

# Adaptive splitting integrators for enhancing sampling efficiency of modified Hamiltonian Monte Carlo methods in molecular simulation

Elena Akhmatskaya,<sup>\*,†,‡</sup> Mario Fernández-Pendás,<sup>†</sup> Tijana Radivojević,<sup>†</sup> and J.M. Sanz-Serna<sup>¶</sup>

*BCAM - Basque Center for Applied Mathematics, IKERBASQUE, Basque Foundation for Science, and Universidad Carlos III de Madrid*

E-mail: [akhmatskaya@bcamath.org](mailto:akhmatskaya@bcamath.org)

## Abstract

The modified Hamiltonian Monte Carlo (MHMC) methods, i.e. importance sampling methods that use modified Hamiltonians within a Hybrid Monte Carlo (HMC) framework, often outperform in sampling efficiency standard techniques such as molecular dynamics (MD) and HMC. The performance of MHMC may be enhanced further through the rational choice of the simulation parameters and by replacing the standard Verlet integrator with more sophisticated splitting algorithms. Unfortunately, it is not easy to identify the appropriate values of the parameters that appear in those algorithms. We propose a technique, that we call MAIA (Modified Adaptive Integration

---

\*To whom correspondence should be addressed

<sup>†</sup>BCAM - Basque Center for Applied Mathematics, Alameda de Mazarredo 14, E-48009 Bilbao, Spain

<sup>‡</sup>IKERBASQUE, Basque Foundation for Science, María Díaz de Haro 3, E-48013 Bilbao, Spain

<sup>¶</sup>Departamento de Matemáticas, Universidad Carlos III de Madrid, Avenida de la Universidad 30, E-28911 Leganés (Madrid), Spain

This material is excerpted from a work that was accepted for publication in Langmuir, copyright © American Chemical Society after peer review. To access the final edited and published work see <http://pubs.acs.org/articlesonrequest/AOR-78SGGfwCHG35ZWZkkSBJ>

Approach), which, for a given simulation system and a given time step, automatically selects the optimal integrator within a useful family of two-stage splitting formulas. Extended MAIA (or e-MAIA) is an enhanced version of MAIA, which additionally supplies a value of the method-specific parameter that, for the problem under consideration, keeps the momentum acceptance rate at a user-desired level. The MAIA and e-MAIA algorithms have been implemented, with no computational overhead during simulations, in MultiHMC-GROMACS, a modified version of the popular software package GROMACS. Tests performed on well-known molecular models demonstrate the superiority of the suggested approaches over a range of integrators (both standard and recently developed), as well as their capacity to improve the sampling efficiency of GSHMC, a noticeable method for molecular simulation in the MHMC family. GSHMC combined with e-MAIA shows a remarkably good performance when compared to MD and HMC coupled with the appropriate adaptive integrators.

## Introduction

The role of numerical integrators in enhancing the performance of Hybrid/Hamiltonian Monte Carlo (HMC) has been a subject of active research in recent years.<sup>1-4</sup> It has been demonstrated that replacing the standard Verlet integrator with a splitting integrator specified by a suitable value of a parameter may significantly improve, for a range of step sizes, the conservation of the Hamiltonian and thus the acceptance rate of the proposals.<sup>1,2</sup> Such integrators however possess shorter stability limits than the familiar Verlet algorithm. In addition, the user is confronted with the problem of how best to choose the value of the parameter. The drawbacks of the use of splitting integrators more sophisticated than Verlet may be alleviated by resorting to the Adaptive Integration Approach, AIA.<sup>5</sup> For a user-chosen time step, this approach automatically identifies an optimal, system-specific integrator, by using information on the highest frequencies of the harmonic interactions present in the system. The term ‘optimal’ refers to the fact that the selected integrator minimizes, within

a family of two-stage integrators, the expectation of the energy error for harmonic forces. When stability is an issue, AIA automatically chooses the Verlet integrator and, as the time step is reduced below the Verlet limit, AIA moves to more accurate integrators.

Another way to improve the performance of HMC is to introduce importance sampling as suggested in Refs. [6-11](#). Taking advantage of the fact that symplectic integrators preserve modified Hamiltonians more accurately than they preserve the true Hamiltonian, the authors proposed to sample with respect to modified/shadow Hamiltonians and to recover the desired distribution by re-weighting. The resulting algorithms are capable of maintaining high acceptance rates and usually exhibit better efficiency than their predecessor HMC. [11-14](#) Moreover, in many applications, the Verlet integrator is sufficient to provide an amount of accepted proposals adequate to generate good statistics, even with parameter settings for which HMC fails. Thus, the methods seem to be less sensitive than HMC to the choice of numerical integrator. However, it was shown that, for importance sampling HMC applied to high-dimensional statistical problems, replacing Verlet with optimised two-stage splitting integrators can improve the observed sampling efficiency by a factor of up to 4. [11](#) That reference, however, does not offer a recipe for the rational choice of the integrator for a given system and step size.

In this paper we propose an adaptive integration approach, MAIA, which extends the ideas of AIA, in order to automatically select, for a given system and step size, the integrator with optimal conservation of the modified Hamiltonian, and therefore with the highest acceptance rate, in modified Hamiltonian Monte Carlo methods. Extended MAIA (or e-MAIA) offers the extra feature of a control on the stochasticity introduced in the momentum refreshment.

# Modified Hamiltonian Monte Carlo methods

The family of modified Hamiltonian Monte Carlo (MHMC) methods consists of HMC algorithms which, instead of sampling from the target canonical distribution

$$\pi(\mathbf{q}, \mathbf{p}) \propto \exp(-\beta H(\mathbf{q}, \mathbf{p})) \quad (1)$$

( $\beta = 1/k_B T$  is the inverse temperature and  $k_B$  the Boltzmann constant), known up to a multiplicative constant, sample from an auxiliary importance canonical density

$$\tilde{\pi}(\mathbf{q}, \mathbf{p}) \propto \exp\left(-\beta \tilde{H}^{[k]}(\mathbf{q}, \mathbf{p})\right), \quad (2)$$

where  $H$  is a Hamiltonian and  $\tilde{H}^{[k]}$  denotes a truncated modified Hamiltonian to be described later. The familiar notations  $\mathbf{q}$  and  $\mathbf{p}$  are used for positions and momenta respectively. Such methods take advantage of two facts in order to enhance sampling efficiency of HMC. First, the closeness of  $\tilde{H}^{[k]}$  to  $H$  makes it possible to implement an importance sampling approach and use samples of  $\tilde{\pi}$  as a means towards computing expectations with respect to  $\pi$ . Second, the fact that the integrator preserves  $\tilde{H}^{[k]}$  better than it does preserve  $H$  leads to a more favourable value of the acceptance probability in the algorithms.

