Constant pressure hybrid Monte Carlo simulations in GROMACS

Mario Fernández-Pendás[1], Bruno Escribano *[1], Tijana Radivojević[1], and Elena Akhmatskaya[1,2]

[1]Basque Center for Applied Mathematics, E-48009 Bilbao, Spain , Tel.: +34 946 567 842, Fax: +34 946 567 843
[2]IKERBASQUE, Basque Foundation for Science, E-48011 Bilbao, Spain

Abstract

Adaptation and implementation of the Generalized Shadow Hybrid Monte Carlo (GSHMC) method for molecular simulation at constant pressure in the NPT ensemble are discussed. The resulting method, termed NPT-GSHMC, combines Andersen barostat with GSHMC to enable molecular simulations in the environment natural for biological applications, namely, at constant pressure and constant temperature. Generalized Hybrid Monte Carlo methods are designed to maintain constant temperature and volume and extending their functionality to preserving pressure is not trivial. The theoretical formulation of NPT-GSHMC was previously introduced. Our main contribution is the implementation of this methodology in the GROMACS molecular simulation package and the evaluation of properties of NPT-GSHMC, such as accuracy, performance, effectiveness for real physical systems in comparison with well-established molecular simulation techniques. Benchmarking tests are presented and the obtained preliminary results are promising. For the first time, the generalized hybrid Monte Carlo simulations at constant pressure are available within the popular open source molecular dynamics software package.

1 Introduction

The isobaric-isoenthalpic and isobaric-isothermal ensembles (also called NPH and NPT ensembles respectively) are the statistical ensembles where a number of particles (N), a pressure (P) as well as either an enthalpy (H) or a temperature (T) are each fixed to particular values. These ensembles play a very important role in chemistry and biology where many processes are carried out at constant pressure. Mathematical techniques called barostats are developed to keep constant pressure during a molecular simulation. In the case of NPT ensembles, barostats are combined with thermostats responsible for temperature maintenance.

Two main kinds of barostats are being mentioned here: those, which introduce an extended variable for the equations of motion (extended ensemble coupling) and those that use an external bath to perform the coupling (weak coupling).

In the classical work by Andersen [1], a method with an extended variable has been proposed. The system is coupled to a fictitious "pressure bath" using an extended Lagrangian, in which the volume acts as an additional variable. The coupling mimics the action of an imaginary external *piston* on a simulated system and the new variable plays a role of the coordinate of a *piston* linked to an external constant reference pressure. The resulting equations of motion produce trajectories which sample the NPH ensemble. The combination with one of the constant-temperature methods (thermostats) allows NPT simulations. The Parrinello-Rahman barostat [2], the

*Contact address: bescribano@bcamath.org

Nosé-Hoover barostat [3, 4, 5] and the Martyna-Tuckerman-Tobias-Klein barostat, MTTK, [6] are all based on the Andersen barostat.

The most popular barostat that uses the external bathing approach is the Berendsen barostat [7]. Instead of modifying the Hamiltonian, as in the previous examples, it proposes a weak coupling to an external bath using the principle of least local perturbation consistent with the required global coupling.

All listed barostat techniques (which in fact do not cover the whole range of methods developed till now, but which are the most relevant to the topic of this paper) are commonly used in MD simulations and freely available in popular MD software packages, such as GROMACS, AMBER, LAMPPS, Desmond [8, 9, 10, 11, 12], etc. Usually several barostats and thermostats are implemented in a modern software suite to cover a range of needs. For example, in GROMACS, the Berendsen, Parrinello-Rahman and MTTK barostats as well as Berendsen, Nosé-Hoover and velocity rescaling thermostats are currently available. Each barostat or thermostat technique has its own limitations and it is a user's responsibility to choose the most appropriate method or their combination for the problem of interest.

It is less common however to see the hybrid Monte Carlo (HMC) method [13] implemented in the popular publicly available molecular dynamics software codes, although the method by construction maintains constant temperature and may serve as a rigorous thermostat. The lack of attention to this method is explained by its reputation of being ineffective for simulation of large complex dynamical systems. Recently several efficient modified hybrid Monte Carlo methods have been introduced [14, 15, 16, 17, 18, 19, 20, 21], which proved to be competitive and often superior to the well established simulation techniques such as thermostated Molecular Dynamics [22, 7, 5]. However, those methods are not well known and mainly used only by their authors. To make the methods available to a broader research community, recently we have implemented the generalized shadow hybrid Monte Carlo methods (GSHMC) [23] in the popular GROMACS package [8, 9]. As its name suggests, with a proper choice of parameters and conditions, the method easily reduces to HMC, generalized HMC, Langevin Monte Carlo or Metropolis adjusted Langevin dynamics, that enriches the GROMACS suite with a set of thermostats. What is missed here, however, is an ability of the listed techniques to maintain the constant pressure during the simulation.

In this work we present the extension of GSHMC to the simulation in the NPT ensemble, which we called NPT-GSHMC, and its implementation in the GROMACS package. Our goal is to make available the efficient, flexible simulation methodology to a broad simulation community. The paper is organised as follows. In section 2 we introduce the method itself: GSHMC in the NPT ensemble. Mathematical formulation is presented and implementation is discussed in details. The results of testing the new implementation on two simple systems in comparison with MD in NPT and GSHMC in NVT are presented in section 3. Conclusions are given in section 4.

2 NPT-GSHMC

2.1 Formulation

The NPT-GSHMC method has been already mathematically formulated in detail in [14]. The method combines the generalized shadow hybrid Monte Carlo (GSHMC) methodology [14] with the Andersen barostat [1]. In this subsection, we briefly summarize the GSHMC algorithm and specify the major steps that should be taken to extend it to simulation at constant pressure.

