain characteristic delays that are subsequently fed into a DSSA implementation[10,11] to simulate the reduced system. The solution of this reduced system is exact, i.e., identical to the original one, but one is able to obtain such solution at a much lower computational cost.

To quickly illustrate this methodology, let us consider the system (1′) with n = 8, reaction rate constants $c_i = 1 (i = 0, 1, 3, 5, 7, 8)$, $c_i = 3 (i = 2, 4, 6)$, $c_{i_r} = 2 (i = 1, 3, 5, 7)$, and $c_{i_r} = 4 (i = 2, 4, 6)$ and initial condition $x_0 = 100$ and $x_i = 0$ for $i = 1 \dots 9$ (example 1). We now want to replace the full system with one reaction

$$S_0 \xrightarrow{c_0,\tau^*} S_9.$$

In the reduced system, the rate constant c_0 is identical to that of the complete system and the delay distribution is the convolution of exponential distributions with parameters equal to the absolute value of the eigenvalues of the associated rate matrix A. Results are shown in Figure 1. Simulating the reduced system was about 20× faster than simulating the full system – without any loss of accuracy.

B. Large chemical systems

Even in cases when the state space is very large, one can still aim at generating a delay distribution in a similar way to Sec. II A. In fact, a very close approximation to the exact solution of the CME might be obtained by only considering a reduced set of eigenvalues. Namely, we would like to obtain those with smallest absolute values, which in turn yield the largest contributions to the convolution of exponential distributions. The problem here is to obtain such eigenvalues in a very efficient manner.

If we consider the inverse of A (a positive-definite matrix by definition), we can obtain Hessenberg reductions though Arnoldi iterations and, by using standard methods, one can further obtain their eigenvalues. These approximate eigenvalues, commonly referred to as Arnoldi estimates or Ritz values, typically converge to those of the full matrix A (Ref. 14). In our case, due to the sparse and regular structure of the state reaction matrix A, we do not expect to encounter convergence issues. Moreover, by applying the Arnoldi iteration, we obtain eigenvalues near the edge of the spectrum of A, ordered by magnitude. Quite conveniently, the Arnoldi iteration can be stopped any time, yielding a desired number of eigenvalues only.

It is worth noting that, even when the full eigenvalue decomposition is desired, using the Arnoldi iteration for large

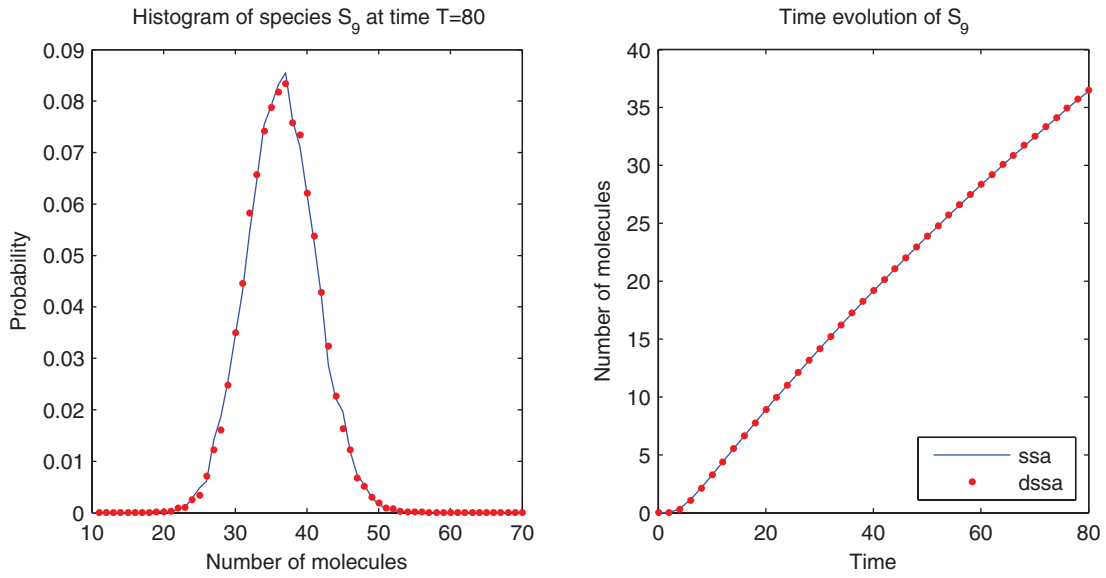


FIG. 1. Linear sequence of unimolecular reactions (example 1). (Left) Histogram for the number of S_9 molecules at time T = 80 obtained from 10^4 SSA (blue) and DSSA (red) simulations where delays were drawn from the 8-exponential iCDF with parameters $\overline{\lambda_k}$, $k = 1, \ldots, 8$ (the eigenvalues of the system's rate matrix A for species S_1–S_8). (Right) Average time evolution of species S_9 in steps of $\Delta t = 2$ until T = 80 for SSA and DSSA.

state reaction matrices will likely result in shorter computational time. The latter will further reduce computational costs when one needs to recalculate the delay distributions, due to specific signaling dynamics (i.e., a system with time-varying reaction rates). As explained in Ref. 15, the key point to consider is that matrix-vector multiplication is much faster than matrix-matrix multiplication. In this case, the matrix-vector multiplication corresponds to the Krylov subspaces spanned by the orthonormal basis obtained with the Arnoldi iteration.

To illustrate the applicability of the Arnoldi iteration, let us now consider system (1′) with n = 8, reaction rate constants $c_0 = c_1 = c_{1_r} = c_{3_r} = c_{4_r} = c_5 = c_{6_r} = c_7 = c_{7_r} = c_8 = 1$, $c_2 = c_4 = 0.1$, $c_{2_r} = c_{5_r} = 2$, $c_6 = 0.5$, and $c_3 = 10$, and initial condition $x_0 = 100$ and $x_i = 0$ for $i = 1 \ldots 9$ (example 2). We will now compare an approximation to the CME using only the smallest absolute eigenvalue, with the exact solution. As can be observed in Figure 2, the approximation stemming from the first Arnoldi estimate already yields a close approximation to the real solution. Generally, with increasing number of eigenvalues the approximation will approach the exact solution. However, in our example, the solution is largely dominated by the first eigenvalue and adding more eigenvalues does not significantly change the solution. Finally, we note that simulating the reduced system was about 70× faster than simulating the full system, without any loss of accuracy.

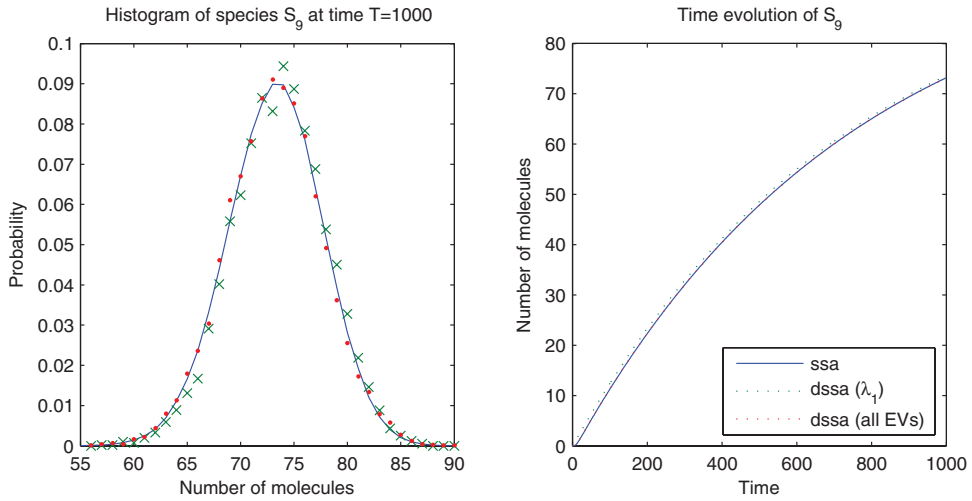


FIG. 2. Linear sequence of reactions with time-scale separation (example 2). (Left) Histogram for the number of S_9 molecules at time T = 1000 for 10^6 SSA simulations (blue) and 10^4 DSSA simulations where delays were drawn from the corresponding 8-exponential iCDF (red) or 1-exponential iCDF for the smallest absolute eigenvalue (green). (Right) Average time evolution of species S_9 in steps of $\Delta t = 10$ for SSA (blue) and DSSA (red/green). The blue and red trajectories are undistinguishable while the green trajectory is above the other two.