Symplectic integrators for the Hamiltonian dynamics with Hamiltonian function  $H(\mathbf{q}, \mathbf{p})$ , while not preserving the value of  $H$  exactly along the computed trajectory, do preserve exactly the value of a so-called modified Hamiltonian<sup>15–17</sup>

$$\tilde{H} = H + \Delta t H_2 + \Delta t^2 H_3 + \dots,$$

where  $\Delta t$  is the integration time step. For an integrator of order  $p$ ,  $\tilde{H} = H + \mathcal{O}(\Delta t^p)$ , so that  $H_2, \dots, H_p$  vanish. In (2),  $\tilde{H}^{[k]}$ ,  $k > p$ , is the truncation of  $\tilde{H}$  given by

$$\tilde{H}^{[k]} = H + \Delta t^p H_{p+1} + \dots + \Delta t^{k-1} H_k.$$

The expectation of the increment of  $H$  in an integration leg satisfies

$$\mathbb{E}_\pi[\Delta H] = \mathcal{O}(D\Delta t^{2p}), \quad (3)$$

( $D$  is the dimension), while

$$\mathbb{E}_{\tilde{\pi}}[\Delta \tilde{H}^{[k]}] = \mathcal{O}(D\Delta t^{2k}), \quad (4)$$

with  $k > p$  (see Ref. <sup>18</sup> and section **MAIA** below) and therefore MHMC algorithms may benefit from high acceptance rates due to the better conservation of  $\tilde{H}^{[k]}$ .

The objective of a modified Hamiltonian Monte Carlo method is to sample from a distribution with probability density function

$$\pi(\mathbf{q}) \propto \exp(-\beta U(\mathbf{q})).$$

This is achieved indirectly, through sampling from the modified distribution (2). In this paper we consider Hamiltonians of the form

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \mathbf{p}^T M^{-1} \mathbf{p} + U(\mathbf{q}),$$

where  $M$  is a positive definite mass matrix and  $U$  is the potential, so that, under the target (1), the variable  $\mathbf{q}$  has the marginal density  $\propto \exp(-\beta U(\mathbf{q}))$ .

Since in MHMC methods the samples are generated with respect to the modified or importance density, the computation of averages with respect to the target density after completion of the sampling procedure requires re-weighting. If  $\Omega_n$ ,  $n = 1, 2, \dots, N$ , are the values of an observable along a sequence of states  $(\mathbf{q}^n, \mathbf{p}^n)$  drawn from  $\tilde{\pi}$  (2), the averages with respect to  $\pi$  (1) are calculated as

$$\langle \Omega \rangle = \frac{\sum_{n=1}^N w_n \Omega_n}{\sum_{n=1}^N w_n},$$

where the importance weights are given by

$$w_n = \exp\left(-\beta(H(\mathbf{q}^n, \mathbf{p}^n) - \tilde{H}^{[k]}(\mathbf{q}^n, \mathbf{p}^n))\right).$$

If the target density  $\pi$  and the importance density  $\tilde{\pi}$  were not close, one would typically encounter high variability among weights, which would lead to large errors in the expectation, as many samples would not contribute significantly to the computation of  $\langle \Omega \rangle$ .

Let us now describe a generic MHMC algorithm. Given a sample  $(\mathbf{q}, \mathbf{p})$  from the distribution  $\tilde{\pi}$ , the next sample  $(\mathbf{q}^{\text{new}}, \mathbf{p}^{\text{new}})$  is defined as follows:

- Obtain the new momentum  $\mathbf{p}^*$  by applying a momentum update procedure that preserves the importance density  $\tilde{\pi}$ .
- Generate a proposal  $(\mathbf{q}', \mathbf{p}')$  by simulating the Hamiltonian dynamics with Hamiltonian function  $H$  and with initial condition  $(\mathbf{q}, \mathbf{p}^*)$  using a symplectic and reversible numerical integrator.
- Choose the next sample  $(\mathbf{q}^{\text{new}}, \mathbf{p}^{\text{new}})$  to be  $(\mathbf{q}', \mathbf{p}')$  (acceptance) with probability

$$\alpha = \min\left\{1, \frac{\tilde{\pi}(\mathbf{q}', \mathbf{p}')}{\tilde{\pi}(\mathbf{q}, \mathbf{p}^*)}\right\}. \quad (5)$$

Otherwise (rejection) set  $(\mathbf{q}^{\text{new}}, \mathbf{p}^{\text{new}}) = (\mathbf{q}, -\mathbf{p}^*)$ , i.e. carry out a momentum flip.

Since

$$\frac{\tilde{\pi}(\mathbf{q}', \mathbf{p}')}{\tilde{\pi}(\mathbf{q}, \mathbf{p}^*)} = \exp\left(-\beta\left(\tilde{H}^{[k]}(\mathbf{q}', \mathbf{p}') - \tilde{H}^{[k]}(\mathbf{q}, \mathbf{p}^*)\right)\right) = \exp\left(-\beta\Delta\tilde{H}^{[k]}(\mathbf{q}, \mathbf{p}^*)\right),$$

one may expect, in view of (3)-(4), fewer rejections/momentum flips, and thus better sampling/more accurate dynamics when sampling with  $\tilde{H}^{[k]}$  instead of  $H$ .

The first methods of the MHMC class were derived for atomistic simulations and differed from each other in the ways of refreshing the momentum, computing modified Hamil-

tonians and integrating the Hamiltonian dynamics. For example, in (Separable) Shadow Hybrid Monte Carlo<sup>6,7</sup> methods, a full momentum update is used, whereas in Targeted Shadow Hybrid Monte Carlo<sup>19</sup> and Generalized Shadow Hybrid Monte Carlo (GSHMC),<sup>8</sup> suitable modifications of the partial momentum update of Horowitz<sup>20</sup> are advocated in order to better mimic the dynamics and enhance sampling. More recent MHMC methods aim at specific applications, such as multi-scale and meso-scale simulations (MTS-GSHMC and meso-GSHMC respectively)<sup>9,10</sup> and computational statistics (Mix&Match Hamiltonian Monte Carlo).<sup>11</sup> As demonstrated in the original papers, for some particular problems, the use of MHMC methods resulted in a sampling efficiency several times higher than that observed with the conventional sampling techniques, such as MD, Monte Carlo (MC) and HMC. Further improvements can be achieved through the use of adaptive integration schemes as indicated in Ref.<sup>5</sup> The proposed in this work MAIA is an adaptive approach which can successfully replace the Verlet integrator in MHMC techniques with more powerful integration schemes.

In this paper we follow the momentum update procedure used in GSHMC,<sup>8</sup> which is based on ideas from Refs.<sup>20,21</sup> We generate trial values

$$\begin{aligned}\mathbf{p}^{\text{trial}} &= \cos \varphi \mathbf{p} + \sin \varphi \mathbf{u} \\ \mathbf{u}^{\text{trial}} &= -\sin \varphi \mathbf{p} + \cos \varphi \mathbf{u}\end{aligned}\tag{6}$$

where  $\varphi \in (0, \pi/2]$  is a parameter and the noise vector  $\mathbf{u}$  is drawn from the normal distribution  $\mathcal{N}(0, \beta^{-1}M)$ . A low value of  $\varphi$  will result in  $\mathbf{p}^{\text{trial}}$  being close to  $\mathbf{p}$  and the behaviour of the algorithm will be close to conventional MD. For  $\varphi$  near  $\pi/2$ ,  $\mathbf{p}^{\text{trial}}$  will be very different from  $\mathbf{p}$ . The proposed trial momentum  $\mathbf{p}^{\text{trial}}$  is accepted ( $\mathbf{p}^* = \mathbf{p}^{\text{trial}}$ ) with probability

$$\alpha_p = \min \left\{ 1, \frac{\hat{\pi}(\mathbf{q}, \mathbf{p}^{\text{trial}}, \mathbf{u}^{\text{trial}})}{\hat{\pi}(\mathbf{q}, \mathbf{p}, \mathbf{u})} \right\}\tag{7}$$

where  $\hat{\pi}$  is the extended p.d.f.

$$\hat{\pi}(\mathbf{q}, \mathbf{p}, \mathbf{u}) \propto \exp(-\beta \hat{H}(\mathbf{q}, \mathbf{p}, \mathbf{u}))$$

corresponding to the extended Hamiltonian

$$\hat{H}(\mathbf{q}, \mathbf{p}, \mathbf{u}) = \tilde{H}^{[k]}(\mathbf{q}, \mathbf{p}) + \frac{1}{2} \mathbf{u}^T M^{-1} \mathbf{u}. \quad (8)$$

In case of rejection we set  $\mathbf{p}^* = \mathbf{p}$ .

