Adiabatic coarse-graining and simulations of stochastic biochemical networks

N. A. Sinitsyn[a,b], Nicolas Hengartner[b], and Ilya Nemenman[a,b,1]

[a]Center for Nonlinear Studies, and [b]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

Edited by William H. Press, University of Texas, Austin, TX, and approved April 10, 2009 (received for review September 18, 2008)

We propose a universal approach for analysis and fast simulations of stiff stochastic biochemical networks, which rests on elimination of fast chemical species without a loss of information about mesoscopic, non-Poissonian fluctuations of the slow ones. Our approach is similar to the Born–Oppenheimer approximation in quantum mechanics and follows from the stochastic path integral representation of the cumulant generating function of reaction events. In applications with a small number of chemical reactions, it produces analytical expressions for cumulants of chemical fluxes between the slow variables. This allows for a low-dimensional, interpretable representation and can be used for high-accuracy, low-complexity coarse-grained numerical simulations. As an example, we derive the coarse-grained description for a chain of biochemical reactions and show that the coarse-grained and the microscopic simulations agree, but the former is 3 orders of magnitude faster.

Computer simulations are often the method of choice to explore an agreement between a model and experimental data in systems biology. Unfortunately, even the simplest biochemical simulations often face serious conceptual and practical problems. First, they usually involve combinatorially many chemical species and reaction processes: for example, a single molecule with n modification sites can exist in 2^n microscopic states (1). Second, although it is widely known that some molecules occur in cells at very low copy numbers (e.g., the DNA), which give rise to important stochastic effects, it is less appreciated that the combinatorial complexity makes this true for many molecular species. Indeed, even for a large total number of molecules, typical abundances of a species may be small if the number of the species is combinatorial. Third, and perhaps the most profound difficulty, is that only very few of the kinetic parameters underlying the networks are experimentally observable.

Although some day, computers may be able to tackle the formidable problem of modeling combinatorially complex biochemical processes and then performing sweeps through parameter spaces in search of an agreement with experiments, this day is far away. More importantly, even if the computing power were available, it would not help in building a comprehensible interpretation of the modeled system and in identifying connections between its microscopic and macroscopic features.

Clearly, such an interpretation can be aided by coarse-graining, that is, by merging or eliminating certain nodes and/or reaction processes (this would be called blocking or decimation in statistical physics). Ideally, one wants to substitute multiple elementary (that is, single-step, Poisson-distributed) biochemical reactions with a few complex processes in a way that retains predictability of the system. Not incidentally, this would help with each of the 3 roadblocks mentioned above by reducing the number of interacting elements, increasing the copy numbers of agglomerated hyperspecies, and combining multiple microscopic rates into a smaller number of effective parameters.

Coarse-graining in biochemistry is well established, and the prime example is the Michaelis–Menten (MM) kinetics (2)

$$E + S \underset{k_{-1}[C]}{\overset{k_1[E][S]}{\rightleftharpoons}} C \xrightarrow{k_2[C]} E + P. \qquad [1]$$

Here k_1, k_2, and k_{-1} are kinetic rates, S, P, E, and C denote the substrate, the product, the enzyme, and the enzyme–substrate complex molecules, respectively, and $[\dots]$ represent the abundances. The enzyme catalyzes the $S \to P$ transformation by merging with S to create an unstable complex C, which then dissociates either back into $E+S$ or forward into $E+P$, leaving E unmodified. If $[S] \gg [E]$, then the enzyme cycles many times before $[S]$ changes appreciably. Thus the enzyme equilibrates resulting in a coarse-grained reaction with the decimated enzyme species:

$$S \xrightarrow{v} P, \quad v = \frac{k_2[S][E]}{[S] + (k_2 + k_{-1})/k_1}. \qquad [2]$$

However, this simple reduction is insufficient when stochasticity is important: Each MM reaction consists of multiple elementary steps, thus approximating the number of the reactions as a Poisson variable (3) is not always valid. While some attempts have been made to extend deterministic coarse-graining to the stochastic domain (4–7), such systematic tools have not been found yet. In this article, we make a step towards the goal.

We start by noting that, in addition to the 3 conceptual problems, a technical one stands in the way of stochastic simulations in systems biology: Molecular species have diverse dynamical time scales, making the systems stiff and difficult to simulate. We propose to use this property to our advantage, finding a coarse-graining procedure exhibiting the following 4 features.

First, like in the deterministic MM case, fast variables must not only be treated differently from the slow ones, but they must be eliminated altogether. Otherwise, the coarse-graining would not decrease the complexity of the interpretation and of numerical simulations, which scale at least linearly with the number of the involved variables.

Second, the distinction between the fast and the slow variables must not be based on reaction rates. For example, for the MM scheme, all 3 reaction rates may be comparable, and coarse-graining is still possible due to the difference in the enzyme and the substrate abundances. Overall, if 2 species of different abundances are coupled by a reaction, then a relatively small change in the high-abundance one can have a dramatic effect on the low-abundance one, leading to the different dynamical time scales. We seek the notion of species-rather than reaction-based adiabaticity as a basis for the coarse-graining. This has an additional advantage: Having higher abundances, coarse-grained variables will be amenable to fast mesoscopic, Langevin-like (8) methods, instead of event-by-event simulations (9).

Third, real biological systems have more than just fast and slow variables; instead a whole spectrum of time scales is usually

Author contributions: N.A.S. and I.N. designed research; N.A.S., N.H., and I.N. performed research; and N.A.S., N.H., and I.N. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ilya@menem.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0809340106/DCSupplemental.

