Perspective: Stochastic algorithms for chemical kinetics

Daniel T. Gillespie,[1,a)] Andreas Hellander,[2,b)] and Linda R. Petzold[2,c)]
[1]*Dan T Gillespie Consulting, 30504 Cordoba Pl., Castaic, California 91384, USA*
[2]*Department of Computer Science, University of California Santa Barbara, Santa Barbara, California 93106, USA*

(Received 5 January 2013; accepted 25 March 2013; published online 1 May 2013)

We outline our perspective on stochastic chemical kinetics, paying particular attention to numerical simulation algorithms. We first focus on dilute, well-mixed systems, whose description using ordinary differential equations has served as the basis for traditional chemical kinetics for the past 150 years. For such systems, we review the physical and mathematical rationale for a discrete-stochastic approach, and for the approximations that need to be made in order to regain the traditional continuous-deterministic description. We next take note of some of the more promising strategies for dealing stochastically with stiff systems, rare events, and sensitivity analysis. Finally, we review some recent efforts to adapt and extend the discrete-stochastic approach to systems that are not well-mixed. In that currently developing area, we focus mainly on the strategy of subdividing the system into well-mixed subvolumes, and then simulating diffusional transfers of reactant molecules between adjacent subvolumes together with chemical reactions inside the subvolumes. [http://dx.doi.org/10.1063/1.4801941]

I. INTRODUCTION

When Ludwig Wilhelmy, in 1850, described the buffered conversion of sucrose into glucose and fructose using a first-order ordinary differential equation (ODE),[1] ODEs were inaugurated as the standard tool for mathematically modeling chemical kinetics. The suitability of ODEs for that task must have seemed obvious to 19th century scientists: after all, ODEs provided the mathematical basis for that most fundamental of all dynamical theories, Newton's second law. But ODEs in chemical kinetics imply a continuous-deterministic time evolution for the species concentrations. That is at odds with the view, which was not universally accepted until the early 20th century, that matter consists of discrete molecules which move and chemically react in a largely random manner.

For a long time, scientists showed little concern with this mismatch. The first serious attempt to model the intrinsically discrete-stochastic behavior of a chemically reacting system was made in 1940 by pioneering biophysicist Max Delbrück;[2] however, no further work in that vein was done until the 1950s. An overview of work through the late 1960s has been given by McQuarrie.[3] In the 1970s, procedures for stochastically simulating chemically reacting systems using large computers began to be devised. But controversies arose over the correct physical basis and mathematical formalism for stochastic chemical kinetics. By the end of the 1970s, that confusion, along with the fact that molecular discreteness and randomness posed no problems for ODEs in describing typical test-tube-size systems, effectively relegated stochastic chemical kinetics to a subject of only academic interest. That changed nearly two decades later, when Arkin, McAdams, and a growing number of other researchers[4–7] showed that in living cells, where reactant species are often present in relatively small molecular counts, discreteness and stochasticity can be important. Since then a great many papers have been published on the theory, computational methods, and applications of stochastic chemical kinetics, practically all aimed toward cellular chemistry.

In Sec. II, we outline our perspective on the stochastic chemical kinetics of systems in which the reactant molecules are "dilute" and "well-mixed"—terms that will be defined shortly. The theoretical and computational picture for that relatively simple scenario has become greatly clarified over just the last dozen years. For situations where the reactant molecules crowd each other or are not well-mixed, situations that are quite common in cellular systems, progress has been made, but the going is slow. Many important questions have not yet been fully answered, and undoubtedly some have yet to be asked. In Sec. III, we discuss some of the issues surrounding one major strategy for simulating systems that are not well-mixed. In Sec. IV we briefly summarize recent accomplishments and current challenges.

II. DILUTE WELL-MIXED CHEMICAL SYSTEMS

A. The chemical master equation and the propensity function

It was primarily through the work of McQuarrie[3] that what is now called the *chemical master equation* (CME) became widely known. For N chemical species $S_1, \ldots, S_N$ whose molecules can undergo M chemical reactions $R_1, \ldots, R_M$, and with $X_i(t)$ denoting the (integer) number of S_i molecules in the system at time t, the CME is a time-evolution equation for $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, the probability that $\mathbf{X}(t) \equiv (X_1(t), \ldots, X_N(t))$ will be equal to $\mathbf{x} = (x_1, \ldots, x_N)$, given that $\mathbf{X}(t_0) = \mathbf{x}_0$ for

a)Electronic mail: gillespiedt@mailaps.org
b)Electronic mail: andreash@cs.ucsb.edu
c)Electronic mail: petzold@cs.ucsb.edu

some $t_0 \le t$:

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(\mathbf{x} - \boldsymbol{\nu}_j) P(\mathbf{x} - \boldsymbol{\nu}_j, t|\mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t|\mathbf{x}_0, t_0)]. \quad (1)$$

Here $\boldsymbol{\nu}_j \equiv (\nu_{1j}, \ldots, \nu_{Nj})$, where ν_{ij} is the (signed integer) change in the S_i molecular population caused by one R_j event. And a_j, now called the *propensity function* for reaction R_j, is defined to be such that, for any infinitesimal time increment dt,

$$a_j(\mathbf{x})\, dt \equiv \text{the probability, given } \mathbf{X}(t) = \mathbf{x}, \text{ that an } R_j \text{ event will occur somewhere inside } \Omega \text{ in } [t, t + dt) \quad (j = 1, \ldots, M). \quad (2)$$

The CME (1) follows rigorously from this definition of a_j and the above definition of P via the laws of probability.

It is sometimes thought that the solution of the CME is a pristine probability function $P(x, t)$ that describes the system independently of any observer-specified initial condition. That this is not so becomes clear when one realizes that if nothing is known about the state of the system at any time before t, then it will not be possible to say anything substantive about the state of the system at t. The solution $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ of the CME (1) actually describes the gradually eroding knowledge that an observer, who last observed the system's state at time t_0, has of the state of the system as time increases beyond t_0.[8]

The key player in the CME (1) from the point of view of physics is the propensity function a_j defined in Eq. (2). As its name suggests, a_j quantifies how likely it is that reaction R_j will fire. Early work on the CME tended to view a_j as merely an ad hoc stochastic extension of the conventional reaction rate in the ODE formalism, with the latter having the more rigorous physical justification. But the situation is actually the other way around. As will become clear later in this section, the ODE formalism is an approximation of the stochastic formalism which is generally accurate only if the system is sufficiently large. Therefore, although there is a very close connection between propensity functions and conventional deterministic reaction rates, the latter, being an approximate special case of the former, cannot be used to derive the former. Nor can propensity functions be justifiably obtained by assuming hypothetical models or rules. An honest derivation of the propensity function must look directly to molecular physics to see how chemical reaction events actually occur, and then adopt a mathematical formalism that accurately characterizes that physical behavior.

B. Physical justification for the propensity function

The implicit assumption in (2) that a chemical reaction is a physical event that occurs practically instantaneously means that, at least for the dilute solution systems that we will be primarily concerned with here, every R_j must be one of two types: either *unimolecular*, in which a single molecule suddenly changes into something else; or *bimolecular*, in which two molecules collide and immediately change into something else. Trimolecular and reversible reactions in cellular chemistry nearly always occur as a series of two or more unimolecular or bimolecular reactions.

Unimolecular reactions of the form $S_1 \to \cdots$ are inherently stochastic, usually for quantum mechanical reasons; there is no formula that tells us precisely when an S_1 molecule will so react. But it has been found that such an R_j is practically always well described by saying that the probability that a randomly chosen S_1 molecule will react in the next dt is equal to some constant c_j times dt. Summing $c_j dt$ over all x_1 S_1 molecules in Ω, in accordance with the addition law of probability, gives Eq. (2) with $a_j(\mathbf{x}) = c_j x_1$.

The bimolecular reaction $S_1 + S_2 \to \cdots$ is more challenging. In 1976, Gillespie[9,10] presented a simple kinetic theory argument showing that, if the reactant molecules comprise a well-mixed dilute gas inside Ω at temperature T, then a propensity function for $S_1 + S_2 \to \cdots$ as defined in (2) exists and is given by

$$a_j(x_1, x_2) = \left(\pi \sigma_{12}^2 \bar{v}_{12} q_j \Omega^{-1}\right) \cdot x_1 x_2 \quad \text{(dilute gas)}. \quad (3a)$$

Here, σ_{12} is the average distance between the centers of a pair of reactant molecules at collision (the sum of their radii for hard sphere molecules); $\bar{v}_{12} = \sqrt{(8k_B T)/(\pi m_{12})}$ is their average relative speed, with m_{12} being their reduced mass; and q_j is the probability that an S_1-S_2 collision will produce an R_j reaction.[11]

The derivation of Eq. (3a)[11] is valid only if the reactant molecules are "well-mixed" and "dilute." The well-mixed requirement means that a randomly selected reactant molecule should no more likely be found in any one subvolume of the system than in any other subvolume of the same size. But note this does not require that there be a perfectly regular placement of the reactant molecules inside Ω, nor that there be a large number of those molecules. If this well-mixed requirement cannot be sustained by the natural motion of the molecules, then it must be secured by external stirring. The dilute requirement means that the average separation between two reactant molecules should be very large compared to their diameters, or equivalently, that the total volume occluded by all the reactant molecules should comprise only a very small fraction of Ω.

Generalizing the dilute gas result (3a) to a solution is obviously a necessary first step toward making the CME applicable to cellular chemistry. But doing that has long seemed problematic. According to the standard theory of diffusion, the root-mean-square displacement of a molecule in time Δt is proportional to $\sqrt{\Delta t}$; this suggests, at least on the basis of the way in which the dilute-gas result (3a) is derived,[11] that the probability for a pair of diffusing molecules to react in the next dt might not have the linear dependence on dt demanded by (2). But in 2009, a detailed physics argument was produced[12] which shows that if the S_1 and S_2 molecules are solute molecules, well-mixed and dilute (in the above sense) in a bath of very many much smaller solvent molecules, then a propensity function for $S_1 + S_2 \to \cdots$ as defined in (2) does

exist, and is given explicitly by

$$a_j(x_1, x_2) = \left(\frac{4\pi\sigma_{12}^2 D_{12}\bar{v}_{12}q_j\Omega^{-1}}{4D_{12} + \sigma_{12}\bar{v}_{12}q_j}\right) \cdot x_1 x_2 \quad \text{(dilute solution)}. \tag{3b}$$

Here, D_{12} is the sum of the diffusion coefficients of the S_1 and S_2 molecules, and the other quantities are as previously defined. Note that the requirement for diluteness in this solution context applies only to the reactant solute molecules, and not to the solvent molecules. In the "fast-diffusion" limit defined by $4D_{12} \gg \sigma_{12}\bar{v}_{12}q_j$, Eq. (3b) reduces to the dilute gas result (3a). At the opposite "diffusion limited" extreme $4D_{12} \ll \sigma_{12}\bar{v}_{12}q_j$, the factor in parentheses in Eq. (3b) reduces to $4\pi\sigma_{12}D_{12}\Omega^{-1}$, which corresponds to a well known deterministic rate result that can be obtained by adapting Smoluchowski's famous analysis of colloidal coagulation.[13] The derivation of Eq. (3b) actually makes use of the Smoluchowski analysis, but does so in a way that takes account of the fact that the standard diffusion equation, on which the Smoluchowski analysis is based, is physically incorrect on small length scales.[12]