In the present study, we use splitting methods to integrate the Hamiltonian dynamics. We use the symbols  $A$  and  $B$  to refer to the split systems

$$\dot{\mathbf{q}} = \mathbf{0}, \quad \dot{\mathbf{p}} = -\nabla U(\mathbf{q})$$

and

$$\dot{\mathbf{q}} = M^{-1} \mathbf{p}, \quad \dot{\mathbf{p}} = \mathbf{0},$$

respectively. These have solution flows

$$\phi_t^A(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, \mathbf{p} - t \nabla U(\mathbf{q})) \quad (9)$$

and

$$\phi_t^B(\mathbf{q}, \mathbf{p}) = (\mathbf{q} + t M^{-1} \mathbf{p}, \mathbf{p}). \quad (10)$$

We limit our attention to the family of two-stage integrators where the map that advances the solution over a step of length  $\Delta t$  is defined as

$$\psi_{\Delta t} = \phi_{b\Delta t}^B \circ \phi_{\Delta t/2}^A \circ \phi_{(1-2b)\Delta t}^B \circ \phi_{\Delta t/2}^A \circ \phi_{b\Delta t}^B. \quad (11)$$

Here  $0 < b < 1/2$  is a parameter that specifies the individual integrator within the family.

These integrators are symplectic, as compositions of flows of Hamiltonian systems, and reversible due to their palindromic structure. When  $b = 1/4$ , a step of length  $\Delta t$  of the integrator in (11) is equivalent to two successive steps of the velocity Verlet integrator each of length  $\Delta t/2$ .

When the error of the integrator is measured with respect to the modified Hamiltonian, the choice of parameter leading to the minimum error method is  $b \approx 0.2306$ ; <sup>11</sup> we shall use the acronym M-ME (modified minimum error) to refer to this value. A minimization procedure parallel to the one used in Ref. <sup>2</sup> suggests the value  $b \approx 0.2380$  <sup>11</sup> that we shall call M-BCSS. Due to well known properties of the Verlet method, the choice  $b = 1/4$  yields the longest linear stability interval achievable in the family (11). Therefore, generally speaking, one would select  $b$  close to  $1/4$  if the value of  $\Delta t$  is so large that stability is the main concern and choose lower values of  $b$ , allowing for more accuracy in the conservation of the modified Hamiltonian, when the user is prepared to operate with smaller values of  $\Delta t$ .

In the next section we introduce algorithms which allow a rational choice of a value  $b$  in order to achieve the best accuracy of integration in a simulation of a given problem.

In our analysis we shall rely on the computationally efficient procedures for the calculation of modified Hamiltonians for the integrators of the family (11) recently proposed in Ref. <sup>11</sup>. The modified Hamiltonians are evaluated through numerical time derivatives  $\nabla \dot{U}(\mathbf{q})$  of the gradient of the potential function  $U$ , which are computed from quantities available during the simulation. This makes feasible the use of those families of integrators in MHMC. For the two-stage family (11) of interest here the 4th order modified Hamiltonian is given by: <sup>11</sup>

$$\tilde{H}^{[4]}(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \mathbf{p}^T M^{-1} \mathbf{p} + U(\mathbf{q}) + \Delta t^2 \left( \lambda \mathbf{p}^T M^{-1} \nabla \dot{U}(\mathbf{q}) + \mu \nabla U(\mathbf{q})^T M^{-1} \nabla U(\dot{\mathbf{q}}) \right), \quad (12)$$

with

$$\lambda = \frac{6b - 1}{24}, \quad \mu = \frac{6b^2 - 6b + 1}{12}.$$

# Adaptive algorithms

In this section we describe two adaptive algorithms: the Modified Adaptive Integration Approach (MAIA) and the extended MAIA (e-MAIA).

## MAIA

MAIA is an algorithm which adapts the parameter  $b$  in (11) to the problem being solved and the value of  $\Delta t$  chosen by the user so as to maximize the expected acceptance rate  $\alpha$  of the proposal  $(\mathbf{q}', \mathbf{p}')$  in (5) or, equivalently, to minimize the expectation of the energy error function

$$\Delta \tilde{H}^{[4]}(\mathbf{q}, \mathbf{p}^*) = \tilde{H}^{[4]}(\mathbf{q}', \mathbf{p}') - \tilde{H}^{[4]}(\mathbf{q}, \mathbf{p}^*).$$

The analysis is based on a detailed study of the one-dimensional harmonic oscillator with potential  $U(q) = (k/2)q^2$  ( $k > 0$  a constant). For an integrator of the family (11), the modified Hamiltonian in (12) takes the form

$$\tilde{H}^{[4]}(q, p) = \frac{1}{2} \frac{p^2}{M} + \frac{1}{2} k q^2 + \Delta t^2 \lambda \frac{k}{M^2} p^2 + \Delta t^2 \mu \frac{k^2}{M} q^2. \quad (13)$$

As it is shown in the section **S1** in the supporting information, if  $\omega = \sqrt{k/M}$  is the angular frequency of the harmonic oscillator and  $h$  denotes the non-dimensional step size  $h = \omega \Delta t$ , then for the expected  $\Delta \tilde{H}^{[4]}$  it holds

$$0 \leq \mathbb{E}[\Delta \tilde{H}^{[4]}] \leq \frac{1}{\beta} \rho(h, b), \quad (14)$$

where  $\rho$  is the function

$$\rho(h, b) = \frac{h^8 \left( b(12 + 4b(6b - 5)) + b(1 + 4b(3b - 2))h^2 - 2 \right)^2}{4(2 - bh^2)(4 + (2b - 1)h^2)(2 + b(2b - 1)h^2)(12 + (6b - 1)h^2)(6 + (1 + 6(b - 1)b)h^2)}. \quad (15)$$

Note that the expectation  $\mathbb{E}[\Delta \tilde{H}^{[4]}]$  is taken with respect to the probability  $\tilde{\pi}$  (2) sampled

by the algorithm. For a model consisting of  $D$ , possibly coupled, harmonic oscillators with angular frequencies  $\omega_i$ ,  $i = 1, \dots, D$ , the bound becomes

$$\mathbb{E}[\Delta\tilde{H}^{[4]}] \leq \sum_{i=1}^D \frac{1}{\beta} \rho(h_i, b),$$

with  $h_i = \omega_i \Delta t$ . Minimization of the right hand-side will therefore ensure optimal conservation of the modified Hamiltonian in the harmonic model.