C. Fast reversible reactions with slow turnover

In order to show the potential speed-ups acquired with our methodology, we also abridged the following four-species system:

$$S_0 \xrightarrow{c_0} S_1 \underset{c_{1_r}}{\overset{c_1}{\rightleftarrows}} S_2 \xrightarrow{c_2} S_3$$

with fast reversible reactions between S_1 and S_2, and a slow production of the final product S_3 ($c_1, c_{1_r} \gg c_2$). For this scenario, the rate matrix A has the form $\begin{pmatrix} -c_1 & c_{1_r} \\ c_1 & -(c_{1_r}+c_2) \end{pmatrix}$. To simplify this scenario further, we assume $c_1 = c_{1_r}$. Thus, the eigenvalues of A can be explicitly stated as $\lambda_{1/2} = (-c_1 - \frac{1}{2}c_2) \pm \sqrt{c_1^2 + (\frac{1}{2}c_2)^2}$.

We ran simulations for several parameter sets (c_1, c_2) with $c_1 \gg c_2$, i.e., the reversible reactions were anywhere between 10 and 1 000 000 times faster than the reaction producing S_3. In all scenarios, the full and the abridged model lead to similar results for species S_0 and S_3 (data not shown) while DSSA simulations (over 100 runs) were up to 1.7×10^3 times faster than SSA simulations (for a simulation time T = 100). See Table I for a summary of speed-ups.

Evidently, it is the smallest absolute eigenvalue that mostly determines the delay distribution. In our scenario with two fast reversible reactions followed by a slower reaction, such eigenvalue is strongly determined by the rate of the latter reaction. In other words, since $|\lambda_1| < |\lambda_2|$ and $\lambda_1 \cong -\frac{1}{2}c_2$, for $c_1 \gg c_2$, the value of the smaller rate c_2 determines the delay distribution: the smaller the value of c_2, the larger the average delays drawn during DSSA simulations. Quite naturally, savings increase with decreasing values of the slow rate (see Table I; last three parameter sets). However, savings ultimately stem from a reduced number of reactions in the abridged model, as compared to the full model (minus the overhead due to delay management). Hence, savings increase alongside the number of times a random walker moves between the internal states (here: S_1 and S_2) prior to arriving at the absorbing state (here: S_3). Obviously, the time until absorption increases alongside values of c_1. Note that the average computational savings are limited by the maximum number of reactions that occur (on average) between the internal (lumped) states in a period of time. This explains why parameter set (1000, 0.1) leads to larger savings than parameter set (100, 0.0001), despite the latter having a larger c_1/c_2 ratio.

TABLE I. Computational savings in terms of average numbers of SSA reactions per single DSSA reaction and speed-up (runtime of SSA over runtime of DSSA runs). Simulations ran until T = 100. Mean values are calculated over 100 simulations.

Parameters ($\mathbf{c}_1$, $\mathbf{c}_2$)	Eigenvalues (λ_1, λ_2)	Avg. no. of DSSA reactions DSSA reaction	Avg. speed-up
(1,0.1)	(−0.049,−2.05)	2.30×10^2	∼2.5
(100,0.1)	(−0.05,−200.05)	1.98×10^4	$\sim 0.2 \times 10^3$
(1000,0.1)	(−0.05,−2000.05)	1.99×10^5	$\sim 1.7 \times 10^3$
(100,0.01)	(−0.005,−200.005)	7.81×10^4	$\sim 0.8 \times 10^3$
(100,0.001)	(−0.0005,−200.0005)	9.66×10^4	$\sim 1.0 \times 10^3$
(100,0.0001)	(−0.00005,−200.00005)	9.88×10^4	$\sim 1.1 \times 10^3$

D. Additional forward and backward bypass reactions

Let $S_j \xleftarrow{b} S_i$, $i - j > 1$, be the additional bypass reaction that converts species S_i back into S_j. As it can be shown, the reduction of the full system to a single reaction system with delays remains exact even when including such backward bypass reactions (cf. Sec. III E). As was true for systems with purely consecutive reversible reactions, the solution of the first-passage time problem is the convolution of exponentials with parameters $\overline{\lambda_k} = -\lambda_k$, the positive values of the eigenvalues of the corresponding rate matrix.

For instance, let us consider example 1 along with additional reactions

$$S_7 \xrightarrow{c_{7b}} S_5, \quad S_5 \xrightarrow{c_{5b}} S_3, \quad S_8 \xrightarrow{c_{8b}} S_4$$

with $c_{8b} = c_{7b} = c_{5b} = 1$ (example 3). By following the same procedure as with consecutive reactions, we can reduce the full system exactly (Figure 3). In this case, simulations of the reduced system are more than 40× faster than to simulate the full system.

However, the situation changes if one wants to consider reactions $S_j \xrightarrow{f} S_i$ $i - j > 1$ (a "forward" bypass). Such reactions change the structure of A significantly (with off-diagonal elements in the lower triangular matrix), and the first-passage solution no longer reduces to the convolution of exponentials with parameters $\overline{\lambda_k} = -\lambda_k$. In such cases, one can perform series of stochastic simulations to obtain a sample distribution of first-passage times that is then fed into the DSSA. Even though the full system can be reduced in such a way, the computational costs are often not very different from those of the full system. This is due to the large number of samples needed to generate a sufficiently smooth first-passage time distribution. However, one gains flexibility when trying to extend the model, as the abridged model with its delay distribution is essentially a module that can be "recycled," reducing the cost of all subsequent simulations significantly.

Let us illustrate this approach by considering the set of reactions from example 1 along with additional reactions

$$S_1 \xrightarrow{c_{1f}} S_5, \quad S_2 \xrightarrow{c_{2f}} S_4, \quad S_4 \xrightarrow{c_{4f}} S_6,$$

where $c_{1f} = c_{2f} = c_{4f} = 1$ (example 4). As can be observed, the solutions of the reduced model are indistinguishable from the exact solution for sufficiently large sample sizes (Figure 4). Figure 5 illustrates the dependence on first-passage time sample sizes for the complete system consisting of all forward and backward bypass reactions with parameters as described above (example 5). As is obvious, the larger the sample, the more accurate the delay distribution and the closer the reduced system is to the exact solution.

