GSHMC: An efficient method for molecular simulation

Elena Akhmatskaya*

Fujitsu Laboratories of Europe Ltd (FLE), Hayes Park Central,
Hayes End Road, Hayes UB4 8FE, United Kingdom

Sebastian Reich†

Universität Potsdam, Institut für Mathematik,
Am Neuen Palais 10, D-14469 Potsdam, Germany

(Dated: January 18, 2008)

Abstract

The hybrid Monte Carlo (HMC) method is a popular and rigorous method for sampling from a canonical ensemble. The HMC method is based on classical molecular dynamics simulations combined with a Metropolis acceptance criterion and a momentum resampling step. While the HMC method completely resamples the momentum after each Monte Carlo step, the generalized hybrid Monte Carlo (GHMC) method can be implemented with a partial momentum refreshment step. This property seems desirable for keeping some of the dynamic information throughout the sampling process similar to stochastic Langevin and Brownian dynamics simulations. It is, however, ultimate to the success of the GHMC method that the rejection rate in the molecular dynamics part is kept at a minimum. Otherwise an undesirable *Zitterbewegung* in the Monte Carlo samples is observed. In this paper, we describe a method to achieve very low rejection rates by using a modified energy, which is preserved to high-order along molecular dynamics trajectories. The modified energy is based on backward error results for symplectic time-stepping methods. The proposed generalized shadow hybrid Monte Carlo (GSHMC) method is applicable to NVT as well as NPT ensemble simulations.

PACS numbers: 47.10.Df, 47.11.Mn, 31.15.-p, 31.25.Qg, 02.50.Ga, 02.70.Uu, 02.70.Ns, 82.20.Wt

*Electronic address: `Elena.Akhmatskaya@uk.fujitsu.com`

†Electronic address: `sreich@math.uni-potsdam.de`

I. INTRODUCTION

A rigorous method for performing constant temperature simulations is provided by the hybrid Monte Carlo (HMC) method [1, 2]. The HMC method combines constant energy molecular dynamics simulations with a Metropolis acceptance criterion and a momentum resampling step. It is crucial that the constant energy molecular dynamics simulations are performed with a volume preserving and time-reversible method. While having the advantage of providing a rigorous sampling technique, practical experience shows, however, that the acceptance rate in the molecular dynamics part of HMC decreases with the size of the molecular system. In particular, HMC simulations become rather inefficient for large biomolecular simulations. Possible rescues include reduction of step-size or increase of accuracy of the molecular simulations by using a higher-order method. Both approaches increase however the computational cost significantly. A different approach has been considered by Hampton and Izaguirre [3], who suggest to make use of the modified equations analysis available for symplectic time-stepping methods such as the Störmer-Verlet method. The fundamental result of [4–6] is that any symplectic integrator (see [7, 8] for a general discussion of symplectic methods) possesses a modified Hamiltonian $\mathcal{H}_{\Delta t}$, which is preserved along the numerical trajectories up to terms $\propto \exp(-c/\Delta t)$, where $c > 0$ is a constant and Δt is the step-size. The shadow hybrid Monte Carlo (SHMC) method [3] samples from a properly defined modified energy and is able to achieve very high acceptance rates in the molecular dynamics part of HMC. Efficient algorithms for computing modified energies can be found in [9] and [10]. However, the momentum resampling step becomes more complex under the SHMC method. In fact, the necessary balance between increased acceptance in the molecular dynamics update and reduced acceptance in the momentum updates limits the efficiency gains of SHMC over HMC [11]. More recently, the S2HMC method has been introduced in [12], which overcomes the efficiency limitation of SHMC at the level of fourth-order modified energies. An extension of S2HMC to higher-order modified energies is currently not available.

In a related paper [13], Akhmatskaya and Reich proposed the targeted shadow hybrid Monte Carlo (TSHMC) method, which combines the idea of modified energies for HMC with a partial momentum update. It is the purpose of the present paper to develop the TSHMC method further by making a link to the generalized hybrid Monte Carlo (GHMC) method

[14, 15]. We call the new method generalized shadow hybrid Monte Carlo (GSHMC). The link to GHMC will allow us in particular to develop an efficient momentum refreshment step for GSHMC based on the work of [15]. This partial momentum update keeps some of the dynamic information throughout the sampling process similar to stochastic Langevin and Brownian dynamics simulations. Furthermore, we develop the GSHMC method for molecular systems in generalized coordinates and for the constant pressure formulation of Andersen [16] in particular. A key prerequisite is the derivation of an appropriate symplectic and time-reversible time-stepping method and the formulation of modified energies. A high acceptance rate in the molecular dynamics part of GSHMC is necessary to avoid an undesirable *Zitterbewegung* due to momentum reversal after a rejected molecular dynamics update. Under the GSHMC method we can achieve this by using modified energies of high enough order. We finally note that the possibility of combining HMC with the constant pressure method of Andersen has been indicated in [2] already.

The paper is organized as follows. We first summarize the GHMC method. We then show how to derive a symplectic and time-reversible time-stepping method for constant energy molecular dynamics in generalized coordinates. This is followed by the introduction of the GSHMC method, the derivation of a fourth-order modified energy, and the discussion of improved momentum refreshment steps. We provide implementation details for GSHMC simulations under an NVT and NPT ensemble. We demonstrate that the constant pressure GSHMC method can be thought of as a rigorous implementation (in the sense of time-stepping artifacts) of the Langevin piston method of Feller et al. [17]. We finally provide numerical results from simulations for argon and a lysozyme protein (2LZM) in water solvent.

II. THE GENERALIZED HYBRID MONTE CARLO METHOD

We consider a molecular system with m degrees of freedom described by generalized coordinates $\mathbf{q} \in \mathbb{R}^m$, potential energy function $V(\mathbf{q})$ and symmetric (possibly non-constant) mass matrix $\mathcal{M}(\mathbf{q}) \in \mathbb{R}^{m\times m}$. The corresponding equations of motion can be derived from the Lagrangian functional

$$L[\mathbf{q}] = \int_{t_0}^{t_1} \mathcal{L}(\dot{\mathbf{q}}(t), \mathbf{q}(t))\, dt \tag{1}$$

with Lagrangian density

$$\mathcal{L}(\dot{\mathbf{q}}, \mathbf{q}) = \frac{1}{2}\dot{\mathbf{q}} \cdot [\mathcal{M}(\mathbf{q})\, \dot{\mathbf{q}}] - V(\mathbf{q}). \tag{2}$$

The associated Euler-Lagrange equations are given by

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{q}} = \frac{d}{dt}\left[\mathcal{M}(\mathbf{q})\,\dot{\mathbf{q}}\right] + \nabla_{\mathbf{q}} V(\mathbf{q}) - \frac{1}{2}\nabla_{\mathbf{q}}\left\{\dot{\mathbf{q}} \cdot \left[\mathcal{M}(\mathbf{q})\,\dot{\mathbf{q}}\right]\right\} = 0. \tag{3}$$

To switch to the Hamiltonian formulation, we first introduce the momentum conjugate to $\mathbf{q}$:

$$\mathbf{p} = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} = \mathcal{M}(\mathbf{q})\,\dot{\mathbf{q}}. \tag{4}$$

The resulting Hamiltonian (energy) is

$$\mathcal{H}(\mathbf{q},\mathbf{p}) = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} \cdot \dot{\mathbf{q}} - \mathcal{L} = \frac{1}{2}\dot{\mathbf{q}} \cdot \left[\mathcal{M}(\mathbf{q})\,\dot{\mathbf{q}}\right] + V(\mathbf{q}) = \frac{1}{2}\mathbf{p} \cdot \left[\mathcal{M}(\mathbf{q})^{-1}\,\mathbf{p}\right] + V(\mathbf{q}) \tag{5}$$

with canonical equations of motion

$$\dot{\mathbf{q}} = +\nabla_{\mathbf{p}}\mathcal{H}(\mathbf{q},\mathbf{p}) = \mathcal{M}(\mathbf{q})^{-1}\,\mathbf{p}, \tag{6}$$

$$\dot{\mathbf{p}} = -\nabla_{\mathbf{p}}\mathcal{H}(\mathbf{q},\mathbf{p}) = -\frac{1}{2}\nabla_{\mathbf{q}}\left\{\mathbf{p} \cdot \left[\mathcal{M}(\mathbf{q})^{-1}\mathbf{p}\right]\right\} - \nabla_{\mathbf{q}} V(\mathbf{q}). \tag{7}$$

We now recall that a Markov process will converge to some distribution of configurations if it is constructed out of updates each of which has the desired distribution as a fixed point, and which taken together are ergodic. The generalized hybrid Monte Carlo (GHMC) algorithm for sampling from the canonical ensemble with density function

$$\rho(\mathbf{q},\mathbf{p}) \propto \exp(-\beta\mathcal{H}(\mathbf{q},\mathbf{p})), \tag{8}$$

$\beta = 1/K_B T$, is defined as the concatenation of a molecular dynamics Monte Carlo (MDMC) and a partial momentum refreshment Monte Carlo (PMMC) step [14, 15]. We now describe both steps in more detail.

A. Molecular dynamics Monte Carlo (MDMC)

This step in turn consists of three parts:

(i) *Molecular dynamics* (MD): an approximate integration of Hamilton's equations of motion (6)-(7) with a time-reversible and volume-preserving method $\Psi_{\Delta t}$ over L steps and step-size Δt. We will derive an appropriate numerical time-stepping method in section III.

The resulting time-reversible and volume-preserving map from the initial to the final state is denoted by $U_\tau : (\mathbf{q},\mathbf{p}) \rightarrow (\mathbf{q}',\mathbf{p}')$, $\tau = L\Delta t$. Recall that a map U_τ is called time-reversible if $U_\tau = U_{-\tau}^{-1}$ and volume-preserving if $\det \frac{\partial U_\tau(\mathbf{q},\mathbf{p})}{\partial(\mathbf{q},\mathbf{p})} = 1$.

(ii) A *momentum flip* $\mathcal{F} : (\mathbf{q}, \mathbf{p}) \to (\mathbf{q}, -\mathbf{p})$.

(iii) *Monte Carlo* (MC): a Metropolis accept/reject test

$$(\mathbf{q}', \mathbf{p}') = \begin{cases} \mathcal{F} \cdot U_\tau(\mathbf{q}, \mathbf{p}) & \text{with probability } \min(1, \exp(-\beta\, \delta H)) \\ (\mathbf{q}, \mathbf{p}) & \text{otherwise} \end{cases}, \tag{9}$$

with

$$\delta H := \mathcal{H}(U_\tau(\mathbf{q}, \mathbf{p})) - \mathcal{H}(\mathbf{q}, \mathbf{p}) = \mathcal{H}(\mathcal{F} \cdot U_\tau(\mathbf{q}, \mathbf{p})) - \mathcal{H}(\mathbf{q}, \mathbf{p}) \tag{10}$$

and $\mathcal{H}$ defined by (5)

Molecular dynamics Monte Carlo (MDMC) satisfies detailed balance since $(\mathcal{F} \cdot U_\tau)^2 = \mathrm{id}$ and U_τ is volume conserving.

B. Partial momentum refreshment Monte Carlo (PMMC)

We first apply an extra momentum flip $\mathcal{F}$ so that the trajectory is reversed upon an MDMC rejection (instead of upon an acceptance). The momenta $\mathbf{p}$ are now mixed with a normal (Gaussian) i.i.d. distributed noise vector $\mathbf{u} \in \mathbb{R}^m$ and the complete partial momentum refreshment step is given by

$$\begin{pmatrix} \mathbf{u}' \\ \mathbf{p}' \end{pmatrix} = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \cdot \mathcal{F} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} \tag{11}$$

where

$$\mathbf{u} = \beta^{-1/2} \mathcal{M}(\mathbf{q})^{1/2} \xi, \qquad \xi = (\xi_1, \ldots, \xi_m)^T, \qquad \xi_i \sim \mathrm{N}(0,1),\, i = 1, \ldots, m, \tag{12}$$

and $0 \le \phi \le \pi/2$. Here $\mathrm{N}(0,1)$ denotes the normal distribution with zero mean and unit variance.