In MAIA, given a physical problem which includes non-harmonic forces and a value of  $\Delta t$ , we estimate the fastest of the angular frequencies,  $\tilde{\omega}$ , of the two-body interactions<sup>5</sup> and compute the nondimensional quantity

$$\tilde{h} = \sqrt{3} \tilde{\omega} \Delta t \tag{16}$$

( $\sqrt{3}$  is a safety factor to be discussed presently). We then find the value of  $b$  that minimizes

$$\max_{0 < h < \tilde{h}} \rho(h, b). \tag{17}$$

Note that  $(0, \tilde{h})$  is the shortest interval that contains all the values  $h_i = \sqrt{3} \omega_i \Delta t$ , where  $\omega_i$  are the frequencies in the problem. In contrast to AIA,<sup>5</sup> where the factor of  $\sqrt{2}$  had to be used to avoid resonances of up to 4th order,<sup>22</sup> in MAIA, the factor  $\sqrt{3}$ , covering resonances of up to 5th order, was found to be appropriate.

The MAIA algorithm can be summarised as follows:

Given a physical system and a value of  $\Delta t$ , the MAIA algorithm determines the value of the parameter  $b$  to be used in (11) in the following way:

1. Find the periods or frequencies of all two-body interactions in the system. Determine the minimum period  $\tilde{T} = 2\pi/\tilde{\omega}$ , with the fastest frequency  $\tilde{\omega}$ , and compute the nondimensional quantity  $\tilde{h}$  in (16).
2. Check whether  $\tilde{h} < 2\sqrt{2}$ , which is the usual stability limit in molecular simulation for

Verlet integrators.<sup>23</sup> If not, there is no value of  $b$  for which the scheme (11) is stable for the attempted step size  $\Delta t$  and the integration is aborted.

3. Find the optimal value of the parameter  $b$  by minimizing (17) with the help of an optimization routine.

When  $\Delta t$  is ‘large’ for the problem at hand, in the sense that stability is the primary concern, MAIA will choose  $b = 1/4$ , i.e. the Verlet integrator. Smaller values of  $\Delta t$  allow MAIA to reduce  $b$  and increase accuracy in the conservation of the modified Hamiltonian (see Figure 1). Figure 1 also shows the advantage of MAIA when compared with the older algorithm AIA,<sup>5</sup> developed for the HMC method, which does not use modified Hamiltonians and samples with respect to the target canonical density.

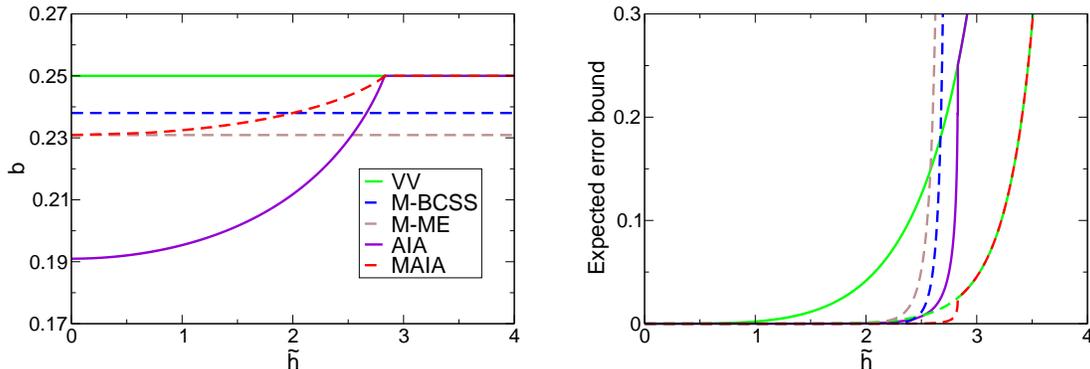


Figure 1: Parameter  $b$  for different integrators as a function of  $\tilde{h}$  (left) and bounds of the expected energy error measured with respect to the true, in solid lines, or modified Hamiltonian, in dashed lines (right). There are two lines for VV, as it may be used to sample from the true (HMC) or the importance density (GSHMC). AIA operates with respect to the true energy and MAIA with respect to its modified counterpart. Clearly the algorithms that operate with modified Hamiltonians possess smaller expected errors. This explains why, in general, VV GSHMC has higher acceptance rates than VV HMC and MAIA improves on AIA.

## e-MAIA

The overall performance of an MHMC method depends not only on the acceptance rate  $\alpha$  of the proposal made after each MD integration leg (see (5)), but also on the momentum update acceptance rate  $\alpha_p$  in (7). The latter may play an important role in the quality of sampling.<sup>8,13</sup> So far, we have looked for the integrator that maximizes  $\alpha$  and our next objective is to find a way to control  $\alpha_p$ .

As we did above, we build the analysis on the use of a harmonic oscillator model. For the scalar harmonic potential, the stationary marginal p.d.f.'s of the (stochastically independent variables)  $p$  and  $u$  (see (6)) are

$$\pi(p) \propto \exp\left(-\beta\left(\frac{1}{2}\frac{p^2}{M} + \Delta t^2 \lambda \frac{k}{M^2} p^2\right)\right), \quad \pi(u) \propto \exp\left(-\frac{\beta}{2} \frac{u^2}{M}\right), \quad (18)$$

respectively, and the extended Hamiltonian in (8) reads

$$\hat{H}(q, p, u) = \tilde{H}^{[4]}(q, p) + \frac{1}{2} \frac{u^2}{M},$$

with  $\tilde{H}^{[4]}$  given in (13). As it was shown in Ref.,<sup>11</sup> the difference in extended Hamiltonian satisfies

$$\begin{aligned} \Delta \hat{H} &= \hat{H}(q, p^{\text{trial}}, u^{\text{trial}}) - \hat{H}(q, p, u) \\ &= \Delta t^2 \lambda \left( \sin^2 \varphi \left( \frac{k}{M^2} u^2 - \frac{k}{M^2} p^2 \right) + 2 \cos \varphi \sin \varphi u \frac{k}{M^2} p \right), \end{aligned} \quad (19)$$

and from here it is found that

$$\mathbb{E}[\beta \Delta \hat{H}] = \Delta t^2 \beta \lambda \sin^2 \varphi \frac{\omega^2}{M} (\mathbb{E}[u^2] - \mathbb{E}[p^2]).$$

From (18) we have

$$\mathbb{E}[p^2] = \beta^{-1} M (1 + 2\Delta t^2 \lambda \omega^2)^{-1}, \quad \mathbb{E}[u^2] = \beta^{-1} M,$$

and then

$$\begin{aligned}\mathbb{E}[\beta\Delta\hat{H}] &= \Delta t^2\lambda\sin^2\varphi\omega^2\left(1 - (1 + 2\Delta t^2\lambda\omega^2)^{-1}\right) \\ &= \frac{2\Delta t^4\lambda^2\sin^2\varphi\omega^4}{1 + 2\Delta t^2\lambda\omega^2}.\end{aligned}$$

In terms of the dimensionless time step  $h = \omega\Delta t$ , one obtains

$$\mathbb{E}[\beta\Delta\hat{H}] = \frac{2h^4\lambda^2\sin^2\varphi}{1 + 2h^2\lambda}. \quad (20)$$

For the model consisting of  $D$  harmonic oscillators with angular frequencies  $\omega_i$ ,  $i = 1, \dots, D$ , the equivalent of (20) is

$$\mathbb{E}[\beta\Delta\hat{H}] = \sum_{i=1}^D \frac{2h_i^4\lambda^2\sin^2\varphi}{1 + 2h_i^2\lambda} \geq D \frac{2\bar{h}^4\lambda^2\sin^2\varphi}{1 + 2\bar{h}^2\lambda}, \quad (21)$$

where  $h_i = \omega_i\Delta t$  are the dimensionless time steps and  $\bar{h} = \bar{\omega}\Delta t$  with  $\bar{\omega}$  equal to the slowest angular frequency among all the oscillators.