Importantly, an alternative to the first-passage time sampling method is the numerical evaluation of the matrix exponential $e^{\tilde{A}t}$ for various time points t. Here, $\tilde{A}$ is the $(n+1) \times (n+1)$ state matrix and includes the transition to S_{n+1}. The last entry of $e^{\tilde{A}t}\,p\,(0)$ corresponds to $F_{T_{n+1}}(t)$, the value

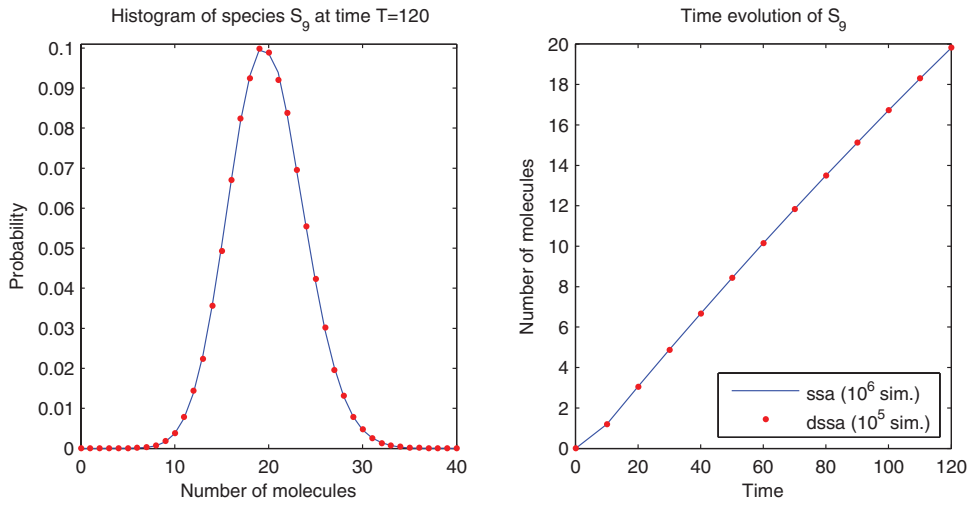


FIG. 3. Sequence of unimolecular reactions with backward bypass reactions (example 3). (Left) Histogram for the number of S_9 molecules at time T = 120 obtained from SSA simulations (blue) and DSSA (red) where delays were drawn from the corresponding 8-exponential iCDF. (Right) Average time evolution of species S_9 in steps of $\Delta t = 10$ using SSA and DSSA.

of the CDF at time t (cf. Sec. III F). Now that we have obtained the CDF, we can draw delays via inverse sampling. We applied this method to example 5 (see Figure 5, black dots) and obtained very accurate solutions at much lower computational costs: computing 100 first-passage times (using the SSA) was roughly 40× slower than computing 100 matrix exponentials (i.e., 100 values of the CDF). Moreover, such matrix exponentials provide us with CDF values at given time points.

E. Systems with additional degradation reactions

Degradation is an essential, ubiquitous process in all biological systems, and ideally we would like to accurately account for it in lumped models. This raises the question whether we can still abridge chemical reaction networks of types (1) and (1′) that include additional degradation of intermediate species by using an appropriate delay distribution. In similitude to systems with forward bypass reactions, an analytic solution of the delay distribution in form of a convolution of exponential distributions is not possible. However, such a distribution can be numerically obtained as described above and in Sec. III F.

We can do so by defining a single additional absorbing state $S_\emptyset$ for all degradation reactions. While it is not necessary to represent $S_\emptyset$ explicitly in the abridged model, one must include the common absorbing state and all degradation reactions in the transition matrix. Since our random walker may now be degraded and therefore never reach "position" S_{n+1}, its first-passage time is now described by a pseudo-distribution function with a cumulative limit $w < 1$ (for $t \to \infty$) while $1 - w$ corresponds to the probability that the

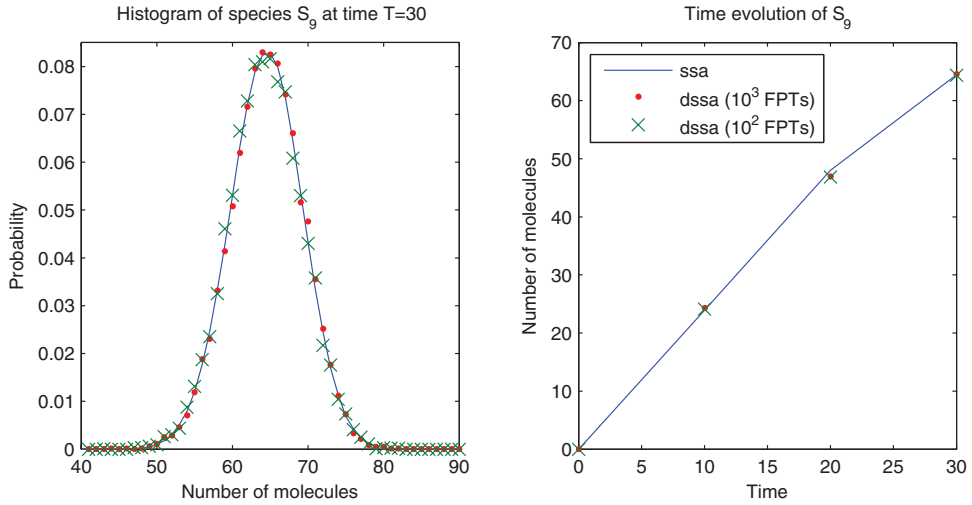


FIG. 4. Sequence of unimolecular reactions with forward bypass reactions (example 4). (Left) Histogram for the number of S_9 molecules at time T = 30 for 10^6 SSA simulations (blue) and 10^4 DSSA simulations (red, green) where delays were obtained from first-passage time distributions based on 10^2 (red) and 10^3 (green) sample times. (Right) Average time evolution of species S_9 in steps of $\Delta t = 10$ for SSA and DSSA.

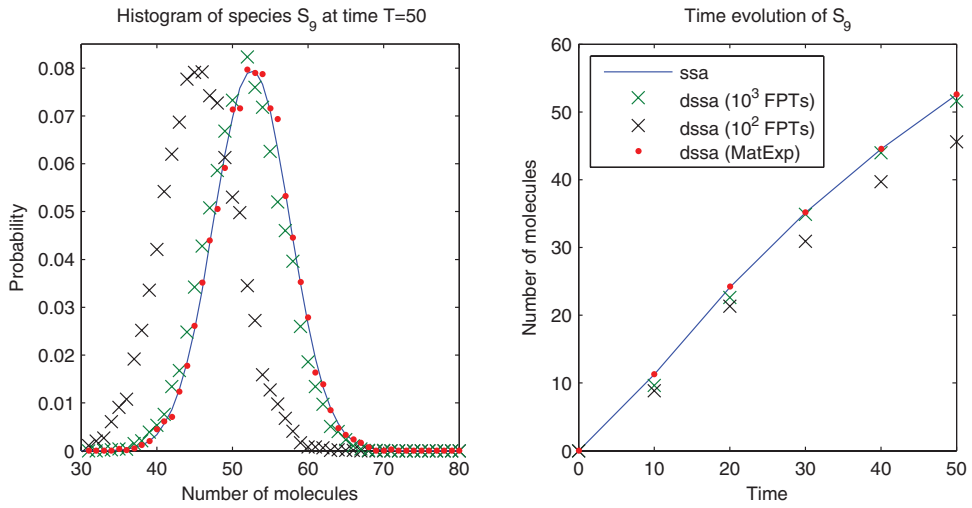


FIG. 5. Sequence of unimolecular reactions with backward and forward bypass reactions (example 5). (Left) Histogram for the number of S_9 molecules at time T = 50 for 10^6 SSA simulations (blue) and 10^4 DSSA simulations where delays were either drawn from first-passage time distributions based on 10^2 (black x) or 10^3 (green x) samples or obtained (via inverse sampling) from the numerical solution of the CDF evaluated at 6401 time points in the interval [0, 640] (black dot). (Right) Average time evolution of species S_9 in steps of $\Delta t = 10$ for SSA and DSSA.

walker is degraded (cf. Sec. III G). Our experimental results show a correct reduction of networks with degradation using this method.

For instance, let us consider example 1 along with additional degradation reactions

$$S_3 \xrightarrow{c_{3b}} S_\emptyset, \quad S_5 \xrightarrow{c_{5b}} S_\emptyset, \quad S_8 \xrightarrow{c_{8b}} S_\emptyset$$

with $c_{3b} = c_{5b} = c_{8b} = 0.1$ (example 6). By following the procedure described above, we can reduce the full system exactly (Figure 6). In this case, simulations of the reduced system are about 20× faster than those of the full system, where we calculated the matrix exponential for various time points t numerically (cf. Sec. III F).