The GSHMC method introduced in [14] is a generalized hybrid Monte Carlo (GHMC) [16, 17] which samples with respect to a modified energy (also called a shadow Hamiltonian). As a modification of GHMC, it consists of two alternating steps: (i) a generation of short molecular dynamics trajectories in the NVE ensemble, i.e. at a constant number of particles N, a constant volume V and a constant energy E, and (ii) a partial momentum update preceding each molecular dynamics trajectory. The decision on accepting / rejecting a proposal in steps (i) and (ii) is made using the appropriate Metropolis function with the true Hamiltonian replaced by the shadow Hamiltonian, $\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p})$, acting as a new modified reference energy. The shadow Hamiltonian is obtained from a truncated Taylor expansion of the usual Lagrangian following the standard Lengendre transform. In this paper we will use

the fourth order shadow Hamiltonian. The ways of calculating $\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p})$ are discussed in [14]. The objective of the GSHMC method is to reduce a number of rejected trajectories through the use of shadow Hamiltonians while retaining dynamical information by only partially refreshing momenta.

The GSHMC algorithm can be summarized as follows:

- Given positions $\mathbf{r}$ and momenta $\mathbf{p}$, evaluate $\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p})$.
- Obtain $\mathbf{u}$ from a Gaussian distribution
$$\mathbf{u} = \beta^{-1/2} M^{1/2} \xi, \tag{1}$$
where $\beta = 1/k_B T$, k_B is the Boltzmann constant, M is the mass matrix and ξ is a noise vector generated from a Gaussian distribution as $\xi = (\xi_1, \ldots, \xi_{3N})^T$, $\xi_i \sim \mathscr{N}(0,1)$, $i = 1, \ldots, 3N$.
- Generate momenta $\mathbf{p}'$ and a vector of auxiliary variables $\mathbf{u}'$ using the momentum update procedure:
$$(\mathbf{u}', \mathbf{p}') = \begin{cases} [R(\phi)(\mathbf{u},\mathbf{p})^T]^T & \text{with probability } P(\mathbf{r},\mathbf{p},\mathbf{u},\mathbf{p}',\mathbf{u}'), \\ (\mathbf{u},\mathbf{p}) & \text{otherwise,} \end{cases} \tag{2}$$
where
$$P(\mathbf{r},\mathbf{p},\mathbf{u},\mathbf{p}',\mathbf{u}') = \min\left\{1, \frac{\exp(-\beta[\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p}') + \frac{1}{2}(\mathbf{u}')^T M^{-1}\mathbf{u}'])}{\exp(-\beta[\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p}) + \frac{1}{2}\mathbf{u}^T M^{-1}\mathbf{u}])}\right\}, \tag{3}$$
and
$$R(\phi) = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}, \tag{4}$$
where ϕ is a parameter taking values $0 < \phi \leq \pi/2$. A prior evaluation of $\mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p}')$ is required for calculating $P(\mathbf{r},\mathbf{p},\mathbf{u},\mathbf{p}',\mathbf{u}')$.

  It is worth noting that $\mathbf{u}$ is totally discarded after each step and it is replaced by a new set of random variables.
- Given $(\mathbf{r},\mathbf{p}')$, integrate the Hamiltonian equations of the system using the symplectic method $\Psi_{\Delta t}$ over L steps with step-size Δt. $\Psi_{\mathscr{T}}(\mathbf{r},\mathbf{p}') = (\mathbf{r}_{\text{new}}, \mathbf{p}_{\text{new}})$, $\mathscr{T} = L\Delta t$.
- Evaluate $\mathscr{H}_{\Delta t}(\mathbf{r}_{\text{new}}, \mathbf{p}_{\text{new}})$.
- Accept the new configuration $(\mathbf{r}_{\text{new}}, \mathbf{p}_{\text{new}})$ with a Metropolis test where the accepting probability is
$$\min\{1, \exp(-\beta\left[\mathscr{H}_{\Delta t}(\mathbf{r}_{\text{new}}, \mathbf{p}_{\text{new}}) - \mathscr{H}_{\Delta t}(\mathbf{r},\mathbf{p}')\right])\}. \tag{5}$$
  - If accepted: choose $(\mathbf{r},\mathbf{p}) = (\mathbf{r}_{\text{new}}, \mathbf{p}_{\text{new}})$ as a new configuration.
  - If not: take $\mathbf{r}$ and negate momenta, i.e. $\mathbf{p} = -\mathbf{p}'$.
- Go to the first step.

As the simulation is performed in the modified ensemble with respect to shadow Hamiltonian, reweighting has to be applied to calculations of statistical averages [14]. Given an observable $\Omega(\mathbf{r},\mathbf{p})$ and its values Ω_i, $i = 1, \ldots, K$, along a sequence of states $(\mathbf{r}_i, \mathbf{p}_i)$, $i = 1, \ldots, K$, we reweight Ω_i to compute averages $\langle\Omega\rangle$ by applying the formula
$$\langle\Omega\rangle_K = \frac{\sum_{i=1}^K w_i \Omega_i}{\sum_{i=1}^K w_i}$$
with weight factors
$$w_i = \exp\left(-\beta(\mathscr{H}(\mathbf{r}_i,\mathbf{p}_i) - \mathscr{H}_{\Delta t}(\mathbf{r}_i,\mathbf{p}_i))\right).$$

To extend this methodology to simulations in the NPT ensemble, the following modifications are required. First, the MD simulations have to be performed in the NPE ensemble rather than in the NVE ensemble. If the barostat chosen in the NPE simulations leads to the modification of Hamiltonian, then the shadow Hamiltonians will be different from those suggested for simulations in NVT ensembles and have to be derived specifically for this case. The integrator used for solving the associated modified equations of motions should be also symplectic as in the original GSHMC method. Below we briefly show how all those problems were addressed in the new NPT-GSHMC method. More details can be found in [14].

The Andersen barostat has been chosen for maintaining constant pressure in MD simulations. The Andersen barostat is based on the introduction of a new extended variable, which physical meaning is the (dynamic) value of the volume of the simulation box. The extended variable is an additional degree of freedom, it must be included in the Lagrangian and new equations of motion are derived. It is also used as a rescaling factor for the positions. Following Andersen's terminology we refer to the extended variable as the *piston*.