It remains to be seen how the propensity function hypothesis (2) will fare if the reactant molecules crowd each other, or if they move by active transport mechanisms along physically confined pathways. Even greater challenges attend relaxing the well-mixed assumption, because without it the "state" of the system can no longer be defined by only the total molecular populations of the reactant species. Those populations must be supplemented with information on the positions of the reactant molecules in order to advance the system in time. That in turn will require tracking the movement of individual molecules in a manner that is physically accurate yet computationally efficient—a very tall order! Some efforts in these directions will be discussed in Sec. III. But here we will assume that the propensity function as defined in Eq. (2) exists, and further that it has the form $c_j x_1$ for the unimolecular reaction $S_1 \to \cdots$, the form $c_j x_1 x_2$ for the bimolecular reaction $S_1 + S_2 \to \cdots$, and the form $c_j \frac{1}{2} x_1(x_1 - 1)$ for the bimolecular reaction $2S_1 \to \cdots$. Also important for later development of the theory is the fact that c_j will be independent of the system volume Ω for unimolecular reactions, and inversely proportional to Ω for bimolecular reactions. The latter property, which can be seen in both Eqs. (3a) and (3b), reflects the obvious fact that it will be harder for two reactant molecules to find each other inside a larger volume.

C. The stochastic simulation algorithm

The difficulty of solving the CME for even very simple systems eventually prompted some investigators to consider the complementary approach of constructing simulated temporal trajectories or "realizations" of $\mathbf{X}(t)$. Averaging over sufficiently many such realizations can yield estimates of any average that is computable from the solution $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ of the CME, and examining only a few realizations often yields insights that are not obvious from $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$. The earliest known constructions of simulated trajectories were made in 1972 by Nakanishi[14] and Šolc and Horsák,[15] and in 1974 by Bunker *et al.*[16] But all these simulation procedures either were designed for specific simple systems, or else made heuristic approximations.

In 1976 Gillespie[9,10] proposed an exact, general-purpose procedure for simulating chemical reactions which is now called the *stochastic simulation algorithm* (SSA). The derivation of the SSA starts by posing the following question: Given the system's state $\mathbf{X}(t) = \mathbf{x}$ at time t, at what time $t + \tau$ will the next reaction in the system occur, and which R_j will that next reaction be? Since, owing to the probabilistic nature of Eq. (2), τ and j must be random variables, the answer to this question can be supplied only by their joint probability density function (PDF), $p(\tau, j \mid \mathbf{x}, t)$. That function is defined so that $p(\tau, j \mid \mathbf{x}, t) \cdot d\tau$ gives the probability, given $\mathbf{X}(t) = \mathbf{x}$, that the next reaction event in the system will occur in the time interval $[t + \tau, t + \tau + d\tau)$ and will be an R_j. Gillespie[9] showed that Eq. (2) together with the laws of probability implies that this PDF is given by

$$p(\tau, j \mid \mathbf{x}, t) = \mathrm{e}^{-a_0(\mathbf{x})\tau} a_j(\mathbf{x}), \tag{4}$$

where $a_0(\mathbf{x}) \equiv \sum_{k=1}^{M} a_k(\mathbf{x})$. The SSA is thus the following computational procedure:

1. In state $\mathbf{x}$ at time t, evaluate (as necessary) $a_1(\mathbf{x}), \ldots, a_M(\mathbf{x})$, and their sum $a_0(\mathbf{x})$.
2. Generate two random numbers τ and j according to the PDF (4).
3. Actualize the next reaction by replacing $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$.
4. Record $(\mathbf{x}, t)$. Return to Step 1, or else end the simulation.

Step 2 of the SSA can be implemented using any of several different exact methods, and Gillespie's paper[9] presented two: the *direct method*, which follows from a straightforward application of the inversion Monte Carlo generating technique to the PDF (4);[17] and the *next reaction method*, which despite its indirectness is equally exact. Of those two methods, the direct method is usually more efficient, and it goes as follows: Draw two unit-interval uniform random numbers u_1 and u_2, and take[17]

$$\tau = \frac{1}{a_0(\mathbf{x})} \ln\left(\frac{1}{1 - u_1}\right), \tag{5a}$$

$$j = \text{the smallest integer satisfying } \sum_{k=1}^{j} a_k(\mathbf{x}) > u_2\, a_0(\mathbf{x}). \tag{5b}$$

Other exact methods for implementing step 2 were subsequently developed by other workers, and they offer computational advantages in various specific situations. The most useful of those appear to be: the *next reaction method* of Gibson and Bruck,[18] which is a major reformulation of the first reaction method; the *first family method* of Lok (described in Ref. 19); the *optimized direct method* of Cao *et al.*;[20] the *sorting direct method* of McCollum *et al.*;[21] the *modified next reaction method* of Anderson,[22] which in this context is the same as the next reaction method[18] but is more flexibly couched in the "random time change representation"

of Kurtz;[23,24] and the *composition-rejection method* of Slepoy *et al.*[25] An in-depth critique of the computational efficiencies of these and a few other methods has been given by Mauch and Stalzer.[26]

Delayed events are difficult to incorporate analytically into the CME, but they can be handled easily by the SSA. Thus, suppose a reaction occurring at time t_1 signals that, independently of any subsequent reactions that might occur, an event E_d will occur at time $t_1 + \tau_d$; e.g., a DNA transcription might start at time t_1 and produce an mRNA a time τ_d later. The event's delay time τ_d could either be a specified value, or it could be a random value that has been sampled from some known PDF. In either case, E_d and its time $t_1 + \tau_d$ are logged into a "delayed-event queue" which will temporarily halt the simulation when a reaction is first called for at some time $t' > t_1 + \tau_d$. Since such a call implies that nothing happens between the time of the last reaction event and $t_1 + \tau_d$, the SSA simply advances the system without change to time $t_1 + \tau_d$, discharges the event E_d, and then resumes the simulation from time $t_1 + \tau_d$ (ignoring the call for a reaction at time t'). Contingencies can also be easily accommodated; thus, if the delayed event E_d will occur at time $t_1 + \tau_d$ only if some other event E_c does not occur first, then if and when E_c occurs in $[t_1, t_1 + \tau_d]$ the SSA simply removes E_d from the queue.

D. Tau-leaping

It often happens that the average time between reactions, which can be shown from Eq. (4) to be $a_0^{-1}(\mathbf{x})$, is so small that simulating every reaction event one at a time is not computationally feasible, no matter what method is chosen to implement step 2 of the SSA. Tau-leaping, introduced in 2001 by Gillespie,[27] aims to give up some of the exactness of the SSA in return for a gain in computational efficiency. It "leaps" the system ahead by a pre-selected time τ which may encompass more than one reaction event (this τ is not the same as the time τ to the next reaction in the SSA). The procedure for doing that is a straightforward consequence of the fact that the Poisson random variable with mean $a\tau$, which we denote by $\mathcal{P}(a\tau)$, gives the (integer) number of events that will occur in the next time τ, provided that the probability of a single event occurring in any infinitesimal time dt is adt where a is any positive constant. Therefore, given $\mathbf{X}(t) = \mathbf{x}$, if τ chosen small enough that

$$a_j(\mathbf{x}) \approx \text{constant in } [t, t+\tau),\ \forall j \quad \text{(1st leap condition)}, \tag{6}$$

then during the interval $[t, t+\tau)$ there will be $\approx \mathcal{P}\left(a_j(\mathbf{x})\tau\right)$ firings of reaction channel R_j. Since each of those firings augments the state by $\boldsymbol{\nu}_j$, the state at time $t+\tau$ will be[27]

$$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \mathcal{P}_j\left(a_j(\mathbf{x})\tau\right)\boldsymbol{\nu}_j, \tag{7}$$

where the $\mathcal{P}_j$ are M independent Poisson random variables. Equation (7) is called the *tau-leaping formula*. Its accuracy depends solely on how well condition (6) is satisfied.

Implementing tau-leaping at first seems straightforward: choose a value for τ, generate M Poisson random numbers with respective means $a_1(\mathbf{x})\tau, \ldots, a_M(\mathbf{x})\tau$, and then evaluate Eq. (7). But in practice, there are problems. One is to determine in advance the largest value of τ that satisfies the leap condition (6). The strategy for doing that in Gillespie's original paper[27] was flawed, and allowed leaps to be taken that could produce substantial changes in propensity functions that have relatively small values. This not only produced inaccurate results, it also occasionally caused the population of some reactant species to go negative.[28] Another problem in implementing tau-leaping is that, while tau-leaping does become exact in the limit $\tau \to 0$, it also becomes infinitely inefficient in that limit: when τ is near zero, the M generated Poisson random numbers in Eq. (7) will usually all be zero, and that results in a computationally expensive leap with no change of state. A way is therefore needed to make tau-leaping segue to the SSA automatically and efficiently as τ becomes comparable to the average time $a_0^{-1}(\mathbf{x})$ to the next reaction. A series of improvements in tau-leaping culminating in the 2006 paper of Cao *et al.*[29] solves these two problems for most practical applications. A tutorial presentation of that improved tau-leaping procedure is given in Ref. 30.

Among several variations that have been made on tau-leaping, three are especially noteworthy: the implicit tau-leaping method of Rathinam *et al.*,[31] which adapts implicit Euler techniques developed for stiff ODEs to a stochastic setting; the R-leaping method of Auger *et al.*,[32] which leaps by a pre-selected total number of reaction firings instead of by a pre-selected time; and the unbiased post-leap rejection procedure of Anderson.[33]

E. Connection to the traditional ODE approach

Tau-leaping is also important because it is the first step in connecting the discrete-stochastic CME/SSA formalism with the continuous-deterministic ODE formalism. The second step along that path focuses on situations in which it is possible to choose a leap time τ that is not only small enough to satisfy the first leap condition (6), but also large enough to satisfy

$$a_j(\mathbf{x})\,\tau \gg 1,\ \forall j \quad \text{(2nd leap condition)}. \tag{8}$$

In that circumstance, we can exploit two well known results concerning $\mathcal{N}(\mu, \sigma^2)$, the normal random variable with mean μ and variance σ^2: First, when $m \gg 1$, the Poisson random variable $\mathcal{P}(m)$ can be well approximated by $\mathcal{N}(m, m)$, at least for reasonably likely sample values of those two random variables. And second, $\mathcal{N}(\mu, \sigma^2) \equiv \mu + \sigma\mathcal{N}(0, 1)$. Applying those two results to the tau-leaping formula (7), again assuming that both leap conditions (6) and (8) are satisfied, yields

$$\mathbf{X}(t+\tau) \doteq \mathbf{x} + \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{x})\,\tau + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{x})}\,\mathcal{N}_j(0,1)\,\sqrt{\tau}. \tag{9a}$$

From one point of view, Eq. (9a) is simply an approximation of the tau-leaping formula (7) that has replaced Poisson random numbers with normal random numbers. But recalling

that $\mathbf{x} \equiv \mathbf{X}(t)$, and noting that because of condition (6) we can regard τ as an infinitesimal dt on macroscopic time scales, we can also write Eq. (9a) as

$$\mathbf{X}(t+dt) - \mathbf{X}(t) \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}(t))\, dt + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}(t))}\, \mathcal{N}_j(0,1)\, \sqrt{dt}. \tag{9b}$$

This equation has the canonical form of a "stochastic differential equation" or "Langevin equation."[34] It is called the *chemical Langevin equation* (CLE).