F. Systems with additional incoming reactions

Another interesting extension to the core systems (1) and (1′) are additional incoming reactions whose products can be any intermediate reactant species S_i ($i = 1...n$). Such abstract reaction networks are commonly used in biological research and, thus, it would be ideal to also reduce their simulation costs. For this purpose, let us consider example 1 with an additional incoming reaction

$$X \xrightarrow{c_i} S_i.$$

Here, we define the reaction to have a reactant species X, but the exact same technique applies to constitutive incoming reactions with no reactant species (i.e., $\emptyset \xrightarrow{c_i} S_i$). In this case, we will again calculate a first-passage time through random walkers, and the abridged scheme will include two delayed

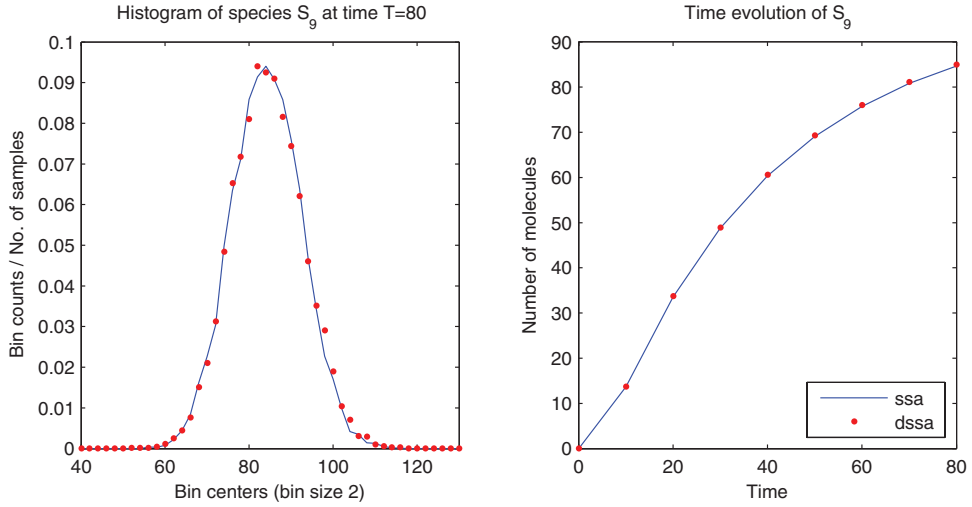


FIG. 6. Sequence of unimolecular reactions with additional degradation reactions (example 6). (Left) Histogram for the number of S_9 molecules at time T = 80 for 10^4 SSA (blue) and DSSA (black dots) simulations where delays were drawn from the inverse of the numerical solution of the CDF (using matrix exponentials). (Right) Average time evolution of species S_9 in steps of $\Delta t = 10$ for SSA and DSSA. For this scenario, we computed $w = 0.194$.

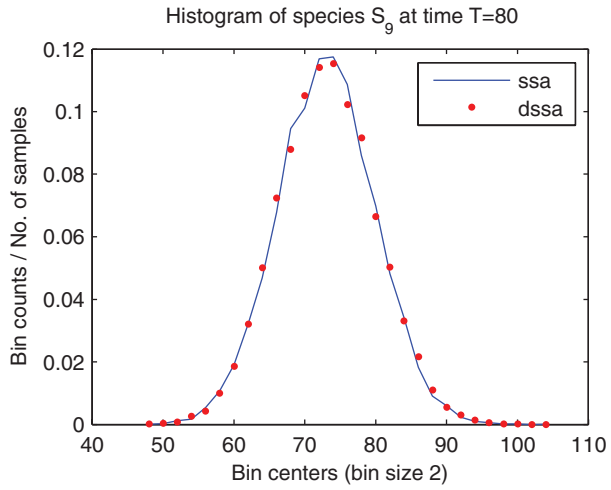


FIG. 7. Histogram (normalized) of species S_9 at time T = 80 from 10 000 simulations of the full (SSA, blue) and the abridged system (DSSA, red).

reactions, namely,

$$S_0 \overset{c_0,\tau^*}{\to} S_{n+1}, \quad X \overset{c_i,\tau'}{\to} S_{n+1}$$

with two distinct delay distributions (τ^*, τ'). As in the original abridged scheme, the second reaction retains the original reaction rate c_i, and the associated delay distribution (τ') represents the first-passage times from S_i to S_{n+1}. This new reaction correctly accounts for random walkers originating at X, and the delay distribution can be numerically calculated through matrix exponentiation (cf. Sec. III F).

Splitting the abridged system into two reactions has additional benefits, as the original system's delay distributions (τ^*) can still be calculated analytically and, therefore, a numerical approach is only necessary for obtaining the additional delay distribution (τ').

Figure 7 shows simulation results for the system described in example 1 extended by the incoming reaction $X \overset{1}{\to} S_3$, where only the delay of the latter is derived numerically.

G. Lumping binary reactions

Our abridgement method works equally well when considering binary reactions at the beginning of the reaction scheme (1′). Also, without loss of generality, the end of the reaction scheme can have two products. For instance, a reaction scheme

$$S_A + S_B \overset{c_0}{\to} S_1 \underset{c_1}{\overset{c_1}{\rightleftarrows}} \dots \underset{c_{n-1_r}}{\overset{c_{n-1}}{\rightleftarrows}} S_n \overset{c_n}{\to} S_C + S_D$$

can be reduced, exactly, to the reaction

$$S_A + S_B \overset{c_0,\tau^*}{\to} S_C + S_D.$$

It is worth noting not all types of reaction systems containing binary reactions can be reduced exactly with a constant delay distribution. However, there are scenarios where an approximate, yet highly accurate, reduction can still be obtained, while considering a constant delay distribution. We will briefly discuss a few scenarios for which a near-exact abridgment can still be achieved with our methodology.

Let us consider the Michaelis-Menten type reaction system

$$S + E \underset{c_{1_r}}{\overset{c_1}{\rightleftarrows}} SE \overset{c_2}{\to} E + P$$

that we want to replace by a delayed reaction

$$S + E \overset{k,\tau^*}{\to} P + E$$

with rate k and delay distribution τ^*, yielding a good approximation of species P dynamics.

To fit into the standard DSSA scheme, where reactions are chosen according to propensities, we need to introduce a rate constant k for this reaction. This rate should be much larger than other reaction rates in the system, and chosen in such a way that the waiting time of each delayed reaction is negligibly small compared to the associated random delay. Alternatively, one can use a modified DSSA approach that draws delays for such reactions whenever reactants are available.[16] If these species are also reactants in other reactions, one would have to calculate additional probabilities for each of the competing reactions to happen, and chose at random.

As already mentioned above, there is no constant delay distribution that yields exact dynamics of P, since the transition rate from "state" $S + E$ to SE depends on the number of molecules of S and E. Instead, each state (S, E) requires a different delay distribution. However, if either $S_0 \ll E_0$ or $E_0 \ll S_0$, the larger initial condition can be used as a factor for the corresponding transition coefficient in matrix A, yielding a delay distribution that is constant and a good approximation. For $E_0 \ll S_0$, the transition matrix has now the following form:

$$A = \begin{pmatrix} -c_1 S_0 & c_{1_r} \\ c_1 S_0 & -(c_{1_r} + c_2) \end{pmatrix}.$$

We illustrate this approach for a set of parameters and initial conditions taken from Wu *et al.*[17] See Figure 8 for reaction rates, initial conditions, and a comparison of histograms of species S and P at system end time of SSA and DSSA.