More specifically, in the classical equations of motion

$$\dot{\mathbf{r}} = -\frac{\partial \mathscr{H}}{\partial \mathbf{r}}, \;\; \dot{\mathbf{p}} = +\frac{\partial \mathscr{H}}{\partial \mathbf{p}} \tag{6}$$

the coordinate vector $\mathbf{r} \in \mathbb{R}^{3N}$ is replaced by a scaled vector $\mathbf{d} \in \mathbb{R}^{3N}$ defined as

$$\mathbf{d} = \mathbf{r}/\mathscr{V}^{1/3}, \tag{7}$$

where $\mathscr{V}$ is the volume of the simulation box.

As the volume $\mathscr{V}$ is allowed to change in order to keep constant pressure, we introduce q as the dynamic value of the volume.

The extended Lagrangian density then reads as

$$\mathscr{L}(\dot{\mathbf{d}}, \dot{q}, \mathbf{d}, q) = \left\{ \frac{1}{2} q^{2/3} \dot{\mathbf{d}} \cdot [M\dot{\mathbf{d}}] - U(q^{1/3}\mathbf{d}) + \frac{\mu}{2}\dot{q}^2 - \alpha q \right\}, \tag{8}$$

where α is the external pressure acting on the system, $\mu > 0$ is the mass of the *piston* and U is the potential energy function. The last two terms of (8) are in fact the kinetic and potential energies associated with the *piston*.

The Hamiltonian $\mathscr{H}$, obtained from (8), is given by

$$\mathscr{H} \quad = \dot{\mathbf{d}} \cdot \nabla_{\dot{\mathbf{d}}} \mathscr{L} + \dot{q} \nabla_{\dot{q}} \mathscr{L} - \mathscr{L} = \frac{1}{2} \mathbf{p}_{\mathbf{r}} \cdot [M^{-1} \mathbf{p}_{\mathbf{r}}] + U(\mathbf{r}) + \frac{1}{2\mu} p^2 + \alpha q, \tag{9}$$

where

$$\mathbf{p}_{\mathbf{d}} = q^{2/3} M \dot{\mathbf{d}}, \;\; p = \mu \dot{q} \tag{10}$$

are the conjugate momenta in the NPE formulation, whereas $\mathbf{p}_{\mathbf{r}} = M\dot{\mathbf{r}} = \mathbf{p}_{\mathbf{d}}/q^{1/3}$ is the NVE momentum vector (6). The associated NPE equations of motion now can be obtained using (6) and (9).

A time-reversible and symplectic method for integrating the NPE equations of motions are suggested in [14] and summarized below:

Given $(\mathbf{d}^n, q^n, \mathbf{p}_{\mathbf{d}}^n, p^n)$ and a step-size Δt we get $\mathbf{d}^{n+1}$ and q^{n+1} from

$$\mathbf{p}_{\mathbf{d}}^n = \frac{1}{2}[(q^{n+1})^{2/3} + (q^n)^{2/3}] M \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) + \frac{\Delta t}{2} \nabla_{\mathbf{d}} U((q^n)^{1/3}\mathbf{d}^n) \tag{11}$$

$$p^n = \mu \left(\frac{q^{n+1} - q^n}{\Delta t} \right) - \frac{\Delta t}{6} (q^n)^{-1/3} \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \cdot \left[M \left(\frac{\mathbf{d}^{n+1} - \mathbf{d}^n}{\Delta t} \right) \right] + \frac{\Delta t}{2} [\nabla_q U((q^n)^{1/3}\mathbf{d}^n) + \alpha]. \tag{12}$$

Then, to complete one step, the values of $\mathbf{p}_{\mathbf{d}}^{n+1}$ and p^{n+1} are explicitly obtained from

$$\mathbf{p}_{\mathbf{d}}^{n+1} = \frac{1}{2}[(q^{n+1})^{2/3} + (q^n)^{2/3}]M\left(\frac{\mathbf{d}^{n+1}-\mathbf{d}^n}{\Delta t}\right) - \frac{\Delta t}{2}\nabla_{\mathbf{d}}U((q^{n+1})^{1/3}\mathbf{d}^{n+1}) \tag{13}$$

$$p^{n+1} = \mu\left(\frac{q^{n+1}-q^n}{\Delta t}\right) - \frac{\Delta t}{6}(q^n)^{-1/3}\left(\frac{\mathbf{d}^{n+1}-\mathbf{d}^n}{\Delta t}\right)\cdot\left[M\left(\frac{\mathbf{d}^{n+1}-\mathbf{d}^n}{\Delta t}\right)\right] - \frac{\Delta t}{2}[\nabla_q U((q^{n+1})^{1/3}\mathbf{d}^{n+1}) + \alpha]. \tag{14}$$

Finally, the expression for the fourth order shadow Hamiltonian associated with the real Hamiltonian $\mathscr{H}$ is provided in [14] by

$$\begin{aligned}\mathscr{H}_{\Delta t}^{[4]} = \mathscr{H} \quad &+\frac{\Delta t^2}{24}\left\{2\mu\dot{Q}Q^{(3)} - \mu\ddot{Q}^2 + 2Q^{2/3}\dot{\mathbf{D}}\cdot[M\mathbf{D}^{(3)}] - Q^{2/3}\ddot{\mathbf{D}}\cdot[M\ddot{\mathbf{D}}]\right\} \\ &+\frac{\Delta t^2}{12}\left\{\left(\frac{4\ddot{Q}}{3Q^{1/3}} - \frac{4\dot{Q}^2}{9Q^{4/3}}\right)\dot{\mathbf{D}}\cdot[M\dot{\mathbf{D}}] - \frac{2}{3Q^{1/3}}\dot{Q}\dot{\mathbf{D}}\cdot[M\ddot{\mathbf{D}}]\right\},\end{aligned} \tag{15}$$

where $Q(t)$ and $\mathbf{D}(t)$ are the interpolation polynomials along numerical trajectories $\{q^n\}$ and $\{\mathbf{d}^n\}$, respectively.

It should be noticed here that the introduction of the Andersen barostat in GSHMC leads also to the modification of the partial momentum update step, namely, updating the *piston* momentum should be also included.