This derivation of the CLE is due to Gillespie.[35] For reasons that can be understood from this derivation, the CLE will not accurately describe "rare events."[36] But if the system admits a $dt \equiv \tau$ that is small enough to satisfy the first leap condition (6) yet also large enough to satisfy the second leap condition (8), then the CLE should give a fair account of the typical (as opposed to the atypical) behavior of the system. As a set of N coupled equations, the CLE (9b) is much less formidable than the CME (1), since the latter is a set of coupled equations indexed by $\mathbf{x}$. Furthermore, since a numerical simulation using Eq. (9a) will, as a consequence of condition (8), step over very many reaction events for each reaction channel, the CLE will be much faster than the SSA.

But the CLE requires both leap conditions to be satisfied, and that is not a trivial requirement since it is easy to find systems for which that cannot be done. But in 2009, Gillespie[37] proved that both leap conditions can always be satisfied, and hence the CLE will always be valid, if the system is made sufficiently "large" in the sense of the thermodynamic limit—where the molecular populations are imagined to go to infinity along with the system volume Ω while the concentrations remain constant. The proof of that result uses the earlier noted fact that real-world elemental reactions R_j have propensity functions that are either of the form $c_j x_i$ with c_j independent of Ω, or $c_j x_i x_k$ with c_j proportional to Ω^{-1}; that implies that, in the thermodynamic limit, all real-world propensity functions $a_j(\mathbf{x})$ grow linearly with the system size. Therefore, roughly speaking, after satisfying the first leap condition (6) by fixing τ sufficiently small, we can satisfy the second leap condition (8) simply by taking the system sufficiently close to the thermodynamic limit.

As the thermodynamic limit is approached, the left side of the CLE (9b) grows linearly with the system size; the first term on the right, being proportional to the propensity functions, also grows linearly with the system size; and the second term on the right, being proportional to the square roots of the propensity functions, grows like the square root of the system size. So in the full thermodynamic limit, the second term on the right of the CLE (9b) becomes negligibly small in comparison with the other terms, and that equation reduces to the ODE[38]

$$\frac{d\mathbf{X}(t)}{dt} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}(t)). \tag{10a}$$

This is the *reaction rate equation* (RRE), the ODE of traditional chemical kinetics. Since as we have just seen, the RRE is generally valid only in the thermodynamic limit, it is more commonly written in terms of the concentration variable[39] $\mathbf{Z}(t) \equiv \mathbf{X}(t)/\Omega$ and the functions $\tilde{a}_j(\mathbf{z})$, the latter being defined as the thermodynamic limit of $\Omega^{-1} a_j(\mathbf{x})$:

$$\frac{d\mathbf{Z}(t)}{dt} \doteq \sum_{j=1}^{M} \boldsymbol{\nu}_j \tilde{a}_j(\mathbf{Z}(t)). \tag{10b}$$

The convergence of the jump Markov process described by the CME (1) first to the continuous Markov process described by the CLE (9b) and then in the thermodynamic limit to the ODE (10b), also follows from results of Kurtz[40] on the approximation of density-dependent Markov chains.

The theoretical structure presented above is summarized in Fig. 1. As we proceed from the top of that figure to the bottom, we move toward approximations that require increasingly larger molecular populations, but are computationally more efficient. The question naturally arises: how many molecules must a system have in order to be reliably described at a particular level in Fig. 1? No general answer to that question can be given, since any answer will depend on the structure and the parameter values of the reaction network.

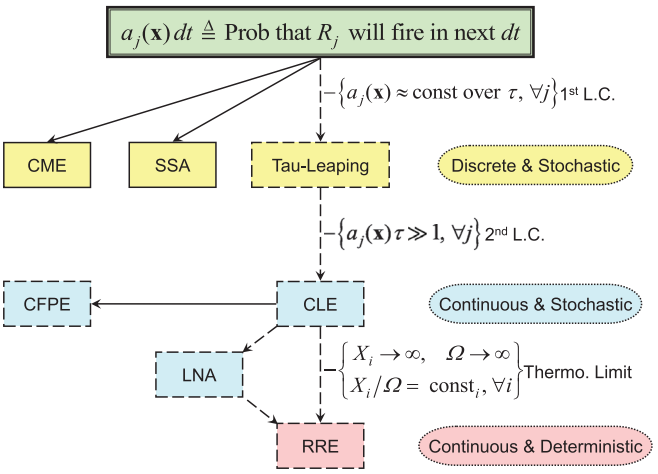

FIG. 1. Stochastic chemical kinetics is premised on the definition (2) of the propensity function in the top box, a definition which must look to molecular physics for its justification. The two solid-outlined boxes in yellow denote mathematically exact consequences of that definition: the chemical master equation (1) and the stochastic simulation algorithm (4). Dashed-outlined boxes denote approximate consequences: tau-leaping (7), the chemical Langevin equation (9), the chemical Fokker-Planck equation (not discussed here but see Ref. 35), and the reaction rate equation (10). The bracketed condition by each dashed inference arrow is the condition enabling that approximation: reading from top to bottom, those conditions are the first leap condition, the second leap condition, and the thermodynamic limit. The rationale for viewing the linear noise approximation (LNA)[41] as an intermediate result between the CLE and the RRE is detailed in Ref. 42. It has been shown[37,42] that for realistic propensity functions, getting "close enough" to the thermodynamic limit will ensure simultaneous satisfaction of the first and second leap conditions, at least for finite spans of time; therefore, the top-to-bottom progression indicated in the figure will inevitably occur as the molecular populations and the system volume become larger. But a given chemical system might be such that the largest value of τ that satisfies the first leap condition will not be large enough to satisfy the second leap condition; in that case, there will be no accurate description of the system below the discrete-stochastic level in the figure.

But it turns out that this question can be rendered practically irrelevant when doing simulations. To see why, suppose we have a tau-leaping implementation that efficiently segues to the SSA, such as the one described in Ref. 29, and using it we have found a τ that satisfies the first leap condition (6). Then the number of firings in the next τ of each reaction channel R_j will be well approximated by a Poisson random variable with mean $a_j(\mathbf{x})\tau$, as in the tau-leaping formula (7). However, the call to the Poisson random number generator for a sample value of each $\mathcal{P}_j\left(a_j(\mathbf{x})\tau\right)$ can be handled on a reaction-by-reaction basis: If $a_j(\mathbf{x})\tau$ is "large" compared to 1 then the Poisson random number generator can return instead a normal random number with mean and variance $a_j(\mathbf{x})\tau$, as would happen for all R_j if we were using the Langevin leaping formula (9a). Or, if $a_j(\mathbf{x})\tau$ is "very large" compared to 1, so that the standard deviation $\sqrt{a_j(\mathbf{x})\,\tau}$ is negligibly small compared to the mean $a_j(\mathbf{x})\tau$, then the Poisson random number generator can return the non-random number $a_j(\mathbf{x})\tau$, as would happen for all R_j if we were using the RRE (10a). In this way, each reaction channel R_j gets assigned to its computationally most efficient level in Fig. 1. It is not necessary for all the reactions channels to be assigned to the same level, nor even for the simulator to be aware of the level to which each reaction has been assigned.

The relationships outlined in Fig. 1 do not assume the applicability of the system-size expansion of van Kampen.[41] But Wallace *et al.*[42] have shown that the most commonly used result of the system-size expansion, namely, van Kampen's linear noise approximation (LNA),[41] does have a place in Fig. 1: A surprisingly easy derivation[42] of the LNA as a linearized approximation of the CLE (9b) positions the LNA midway between the CLE and the RRE. That positioning is consistent with findings of Grima *et al.*[43] which indicate that the CLE is indeed more accurate than the LNA. The LNA describes the "initial departure" of the CLE from the deterministic RRE as we back away from the thermodynamic limit to a large but finite system. That initial departure is the appearance of normal fluctuations about the deterministic RRE solution, the variances of which are given explicitly by the LNA.

Concern about the accuracy of the CLE (9b) prompts the question: Is there a formula for $\mathbf{X}(t+dt)-\mathbf{X}(t)$ that is exactly equivalent to the CME/SSA? The answer is yes. It is the $\tau = dt$ version of the tau-leaping formula (7), which with $\mathbf{x} = \mathbf{X}(t)$ reads

$$\mathbf{X}(t+dt)-\mathbf{X}(t) = \sum_{j=1}^{M} \mathcal{P}_j(a_j(\mathbf{X}(t))dt)\boldsymbol{\nu}_j.$$

This is so because the only constraint on the accuracy of Eq. (7) is that τ be "small enough," and that constraint is always satisfied by a true infinitesimal dt (the dt in the CLE is a "macroscopic" infinitesimal). However, this equation is practically useless for computation: since the Poisson random variable $\mathcal{P}(m)$ takes only integer values, then $\mathcal{P}(a\,dt)$ for any finite a will almost always take the value 0, so the right side of the above equation will almost always be exactly 0. In contrast, the right side of the CLE (9b) will almost always give some small but non-zero value for $\mathbf{X}(t+dt)-\mathbf{X}(t)$. However, integrating the above equation from t_0 to t yields a more useful result:

$$\begin{aligned}\mathbf{X}(t)-\mathbf{X}(t_0) &= \sum_{j=1}^{M}\int_{t_0}^{t}\mathcal{P}_j(a_j(\mathbf{X}(t'))dt')\,\boldsymbol{\nu}_j \\ &= \sum_{j=1}^{M} Y_j\left(\int_{t_0}^{t} a_j(\mathbf{X}(t'))dt'\right)\boldsymbol{\nu}_j, \qquad (11)\end{aligned}$$

where the Y_j are "scaled, independent, unit-rate Poisson processes."[44] Equation (11) is the *random time change representation* of Kurtz;[23,24] it is fully equivalent to the CME, in essentially the same way that a Langevin equation is fully equivalent to a Fokker-Planck equation. As shown by Anderson,[22] Eq. (11) provides an alternate way of viewing the next reaction method of Gibson and Bruck;[18] furthermore, it is the basis for Anderson's modified next reaction method[22] and post-leap rejection tau-leaping procedure.[33]

F. Stiff systems and the slow-scale SSA

Many real-world chemical systems include a mixture of fast and slow reaction channels which share one or more species. Perhaps the simplest example is

$$S_1 \underset{c_2}{\overset{c_1}{\rightleftarrows}} S_2 \xrightarrow{c_3} S_3 \quad \text{with } c_2 \gg c_3. \qquad (12)$$

Here, successive R_3 firings will be separated by very many relatively uninteresting R_1 and R_2 firings; yet the latter will consume most of the time in a regular SSA run. This inefficiency cannot be overcome by ordinary tau-leaping, because leap times that satisfy the first leap condition will typically be on the order of the average time between the fastest reactions. An analogous computational inefficiency plagues deterministic chemical kinetics, where it is known as "stiffness."