H. Application: Chains of reactions in mRNA decay

We will finally apply our methodology to a model for the detailed turnover process of MFA2pG mRNAs presented in Ref. 12, Figure 1, and Table I. Figure 9 shows the original mRNA degradation model with associated kinetic parameter values and our abridged model. The latter assumes that we are only interested in the dynamics of fragment I2. It has the form of system (1′) plus additional degradation reactions for the branching points (at species C and D). Figure 10 shows the time evolution of species I2 for the full and the abridged model. As expected, both solutions are indistinguishable from each other.

Note that if we would like to observe the total of all 3′ fragments (L + I1 + I2), 5′ fragments (G + M), or numbers of full-length mRNA (A + B + BC1 +⋯+ BC5 + C + D + E + F) as in Ref. 12, exact lumping in one reaction would

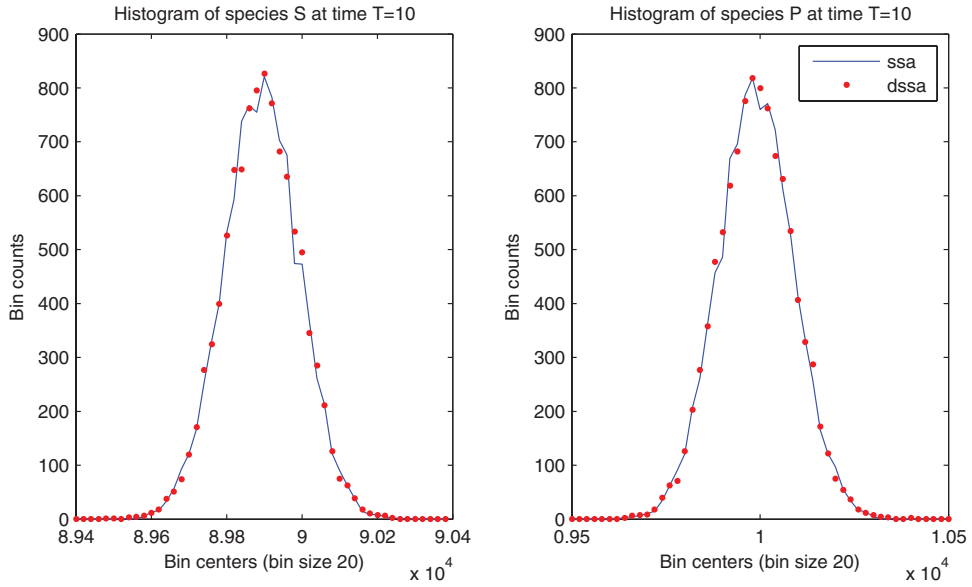
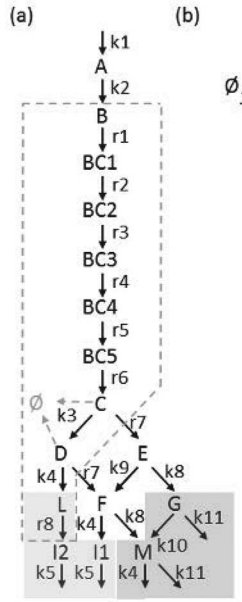


FIG. 8. Histograms of species S and P at time T = 10 from 10 000 simulations of the full (SSA, blue) and the abridged system (DSSA, red) for the parameter set $c_1 = 1$, $c_2 = 10$, $c_3 = 10$ and initial conditions $S_0 = 10^5$, $E_0 = 10^2$, $ES_0 = 0$, $P = 0$. Here, we use $k = 10^{15}$.

only be applicable to the combined cytosolic translocation plus poly(A) shortening process (reactions A → B →…→C) due to subsequent branching points in the reaction network. For this sub-process, the corresponding rate matrix is a lower triangular matrix with diagonal elements of the form $c_i - \lambda_i$ and, hence, the delay distribution is nothing but the convolution of the exponential distributions with parameters c_i (here: k2, r1, r2, …, r6) – saving us the calculation of the eigenvalues.



Param.	Rates
k1	1
k2	0.2
k3	0.0077
k4	1
k5	0.0009
k8	0.0012
k9	0.0077
k10	0.077
k11	0.001201
r1	0.011
r2	0.02201
r3	0.02202
r4	0.02203
r5	0.02204
r6	0.023
r7	0.00222
r8	0.0022201

FIG. 9. (a) and (c) Complete model and kinetic parameters used in modelling MFA2pG-mRNA degradation as stated in Ref. 12. 3′ fragments (L + I1 + I2) and 5′ fragments (G + M) are highlighted in grey, all other species are considered full length mRNA. (b) Abridged model. The dotted line in (a) refers to the process that is represented by a delay distribution in our abridged model. Moreover, reactions C → E and D → F are lumped into a single degradation reaction. The probability for such a degradation is $1 - w$ (cf. Sec. III G).

III. METHODS

A. Determinant of A

The rate matrix A of system (1) is a tridiagonal $n \times n$ matrix with diagonal entries $a_{1,1} = -c_1$ and $a_{i,i} = -(c_{i-1_r} + c_i)$ for $i = 2, \ldots, n$, and subdiagonal and superdiagonal elements $a_{i+1,i} = c_i$ and $a_{i,i+1} = c_{i_r}$ for $i = 1, \ldots, n-1$, respectively. The determinant of a tridiagonal matrix is given by the extended continuant[18]

$$|A|_n = a_{n,n}\,|A|_{n-1} - a_{n,n-1}a_{n-1,n}\,|A|_{n-2}\,,$$

where $|A|_j$ is the jth principal minor of the first j rows/columns of A. For our example, we can show inductively that the determinant of A is given by the following expression:

$$\det(A) = |A|_n = (-1)^n \prod_{i=1}^{n} c_i \tag{4}$$

with $|A|_0 = 1$ and $|A|_1 = -c_1$. The inductive step

$$\begin{aligned} |A|_n &= -(c_{n-1_r} + c_n)|A|_{n-1} - c_{n-1}c_{n-1_r}|A|_{n-2} \\ &= -(c_{n-1_r} + c_n)(-1)^{n-1}\prod_{i=1}^{n-1} c_i \\ &\quad -c_{n-1}c_{n-1_r}(-1)^{n-2}\prod_{i=1}^{n-2} c_i \\ &= -c_n(-1)^{n-1}\prod_{i=1}^{n-1} c_i = (-1)^n \prod_{i=1}^{n} c_i \end{aligned}$$

shows the validity of (4). Note that A is nonsingular since $c_i > 0$, $i = 1, \ldots, n$ and, hence, $\det(A) \neq 0$.

B. Eigenvalues of A

It follows from (4) that A is not singular, and consequently none of its eigenvalues λ_i is zero. We can further

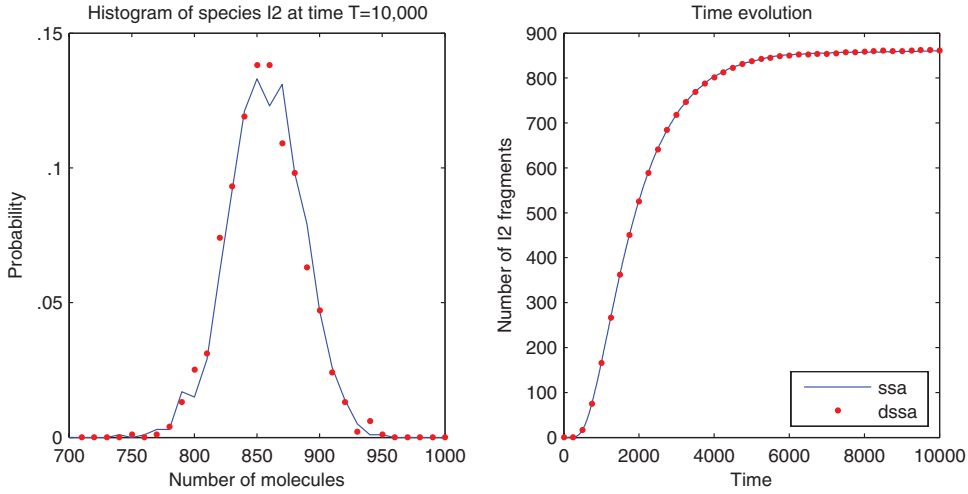


FIG. 10. (Left) Histogram for the number of I2 molecules at time T = 10 000 for SSA simulations (blue) and DSSA (black) where delays were drawn from the inverse of the numerical solution of the CDF (using matrix exponentials). (Right) Average time evolution of species I2 for SSA (blue line) and DSSA (black dots). For this scenario, we computed $w = 0.7745$, i.e., a 77.45% probability that the former full length mRNA is degraded via this pathway.