The complete algorithm for the NPT-GSHMC method now can be summarized as follows:

- Given positions $\mathbf{d}$ and associated momenta $\mathbf{p}_{\mathbf{d}}$ evaluate a shadow Hamiltonian (15).
- Draw the noise vector $\mathbf{u}$ from the Gaussian distribution as in (1) and take $u = \beta^{-1/2}\mu^{1/2}\xi$, $\xi \sim \mathscr{N}(0,1)$.
- Generate momenta $\mathbf{p}'_{\mathbf{d}}$ and the vector $\mathbf{u}'$ as in (2). For the refreshment of the *piston* momentum p a similar procedure is followed:

$$\begin{aligned} u' &= -\sin(\phi)p + \cos(\phi)u, \\ p' &= \cos(\phi)p + \sin(\phi)u. \end{aligned} \tag{16}$$

  The probability of acceptance (3) in this case is slightly modified and replaced by

$$P(\mathbf{d},q,\mathbf{p}_{\mathbf{d}},p,\mathbf{u},u,\mathbf{p}'_{\mathbf{d}},p',\mathbf{u}',u') = \min\left\{1, \frac{\exp(-\beta[\mathscr{H}_{\Delta t}(\mathbf{d},q,\mathbf{p}'_{\mathbf{d}},p') + \frac{1}{2}(\mathbf{u}')^T M^{-1}\mathbf{u}' + \frac{1}{2\mu}(u')^2])}{\exp(-\beta[\mathscr{H}_{\Delta t}(\mathbf{d},q,\mathbf{p}_{\mathbf{d}},p) + \frac{1}{2}\mathbf{u}^T M^{-1}\mathbf{u} + \frac{1}{2\mu}u^2])}\right\}. \tag{17}$$

  The shadow Hamiltonian corresponding to the newly generated momenta has to be evaluated using (15) in order to complete Metropolis test.

- The system is integrated following the new symplectic method (11)–(14) for L steps with the step-size Δt to obtain $(\mathbf{d}_{\text{new}}, \mathbf{p}_{\mathbf{d}_{\text{new}}}, q_{\text{new}}, p_{\text{new}})$.
- The new shadow Hamiltonian (15) is evaluated at the positions and momenta $(\mathbf{d}_{\text{new}}, \mathbf{p}_{\mathbf{d}_{\text{new}}}, q_{\text{new}}, p_{\text{new}})$ obtained from the integration.
- Accept the new configuration $(\mathbf{d}_{\text{new}}, \mathbf{p}_{\mathbf{d}_{\text{new}}}, q_{new}, p_{new})$ with probability

$$\min\{1, \exp(-\beta[\mathscr{H}_{\Delta t}(\mathbf{d}_{\text{new}}, \mathbf{p}_{\mathbf{d}_{\text{new}}}, q_{\text{new}}, p_{\text{new}}) - \mathscr{H}_{\Delta t}(\mathbf{d}, \mathbf{p}'_{\mathbf{d}}, q, p')])\}. \tag{18}$$

  - If accepted: choose $(\mathbf{d}, \mathbf{p}_{\mathbf{d}}, q, p) = (\mathbf{d}_{\text{new}}, \mathbf{p}_{\mathbf{d}_{\text{new}}}, q_{\text{new}}, p_{\text{new}})$ as a new configuration.
  - If not: take $\mathbf{d}$ and q, and negate momenta, i.e. $\mathbf{p}_{\mathbf{d}} = -\mathbf{p}'_{\mathbf{d}}$ and $p = -p'$.

- Go to the first step.

A change of variable option aiming to increase a momenta acceptance rate is implemented in this algorithm as explained in [14].

At the end of simulation, re-weighting of expectation values should be performed to recover the Boltzmann distribution.

In the next subsection, the implementation of this algorithm is explained in detail.

2.2 Implementation in GROMACS

The NPT-GSHMC method has been implemented in the GROMACS software package [8, 9], the modified version 4.5.4. This version has been chosen in order to complement from the implementation of the NVT-GSHMC [23] and also for the straightforward comparison of the accuracy and performance of both hybrid Monte Carlo methodologies. Extending the implementation of NVT/NPT-GSHMC to the latest, GPU accelerated, version of GROMACS is planned for the nearest future.

GROMACS is a popular MD software package available under the GNU Lesser General Public License. It is written in the C programming language, highly optimized for maximal computational efficiency and fully parallelized using the MPI protocol. The package is mainly used for performing molecular dynamics simulations. It supports most important algorithms expected from a modern molecular dynamics implementation, including popular barostats, such as Berendsen, Parrinello-Rahman, and MTTK, and thermostats, such as Berendsen, Nosé-Hoover and V-Rescale [24].

The generalized shadow hybrid Monte Carlo (GSHMC) method [14], recently implemented in GROMACS [23] (we shall call the resulting code here GROMACS-GSHMC), provides a rigorous method for performing constant temperature simulations and can be served as a thermostat itself. In order to achieve constant temperature and constant pressure simulation, one also needs to have the Andersen barostat at hand as well as the implemented specific features of the NPT-GHSMC method explained in the previous subsection.

The Andersen barostat is not available in the released version of GROMACS though the MTTK, an Andersen-based barostat [6], is already implemented there. This barostat must be combined with a Nosé-Hoover thermostat [24] for running simulations in the NPT ensemble and it does not allow using a different thermostat, for example GSHMC. Thus it cannot serve our purposes and it was necessary to implement the original formulation of Andersen barostat in GROMACS-GSHMC. In practice it means the implementation of a new symplectic and time-reversible integrator (11)-(14). For simplicity and consistency, the new integrator was introduced as a modification of the existing velocity Verlet algorithm.

Other modifications included:

- Evaluation of an NPT shadow Hamiltonian (15).
- Adding a new momentum update procedure (16), specific to the NPT-GSHMC algorithm.
- Adding new options to the *.mdp* configuration file.

Symplectic integrator: The symplectic time-reversible integrator has been extended to the case of the Andersen equations of motion. The updating scheme is the following (for further details the reader can consult [25]).