The *slow-scale stochastic simulation algorithm* (ssSSA)[45] is a way of handling this problem that combines efficacy with a clear theoretical justification. It begins by defining as "fast reactions" those which occur much more frequently than all the other reactions, which are designated "slow." In reactions (12), that criterion designates R_1 and R_2 as fast reactions, and R_3 as a slow reaction.[46] Next the ssSSA identifies as "fast species" all those whose populations get changed by a fast reaction, and all the other species as "slow." In reactions (12), S_1 and S_2 are fast species, and S_3 is a slow species. The ssSSA then defines the "virtual fast process" (VFP) to be the fast species populations evolving under only the fast reactions; in reactions (12), the VFP is X_1 and X_2 evolving under R_1 and R_2. Unlike the real fast process, where fast species populations can also get changed by slow reactions (in this case R_3), the VFP will always have a Markovian master equation. For the ssSSA to be applicable, the solution of the VFP master equation must have an asymptotic ($t \to \infty$) steady-state which is effectively reached in a time that is small compared to the average time between slow reactions. In contrast to several other approaches to the stochastic stiffness problem (implicit tau-leaping, hybrid methods, etc.), the ssSSA does not require the fast species to have large molecular populations.

The goal of the ssSSA is to skip over uninteresting fast reactions and simulate only the slow ones, using modified versions of their propensity functions. What allows that to be done in a provably accurate manner is a result called the slow-scale approximation lemma.[45] It says that, under the conditions described above, replacing the slow reaction propensity functions with their averages over the fast species, as computed from the asymptotic VFP, will yield a set of modified propensity functions for the slow reactions that can be used in the SSA to simulate the evolution of the slow-species populations.

If the separation between the fast and slow timescales is sufficiently large, a substantial increase in simulation speed can usually be achieved with the ssSSA. The main challenge in implementing it is computing the required averages with respect to the asymptotic VFP. That can be done exactly for reactions (12), but approximations are usually required for more complicated reaction sets.[45] Often solutions of the equilibrium RRE corresponding to the asymptotic VFP will suffice,[47] although more sophisticated moment closure approximations for the asymptotic VFP will usually be more accurate.[48,49] In some circumstances it may be easier to estimate the required averages by making brief SSA runs of the VFP.[50] A software implementation of the ssSSA which automatically and adaptively partitions the system and efficiently computes the modified slow propensity functions for general mass action models is available.[51]

G. Rare events

In biochemical systems, rare events are important because their occurrence can have major consequences. But the standard SSA is ill-suited to quantifying rare events, since witnessing just one will, by definition, require an impractically long simulation run. A promising way around this difficulty called the *weighted SSA* (wSSA) was introduced in 2008 by Kuwahara and Mura.[52] Instead of pursuing a traditional mean first passage time, they innovatively focused on the probability $p(\mathbf{x}_0, \mathcal{E}; t)$ that the system, starting at time 0 in a specified state $\mathbf{x}_0$, will first reach any state in a specified set $\mathcal{E}$ before a specified time $t > 0$. In other words, instead of trying to estimate the very long time it would take for the rare event to happen, the wSSA tries to estimate the very small probability that the rare event will happen in a time t of practical interest.

To compute $p(\mathbf{x}_0, \mathcal{E}; t)$, the wSSA employs a Monte Carlo procedure called importance sampling. More specifically, it uses the SSA to advance the system to the target time t with the direct method (5), except that in the j-selection procedure (5b) the propensity functions $a_j(\mathbf{x})$ are replaced with a modified set of propensity functions $b_j(\mathbf{x})$ which are biased toward the target states $\mathcal{E}$. Correction for that bias is achieved by weighting the resulting realization by a product of the weights $(a_j/a_0)/(b_j/b_0)$ for each reaction in the realization. Realizations that fail to reach $\mathcal{E}$ by time t are assigned a weight of 0. The average of these weighted realizations then estimates the probability $p(\mathbf{x}_0, \mathcal{E}; t)$. Kuwahara and Mura[52] showed that, with appropriate weighting, their wSSA can achieve substantial improvements over runs made using the unweighted SSA. Gillespie *et al.*[53] later introduced some refinements and clarifications, notably computing also the variance of the weighted trajectories, which quantifies the uncertainty in the estimate of $p(\mathbf{x}_0, \mathcal{E}; t)$ and helps in finding optimally biased propensities.

Choosing the biased propensity functions is the major challenge of the wSSA, because it is often not clear which reactions should be biased nor how strongly, and making suboptimal choices can result in being even less efficient than the SSA. The original biasing scheme proposed by Kuwahara and Mura[52] took $b_j(\mathbf{x}) = \gamma_j a_j(\mathbf{x})$, with the constants γ_j being chosen by intuition and trial-and-error. Subsequent innovations by Roh and Daigle and their collaborators[54–56] have yielded a greatly improved version of the wSSA called the *state-dependent doubly weighted SSA* (sdwSSA). The sdwSSA: (i) allows the proportionality constants γ_j to be state dependent; (ii) biases not only the j-selection procedure but also the τ-selection procedure, replacing $a_0(\mathbf{x})$ in (5a) with $b_0(\mathbf{x})$; and (iii) uses the multi-level cross-entropy method of Rubinstein[57] to develop a robust variance-estimation procedure that automatically determines the optimal biasing propensities $b_j(\mathbf{x})$ with minimal input from the user.[56]

H. Sensitivity analysis

A commonly used measure of the sensitivity of the average of some function f of state $\mathbf{x}$ (e.g., the molecular population of a particular species) to a parameter c (e.g., the rate constant of a particular reaction) at a specified time $t > 0$ given $\mathbf{X}(0) = \mathbf{x}_0$ is the change in that average when c is changed by some small amount ε, divided by ε:

$$\text{sens}\{f(t;\mathbf{x}_0, c), \varepsilon\} \equiv \frac{\langle f(t;\mathbf{x}_0, c+\varepsilon)\rangle - \langle f(t;\mathbf{x}_0, c)\rangle}{\varepsilon}.$$

The most obvious way to estimate this quantity would be to use a finite difference approximation, i.e., make two independent sets of SSA runs to time t, one set using the parameter value c and the other using the parameter value $c + \varepsilon$, and compute the two averages $\langle f(t; \mathbf{x}_0, c)\rangle$ and $\langle f(t; \mathbf{x}_0, c + \varepsilon)\rangle$. But since ε needs to be small to localize the sensitivity at c, the difference between those two averages will usually be much smaller than the statistical uncertainties in their estimates for runs of reasonable length. As a consequence, the relative uncertainty in the estimate of the numerator on the right will usually be too large to be informative.

One way of dealing with this problem, the origin of which dates from the early days of Monte Carlo work, is called the *common random numbers* (CRN) procedure. It generates the SSA trajectories for c and $c + \varepsilon$ in pairs, using the same uniform random number string $\{u_i\}$ for each pair member, and then computes the average of the difference $[f(t, c + \varepsilon) - f(t, c)]$ over the paired trajectories. The positive correlation between the paired trajectories caused by using the same string of random numbers to generate them gives $[f(t, c + \varepsilon) - f(t, c)]$ a smaller variance about its average $\langle f(t, c + \varepsilon) - f(t, c)\rangle \equiv \langle f(t, c + \varepsilon)\rangle - \langle f(t, c)\rangle$ than in the independent run case. That in turn yields a more accurate estimate of the sensitivity for a given number of runs. But unless t is very small, the paired trajectories eventually get out of sync with

each other. When that happens the correlation gradually dies off, and the CRN estimate of sens$\{(f(t; \mathbf{x}_0, c), \varepsilon)\}$ eventually becomes no more accurate than what would be obtained with the independent trajectories approach.

Rathinam *et al.*[58] have developed a significant improvement in this procedure called the *common reaction path* (CRP) method. In generating the paired c and $c + \varepsilon$ trajectories, they use a variation of Anderson's modified next reaction method, in which paired trajectories use the same streams of unit exponential random number for each of the unit-rate Poisson processes Y_j $(j = 1, \ldots, M)$ in the random time change representation (11). That results in a significantly tighter correlation between paired trajectories than in the CRN procedure, and hence a significantly more accurate estimation of sens$\{(f(t; \mathbf{x}_0, c), \varepsilon)\}$ for the same computational effort. Anderson[59] has introduced a different variation of the modified next reaction method, called the *coupled finite difference* (CFD) procedure. It exploits the additivity of the Y_js in Eq. (11) to split them into sub-processes that are shared by paired trajectories in a way that usually gives even tighter and longer lasting correlations, and thus an even more accurate estimate of the sensitivity. References 58 and 59 give detailed descriptions of the CRP and CFD sensitivity estimation procedures.

III. BEYOND WELL-MIXED SYSTEMS

Many situations require relaxing the assumption of a well-mixed reaction volume. Compartmentalization and localization of reactions to cellular membranes are ubiquitous mechanisms for cellular regulation and control. Even in cases where the geometry does not call for spatial resolution, short-range correlations can give rise to effects that can only be captured in simulations with spatial resolution.[60] And models increasingly call not only for spatial resolution, but also stochasticity.[61–64] A striking example of that is the oscillation of Min proteins in the bacterium *E. coli*, where a deterministic partial differential equation model could replicate wild type behavior but not the behavior of known mutants.[65]

A. The reaction-diffusion master equation and simulation algorithm

A popular extension of the CME (1) to the spatially inhomogeneous case, which dates back at least to the 1970s,[66] is the reaction-diffusion master equation (RDME). The original idea of the RDME was to subdivide the system volume Ω into K uniform cubic subvolumes or "voxels" Ω_k $(k = 1, \ldots, K)$, each of edge length h, in such a way that within each voxel the reactant molecules can be considered to be well-mixed. Chemical reactions are then regarded as occurring completely inside individual voxels. The M nominal reactions $\{R_j\}$ thus get replaced by KM reactions $\{R_{jk}\}$, where R_{jk} is reaction R_j inside voxel Ω_k. The propensity function a_{jk} for R_{jk} is the propensity function a_j for R_j, but now referred to the voxel volume $|\Omega_k| = h^3$, and regarded as a function of $\mathbf{x}_k = (x_{1k}, \ldots, x_{Nk})$ where x_{ik} is the current number of S_i molecules in Ω_k. The state-change vector $\boldsymbol{\nu}_{jk}$ for R_{jk} is the $\boldsymbol{\nu}_j$ for R_j, but confined to the space of $\mathbf{x}_k$.

