Generalized ensemble and tempering simulations: A unified view

Walter Nadler[1,*] and Ulrich H. E. Hansmann[1,2,†]

[1]*Department of Physics, Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan 49931-1295, USA*
[2]*John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany*

From the underlying master equations we derive one-dimensional stochastic processes that describe generalized ensemble simulations as well as tempering (simulated and parallel) simulations. The representations obtained are either in the form of a one-dimensional Fokker-Planck equation or a hopping process on a one-dimensional chain. In particular, we discuss the conditions under which these representations are valid approximate Markovian descriptions of the random walk in order parameter or control parameter space. They allow a unified discussion of the stationary distribution on, as well as of the stationary flow across, each space. We demonstrate that optimizing the flow is equivalent to minimizing the first passage time for crossing the space and discuss the consequences of our results for optimizing simulations. Finally, we point out the limitations of these representations under conditions of broken ergodicity.

I. INTRODUCTION

The effective simulation of complex thermal systems like proteins and glasses is a constant challenge in contemporary computational physics. Markov chain Monte Carlo simulation techniques for these systems have undergone remarkable advances in the last decades. Two main classes that have evolved are generalized ensemble and parallel tempering methods.

In the generalized ensemble (GE) approach [1] the goal is to sample the state space of a physical system so that particularly rare but important states—e.g., low energy or barrier states—are encountered frequently. A variety of weight functions have been tested, as well as methods for iteratively improving these functions [2].

A persistent problem of such simulations is that relaxation is slow due to barriers and bottlenecks, and for a long time it was not clear whether and how they can be controlled by using particular weight functions on the usual order parameter spaces. Parallel tempering (PT)—sometimes also called replica exchange method—promised a way out of this dilemma [3–5]. Here simulations are performed in parallel at different values of a control parameter, most often the temperature. At certain times the current conformations of replicas at neighboring control parameter values are exchanged according to a generalized Metropolis rule. Thereby an individual replica can perform an additional random walk in control parameter space and, due to shorter relaxation times in some control parameter regime, explore state space more evenly.

However, the problem of slow relaxation arises also this time in the form of a slow and possibly uneven random walk through control parameter space. At least part of this problem is related to finding an efficient discretization of control parameter space.

Increasing the flow through order parameter space in GE sampling as well as through control parameter space in PT

was always an incentive. Usually it was discussed only informally, and only recently Trebst and co-workers [6–9] have made an attempt to look at that problem systematically. Instead of concentrating on the stationary distributions that arise from the sampling, they concentrated on the stationary flow across order parameter and control parameter space. In order to optimize the flow they derived weight functions and control parameter discretization schemes.

In this contribution we want to give that approach a more fundamental basis. We will first concentrate on the underlying master equations describing GE and PT simulations. From them we will derive in a systematic way the one-dimensional stochastic equations that form the basis for a flow analysis in order parameter and control parameter space. These equations will also allow us to investigate connections between flow analysis and another concept describing the dynamics of stochastic processes, the first passage time (FPT). The one-dimensional representations are valid approximations for the underlying simulations only under certain conditions. If they are violated, optimization schemes may still fail. For parallel tempering we will find a criterion from which the validity can be determined.

We will focus in the next section on generalized ensemble sampling, while parallel tempering is the focus of Sec. III. We will close with a discussion of the effects of broken ergodicity on our results and an outlook.

II. GENERALIZED ENSEMBLE SAMPLING

Markov chain Monte Carlo simulations utilize a certain move set in combination with an acceptance probability, most often of Metropolis form [10], to impose stochastic dynamics on a physical system. A move from state s to s' is accepted with the probability

$$p_M(s \rightarrow s') = \min(1, w(s')/w(s)). \tag{1}$$

The original choice for the weight function $w(s)$ is the thermal or Boltzmann weight

---

*Electronic address: wnadler@mtu.edu

†Electronic address: hansmann@mtu.edu

$$w(s) \propto \exp[-\beta E(s)], \tag{2}$$

resulting in canonical sampling at inverse temperature $\beta = 1/k_B T$. However, other choices are possible as well, depending on which particular aspect of state space should be emphasized. We will assume in the following that the weight function is based only on the energy $E(s)$ of the state s, and we will use $w(s) = w[E(s)] = w(E)$ interchangeably.

In order to get a deeper understanding of simulations based on Eq. (1) we should look at their description via a master equation in state space. $P(s,t)$ is the probability to be in state s at time t, and its evolution in discrete computer time is given by

$$P(s, t+1) = \sum_{s' \neq s} P(s', t) W_s(s' \to s) + P(s, t)$$
$$\times \left[1 - \sum_{s' \neq s} W_s(s \to s') \right], \tag{3}$$

where the sums are over all possible states. The transition probabilities $W(s \to s')$ in this equation are

$$W_s(s \to s') = \frac{\psi(s \to s')}{\sum_{s''} \psi(s \to s'')} p_M(s \to s'), \tag{4}$$

where $\psi(s \to s')$ is the characteristic function of the move set with $\psi(s \to s') = 1$ if the move from s to s' is allowed and zero otherwise. Provided the move set is balanced and ergodic, the *stationary distribution* reached by this Markov chain is $P_0(s) \propto w(s)$ because of detailed balance. Note however that in addition to $\psi(s \to s') = \psi(s' \to s)$, for (s, s') with $\psi(s \to s') = 1$ the property $\sum_{s''} \psi(s \to s'') = \sum_{s''} \psi(s' \to s'')$ must be fulfilled. Particularly the latter property—i.e., every state must connect to the same number of other states—is sometimes overlooked.