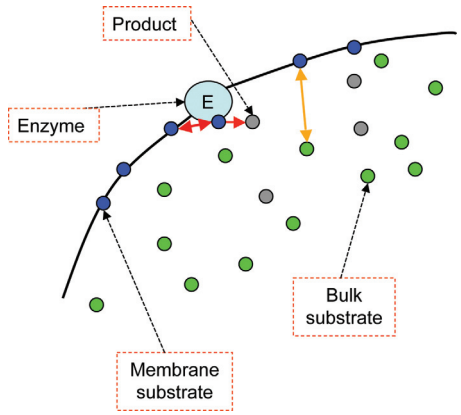

Fig. 1. The model system. Circles represent molecules and are labeled. Arrows stand for reactions: (1, 2) adsorption and dissociation of S (orange); (3) multistep MM conversion $S \to P$ (red).

present. A coarse-graining procedure must be applicable to such systems, eliminating time scales hierarchically.

While many adiabatic approaches have been explored, including refs. 3, 7, and 10–13, none possesses all 3 of the above features, leaving room for substantial improvements. The method we propose here reaches the goal by building upon the stochastic path integral (SPI) technique from mesoscopic physics (14, 15). To make the SPI applicable to biology, significant modifications are needed. First, we extend the technique to discrete degrees of freedom, such as a single enzyme, in addition to its usual mesoscopic domain. Second, we explain how to use the SPI for a network of multiple reactions, reducing them to a few complex reaction links. Finally, we show how the procedure can be turned into an efficient algorithm for coarse-grained simulations, preserving statistical characteristics of the original dynamics. As required, the algorithm is akin to the Langevin (8) or τ-leaping (16) schemes, but it simulates complex reactions in a single step. This development of a fast, yet precise numerical algorithm is the most important practical contribution of our work.

For pedagogical reasons, we develop the method using a model system that is simple enough for a detailed analysis, yet is complex enough to support our goals.

Model

Consider an enzyme on a cell membrane, Fig. 1. S_B substrate molecules are distributed over the bulk cell volume. Each can be reversibly adsorbed by the membrane, forming the species S_M. The enzyme interacts only with S_M. The enzyme–substrate complex C can split either into $E + S_M$ or into $E + P$. The latter reaction is observable; for example, a GFP tag sparks each time a product molecule is created (17). Finally, we assume that $C \to E + P$ is irreversible. This is a simple model of receptor signaling, such as in vision or immune system, or of a reaction-diffusion enzyme, where the membrane/bulk play the roles of the nearby/far away regions around the enzyme.

The full set of elementary reactions is

1. adsorption of the bulk substrate, $S_B \to S_M$ (rate $q_0 S_B$);
2. reemission of the substrate into the bulk, $S_M \to S_B$ (rate qS_M);
3. MM conversion of S_M into P, consisting of
   (a) complex formation, $S_M + E \to C$, (rate $k_1 S_M$);
   (b) complex backward decay, $C \to S_M + E$ (rate k_{-1});
   (c) product emission $C \to E + P$ (rate k_2).

Note that here and in the rest of the article, we do not distinguish between a species name and the number of its molecules.

Our goal is to coarse-grain the above system of 5 reaction processes into a single complex reaction $S_B \to P$, as in Fig. 2C,

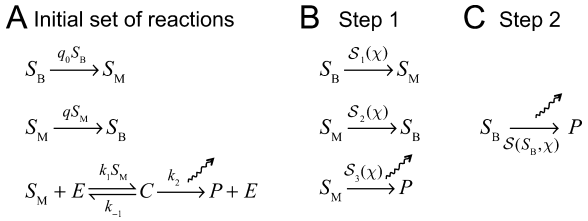

Fig. 2. Coarse-graining of the model system. (*A*) The original set of reactions. (*B*) The reactions after the first coarse-graining step: the MM mechanism has been replaced by a single complex reaction, and all the remaining reactions are now characterized by slowly varying CGFs. (*C*) The final reaction that describes the system at time scales $\delta t \gg \tau_M$. The wavy line corresponds to a spark of the tracer molecule (17), which counts $S_B \to P$ transformations.

eliminating all intermediate species and processes while preserving their effects on the statistics of the complex reaction on time scales appropriate for its dynamics.

Results

There are 3 effective time scales in our model. One is the scale τ_B of the variation of the bulk substrate abundance. We assume that $S_B \gg S_M$. Therefore, this scale is the slowest, and we will be interested in studying the response of the system to changes in S_B on this scale. A faster time scale, τ_M is given by the dynamics of S_M. Finally, the fastest scale, τ_E, is set by single reaction events, that is, the characteristic time between enzyme–substrate binding/unbinding. Overall, $\tau_E \ll \tau_M \ll \tau_B$. We emphasize that all species in the problem are connected by reactions that happen with similar rates, and the separation of the time scales is a result of the particle abundances only: It takes longer to change a high-abundance species.

The hierarchy of times allows us to coarse-grain the system in 2 steps, as in Fig. 2. First, we remove the variable with the fastest dynamics: the binary substrate–enzyme complex C. This replaces the 3 steps of the MM mechanism with a single reaction $S_M \to P$, Fig. 2B. Additionally, we represent the other reactions in a more convenient form. In the second step, we eliminate S_M, which changes on the scale τ_M. This results in the characterization of the average $S_B \to P$ flux and its fluctuations, treating S_B as a time-dependent input parameter (Fig. 2C).

Preliminaries. Because we are interested in adiabatic properties, single reaction events are not important, and we introduce δQ_μ—the mesoscopic number of reaction events for the reaction type μ ($\mu = 1, 2, 3$ corresponds to adsorption, detachment, and the MM reaction, respectively). Then $P(\delta Q_\mu|T)$ is the probability distribution of the number of events of type μ in a time window of duration T. Hierarchical coarse-graining of the reaction network would require convolutions of such distributions, which are easier to perform working with the corresponding moment generating functions (MGFs):*

$$\mathcal{Z}_\mu(\chi, T) = e^{\mathcal{S}_\mu(\chi,T)} = \sum_{\delta Q_\mu=0}^{\infty} P(\delta Q_\mu|T) e^{i\delta Q_\mu \chi}. \qquad [3]$$

where $\mathcal{S}_\mu(\chi)$ is the cumulant generating function (CGF). Then the cumulants of order a of $P(\delta Q_\mu|T)$ are

$$c_{\mu,a} = (-i)^a \left.\frac{\partial^a}{\partial \chi^a}\right|_{\chi=0} \mathcal{S}_\mu(\chi, T), \qquad [4]$$

In particular, the average fluxes are $c_{\mu,1}$, and the variances are $c_{\mu,2}$.