characterize the eigenvalues by applying Gerschgorin's circle theorem[14] to bound the spectrum of A. Applied to the columns of A, the theorem states that every eigenvalue must lie within at least one of the Gerschgorin discs centred at $a_{i,i}$ and with radius $r_i = \sum_{j\neq i} |a_{j,i}|$, $i = 1, \ldots, n$. Moreover, it should be noted that $a_{i,i} = -\sum_{j\neq i} a_{j,i}$, and therefore, the non-zero eigenvalues of A have negative real parts. Additionally, A is a tridiagonal matrix with real off-diagonal values that satisfy $a_{i,i+1}$, $a_{i+1,i} > 0$, for $i = 1, \ldots, n-1$. Any such matrix is known to be similar to a Hermitian matrix and therefore has real and simple eigenvalues.[19] Putting together all the previous facts, we can conclude that the rate matrix A from system (1) is diagonalizable and has non-zero real negative simple eigenvalues.

C. Resolvent of A

The term $P(s) = (sI - A)^{-1}$ is called the resolvent of A. Since we want to find the solution of $p(S_n, t|S_1, 0)$, with $p(t) = e^{At}p(0)$ and $p(0) = [1\,0\ldots0]^T$, we only need the $(n, 1)$ entry of the resolvent $(sI - A)^{-1}$. This entry can be expressed using Cramer's rule as

$$P_n(s) = (sI - A)^{-1}_{n,1} = (-1)^{n+1} \frac{M_{1,n}}{det\,(sI - A)},$$

where $M_{i,j}$ is the (i, j) minor, defined as the determinant of the matrix resulting from removing the ith row and jth column from $(sI - A)$. Note that $(sI - A)$ is also a tridiagonal matrix with diagonal entries $(sI - A)_{1,1} = s + c_1$ and $(sI - A)_{i,i} = s + \left(c_{i-1_r} + c_i\right)$ for $i = 2, \ldots, n$ and subdiagonal and superdiagonal elements $(sI - A)_{i+1,i} = -c_i$ and $(sI - A)_{i,i+1} = -c_{i_r}$ for $i = 1, \ldots, n-1$, respectively. The outcome of removing the 1st row and nth column of a $n \times n$ tridiagonal matrix is an upper diagonal $(n-1) \times (n-1)$ matrix whose diagonal corresponds to the subdiagonal of the original one. Consequently, $M_{1,n} = (-1)^{n-1} \prod_{k=1}^{n-1} c_k$. The term $det(sI - A)$ is simply the characteristic polynomial of A which can be expressed as $\prod_{k=1}^{n} (s - \lambda_k)$. Each eigenvalue λ_k has multiplicity 1 as has been proved before (cf. Sec. III B).

D. First-passage time PDF

Rather than in $P_n(s)$, we are actually interested in an expression for $c_n P_n(s)$ such that applying the inverse Laplace transform yields $c_n p_n(t)$ (due to linearity of the transform), which is precisely the time distribution we are looking for (see Eq. (3)). Using our previous results, we note

$$\begin{aligned} F_n(s) = c_n P_n(s) &= c_n (-1)^{n+1} (-1)^{n-1} \frac{\prod_{k=1}^{n-1} c_k}{\prod_{k=1}^{n} (s - \lambda_k)} \\ &= (-1)^n \frac{(-1)^n \prod_{k=1}^{n} c_k}{\prod_{k=1}^{n} (s - \lambda_k)} = (-1)^n \frac{\det(A)}{\prod_{k=1}^{n} (s - \lambda_k)} \\ &= (-1)^n \frac{\prod_{k=1}^{n} \lambda_k}{\prod_{k=1}^{n} (s - \lambda_k)} = \prod_{k=1}^{n} \frac{-\lambda_k}{(s - \lambda_k)}. \end{aligned}$$

Let $\overline{\lambda_k} = -\lambda_k$ be the positive value of eigenvalue λ_k (we had proved $\lambda_k < 0$) and denote $E_k(s) = \overline{\lambda_k}/(s + \overline{\lambda_k})$ the kth product term of $F_n(s)$. By applying the inverse Laplace transform to each term individually, we obtain

$$\mathcal{L}^{-1}\{E_k(s)\} = e_k(t) = \overline{\lambda_k} e^{-\overline{\lambda_k} t}, \quad \text{for } k = 1, \ldots, n,$$

that is, the probability density function of an exponential distribution with parameter $\overline{\lambda_k}$.

Since $\mathcal{L}^{-1}\{F(s)\,G(s)\} = f(t) \star g(t)$ for two functions $f(t)$ and $g(t)$ with their respective Laplace transforms $F(s)$ and $G(s)$, we obtain for the inverse Laplace transform of $F_n(s)$

$$\mathcal{L}^{-1}\{F_n(s)\} = f_{T_{n+1}} = e_1(t) \star e_2(t) \star \ldots \star e_n(t).$$

In summary, the probability distribution of the first-passage time for our system (1) is the convolution of exponential distributions with parameters $\overline{\lambda_k} = -\lambda_k$.

Note that an alternative argument for this conclusion is that each product term $E_k(s)$ corresponds to the moment-generating function of an exponential distribution and the

product of such moment-generating functions corresponds to the convolution of the respective probability density functions.

Now that we know the specific form of the first-passage time distribution, we only need to know how to draw random numbers from it. Fortunately, an expression for the PDF of a convolution of n exponential distributions has been analytically derived (Ref. 20 as cited in Ref. 21) from where we can derive the CDF

$$F_{T_{n+1}} = \sum_{k=1}^{n} \left(\prod_{l=1, l\neq k}^{n} \frac{\overline{\lambda_l}}{(\overline{\lambda_l} - \overline{\lambda_k})} \right) (1 - e^{-\overline{\lambda_k} t}).$$

During simulations we can then draw time delays by drawing uniform random numbers and evaluating the corresponding (pre-calculated) inverse CDF (iCDF) at such values.

E. Backward bypass reactions

Extending the basic system (1), we will now include backward bypass reactions and show that the new system can still be lumped by using delays that correspond to the convolution of exponentials with $\overline{\lambda_k} = -\lambda_k$. We can reason with one single backward bypass with the understanding that it can be generalized to additional ones. Note that backward bypass reactions add additional entries in the upper triangular part of the rate matrix A (*upper Hessenberg matrix*).

Let $S_j \overset{b}{\leftarrow} S_i$, $i - j > 1$ be the additional bypass reaction that converts species S_i back into S_j. The new rate matrix A' can be described in terms of A (the one without the backward bypass) as

$$A' = [A_1, A_2, \ldots, A_i + B, \ldots, A_n],$$

where A_k is the kth column vector of A, and B is the column vector that accounts for the backward reaction, i.e., $b_j = +b$ and $b_i = -b$. Consequently,

$$\det(A') = \det(A) + \det(A_B),$$
$$A_B = [A_1, A_2, \ldots, A_{i-1}, B, A_{i+1}, \ldots, A_n].$$

Applying Laplace's formula to expand $\det(A_B)$ along the ith column we obtain its expression in terms of two of its minors, namely,

$$\det(A_B) = (-1)^{2i}(-b) M_{i,i} + (-1)^{j+i} b M_{j,i}.$$

It can be shown that these two terms cancel each other out, and therefore $\det(A_B) = 0$ implying $\det(A') = \det(A)$.