Equation (3) is an exact description of the simulation process and it is also the basis for more thorough investigations in the mathematics of Metropolis simulations [11–14]. However, from a physicist's point of view a *reduced* description in terms of *slow order parameters* is of more interest. Prominent among the order parameters chosen is the energy itself. Using adiabatic elimination of fast degrees of freedom (see the Appendix), an approximate master equation in energy space can be formulated:

$$P(E, t+1) = \sum_{E' \neq E} P(E', t) W_E(E' \to E) + P(E, t)$$
$$\times \left[1 - \sum_{E' \neq E} W_E(E \to E') \right]. \tag{5}$$

The effective transition probabilities $W_E(E \to E')$ can be derived from Eqs. (3) and (4), and are given in that appendix, too. Note however that, in contrast to Eq. (3), Eq. (5) is an approximation, valid only if all other degrees of freedom relax much faster than the energy. Nevertheless, even if relaxation orthogonal to the energy is slow, Eq. (5) can still be viewed as a *Markovian* approximation to the fully non-Markovian process. Due to the degeneracy of states with

energy, the stationary distribution of Eq. (5) is now

$$P_0(E) \propto n(E) w(E). \tag{6}$$

Here, $n(E)$ is the density of states and we have assumed that the weight function for the Metropolis algorithm is based on energies only, as mentioned above in the discussion of Eq. (1).

Coarse graining time and energy leads to a form of the master equation that is continuous in both variables:

$$\frac{\partial}{\partial t} P(E, t) = \int_{E'} P(E', t) R_E(E' \to E) - P(E, t) \int_{E'} R_E(E \to E'), \tag{7}$$

where the transition probabilities have been replaced by the rates $R_E(E \to E')$. Note that for various systems state space and energy E could have been continuous from the start; i.e., the sums in Eqs. (3) and (5) could have already been integrals. The continuous form of the master equation in energy space is now the starting point for our final approximation. If the transition rates $R_E(E \to E')$ are strongly peaked around $E' \approx E$, a second-order partial differential equation can be derived from Eq. (7) by various techniques: e.g., Kramers-Moyal expansion [15–17]. This Fokker-Planck equation [16] for $P(E, t)$ is given by

$$\frac{\partial}{\partial t} P(E, t) = \frac{\partial}{\partial E} D(E) \left[\frac{\partial}{\partial E} - F(E) \right] P(E, t), \tag{8}$$

and we have written it already in a form that separates static and dynamic properties. $D(E)$ is the energy-dependent diffusion coefficient that describes the local mobility and $F(E)$ is the drift term which in one dimension can always be derived from a potential, $F(E) = -(d/dE)U(E)$. In particular the stationary distribution of Eq. (8) is fully determined by this potential only,

$$P_0(E) \propto \exp[-U(E)]. \tag{9}$$

We emphasize again that the transition from the master equation (7) to the Fokker-Planck equation (8) is possible only if the transition rates $R_E(E \to E')$—i.e., the underlying transition probabilities $W_E(E \to E')$—are *sufficiently local* in energy. Only in that limit is the Fokker-Planck equation an effective description of the more general equation (7). Nevertheless, even if there are nonlocal contributions to the transition rates, Eq. (8) can be viewed as the best *local* approximation to Eq. (7).

Equation (8) can be written in a more compact form using the stationary distribution $P_0(E)$,

$$\frac{\partial}{\partial t} P(E, t) = \left[\frac{\partial}{\partial E} D(E) P_0(E) \frac{\partial}{\partial E} P_0(E)^{-1} \right] P(E, t). \tag{10}$$

In this form the fact that $P_0(E)$ is the stationary distribution can be seen immediately from the vanishing of the rightmost derivative on the right-hand side (RHS) if $P(E, t)$ is replaced by $P_0(E)$. This equation will be the basis for our further analysis of distribution and flow in energy space.

The stationary distribution in energy space, $P_0(E)$, is actually the histogram $H(E)$ of the energy distribution that is

observed in an actual simulations. It is still given by Eq. (6)—i.e., by the Metropolis weight function $w(E)$ multiplied by the density of states, $n(E)$. By an appropriate choice of the weight function *any* histogram can be produced in the simulation. The usual choice of Boltzmann weights leads to the canonical distribution $H(E) \propto n(E)\exp(-\beta E)$, while the choice $w(E) \propto 1/n(E)$ leads to a flat histogram $H(E)=$ const [2,18]. A flat histogram would be most appropriate to describe the properties of the system in question over a wide temperature range with equal accuracy, provided the Monte Carlo error at each energy is the same. Various methods have been discussed to actually obtain approximations to $n(E)$ by iteratively improving simulations [2,19]. However, all of them are still plagued by the problem that equilibration in the system can be slow, in particular when a wide energy range is considered [20,21]. Moreover, it turned out that, even if a flat histogram is reached, the error distribution is not flat at all [6,22].

It was the important new step by Trebst *et al.* [6] to look systematically at the flow in energy space in simulations. Instead of monitoring the histogram $H(E)$, corresponding to the stationary distribution, they added a label to the system and changed its value whenever it reached minimal and maximal values in the energy—i.e. E_{min} and E_{max}. By counting just these labels at each energy E, the distributions of systems moving up and down in energy, denoted by $n_{up}(E)$ and $n_{down}(E)$, respectively, can be monitored. The original histogram is recovered from $H(E)=n_{up}(E)+n_{down}(E)$. However, in this way it is also possible to measure the fraction of systems moving up,

$$f_{up}(E) = \frac{n_{up}(E)}{n_{up}(E) + n_{down}(E)}, \tag{11}$$

and, correspondingly, that of systems moving down in energy, $f_{down}(E)$. Note that $f_{up}(E)+f_{down}(E)=1$. Both distributions actually are the stationary distributions of probability flow in the systems with boundary conditions $f_{up}(E_{min})=1$, $f_{up}(E_{max})=0$ and $f_{down}(E_{min})=0$, $f_{down}(E_{max})=1$, respectively.

This flow in energy space can now be analyzed using the above Fokker-Planck equations. Equations (8) and (10) are actually continuity equations for the probability flow,

$$\frac{\partial}{\partial t}P(E,t) = \frac{\partial}{\partial E}J(E,t), \tag{12}$$

with $J(E,t)$ the probability current. The current between E_{min} and E_{max} can now be determined as the stationary solution of Eq. (12),

$$J = \left[D(E)P_0(E)\frac{\partial}{\partial E}P_0(E)^{-1}\right]P_J(E) \equiv \text{const}, \tag{13}$$

with $P_J(E)$ being the stationary distribution for the flow under the above boundary conditions. Note that the stationary distribution $P_0(E)$ discussed before is actually the solution of Eq. (13) for zero flow. Integrating that equation we obtain

$$\frac{P_J(E)}{P_0(E)} - \frac{P_J(E_{min})}{P_0(E_{min})} = J\int_{E_{min}}^{E} dE' \frac{1}{D(E')P_0(E')}. \tag{14}$$

Using any of the above boundary conditions the total flow across energy space is then given by

$$|J| = \langle[D(E)P_0(E)]^{-1}\rangle^{-1}, \tag{15}$$

where we used the notation $\langle\cdot\rangle = \int_{E_{min}}^{E_{max}} dE$.