If $\mathbf{p}$ and $\mathbf{u}$ are both distributed according to the same normal (Gaussian) distribution, then so are $\mathbf{p}'$ and $\mathbf{u}'$. This special property of Gaussian random variables under an orthogonal transformation (11) makes it possible to conduct the partial momentum refreshment step without a Metropolis accept/reject test. See [15] for details.

C. Special cases of GHMC

Several well-known algorithms are special cases of GHMC:

- The standard hybrid Monte Carlo (HMC) algorithm of Duane, Kennedy, Pendleton and Roweth [1] is the special case where $\phi = \pi/2$. The momentum flips may be ignored in this case since $\mathbf{p}' = \mathbf{u}$ in (11) and the previous value of $\mathbf{p}$ is entirely discarded. According to theoretical results in [15], this choice is optimal for sampling purposes and long MD trajectories. However, one has to keep in mind that the theoretical setting of [15] is unlikely to apply for biomolecular simulations and that a different choice of ϕ could be more appropriate for such simulations.
- The choice $\phi = 0$ corresponds to constant energy molecular dynamics under the assumption that the propagator U_τ conserves energy exactly.
- The Langevin Monte Carlo algorithm of Horowitz [14] corresponds to $L = 1$; i.e., a single MD time-step with $\tau = \Delta t$, and ϕ arbitrary. The single step ($L = 1$) may be replaced by a small number of MD steps and $\tau = L\Delta t$. Langevin Monte Carlo recovers stochastic Langevin molecular dynamics [18]

  $$\dot{\mathbf{q}} = \mathcal{M}^{-1}(\mathbf{q})\,\mathbf{p}, \qquad \dot{\mathbf{p}} = -\frac{1}{2}\nabla_{\mathbf{q}}\left\{\mathbf{p}\cdot\left[\mathcal{M}(\mathbf{q})^{-1}\mathbf{p}\right]\right\} - \nabla_{\mathbf{q}}V(\mathbf{q}) - \gamma\mathbf{p} + \sigma\dot{\mathbf{W}}, \tag{13}$$

  provided $\phi = \sqrt{2\gamma\tau} \ll 1$. Here $\gamma > 0$ is a constant, $\mathbf{W}(t)$ is an m-dimensional Wiener process, and σ is determined by the standard fluctuation-dissipation relation [18]. Indeed, we find that (11) without the momentum flip $\mathcal{F}$ reduces to

  $$\mathbf{p}' \approx (1-\gamma\tau)\,\mathbf{p} + (2\gamma\tau)^{1/2}\mathbf{u} \tag{14}$$

  for $\phi = \sqrt{2\gamma\tau} \ll 1$ and one may view the GHMC algorithm as a mean to perform stochastic molecular dynamics (instead of using GHMC as a pure sampling device).

III. A SYMPLECTIC AND TIME-REVERSIBLE PROPAGATOR

To implement the generalized hybrid Monte Carlo method for Hamiltonian systems of the form (6)-(7), we need to find a time-reversible and volume-preserving approximation to the exact solution flow map. The essential idea is to replace exact time derivatives $\dot{\mathbf{q}}$ in the Lagrangian density (2) by (forward and backward) finite difference approximations

$$\delta_t^+ \mathbf{q}^n = \frac{\mathbf{q}^{n+1} - \mathbf{q}^n}{\Delta t}, \qquad \delta_t^- \mathbf{q}^n = \frac{\mathbf{q}^n - \mathbf{q}^{n-1}}{\Delta t}, \tag{15}$$

and to start from a discrete approximation

$$L_{\Delta t}[\{\mathbf{q^n}\}] = \sum_n \mathcal{L}_{\Delta t}(\delta_t^+ \mathbf{q}^n, \delta_t^- \mathbf{q}^n, \mathbf{q}^n)\, \Delta t \tag{16}$$

to the Lagrangian functional (1) with

$$\mathcal{L}_{\Delta t}(\delta_t^+ \mathbf{q}^n, \delta_t^- \mathbf{q}^n, \mathbf{q}^n) = \frac{1}{4} \left\{ \delta_t^+ \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^n)\, \delta_t^+ \mathbf{q}^n \right] + \delta_t^- \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^n)\, \delta_t^- \mathbf{q}^n \right] \right\} - V(\mathbf{q}^n). \tag{17}$$

Following the discrete variational principle (see, e.g., [8]), we find the associated discrete equations of motion from $\partial L_{\Delta t}/\partial \mathbf{q}^n = 0$ and obtain the generalized leapfrog scheme

$$\begin{aligned} 0 =& \delta_t^+ \left\{ \frac{1}{2} \left[\mathcal{M}(\mathbf{q}^n) + \mathcal{M}(\mathbf{q}^{n-1}) \right] \delta_t^- \mathbf{q}^n \right\} + \nabla_{\mathbf{q}} V(\mathbf{q}^n) \\ & - \frac{1}{4} \nabla_{\mathbf{q}} \left\{ \delta_t^+ \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^n)\, \delta_t^+ \mathbf{q}^n \right] + \delta_t^- \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^n)\, \delta_t^- \mathbf{q}^n \right] \right\}. \end{aligned} \tag{18}$$

This scheme is time-reversible since replacing $\mathbf{q}^{n+1}$ by $\mathbf{q}^{n-1}$ and Δt by $-\Delta t$ leaves the scheme unchanged.

We now convert this scheme into an equivalent (in terms of $\mathbf{q}$-propagation) symplectic one-step method by noting that

$$\sum_n \mathcal{L}_{\Delta t}(\delta_t^+ \mathbf{q}^n, \delta_t^- \mathbf{q}^n, \mathbf{q}^n)\, \Delta t = \sum_n \mathcal{L}_{\Delta t}^{n+1/2} \tag{19}$$

with

$$\mathcal{L}_{\Delta t}^{n+1/2} = \frac{1}{2} \left\{ \delta_t^+ \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^n) + \mathcal{M}(\mathbf{q}^{n+1}) \right] \delta_t^+ \mathbf{q}^n - \left[V(\mathbf{q}^n) + V(\mathbf{q}^{n+1}) \right] \right\} \Delta t. \tag{20}$$

The discrete approximation $\mathcal{L}_{\Delta t}^{n+1/2}$ is now used as a generating function (see, e.g., [8]) to yield a symplectic (and hence volume-preserving) time-stepping method

$$\Psi_{\Delta t} : (\mathbf{q}^n, \mathbf{p}^n) \to (\mathbf{q}^{n+1}, \mathbf{p}^{n+1}) \tag{21}$$

via

$$\begin{aligned} \mathbf{p}^{n+1} =& + \nabla_{\mathbf{q}^{n+1}} \mathcal{L}_{\Delta t}^{n+1/2} \\ =& \frac{1}{2} \left(\mathcal{M}(\mathbf{q}^n) + \mathcal{M}(\mathbf{q}^{n+1}) \right) \delta_t^+ \mathbf{q}^n + \frac{\Delta t}{2} \nabla_{\mathbf{q}} \left\{ \delta_t^+ \mathbf{q}^n \cdot \left[\mathcal{M}(\mathbf{q}^{n+1})\, \delta_t^+ \mathbf{q}^n \right] - V(\mathbf{q}^{n+1}) \right\} \end{aligned} \tag{22}$$

and

$$\begin{aligned}\mathbf{p}^n =& -\nabla_{\mathbf{q}^n}\mathcal{L}_{\Delta t}^{n+1/2} \\ =& \frac{1}{2}\left(\mathcal{M}(\mathbf{q}^n)+\mathcal{M}(\mathbf{q}^{n+1})\right)\delta_t^+\mathbf{q}^n - \frac{\Delta t}{2}\nabla_{\mathbf{q}}\left\{\delta_t^+\mathbf{q}^n\cdot\left[\mathcal{M}(\mathbf{q}^n)\,\delta_t^+\mathbf{q}^n\right]-V(\mathbf{q}^n)\right\}.\end{aligned} \tag{23}$$

Given $(\mathbf{q}^n,\mathbf{p}^n)$, the map $\Psi_{\Delta t}$ is implemented numerically by first solving (23) for $\mathbf{q}^{n+1}$. The new momentum $\mathbf{p}^{n+1}$ is then given explicitly by (22). We finally note that the generating function (20) was first proposed by MacKay in [19] for deriving symplectic methods for systems with general Lagrangian density $L(\dot{\mathbf{q}},\mathbf{q})$.

The generalized Störmer-Verlet method is second-order in time and the average energy fluctuation $\langle\delta\mathcal{H}\rangle$ satisfies

$$\langle\delta\mathcal{H}\rangle = \mathcal{O}(m\Delta t^4), \tag{24}$$

where m is the number of degrees of freedom and $\delta\mathcal{H}$ is given by (10) [3, 20]. Following the analysis of [15, 20], the average Metropolis acceptance rate for the MDMC step is given by

$$P_{\text{acc}} = \text{erfc}\left(\frac{1}{2}\sqrt{\beta\langle\delta\mathcal{H}\rangle}\right) \tag{25}$$

and the acceptance rate deteriorates with increasing system size m.

IV. GENERALIZED SHADOW HYBRID MONTE CARLO (GSHMC) METHOD

The basic idea of the GSHMC method is to implement the GHMC method with respect to an appropriately modified reference energy $\mathcal{H}_{\Delta t}$. This reference energy is chosen such that the acceptance rate (25) in the MDMC part of the GHMC algorithm is increased. This goal can indeed be achieved by making use of backward error analysis and the implied existence of modified energies, which are preserved to high accuracy by the time-stepping method [3, 13]. The price we pay for this increased acceptance rate is that (i) the PMMC step becomes more complex and that (ii) computed samples need to be reweighted after the simulation to become consistent with the desired canonical distribution function (8).

We provide the details of the GSHMC method in several steps. First we describe the MDMC step when implemented with respect to a reference energy $\mathcal{H}_{\Delta t} = \mathcal{H}+\mathcal{O}(\Delta t^p)$, $p \geq 4$. This step is a rather trivial modification of the GHMC method. We then explicitly derive a fourth-order modified energy $\mathcal{H}_{\Delta t}^{[4]}$ for the generalized Störmer-Verlet method of Section III. We finally discuss the necessary modifications to the momentum refreshment Monte Carlo step, which are non-trivial but vital to the success of the GSHMC method.

A. Modified MDMC step

The MDMC step of Section II A remains as before with only (10) replaced by

$$\delta H = \mathcal{H}_{\Delta t}(U_\tau(\mathbf{q},\mathbf{p})) - \mathcal{H}_{\Delta t}(\mathbf{q},\mathbf{p}). \tag{26}$$

In the remaining part of the subsection we derive a fourth-order reference energy $\mathcal{H}_{\Delta t} = \mathcal{H}_{\Delta t}^{[4]}$ for the generalized Störmer-Verlet method of Section III. A generalization to sixth-order and higher can be found in the Appendix.