Using  $\alpha_p \leq \exp(-\beta\Delta\hat{H})$  (see (7)), from the inequality (21) and for a concrete choice of the angle  $\varphi_p$ , we can find the approximation

$$-\frac{\log \mathbb{E}[\alpha_p]}{D} \approx \frac{2\bar{h}^4\lambda^2\sin^2\varphi_p}{1 + 2\bar{h}^2\lambda}. \quad (22)$$

It has to be remarked that the fastest oscillation frequency features in the analyses of MAIA and in its predecessor AIA<sup>5</sup> but the *slowest* frequency is used in (22).

From (22), the expected acceptance rate in the momentum update may be controlled by three parameters: the parameter  $\lambda = \lambda(b)$  that depends on the specific integrator being used, the parameter  $\bar{h}$ , which for a given problem is a function of  $\Delta t$ , and the angle  $\varphi$ . This fact motivates the algorithm that we call extended MAIA or e-MAIA. For a user-chosen  $\Delta t$ , e-MAIA first finds an integrator within the family of two-stage schemes that maintains the smallest expected modified energy error in the molecular dynamics part of the MHMC algorithm and then adjusts the value of  $\varphi$  so as to achieve a desired acceptance rate for the

momentum update step. As it is explained above, the acceptance rates in the momentum update step depend on the choice of an angle  $\varphi$  whereas the MAIA analysis does not depend on  $\varphi$ . This means that, for some fixed values of  $\varphi$  and  $\Delta t$ , the nominated by MAIA integrator may be not favourable for maintaining an appropriate acceptance rate in the momenta. The goal of e-MAIA is to provide an adaptive choice of the angle  $\varphi$  to achieve a target, user-specified acceptance rate in the momentum update step while keeping the highest acceptance rate for positions.

While a high acceptance rate in the MD part has a positive effect on sampling with modified Hamiltonians, a too frequent acceptance of momentum (close to 100 %) could lead to the negative effect of the thermalization of the simulation, deteriorating its accuracy.<sup>13,24</sup> The rationale for introducing e-MAIA is the possibility of simultaneously adapting the parameters  $b$  and  $\varphi$  to control both the acceptance probabilities  $\alpha$  and  $\alpha_p$  of the MD integration legs and the momentum updates.

The algorithm e-MAIA is as follows:

1. For a given physical problem, choose a time step  $\Delta t$  for the integration of the equations of motion, a target acceptance rate  $AR_p$  for the momentum update and an initial value  $\varphi_0$  of the angle  $\varphi$ .
2. Find the slowest and the fastest angular frequencies in the harmonic interactions,  $\bar{\omega}$  and  $\tilde{\omega}$  respectively.
3. The integrator parameter  $b^*$  is obtained as in MAIA by optimization of the function  $\rho$  in (15). This choice of  $b^*$  guarantees the highest possible acceptance rate for harmonic interactions in the MD step.
4. The function which bounds the expected extended Hamiltonian error is given by (see (22))

$$\tau(\bar{h}, b^*, \varphi) = \frac{2\bar{h}^4 \lambda^{*2} \sin^2 \varphi}{1 + 2\bar{h}^2 \lambda^*},$$

where  $\lambda^*$  is the value of  $\lambda$  when  $b = b^*$  and

$$\bar{h} = \bar{\omega}\Delta t. \quad (23)$$

The angle  $\varphi^*$  is chosen as

$$\varphi^* = \arg \min_{\varphi \in (0, \pi/2]} \theta(\varphi), \quad (24)$$

with

$$\theta(\varphi) = \left| -\frac{\log(AR_p)}{D} - \tau(\bar{h}, b^*, \varphi) \right|.$$

5. If the selected  $\varphi^*$  is smaller than  $\varphi_0$ , then either decrease the target  $AR_p$  and go to step 4 or, alternatively, define the function

$$\sigma(h, b, \varphi_0) = \rho(h, b) + \tau(h, b, \varphi_0) \quad (25)$$

and choose  $b^{**}$  that minimizes  $\max_{0 < h < \bar{h}} \sigma(h, b, \varphi_0)$ . (The fastest oscillation is used again for the momentum update part since in this case we are constructing an upper bound of the expected energy error.)

We stress that for very small values of  $\varphi$ , an MHMC method loses its extra sampling abilities and behaves similarly to standard molecular dynamics. In e-MAIA this possibility is eliminated in step 5 of the algorithm in two optional ways. One way is to keep decreasing the target  $AR_p$  till  $\varphi^*$  rises above  $\varphi_0$ . Another option is to optimize the joint bound function constructed for both expected errors,  $\mathbb{E}[\beta\Delta\tilde{H}]$  and  $\mathbb{E}[\beta\Delta\hat{H}]$ . Though this sacrifices the position acceptance rates, the expected loss is small provided that  $\varphi_0 \ll \pi/2$ .

The reader should notice that, whereas MAIA in principle works for any method that samples with respect to modified Hamiltonians, e-MAIA only works for those MHMC methods which perform the momentum update step in the way described in this paper (6).

# Implementation

## MultiHMC-GROMACS

MAIA and e-MAIA have been implemented in the MultiHMC-GROMACS<sup>5,25,26</sup> software package. MultiHMC-GROMACS is a modified version of the popular molecular simulation code GROMACS<sup>27,28</sup> and offers a set of recently developed algorithms that aim at improving the accuracy and sampling performance of the original GROMACS without sacrificing computational efficiency and parallel scaling properties. From the user’s point of view, there is no difference in setting up a simulation with MultiHMC-GROMACS and doing so with GROMACS, except for the need to specify a few additional parameters in the MultiHMC-GROMACS input file *.mdp*.

A detailed description of the MultiHMC-GROMACS package may be found elsewhere;<sup>5,25,26</sup> here we briefly review those features that are relevant to the present study.

- Hybrid Monte Carlo methods.

Hybrid Monte Carlo (HMC),<sup>29</sup> Generalized Hybrid Monte Carlo (GHMC)<sup>20,21</sup> and Generalized Shadow Hybrid Monte Carlo (GSHMC)<sup>8</sup> are all available in MultiHMC-GROMACS. The implementation of GSHMC in GROMACS has been discussed in detail in Refs.<sup>25,26</sup> The other methods, i.e. HMC and GHMC, are implemented as special cases of GSHMC as described in Ref.<sup>5</sup>

- Modified Hamiltonians.

Two types of modified Hamiltonians are currently available in MultiHMC-GROMACS: the shadow Hamiltonian in a Lagrangian formulation<sup>8</sup> for the Verlet/leapfrog integrator and the modified Hamiltonian (12)<sup>11</sup> for two-stage integrators of the family (11).

- Two-stage integrators.