We begin with performing a half step for the velocities:

- $\dot{\mathbf{r}}^{n+1/2} = \dot{\mathbf{r}}^{n} + \dfrac{\Delta t}{2}\dfrac{1}{M} f^{n}$, with the force f^n corresponding to the velocity $\dot{\mathbf{r}}^n$,
- $\dot{\mathbf{d}}^{n+1/2} = \dfrac{\dot{\mathbf{r}}^{n+1/2}}{(q^n)^{1/3}}$,
- $\dot{q}^{n+1/2} = \dot{q}^{n} + \dfrac{\Delta t}{2}\dfrac{1}{\mu}(P-\alpha)$, with the pressure P evaluated taking the old positions $\mathbf{r}^n$, the old volume q^n but the already updated velocities $\dot{\mathbf{r}}^{n+1/2}$.

Then perform a full step for the positions:

- $q^{n+1/2} = q^{n} + \dfrac{\Delta t}{2}\dot{q}^{n+1/2}$,
- $\mathbf{r}^{n+1} = \mathbf{r}^{n} + \Delta t \dfrac{(q^n)^{2/3}}{(q^{n+1/2})^{2/3}}\dot{\mathbf{r}}^{n+1/2}$,

- $\mathbf{d}^{n+1} = \mathbf{d}^n + \Delta t \dot{\mathbf{d}}^{n+1/2}$,
- $q^{n+1} = q^{n+1/2} + \frac{\Delta t}{2} \dot{q}^{n+1/2}$.

Now two rescaling steps follow:

- $\dot{\mathbf{r}}^{n+1/2} = \dot{\mathbf{r}}^{n+1/2} \frac{(q^n)^{1/3}}{(q^{n+1})^{1/3}}$,
- $\mathbf{r}^{n+1} = \mathbf{r}^{n+1} \frac{(q^{n+1})^{1/3}}{(q^n)^{1/3}}$.

And finally we complete the full step for the velocities:

- $\dot{q}^{n+1} = \dot{q}^{n+1/2} + \frac{\Delta t}{2} \frac{1}{\mu} (P - \alpha)$, with the pressure P evaluated taking the new positions $\mathbf{r}^{n+1}$, the new volume q^{n+1} but the half-step velocities $\dot{\mathbf{r}}^{n+1/2}$,
- $\dot{\mathbf{r}}^{n+1} = \dot{\mathbf{r}}^{n+1/2} + \frac{\Delta t}{2} \frac{1}{M} f^{n+1}$, with the force f^{n+1} evaluated in the previous velocity $\dot{\mathbf{r}}^{n+1/2}$,
- $\dot{\mathbf{d}}^{n+1} = \frac{\dot{\mathbf{r}}^{n+1}}{(q^{n+1})^{1/3}}$.

Since in the updating scheme above the dynamic value of volume q is changing, one has to make sure that the simulation box is also changing to fit to this volume. In the code it is done by re-scaling the box dimensions with the new value of the dynamic volume in the function *update_box()*. This implementation only applies to the case of a simulation box changing isotropically.

It is important to mention that in order to make the integration scheme (11)-(14) working, the values of pressure and forces have to be updated every time step. This is done in the original version of the GROMACS code. However, the frequency of the pressure updates has to be specified by a user in the GROMACS parameter file. For using the NPT-GSHMC within the GROMACS code, such a parameter should be always set to 1. Such a choice does not introduce an important computational overhead as can be seen from the numerical tests in the following section.

It is also noteworthy that GROMACS works with velocities instead of momenta. That is why the theoretical formulation (10) is slightly modified in the above scheme taking into account the relation between velocities and momenta.

The current version of the GROMACS software offers the velocity Verlet integrator. The new integrator (11)-(14) is placed in the same part of the code. The both GROMACS routines for updating positions and velocities need to be modified in the function *update_coords()*, but the modifications are straightforward, mainly related to a change of parameters of the subroutines.

There is also another important issue to consider: in GROMACS, when dealing with pressures, a rescaling factor is used. It has to be included in the time integration of equations of motion for the volume q, and in the calculation of the additional energy terms.

Shadow Hamiltonian: In order to introduce the NPT Shadow Hamiltonians (15) in the GROMACS-GSHMC code, the shadow Hamiltonian implemented in [23] can be taken as a starting point. As GSHMC (even HMC) methods are not a part of the released GROMACS version, the shadow Hamiltonian appears in a new piece of the code in the subroutine *shadow()* which is called from the function *do_md()* in its main time-step loop. As stated in section 2.1, in the NPT ensemble one has to consider a different shadow Hamiltonian (15) where the extended variable (the *piston* volume) introduces new terms. However, this modification does not entail a great complexity since the NPT shadow Hamiltonians are calculated in a similar way as the NVT shadow Hamiltonians.

Both types of shadow Hamiltonians are currently available in the code and can be chosen at runtime according to the parameters of the simulation.

Momentum refreshment: In comparison with GROMACS-GSHMC, the momentum refreshment procedure for the NPT-GSHMC requires also the update of the momentum p for the *piston*. This is a relatively simple extension of the previous implementation. The algorithmic details can be found in section 2.1.

Parameter file: GROMACS needs to receive two new parameters through the *.mdp* parameter file, the *piston* mass μ and the reference pressure α. Additionally, the Andersen barostat has to be recognized as a pressure coupling method. These modifications were done in the standard way described in the GROMACS Developer's Guide [24]. The specific parameters in the *.mdp* file look like this:

```
; Generalized Shadow Hybrid Monte Carlo =
GSHMC                   = yes
parameter_phi           = 0.32
nr_mom_updates          = 1
variable_change         = no
nr_MD_steps             = 100
hamiltonian_order       = 4
canonical_temperature   = 310
momentum_flip           = yes

; Andersen barostat =
Pcoupl                  = Andersen
Pcoupltype              = isotropic
mu_mass                 = 100
alpha_press             = 1
```

More details on GSHMC input parameters can be found in [23].

3 Results

We test the new NPT-GSHMC method by comparing it with the NVT-GSHMC implementation [23] and NPT-MD which uses the velocity rescale thermostat [22], the Parrinello-Rahman barostat [2] and the position leapfrog integrator (as required by the chosen barostat). The same code with the appropriate choice of parameters for each case is used for running all three simulations.