Given a numerical trajectory $\{\mathbf{q}^i\}_{i=-k}^{L+k}$, we construct for t_n, $n \in \{0, L\}$, the unique interpolation polynomial $\mathbf{Q}(t) \in \mathbb{R}^m$ of order $p \leq 2k$, $k \geq 2$, such that

$$\mathbf{Q}(t_i) = \mathbf{q}^i, \qquad i = n-k, \ldots, n, \ldots, n+k \tag{27}$$

[21]. We then make use of standard Taylor expansion, i.e.

$$\mathbf{q}^{n\pm 1} = \mathbf{Q}(t_n) \pm \Delta t \dot{\mathbf{Q}}(t_n) + \frac{\Delta t^2}{2}\ddot{\mathbf{Q}}(t_n) \pm \frac{\Delta t^3}{6}\mathbf{Q}^{(3)}(t_n) + \cdots, \tag{28}$$

in the discrete Lagrangian density (17) to obtain

$$\begin{aligned}
\mathcal{L}_{\Delta t} =& \frac{1}{4}\left(\dot{\mathbf{Q}} + \frac{\Delta t}{2}\ddot{\mathbf{Q}} + \frac{\Delta t^2}{6}\mathbf{Q}^{(3)}\right)\cdot\left[\mathcal{M}(\mathbf{Q})\left(\dot{\mathbf{Q}} + \frac{\Delta t}{2}\ddot{\mathbf{Q}} + \frac{\Delta t^2}{6}\mathbf{Q}^{(3)}\right)\right] \\
&+ \frac{1}{4}\left(\dot{\mathbf{Q}} - \frac{\Delta t}{2}\ddot{\mathbf{Q}} + \frac{\Delta t^2}{6}\mathbf{Q}^{(3)}\right)\cdot\left[\mathcal{M}(\mathbf{Q})\left(\dot{\mathbf{Q}} - \frac{\Delta t}{2}\ddot{\mathbf{Q}} + \frac{\Delta t^2}{6}\mathbf{Q}^{(3)}\right)\right] - V(\mathbf{Q}) + \mathcal{O}(\Delta t^3) \\
=& \mathcal{L}(\dot{\mathbf{Q}},\mathbf{Q}) + \Delta t^2\, \delta\mathcal{L}^{[4]}(\mathbf{Q}^{(3)}, \ddot{\mathbf{Q}}, \dot{\mathbf{Q}}, \mathbf{Q}) + \mathcal{O}(\Delta t^4)
\end{aligned} \tag{29}$$

with

$$\delta\mathcal{L}^{[4]}(\mathbf{Q}^{(3)}, \ddot{\mathbf{Q}}, \dot{\mathbf{Q}}, \mathbf{Q}) = \frac{1}{24}\left\{3\ddot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\ddot{\mathbf{Q}}\right] + 4\dot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(3)}\right]\right\} \tag{30}$$

and with all quantities involving the interpolation polynomial $\mathbf{Q}(t)$ evaluated at $t = t_n$.

We note that the truncated expansion

$$\mathcal{L}_{\Delta t}^{[4]} = \frac{1}{2}\dot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\dot{\mathbf{Q}}\right] - V(\mathbf{Q}) + \frac{\Delta t^2}{24}\left\{3\ddot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\ddot{\mathbf{Q}}\right] + 4\dot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(3)}\right]\right\} \tag{31}$$

can be viewed as a new (higher-order) Lagrangian density with associated (higher-order) Euler-Lagrange equations. We derive the associated conserved energy according to the formula

$$\mathcal{H}_{\Delta t}^{[4]} = \frac{\partial \mathcal{L}_{\Delta t}^{[4]}}{\partial \dot{\mathbf{Q}}}\cdot\dot{\mathbf{Q}} + \frac{\partial \mathcal{L}_{\Delta t}^{[4]}}{\partial \ddot{\mathbf{Q}}}\cdot\ddot{\mathbf{Q}} - \frac{d}{dt}\frac{\mathcal{L}_{\Delta t}^{[4]}}{\partial \ddot{\mathbf{Q}}}\cdot\dot{\mathbf{Q}} + \frac{\partial \mathcal{L}_{\Delta t}^{[4]}}{\partial \mathbf{Q}^{(3)}}\cdot\mathbf{Q}^{(3)} - \frac{d}{dt}\frac{\partial \mathcal{L}_{\Delta t}^{[4]}}{\partial \mathbf{Q}^{(3)}}\cdot\ddot{\mathbf{Q}} + \frac{d^2}{dt^2}\frac{\partial \mathcal{L}_{\Delta t}^{[4]}}{\partial \mathbf{Q}^{(3)}}\cdot\dot{\mathbf{Q}} - \mathcal{L}_{\Delta t}^{[4]}. \tag{32}$$

An explicit expression is provided by

$$\begin{aligned}\mathcal{H}_{\Delta t}^{[4]} =& \frac{1}{2}\dot{\mathbf{Q}} \cdot \left[\mathcal{M}(\mathbf{Q})\,\dot{\mathbf{Q}}\right] + V(\mathbf{Q}) \\ &+ \frac{\Delta t^2}{24}\left\{4\dot{\mathbf{Q}} \cdot \left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(3)}\right] - 6\dot{\mathbf{Q}} \cdot \frac{d}{dt}\left[\mathcal{M}(\mathbf{Q})\,\ddot{\mathbf{Q}}\right] + 4\dot{\mathbf{Q}} \cdot \frac{d^2}{dt^2}\left[\mathcal{M}(\mathbf{Q})\,\dot{\mathbf{Q}}\right]\right\} \\ &+ \frac{\Delta t^2}{24}\left\{3\ddot{\mathbf{Q}} \cdot \mathcal{M}(\mathbf{Q})\,\ddot{\mathbf{Q}} - 4\ddot{\mathbf{Q}} \cdot \frac{d}{dt}\left[\mathcal{M}(\mathbf{Q})\,\dot{\mathbf{Q}}\right]\right\}. \qquad (33)\end{aligned}$$

It can be shown that $\mathcal{H}_{\Delta t}^{[4]}$ is preserved to fourth-order along trajectories of (23)-(22) and (18), respectively, provided $k = 2$ and $p = 4$ in (27). This procedure can be generalized and we obtain modified energies $\mathcal{H}_{\Delta t}^{[2k]}$ for any $k \geq 2$. See the Appendix for the case $k = 3$. These modified energies $\mathcal{H}_{\Delta t}^{[2k]}$, with an appropriate order $p = 2k \geq 4$, will be used in the GSHMC method as the reference energy function $\mathcal{H}_{\Delta t}$.

It should be noted that the thus constructed value of a modified energy $\mathcal{H}_{\Delta t}$ at time $t = t_n$ depends only on the positions $\mathbf{q}^n$ and the momenta $\mathbf{p}^n$ at $t = t_n$. This follows from the uniqueness of the numerical trajectory $\{\mathbf{q}^i\}_{i=n-k}^{n+k}$ and, hence, of the interpolation polynomial $\mathbf{Q}(t)$ on a given pair $(\mathbf{q}^n, \mathbf{p}^n)$.

Using modified energies, the estimate (24) gets replaced by

$$\langle \delta\mathcal{H} \rangle = \mathcal{O}(m\,\Delta t^{4k}), \qquad (34)$$

with $\delta\mathcal{H}$ now being given by (26) and $\mathcal{H}_{\Delta t} = \mathcal{H}_{\Delta t}^{[2k]}$. Hence an increase in system size m can be counterbalanced by an increase in the order $p = 2k$ of the modified energy to keep the product of m and Δt^{4k} roughly constant. In other words, modified energies offer a rather inexpensive way to increase the acceptance rate (25) of the MDMC step.

B. Modified PMMC step

The original TSHMC method has been based on a simple momentum proposal step of the form

$$\mathbf{p}' = \mathbf{p} + \sigma\,\mathbf{u}. \qquad (35)$$

Here $\sigma > 0$ is a free parameter and $\mathbf{u}$ is defined by (12). Smaller values of σ lead to smaller perturbations in the momenta. The new set of momenta $\mathbf{p}'$ is accepted/rejected according to an appropriate Metropolis criterion [13].

It has been found that increased values of σ lead to an increased rejection rate. In this section, a modified momentum update is proposed for GSHMC to reduce this undesirable increase in the rejection rate. This modification is indeed found to significantly improves the efficiency of GSHMC as a sampling tool.

The idea of the modification is to combine the GHMC momentum update (11) with the fact that in GSHMC one samples with respect to a modified energy function $\mathcal{H}_{\Delta t}$. This idea can be realized by implementing the PMMC step of Section II B as a Markov chain Monte Carlo step with respect to the reference energy $\mathcal{H}_{\Delta t}$. Specifically, we define $\mathbf{u}$ as in (12) and propose a new set of momenta $\mathbf{p}'$ and auxiliary variables $\mathbf{u}'$ by (11). The set of momenta $\mathbf{p}'$ and the vector $\mathbf{u}'$ are accepted according to the Metropolis test

$$(\mathbf{u}', \mathbf{p}') = \begin{cases} \left[R(\phi)(\mathbf{u}, \mathbf{p})^T\right]^T & \text{with probability } P(\mathbf{q}, \mathbf{p}, \mathbf{u}, \mathbf{p}', \mathbf{u}') \\ (\mathbf{u}, \mathbf{p}) & \text{otherwise} \end{cases}, \tag{36}$$

where

$$P(\mathbf{q}, \mathbf{p}, \mathbf{u}, \mathbf{p}', \mathbf{u}') = \min\left(1, \frac{\exp\left(-\beta\left[\mathcal{H}_{\Delta t}(\mathbf{q}, \mathbf{p}') + \frac{1}{2}(\mathbf{u}')^T\mathcal{M}(\mathbf{q})^{-1}\mathbf{u}'\right]\right)}{\exp\left(-\beta\left[\mathcal{H}_{\Delta t}(\mathbf{q}, \mathbf{p}) + \frac{1}{2}\mathbf{u}^T\mathcal{M}(\mathbf{q})^{-1}\mathbf{u}\right]\right)}\right) \tag{37}$$

and

$$R(\phi) = \begin{bmatrix} \cos(\phi) & \sin(\phi) \\ \sin(\phi) & -\cos(\phi) \end{bmatrix}. \tag{38}$$

It should be noted that the updated variable $\mathbf{u}'$ is entirely discarded after each momentum refreshment step and is replaced by a new set of random variables (12). The Monte Carlo step is therefore best understood by interpreting the update as a 'classical' hybrid Monte Carlo method with $\mathbf{u}$ taking the role of 'momentum' and $\mathbf{p}$ the role of 'positions'. Note that the 'real' positions $\mathbf{q}$ are not changed. Note furthermore that (11) is a linear map from $(\mathbf{p}, \mathbf{u})$ to $(\mathbf{p}', \mathbf{u}')$. This map is characterized by the 2×2 matrix (38). Since $\det(R) = -1$ and $R^2 = I$, the proposal step (11) satisfies detailed balance. Hence (12) and (11) together with (36) sample from a canonical distribution with density function

$$\rho_{\text{ext}}(\mathbf{q}, \mathbf{p}, \mathbf{u}) \propto \exp\left(-\beta\left[\mathcal{H}_{\Delta t}(\mathbf{q}, \mathbf{p}) + \frac{1}{2}\mathbf{u}^T\mathcal{M}(\mathbf{q})^{-1}\mathbf{u}\right]\right). \tag{39}$$

The angle ϕ in (38) is chosen such that the rejection rate in the momentum refreshment step is below 10%. A much higher rejection rate would imply that the system gets 'thermalized' too infrequently. A fixed rejection rate implies that larger systems require a smaller value of ϕ, which seems acceptable once we take into account that large NVE simulations behave almost like an NVT ensemble.

To further decrease the rejection rate one can repeat the refreshment step before continuing with the molecular dynamics part of GSHMC. Hence the complete GSHMC cycle consists then of a molecular dynamics Monte Carlo step, a momentum flip, a Monte Carlo momentum refreshment step, followed by another Monte Carlo momentum refreshment step. In other words, GSHMC becomes the concatenation of four Markov processes (here we counted the momentum flip as an independent Markov process) with identical invariant distribution functions (here the canonical distribution with respect to a modified Hamiltonian $\mathcal{H}_{\Delta t}$). Of course, this approach can be further modified by additional (relatively inexpensive) momentum update steps.