*The usual definition of the MGF is without i in the exponent. The same is true for our CGF, $\mathcal{S}_\mu$. We chose this nomenclature to emphasize that we use the functions for calculations of moments and cumulants.

Step 1: Generating Function Representation. This step can be viewed as a generalization of τ-leaping (16), which simulates elementary reactions, for example membrane binding in Fig. 2*A*, by choosing a time step δt, over which the number of the reactions is large, yet the slowly varying reaction rates are quasi-stationary. In τ-leaping, one then approximates $P(\delta Q_\mu|\delta t)$ as Poissons. Similarly, for $\tau_E \ll \delta t \ll \tau_M$, we can approximate CGFs of membrane binding/unbinding as those of Poisson processes, $\mathcal{S}_\mu(\chi) = r_\mu(t)(e^{i\chi} - 1)\delta t$, and the rates are $r_1 = q_0 S_B(t)$ and $r_2 = qS_M(t)$.

Unfortunately, not all biochemical processes are so simple. For example, for a single MM enzyme in Fig. 1, the instantaneous rate of the product creation is a fast varying function of time, switching between zero and k every time the complex forms. Therefore, one cannot treat the product creation, $P(\delta Q_3|\delta t)$, as a homogeneous Poisson process, and τ-leaping is inapplicable. Still, we would like to avoid the Gillespie (9) or similar techniques, which track individual reaction events and are slow.

As an alternative, we derive an approximation for the non-Poisson distribution of δQ_3 by characterizing its CGF, $\mathcal{S}_3$. To this end, we eliminate the binary substrate-enzyme complex C and reduce the MM reaction triplet to a single process, whose dynamics are quasi-stationary over times much longer than a single reaction event. The details are in *Materials and Methods*, Eq. **21**, and the obtained expression is valid for times δt, $\tau_E \ll \delta t \ll \tau_M$, so that many enzyme turnovers happen, but the effect on the abundance of S_M is still relatively small.

This completes Step 1 of the coarse-graining in which each reaction, or a small complex of reactions, is subsumed by a quasi-stationary CGF $\mathcal{S}_\mu$ of the distribution of the number of its events. Importantly, in this step, we remove the only species that exists, at most, in a single copy, thus simplifying analysis.

To illustrate the simplification, using Eq. **21**, we write the first few cumulants of the number of MM products:

$$c_{3,1} = \frac{k_1 k_2 S_M}{K}\delta t, \quad K = k_1 S_M + k_2 + k_{-1}, \qquad [5]$$

$$c_{3,2} = c_{3,1}F, \quad F = 1 - 2Q/K, \quad Q = c_{3,1}/\delta t, \qquad [6]$$

$$c_{3,3} = c_{3,1}[1 - 6Q(K - 2Q)/K^2], \qquad [7]$$

$$c_{3,4} = c_{3,1}[1 - 2Q(7K^2 - 36KQ + 60Q^2)/K^3]. \qquad [8]$$

The coefficient F is called the Fano factor (see below). To the extent that $F \neq 1$, this complex reaction is non-Poisson [supporting information (SI) Fig. S1].

Knowing $c_{3,a}$ allows for a numerical simulation procedure

$$\delta Q_3(t) = \eta_3(t, \delta t), \qquad [9]$$

$$S_M(t + \delta t) = S_M(t) - \delta Q_3(t) + J_M(t)\delta t, \qquad [10]$$

$$P(t + \delta t) = P(t) + \delta Q_3(t), \qquad [11]$$

where $\eta_3(t)$ is a random variable with the cumulants given by Eqs. **5**–**8**, and $J_M(t)$ represents currents exogenous to the MM reaction, such as changes in S_M due to membrane binding/unbinding. Here we treat the reaction in a quasistationary, mesoscopic manner by drawing a (random) number of reaction events within δt directly, assuming that all parameters defining the reaction are constants over this time. The price for the coarse-graining is that this reaction is non-Poisson and is characterized by a prescribed sequence of cumulants.

In principle, generation of such random variables is an ill-posed task because, once we allow for a nonzero third cumulant, the remaining higher-order cumulants cannot be all zero, and the random variable depends on assumptions made about them. Fortunately, in our case, the situation is simpler because all $c_{3,k} \propto \delta t$. Thus, higher cumulants have a progressively smaller effect, $\propto (\delta t)^{1/k}$, on a number drawn from the distribution—our random variable is near-Gaussian. Then the Gram–Charlier (GC) series expansion (18) aided either by the importance or the rejection sampling (19, 20) reduces the simulation scheme, Eqs. **9**–**11**, to a simple Langevin simulation with a small penalty, as in *Materials and Methods*; see Fig. 5 for illustration of the precision of these tools.

Step 2: Coarse-Graining Membrane Reactions. In Step 2 of the coarse-graining, we start with the CGFs S_μ, $\mu = 1, 2, 3$, of the slowly varying reactions. Using the SPI technique, we then express the CGF of δQ, the number of the entire coarse-grained reactions $S_B \to P$ in Fig. 2*C* over time T, in terms of the component CGFs (this is where working with the CGFs instead of the distributions is the most advantageous). We then simplify the expression to account for the time scale separation between τ_B and τ_M, see *Materials and Methods*, Eq. **31**. This formally completes the coarse-graining. That is, we find the CGF of the $S_B \to P$ particle flux for times $T \lesssim \tau_B$, much longer than τ_E and τ_M.