The diffusion of an S_i solute molecule in a sea of many smaller solvent molecules is generally assumed to be governed by the Einstein diffusion equation,

$$\frac{\partial p(\mathbf{r}, t)}{\partial t} = D_i \nabla_{\mathbf{r}}^2 p(\mathbf{r}, t), \tag{13}$$

where p is the position PDF of the S_i molecule and D_i is its diffusion coefficient. But the RDME actually models the diffusion of an S_i molecule from voxel Ω_k to adjacent voxel Ω_l as a "diffusive transfer reaction" R_{ikl}^{d}, whose propensity function a_{ikl}^{d} is $d_{ikl}x_{ik}$ where d_{ikl} is a constant, and whose state-change vector $\boldsymbol{\nu}_{ikl}^{\mathrm{d}}$ decreases x_{ik} by 1 and increases x_{il} by 1. A variety of arguments show[67,68] that this modeling of the diffusive transfer of an S_i molecule to an adjacent voxel will, for sufficiently small h, approximate the behavior dictated by Eq. (13) provided the constant d_{ikl} is taken to be

$$d_{ikl} = \frac{D_i}{h^2}. \tag{14}$$

If B is the total number of planar surface boundary elements shared by two adjacent voxels, then there will be a total of $2NB$ diffusive transfer reactions.

Since diffusion is here being modeled as sudden jumps in the system's state of the same mathematical type as the dynamics of the state jumps induced by chemical reactions, the RDME is just the well-mixed CME (1) with the following reinterpretation of its symbols: the N-dimensional state vector $\mathbf{x} = \{x_i\}$ in Eq. (1) is now regarded as the KN-dimensional state vector $\{x_{ik}\}$; and the M propensity functions $\{a_j\}$ and their associated state-change vectors $\{\boldsymbol{\nu}_j\}$ in Eq. (1) are now regarded as those for the KM chemical reactions $\{R_{jk}\}$ and the $2NB$ diffusive transfer reactions $\{R_{ikl}^{\mathrm{d}}\}$, with all of those reactions being treated on an equal footing. The algorithm for exactly simulating the system described by the RDME is therefore the SSA described in Sec. II C, but with these same reinterpretations of $\mathbf{x}$, a_j, and $\boldsymbol{\nu}_j$.

Uniform Cartesian meshes are attractive and efficient to use for relatively simple geometries, such as those that can be logically mapped to rectangles. For cellular geometries with curved inner and outer boundaries or subcellular structures, however, it is challenging to impose a Cartesian grid that respects the boundaries without a very fine mesh resolution. Using other types of meshes and discretizations, complex geometries can be accommodated in RDME simulations. By defining the jump probability rate constants on a general mesh based on a numerical discretization of Eq. (13), the probability to find a (mesoscopic) molecule inside a certain voxel at a given time will approximate that of a Brownian particle. Also, in the thermodynamic limit, the mean value of the concentration of the molecules in a voxel will converge to the solution of the classical (macroscopic) diffusion equation. The latter follows from classical results of Kurtz.[69]

Isaacson and Peskin proposed to discretize the domain with a uniform Cartesian mesh but allowed for a general, curved boundary by using an embedded, or cut-cell, boundary method.[70,71] Engblom *et al.*[72] took another approach, and used an unstructured triangular or tetrahedral mesh and a finite element method to discretize the domain and to compute the rate constants. The use of unstructured meshes greatly

simplifies resolution of curved boundaries, but also introduces additional numerical considerations. Mesh quality becomes an important factor for how accurately the jump constants d_{ikl} can be computed. An in-depth discussion of the criteria imposed on the unstructured mesh and the discretization scheme can be found in Engblom *et al.*[72] In Drawert *et al.*[73] the benefits and drawbacks of structured Cartesian versus unstructured triangular and tetrahedral meshes are further illustrated from a software and simulation perspective. Figure 2 shows parts of a Cartesian mesh (a) and a triangular mesh (b) in 2D. For the unstructured triangular mesh, molecules are assumed to be well-mixed in the "dual" elements, which are shown in pink. The same interpretation holds for a Cartesian mesh, where the dual is the staggered grid with respect to the primal mesh.

Perhaps the most challenging aspect of a RDME simulation today is to choose the mesh resolution. Common practice to decide whether a given mesh is appropriate is to repeat simulations with both coarser and finer meshes to determine whether the solution seems to change significantly with respect to some output of interest. Such a trial and error approach is not a satisfactory solution, since it is both time consuming and difficult to determine to what extent the accuracy of realizations of the stochastic process changes. Development of both *a priori* and *a posteriori* error estimates for the effects of discretization errors in a general reaction-diffusion simulation would help make RDME simulations more robust from the perspective of non-expert users, and make possible, e.g., adaptive mesh refinement. One step in this direction has been taken by Kang *et al.*[74] The situation is complicated by the fact that the standard formulation of the RDME breaks down for very small voxel sizes. This issue will be discussed in more detail in Sec. III C.

B. Algorithms for spatial stochastic simulation

Variations on the several methods for implementing step 2 of the SSA in Sec. II C have been developed for reaction-diffusion systems that exploit the relatively sparse dependencies among the reactions and/or the simple linear form of the diffusive transfer propensity functions. Since the total number of reactions (chemical plus diffusive transfer) will be very large if there are many voxels, the direct method[9] will usually be very inefficient, because of the linear computational complexity of its j-selection step (5b) in the number of voxels. Elf and Ehrenberg[62] proposed the next subvolume method, which combines ideas of Lok's first family method (described in Ref. 19) and Gibson and Bruck's next reaction method,[18] in an algorithm specifically tailored for reaction-diffusion systems. Reactions are grouped into families according to their voxels, and one samples first the firing time of the next voxel family, then whether it was a chemical reaction or a diffusive transfer reaction, and finally which particular reaction. The methodology of the next reaction method is used to select the next firing voxel, so the complexity in selecting the next event increases only logarithmically with the number of voxels.

As the voxel size h is made smaller, the growing number of voxel boundaries together with the increasing values of d_{ikl} in Eq. (14) conspire to increase the number of

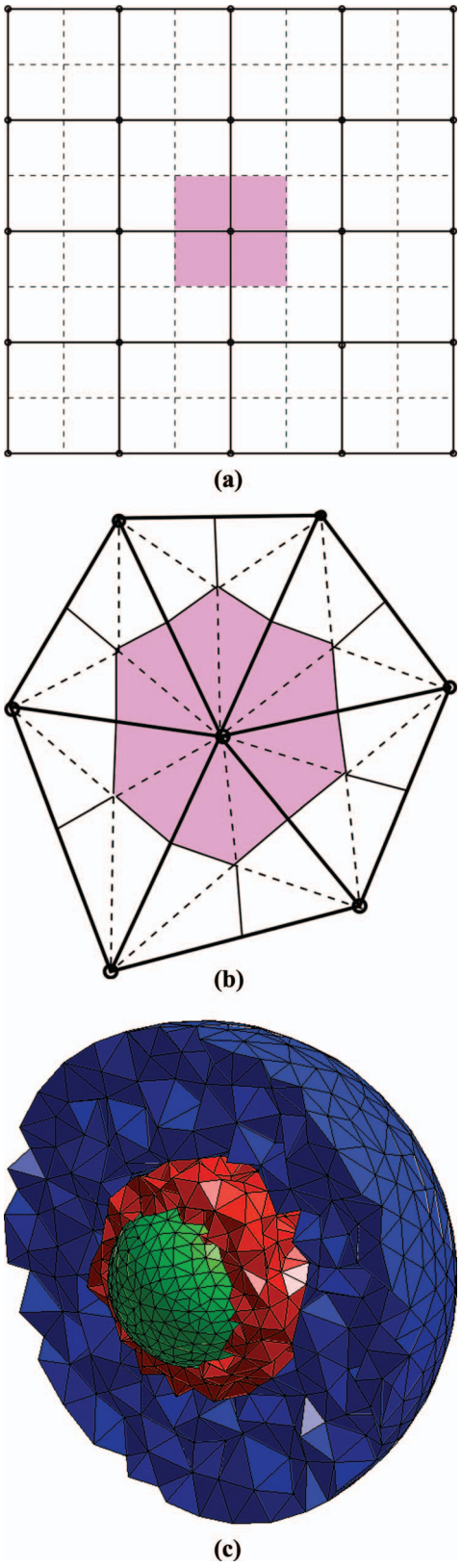

FIG. 2. Parts of a Cartesian mesh (a) and an unstructured triangular mesh (b). Molecules are assumed to be well-mixed in the local volumes that make up the dual elements of the mesh (depicted in pink color). For the Cartesian grid (a), the dual is simply the staggered grid. The dual of the triangular mesh in (b) is obtained by connecting the midpoints of the edges and the centroids of the triangles. (c) shows how a model of a eukaryotic cell with a nucleus (green) can be discretized with a mesh made up of triangles and tetrahedra. The figure is adapted from Ref. 70, where a model of nuclear import was simulated on this domain using the URDME software.

TABLE I. Simulation times for a spatial stochastic system simulated to a final time of 200 s with the next-subvolume method, as implemented in URDME.

$h_{\max}$[a]	No. of voxels	t (s)[b]	No. of events/s (s^{-1})	No. of events
2×10^{-7}	1555	56	3.0×10^{6}	1.7×10^{8}
1×10^{-7}	10 759	333	1.8×10^{6}	6.0×10^{8}
5×10^{-8}	80 231	2602	0.8×10^{6}	22.8×10^{8}

[a] $h_{\max}$ is the maximum local mesh size allowed in the unstructured mesh; it corresponds to h in Eq. (14).
[b] t is the execution time; it grows rapidly with increasing mesh resolution.

diffusive transfer events that occur over a fixed interval of system time. This is illustrated in Table I, where we show the execution time for next subvolume method as implemented using the URDME software package,[72] when applied to a 3D simulation of Min oscillations in the rod-shaped bacterium *E. coli*.[65] Although the time to generate each individual event (fourth column) scales well with the number of voxels (second column), the total time to simulate the system (third column) grows rapidly. For the finest mesh resolution, 99.98% of all events are diffusive transfers. Thus, the stochastic simulation of reaction-diffusion systems with ever finer spatial resolution eventually becomes dominated by transfers of individual molecules between adjacent voxels. This will be so whether or not the reactions are diffusion-limited. Algorithm development to increase the speed of simulations of reaction-diffusion systems has therefore focused on ways to reduce the cost of diffusive transfers through approximations that aggregate the diffusion events in order to update the system's state.