In order to optimize the weight function $w(E)$ used for the simulation to reach maximal flow across energy space, Trebst *et al.* maximized Eq. (15) under the constraint of keeping the distribution $P_0(E)$ normalized, which is done by adding a Lagrange multiplier

$$\frac{\delta}{\delta P_0(E)}[\langle[D(E)P_0(E)]^{-1}\rangle^{-1} + \lambda\langle P_0(E)\rangle] = 0. \tag{16}$$

Stationary flow is not the only concept that can be used to investigate the stochastic dynamics in a system. Another often used concept is the mean FPT [23]. It is the average time a particle starting at one end of a diffusion space needs to reach the other end for the first time. The total mean first passage time for crossing energy space from E_{min} to E_{max} in both directions,

$$\tau = \tau(E_{min} \to E_{max}) + \tau(E_{max} \to E_{min}), \tag{17}$$

can be derived from Eqs. (8) and (10) and is given by [24]

$$\begin{aligned}
\tau &= \int_{E_{min}}^{E_{max}} dE \frac{1}{D(E)P_0(E)} \int_{E_{min}}^{E} dE' P_0(E') \\
&+ \int_{E_{min}}^{E_{max}} dE \frac{1}{D(E)P_0(E)} \int_{E}^{E_{max}} dE' P_0(E') \\
&= \langle[D(E)P_0(E)]^{-1}\rangle\langle P_0(E)\rangle.
\end{aligned} \tag{18}$$

Note that $P_0(E)$ does not have to be normalized here. Minimization of τ with respect to $P_0(E)$ leads to

$$\frac{\delta}{\delta P_0(E)}[\langle[D(E)P_0(E)]^{-1}\rangle\langle P_0(E)\rangle] = 0. \tag{19}$$

Both variational equations (16) and (19) lead to the same solution

$$P_{0,opt} \propto \frac{1}{\sqrt{D(E)}}, \tag{20}$$

which is already derived in [6] from Eq. (16). This current-optimized stationary distribution leads to a symmetric form of the Fokker-Planck equation,

$$\frac{\partial}{\partial t}P(E,t) = \left[\frac{\partial}{\partial E}\sqrt{D(E)}\frac{\partial}{\partial E}\sqrt{D(E)}\right]P(E,t). \tag{21}$$

In order to reach such a current optimized histogram $H_{opt}(E)$ in an actual simulation, the weight function has to be chosen accordingly. For this purpose it is necessary to obtain the local diffusion coefficient from a simulation using some initial weight function $w(E)$. Differentiating Eq. (14), $D(E)$ is obtained from

$$D(E) = \left[P_0(E) \frac{d}{dE} \frac{P_J(E)}{P_0(E)} \right]^{-1} = [H(E)f'(E)]^{-1}, \quad (22)$$

where we used the fact that

$$f(E) = \frac{P_J(E)}{P_0(E)} \quad (23)$$

holds. Both $f_{up}(E)$ and $f_{down}(E)$ can be used for that purpose. However, some smoothening may have to be performed to obtain a smooth derivative, as discussed in [6]. Since $P_{0,opt}(E) = n(E)w_{opt}(E)$, the optimized weight function is given by

$$w_{opt}(E) = \frac{1}{n(E)\sqrt{D(E)}}. \quad (24)$$

Using the fact that $n(E)$ is obtained from the actual simulation by $n(E) = H(E)/w(E)$, we arrive at the iteration formula

$$w_{opt}(E) = w(E) \sqrt{\frac{f'(E)}{H(E)}}. \quad (25)$$

At the fixed point of this iteration $f'(E) = H(E)$, leading to $H(E) = 1/\sqrt{D(E)}$ as required; see Eq. (22).

In an actual simulation situation one iteration might not suffice. This is discussed in detail in [6–8]. In the following we will attempt to apply the same approach to tempering simulations.

III. PARALLEL TEMPERING

Generalized ensemble sampling can be extended by adding movement in *control parameter space*. In addition to moves among different states of the system at a particular value of the control parameter, moves along one or more control parameter directions are possible. The motivation behind such an extension is that relaxation in some control parameter regime may be faster, thereby facilitating relaxation in the whole *state+control parameter space* [3].

Usually, the motion along control parameter directions is not made continuous. Instead, a list of monotonically increasing or decreasing values β_n is chosen, thereby introducing control parameter hopping. The most commonly used control parameter is temperature, although other choices are possible, too [25]. Simulations at a particular control parameter value are performed using the Metropolis criterion (1) with a weight function $w(\beta, s)$, as before. Although Boltzmann weights are usually chosen, leading to canonical simulations at each parameter value, in principle any weight function is possible.

Parallel tempering is the parallelized extension of a method called simulated tempering which we will sketch only briefly. In simulated tempering [26,27], a single instance of the system is simulated only. After certain times, an attempt is made to change the control parameter value. In order to ensure that the stationary distribution at each control parameter value is given by $w(\beta, s)$, the Metropolis criterion

$$p_M((\beta, s) \rightarrow (\beta', s)) = \min\left(1, \frac{w(\beta', s)}{w(\beta, s)} \right) \quad (26)$$

is used for the acceptance of a control parameter change. In order to ensure appreciable exchange between different control parameter values using Boltzmann weights (2), the weight function has to be adjusted by a control parameter dependent function $g(\beta)$,

$$w(\beta, s) = \exp[-\beta E(s) + g(\beta)]. \quad (27)$$

This function $g(\beta)$ actually determines the distribution of the single system among the control parameter values and, up to an additive constant, the optimal choice is the free energy $g(\beta) = \beta F(\beta)$ of the system analyzed [27]. However, the problem of determining $g(\beta)$ has hampered this approach.