The full expression for CGF is cumbersome and nonilluminating. Fortunately, we only look for the first few cumulants of $P(\delta Q|T)$, and these are obtained by differentiating the CGF as in Eq. **4**. The expressions for the first 3 cumulants, c_1, c_2, and c_3 are in *SI Text*. Then, similar to the MM reaction, we can simulate the whole 5-reaction network in 1 Langevin-like step:

$$\delta Q(t) = \eta(t, T), \qquad [12]$$

$$S_B(t + T) = S_B(t) - \delta Q(t) + J_B(t)T, \qquad [13]$$

$$P(t + T) = P(t) + \delta Q(t), \qquad [14]$$

where η is a random variable with the cumulants as in Eqs. **S1**–**S4** in *SI Text*, and $J_B(t)$ is an external current, such as production or decay of the bulk substrate in other cellular processes.

Fano Factor in Counting Experiments. When experimentally measuring the number of created products, one can estimate the Fano factor, $F = c_2/c_1$. The factor is zero for deterministic systems and one for a Poisson process, providing a quantification of the importance of stochastic effects.

Traditionally, to compare experimental data about F to a mathematical model, one would simulate the model using the Gillespie algorithm (9), which takes a long time to converge. In contrast, our coarse-grained quasistationary approach yields an analytic expression for the Fano factor of the $S_B \to P$ transformation, see Eq. **S3** in *SI Text*. Similar analytical shortcuts should be possible for other kinetic schemes. In Fig. 3, we compare the analytical expression to stochastic simulations for the full set of reactions in Fig. 2*A*, seeing an excellent agreement.

Note that $F \neq 1$, indicating a non-Poissonian nature. The backwards decay of C adds extra randomization, thus larger values of k_{-1} increase F. At the other extreme, when $k_{-1} = 0$, the Fano factor may be as small as 1/2, so that the entire $S_B \to P$ chain is equal to a sequence of 2 Poisson events with similar rates. Finally,

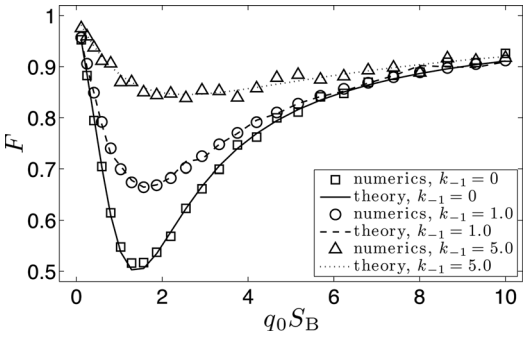

Fig. 3. Comparison of the analytically calculated Fano factor for the $S_B \to P$ reaction to Monte Carlo simulations with the Gillespie algorithm (9). We use $q = 0.02$, $k_1 = 0.05$, $k_2 = 1$, and $T = 10,000$. Each numerical data point averages 10,000 simulation runs.

Table 1. Comparison of cumulants of the product flux for the full system calculated using the Gillespie simulations, the coarse-grained simulations at Step 1 and Step 2, and the analytical predictions

Cumulant	Gillespie	CG (step 1)	CG (step 2)	Analytics
c_1	418.7 (1)	420.0 (1)	418.9 (1)	418.9
c_2/c_1	0.771 (1)	0.764 (2)	0.768 (1)	0.767
c_3/c_1	0.50 (3)	0.46 (8)	0.48 (3)	0.472
Time	1 h 14 min	1 min 17 s	1 s	N/A

Numbers in parentheses are the estimated errors in the last significant digits.

when $q = 0$, i.e., the substrates are removed from the membrane only via $S_M \to P, F = 1$. This is because then the only stochasticity in the problem is from Poisson membrane binding, and all bound substrates will eventually get converted to P.

Computational Complexity of Coarse-Grained Simulations. We expect our approach to be particularly useful for simulations in systems biology because it is much faster than the traditional Gillespie algorithm (9). Indeed, for our model, the computational complexity of a single Gillespie simulation run is $O(MT/\tau_E)$, where $M = 5$ is the number of reactions in the system, and T is the duration of the simulated dynamics. In contrast, the complexity of a coarse-grained run is $O[M^0(T/\tau_E)^0]$ because we have eliminated the internal species and simulate the dynamics in steps of $\approx T$ instead of $\approx \tau_E$. However, the Gillespie algorithm is (statistically) exact, while ours relies on quasi-stationarity. Thus the coarse-grained simulation must be benchmarked against the Gillespie algorithm to gauge its practical utility in reducing the simulation time while retaining a high accuracy.

Similarly, it would have been useful to benchmark against the Slow Scale Stochastic Simulation Algorithm (ssSSA) (10), an adiabatic simulation method, on which most others are based. However, using $k_1 = 0.02, k_{-1} = 2, k_2 = 1, q = 0.01$, and $q_0 S_B = 1.5$, we get $S_M \approx 100$, and all of the reaction rates are ≈ 1. Thus none of the 5 elementary reactions is fast and ssSSA and its derivatives, which are based on the fast reaction approach, are not applicable. Therefore, we benchmark against the Gillespie algorithm only.

All simulations were performed by using Fortran 90, on a single CPU (1.83 GHz, Windows 2000). In *SI Text* we provide the benchmark results for the single MM enzyme (reaction 3), where the coarse-graining achieves factor of 40 speedup. Here, we focus on the full model system viewed at different coarseness.

Coarse-grained, Step 1: Total time of the evolution is $T = 1,000$, and $S_M(t = 0) = 120$. Then the relaxation time of a typical fluctuation of S_M is $\tau_M \approx 1/[q + (\partial k_{MM}/\partial S_M)] \approx 80$, where k_{MM} is the rate of the MM reaction for a given S_M, and this sets the scale $1 \approx \tau_E \ll \delta t = 20 \ll \tau_M \approx 80$. We simulate the reactions that survive Step 1 (membrane binding/unbinding and MM transformation) by approximating their distributions with the GC series with 3 known cumulants, and we perform 10^6 simulation runs, which is sufficient for convergence of the third cumulant of the $S_B \to P$ aggregate reaction. As shown in Table 1, the coarse-grained approach speeds simulations 60-fold relative to the Gillespie one with little apparent accuracy loss.