As testing systems, we choose a coarse-grained model of the VSTx1 toxin in a POPC bilayer [26] and an atomistic model of the protein villin [27]. From here on these systems are denoted as toxin and villin. In the coarse-grained system, four heavy particles on average are represented as one sphere [28, 29], which produces a total number of 7810 particles. The integrator step-size is set to 20 fs for optimal accuracy and 30 fs for optimal sampling. For both Coulomb and Van der Waals interactions the shift algorithm is used. Both potential long range energies are shifted to 0 kJmol^{-1} at a radius of 1.2 nm and the forces are updated from those potentials. Periodic boundary conditions are considered in all directions. No specific constraints are taken into account but the ones defined in the topology files. For the NPT-GSHMC particular case we use the *.mdp* parameters shown above. The villin protein is composed of 389 atoms and the system is solvated with 3000 water molecules. The integrator step-size for this system is set to the standard 1 fs in all cases. Coulomb and Van der Waals interactions and periodic boundary conditions are considered as in the previous system. However, for villin, the bonds with H-atoms are converted to constraints and the constraint algorithm used is LINCS. The specific NPT-GSHMC parameters are the same as taken for the coarse-grained system but with the angle ϕ equal to 0.4 and the reference temperature equal to 300 K.

Table 1: Toxin-bilayer system statistical averages.

	averages				Acc. Rates	
simulation	d (nm)	T (K)	P (bar)	U_{pb} (kJ mol^{-1})	A_r(%)	A_p(%)
NPT-GSHMC	2.3±0.4	308.5±0.3	1.2±0.5	-16.3±2.0	97	83
NVT-GSHMC	2.4±0.3	308.4±0.1	–	-14.9±0.6	100	85
NPT-MD	2.4±0.4	309.9±0.1	0.6±0.4	-15.8±0.2	–	–

Table 2: Villin protein system statistical averages.

	averages			Acc. Rates	
simulation	T (K)	P (bar)	U_{dih} (kJ mol^{-1})	A_r(%)	A_p(%)
NPT-GSHMC	299.5±0.7	1.4±0.9	276±1	95	94
NVT-GSHMC	299.9±0.9	–	276±3	100	97
NPT-MD	299.9±0.7	1.7±0.4	282±1	–	–

3.1 Accuracy

In order to test the accuracy of the new method, we calculate averages for several thermodynamic observables in similar simulations with the three methods. As it was discussed in the previous section, the simulations involving the GSHMC method need to re-weight statistical averages to compensate for the disturbance introduced by the use of shadow Hamiltonians [14].

In the case of the toxin system, 30 ns simulations were performed with an integrator step size of 20 fs, with the target temperature of 310 K and the target pressure of α=1 bar. It should be noted that the efficiency and precision of all three methods can vary according to several tuning parameters. In Table 1 typical results for all methods are shown with a set of parameters chosen for optimizing the accuracy of results.

We choose to monitor four properties of the toxin system, (i) the distance traveled by the toxin from the centre of the membrane to the preferable location at the surface of the membrane d, (ii) the temperature T, (iii) the pressure P and (iv) the Coulomb energy between the protein and the bilayer U_{pb}. For the GSHMC methods, the re-weighted averages are given. All calculated properties are in a good agreement. Error estimates correspond to the standard deviation as provided by GROMACS [8, 9].

The Coulomb potential energy between the protein and the bilayer has been measured before for similar coarse-grained simulations [20], with resulting values close to -16 kJ/mol, which is consistent with our results (see Table 1). The Andersen barostat shows slightly more accurate pressure than the Parrinello-Rahman in our tests. These particular systems exhibit pressure oscillations of considerable amplitude, so we consider that the reported values for both barostats are sufficiently accurate. The NVT-GSHMC has no pressure coupling so the measured average is disregarded.

Table 2 shows the test results for the villin system. Simulations were run for 1 ns with a step size of 1 fs, the target temperature of 300 K and the reference pressure of 1 bar. The observed average temperatures, T, and dihedral potential energies, U_{dih}, agree well for all simulation methods. Similar average values of pressure are achieved with both NPT simulations, NPT-MD and NPT-GSHMC.

3.2 Sampling

The GSHMC method and the Andersen barostat have several tuning parameters that can affect their performance. The two most important parameters in the case of GSHMC are the length of the MD trajectories and the angle in the momentum update procedure, ϕ. When the length of trajectories is too long, the gain over MD in terms of

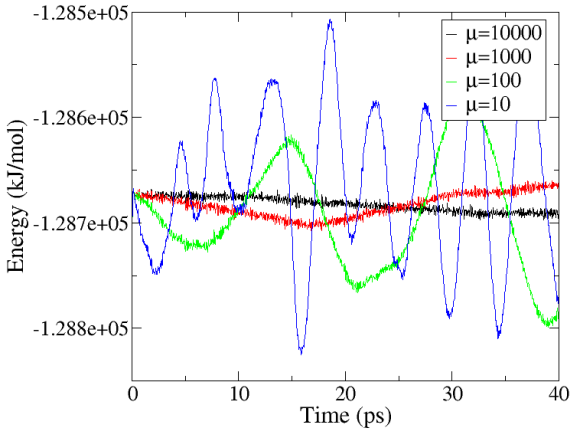


Figure 1: Total energy oscillations in the villin system using NPT-GSHMC with varying *piston* masses μ.

sampling efficiency is less noticeable. But if the length is too short, then the computational time spent on frequent calculations of shadow Hamiltonians becomes too expensive.

The value of ϕ must be between 0 and $\pi/2$. If it is too small then the temperature coupling might be too weak, but larger values interfere with the dynamics and can yield very low acceptance rates. The optimal values for the length of trajectories and ϕ are usually found through trial and error. However the reasonable default ranges for both parameters are not difficult to define and in the case of biomolecular simulation, they are 500-1500 integraton steps and 0.1-0.4 radians respectively.

Other parameters such as a step-size used in the integrator, the order of shadow Hamiltonians or the type of momentum flip upon rejection are discussed elsewhere [23, 30].