Inspired by the work of Sweet et al. [12], we finally mention an additional strategy for increasing the acceptance rate of the PMMC step. We replace (11) by

$$\begin{pmatrix} \mathbf{u}' \\ \bar{\mathbf{p}}' \end{pmatrix} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ \sin(\phi) & -\cos(\phi) \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \bar{\mathbf{p}} \end{pmatrix}, \tag{40}$$

where $\bar{\mathbf{p}}'$ is defined through an appropriate change of variables $\bar{\mathbf{p}} = \psi(\mathbf{q}, \mathbf{p}, \Delta t)$. It is assumed that the map ψ is invertible in the momentum vector $\mathbf{p}$. The new momentum vector $\mathbf{p}'$, implicitly defined by $\bar{\mathbf{p}}' = \psi(\mathbf{q}, \mathbf{p}', \Delta t)$, is then accepted with probability (37).

See [12] for an appropriate choice of ψ in case of a constant mass matrix. More specifically, given $(\mathbf{q}, \mathbf{p})$, we perform a single time step forward and backward in time. The results are denoted by $(\mathbf{q}^+, \mathbf{p}^+)$ and $(\mathbf{q}^-, \mathbf{p}^-)$, respectively. We define

$$\bar{\mathbf{p}} = \psi(\mathbf{q}, \mathbf{p}, \Delta t) := \mathbf{p} - \frac{\Delta t}{24}\left(\nabla_{\mathbf{q}} V(\mathbf{q}^+) - \nabla_{\mathbf{q}} V(\mathbf{q}^-)\right)\right). \tag{41}$$

Note that, contrary to the S2HMC method [12], the modified PMMC step (40)-(41) can be used together with any choice of the reference Hamiltonian $\mathcal{H}_{\Delta t}$ in (37) and also for systems with non-constant mass matrix.

C. Reweighting

Given an observable $\Omega(\mathbf{q}, \mathbf{p})$ and its values Ω_i, $i = 1, \ldots, K$, along a sequence of states $(\mathbf{q}_i, \mathbf{p}_i)$, $i = 1, \ldots, K$, computed by the GSHMC method, we need to reweight Ω_i to compute averages $\langle\Omega\rangle_K$ according to the desired canonical distribution (8). In particular, one needs to apply the formula

$$\langle\Omega\rangle_K = \frac{\sum_{i=1}^{K} w_i \, \Omega_i}{\sum_{i=1}^{K} w_i} \tag{42}$$

with

$$w_i = \exp(-\beta\{\mathcal{H}(\mathbf{q}_i, \mathbf{p}_i)) - \mathcal{H}_{\Delta t}(\mathbf{q}_i, \mathbf{p}_i)\}). \tag{43}$$

V. APPLICATIONS

A. Constant temperature and volume (NVT) GSHMC

The starting point of any (classical) molecular simulation is a system of N particles, which interact through both long and short range forces via Newton's second law. We write the equations of motion in the form

$$\dot{\mathbf{r}} = M^{-1}\mathbf{p}_r, \qquad \dot{\mathbf{p}}_r = -\nabla_{\mathbf{r}} V(\mathbf{r}), \tag{44}$$

where $\mathbf{r} \in \mathbb{R}^{3N}$ is the vector of atomic positions, $\mathbf{p}_r \in \mathbb{R}^{3N}$ the associated momentum vector, $M \in \mathbb{R}^{3N\times 3N}$ is the (constant) symmetric mass matrix and $V : \mathbb{R}^{3N} \to \mathbb{R}$ is the empirical potential energy function. The equations of motion (44) are equivalent to the Euler-Lagrange equations

$$M\,\ddot{\mathbf{r}} + \nabla_{\mathbf{r}} V(\mathbf{r}) = 0 \tag{45}$$

for the Lagrangian density

$$\mathcal{L} = \frac{1}{2}\dot{\mathbf{r}} \cdot [M\,\dot{\mathbf{r}}] - V(\mathbf{r}). \tag{46}$$

We find that (46) fits into the general form (2) with constant mass matrix $\mathcal{M}(\mathbf{q}) = M$, $\mathbf{q} = \mathbf{r}$, and $m = 3N$.

Because the mass matrix M is now constant, the symplectic time-stepping method $\Psi_{\Delta t}$, defined by (22)-(23) becomes equivalent to the standard Störmer-Verlet method (see, e.g., [7, 8])

$$\mathbf{p}_r^{n+1/2} = \mathbf{p}_r^n - \frac{\Delta t}{2}\nabla_{\mathbf{r}} V(\mathbf{r}^n), \tag{47}$$

$$\mathbf{r}^{n+1} = \mathbf{r}^n + M^{-1}\mathbf{p}_r^{n+1/2}, \tag{48}$$

$$\mathbf{p}_r^{n+1} = \mathbf{p}_r^{n+1/2} - \frac{\Delta t}{2}\nabla_{\mathbf{r}} V(\mathbf{r}^{n+1}), \tag{49}$$

and the expression for the modified energy $\mathcal{H}_{\Delta t}^{[4]}$ reduces to

$$\mathcal{H}_{\Delta t}^{[4]} = \frac{1}{2}\dot{\mathbf{R}} \cdot \left[M\, \dot{\mathbf{R}} \right] + V(\mathbf{R}) + \frac{\Delta t^2}{24} \left\{ 2\dot{\mathbf{R}} \cdot \left[M\, \mathbf{R}^{(3)} \right] - \ddot{\mathbf{R}} \cdot M\, \ddot{\mathbf{R}} \right\}, \tag{50}$$

where $\mathbf{R}(t)$ denotes now the interpolating polynomial and replaces $\mathbf{Q}(t)$ in (33).

The application of the GSHMC method, as described in Section IV, is now straightforward. Numerical results will be presented in Section VII.

We finally note that the equations of motion (45) subject to holonomic constraints (such as bond stretching and bending constraints) can be treated numerically by the SHAKE extension [22] of the standard Störmer-Verlet/leapfrog method. The associated modified energies remain unaffected by that extension and the fourth-order modified energy, in particular, is still provided by the expression (50).

B. Constant temperature and pressure (NPT) GSHMC

We first summarize the constant energy and pressure formulation of Andersen [16]. We then discuss a symplectic and time-reversible integration method and derive its fourth-order modified energy. This provides the essential building block to extend the GSHMC method to molecular simulations in an NPT ensemble.

1. *Constant pressure molecular dynamics*

Given a classical molecular system described by (44), the constant pressure and energy (NPE) formulation of Andersen is derived as follows. The coordinate vector $\mathbf{r} \in \mathbb{R}^{3N}$ in (45) is replaced by a scaled vector $\mathbf{d} \in \mathbb{R}^{3N}$ defined by

$$\mathbf{d} = \mathbf{r}/\mathcal{V}^{1/3} \tag{51}$$

where $\mathcal{V}$ is the volume of the simulation box. Consider now the extended Lagrangian density

$$\mathcal{L}(\dot{\mathbf{d}}, \dot{q}, \mathbf{d}, q) = \left\{ \frac{1}{2} q^{2/3}\, \dot{\mathbf{d}} \cdot \left[M\, \dot{\mathbf{d}} \right] - V(q^{1/3}\mathbf{d}) + \frac{\mu}{2}\dot{q}^2 - \alpha q \right\}. \tag{52}$$

We interpret q as the (dynamic) value of the volume $\mathcal{V}$ and call this additional degree of freedom the ‘piston’ degree of freedom. The constant α corresponds to the external pressure acting on the system and $\mu > 0$ is the mass of the ‘piston’.

Upon defining $\mathbf{q} = (\mathbf{d}^T, q)^T \in \mathbb{R}^m$, $m = 3N + 1$, we find that (52) fits into the general form (2) with non-constant mass matrix

$$\mathcal{M}(\mathbf{q}) = \begin{bmatrix} q^{2/3}M & 0 \\ 0 & \mu \end{bmatrix}. \tag{53}$$

The associated NPE equations of motion are now easily derived using (3). See also Andersen's original publication [16]. The conserved energy $\mathcal{H}$ can be derived from the Lagrangian density (52) according to the standard formula (5), i.e.,

$$\begin{aligned}\mathcal{H} =&\dot{\mathbf{d}} \cdot \nabla_{\dot{\mathbf{d}}}\mathcal{L} + \dot{q}\,\nabla_{\dot{q}}\mathcal{L} - \mathcal{L}\\ =&\frac{1}{2}q^{2/3}\dot{\mathbf{d}} \cdot \left[M\,\dot{\mathbf{d}}\right] + \frac{\mu}{2}\dot{q}^2 + V(q^{1/3}\mathbf{d}) + \alpha q\\ =&\frac{1}{2}q^{-2/3}\mathbf{p}_d \cdot \left[M^{-1}\mathbf{p}_d\right] + \frac{1}{2\mu}p^2 + V(q^{1/3}\mathbf{d}) + \alpha q\\ =&\frac{1}{2}\mathbf{p}_r \cdot \left[M^{-1}\mathbf{p}_r\right] + V(\mathbf{r}) + \frac{1}{2\mu}p^2 + \alpha q,\end{aligned} \tag{54}$$

where

$$\mathbf{p}_d = q^{2/3}M\dot{\mathbf{d}}, \qquad p = \mu\dot{q} \tag{55}$$

are the conjugate momenta in the NPE formulation and $\mathbf{p}_r = M\,\dot{\mathbf{r}} = \mathbf{p}_d/q^{1/3}$ is the classical momentum vector of the NVE formulation (44).

2. A time-reversible and symplectic implementation

We use the previously developed discrete variational principle to derive a symplectic time-stepping method and obtain the generalized leapfrog method

$$\delta_t^+ \left\{\frac{1}{2}\left[(q^n)^{2/3} + (q^{n-1})^{2/3}\right] M\,\delta_t^-\mathbf{d}^n\right\} = -\nabla_{\mathbf{d}}V((q^n)^{1/3}\mathbf{d}^n) \tag{56}$$

and

$$\mu\,\delta_t^+\delta_t^-q^n = \frac{(q^n)^{-1/3}}{6}\left\{\delta_t^+\mathbf{d}^n \cdot \left[M\,\delta_t^+\mathbf{d}^n\right] + \delta_t^-\mathbf{d}^n \cdot \left[M\,\delta_t^-\mathbf{d}^n\right]\right\} - \alpha - \nabla_q V((q^n)^{1/3}\mathbf{d}^n). \tag{57}$$

The equivalent generalized Störmer-Verlet formulation is defined as follows. Given $(\mathbf{d}^n, q^n, \mathbf{p}_d^n, p^n)$, we first find $\mathbf{d}^{n+1}$ and q^{n+1} from the equations

$$\mathbf{p}_d^n = \frac{1}{2}\left[(q^{n+1})^{2/3} + (q^n)^{2/3}\right] M\left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t}\right) + \frac{\Delta t}{2}\nabla_{\mathbf{d}}V((q^{1/3})^n\mathbf{d}^n) \tag{58}$$

and

$$\begin{aligned} p^n =&\mu \left(\frac{q^{n+1} - q^n}{\Delta t} \right) - \frac{\Delta t}{6} (q^n)^{-1/3} \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \cdot \left[M \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \right] \\ &+ \frac{\Delta t}{2} \left[\nabla_q V((q^n)^{1/3} \mathbf{d}^n) + \alpha \right]. \end{aligned} \tag{59}$$

The values for $\mathbf{p}_d^{n+1}$ and p^{n+1} are explicitly given by

$$\mathbf{p}_d^{n+1} = \frac{1}{2} \left[(q^{n+1})^{2/3} + (q^n)^{2/3} \right] M \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) - \frac{\Delta t}{2} \nabla_{\mathbf{d}} V((q^{n+1})^{1/3} \mathbf{d}^{n+1}) \tag{60}$$

and

$$\begin{aligned} p^{n+1} =&\mu \left(\frac{q^{n+1} - q^n}{\Delta t} \right) + \frac{\Delta t}{6} (q^{n+1})^{-1/3} \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \cdot \left[M \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \right] \\ &- \frac{\Delta t}{2} \left[\nabla_q V((q^{n+1})^{1/3} \mathbf{d}^{n+1}) + \alpha \right]. \end{aligned} \tag{61}$$

This completes one time step.