Coarse-grained, Step 2: We do similar benchmarking for the system represented as a single coarse-grained reaction $S_B \to P$. Here, we use a single time step equal to T, $\tau_M \ll T \ll \tau_B$. Table 1 shows that simulating all 5 reactions in a single step results in a dramatic 4,000-fold speedup. For all cumulants, coarse grained simulations and analytic results differ from exact Gillespie values by, at most, a percent, which hardly matters. Yet the reduction of the simulation time by the factor of $10^3 \ldots 10^4$ is certainly tangible.

It is important to understand where the 60- and the 4,000-fold improvements are from. Both relate to (*i*) simulating fewer than five original reactions, (*ii*) not spending any time on simulating backwards processes, and (*iii*) making time steps of 20 and 1,000 respectively, compared with $\tau_E \approx 1$. Clearly, the latter (the adiabaticity) contributes the most, at least in our example. In their turn, the 2 levels of adiabaticity exist because there is 1 enzyme, 100 membrane substrates, and many thousands of bulk substrates, so that the typical fractional change of the membrane and the bulk substrates is very small over the respectively chosen time steps. For other kinetic schemes, improvements should also scale with the ratio of time scales or abundances.

Generalization to a Network of Reactions. As discussed in detail in the original literature (14), in the SPI formalism, a network of M reactions with N chemical species (Fig. 4) is generally described by $2MN$ ordinary differential equations specifying the saddle point solution of the corresponding path integral. *Materials and Methods* provides a particular example, and we refer the readers to the original literature for generalizations. Here, we build on the ref. 14 and focus on developing a relatively simple, yet general coarse-graining procedure for more complex reaction networks.

At intermediate time scales, δt, many fast species connecting slow ones can be considered statistically independent. Therefore, in the SPI, every separate chain of such species adds to the effective Hamiltonian. Namely, we enumerate slow chemical species by $\mu, \nu, \ldots$. Fast chains connecting them can be marked by pairs of indexes, e.g., $\mu\nu$ (Fig. 4). An entire such chain will contribute a single effective Hamiltonian term, $H_{\mu\nu}(\{N\}, \{\chi\}, \{\chi_C\})$, to the full CGF of the slow fluxes, where $\{N\}$, $\{\chi\}$, and $\{\chi_C\}$ are the slow species abundances and the conjugate counting variables. If necessary, the geometric correction to the CGF, $\mathcal{S}^{\mu\nu}_{\text{geom}}(\{N\}, \{\chi\}, \{\chi_C\})$, can be written out as well (15). Overall,

$$\mathcal{S}(\{\chi_C\}, T) = \sum_{\mu<\nu} \mathcal{S}^{\mu\nu}_{\text{geom}}(\{N(t)\}, \{\chi(t)\}, \{\chi_C\}, T) + \int_0^T dt \left[\sum_\mu i\chi_\mu \dot{N}_\mu + \sum_{\mu<\nu} H_{\mu\nu}(\{N(t)\}, \{\chi(t)\}, \{\chi_C\}) \right]. \quad [15]$$

Again, as in *Materials and Methods*, summation over all CGFs in Eq. **15** is a result of the convolution of conditional distributions of the slow fluxes.

This provides for the following coarse-graining procedure. First, one finds a time scale δt, small enough for the slow species to be considered stationary and yet fast enough for the fast ones to equilibrate. If the fast species consist only of a few degrees of freedom, like in the case of a single enzyme, one derives the CGF of the transformations mediated by these species similar to *Materials and Methods*. If instead the fast species are mesoscopic, one uses the SPI technique to derive the CGF by analogy with Step 2.

At the next step, the CGFs of the fast species are incorporated into the SPI over the abundances of the slow ones. For this, one writes down the the full effective Hamiltonian, Eq. **15**, assumes adiabatic evolution, and solves the ensuing saddle point equations. The extremum of the effective Hamiltonian determines the

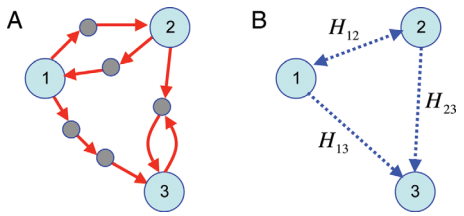

Fig. 4. Schematic coarse-graining of a network of reactions. (*A*) The network has $M = 10$ reactions (red arrows) and $N = 8$ species, of which 3 are slow (large circles), and 5 are fast (small circles). (*B*) Dynamics of each fast node can be integrated out, leaving effective, pairwise fluxes among the slow nodes (blue arrows), labeled by the corresponding effective Hamiltonians $H_{\mu\nu}$. Note that, for reversible pathways, the flux may be positive or negative (2-sided arrow), and it is nonnegative otherwise (1-sided arrows).

CGF of the coarse-grained process. For hierarchies of time scales, this reduction procedure is then repeated. A limitation of the procedure is in requiring the knowledge of typical species concentrations and the associated time scales. One can identify those by a few preliminary Gillespie simulations.

Discussion

Rigorous mathematical techniques are finding applications in biology. Here we present one such example, where adiabatic approach, paired with the SPI formalism of statistical physics, allows one to coarse-grain stochastic biochemical networks. We eliminate fast variables, reducing a network with a separation of time scales to a handful of slow species coupled by complex interactions with properties that account for the decimated nodes. The simplified system is smaller, nonstiff, and easier to analyze, resulting in orders-of-magnitude improvement in the speed of its simulations. This has a potential for a wide impact in systems biology, at least for systems with diverse time scales.