The two-stage integrators of the family (11) have been implemented in MultiHMC-GROMACS as a concatenation of alternating updates of velocities and positions in the

routine *do\_md()* in *md.c*. The v-rescale thermostat<sup>30</sup> available in GROMACS has been adapted to work with the two-stage schemes.<sup>5</sup> In addition, the two-stage integrators are optionally coupled with the SHAKE algorithm for simulation of constrained dynamics.<sup>5</sup>

At present MultiHMC-GROMACS may carry out simulations with the BCSS integrator,<sup>2</sup> the McLachlan’s method,<sup>1</sup> their counterparts for modified Hamiltonians<sup>11</sup> and the velocity Verlet two-stage integrator. These integrators are given by specific choices of the parameter  $b$  in (11). In addition, the MultiHMC-GROMACS code implements the AIA integrator,<sup>5</sup> where the algorithm tunes  $b$  to the particular simulation being carried out.

## Implementing MAIA and e-MAIA

Similarly to AIA,<sup>5</sup> MAIA and e-MAIA have been implemented in the GROMACS preprocessing module *grompp*. The preprocessing module is run only once before any simulation and, thus, does not introduce computational overheads in the simulation itself.

In the original GROMACS package, the module *grompp* reads the input files that contain the essential information about the simulated system, such as topology and structure, and makes the necessary adaptation of that information for its use in the molecular dynamics module *mdrun*. The module *grompp* also checks the *.mdp* file for choices of input parameters for the simulation algorithms and, if necessary, generates warnings that allow users to reconsider the chosen setup. The *grompp* module finishes by producing the file *.tpr* to be used as input in the *mdrun* module for running molecular dynamics simulation.

In addition to these functionalities, a more advanced analysis of the harmonic interactions is included in *grompp* in MultiHMC-GROMACS. As it has been explained in Ref.,<sup>5</sup> the fastest harmonic interaction predetermines a maximal step size allowed for the stable numerical integration of the equations of motion. On the other hand, the slowest harmonic interactions are used in the e-MAIA algorithm to identify the best choice of the parameter  $\varphi$ . In MultiHMC-GROMACS, *grompp* searches for the periods corresponding to the fastest

and slowest oscillations,  $\tilde{T}$  and  $\bar{T}$  respectively. The value  $\tilde{T}$  is used to define the upper limit of the dimensionless time step,  $\tilde{h} = \sqrt{3}(2\pi/\tilde{T})\Delta t$ , following MAIA algorithm. The optimal value of the parameter  $b$  for a MAIA or e-MAIA integrator is then found as the argument that minimizes the maximum of  $\rho$  (15) for the range of dimensionless time steps from zero to  $\tilde{h}$ . As in Ref.,<sup>5</sup> the minimization is performed with a particle swarm optimization algorithm driven by a golden section search.<sup>31</sup> The value  $\bar{T}$  is used to determine the angle  $\varphi$ , as explained in the e-MAIA algorithm.

Both  $b$  and  $\varphi$  are stored in the *input record* structure introduced by GROMACS for keeping all the input data during the whole simulation. Thus,  $b$  and  $\varphi$  can be accessed from every routine in the package.

The flow chart in Figure S1 summarizes MAIA and e-MAIA algorithms.

## Numerical experiments

In order to evaluate the efficiency of the proposed (e-)MAIA algorithm, we first compared its performance with that of several integration schemes which potentially can compete with it. Then we estimated the performance of GSHMC combined with (e-)MAIA in comparison with other popular sampling methods. More precisely:

- e-MAIA was compared with fixed parameters integrators specifically derived for MHMC methods. The counterpart of BCSS<sup>2</sup> for modified Hamiltonians (M-BCSS) and the equivalent to the scheme of McLachlan<sup>1</sup> that minimizes the errors of modified Hamiltonians (M-ME) were included in the comparison. Both M-BCSS and M-ME have been recently derived<sup>11</sup> and implemented in MultiHMC-GROMACS. All three integrators were combined with the GSHMC method. In addition, e-MAIA was compared with integrators successfully used for molecular simulation in MD, HMC and GSHMC. The velocity Verlet and AIA combined with GSHMC were selected in this case.
- e-MAIA was compared with MAIA when both were implemented within GSHMC.

- GSHMC was compared with HMC and MD. For each tested sampling method the most efficient integrator was used: e-MAIA was chosen for GSHMC and AIA was employed in MD and HMC.

To provide a fair comparison, the following issues have been taken into account while producing the numerical results. To equalise the time spent on force calculations using Verlet and two-stage integrators, Verlet was always run with half a step size and twice a number of steps. Also, in the simulations with HMC and GSHMC, the number of Metropolis tests was kept constant regardless of the acceptance/rejection output. The computational overhead due to evaluation of modified Hamiltonians in GSHMC<sup>8</sup> was taken into account by normalising calculated integrated autocorrelation functions with respect to computational times.

The tests were performed using two benchmark systems (see below), both run over a range of time steps  $\Delta t$ . The aim was to monitor the evolution of the parameters  $b$  and  $\varphi$  (6) automatically chosen for each  $\Delta t$  in (e-)MAIA, and estimate their effect on the overall sampling performance of GSHMC. In all plots in section **Results**, values of time steps correspond to two-stage integrators and assume twice smaller time steps for velocity Verlet.

Different lengths of MD trajectories  $L$  in GSHMC simulations were also tested. This parameter may play an important role in the sampling efficiency of GSHMC simulations when too small or too large values are chosen.<sup>13</sup> However, for the sake of clarity, in all tests presented in this work, the length of MD trajectories was fixed to 2000 steps when two-stage integrators were used and to 4000 otherwise. These values were found to be good choices for both GSHMC and HMC with different integration schemes and this is also confirmed by findings in Ref.<sup>5</sup>

With the obvious exception of e-MAIA, the angle used for the momentum refreshment (6) was set to 0.2 for all tests unless stated otherwise.

Each individual test has been repeated 10 times and every result reported in this paper was obtained by averaging over the multiple runs to reduce statistical errors.

## Benchmarks and simulation setup

The numerical experiments were performed using two benchmark systems: the coarse-grained VSTx1 toxin in a POPC bilayer<sup>32</sup> and the atomistic 35-residue villin headpiece protein subdomain.<sup>33,34</sup> In what follows we shall refer to these systems as toxin and villin respectively.

Toxin is a coarse-grained system of 7810 particles. Four heavy particles were represented on average as one sphere.<sup>35,36</sup> Coulomb and Van der Waals interactions were solved using the shift algorithm. Both potential energies were shifted to 0 kJ mol<sup>-1</sup> at the radius of 1.2 nm. Periodic boundary conditions were considered in all directions. The target temperature was chosen to be 310 K and it was controlled in MD simulations by the v-rescale algorithm whereas no additional thermostat was required in HMC and GSHMC. No constraints were defined for this system. The total length of all simulations was 20 ns, which was sufficient for equilibration of the system for those choices of time steps that provided a stable integration.

The villin protein is a 9389 atoms system, composed of 389 atoms solvated with 3000 water molecules. Coulomb interactions were solved with the PME algorithm of order 6. Van der Waals interactions were considered as in the toxin system but with a radius of 0.8 nm. Periodic boundary conditions were defined in all directions. As in the previous study,<sup>5</sup> the bonds involving hydrogens were constrained using SHAKE/RATTLE. The use of these constraints does not affect the accuracy of simulations but allows for longer time steps which, due to the atomistic nature of the system, still are significantly shorter than in the coarse-grained toxin system. The temperature of 300 K was maintained using the v-rescale algorithm in MD and through the Metropolis tests in HMC and GSHMC.