This problem is solved in parallel tempering. Here, copies of the system are simulated in parallel at each of the various control parameter values. At certain times *exchanges* of replicas with neighboring control parameter values are attempted. In order to ensure that the stationary distribution at a control parameter value β is given by $w(\beta, s)$, the generalized Metropolis criterion

$$p_M((\beta, s) \rightarrow (\beta', s')) = \min\left(1, \frac{w(\beta, s')w(\beta', s)}{w(\beta, s)w(\beta', s')} \right) \quad (28)$$

has to be used for the acceptance of such an attempt. It can be seen easily that any function $g(\beta)$ in the weight function (27), in particular the one that was necessary to ensure equilibration among control parameter values in simulated tempering, simply drops out in the parallel form. Thereby, the problem of determining the free energy in order to optimize the simulation vanishes. In the case of Boltzmann weights Eq. (28) reduces to

$$p_M((\beta, s) \rightarrow (\beta', s')) = \min(1, \exp(\Delta\beta\Delta E)), \quad (29)$$

with $\Delta\beta = \beta' - \beta$ and $\Delta E = E' - E$.

We will concentrate on temperature hopping in the following and choose the list of inverse temperatures $\beta_0 > \beta_1 > \cdots > \beta_N$. In a parallel implementation simulations at a particular value β_n are usually run on a particular node of the parallel computer, conveniently labeled n. In order to simplify the notation, we will therefore abbreviate β_n by n whenever possible and also use "node" synonymously with "control parameter value."

It is sufficient to follow only a single replica through state and control parameter space, since all replicas are equivalent. For times between replica exchanges, simulations are performed at each node independently, and the time evolution of the distribution function $P[(\beta_n, s), t]$ at a particular node n is described by the master equation (3) with the respective transition probabilities determined by the appropriate temperature β_n. For times $t = mT$, $m = 1, 2, \ldots$, replica exchange is attempted. For this time step the master equation in state and temperature space is

$$P((\beta_n,s),t+1) = \sum_{s'} \{P((\beta_{n-1},s'),t)W_s((\beta_{n-1},s') \to (\beta_n,s))$$
$$+ P((\beta_{n+1},s'),t)W_s((\beta_{n+1},s') \to (\beta_n,s))\}$$
$$+ P((\beta_n,s),t)$$
$$\times \sum_{s'} \{1 - W_s((\beta_n,s) \to (\beta_{n-1},s'))$$
$$- W_s((\beta_n,s) \to (\beta_{n+1},s'))\}. \tag{30}$$

The transition probabilities are given by

$$W_s((\beta,s) \to (\beta',s')) = \frac{1}{\Omega N} p_M((\beta,s) \to (\beta',s')) \quad \text{for } s \neq s'. \tag{31}$$

Here, Ω is the normalization by state space and reflects the fact that any conformations can be exchanged, which is different from the case of GE sampling where the move set was restricting possible conformation changes; see Eq. (4). N takes into account that just for one random neighboring pair of nodes a replica exchange is attempted at time $t=mT$. However, other strategies are possible, too, that would lead to a different normalization constant. Due to the exchange of replicas in parallel tempering the transition probabilities are symmetric—i.e.,

$$W_s((\beta,s) \to (\beta',s')) = W_s((\beta',s') \to (\beta,s)). \tag{32}$$

Elimination of fast degrees of freedom orthogonal to the energy is possible in the same way as it was discussed in the last section. This leaves us with

$$P((\beta_n,E),t+1) = \sum_{E'} \{P((\beta_{n-1},E'),t)$$
$$\times W_E((\beta_{n-1},E') \to (\beta_n,E))$$
$$+ P((\beta_{n+1},E'),t)W_E((\beta_{n+1},E') \to (\beta_n,E))\}$$
$$+ P((\beta_n,E),t)$$
$$\times \sum_{E'} \{1 - W_E((\beta_n,E) \to (\beta_{n-1},E'))$$
$$- W_E((\beta_n,E) \to (\beta_{n+1},E'))\}. \tag{33}$$

The symmetry (32) naturally leads to a similar symmetry for the transition rates between nodes in reduced—i.e., energy—space,

$$W_E((\beta,E) \to (\beta',E')) = W_E((\beta',E') \to (\beta,E)). \tag{34}$$

In order to finally derive an effective master equation for hopping in temperature space, we have to additionally assume fast relaxation in energy space—i.e., on times scales $t<T$. This means that at any particular node we assume to have reached the respective equilibrated distribution $P_0(\beta,E)$. Using similar reasoning as in Appendix A for GE sampling, this last approximation leads to the final form of the master equation in temperature space on a coarse-grained time scale $t \to t/TN$,

$$P(\beta_n,t+1) = P(\beta_{n-1},t)W_\beta(\beta_{n-1} \to \beta)$$
$$+ P(\beta_{n+1},t)W_\beta(\beta_{n+1} \to \beta) + P(\beta_n,t)$$
$$\times [1 - W_\beta(\beta \to \beta_{n-1}) - W_\beta(\beta \to \beta_{n+1})]. \tag{35}$$

In a way similar to the derivation of Eq. (A3), effective transition probabilities can be derived from the equilibrated distributions at a node,

$$W_\beta(\beta \to \beta') = \int dE \int dE' P_0(\beta,E)$$
$$\times p_M((E,\beta) \to (E',\beta'))P_0(\beta',E'). \tag{36}$$

We will discuss in Appendix B the properties of these effective transition probabilities in particular situations. Finally, the symmetry of the transition probabilities

$$W_\beta(\beta \to \beta') = W_\beta(\beta' \to \beta) \tag{37}$$

holds analogously here, too.

Due to this last property, the stationary distribution for PT can be derived easily from Eq. (35) and is given simply by

$$P_0(\beta) = \text{const}; \tag{38}$$

i.e., in an equilibrated simulation every single replica appears on each control parameter node with the same probability $1/(N+1)$, with $N+1$ being the number of nodes. This result is an important simplification of the situation over the case of simulated annealing and over GE sampling, and is due to the construction of replica exchange.