The time-reversible and symplectic method (58)-(61) allows for the implementation of a hybrid Monte Carlo methods as proposed in [2]) and described in more detail in [23]. We now derive a fourth-order modified energy for the GSHMC method.

Let $Q(t)$ and $\mathbf{D}(t)$ denote the interpolation polynomials along numerical trajectories $\{q^n\}$ and $\{\mathbf{d}^n\}$, respectively. Then the associated fourth-order modified energy, defined by (33), is given by

$$\begin{aligned} \mathcal{H}_{\Delta t}^{[4]} =&\mathcal{H} + \frac{\Delta t^2}{24} \left[2\mu \dot{Q} Q^{(3)} - \mu \ddot{Q}^2 \right] \\ &+ \frac{\Delta t^2}{24} \left\{ 4\dot{\mathbf{D}} \cdot \left[Q^{2/3} M \, \mathbf{D}^{(3)} \right] - 6\dot{\mathbf{D}} \frac{d}{dt} \left[Q^{2/3} M \, \ddot{\mathbf{D}} \right] + 4\dot{\mathbf{D}} \cdot \frac{d^2}{dt^2} \left[Q^{2/3} M \, \dot{\mathbf{D}} \right] \right\} \\ &+ \frac{\Delta t^2}{24} \left\{ 3\ddot{\mathbf{D}} \cdot \left[Q^{2/3} M \, \ddot{\mathbf{D}} \right] - 4\ddot{\mathbf{D}} \cdot \frac{d}{dt} \left[Q^{2/3} M \, \dot{\mathbf{D}} \right] \right\} \\ =&\mathcal{H} + \frac{\Delta t^2}{24} \left\{ 2\mu \dot{Q} Q^{(3)} - \mu \ddot{Q}^2 + 2Q^{2/3} \dot{\mathbf{D}} \cdot \left[M \, \mathbf{D}^{(3)} \right] - Q^{2/3} \ddot{\mathbf{D}} \cdot \left[M \, \ddot{\mathbf{D}} \right] \right\} \\ &+ \frac{\Delta t^2}{12} \left\{ \left(\frac{4\ddot{Q}}{3Q^{1/3}} - \frac{4\dot{Q}^2}{9Q^{4/3}} \right) \dot{\mathbf{D}} \cdot \left[M \, \dot{\mathbf{D}} \right] - \frac{2}{3Q^{1/3}} \dot{Q} \dot{\mathbf{D}} \cdot \left[M \, \ddot{\mathbf{D}} \right] \right\} \end{aligned} \tag{62}$$

with $\mathcal{H}$ given by (54).

3. A modified PMMC step

The one-step formulation (58)-(59) together with (60)-(61) will be used in the GSHMC method. After each completed NPE molecular dynamics sub-step, we refresh the momenta $\mathbf{p}_d$ and p as described in Section IV.

Following the Langevin piston method of Feller et al. [17], one can also apply the following simplified momentum update. We always keep the particle momentum $\mathbf{p}_d$ and only refresh the "piston" momentum p, i.e., we replace (11) by

$$\mathbf{u}_d' = \mathbf{u}_d, \tag{63}$$

$$\mathbf{p}_d' = -\mathbf{p}_d, \tag{64}$$

$$u' = \sin(\phi)\, p + \cos(\phi)\, u \tag{65}$$

$$p' = -\cos(\phi)\, p + \sin(\phi)\, u, \tag{66}$$

with

$$u = \beta^{-1}\mu^{1/2}\xi, \qquad \xi \sim \mathrm{N}(0,1). \tag{67}$$

The probability (37) is replaced by

$$P(\mathbf{d}, q, \mathbf{p}_d, p, u, p', u') = \min\left(1, \frac{\exp\left(-\beta\left[\mathcal{H}_{\Delta t}(\mathbf{d}, q, \mathbf{p}_d, p') + \frac{1}{2\mu}(u')^2\right]\right)}{\exp\left(-\beta\left[\mathcal{H}_{\Delta t}(\mathbf{d}, q, \mathbf{p}_d, p) + \frac{1}{2\mu}u^2\right]\right)}\right), \tag{68}$$

where $\mathcal{H}_{\Delta t}$ is an appropriate modified energy, e.g., $\mathcal{H}_{\Delta t} = \mathcal{H}_{\Delta t}^{[4]}$ with $\mathcal{H}_{\Delta t}^{[4]}$ given by (62).

Given a collision frequency γ for the Langevin piston method [17], we choose ϕ and $\tau = L\Delta t$ such that $\phi = \sqrt{2\gamma\tau} \ll 1$ and the resulting GSHMC method can be viewed as a rigorous implementation of the Langevin piston method in the sense of section II C under the assumption of ergodicity of the induced Markov process. Note that, on the contrary, the Langevin piston method combined with the Brunger, Brooks, Karplus (BBK) time-stepping algorithm [24] leads to statistical errors proportional to Δt^2. In particular, one needs to require that $\gamma\Delta t$ is small.

VI. ALGORITHMIC SUMMARY OF THE GSHMC METHOD

We summarize the algorithmic implementation of the GSHMC method for the fourth-order modified energy (33) as follows:

A. MDMC step of GSHMC

Given an accepted MC sample with generalized position vector $\mathbf{q}$ and momentum vector $\mathbf{p}$, we determine the associated modified energy $\mathcal{H}_{\Delta t}^{[4]}(\mathbf{q}, \mathbf{p})$ by integrating the equations of motion two steps forward and backward in time using (22)-(23) in order to construct the required interpolation polynomial $\mathbf{Q}(t)$ as defined in section IV A.

The equations of motion are then solved forward in time over L time steps using the symplectic and time-reversible method (22)-(23). Denote the result by $(\mathbf{q}', \mathbf{p}')$.

An additional two time steps are performed to evaluate the associated modified energy $\mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}', \mathbf{p}')$ and the proposal step $(\mathbf{q}', \mathbf{p}')$ is accepted with probability

$$\min\left(1, \exp(-\beta\{\mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}', \mathbf{p}') - \mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}, \mathbf{p})\}\right). \tag{69}$$

In case of rejection, we continue with $(\mathbf{q}', \mathbf{p}') = (\mathbf{q}, -\mathbf{p})$.

B. PMMC step of GSHMC

Using a change of variables as, for example, defined by (41), we first compute $\bar{\mathbf{p}}' = \psi(\mathbf{q}', \mathbf{p}', \Delta t)$. The momentum vector $\bar{\mathbf{p}}'$ is now mixed with a noise vector $\mathbf{u}$ distributed according to (12). We formally set $\mathbf{q}'' = \mathbf{q}'$ and define

$$\begin{pmatrix} \mathbf{u}' \\ \bar{\mathbf{p}}'' \end{pmatrix} = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \bar{\mathbf{p}}' \end{pmatrix}. \tag{70}$$

The proposal momentum vector $\mathbf{p}''$, implicitely defined by $\bar{\mathbf{p}}'' = \psi(\mathbf{q}'', \mathbf{p}'', \Delta t)$, is accepted with probability

$$\min\left(1, \frac{\exp\left(-\beta\left[\mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}'', \mathbf{p}'') + \frac{1}{2}(\mathbf{u}')^T \mathcal{M}(\mathbf{q}'')^{-1}\mathbf{u}'\right]\right)}{\exp\left(-\beta\left[\mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}', \mathbf{p}') + \frac{1}{2}\mathbf{u}^T \mathcal{M}(\mathbf{q}')^{-1}\mathbf{u}\right]\right)}\right), \tag{71}$$

where two time steps forward and backwards need to be performed in order to evaluate $\mathcal{H}^{[4]}_{\Delta t}(\mathbf{q}'', \mathbf{p}'')$. In case of rejection, we continue with $(\mathbf{q}'', \mathbf{p}'') = (\mathbf{q}', \mathbf{p}')$.

A single GSHMC step is now completed. We store the accepted MC sample as $(\mathbf{q}_{i+1}, \mathbf{p}_{i+1}) = (\mathbf{q}'', \mathbf{p}'')$ and evaluate the associated weight factor w_{i+1} using (43).

C. Comments

We summarize here a few general comments on the GSHMC method.

(i) Note that different angles ϕ can be assigned to different components of $\mathbf{u}$ and $\bar{\mathbf{p}}'$ in (70). This freedom has been used in section V B 3.

(ii) Note also that the summmary of the GSHMC method has been formulated such that the number of necessary momentum flips is minimized. This is in contrast to the (entirely equivalent) presentation used so far, which has been based on the detailed balance requirement.

(iii) The number of additional force evaluations for GSHMC with $\bar{\mathbf{p}} = \mathbf{p}$ over standard HMC amounts to $p - 2$, where p is the order of the modified energy. For example, GSHMC with (33) requires two additional force evaluations per complete Monte Carlo step.

The change of variables (41) requires additional force evaluations [12].

(iv) The time step Δt and the angle ϕ should be chosen such that the probability of having both the MDMC as well as the PMMC step being simultaneously rejected is less than 1%. This is because we obtain $\mathbf{q}_{i+1} = \mathbf{q}_i$ and $\mathbf{p}_{i+1} = -\mathbf{p}_i$ in such a case, which leads to the undesired *Zitterbewegung* in the MC samples.

This requires, in general, a decrease of ϕ in (70) as the system size, $d = 3N$, increases. Furthermore, the discussion in [16] on a dynamically consistent collision frequency γ for a small volume of liquid surrounded by a much larger volume suggests that $\phi \propto \gamma^{1/2} \propto 1/N^{1/3}$, where N is the number of atoms.

(v) In case the PMMC step is performed with a change of variables as defined, for example, by (41), we refer to the resulting method as the GS2HMC method (in analogy to the S2HMC method of [12]).

In case of $\bar{\mathbf{p}} = \mathbf{p}$, we continue using the acronym GSHMC.

VII. NUMERICAL RESULTS

In this section, we perform three sets of experiments. The first set is based on an NVT simulation of argon and assesses rejection rates for several MC methods in the context of sampling. The second set of experiments is based on an NPT simulation of argon. Here we compare the GSHMC algorithm and the Langevin piston method of Feller et al. [17] and assess the performance of GSHMC in the context of stochastic dynamics simulations. We finally implement GSHMC for a larger biomolecular system, the bacteriophage T4 lysozyme protein, and compare the sampling efficiency of GSHMC to constant temperature MD using the Berendsen thermostat [25].

A. Argon

We perform simulations for argon in a periodic box under an NVT and NPT, respectively, ensemble. We now present numerical results for both ensembles. We begin with the NVT simulations.

1. NVT simulations

We perform NVT simulations at a temperature of $T = 120$ K using the following two settings:

(A) $N = 5^3$, $L = 20.1$ Å,

(B) $N = 8^3$, $L = 31.96$ Å.

We implement the GSHMC method with three values of the angle ϕ ($\pi/2$, $\pi/4$, $\pi/8$) in the PMMC step. We also implement the GSHMC method with the modified momentum refreshment step, as defined by (41), with $\phi = \pi/2$. We refer to this implementation as GS2HMC.

Results are compared to implementations of the standard HMC method and the newly proposed S2HMC method of [12].

All Monte Carlo (MC) implementations use $\tau = L\,\Delta t = 2.17$ ps and generate a total of $K = 10^4$ Monte Carlo samples to compute expectation values according to (42). Simulations are performed for four different values of Δt ($\tau/50 \approx 43.4$ fs, $\tau/75 \approx 28.9$ fs, $\tau/100 \approx 21.7$ fs, $\tau/200 \approx 10.9$ fs).