Fortunately, such systems are more common than one would expect. Consider, for example, the system briefly mentioned in the Introduction: A molecule must be modified on n sites in an arbitrary order to get activated. The kinetic diagram for this system is an n-dimensional hypercube, and the number of states of the molecule with m modified sites is $\binom{n}{m}$. Therefore, if the total number of molecules is N, then a typical state with m modifications has $N_m \approx N/\binom{n}{m}$ molecules. This number may be small, ensuring the need for a stochastic analysis. More importantly, it is quite different from either N_{m-1} or N_{m+1}, e.g., $N_m/N_{m+1} = (m+1)/(n-m)$, and the different abundances result in different time scales.

The coarse-graining simplifies interpretation of biological systems. For example, the Fano factor of the $S_B \to P$ reaction, Fig. 3, may approach unity, suggesting a simple, yet rigorous, replacement of the entire reaction by a simple Poisson step. Then the list of relevant parameters is smaller than suggested ab initio, improving interpretability and decreasing the effective number of biochemical features that must be measured experimentally.

Although orders-of-magnitude improvement in simulation speed is impressive, we are still far from coarse-graining cellular-scale networks. However, the following properties of our approach suggest that we may be on the right track:

- We eliminate fast variables not just treat them differently.
- We can operate with arbitrarily long series of cumulants of the number of reaction events, which allows for hierarchical applications and for keeping track of even rare fluctuations.
- Unlike most other adiabatic methods, ours does not depend on existence of fast reactions (for the MM system, this results in the ability to coarse-grain for any reaction rates rather than only for the linear and the saturated regimes).
- Standard adiabatic approximations, well developed in classical and quantum physics, can be applied easily in the SPI context.
- Unlike some other coarse-graining techniques, the SPI approach can deal with copy numbers of order unity.
- With the SPI, large networks of stochastic reactions can be reduced to a set of deterministic differential equations.
- Finally, the SPI is rigorous, mathematically justifiable, and allows for controlled approximations.

In the forthcoming publications, we expect to show how these advantageous properties of the adiabatic SPI technique allow to coarse-grain standard biochemical network motifs.

Materials and Methods

Coarse-Graining the MM Reaction. Consider the $S_M \to P$ reaction, described mathematically as in Eq. **1**:

$$S_M + E \underset{k_{-1}}{\overset{k_1 S_M}{\rightleftharpoons}} C \xrightarrow{k_2} E + P. \qquad [16]$$

The probabilities of transitions between bound, P_b, and unbound, $P_u = 1 - P_b$, states of the enzyme are given by a 2-state Markov process

$$\frac{d}{dt}\begin{bmatrix} P_u \\ P_b \end{bmatrix} = -\begin{bmatrix} k_1 S_M & -k_{-1} - k_2 \\ -k_1 S_M & k_{-1} + k_2 \end{bmatrix}\begin{bmatrix} P_u \\ P_b \end{bmatrix}. \qquad [17]$$

Using Eq. **17** and the definition of $\mathcal{Z}_u$, Eq. **3**, one can show that $\mathcal{Z}_3(\chi, \delta t)$ satisfies a Schrödinger-like equation with a χ-dependent Hamiltonian, leading to a formal solution (5, 15, 21)

$$\mathcal{Z}_3(\chi, \delta t) = 1^+ \left(e^{-\hat{H}_{MM}(\chi,t)\delta t}\right) p(t_0), \qquad [18]$$

where $1^+ = (1, 1)$, $p(t_0)$ is the probability of the initial states, and

$$\hat{H}_{MM}(\chi) = \begin{bmatrix} k_1 N_s & -k_{-1} - k_2 e^{i\chi} \\ -k_1 N_s & k_{-1} + k_2 \end{bmatrix}. \qquad [19]$$

Similar Hamiltonians can be derived for a wide class of kinetic schemes (5, 15, 21, 22), allowing for a straightforward extension of our methods.

The solution, Eq. **18**, can be simplified if the MM reaction is considered in a quasi-steady-state approximation, that is P_u is equilibrated at a current value of the other parameters. This means that the time scale of interest, $\delta t \approx \tau_M$, is much larger than a characteristic time of a single enzyme turnover, τ_E, so we can consider $\delta t \to \infty$ in Eq. **18**. Then only the eigenvalue $\lambda_0(\chi)$ of $\hat{H}_{MM}(\chi)$ with the smallest real part is relevant, and $\mathcal{Z}_3(\chi, \delta t) = e^{-\lambda_0(\chi)\delta t}$.

It is possible to incorporate a slow time dependence of the parameters into this answer. By analogy with the quantum mechanical Berry phase (6, 15), the lowest order nonadiabatic correction can be expressed as a geometric phase

$$\mathcal{Z}_3(\chi) = e^{\mathcal{S}_3(\chi)} = e^{\int_c A\cdot dk - \int dt \lambda_0(\chi,t)}, \qquad [20]$$

where $A = \langle u_0(\chi)|\partial_k u_0(\chi)\rangle$, k is the vector in the parameter space that draws a contour c during the parameter evolution, and $\langle u_0(\chi)|$ and $|u_0(\chi)\rangle$ are the left and the right eigenvectors of $\hat{H}_{MM}(\chi, t)$ corresponding to the instantaneous eigenvalue $\lambda_0(\chi, t)$. The first term is the geometric phase, which is responsible for various ratchet-like fluxes (6).

The geometric phase gives rise to magnetic field-like corrections to the evolution of the slow variables. However, these corrections are proportional to (small) time derivatives of these variables, and they often can be neglected. In our model, the geometric effects are negligible when $\tau_E/\tau_M \approx 1/S_M \ll 1$, and we deemphasize them.