Flow between the control parameter nodes can now be analyzed, too. Here, replicas reaching node 0 or N are labeled and the respective distributions over the nodes can be monitored. Using the same notation as before, $n_{up}(i)$ being the number of replicas at node i that came from node 0 and $n_{down}(i)$ being the number of replicas at node i that came from node N, we can measure the fraction of replicas moving up

$$f_{up}(i) = \frac{n_{up}(i)}{n_{up}(i) + n_{down}(i)}, \tag{39}$$

and a corresponding quantity for those moving down, $f_{down}(i)$. Both distributions are stationary distributions of probability flow between temperature nodes, with boundary conditions $f_{up}(0)=1, f_{up}(N)=0$ and $f_{down}(0)=0, f_{down}(N)=1$, respectively. As before, they can be analyzed by looking at the stationary solution of the underlying stochastic equation (35). It can be written as

$$P(\beta_n,t+1) - P(\beta_n,t) = J(\beta_n,t) - J(\beta_{n-1},t), \tag{40}$$

which is the discrete form of the continuity equation (12). The discrete case current $J(\beta_n,t)$ is given by

$$J(\beta_n,t) = P(\beta_{n+1},t)W_\beta(\beta_{n+1} \to \beta_n) - P(\beta_n,t)W_\beta(\beta_n \to \beta_{n+1}). \tag{41}$$

Consequently, the stationary current J is determined from

$$J = P_J(\beta_{n+1}) W_\beta(\beta_{n+1} \to \beta_n) - P_J(\beta_n) W_\beta(\beta_n \to \beta_{n+1}) = \text{const},$$

(42)

with $P_J(\beta_n)$ being the stationary distribution for flow under the above boundary conditions. Taking into account the symmetry properties of the transition probabilities, the stationary distribution for constant current between nodes β_0 and β_N is given by

$$P_J(\beta_n) = J^{-1} \sum_{i=0}^{n-1} \frac{1}{W_\beta(\beta_i \to \beta_{i+1})},$$

(43)

with the current

$$J = \left[\sum_{i=0}^{N-1} \frac{1}{W_\beta(\beta_i \to \beta_{i+1})} \right]^{-1}.$$

(44)

Due to the simple form of the stationary distribution among nodes (38), the analytic forms for $P_J(i)$ and J are considerably simpler than for GE sampling.

Again, as before with GE sampling, a concept different from stationary flow can also be analyzed, the total mean first passage time to cross the network of nodes. For the general hopping process (35), this is given by [15,17]

$$\begin{aligned}
\tau &= \tau(0 \to N) + \tau(N \to 0) \\
&= \sum_{i=0}^{N-1} \frac{1}{P_0(\beta_i) W_\beta(\beta_i \to \beta_{i+1})} \sum_{j=0}^{i} P_0(\beta_i) \\
&\quad + \sum_{i=1}^{N} \frac{1}{P_0(\beta_i) W_\beta(\beta_i \to \beta_{i-1})} \sum_{j=i}^{N} P_0(\beta_i) \\
&= \left(\sum_{i=0}^{N-1} \frac{1}{P_0(\beta_i) W_\beta(\beta_i \to \beta_{i+1})} \right) \left(\sum_{i=0}^{N} P_0(\beta_i) \right).
\end{aligned}$$

(45)

This final result is practically a discrete version of Eq. (18). Taking into account that the stationary distribution in PT is constant, we finally obtain a result that is just the inverse of Eq. (44),

$$\tau = \sum_{i=0}^{N-1} \frac{1}{W_\beta(\beta_i \to \beta_{i+1})}.$$

(46)

Here is an important difference from GE sampling. There the local diffusion coefficient was fixed by the move set and, as turned out in actual simulations, mostly independent from the chosen weight function. The stationary distribution, however, was free to be chosen by varying the weight function.

In the case of PT this is exactly the opposite. Here the stationary distribution is fixed by construction of the replica exchange process. However, by adjusting the control parameter intervals, the transition probabilities can be chosen relatively freely. Hence, we are interested in obtaining the optimal distribution of transition probabilities that maximizes the flow across the control parameter space. Since we are mainly interested in local variations of the optimized transition probabilities—i.e., in deviations from an average value

(which will be actually determined afterwards), we keep the average transition probability constant via a Lagrangian multiplier—i.e.,

$$\frac{\delta}{\delta W_\beta(\beta_j \to \beta_{j+1})} \left[\left(\sum_{i=0}^{N-1} \frac{1}{W_\beta(\beta_i \to \beta_{i+1})} \right)^{-1} + \lambda \sum_{i=0}^{N-1} W_\beta(\beta_i \to \beta_{i+1}) \right] = 0.$$

(47)

An equivalent equation results for optimizing τ.

It is easily seen that optimizing the current as well as the mean first passage time simply gives a constant transition probability between neighboring nodes the whole range of control parameter values,

$$W^{opt} = \text{const},$$

(48)

as optimal solution. Consequently, from Eq. (43) we can conclude that the optimal flow distribution among the nodes is linear in the node number,

$$P_J(\beta_n) = n/N.$$

(49)

Therefore, the temperature spacing is optimal if such a flow distribution *together* with constant transition probabilities can be obtained in an actual simulation. The linear dependence of the flow distribution on the node number, Eq. (49), was obtained for PT already in [7–9] by mapping PT onto the Fokker-Planck equation for GE sampling, Eq. (8), and assuming a particular temperature dependence for the local diffusion coefficient. Here, however, we see that it follows directly from the hopping description of PT. In addition, we obtain its equivalence to constant transition probabilities, Eq. (48).

The iteration scheme used in Refs. [7–9] for assigning temperatures to nodes appeared to exhibit fast convergence to the optimal behavior of [49]. We can rephrase it here without having to recur to some intermediate local diffusivity (and we stick to our use of inverse temperatures as example control parameters).

(i) A particular set of control parameters $\beta_0 > \beta_1 > \ldots > \beta_N - 1 > \beta_N$ gives rise to a flow distribution $f_{up}(0) = 1 \geq f_{up}(1) \geq \ldots \geq f_{up}(N-1) \geq f_{up}(N) = 0$.

(ii) These later values give rise to stepwise defined function $g[f]$, with $g[f(i)] = \beta_i$, in particular $g[1] = \beta_0$ and $g[1] = \beta_N$, and linear interpolation in between these values.

(iii) The new control parameter values are determined from this function by $\beta_i' = g[i/N]$, $i = 1, \ldots, N-1$, keeping β_0 and β_N fixed.

This procedure is actually illustrated quite nicely in Fig. 2 of Ref. [7], and we can refrain from repeating it here.