We state rejection rates for the MDMC step and the PMMC step (where applicable) in table I for setting A and in table II for setting B, respectively. We observe an increase in rejection rates for all methods for increasing system size d and step-size Δt. The acceptance rate for the MDMC step is similar for all GSHMC and S2HMC implementations and is consistently better than the corresponding rate of standard HMC. The acceptance rate of PMMC step in GSHMC improves with smaller values of ϕ. The GS2HMC method almost

MDMC/PMMC rejections	$\Delta t \approx 43.4$ fs	$\Delta t \approx 28.9$ fs	$\Delta t \approx 21.7$ fs	$\Delta t \approx 10.9$ fs
GSHMC method, $\phi = \pi/2$	20% / 23%	2% / 12%	<1% / 6%	<1% / 2%
GSHMC method, $\phi = \pi/4$	22% / 17%	2% / 8%	<1% / 4%	<1% / 1%
GSHMC method, $\phi = \pi/8$	21% / 9%	2% / 5%	<1% / 2%	<1% / <1%
GS2HMC method, $\phi = \pi/2$	19% / <1%	2% / <1%	<1% / <1%	<1% / <1%
S2HMC method	20% / NA	1% / NA	<1% / NA	<1% / NA
HMC method	22% / NA	9% / NA	6% / NA	2% / NA

TABLE I: Rejection rates for MDMC and PMMC steps, respectively, for all tested methods under the experimental setting A.

MDMC/PMMC rejections	$\Delta t \approx 43.4$ fs	$\Delta t \approx 28.9$ fs	$\Delta t \approx 21.7$ fs	$\Delta t \approx 10.9$ fs
GSHMC method, $\phi = \pi/2$	33% / 37%	3% / 19%	<1% / 10%	<1% / 3%
GSHMC method, $\phi = \pi/4$	33% / 27%	3% / 12%	<1% / 7%	<1% / 3%
GSHMC method, $\phi = \pi/8$	32% / 15 %	3% / 7%	<1% / 4%	<1% / 1%
GS2HMC method, $\phi = \pi/2$	32% / <1%	3% / <1%	<1% / <1%	<1% / <1%
S2HMC method	33% / NA	2% / NA	<1% / NA	<1% / NA
HMC method	99% / NA	15% / NA	10% / NA	3% / NA

TABLE II: Rejection rates for MDMC and PMMC steps, respectively, for all tested methods under the experimental setting B.

reaches the perfect behavior of S2HMC and HMC in terms of momentum resampling. One should note, however, that the transformation step (41) requires additional force evaluations.

We also give expectation values of total energy, E, diffusion constant,

$$D = \frac{1}{6Nt} \|\mathbf{r}(t) - \mathbf{r}(0)\|^2, \tag{72}$$

and pressure, P, as well as their standard deviation range (corresponding to the 95% confidence interval of normally distributed data) for the experimental setting A and $\Delta t = \tau/75 \approx 28.9$ fs in table III. All methods lead to comparable results in terms of total energy, E, implying that all methods correctly sample from the canonical ensemble. More remarkably, the diffusion constant, D, increases significantly for smaller values of ϕ in the PMMC step of

	energy E [120 k_b K]	diffusion D [Å^2 ps^{-1}]	pressure P [kN/cm^2]
GSHMC method, $\phi = \pi/2$	-442.6 ± 33.6	0.2873 ± 0.0564	0.5904 ± 0.7302
GSHMC method, $\phi = \pi/4$	-442.7 ± 32.8	0.4782 ± 0.1275	0.5881 ± 0.7204
GSHMC method, $\phi = \pi/8$	-442.0 ± 31.2	0.7742 ± 0.1465	0.5958 ± 0.7049
GS2HMC method, $\phi = \pi/2$	-441.0 ± 33.2	0.2927 ± 0.0205	0.6515 ± 0.7317
S2HMC method	-441.9 ± 32.6	0.2877 ± 0.0668	0.6630 ± 0.7266
HMC method	-438.0 ± 33.8	0.2691 ± 0.0219	0.6571 ± 0.7344

TABLE III: Expectation values and their standard deviation range for total energy, E, diffusion constant, D, and pressure, P, from numerical experiments using setting A and $\Delta t \approx 28.9$ fs.

GSHMC. This confirms the fact that HMC methods influence the dynamical properties of a molecular system. Pressure, P, fluctuates largely for all methods, which is not unexpected for a small molecular system such as that of setting A.

2. NPT simulations

We now simulate $N = 125$ argon atoms at constant temperature $T = 120$ K and constant pressure $P = 0.65 \cdot 10^7$ N m^{-2}.

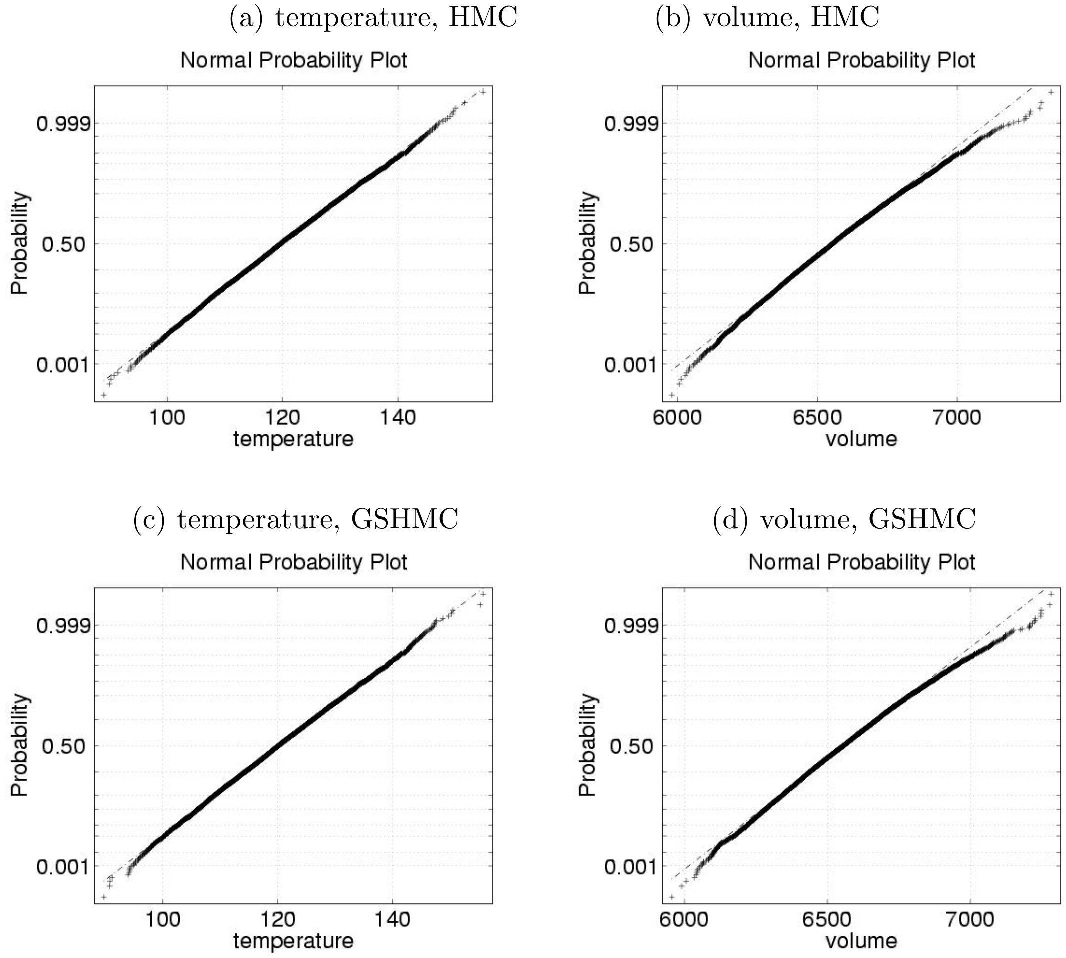
We implement a standard constant pressure and temperture HMC algorithm (see, e.g., [23]) and compare the results to the corresponding GSHMC implementation of section V B with $\phi = \pi/2$.

The simulation parameters are as follows. Both methods are implemented with a step-size of $\Delta t = 10.9$ fs, samples are taken at in intervals of $\tau = L\,\Delta t = 2.17$ ps, i.e., $L = 200$, and the total number of samples is $K = 10^4$. The mass of the piston degree of freedom is set equal to $\mu = 6$, and $\alpha = 0.65 \cdot 10^7$ N m^{-2}.

We compare pressure, P, temperature, T, and total energy, E. Mean values and their standard deviation range can be found in table IV. We also verify that the volume and temperature fluctuations are Gaussian distributed. We display the results for the GSHMC and HMC method in figure 1. Both methods lead to very similar distributions. The temperature distribution is almost ideal while the volume fluctuations display some non-Gaussian behavior in the tails. The effect can be attributed to the finite size of the sample.

	pressure [$\times 10^7$ N m^{-2}]	temperature [K]	energy [120 k_B K]
GSHMC method $\phi = \pi/2$	0.6492 ± 0.8450	120 ± 17	-330 ± 49
HMC method	0.6342 ± 0.8404	120 ± 17	-331 ± 49

TABLE IV: Mean values and their standard deviation range for pressure, P, temperature, T, and total energy, E, for GSHMC and HMC implementation of Andersen's constant pressure formulation.



(a) temperature, HMC

(b) volume, HMC

(c) temperature, GSHMC

(d) volume, GSHMC

FIG. 1: Normal probability plots for volume and temperature fluctuations from HMC and GSHMC implementation of Andersen's constant pressure formulation. Straight lines indicate a Gaussian distribution of data.

	pressure [$\times 10^7$ N m^{-2}]	temperature [K]	energy [120 k_B K]
GSHMC method $\phi = \sqrt{2\gamma\Delta t}$	0.6500 ± 0.8425	118 ± 14	-340 ± 11
Langevin piston, BBK algorithm	0.6477 ± 0.8580	123 ± 18	-314 ± 45

TABLE V: Mean values and their standard deviation range for pressure, P, temperature, T, and total energy, E, for GSHMC and Langevin piston BBK simulation of the NPT ensemble.

We also implement the constant pressure and temperature GSHMC algorithm using the partial momentum update (63)-(66) and compare the results to the Langevin piston method of Feller et al [17]. The Langevin piston equations of motion are implemented using the Brunger, Brooks, Karplus (BBK) algorithm [24].

The simulation parameters are now as follows. Both methods are implemented with a step-size of $\Delta t = 21.7$ fs, samples are taken at in intervals of $\tau = L\,\Delta t = 0.217$ ps, i.e., $L = 10$, and the total number of samples is $K = 2 \times 10^4$. The mass of the piston degree of freedom is set equal to $\mu = 6$, $\alpha = 0.65 \cdot 10^7$ N m^{-2}, and the collision frequency in the Langevin piston is set equal to $\gamma = 0.1152$ ps^{-1}. The angle, ϕ, in (65)-(66) is determined according to $\phi = \sqrt{2\,\Delta t\,\gamma} \approx 0.2236$. Both methods are started from an equilibrated configuration.

We compare pressure, P, temperature, T, and total energy, E. Mean values and their standard deviation range can be found in table V. Note that both methods couple to a constant temperature 'heat bath' only through the piston degree of freedom. The results from both methods are in agreement (to within the expected errors given the simulation length, the system size, and the weak coupling to the 'heat bath') with the desired NPT ensemble.

B. Lysozyme protein in water

A larger molecular system, the bacteriophage T4 lysozyme protein (pdb entry 2LZM), is simulated to compare the sampling efficiency of GSHMC and constant temperature MD. A united atoms representation is used to eliminate all hydrogen atoms from the protein, and water is modeled using the SPC model [26]. The total number of atoms is 23207, which are placed in a rhombic dodecahedron simulation box. Both simulation approaches, MD and GSHMC, use GROMACS 3.2.1 [27] to perform the molecular dynamics steps. Specifically, a switch cut-off scheme is used for Lennard-Jones interactions. Coulomb interactions are treated using a particle-mesh Ewald summation (PME) method [28, 29]. The full direct and reciprocal space parts are calculated in each step and a lattice spacing of 0.1 nm is applied. All bonds are constrained using the SHAKE method [22] with a relative tolerance of 10^{-12} allowing for a step-size of $\Delta t = 2$ fs.