The eigenvalues $\lambda_k{}'$ of A' can be characterized in a similar way to what was done previously. It can be shown that they are non-zero, real, and negative. This allows us to follow an analogous reasoning for the resolvent of A' and for the probability distribution of the first-passage time. Note that the new minor $M'_{1,n}$ is not affected by the backward reaction since its value depends exclusively on the subdiagonal which is not modified. Therefore,

$$F'_n(s) = c_n P'_n(s) = c_n(-1)^{n+1}(-1)^{n-1} \frac{\prod_{k=1}^{n-1} c_k}{\prod_{k=1}^{n}(s - \lambda'_k)}$$
$$= (-1)^n \frac{\det(A)}{\prod_{k=1}^{n}(s-\lambda'_k)} = (-1)^n \frac{\det(A')}{\prod_{k=1}^{n}(s-\lambda'_k)}$$
$$= (-1)^n \frac{\prod_{k=1}^{n} \lambda'_k}{\prod_{k=1}^{n}(s - \lambda'_k)} = \prod_{k=1}^{n} \frac{-\lambda'_k}{(s - \lambda'_k)}.$$

In conclusion, a system with consecutive reversible reactions plus additional backward bypass reactions and rate matrix A' can be exactly reduced to a single reaction with associated delay distribution that is the convolution of exponential distributions with parameters $\overline{\lambda'_k} = -\lambda'_k$.

F. Numerical solution of $F_{T_{n+1}}$

In cases where an analytic solution of $f_{T_{n+1}}$ is not available (systems including forward bypass reactions or degradations), we can still obtain the distribution numerically by solving the $(n+1)$-dimensional equation $\frac{d}{dt} p(t) = \tilde{A} p(t)$.

Here, $p(t) = [p(S_1, t|S_1, 0) p(S_2, t|S_1, 0) \ldots p(S_{n+1}, t|S_1, 0)]^T$ and $\tilde{A}$ is the transition matrix for the system including S_{n+1}. The solution has again the form $p(t) = e^{\tilde{A}t} p(0)$, $p(0) = [1\, 0 \ldots 0]^T$, and its last entry corresponds to the CDF $F_{T_{n+1}}(t)$ of our delay distribution. For efficiently calculating matrix exponentials, we use Roger Sidje's software *Expokit*.[22] Even though this software is very efficient, calculating matrix exponentials is still a computationally expensive task. How many and what sample points to choose is not straightforward. However, the CDF will be monotonically non-decreasing towards its maximum (normally this is 1, unless the system includes degradations (cf. Sec. III G)) and one can stop sampling when either the difference to the known maximum or the difference between two consecutive time points is below a certain threshold.

G. Degradation

For taking degradation of intermediate species into account, we have to modify the DSSA. First, we construct the pseudo-CDF of the delay distribution (first-passage distribution) by calculating the matrix exponential of the full transition matrix at various time points t (cf. Sec. III F). For $t \to \infty$, the CDF will now converge to a value w such that $1 - w$ is the degradation probability of any molecule in the abridged model that undergoes any delayed reaction. To decide if a molecule is eventually degraded or ends up as species S_{n+1}, we draw a uniform random number $r \in U(0, 1)$. If $r > w$, we assume that degradation will occur during the delay, otherwise we use the random number to sample a delay from the corresponding iCDF of the delay distribution.

H. Simulations

All examples were verified with numerical results from discrete simulation methods. We followed a standard method: for each system, we ran multiple simulations and obtained

the state value at a particular time point, upon which we compared associated histograms. Full/original systems were simulated with the SSA,[3] and reduced systems were simulated with the direct DSSA.[11]

For determining the eigenvalues, we use MATLAB's *eig* command, an iterative eigensolver employing an efficient version of the QR factorization[23] – a standard method for eigenvalue calculations of real matrices. For nonsymmetric tridiagonal eigenvalue problems – e.g., matrices stemming from linear first-order reaction schemes without backward or forward bypass reactions – a tailored algorithm has been proposed by Bini *et al.*[24] This algorithm is robust and computes eigenvalues of a real n×n nonsymmetric tridiagonal matrix T in $O(n^2)$ operations while the QR method requires $O(n^3)$ operations.[23,24]

In general, an eigensolver's performance does not only depend on the matrix (e.g., its size, sparseness, or shape), but also on the computing platform and underlying software libraries (cf. Demmel *et al.*[25] for a discussion and performance comparison of eigensolvers for symmetric tridiagonal matrices). In practice, in the case of constant reaction rates, the delay distribution has to be calculated only once, prior to running DSSA simulations. In this sense, the time it takes to calculate eigenvalues even for large matrices is rather negligible. For instance, using MATLAB's *eig* function on a laptop with i7-2820QM CPU at 2.3 GHz, calculating all eigenvalues of a random, real valued, dense matrix with n = 1000 takes around 1 s while it takes around 0.75 s for a random tridiagonal matrix of same size.

For Arnoldi estimates, we use MATLAB's *eigs* command, a highly refined implicitly restarted Arnoldi (IRA) method that is supposed to work particularly well with large sparse matrices. For small numbers of required eigenvalues *eigs* is faster than *eig*. For the very same random, dense matrix with n = 1000 for which *eig* takes around 1 s to calculate all eigenvalues, *eigs* calculates the four largest magnitude eigenvalues in about half the time and only the largest eigenvalue in roughly 0.03 s (for the random tridiagonal matrix mentioned above, calculating the four largest magnitude eigenvalues takes about 0.1 s, while calculating only the largest magnitude eigenvalue takes about 0.06 s). A detailed complexity analysis of a further accelerated IRA in terms of matrix order, number of non-zero entries, number of block Arnoldi steps, degree of the Chebychev polynomials, and number of required eigenvalues is provided in Nishida *et al.*[26] While Arnoldi iterations are less important in case of constant rate matrices when the delay distribution is calculated only once, their application can lead to considerable savings in scenarios where the rate matrices are (slowly) time varying (work in progress).

Note that pre-calculation of the iCDF of the delay distribution is done for a discrete number of sample points in [0,1]. This introduces some minor but negligible errors when choosing uniform numbers that fall in-between those pre-defined sample points. However, such error can be minimized by increasing the number of sample points (e.g., here, we used a distance between two sample points of 0.001).

Calculated speed-ups of our method are based on MATLAB implementations of the standard SSA and the direct DSSA (for 1000 simulations each). Note that the time for running the DSSA includes one-time pre-calculation of the iCDF of our delay distribution. Even though both implementations have been performance-optimized, the calculated speed-ups represent only rough estimates of the expected computational savings.

IV. DISCUSSION

Biological processes often involve reactions and mechanisms that may not happen instantaneously, and are best described in a model by means of time delays. For instance, they are commonly used to represent eukaryotic transcription and translation, which imply other spatiotemporal processes often not explicitly modelled (e.g., diffusion and translocation into and out of the nucleus, RNA polymerase activation, splicing, protein synthesis, and protein folding). Additionally, distributed delays can also be incorporated into temporal models to capture essential spatial information, where molecules are allowed to translocate between different cellular compartments and undergo chemical reactions, at a small fraction of the original computational cost.[16]

In this paper, we present yet another use for time delays, within a discrete stochastic setting. Namely, as a novel methodology for exact model reduction, depending on the type of reactions considered. We understand delays as a phenomenological product of many reactions taking time to be completed. Then, instead of describing complex networks of reactions, we lump all processes into a delay distribution that can be calculated, depending on the network structure, either analytically or numerically. We show what method to use for different types of reaction schemes and illustrate their applicability in representing chains of chemical reactions accurately, at much lower computational costs.