It is important to note at this point that the above results, in particular the *equivalence* of constant transition probabilities and a flow distribution that is linear in the node number, depend on the validity of the underlying one-dimensional representation of the simulation process. It turns out that, while the flow distribution (49) is readily attainable in actual simulations, this is usually not accompanied by acceptance probabilities that are constant over the whole system [7–9].

In the following final section we will discuss reasons for such a discrepancy.

IV. SUMMARY, DISCUSSION, AND OUTLOOK

The goal of most recent advances in Markov chain Monte Carlo sampling is to analyze and increase the flow through state space. To do so, heuristic equations have been used to describe the flow in reduced state space along a slow order parameter—e.g., the energy—in generalized ensemble sampling and among nodes performing simulations at various control parameter values in parallel tempering (PT) [6–8]. In this contribution we have derived such one-dimensional stochastic equations for GE and PT sampling from the underlying master equations. Using these stochastic equations, weight functions for GE sampling and strategies for finding optimal control parameter values for PT can be devised that optimize the flow through order parameter and control parameter space, respectively. We have also demonstrated that optimization of flow is equivalent to minimizing the first passage time to cross the system.

All considerations in the previous sections were based on the assumptions that the Fokker-Planck equation (8) or hopping equation (35) are a valid Markovian representation of the underlying more complex dynamics. That, however, is true only if the approximations discussed there apply.

In the case of GE sampling these approximations are such that relaxations in the degrees of freedom orthogonal to the energy are fast, together with the locality of transitions. Since the ultimate goal of deriving Eq. (8) is the optimization of flow through state space, a violation of the latter condition is not detrimental. Nonlocality in the energy of the move set usually leads to faster relaxation since state space is connected more densely. This is, for example, one reason for the success of the Swendsen-Wang algorithm [28]. Although in such a situation Eq. (8) may not be able to capture the full dynamics correctly, it nevertheless is still able to identify local bottlenecks. Optimizing the flow according to the methods discussed can handle these local bottlenecks in the transition between neighboring energy values.

However, slow relaxation orthogonal to the energy leads to a more complicated situation. Now it may not be possible anymore to reach all values of the additional degrees of freedom by moves local in the energy. Instead, detours via other—usually high-energy—areas of the state space have to be performed. This leads to the comblike structure of the accessible state space sketched in Fig. 1. It describes the situation that free energy basins at constant energy are disconnected.

Actually, it is this feature of state space partitioning that led us to the requirement of large flow between low- and high-energy areas of state space in the first place. Only for a large flow can the "teeth" of the comblike structure in Fig. 1 be sampled adequately. Nevertheless, it also leads to the situation that the effective one-dimensional Fokker-Planck equation (8) is only the best Markovian one-dimensional approximation of an underlying effectively higher-dimensional process.

This situation is even clearer in the case of PT. If the relaxation at a particular control parameter value is faster
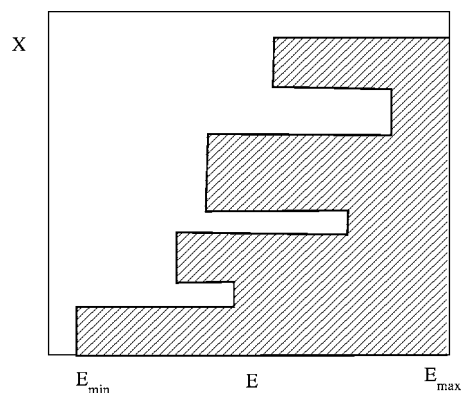

FIG. 1. Sketch of state space for generalized ensemble sampling in the case of broken ergodicity; X denotes any degree of freedom orthogonal to the control parameter (energy).

than the time scale of hopping in control parameter space, then the requirements for the analysis performed in the previous section are fulfilled. However, if that is not the case, the state space at such a node partitions into disjoint free energy basins that do not communicate. Viewed over the whole control parameter range, we are dealing with a hierarchical network of free energy basins as sketched in Fig 2. Such a situation has been aptly termed *broken ergodicity* [29,30] and was discussed in the field of glassy dynamics several years ago.

In principle, the topology of the treelike control+state space depends on the time scale of the control parameter hopping. If relaxation is possible at all nodes, then no bifurcations occur and the system is just a one-dimensional hopping chain as was analyzed in the previous section. However, this is true only in the limit of infinite time between replica exchange steps, $T \rightarrow \infty$. Practically the topology of the branching will be the same over a wide range of time scales and only the position of the branching nodes may vary.

For a truly one-dimensional system the optimized transition probabilities are constant and, equivalently, the optimal flow stationary distribution is $f(n) \propto n$, Eq. (49). Under conditions of broken ergodicity, the situation is more complicated. Now, several transition rates between neighboring nodes may have to be taken into account, describing exchange along the different branches depicted in Fig. 2. Moreover, *observed* acceptance rates are weighted averages of
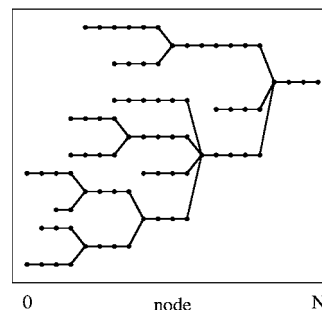

FIG. 2. Sketch of bifurcations and multifurcations in the case of broken ergodicity for parallel tempering; for certain nodes, the system partitions into several disjoint free energy wells.

these various transition probabilities. Therefore, the equivalence of constant observed acceptance ratios to a flow distribution linear in the node number may no longer hold. This discrepancy has been observed already in Refs. [7–9]. While satisfying Eq. (49) was possible by an appropriate choice of node temperatures, constant acceptance rates were not obtained concomitantly. Such a result can be used as a clear signature of broken ergodicity occurring in PT. In contrast to GE sampling, where such a clear criterion is not available yet, PT has a particular advantage here.