Reading the value of $\lambda_0(\chi)$ from ref. 6, we write the CGF of $P(\delta Q_3|\delta t)$, $\tau_E \ll \delta t \lesssim \tau_M$ in the adiabatic limit:

$$\mathcal{S}_3(\chi, \delta t) = \mathcal{S}_{geom}(\chi, S_M, \dot{S}_M) + \frac{\delta t}{2}\Big[-(k_{-1} + k_2 + S_M k_1) + \sqrt{(k_{-1} + k_2 + S_M k_1)^2 + 4S_M k_1 k_2 (e^{i\chi} - 1)}\Big]. \qquad [21]$$

Simulations with Near-Gaussian Distributions. A probability distribution $P(\delta Q)$ with known cumulants $c_1, c_2, c_3, \ldots$, can be approximated as a limited GC expansion (18)

$$P(\delta Q) \approx \Psi(\delta Q, c_1, c_2)\left[1 + \frac{c_3(y^3 - y)}{6c_2^{3/2}} + \frac{c_4(y^4 - 6y^2 + 3)}{24c_2^2} + \frac{c_3^2(y^6 - 15y^4 + 45y^2 - 15)}{72c_2^3} + \cdots\right], \qquad [22]$$

where $y = (\delta Q - c_1)/\sqrt{c_2}$ and $\Psi(\delta Q, c_1, c_2)$ is the Gaussian density with the mean c_1 and the variance c_2. The leading term in Eq. **22** is a standard Gaussian approximation, and the subsequent terms account for skewness, kurtosis, etc. If all cumulants scale similarly (the near-Gaussian case), then the terms in the series become progressively smaller, ensuring rapid convergence.

Generation of random samples from the non-Gaussian GC series is still a difficult task. However, if, instead of the random numbers per se, the goal is to calculate the expectation of some function $f(\delta Q)$ over the distribution P, $\langle f(\delta Q)\rangle_P$, then the importance sampling (19) can be used. Specifically, we generate a Gaussian random number δQ from $\Psi(\delta Q, c_1, c_2)$ and define its

importance factor according to its relative probability in the normal distribution and the considered GC series $\eta = P(\delta Q)/\Psi(\delta Q, c_1, c_2)$. After generating N such random numbers δQ_ν, $\nu = 1, \ldots, N$, we get

$$\langle f(\delta Q)\rangle_P = \frac{\sum_{\nu=1}^{N} \eta_\nu f(\delta Q_\nu)}{\sum_{\nu=1}^{N} \eta_\nu}. \quad [23]$$

If a current random number draw represents just 1 reaction in a larger reaction network, then the overall importance factor of a Monte Carlo realization is a product of the factors for each of the random numbers drawn within it.

This reduces the complexity of simulations to that of a simple Gaussian, Langevin process with a small burden of (*i*) evaluating an algebraic expression for the GC series, and (*ii*) keeping track of the importance factor. Yet this small computational investment allows one to account for an arbitrary number of cumulants of the involved variables. To illustrate this, in Fig. S2, we compare the GC, importance-sampling simulations of the MM reaction flux to the exact results in *Results: Step 1*. The third and the fourth cumulants make the two almost indistinguishable.

Here, we sound a note of caution: the GC series produces approximations that are not necessarily positive and hence are not, strictly speaking, probability distributions. However, the leading Gaussian term decreases so fast that this may not matter in practice. In fact, in our simulations, we simply rejected random numbers that had negative importance corrections. However, this is inadequate for lengthy simulations, where the probability that one of random numbers in a long chain of events falls into such badly approximated region approaches 1. Then other means of generating random numbers, such as the well-known acceptance–rejection method (20) should be used. Because the distributions of interest are near-Gaussian, a Gaussian with a slightly larger variance is an envelope function for the GC approximation to the true distribution. Then the average random number acceptance probability scales as the ratio of the true and the envelope standard deviations, and it is almost 1. Then the rejection approach requires just a bit more than 1 normal and 1 uniform random numbers to generate a sample from the GC series. Importantly, in this case, the negativity of the series is not a problem because it leads to a rejection of a single, highly improbable sample, rather than of an entire sampling trajectory.

Coarse-Graining All Membrane Reactions. To perform the coarse-graining that connects Fig. 2 *B* and *C*, we look for the MGF of the total number of products Q_P produced over time $T \approx \tau_B$:

$$\mathcal{Z}(\chi_C) = e^{\mathcal{S}(\chi_C)} = \sum_{Q_P=0}^{\infty} P(Q_P|T) e^{iQ_P\chi_C}. \quad [24]$$

For this, we discretize the time into intervals t_k of duration δt, and introduce random variables $\delta Q_\mu(t_k)$ ($\mu = 1, 2, 3$), which denote the numbers of each of the 3 different reactions in Fig. 2*B* (membrane binding, unbinding, and MM conversion) during each time interval. The probability distributions of $\delta Q_\mu(t_k)$ are given by inverse Fourier transforms of the corresponding MGFs:

$$P(\delta Q_\mu(t_k)) = \frac{1}{2\pi}\int d\chi_\mu(t_k) e^{-i\chi_\mu(t_k)\delta Q_\mu(t_k) + H_\mu(\chi_\mu(t_k), S_B(t_k))\delta t}, \quad [25]$$

where the CGF are $\mathcal{S}_\mu(\chi, S_B) = H_\mu(\chi, S_B)\delta t$. Following refs. 14 and 15, and recalling that $Q_P = \sum_k \delta Q_3(t_k)$, we write the MGF of the total number of products created during time interval $(0, T)$ as the path integral over all possible trajectories of $\delta Q_\mu(t_k)$ and $S_M(t_k)$:

$$\begin{aligned} e^{\mathcal{S}} = \langle e^{i\chi_C Q_P}\rangle = \int DS_M(t_k) \prod_{k,\mu}\int D\delta Q_\mu(t_k) \\ \times P[\delta Q_\mu(t_k)] e^{i\chi_C \sum_{t_k} \delta Q_3(t_k)} \\ \times \delta[S_M(t_{k+1}) - S_M(t_k) - \delta Q_1(t_k) + \delta Q_2(t_k) + \delta Q_3(t_k)]. \end{aligned} \quad [26]$$