So far, our methodology has been shown to work for linear sequences with a final irreversible reaction, with additional forward and backward bypass reactions, constitutive creation, and degradations of intermediate species. Note that reactions that do not interfere with our abridgement scheme, i.e., that do not involve species inside the abridged reaction block (neither as reactants nor as products), are simply carried over into the new model.

Our method's applicability to more complex scenarios such as systems that include general binary reactions is under current investigation. However, we already present accurate abridgement of Michaelis-Menten reactions, as well as reaction chains initiated by binary reactions. Future work on generalized higher order reactions abridgement may be aided by consideration of first-passage time distributions of simple bimolecular reactions, such as that presented in Ref. 27. The current scheme could in principle work when solely considering dynamics at steady state. However, further work would be required to represent systems away from equilibrium, or general chemical reaction networks where reactants or products can participate in other reactions in a competitive way.

We also showed how schemes such as $S_1 \underset{c_{1_r}}{\overset{c_1}{\rightleftarrows}} S_2 \overset{c_2}{\rightarrow} S_3$ can be accurately abridged by $S_1 \overset{k,\tau^*}{\rightarrow} S_3$ with increasing degrees

of accuracy, since the initial unidirectional reaction is not a strict requirement for our abridgment method. The latter is achieved by using an appropriate delay distribution and a very high (artificially introduced) rate k, where the scheme becomes exact as $k \to \infty$. Depending on the reaction rates, such abridgments may yield incredible computational savings ($1000\times$ and higher) without any loss of accuracy. Alternatively, one can use a modified DSSA approach that draws delays for such reactions whenever reactants are available. Also, our latest investigations suggest that an exact abridgement with the standard DSSA approach could be possible for reaction schemes with initial and/or final reversible reactions. However, abridgment of fully reversible reaction schemes is work in progress and beyond the scope of this paper.

Unlike the abridgment method by Gillespie *et al.*,[6] which is based on using a modified rate constant only, our abridgment is valid for all possible rates and initial conditions. However, our abridgment underlies computational costs for calculating the delay distribution and, more importantly, an additional overhead for handling delayed reactions as part of the DSSA algorithm. By the same token, in some instances, running SSA simulations may involve significantly larger amounts of time than DSSA simulations. For instance, there can be scenarios where the SSA continues calculating waiting times and selected reactions of long sequences of reactions, while the DSSA need only update all pending delays of queued lumped reactions. Hence, its speed-up depends on the difference in the number of reactions performed by the non-abridged and the abridged system.

Now, one may think obtaining direct analytic solutions of the CME would be simpler but this approach is limited to monomolecular reaction systems. In order to derive an analytic solution, one needs to calculate the probability distribution describing the CME for each time point individually, as a convolution of multinomial and product Poisson distributions.[28] In fact, calculating the parameters of the multinomial distributions requires solving matrix exponentials for each time point. The latter can only be solved analytically when considering very small systems, for which lumping would not even be necessary. In all other cases, analytical approaches would suffer from the same limitations as our own methodology, and in turn would not offer a solution for reducing the problem exactly (potentially reducing computational costs greatly). Moreover, a major benefit of our method is its modularity, namely, once a delay distribution is calculated, it can be recycled and used in larger chemical reaction systems, for which direct solutions of the CME are unfeasible.

However, it should be noted that real chemical reaction systems do not consist only of first-order reactions. Moreover, reactions that appear to be first-order often follow the so-called Lindemann mechanism. That is, the unimolecular reaction $A \to B$ is in fact an abridged version of the process $A + M \rightleftarrows A^* + M$, and $A^* \to B$. Obviously, our method is exact (in the mathematical sense) only for truly unimolecular reactions. However, it can be considered quasi-exact for systems for which the pseudo-first-order limit of all their bimolecular reactions holds.

Finally, a major advantage of our method as compared to abridgment methods that rely on time-scale separation (such as the methods by Mastny *et al.*[7] and Thomas *et al.*[8]) is that it does not require any time-scale separation conditions to be accurate. Thus, our methodology largely increases the range of reducible biochemical models. Such reduction is exact when dealing with unimolecular and/or backward bypass reactions and, as discussed above, exact when dealing with constitutive creation, degradation, or forward bypass reactions, depending on sample sizes.

ACKNOWLEDGMENTS

M.B. would like to thank Carlos Marijuán for helpful discussions on eigenvalues.

[1] B. Munsky and M. Khammash, J. Chem. Phys. **124**(4), 044104 (2006).
[2] T. G. Kurtz, J. Chem. Phys. **57**(7), 2976 (1972).
[3] D. T. Gillespie, Annu. Rev. Phys. Chem. **58**, 35 (2007).
[4] J. A. Borghans, R. J. de Boer, and L. A. Segel, Bull. Math. Biol. **58**(1), 43 (1996).
[5] M. Frenklach, Chem. Eng. Sci. **40**(10), 1843 (1985).
[6] D. T. Gillespie, Y. Cao, K. R. Sanft, and L. R. Petzold, J. Chem. Phys. **130**(6), 064103 (2009).
[7] E. A. Mastny, E. L. Haseltine, and J. B. Rawlings, J. Chem. Phys. **127**(9), 094106 (2007).
[8] P. Thomas, A. V. Straube, and R. Grima, BMC Syst. Biol. **6**(1), 39 (2012); P. Thomas, R. Grima, and A. V. Straube, Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **86**(4 Pt 1), 041110 (2012).
[9] T. T. Marquez-Lago and J. Stelling, Biophys. J. **98**(9), 1742 (2010).
[10] M. Barrio, K. Burrage, A. Leier, and T. Tian, PLoS Comput. Biol. **2**(9), e117 (2006).
[11] X. Cai, J. Chem. Phys. **126**(12), 124108 (2007).
[12] D. Cao and R. Parker, RNA **7**(9), 1192 (2001).
[13] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier Science, 2007).
[14] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra* (SIAM, 1997).
[15] G. Strang, *Introduction to Linear Algebra* (Wellesley Cambridge, 2009).
[16] T. T. Marquez-Lago, A. Leier, and K. Burrage, BMC Syst. Biol. **4**, 19 (2010).
[17] S. Wu, J. Fu, Y. Cao, and L. Petzold, J. Chem. Phys. **134**(13), 134112 (2011).
[18] T. Muir, *A Treatise on the Theory of Determinants* (Dover, 1960).
[19] K. Veselic, Linear Algebra Appl. **27**, 167 (1979).
[20] H. Jasiulewics and W. Kordecki, Demonstratio Math. **36**(1), 231 (2003).
[21] M. Akkouchi, J. Chungcheong Math. Soc. **21**(4), 501–510 (2008).
[22] R. B. Sidje, ACM Trans. Math. Softw. **24**(1), 130 (1998).
[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins University Press, Baltimore, 1996).
[24] D. A. Bini, L. Gemignani, and F. Tisseur, SIAM J. Matrix Anal. Appl. **27**(1), 153 (2005).
[25] J. W. Demmel, O. A. Marques, B. N. Parlett, and C. Vomel, SIAM J. Sci. Comput. **30**(3), 1508 (2008).
[26] A. Nishida, R. Suda, and Y. Oyanagi, in *Proceedings of the Fourth IMACS International Symposium on Iterative Methods in Scientific Computation*, edited by D. R. Kincaid and A. C. Elster (IMACS, Austin, TX, 1998), Vol. 5, p. 45.
[27] P. Keller and A. Valleriani, J. Chem. Phys. **137**(8), 084106 (2012).
[28] T. Jahnke and W. Huisinga, J. Math. Biol. **54**(1), 1 (2007).