Naturally, the question arises how to optimize flow in PT under conditions of broken ergodicity. Control via the choice of node temperatures is somewhat limited in such a situation, since changes may affect different transition probabilities differently. Nevertheless, even in such a situation the choice of a flow distribution $f(n) \propto n$ still assures that the flow of replicas along the main branch–i.e., between the lowest and the highest temperature node—is still optimal. In contrast, the effect of making apparent acceptance ratios constant, if possible at all, is unclear and depends on knowledge of the particular structure of ergodicity breaking. In the final analysis, however, optimizing flow under such conditions means that, in addition to the flow between the lowest and highest nodes, also flow among side branches has to be considered. In order to assure that flow among all side branches is optimized, too, a more detailed flow analysis—i.e., determining the flow matrix between all individual nodes—would have to be performed.

We believe that an additional advantage of PT is that such branching situations can be analyzed directly, without resorting to an actual system. In a way, it will be possible to *simulate the simulations* to investigate possible flow behaviors. By analyzing the master equations modeling the hierarchical broken ergodicity networks, conclusions about the behavior of actual simulations can be drawn [31]. Thereby, the present results open the way to investigate the effects of broken ergodicity in GE and PT sampling in a more systematic way.

APPENDIX A: ADIABATIC ELIMINATION OF FAST DEGREES OF FREEDOM

As a first step to eliminate the fast degrees of freedom in Eq. (3), we have to single out the slow degree of freedom by appropriate labeling. Therefore, we replace the state label s by a more detailed one that consists of the energy E—i.e., the slow degree of freedom—and an additional label s_E that designates all microstates with energy E, $s \equiv (E, s_E)$. Equation (3) is then replaced by

$$P((E,s_E),t+1) = \sum_{E' \neq E} \sum_{s_{E'}} P((E',s_{E'}),t)$$
$$\times W_s((E',s_{E'}) \to (E,s_E)) + P((E,s_E),t)$$
$$\times \left\{ 1 - \sum_{E' \neq E} \sum_{s_{E'}} W_s((E,s_E) \to (E',s_{E'})) \right\},$$
(A1)

where the sums are over all possible energies E and states s_E for each energy. If relaxation among the microstates for a particular energy is fast, each of these states will assume the same probability, and we can approximate

$$P[(E,s_E),t] \approx \frac{1}{N(E)} P(E,t),$$
(A2)

with $N(E)$ being the number of states—i.e., $\sum_{s_E} = N(E)$, introduced to ensure correct normalization, $\sum_E P(E,t) = 1$. This approximation is the crucial step, since it allows us to sum over all microstates s_E for each energy in Eq. (A1). After carrying out such a summation, we arrive at the master equation for the energy, Eq. (5), with the transition probabilities $W_E(E \to E')$ given by

$$W_E(E \to E') = \frac{1}{N(E)} \sum_{s_E} \sum_{s_{E'}} W_s((E,s_E) \to (E',s_{E'})).$$
(A3)

Note the asymmetry with respect to E and E' that is due to the nonconstant density of states.

A more rigorous and systematic treatment, which would also allow the derivation of corrections, would involve projection operator techniques as they are used, e.g., in Chap. 6.4 of Ref. [17]. However, since the above approach suffices for our purposes, we refrain from embarking on such a more detailed elaboration.

APPENDIX B: TRANSITION PROBABILITIES FOR PARALLEL TEMPERING

In GE sampling, a general analysis of the effective transition probabilities $W(E \to E')$ is not easy since they depend strongly on the—usually unknown—distributions orthogonal to the energy, combined with a possibly sparse move set. In contrast to GE sampling, PT allows more insight into the effective transition probabilities that govern the temperature hopping. Since transitions are possible between any energy values, the complications due to a particular sparse move set do not arise. The only requirement is the assumption of equilibration at each temperature. In this limit, we can calculate the effective transition probability by

$$W(\beta \to \beta') = \int dE \int dE' P_0(\beta,E)$$
$$\times p_M((E,\beta) \to (E',\beta')) P_0(\beta',E'), \quad \text{(B1)}$$

with $P_0(\beta,E)$ being the equilibrated distribution at β and p_M given by Eq. (28).

There have been several approaches to evaluate that formula. Predescu *et al.* [32] and Kofke [33] emphasize the

importance of taking into account the asymmetry of an actual distribution, having a low-energy cutoff and an exponential tail at high energies. Nevertheless, Kone and Kofke [34] later use an approximation based on a Gaussian approximation— i.e., symmetric without cutoff and nonexponential tail— together with the assumption of constant specific heat over the entire range. All authors limit themselves to unimodal distributions in their analysis and do not question the peak is quadratic in energy.

However, the distributions change dramatically at critical values of the control parameter—i.e., at first- and second-order phase transitions. While at second-order phase transitions the functional form of the peak changes from quadratic to quartic, at first-order phase transitions the energy distributions become bimodal. With respect to the distribution tails, on the other hand, since the goal is anyhow to optimize the transition probabilities, one should try to avoid control parameter intervals so large that the explicit structure of the tails become relevant. Moreover, Eq. (B1) is anyhow only valid in the limit of fast relaxation at a node. So it is from the outset only an approximation to the actually observed transition rate. Since in order to optimize the flow large values of the transition probabilities are sought for, a quantitative analysis makes sense only for the cases where the overlap is appreciable—i.e., for small temperature differences.

We therefore add here our approach to evaluate Eq. (B1) using the first-order approximations to these distributions— i.e. Gaussians,

$$P_0(\beta, E) \propto \exp\left[-\frac{[E - \bar{E}(\beta)]^2}{2\sigma^2(\beta)} \right], \tag{B2}$$

with $\bar{E}(\beta)$ the average energy and $\sigma^2(\beta) = \overline{[E - \bar{E}(\beta)]^2}$ the energy fluctuations at β. However, we avoid the unrealistic and very limiting assumption of a constant specific heat [34].