To that end, Rossinelli *et al.*[75] have proposed a combination of tau-leaping, hybrid tau-leaping and deterministic diffusion. Iyengar *et al.*[76] and Marquez-Lago and Burrage[77] have introduced and compared additional implementations of spatial tau-leaping. But the efficiency of explicit tau-leaping in a spatial context is severely limited by the fact that it is necessary to generate one Poisson random number in every outgoing direction (edge) of each vertex in the mesh in order for the method to conserve total copy number. This tends to negate the performance benefit from aggregating diffusion events in the voxels. Lampoudi *et al.*[78] proposed to deal with this issue by using a multinomial simulation algorithm that aggregates molecular transfers into and out of voxels between each chemical reaction event, simulating only the net diffusional transfers. Koh and Blackwell[79] likewise simulate only net diffusional transfers between voxels, but then use tau leaping, instead of the SSA, with a leap time that is sensitive to both chemical reactions and the diffusive transfers. Another innovative approach is the diffusive finite state projection (DFSP) algorithm of Drawert *et al.*[80] The DFSP is conceptually related to the multinomial method, but it achieves better efficiency and flexibility through numerical solution of local (in space) approximations to the diffusion equation (13). For systems that exhibit scale separation, hybrid methods can achieve good speedups over pure stochastic simulations. Ferm *et al.*[81] propose a space-time adaptive hybrid algorithm where deterministic diffusion, explicit tau-leaping, and the next subvolume method are combined in such a way that the appropriate method is dynamically chosen in each voxel based on the expected errors in the different methods. In both of these last strategies, operator splitting is used to decouple diffusion and reactions and to propagate the hybrid system in time.

While approximate and hybrid methods hold promise for making spatial stochastic simulation feasible for large systems, many challenges remain to be met before such methods are robust enough to be an alternative to exact algorithms for most users. In particular, goal-oriented error estimation strategies and (spatial and temporal) adaptivity have yet to be developed. Another challenging aspect is parallel implementations of the simulation algorithms; frequent diffusive transfers between neighboring voxels severely limit the performance advantage of parallel implementations, because they introduce extensive communication.

C. The RDME on small length scales

The popularity of the well-mixed, dilute SSA stems in large part from the ease with which it can be implemented and from its robustness; there are no parameters that need to be tuned, and it is exact. The error in a simulation hence stems from modeling error only, which makes the method easy to use and to interpret. Spatial simulations based on the RDME inherit the ease of implementation of the well-mixed, dilute SSA but unfortunately not its robustness. As the voxel size is decreased, the accuracy first improves, thanks to smaller discretization error in the diffusion. But as the voxel size approaches the diameter of a reactant molecule, the RDME will give unphysical results for systems with bimolecular reactions, since in that case the RDME's requirement that the two reactant molecules must be in the same voxel in order to react leads to too slow association kinetics.

It can be seen from the discussion in Sec. III A that a major assumption of the RDME is that the standard CME and SSA must apply inside each voxel. That means, in particular, that propensity functions must exist for bimolecular reactions inside each voxel. That requirement will be problematic if the voxel size is comparable to the sizes of the reactant molecules. The physics derivation of the bimolecular propensity function (3b) for any system volume Ω assumes that the reactant molecules are "dilute," in the sense that the total volume occluded by all the reactant molecules is negligibly small compared to Ω. Therefore, the straightforward strategy of taking the bimolecular propensity function inside voxel k to be Eq. (3b) with the replacements $\Omega \to \Omega_k$ and $x_i \to x_{ik}$ will be physically justified only if the total volume occluded by all reactant molecules inside Ω_k is a negligibly small fraction of the voxel volume h^3. If that is not true, the form of the bimolecular propensity function (3b), and in particular its dependence on the variables Ω_k and x_{ik}, will change in some unclear but potentially dramatic way. Accounting for excluded volume effects has been shown to be straightforward for a one-dimensional system of hard-rod molecules: the total volume of the system simply needs to be decreased by the volume actually occupied by the molecules.[82] But the correction to the bimolecular propensity function for two-dimensional hard-disk molecules is not that simple,[83] and presumably that is also true for three-dimensional molecules. Grima[84] has studied the effects of crowded cellular conditions

in two dimensions for the reversible dimerization reaction by constructing a master equation in which the propensity functions have been renormalized using concepts from the statistical mechanics of hard sphere molecules.

One way to view the RDME is as a coarse-grained approximation to the continuous Smoluchowski diffusion-limited reaction (SDLR) model[13] which underlies particle tracking simulation methods such as Green's function reaction dynamics.[85] There, two molecules are assumed to move according to Eq. (13) and to react with a certain probability at the contact point between the two hard spheres. The distance at which they react is determined by the sum of the molecules' reaction radii ρ. The probability of a bimolecular reaction is governed by the diffusion equation supplemented with a partially absorbing boundary condition: given an initial relative position r_0 at time t_0, the PDF p of the new relative position (in a spherical coordinate system $\mathbf{r} = (r, \theta, \phi)$) is taken to be the solution of Eq. (13) subject to the initial condition $p(\mathbf{r}, t_0) = \delta(\mathbf{r} - \mathbf{r}_0)$ and the boundary conditions:

$$\lim_{r\to\infty} p(r,t) = 0, \quad 4\pi\rho^2 D \left.\frac{\partial p(r,t)}{\partial r}\right|_{r=\rho} = k_r\, p(\rho,t).$$

Here, D is the sum of the diffusion coefficients of the reacting molecules. And k_r is an assumed microscopic "association rate," which the physics-based derivation of Eq. (3b) shows is given by $k_r = \pi\sigma_{12}^2 \bar{v}_{12} q_j$, where $\sigma_{12} = \rho$.

Motivated by the observation that for highly diffusion limited reactions, the error in RDME simulations incurred by too small voxels can be substantial,[86,87] recent work on the RDME has tried to understand to what extent and in what sense the RDME approximates the SDLR model on short length scales, where the assumption $h \gg \rho$ does not hold. Isaacson[88] considered a bimolecular reaction and expanded the RDME to second order in the molecules' reaction radius to show that, for a given value of ρ, the second order term in the expansion diverges as h^{-1} compared to the corresponding term in an expansion of the solution of the Smoluchowski equation. He suggested that in order for the RDME to better approximate the microscopic model, it is necessary to "appropriately renormalize the bimolecular reaction rate and/or extend the reaction operator to couple in neighboring voxels." Isaacson and Isaacson[89] demonstrated that for a given value of h, the RDME can be viewed as an asymptotic approximation to the SDLR model in ρ.

Hellander *et al.*[90] gave an alternative explanation of the RDME breakdown based on the mean binding times of two particles performing a random walk on the lattice in 2 and 3 dimensions. Figure 3 is adapted from Ref. 90, and shows a schematic representation of the RDME's behavior as a function of the mesh size. For $h < \rho$, i.e., voxels smaller than the molecular reaction radius, the RDME makes little sense physically. In the other extreme, above $h_{\max}$, discretization errors due to large voxels will be unacceptably high. For $h_{\min} < h < h_{\max}$ (green region) the RDME will work well, but for $h < h_{\min}$ it can yield increasingly unphysical results. For $h < h^*$ the conventional RDME and the SDLR model cannot be made consistent in the sense that the mean binding time between two particles in the RDME converges to that of the microscopic model.[90] The values of $h_{\min}$, $h_{\max}$, and h^* are model and geometry dependent, but in the limit of perfect diffusion control, a box-geometry, and a uniform Cartesian mesh, the critical voxel sizes take the values $h^* = \pi\rho$ (3D) and $h^* \approx 5.2\rho$ (2D).

Two main approaches have been proposed to improve on the robustness of simulation with the RDME when small length scales need to be considered. The first relies on modification of the bimolecular association reaction rate k_a, as suggested by Isaacson and Isaacson.[89] For a given k_a and a Cartesian discretization, Erban and Chapman[86] derived a new rate expression by requiring that the spatially independent steady-state distribution for a model problem solved with the RDME be invariant under changes to the voxel size. Fange *et al.*[87] derived mesh dependent propensities in both 2D and 3D based on the ansatz that the equilibration time for a reversible bimolecular reaction should be the same in the SDLR model and the RDME. Furthermore, they allow for reactions between molecules in neighboring voxels. In this way, they obtain good agreement in numerical experiments between the two models, for mesh resolutions close to the reaction radius ρ. A third set of corrected rate functions was obtained by Hellander *et al.*[90] in both 2D and 3D. A problematic aspect of relying on mesh-dependent rate functions is that different approaches lead to different expressions, and they are dependent on the nature of the voxels, the geometry, and the test problem. Another approach to the problem was recently taken by Isaacson,[91] where he constructs a new and convergent form of the RDME based on a discretization of a particle tracking model.[92] Here, the mesoscopic model is formulated in such a way as to converge to a specific microscopic, continuum model per construction.

The other main approach that has been proposed to make simulations more robust is the use of mesoscopic-microscopic hybrid methods, that switch to the microscopic model whenever microscale resolution is required. Hellander *et al.*[93] use an RDME (mesoscopic) model in combination with a Smoluchowski GFRD (microscopic) model. The microscopic

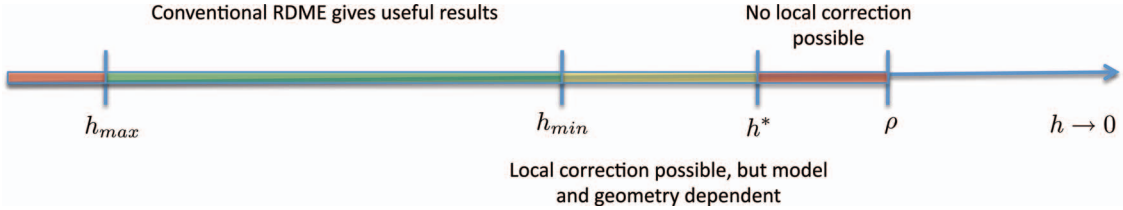

FIG. 3. Schematic representation of the RDME's behavior as a function of the voxel size h. For $h < h^*$, no local correction to the conventional mesoscopic reaction rates exists that will make the RDME consistent with the Smoluchowski model for the simple problem of diffusion to a target. Figure adapted from Ref. 90.

model is used only in the regions of the domain that require a high spatial resolution, where the RDME is not well-defined (such as for binding to a curved membrane), or for those reactions that are strongly diffusion limited. With only a small fraction of the molecules treated microscopically, the hybrid method is capable of accurately resolving the features of the model in Ref. 60. Flegg *et al.*[94] studied pure diffusion of particles and focused on the accurate treatment of the transition between the mesoscopic and microscopic methods at an interface in 1D.

In contrast to the approach where the RDME is modified in different ways to support microscopically small voxels, hybrid methods have the potential to greatly speed up simulation of models with scale separation in species molecular populations if microscopic resolution is needed only for parts of the model, or in parts of the computational domain. For the majority of the model, relatively large voxels can be used. However, several issues must be resolved before a mesoscopic-microscopic hybrid method can become a general purpose tool. For example, in Ref. 93 a static partitioning is chosen *a priori*, which requires information on the degree of diffusion control for a reaction. Criteria to partition a system automatically and adaptively are needed.