The Andersen barostat introduces two additional parameters: the mass of the *piston* μ and the reference or target pressure α. The reference pressure is used in the integrator for updating the *piston* velocity (see Section 2.2), as well as in the additional potential energy term in the Hamiltonian (9). When simulating biological experiments this pressure is commonly set to 1 bar.

μ represents the inertial mass of the extended coordinate and has a strong influence on the performance of the barostat. Figure 1 shows the effect of μ on the amplitude and frequency of the total energy of the villin system. Small *piston* masses can lead to wild oscillations in volume that could not only cause stability problems but also keep simulation from reaching its target pressure. But if the *piston* mass is too big then the volume of the box barely changes and an NVT simulation is recovered with a pressure that very slowly tends to α. For a complete discussion on an optimal choice of μ see [1] and [25].

One of the most important advantages of using GSHMC instead of standard MD is the noticeable improvement in sampling efficiency [14, 23, 20]. To test the efficiency gain of the new NPT-GSHMC method, the distance traveled by the toxin towards the POPC bilayer in the coarse-grained system was measured. In Figure 2 the time evolution of this distance and the corresponding autocorrelation functions are shown for the three methods. In general, GSHMC methods are expected to decorrelate faster and hence sample better. In this case both NVT and NPT-GSHMC arrive together at the $\sim$2.4 nm distance (the position of the bilayer) in approximately half the time required by NPT-MD. This performance is consistent with our previous work [20].

A better way to measure the sampling efficiency is to calculate the integrated autocorrelated function *IACF* for distance d during the equilibration phase of the simulation (see for example [17]). Lower values of IACF indicate lower correlations and hence better sampling. The values obtained for this case are shown in Table 3, which correspond to the IACF for the first 5000 ps of simulation. In this particular case the integrator step-size

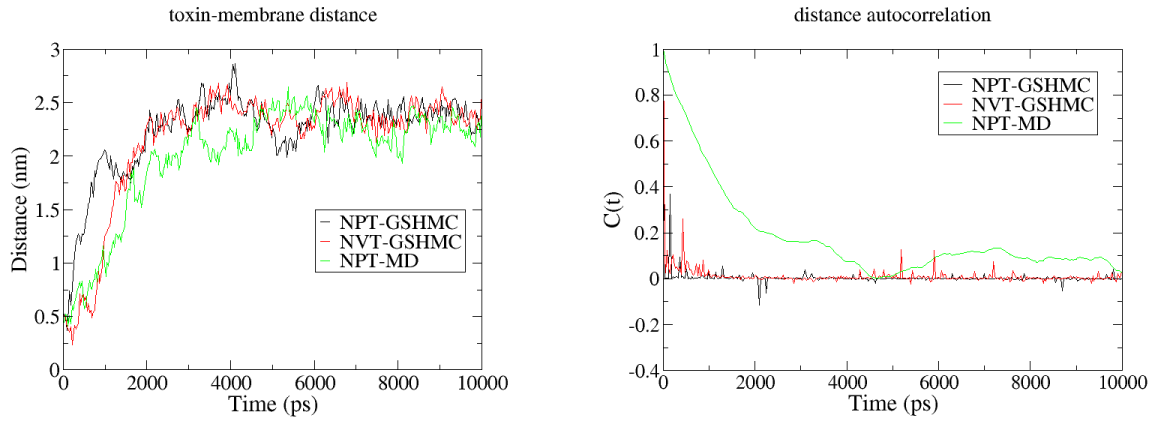


Figure 2: Comparison for the time evolution of the distance traveled by the toxin towards the membrane bilayer with the three different methods (left) and the autocorrelation function for said distance (right).

Table 3: Integrated autocorrelation for toxin-bilayer distance.

	NPT-GSHMC	NVT-GSHMC	NPT-MD
IACF	1.9	4.7	21.4

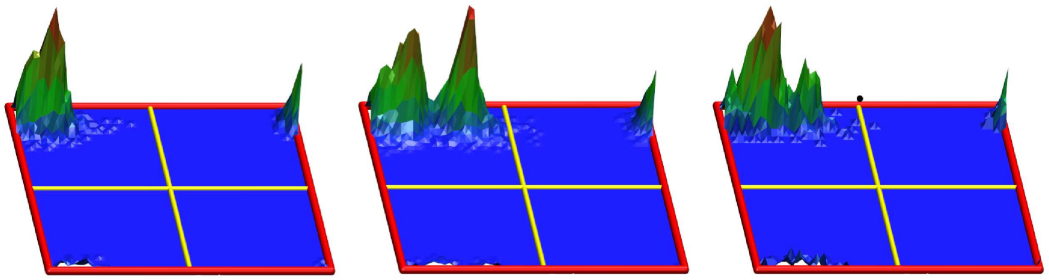
was set to 30 fs for optimal sampling efficiency. It is clear that both GSHMC methods outperform MD.

Another way to further test the sampling efficiency of the new method is to plot Ramachandran histograms [31] for the amino acid residues in the villin system. These histograms show how the $\phi - \psi$ phase space of a particular residue is explored during the simulation. As a representative example, Figure 3 compares the resulting plots for the Met13 residue extracted from a 1 ns simulations using the all three simulation techniques. One can immediately see that both GSHMC methods are exploring a larger portion of the configurational space compared with MD. Most other residues show a similar improvement in sampling efficiency and several examples have been included in the Supplementary Material.

As a final comparison, it is necessary to weight the computational expense introduced by the Andersen barostat. A way to do so is by comparing the computational times necessary to complete a 30 ns simulation of the toxin system and a 1 ns simulation of the villin system using an 8 processor node. The results in this case confirm what was measured previously for our GSHMC implementation [23]. The NVT-GSHMC method introduces on average an additional 2-4% computational overhead compared to the NPT-MD simulation. The NPT-GSHMC takes approximately the same computational time, which comes to show that our Andersen barostat implementation introduces almost no overhead and is fully compatible with the MPI parallelization in GROMACS. See Table 4 for a comparison of computational times.