The system is initially equilibrated for 1 ns using standard MD techniques. The MD and GSHMC simulations are then performed for another 1 ns at a temperature of 300 K. In the traditional MD approach the temperature is coupled to a heat bath of 300 K using the Berendsen thermostat with a coupling time constant of 0.1 ps [25].

To find the optimal settings for GSHMC production stage we investigate the effect of different simulation parameters on the sampling efficiency of GSHMC. A set of comparatively short simulations is performed using three different step-sizes Δt (1, 2 and 4 fs), two different MD simulation lengths τ (2 and 4 ps), five values of the angle ϕ ($\pi/24$, $\pi/12$, 0.3, 0.5, $\pi/2$) and two values of the order p (4, 6) for the modified Hamiltonian $\mathcal{H}^{[p]}$. The results of this study are shown in figures 2 and 3.

Since we found that acceptance rate for MDMC step was consistently high (98-100%) for all tested parameters, we present here the results for the acceptance rate in the PMMC step only. Figure 2 demonstrates the effect of step-size and MD simulation length on the momentum acceptance rates whereas figure 3 shows how the momentum acceptance rate depends on the angle ϕ. The momentum acceptance rate was found to be essentially independent of the order (here 4th and 6th order) of the modified energies.

It can be concluded from figures 2 and 3 that smaller step-sizes, larger MD simulation lengths, and smaller values of ϕ induce a higher acceptance rate in the PMMC step. A nearly optimal choice of the parameter ϕ and the step-size Δt is crucial for the performance of GSHMC. Choosing $\phi = \pi/2$ is found to be not efficient for this large system.

We have to stress that the PMMC step is cheap compared with the MDMC step. To decrease the rejection rate of the PMMC step one can repeat the step a desired number of times. This strategy is efficiently implemented in parallel in our code.

In addition, we consider the evolution of the mean-square displacement of the centre-of-mass (c.o.m.) of the protein for GSHMC simulations using two different values of ϕ: $\phi = \pi/24$ and $\phi = \pi/12$. We find that the c.o.m. mobility of the protein in GSHMC simulation increases with an increasing of ϕ. This is shown in figure 4.

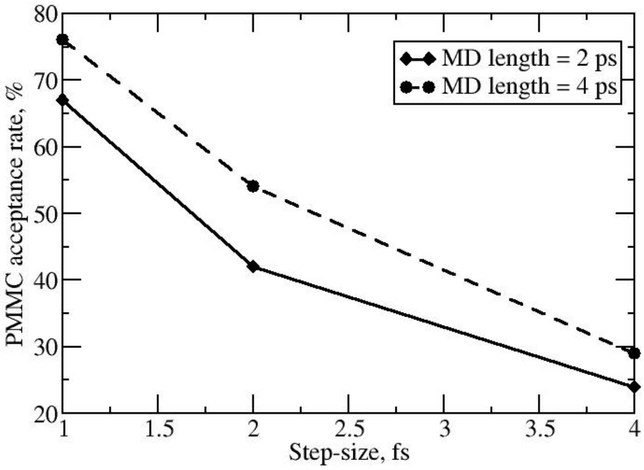


FIG. 2: PMMC acceptance rate vs. MD step-size Δt and MD length τ for fixed angle $\phi = \pi/24$.

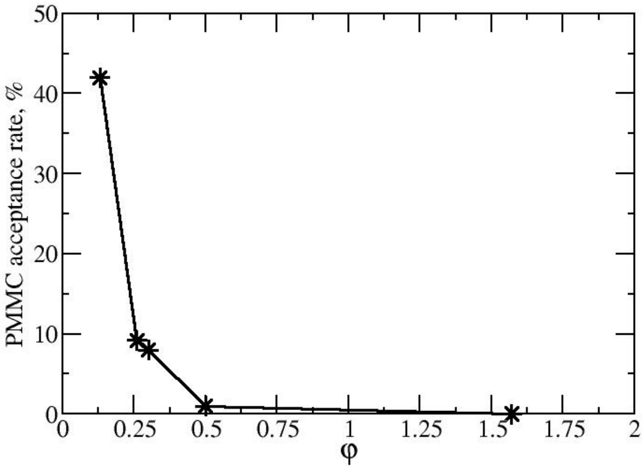


FIG. 3: PMMC acceptance rate vs. ϕ for fixed step-size $\Delta t = 2$ fs and MD simulation length $\tau = 2$ ps.

To perform a comparison between GSHMC and MD simulations we run the GSHMC simulation with a step-size of $\Delta t = 2$ fs, the number of MD steps in MDMC equal to $L = 1000$, and $\phi = \pi/12$ on ten processors of a PC cluster. We use a sixth-order modified energy.

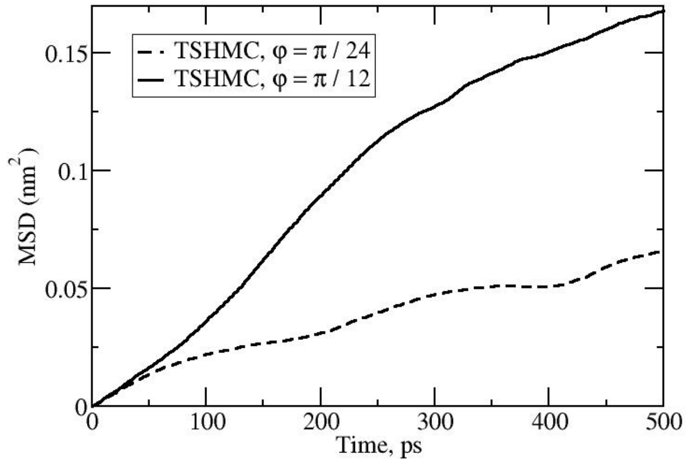


FIG. 4: Mean-square displacements of the protein centre-of-mass vs. ϕ. The mean trajectory for $\phi = \pi/24$ is depicted by a dashed line whereas the trajectory for $\phi = \pi/12$ is presented by a solid line. The step-size is $\Delta t = 2$ fs and the MD simulation length is $\tau = 2$ ps.

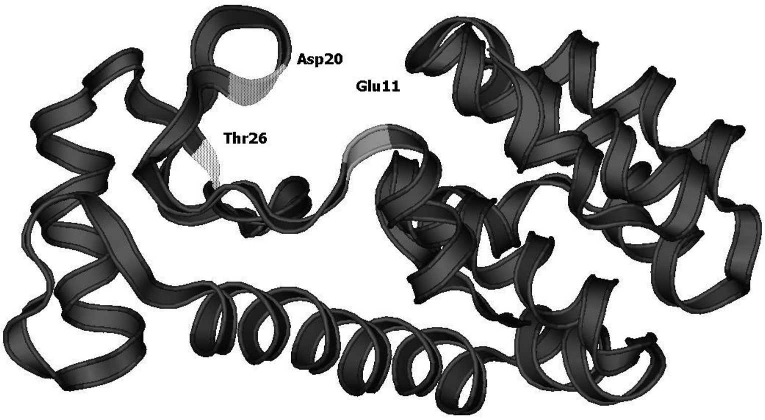


FIG. 5: VMD [30] ribbon diagram of 2LZM illustrating locations of catalytic residues Glu11, Asp20, and Thr26.

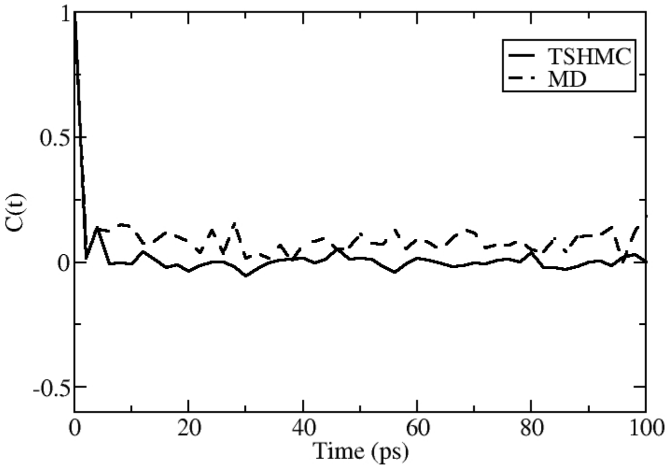


FIG. 6: Autocorrelation function of main chain torsion angle Φ of residue Thr26.

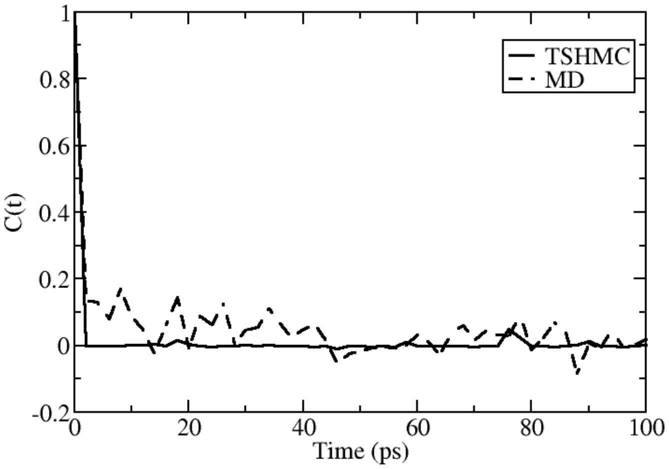


FIG. 7: Autocorrelation function of main chain torsion angle Ψ of residue Thr26.

To compare the sampling efficiency of different sampling methods with respect to an observable Ω, we evaluate the integrated autocorrelation function values of a time series $\{\Omega_i\}_{i=1}^{K}$, where K is the number of samples [15]. The integrated autocorrelation function

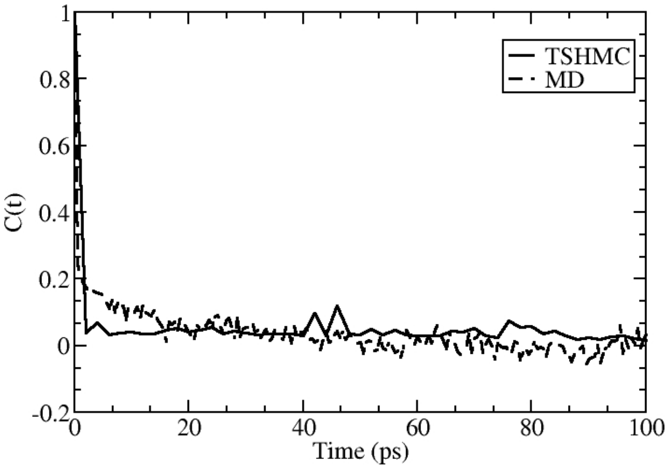


FIG. 8: Autocorrelation function of side chain torsion angle χ_1 of residue Thr26.

value is defined by

$$A_\Omega = \sum_{l=1}^{K'} C(\tau_l), \tag{73}$$

where $C(\tau_l)$, $l = 0, \ldots, K' < K$ is the standard autocorrelation function for the time series $\{\Omega_i\}_{i=1}^{K}$ with the normalization $C(\tau_0) = C(0) = 1$. The integrated autocorrelation function value provides a good measure for the efficiency of a sampling method since, on average, $1 + 2A_\Omega$ correlated measurements Ω_i are needed to reduce the variance by the same amount as a single truly independent measurement of Ω [15].

We present the autocorrelation functions for the dihedrals of Asp20, Glu11 and Thr26 residues in figure 5. These dihedrals are known to be critical catalytic residues in lysozyme. In fact, it has been reported that the catalytic activity of most lysozymes is largely due to three amino acids. In the case of the bacteriophage T4 lysozyme, catalysis takes place due to the concerted action of Glu11, Asp20, and Thr26 with the substrate [31–35].

The autocorrelation functions $C(\tau_l)$ for the main chain torsion angles Φ, Ψ, and a side chain torsion angle χ_1 of the Thr26 residue are shown in figures 6, 7, and 8, respectively, for $\tau_l \leq 100$ ps.