Similarly to Ref.,<sup>5</sup> an exhaustive study of the complete folding process of the villin protein is out of the scope of this work. Instead, the aim is to show the beneficial effect of (e-)MAIA on the accuracy and performance of the simulation of a constrained atomistic system. For our purposes, the length of 5 ns for each villin experiment was found sufficient.

Earlier reported studies of the villin system have suggested that the use of a weak coupling thermostat and a barostat may lead to a better agreement with experiments.<sup>37</sup> Barostats

are not considered, however, in this study, since the primary target of the algorithms presented here are modified Hamiltonian Monte Carlo methods, which if no extra variations are introduced,<sup>8,26</sup> sample in the NVT ensemble.

## Results

### Toxin

We start by measuring the acceptance rates of positions and momenta in the GSHMC simulations with different integration schemes. For the sake of clarity, we excluded from the plots the results for the MAIA algorithm, leaving only the data for e-MAIA. This makes sense because the position acceptance rates for MAIA and e-MAIA are always very similar (see step 3 of the e-MAIA algorithm), while e-MAIA has an obvious advantage over MAIA as far as the acceptance rates for momenta are concerned. We shall provide more details on this issue later.

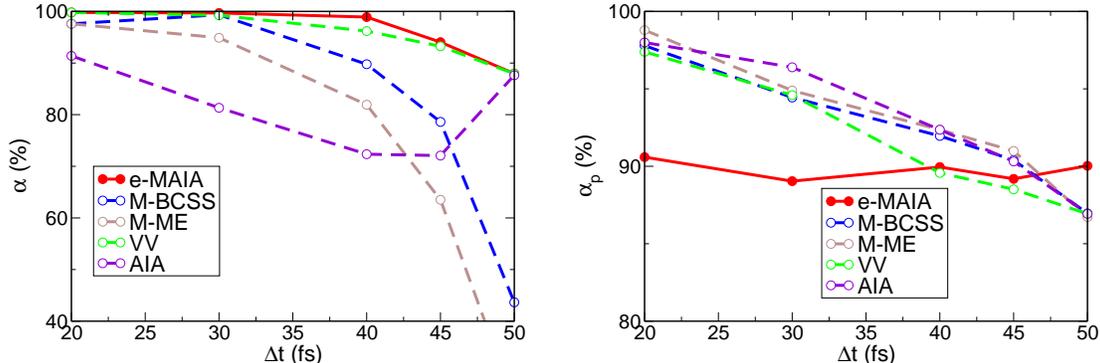


Figure 2: Toxin. Acceptance rates for positions (left) and momenta (right) observed in GSHMC simulations when using M-BCSS, M-ME, VV, AIA (all dashed lines) and e-MAIA (solid line). e-MAIA maintains the target  $AR_p$  of 90 % for each value of  $\Delta t$  (right).

The primary objective of the MAIA algorithm is to maximize the acceptance of position proposals in an MHMC method by minimizing the expected errors in modified Hamiltonians.

Then, the first natural test for MAIA is to check whether the position acceptance rates observed in GSHMC simulations combined with MAIA are not below those observed with other two-stage integrators. In Figure 2, the effect of various integrators such as e-MAIA, the modified versions of BCSS<sup>2</sup> (M-BCSS) and ME<sup>1</sup> (M-ME),<sup>11</sup> the standard VV, and AIA,<sup>5</sup> on the acceptance rates in GSHMC simulations is investigated. The trends presented in the left plot are in good agreement with the theoretical prediction in Figure 1 (right panel). Indeed, the acceptance rates obtained with the modified adaptive approach e-MAIA, over the range of time steps considered, are never lower than the ones provided by the other integrators tested. For small time steps, all integrators, except AIA, guarantee high acceptance rates, but the situation changes as the time step increases and the shorter stability intervals of M-BCSS and M-ME result in acceptance rates well below those achieved with e-MAIA and VV. The low acceptance rates for AIA are not surprising, since this method was developed for sampling with respect to the true Hamiltonian and provides the lowest expected errors in Hamiltonian rather than in modified Hamiltonian. However, for the largest time step of 50 fs, the parameter  $b$  in AIA becomes equal to  $1/4$  and thus AIA is equivalent to VV (see Figure 1 (left)). The same applies to MAIA/e-MAIA for the longest time step, as can be also seen in Figure 1 (left). This simply reflects the fact that the velocity Verlet integrator possesses the longest stability interval among the two-stage integrators and the adaptive methods AIA and MAIA select velocity Verlet when the time step goes beyond the stability limit of other two-stage integrators.

The acceptance rates for momenta are shown in the right panel of Figure 2. For e-MAIA, we fixed the target acceptance  $AR_p$  to 90 % bearing in mind that too high (near 100 %) acceptance rates may deteriorate accuracy, whereas low acceptance rates normally reduce the sampling efficiency of GSHMC. With this target set, e-MAIA chose an appropriate value of  $\varphi$  for each time step being tested. The simulations with other integrators were run with the fixed value  $\varphi = 0.2$ . This was selected to achieve a good performance for the longest time steps. Obviously, with any integrator, the parameter  $\varphi$  can be adapted, by trial and error,

to each simulation and time step, but we have to stress that, in practice, tuning blindly the value of  $\varphi$  is rather time consuming and not necessarily results in the optimal choice of  $\varphi$ . That is why the ability of e-MAIA to optimize automatically such a choice is very welcome. As follows from Figure 2 (right), for all tested time steps, e-MAIA maintained well the target  $AR_p$  by varying  $\varphi$ . The other integrators being combined with GSHMC led to very high, unwelcome acceptance rates for most time steps tested.

We shall see next how the trends observed above for the acceptance rates impact the sampling efficiency of GSHMC. In the case of toxin, this efficiency was measured in terms of the integrated autocorrelation function IACF of the toxin drift  $d$  to the preferred interfacial location over the "convergence period". The IACF is defined as

$$\text{IACF}^\Omega = \sum_{l=0}^{K'} \text{ACF}(\tau_l), \quad (26)$$

where  $\text{ACF}(\tau_l)$ ,  $l = 0, \dots, K' < K$  is the standard autocorrelation function for the time series  $\Omega_k$  of  $K$  samples,  $k = 1, \dots, K$ .<sup>21,38</sup> For GSHMC, the ACF's are calculated taking into account the weights collected during simulations.<sup>11</sup> We notice that in all performed simulations the normalized weights are close to 1 due to small differences between modified and true Hamiltonians observed in the simulations as well as the choice of temperatures (common for molecular simulations of biological systems) leading to  $\beta < 1$ . This means that the metrics designed for weighted and non-weighted methods would not generate too different data. This however is not expected in a general case and is not common in statistical applications.<sup>39</sup> The IACF gives a quantitative measure of the time required, on average, to generate an uncorrelated sample. Low values of measured IACFs imply low correlations between samples and thus more efficient sampling.