4 Conclusions

The GSHMC method has been adapted to the NPT ensemble using an Andersen barostat and implemented in the open source software GROMACS. The implementation has been tested against the NPT-MD and NVT-GSHMC simulation methods available in the GROMACS-GSHMC suite [23]. NPT-GSHMC shows the same level of accuracy as demonstrated by NPT-MD and NVT-GSHMC in calculation of the thermodynamic properties of the tested



NPT-MD NPT-GSMHC NVT-GSHMC

Figure 3: Ramachandran plots for the Met13 dihedral in villin. Left: NPT-MD; Middle: NPT-GSHMC; Right: NVT-GSHMC

Table 4: Comparison of computational times for all methods.

	toxin		villin	
simulation	time (s)	ns/day	time (s)	ns/day
NPT-GSHMC	5766	749	11222	7.69
NVT-GSHMC	5747	751	11550	7.48
NPT-MD	5645	765	11087	7.79

systems, such as the toxin in a POPC bilayer and the protein villin at constant pressure and temperature.

The NPT-GSHMC method has been also proven to achieve a comparable sampling efficiency to NVT-GSHMC, as was expected from the theoretical formulation. The introduction of a barostat does not limit the benefits over MD that were previously obtained by the use of GSHMC.

The method does not introduce any noticeable computational load and is fully compatible with the highly optimized parallelization for multiple processors and threads already available in GROMACS.

In summary, all advantages offered by the generalized shadow hybrid Monte Carlo methods, such as rigorous temperature control, sampling efficiency, are now available in GROMACS for simulation of real life experiments at constant pressure and constant temperature without a loss of computational efficiency.

Acknowledgements

The authors would like to thank the financial support from MTM2011-24766 and MTM2010-18318 funded by MICINN (Spain). This work has been possible thanks to the support of the computing infrastructure of the i2BASQUE academic network and the SGI/IZO-SGIker UPV/EHU. TR would like to thank the Spanish Ministry of Education for funding through the fellowship FPU12/05209. This research is supported by the Basque Government through the BERC 2014-2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa accreditation SEV-2013-0323.

References

[1] Andersen HC (1980) Molecular dynamics simulations at constant pressure and/or temperature. J Chem Phys 72:2384-2393.

[2] Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52:7182-7190.

[3] Nosé S (1984) A unified formulation of the constant temperature molecular-dynamics methods. J Chem Phys 81(1):511-519.

[4] Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A 31(3):1695-1697.

[5] Evans DJ, Holian BL (1985) The Nose–Hoover thermostat. J Chem Phys 83:4069.

[6] Martyna GJ, Tuckerman ME, Tobias DJ, Klein ML (1996) Explicit reversible integrators for extended systems dynamics. Molec Phys 87(5):1117-1157.

[7] Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684-3690.

[8] Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4(3):435-447.

[9] Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Comm 91:43-56.

[10] Salomon-Ferrer R, Case DA, Walker RC (2013) An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci 3:198-210.

[11] Plimpton S (1995) Fast Parallel Algorithms for Short-Range Molecular Dynamics. J Comput Phys 117:1-19.

[12] Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossváry I, Moraes MA, Sacerdoti FD, Salmon JK, Shan Y, Shaw DE (2006) Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida, November 11–17.

[13] Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett B 195:216-222.

[14] Akhmatskaya E, Reich S (2008) GSHMC: An efficient method for molecular simulation. J Comput Phys 227:4934-4954.

[15] Akhmatskaya E, Reich S, Nobes R (2011) Method, apparatus and computer program for molecular simulation. US patent (granted), US007908129.-

[16] Horowitz AM (1991) A generalized guided Monte Carlo algorithm. Phys Lett B 268:247-252.

[17] Kennedy AD, Pendleton B (2001) Cost of the Generalised Hybrid Monte Carlo Algorithm for Free Field Theory. Nucl Phys B 607:456-510.

[18] Izaguirre JA, Hampton SS (2004) Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. J Comput Phys 200:581-604.

[19] Akhmatskaya E, Reich S (2010) New Hybrid Monte Carlo Methods for Efficient Sampling: from Physics to Biology and Statistics. Proceedings of the Joint International Conference of the Supercomputing in Nuclear Application and Monte Carlo, Tokyo, Japan, October 17–21.

[20] Wee CL, Sansom MS, Reich S, Akhmatskaya E (2008) Improved sampling for simulations of interfacial membrane proteins: application of generalized shadow hybrid Monte Carlo to a peptide toxin/bilayer system. J Phys Chem B 112(18):5710-5717.

[21] Faller R, de Pablo JJ (2002) Constant pressure hybrid Molecular Dynamics-Monte Carlo simulations. J Chem Phys 116:55-59.

[22] Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126:014101.

[23] Escribano B, Akhmatskaya E, Mujika JI (2013) Combining stochastic and deterministic approaches within high efficiency molecular simulations. Central European Journal of Mathematics 11(4):787-799.

[24] GROMACS Programmer's Guide, available at `http://www.gromacs.org/Developer_Zone/Programming_Guide/Programmer`

[25] Kolb A, Dünweg B (1999) Optimized constant pressure stochastic dynamics. J Chem Phys 111:4453-4459.

[26] Jung HJ, Lee JY, Kim SH, Eu YJ, Shin SY, Milescu M, Swartz KJ, Kim JL (2005) Solution Structure and Lipid Membrane Partitioning of VSTx1, an Inhibitor of the KvAP Potassium Channel. J Biochemistry 44(16):6015-6023.

[27] Bazari WL, Matsudaira P, Wallek M, Smeal T, Jakes R, Ahmed Y (1988) Villin sequence and peptide map identify six homologous domains. Proceedings of the National Academia of Sciences USA 85(14):4986–4990.

[28] Wallace E, Sansom M (2007) Carbon Nanotube/Detergent Interactions via Coarse-Grained Molecular Dynamics. Nano Lett 7(7):1923-1928.

[29] Shih A, Arkhipov A, Freddolino P, Schulten K (2006) Coarse Grained Protein-Lipid Model with Application to Lipoprotein Particles. J Phys Chem B 110(8):3674-3684.

[30] Wagoner JA, Pande VS (2012) Reducing the effect of Metropolization on mixing times in molecular dynamics simulations. J Chem Phys 137:214105.

[31] Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Molec Biol 7:95-99.