IV. ACCOMPLISHMENTS AND CHALLENGES

The area of discrete stochastic simulation of chemically reacting systems has seen many advances in algorithms and theory. During the past decade, fast exact and approximate algorithms have been developed for well-mixed systems, multiscale issues have been addressed, and efficient and robust algorithms have been developed for characterizing rare events and estimating parameters. Spatial stochastic simulation is rapidly becoming better understood, and algorithms and software for spatial stochastic simulation on unstructured meshes have begun to appear. There is a great deal left to be done to accommodate models of all of the mechanisms that occur, for example, in cell biology. At the same time, we have come a long way and have developed an extensive collection of algorithms and theory.

One of the major challenges we see at this point is to make these advances available to practitioners in a form that will allow both flexibility and ease of use. We envision an integrated software environment that makes it easy to build a model, scale it up to increasing complexity including spatial simulation, explore the parameter space, and seamlessly deploy the appropriate computing resources as needed. To this end, we have recently begun to develop a new software platform, StochSS (Stochastic Simulation Service, *www.stochss.org*). It will incorporate, in addition to ODE and PDE solvers, well mixed stochastic simulations via the algorithms of StochKit2,[95] and spatial stochastic simulations via the algorithms of URDME.[72] StochSS will also enable the use of a wide range of distributed cloud and cluster computing resources to make the generation of large ensembles of realizations and large parameter sweeps possible, greatly facilitating a careful statistical analysis for even the most expensive simulations.

ACKNOWLEDGMENTS

The work of D.T.G. was funded by the University of California, Santa Barbara under professional services Agreement No. 130401A40, pursuant to National Institutes of Health (NIH) Award No. R01-EB014877-01. The work of A.H. and L.R.P. was funded by National Science Foundation (NSF) Award No. DMS-1001012, ICB Award No. W911NF-09-0001 from the U.S. Army Research Office, NIBIB of the NIH under Award No. R01-EB014877-01, and (U.S.) Department of Energy (DOE) Award No. DE-SC0008975. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of these agencies.

[1] L. Wilhelmy, Ann. Phys. Chem. **81**, 413 (1850); **81**, 499 (1850).
[2] M. Delbrück, J. Chem. Phys. **8**, 120 (1940).
[3] D. McQuarrie, J. Appl. Probab. **4**, 413 (1967).
[4] H. McAdams and A. Arkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 814 (1997).
[5] A. Arkin, J. Ross, and H. McAdams, Genetics **149**, 1633 (1998).
[6] M. Elowitz, A. Levine, E. Siggia, and P. Swain, Science **297**, 1183 (2002).
[7] L. Weinberger, J. Burnett, J. Toettcher, A. Arkin, and D. Schaffer, Cell **122**, 169 (2005).
[8] If the observer knows the initial state $\mathbf{x}_0$ only as a random variable with some PDF $Q_0(\mathbf{x}_0)$, then the CME straightforwardly transforms, without change, to an equation for $\sum_{\mathbf{x}_0} P(\mathbf{x}, t|\mathbf{x}_0, t_0)Q_0(\mathbf{x}_0)d\mathbf{x}_0$. But that function is not an "unconditioned" PDF $P(x, t)$, because it is obviously conditioned on the initial distribution Q_0. The essential role of an observer in the CME is also illustrated by the fact that in many cases, $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$ becomes independent of t as $t \to \infty$ while $\mathbf{X}(t)$ does not; in that case, the only thing that eventually stops changing with time is best prediction of $\mathbf{X}$ that can be made by an observer who last observed $\mathbf{X}$ at time t_0.
[9] D. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[10] D. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[11] The derivation of Eq. (3a) that was given in Ref. 9 can be briefly summarized as follows: $\pi\sigma_{12}^2 \cdot \bar{v}_{12}dt$ is the average "collision volume" that a randomly chosen S_2 molecule sweeps out relative to the center of a randomly chosen S_1 molecule in time dt. Dividing that collision volume by the system volume Ω gives, because the system is dilute and well-mixed, the probability that the center of the S_1 molecule lies inside the collision volume, and hence the probability that the two molecules will collide in the next dt. That collision probability multiplied by q_j gives the probability that the two molecules will react in the next dt. And finally, that single-pair reaction probability summed over all x_1x_2 distinct reactant pairs gives the probability defined in Eq. (2). If the two reactant molecules are of the same species, say S_1, then the number of distinct reactant pairs will instead be $x_1(x_1 - 1)/2$. Determining the collision-conditioned reaction probability q_j, which will always be some number between 0 and 1, requires additional physical reasoning. The best known example is for the model in which an R_j reaction will occur between the two colliding molecules if and only if their "collisional kinetic energy," suitably defined, exceeds some threshold value E_{th}. In that case it can be shown [see for instance D. Gillespie, Physica A **188**, 404 (1992)] that $q_j = \exp(-E_{\text{th}}/k_{\text{B}}T)$, which is the famous Arrhenius factor.
[12] D. Gillespie, J. Chem. Phys. **131**, 164109 (2009). As discussed in Sec. VI of this reference, the analysis leading to the result (3b) can be regarded as a refined, corrected, and stochastically extended version of the analysis of F. Collins and G. Kimball, J. Colloid Sci. **4**, 425 (1949). A derivation of Eq. (3b) that is further improved over the one given in the 2009 paper can be found in Secs. 3.7 and 4.8 of D. Gillespie and E. Seitaridou, *Simple Brownian Diffusion* (Oxford University Press, 2012).
[13] M. Smoluchowski, Z. Phys. Chem. **92**, 129 (1917).
[14] T. Nakanishi, J. Phys. Soc. Jpn. **32**, 1313 (1972); **40**, 1232 (1976).
[15] M. Šolc and I. Horsák, Collect. Czech. Chem. Commun. **37**, 2994 (1972); **38**, 2200 (1973); **40**, 321 (1975).
[16] D. Bunker, B. Garrett, T. Kleindienst, and G. Long III, Combust. Flame **23**, 373 (1974).
[17] The inversion Monte Carlo method exploits the theorem that, if F_Y is the cumulative distribution function (CDF) of the random variable Y, then $F_Y(Y)$ will be the unit-interval uniform random variable U. Thus, if u is a random

sample of U, then solving (inverting) $F_Y(y) = u$ will yield a random sample y of Y. Writing Eq. (4) as the product of $a_0(\mathbf{x})e^{-a_0(\mathbf{x})\tau}$ and $a_j(\mathbf{x})/a_0(\mathbf{x})$ reveals that τ and j are statistically independent random variables with those respective PDFs. Integrating (summing) those PDFs over their respective arguments gives their CDFs, and Eqs. (5a) and (5b) then follow from the inversion method. The generic forms of Eqs. (5a) and (5b), being such straightforward consequences of applying the inversion Monte Carlo method to the aforementioned PDFs of τ and j, were known long before Gillespie's 1976 paper;[9] e.g., see page 36 of J. Hammersley and D. Handscomb, *Monte Carlo Methods* (Methuen, 1964). The main contributions of Gillespie's 1976 paper[9] were: (i) proving from simple kinetic theory that bimolecular reactions in a dilute gas, like unimolecular reactions, are describable by propensity functions as defined in (2); and (ii) proving that propensity functions as defined in (2) imply that the "time to next reaction" and the "index of next reaction" are random variables distributed according to the joint PDF (4).

[18] M. Gibson and J. Bruck, J. Phys. Chem. **104**, 1876 (2000).

[19] D. Gillespie, Annu. Rev. Phys. Chem. **58**, 35 (2007).

[20] Y. Cao, H. Li, and L. Petzold, J. Chem. Phys. **121**, 4059 (2004).

[21] J. McCollum, G. Peterson, C. Cox, M. Simpson, and N. Samatova, Comput. Biol. Chem. **30**, 39 (2006).

[22] D. Anderson, J. Chem. Phys. **127**, 214107 (2007).

[23] T. Kurtz, Ann. Probab. **8**, 682 (1980); S. Ethier and T. Kurtz, *Markov Processes: Characterization and Convergence* (Wiley, 1986).

[24] D. Anderson and T. Kurtz, "Continuous time Markov chain models for chemical reaction networks," in *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, edited by H. Koeppl *et al*. (Springer, 2011).

[25] A. Slepoy, A. Thompson, and S. Plimpton, J. Chem. Phys. **128**, 205101 (2008).

[26] S. Mauch and M. Stalzer, IEEE/ACM Trans. Comput. Biol. Bioinf. **8**, 27 (2011).

[27] D. Gillespie, J. Chem. Phys. **115**, 1716 (2001).

[28] An early conjecture was that the negativity problem in tau-leaping was caused by the unbounded Poisson random variables in the tau-leaping formula (7) occasionally allowing too many reaction firings in a single leap. That conjecture led to proposals to replace the Poisson random variables in Eq. (7) with binomial random variables, which are bounded. But it was subsequently determined that the principal causes of the negativity problem lay elsewhere. First was the flawed implementation of the leap condition (6) that was used in Ref. 27: the small-valued propensity functions that it failed to protect often have a reactant with a small molecular population which can easily be driven negative. Second, since the firing numbers of the individual reactions in the tau-leaping formula (7) are generated independently of each other, two or more reaction channels that decrease the population of a common species could inadvertently collude to overdraw that species. Neither of these two problems is fixed by the ad hoc substitution of binomial random variables for the Poisson random variables in Eq. (7). But both problems are effectively dealt with in the heavily revised tau-leaping procedure described in Ref. 29, which uses the theoretically appropriate Poisson random variables.

[29] Y. Cao, D. Gillespie, and L. Petzold, J. Chem. Phys. **124**, 044109 (2006).

[30] D. Gillespie, "Simulation methods in systems biology," in *Formal Methods for Computational Systems Biology*, edited by M. Bernardo, P. Degano, and G. Zavattaro (Springer, 2008), pp. 125–167, Sec. 3 gives a tutorial on tau-leaping, Secs. 4 and 5 give a tutorial on the ssSSA.

[31] M. Rathinam, L. Petzold, Y. Cao, and D. Gillespie, J. Chem. Phys. **119**, 12784 (2003); **121**, 12169 (2004); Multiscale Model. Simul. **4**, 867 (2005).

[32] A. Auger, P. Chatelain, and P. Koumoutsakos, J. Chem. Phys. **125**, 084103 (2006).

[33] D. Anderson, J. Chem. Phys. **128**, 054103 (2008).

[34] The view we have taken here of dt as an independent real variable on the interval $[0, \varepsilon)$, where ε is an arbitrarily small positive number, means that $\sqrt{dt}$ in Eq. (9b) is perfectly well defined. For an explanation of the connection between the factor $\mathcal{N}_j(0,1)\sqrt{dt}$ in Eq. (9b) and "Gaussian white noise," see D. Gillespie, Am. J. Phys. **64**, 225 (1996); **64**, 1246 (1996).