Computed integrated autocorrelation function values, A_Ω, are based on autocorrelation functions $C(\tau_l)$ and $\tau_l \leq 500$ ps. Ratios of integrated autocorrelations function values for the main chain torsion angles Φ, Ψ and side chain torsion angles χ_1, χ_2, χ_3 for residues

$A_\Omega^{\mathrm{MD}}/A_\Omega^{\mathrm{GSHMC}}$	Asp20	Thr26
Φ	3.8	14.0
Ψ	3.4	4.5

TABLE VI: Comparison between GSHMC and MD in efficiency for sampling of main chain torsion angles of important catalytic residues. $A_\Omega^{\mathrm{MD}}/A_\Omega^{\mathrm{GSHMC}}$ is the ratio of integrated autocorrelation function values obtained from MD and GSHMC simulations.

$A_\Omega^{\mathrm{MD}}/A_\Omega^{\mathrm{GSHMC}}$	Glu11	Asp20	Thr26
χ_1	5.54	1.0	2.69
χ_2	7.11	1.56	NA
χ_3	3.76	NA	NA

TABLE VII: Comparison between GSHMC and MD in efficiency for sampling of side chain torsion angles of important catalytic residues. $A_\Omega^{\mathrm{MD}}/A_\Omega^{\mathrm{GSHMC}}$ is a ratio of integrated autocorrelation function values obtained from MD and GSHMC simulations.

Asp20, Glu11 and Thr26, as observed during GSHMC and MD simulations, are presented in table VI and table VII, respectively. As shown in tables VI and VII, GSHMC requires less (up to 14 times!) iterations (MD steps) than standard MD to achieve one statistically independent sample for all torsion angles of catalytic residues Asp20, Glu11 and Thr26.

VIII. SUMMARY

We have presented a more efficient implementation of the GHMC method, which is based on the use of modified energies. The resulting GSHMC/GS2HMC methods allow the user to either perform pure sampling or stochastic dynamics simulations.

In the case of sampling, the GS2HMC method has the advantage of keeping the acceptance rate in the PMMC step high without having to make ϕ smaller as the system size increases. However, the transformation step (41) requires additional force field evaluations. Repeated application of the PMMC step with a reduced value of ϕ and $\bar{\mathbf{p}} = \mathbf{p}$, i.e. no transformation, provides a viable alternative.

The GS2HMC method behaves similarly to the recently proposed S2HMC method. An advantage of GS2HMC over S2HMC is that it can be combined with higher-order (higher than fourth order) modified energies and that it can be used with partial momentum refreshment. To take full advantage of higher-order modified energies, the force field evaluations have to be performed accurately enough and sufficiently smooth cut-off functions need to be implemented.

For small values of $\phi = \sqrt{2\gamma\Delta t}$, i.e. stochastic dynamics simulations, the GSHMC method without the transformation (41) is to be recommended since the acceptance rate in the PMMC step of GSHMC is high for small values of ϕ and since GSHMC is cheaper to implement than GS2HMC.

Numerical experiments have demonstrated that GSHMC/GS2HMC are suitable for NVT as well as NPT simulations. In particular, we have shown that GSHMC/GS2HMC outperform both classical MD as well as standard HMC in terms of sampling. Furthermore, GSHMC provides a statistically rigorous simulation tool for stochastic dynamics in an NVT or NPT ensemble.

We finally wish to mention that the GSHMC method can be used to solve statistical inference problems in the same manner as the standard HMC method can be applied to such problems (see, e.g., [36, 37]). In particular, in a Bayesian framework, all inference problems can be reduced to the evaluation of certain expectation values with respect to the *posterior distribution* of unknown variables. This target posterior distribution can always be written out explicitly, up to a normalization constant, as

$$\pi(\mathbf{q}) \propto f(\mathbf{y}|\mathbf{q})\,\pi_0(\mathbf{q}) \equiv \exp(-V(\mathbf{q})) \tag{74}$$

where f is the probabilistic model that connects data $\mathbf{y}$ with unknown parameters $\mathbf{q}$, π_0 is the prior distribution in $\mathbf{q}$ (which is often assumed to be Gaussian), and

$$V(\mathbf{q}) = -\log f(\mathbf{y}|\mathbf{q}) - \log \pi_0(\mathbf{q}). \tag{75}$$

In order to use the GSHMC to sample the posterior distribution (74), we introduce an auxiliary 'momentum' variable $\mathbf{p}$, a (constant) symmetric mass matrix M, and the 'guide Hamiltonian'

$$\mathcal{H} = \frac{1}{2}\mathbf{p} \cdot [M^{-1}\mathbf{p}] + V(\mathbf{q}) \tag{76}$$

with associated Newtonian equations of motion

$$\dot{\mathbf{q}} = M^{-1}\mathbf{p}, \qquad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}} V(\mathbf{q}). \tag{77}$$

These equations can be integrated in time by a symplectic and time-reversible method such as Störmer-Verlet. The resulting propagator U_τ, with appropriate reference Hamiltonian $\mathcal{H}_{\Delta t}$, is then to be used in the MDMC part of the GSHMC method. The PMMC part and the re-weighting procedure for expectation values remain unchanged.

Appendix

We derive the sixth-order modified energy. Following the approach of section IV A we first derive a modified Lagrangian density to sixth order:

$$\begin{aligned}\mathcal{L}_{\Delta t} =& \frac{1}{4}\left(\sum_{i=1}^{\infty}\frac{\Delta t^{i-1}}{i!}\mathbf{Q}^{(i)}\right)\cdot\left[\mathcal{M}(\mathbf{Q})\left(\sum_{i=1}^{\infty}\frac{\Delta t^{i-1}}{i!}\mathbf{Q}^{(i)}\right)\right] \\ &+\frac{1}{4}\left(\sum_{i=1}^{\infty}\frac{(-1)^{i-1}\Delta t^{i-1}}{i!}\mathbf{Q}^{(i)}\right)\cdot\left[\mathcal{M}(\mathbf{Q})\left(\sum_{i=1}^{\infty}\frac{(-1)^{i-1}\Delta t^{i-1}}{i!}\mathbf{Q}^{(i)}\right)\right] - V(\mathbf{Q}), \\ =& \mathcal{L} + \Delta t^2\,\delta\mathcal{L}^{[4]} + \Delta t^4\,\delta\mathcal{L}^{[6]} + \mathcal{O}(\Delta t^6)\end{aligned} \tag{78}$$

where $\mathcal{L}$ is given by (2), $\delta\mathcal{L}^{[4]}$ by (30), and $\delta\mathcal{L}^{[6]}$ by

$$\delta\mathcal{L}^{[6]} = \frac{1}{720}\left\{6\,\dot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(5)}\right] + 15\,\ddot{\mathbf{Q}}\cdot\left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(4)}\right] + 20\,\mathbf{Q}^{(3)}\cdot\left[\mathcal{M}(\mathbf{Q})\,\mathbf{Q}^{(3)}\right]\right\}. \tag{79}$$

Hence, we define the sixth-order modified Lagrangian density by

$$\mathcal{L}_{\Delta t}^{[6]} = \mathcal{L} + \Delta t^2\,\delta\mathcal{L}^{[4]} + \Delta t^4\,\delta\mathcal{L}^{[6]} \tag{80}$$

and higher-order modified Lagrangian can be found by including higher-order terms in the expansion (78). The sixth-order modified energy is now given by

$$\mathcal{H}_{\Delta t}^{[6]} = \sum_{i=1}^{5}\left\{\sum_{j=0}^{i-1}(-1)^j\left[\frac{d^j}{dt^j}\frac{\partial\mathcal{L}_{\Delta t}^{[6]}}{\partial\mathbf{Q}^{(i)}}\right]\cdot\mathbf{Q}^{(i-j)}\right\} - \mathcal{L}_{\Delta t}^{[6]} \tag{81}$$

with the generalization to higher-order again being straightforward.

[1] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).

[2] B. Mehlig, D. Heermann, and B. Forrest, Phys. Rev. B **45**, 679 (1992).

[3] J. Izaguirre and S. Hampton, J. Comput. Phys. **200**, 581 (2004).

[4] G. Benettin and A. Giorgilli, J. Stat. Phys. **74**, 1117 (1994).

[5] E. Hairer and C. Lubich, Numer. Math. **76**, 441 (1997).

[6] S. Reich, SIAM J. Numer. Anal. **36**, 475 (1999).

[7] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics* (Cambridge University Press, Cambridge, 2005).

[8] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration* (Springer-Verlag, Berlin Heidelberg, 2002).

[9] B. Moore and S. Reich, Numer. Math. **95**, 625 (2003).

[10] R. Skeel and D. Hardy, SIAM J. Sci. Comput. **23**, 1172 (2001).

[11] C. Sweet, S. Hampton, and J. Izaguirre, Tech. Rep. TR-2006-09, University of Notre Dame (2006).

[12] C. Sweet, S. Hampton, R. Skeel, and J. Izaguirre, Tech. Rep., University of Notre Dame (2007).

[13] E. Akhmatskaya and S. Reich, in *New Algorithms for Macromolecular Simulations*, edited by B. L. et al (Springer-Verlag, Berlin, 2006), vol. 49 of *Lecture Notes in Computational Science and Engineering*, pp. 145–158.

[14] A. Horowitz, Phys. Lett. B **268**, 247 (1991).

[15] A. Kennedy and B. Pendleton, Nucl. Phys. B **607**, 456 (2001).

[16] H. Andersen, J. Chem. Phys. **72**, 2384 (1980).

[17] S. Feller, Y. Zhang, R. Pastor, and B. Brooks, J. Chem. Phys. **103**, 4613 (1995).

[18] M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).

[19] R. MacKay, in *The dynamics of numerics and the numerics of dynamics*, edited by D. Broomhead and A. Iserles (Clarendon Press, Oxford, 1992), pp. 137–193.

[20] S. Gupta, A. Irbäck, F. Karsch, and B. Pterersson, Phys. Lett. B **242**, 437 (1990).

[21] R. Burden and J. Faires, *Numerical Analysis* (Brooks Cole, 2004), 8th ed.

[22] J. Ryckaert, G. Ciccotti, and H. Berendsen, J. Comput. Phys. **23**, 327 (1977).

[23] R. Faller and J. de Pablo, J. Chem. Phys. **116**, 55 (2002).

[24] A. Brünger, C. Brooks, and M. Karplus, Chem. Phys. Lett. **105** (1984).

[25] H. Berendsen, J. Postma, W. van Gunsteren, A. DiNola, and J. Haak, J. Chem. Phys. **81**, 3684 (1984).

[26] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (D. Reidel Publishing Company, Dordrecht, 1981), pp. 331–342.

[27] E. Lindahl, B. Hess, and D. Spoel, J. Mol. Modeling **7**, 305 (2001).

[28] T. Darden, D. York, and L. Pedersen, J. Comput. Phys. **98**, 10089 (1993).

[29] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Comput. Phys. **103**, 8577 (1995).

[30] W. Humphries, A. Dalke, and K. Schulten, J. Molec. Graphics p. 33 (1996).

[31] W. Anderson, M. Grütter, S. Remington, L. Weaver, and B. Matthews, J. Mol. Biol **147**, 523 (1981).

[32] L. Hardy and A. Poteete, Biochemistry **30**, 9457 (1991).

[33] R. Kuroki, L. Weaver, , and B. Matthews, Science **262**, 2030 (1993).

[34] R. Kuroki, L. Weaver, and B. Matthews, Nat. Struct. Biol. **2**, 1007 (1995).

[35] R. Kuroki, L. Weaver, and B. Matthews, Proc. Natl. Acad. Sci. **96**, 8949 (1999).

[36] R. Neal, *Bayesian learning for neural networks* (Springer-Verlag, New York, 1996).

[37] J. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer-Verlag, New York, 2001).