The δ-function in Eq. **26** expresses the conservation law for the slowly changing number of substrate molecules S_M. We rewrite it as

$$\delta(\cdots) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} d\chi_M(t_k) \exp\{i\chi_M(t_k)\cdots\}, \quad [27]$$

and we substitute the expression together with Eq. **25** into Eq. **26**. Then the integration over $\delta Q_\mu(t)$ produces new δ-functions over χ_μ, which, in turn, are removed by integration over $\chi_\mu(t_k)$. This leads to an expression for the MGF:

$$e^{\mathcal{S}} = \int DS_M D\chi_M e^{\int_0^T dt[i\chi_M \dot{S}_M + H(S_M, \chi_M, \chi_C)]}, \quad [28]$$

$$\begin{aligned} H &= H_1(-\chi_M, S_M, t) + H_2(\chi_M, S_M, t) + H_3(\chi_M + \chi_C, S_M, t) \\ &= q_0 S_B e_{-\chi_M} + S_M q e_{\chi_M} + \frac{1}{2}\Big[-(k_{-1} + k_2 + S_M k_1) \\ &\quad + \sqrt{(k_{-1} + k_2 + S_M k_1)^2 + 4S_M k_1 k_2 (e^{i\chi_M + \chi_C} - 1)}\Big]. \end{aligned} \quad [29]$$

where $e_{\pm\chi_M} = e^{\pm i\chi_M} - 1$. The original SPI work (14) assumed all component reactions to be Poisson. However, here H_3 is the CGF of the entire complex, non-Poisson MM reaction, which we read as the coefficient in front of δt in Eq. **21**. This ability to include subsystems with small number of degrees of freedom, such as the MM enzyme, opens doors to application of the method to a wide variety of coarse-graining problems.

The meaning of Eq. **29** is simple. To evaluate statistics of slow fluxes, one needs to convolve their distributions conditional on the slow species abundances with the distributions of these abundances, which themselves depend on the fluxes in the previous moments of time. As always, complicated convolutions result in simple summation/integration for CGFs.

Because $S_M \gg 1$, this path integral is dominated by the classical solution of the equations of motion, which, near the steady state, are

$$\dot{S}_M = \dot{\chi}_M = 0, \quad \frac{\partial H}{\partial \chi_M} = \frac{\partial H}{\partial S_M} = 0. \quad [30]$$

Let $\chi_{cl}(\chi_C)$ and $S_{M,cl}(\chi_C)$ solve Eq. **30**. Then the cumulants generating function in the quasi-steady-state approximation is

$$\mathcal{S}(\chi_C, T) = TH(S_{M,cl}(\chi_C), \chi_{cl}(\chi_C), \chi_C) \quad [31]$$

This Born–Oppenheimer-like procedure completes the coarse-graining by deriving the CGF for the number of product creations over long times.

ACKNOWLEDGMENTS. We thank F. Alexander, G. Bel, W. Hlavacek, B. Munsky, and M. Wall for useful discussions and the anonymous referees for their insightful comments. This work was supported in part by the U.S. Department of Energy under Contract No. DE-AC52-06NA25396.

1. Hlavacek W, et al. (2006) Rules for modeling signal-transduction systems. *Sci STKE* 344.
2. Michaelis L, Menten M (1913) The kinetics of invertase activity. *Biochem Z* (in German) 49:333.
3. Rao C, Arkin A (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *J Chem Phys* 118: 4999.
4. Givon D, Kupferman R, Stuart A (2004) Extracting macroscopic dynamics: Model problems and algorithms. *Nonlinearity* 17:55.
5. Gopich I, Szabo A (2006) Theory of the statistics of kinetic transitions with application to single-molecule enzyme catalysis. *J Chem Phys* 124:154712.
6. Sinitsyn N, Nemenman I (2007) Berry phase and pump effect in stochastic chemical kinetics. *EPL* 77:58001.
7. Lan Y, Elston T, Papoian G (2008) Elimination of fast variables in chemical Langevin equations. *J Chem Phys* 129:214115.
8. Zinn-Justin J (2002) *Quantum Field Theory and Critical Phenomena* (Oxford Univ Press, New York).
9. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340.
10. Cao Y, Gillespie D, Petzold L (2005) The slow-scale stochastic simulation algorithm. *J Chem Phys* 122:014116.
11. Mastny E, Haseltine E, Rawlings J (2007) Two classes of quasi-steady-state model reductions for stochastic kinetics. *J Chem Phys* 127:094106.
12. E W, Liu D, Vanden-Eijnden E (2007) Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. *J Comput Phys* 221:158–180.
13. Rathinam M, El-Samad H (2007) Reversible-equivalent-monomolecular tau: A leaping method for "small number and stiff" stochastic chemical systems. *J Comput Phys* 224:897–923.
14. Jordan A, Sukhorukov E, Pilgram S (2004) Fluctuation statistics in networks: A stochastic path integral approach. *J Math Phys* 45:4386.
15. Sinitsyn N, Nemenman I (2007) Universal geometric theory of mesoscopic stochastic pumps and reversible ratchets. *Phys Rev Lett* 99:220408.
16. Gillespie D (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* 115:1716–1733.
17. English B, et al. (2006) Ever-fluctuating single enzyme molecules: Michaelis–Menten equation revisited. *Nat Chem Biol* 2:87.
18. Blinnikov S, Moessner R (1998) Expansions for nearly Gaussian distributions. *Astron Astrophys Suppl Ser* 130:193.
19. Srinivasan R (2002) *Importance Sampling—Applications in Communications and Detection* (Springer, Berlin).
20. von Neumann J (1951) Various techniques used in connection with random digits. Monte Carlo methods. *Natl Bureau Standards* 12:36.
21. Bagrets D, Nazarov Y (2003) Full counting statistics of charge transfer in Coulomb blockade systems. *Phys Rev B* 67:085316.
22. Sukhorukov E, Jordan A (2007) Stochastic dynamics of a Josephson junction threshold detector. *Phys Rev Lett* 98:136803.