Assuming $\beta < \beta'$, using a step function approximation of error functions that result from the inner integrals, and performing a symmetric evaluation we obtain

$$W(\beta \rightarrow \beta') \approx \frac{1}{4} \exp\left[\Delta\beta\Delta E + \frac{1}{2}\Delta\beta^2(\sigma + \sigma') \right]$$

$$\times \left[2 + \mathrm{erf}\left(\frac{\Delta E + \Delta\beta(\sigma^2 + \sigma'^2)}{\sqrt{2}\sigma} \right) \right.$$

$$\left. + \mathrm{erf}\left(\frac{\Delta E + \Delta\beta(\sigma^2 + \sigma'^2)}{\sqrt{2}\sigma'} \right) \right]$$

$$+ \frac{1}{4}\left[\mathrm{erfc}\left(\frac{\Delta E}{\sqrt{2}\sigma} \right) + \mathrm{erfc}\left(\frac{\Delta E}{\sqrt{2}\sigma'} \right) \right], \tag{B3}$$

where we have used the abbreviations $\Delta\beta = \beta - \beta'$ and ΔE

$= \bar{E}(\beta) - \bar{E}(\beta')$. We can now employ the fact that the specific heat is given by

$$c = \frac{d}{dT}\bar{E}(T) = -\beta^2 \frac{d}{d\beta}\bar{E}(\beta), \tag{B4}$$

as well as by

$$c = \beta^2 \overline{[E - \bar{E}(\beta)]^2} = \beta^2 \sigma^2(\beta), \tag{B5}$$

to obtain a relationship between the derivative of the average energy and the energy fluctuations:

$$-\frac{d}{d\beta}\bar{E}(\beta) = \overline{[E - \bar{E}(\beta)]^2}. \tag{B6}$$

Using

$$\bar{E}'(\beta) \equiv \frac{d}{d\beta}\bar{E}(\beta) = -\sigma^2(\beta), \tag{B7}$$

this relation can be employed to approximate the difference of the average energies by

$$\Delta E = \bar{E}(\beta) - \bar{E}(\beta') \approx \frac{1}{2}[\bar{E}'(\beta) + \bar{E}'(\beta')]\Delta\beta$$

$$= -\frac{1}{2}(\sigma^2 + \sigma'^2)\Delta\beta. \tag{B8}$$

Note that we have assumed $\beta < \beta'$—i.e., $\Delta\beta < 0$—and, consequently, $\Delta E > 0$ in the evaluation of Eq. (B1). This result shows that the exponent in Eq. (B3) cancels and, using $\mathrm{erf}(-x) = -\mathrm{erf}(x)$, we have the final approximate expression for the transition probability:

$$W(\beta \rightarrow \beta') \approx \frac{1}{2}\left[\mathrm{erfc}\left(\frac{\Delta E}{\sqrt{2}\sigma} \right) + \mathrm{erfc}\left(\frac{\Delta E}{\sqrt{2}\sigma'} \right) \right]$$

$$= \left[\mathrm{erfc}\left(\frac{|\Delta\beta|(\sigma^2 + \sigma'^2)}{2\sqrt{2}\sigma} \right) \right.$$

$$\left. + \mathrm{erfc}\left(\frac{|\Delta\beta|(\sigma^2 + \sigma'^2)}{2\sqrt{2}\sigma'} \right) \right]. \tag{B9}$$

For small values of $\Delta E/\sigma$ this can be further approximated to

$$W(\beta \rightarrow \beta') \approx 1 - \frac{1}{\sqrt{2\pi}}\left(\frac{\Delta E}{\sqrt{2}\sigma} + \frac{\Delta E}{\sqrt{2}\sigma'} \right) + \frac{1}{6\sqrt{2\pi}}$$

$$\times \left[\left(\frac{\Delta E}{\sqrt{2}\sigma} \right)^3 + \left(\frac{\Delta E}{\sqrt{2}\sigma'} \right)^3 \right] + O(\Delta E^5). \tag{B10}$$

[1] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).

[2] B. A. Berg, J. Stat. Phys. **82**, 323 (1996).

[3] C. J. Geyer and A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995).

[4] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).

[5] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[6] S. Trebst, D. A. Huse, and M. Troyer, Phys. Rev. E **70**, 046701 (2004).

[7] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).

[8] S. Trebst, D. A. Huse, E. Gull, H. G. Katzgraber, U. H. E. Hansmann, and M. Troyer, in *Developments in Computer Simulation Studies in Condensed Matter Physics XIX, Springer Proceedings in Physics*, Vol. 115, edited by D. P. Landau, S. P. Lewis, and H.-B. Schüttler (Springer, Berlin, 2007).

[9] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, J. Stat. Mech.: Theory Exp. (2006) P03018.

[10] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[11] P. Diaconis, Proc. Natl. Acad. Sci. U.S.A. **93**, 1659 (1996).

[12] P. Diaconis and L. Saloff-Coste, J. Comput. Syst. Sci. **57**, 20 (1998).

[13] L. J. Billera and P. Diaconis, Stat. Sci. **16**, 335 (2001).

[14] P. Diaconis and J. W. Neuberger, Exp. Math. **13**, 207 (2004).

[15] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Amsterdam, North Holland, 1981).

[16] H. Risken, *The Fokker-Planck Equation: Methods of Solutions and Applications*, Springer Series in Synergetics (Springer, Berlin, 1984).

[17] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer, Berlin, 1985).

[18] B. A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991).

[19] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[20] P. Dayal, S. Trebst, S. Wessel, D. Wurte, M. Troyer, S. Sabhapandit, and S. N. Coppersmith, Phys. Rev. Lett. **92**, 097201 (2004).

[21] S. Alder *et al.*, J. Stat. Mech.: Theory Exp. (2004) P07008.

[22] B. J. Schulz, K. Binder, and M. Mller, Int. J. Mod. Phys. C **13**, 477 (2002).

[23] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, Cambridge, England, 2001).

[24] K. Schulten, Z. Schulten, and A. Szabo, J. Chem. Phys. **74**, 4426 (1981).

[25] W. Kwak and U. H. E. Hansmann, Phys. Rev. Lett. **95**, 138102 (2005).

[26] M. Lubartsev and V. Vorontzov, J. Chem. Phys. **96**, 1776 (1992).

[27] E. Marinari and R. Parisi, Europhys. Lett. **69**, 2292 (1992).

[28] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987).

[29] R. Palmer, Adv. Phys. **31**, 669 (1982).

[30] D. L. Stein and C. M. Newman, Phys. Rev. E **51**, 5228 (1995).

[31] W. Nadler and U. H. E. Hansmann (unpublished).

[32] C. Predescu, M. Predescu, and C. Ciabanu, J. Chem. Phys. **120**, 4119 (2004).

[33] D. A. Kofke, J. Chem. Phys. **117**, 6911 (2004).

[34] A. Kone and D. A. Kofke, J. Chem. Phys. **122**, 206101 (2005).