[35] D. Gillespie, J. Chem. Phys. **113**, 297 (2000). The CLE (9b) can be shown to be mathematically equivalent to the Fokker-Planck equation that is obtained by first making a formal Taylor series expansion of the right side of the CME (1), which yields the so-called Kramers-Moyal equation, and then truncating that expansion after the second derivative term. However, that way of "obtaining" the CLE, which was known well before this 2000 paper, does not qualify as a derivation, because it does not make clear under what conditions the truncation will be accurate. In contrast, the derivation of the CLE given here provides a clear and testable criterion for the accuracy of its approximations, namely, the extent to which both leap conditions are satisfied. But see also the proviso in Ref. 36.

[36] The approximation $\mathcal{P}(m) \approx \mathcal{N}(m, m)$ that was made in deriving the CLE from the tau-leaping formula (7), while accurate for likely values of those two random variables, is very inaccurate in the tails of their probability densities, even when $m \gg 1$. Although both tails are "very near 0," they differ by many orders of magnitude. Since rare events arise from the "unlikely" firing numbers under those tails, it follows that the CLE will *not* accurately describe the *atypical* behavior of a chemical system, even if both leap conditions are well satisfied.

[37] D. Gillespie, J. Phys. Chem. B **113**, 1640 (2009).

[38] The foregoing chain of reasoning can be summarized from the perspective of computational mathematics as follows: The definition (2) of the propensity function implies, for any time step τ that is small enough to satisfy the first leap condition (6), the tau-leaping formula (7). If the second leap condition (8) is also satisfied, the tau-leaping formula becomes Eq. (9a), which is the forward Euler formula for a stochastic differential equation. And that formula becomes in the thermodynamic limit, where its diffusion term will be negligibly small compared to its drift term, the forward Euler formula for an ordinary differential equation.

[39] A common misconception is that, while the molecular population X is obviously a discrete variable, the molecular concentration $Z \equiv X/\Omega$ is a "continuous" variable. The error in that view becomes apparent when one realizes that simply by adopting a unit of length that gives Ω the value 1, Z becomes numerically equal to X. But even if that is not done, a sudden change in X, say from 10 to 9, will always result a discontinuous 10% decrease in Z. The molecular concentration Z is no less discrete, and no more continuous, than the molecular population X.

[40] T. Kurtz, Stochastic Proc. Appl. **6**, 223 (1978).

[41] N. van Kampen, Adv. Chem. Phys. **34**, 245 (1976); *Stochastic Processes in Physics and Chemistry* (North-Holland, 1992).

[42] E. Wallace, D. Gillespie, K. Sanft, and L. Petzold, IET Syst. Biol. **6**, 102 (2012).

[43] R. Grima, P. Thomas, and A. Straube, J. Chem. Phys. **135**, 084103 (2011).

[44] The last step in Eq. (11) can be heuristically understood as follows: The integral $\int_{t_0}^{t} \mathcal{P}_j\left(a_j\left(\mathbf{X}(t')\right)dt'\right)$ is essentially a sum of independent Poisson random variables, each indexed by t' and having mean $a_j(\mathbf{X}(t'))\,dt'$. An established result in random variable theory is that the sum of K independent Poisson random variables with means $m_1, \ldots, m_K$ is a Poisson random variable with mean $\sum_{k=1}^{K} m_k$. Therefore, the integral $\int_{t_0}^{t} \mathcal{P}_j(a_j(\mathbf{X}(t'))\,dt')$ in Eq. (11) is a Poisson random variable with mean $\int_{t_0}^{t} a_j(\mathbf{X}(t'))\,dt'$. For an explanation of how that Poisson random variable can be viewed as a "unit-rate Poisson process" Y_j that is "scaled" by $\int_{t_0}^{t} a_j\left(\mathbf{X}(t')\right)dt'$, see Refs. 23 and 24. Note that all the integrals here can be written as finite algebraic sums, since $\mathbf{X}(t')$ stays constant between successive reactions.

[45] Y. Cao, D. Gillespie, and L. Petzold, J. Chem. Phys. **122**, 014116 (2005); **123**, 144917 (2005). A more concise tutorial presentation of the ssSSA, featuring illustrative applications to reactions (11) and the classical enzyme-substrate reactions $E + S \rightleftharpoons ES \rightarrow E + P$, is given in Ref. 30. A more thorough examination of the relation between the ssSSA's treatment of the enzyme-substrate system and the classical Michaelis-Menten approximation is given in K. Sanft, D. Gillespie, and L. Petzold, IET Syst. Biol. **5**, 58 (2011).

[46] Under the condition $c_2 \gg c_3$, R_1 will be a "fast" reaction regardless of the size of c_1. That is because the conversion of an S_1 molecule into an S_3 molecule necessarily takes exactly one more R_1 reaction than R_2 reaction. Thus, in most time intervals there will be at least as many R_1 firings as R_2 firings; hence, if R_2 is fast, then R_1 must be also. This illustrates the important fact that fast and slow reaction channels cannot always be identified solely on the basis of the magnitudes of their rate constants.

[47] Y. Cao, D. Gillespie, and L. Petzold, J. Comput. Phys. **206**, 395 (2005).

[48] L. Ferm, P. Lotstedt, and A. Hellander, J. Sci. Comput. **34**, 127 (2008).

[49] A. Singh and J. Hespanha, IEEE Trans. Autom. Control. **56**, 414 (2011).

[50] W. E, D. Liu, and E. Vanden-Eijnden, J. Chem. Phys. **123**, 194107 (2005); D. Gillespie, L. Petzold, and Y. Cao, *ibid.* **126**, 137101 (2007).

[51] K. Sanft, Ph.D. thesis, University of California, Santa Barbara, 2012.

[52] H. Kuwahara and I. Mura, J. Chem. Phys. **129**, 165101 (2008).

[53] D. Gillespie, M. Roh, and L. Petzold, J. Chem. Phys. **130**, 174103 (2009).

[54] M. Roh, D. Gillespie, and L. Petzold, J. Chem. Phys. **133**, 174106 (2010).

[55] B. Daigle, Jr., M. Roh, D. Gillespie, and L. Petzold, J. Chem. Phys. **134**, 044110 (2011).

[56]M. Roh, B. Daigle, Jr., D. Gillespie, and L. Petzold, J. Chem. Phys. **135**, 234108 (2011).
[57]R. Rubinstein and D. Kroese, *The Cross-Entropy Method* (Springer, 2004).
[58]M. Rathinam, P. Sheppard, and M. Khammash, J. Chem. Phys. **132**, 034103 (2010).
[59]D. Anderson, SIAM J. Numer. Anal. **50**, 2237 (2012).
[60]K. Takahashi, S. Tănase-Nicola, and P. ten Wolde, Proc. Natl. Acad. Sci. U.S.A. **107**, 2473 (2010).
[61]K. Doubrovinski and M. Howard, Proc. Natl. Acad. Sci. U.S.A. **102**, 9808 (2005).
[62]J. Elf and M. Ehrenberg, Syst. Biol. **1**, 230 (2004).
[63]R. Metzler, Phys. Rev. Lett. **87**, 068103 (2001).
[64]M. Sturrock, A. Hellander, M. Matzavinos, and M. Chaplain, J. R. Soc. Interface **10**, 20120988 (2013).
[65]D. Fange and J. Elf, PLoS Comput. Biol. **2**(6), e80 (2006).
[66]C. Gardiner, K. McNeil, D. Walls, and I. Matheson, J. Stat. Phys. **14**, 307 (1976).
[67]See, for example, D. Gillespie and E. Seitaridou, *Simple Brownian Diffusion* (Oxford University Press, 2012), Chap. 5. That chapter's *revised* Sec. 5.6 (which can be downloaded from the book's webpage on the publisher's website) shows that below a certain value of h, no propensity function can give a physically accurate modeling of diffusive transfers of molecules between voxels.
[68]Equation (14) also follows by applying the "centered finite difference method" in numerical analysis to the diffusion Eq. (13).
[69]T. Kurtz, J. Appl. Probab. **7**, 49 (1970).
[70]S. Isaacson and C. Peskin, SIAM J. Sci. Comput. **28**, 47 (2006).
[71]S. Isaacson, D. McQueen, and C. Peskin, Proc. Nat. Acad. Sci. U.S.A. **108**, 3815 (2011).
[72]S. Engblom, L. Ferm, A. Hellander, and P. Lötstedt, SIAM J. Sci. Comput. **31**, 1774 (2009).
[73]B. Drawert, S. Engblom, and A. Hellander, BMC Syst. Biol. **6**, 76 (2012).
[74]H.-W. Kang, L. Zheng, and H. Othmer, J. Math. Biol. **65**, 1017 (2012).
[75]D. Rossinelli, B. Bayati, and P. Koumoutsakos, Chem. Phys. Lett. **451**, 136 (2008).
[76]K. Iyengar, L. Harris, and P. Clancy, J. Chem. Phys. **132**, 094101 (2010).
[77]T. Marquez-Lago and K. Burrage, J. Chem. Phys. **127**, 104101 (2007).
[78]S. Lampoudi, D. Gillespie, and L. Petzold, J. Chem. Phys. **130**, 094104 (2009).
[79]W. Koh and K. Blackwell, J. Chem. Phys. **134**, 154103 (2011).
[80]B. Drawert, M. Lawson, L. Petzold, and M. Khammash, J. Chem. Phys. **132**, 074101 (2010).
[81]L. Ferm, A. Hellander, and P. Lötstedt, J. Comput. Phys. **229**, 343 (2010).
[82]D. Gillespie, S. Lampoudi, and L. Petzold, J. Chem. Phys. **126**, 034302 (2007).
[83]S. Lampoudi, D. Gillespie, and L. Petzold, J. Comput. Phys. **228**, 3656 (2009).
[84]R. Grima, J. Chem. Phys. **132**, 185102 (2010).
[85]J. Zon, S. van Zon, and P. Rein ten Wolde, Phys. Rev. Lett. **94**, 128103 (2005).
[86]R. Erban and J. Chapman, Phys. Biol. **6**, 046001 (2009).
[87]D. Fange, O. G. Berg, P. Sjöberg, and J. Elf, Proc. Natl. Acad. Sci. U.S.A. **107**, 19820 (2010).
[88]S. Isaacson, SIAM J. Appl. Math. **70**, 77 (2009).
[89]S. Isaacson and D. Isaacson, Phys. Rev. E **80**, 066106 (2009).
[90]S. Hellander, A. Hellander, and L. Petzold, Phys. Rev. E **85**, 042901 (2012).
[91]S. Isaacson, "A convergent reaction-diffusion master equation," preprint arXiv:1211.6772v1 (2012).
[92]M. Doi, J. Phys. A: Math. Gen. **9**, 1465 (1976); **9**, 1479 (1976).
[93]A. Hellander, S. Hellander, and P. Lötstedt, Multiscale Model. Simul. **10**, 585 (2012).
[94]M. Flegg, S. Chapman, and R. Erban, J Roy. Soc. Interface **9**, 859 (2012).
[95]K. Sanft, S. Wu, M. Roh, J. Fu, R. Lim, and L. Petzold, Bioinformatics **27**, 2457 (2011).