Figure 3 (left) presents the IACFs (normalized with respect to computational time) obtained from GSHMC simulations using different integrators and time steps. Clearly, the simulations with e-MAIA provided the lowest values of IACFs and thus the best sampling

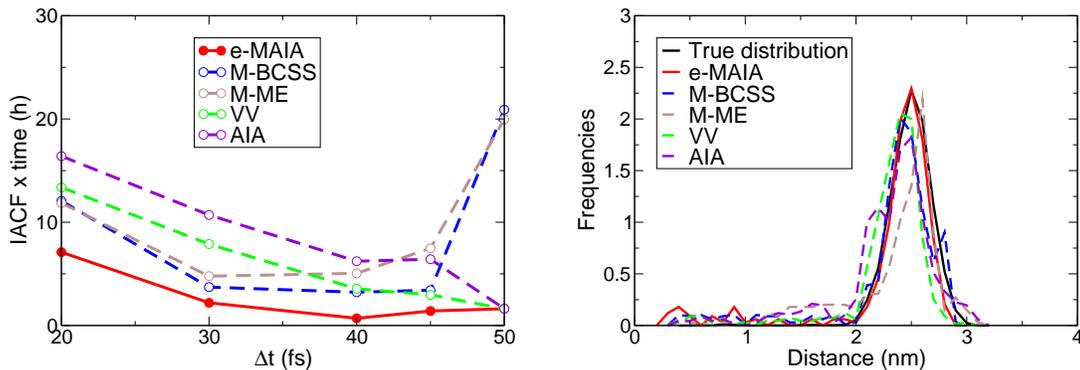


Figure 3: Toxin. Sampling efficiency of GSHMC combined with the integrators used in Figure 2. On the left, IACF of the drift,  $d$ , of the toxin to the preferred interfacial location evaluated as a function of  $\Delta t$  in GSHMC tests. On the right, the distribution of  $d$  observed in GSHMC simulations with various integrators using a time step of 30 fs. The solid black line (right) presents the “true” distribution produced with a ten times longer simulation (200 ns).

for all choices of time step. All methods showed the good performance at  $\Delta t = 40$  fs and, for this time step, the simulations with e-MAIA resulted in an efficiency (as measured by IACF) from 5 (vs. M-BCSS, VV) to 9 (vs. AIA) times higher than the simulations with other integration schemes. For the largest time step,  $\Delta t = 50$  fs, the performance achieved using e-MAIA was 12 times better than in the simulations with M-BCSS and M-ME, but it did not differ anymore from those observed in the simulations with VV and AIA, because for this long time step both AIA and e-MAIA chose velocity Verlet as integrator.

The right panel of Figure 3 compares the distributions of the distance  $d$  between the c.o.m. of the toxin and the c.o.m. of the bilayer, collected from simulations with  $\Delta t = 30$  fs with different integrators, against the “true” distribution obtained from an MD simulation with velocity Verlet, over a time interval of length 200 ns, i.e. ten times longer. As for all tests in this section, the plots have results averaged over 10 repetitive runs. The curve corresponding to the simulation with e-MAIA (in red) shows the best match with the “true” distribution (in black).

The performances of e-MAIA and MAIA are compared in Figure 4. We chose the target

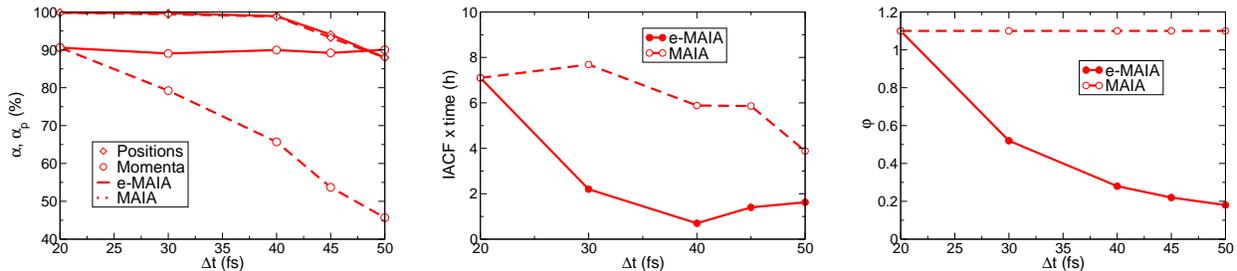


Figure 4: Toxin. e-MAIA (solid) vs. MAIA (dashed). Acceptance rates for positions and momenta (left), IACFs (center) and the angle  $\varphi$  found by e-MAIA as a function of the time step (right) observed in GSHMC simulations. The angle  $\varphi$  used in MAIA was 1.1 and the target  $AR_p$  for e-MAIA was 90 %.

$AR_p$  in e-MAIA to be 90 % and the angle  $\varphi$  in MAIA to be equal to 1.1, which was the value found by e-MAIA for achieving the target  $AR_p = 90$  % in GSHMC simulations at the smallest time step tested,  $\Delta t = 20$  fs. Figure 4 reveals that, even though both e-MAIA and MAIA find the same integrator parameter  $b$ , leading to similar acceptance rates for positions, a good choice of the angle  $\varphi$  may visibly improve the sampling performance of GSHMC. The improvement is by factors of 8 and 2 for  $\Delta t = 40$  fs and  $\Delta t = 50$  fs respectively. The evolution, as the time step increases, of the optimal parameter  $\varphi$  as calculated by e-MAIA is also shown in Figure 4.

To finalize the numerical experiments on the toxin benchmark, we compared, using the normalized IACF metrics, the performance of three sampling methods, MD, HMC and GSHMC. For each method, the best performing integrator was selected. Thus, GSHMC was combined with e-MAIA, based on the findings discussed above, whereas the AIA integrator was used for HMC and MD, according to the recommendations in Ref.<sup>5</sup> Figure 5 (left) demonstrates the superiority of GSHMC over the other two methods, regardless the choice of time step. For the optimal choice of time step for this system, namely,  $\Delta t = 40$  fs, the sampling efficiency of GSHMC is 4 times higher than that of HMC and 11 times better than that of MD. For the longest time step,  $\Delta t = 50$  fs, the difference is even more dramatic and expressed in improvement factors of 17 and 30 over HMC and MD respectively. Plotted

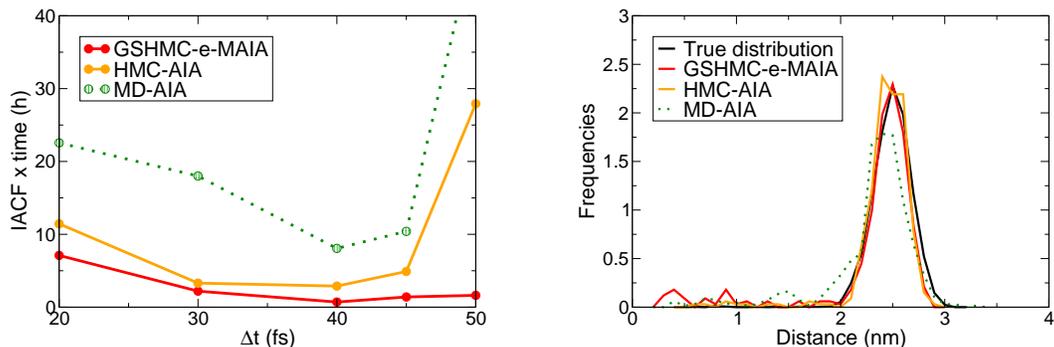


Figure 5: Toxin. Sampling efficiency: GSHMC (e-MAIA) vs. HMC (AIA) vs. MD (AIA). The best integrator for each sampling method was employed. Sampling efficiency was measured by means of IACFs (left) and the distribution of the distance between the toxin and the membrane bilayer (right). The solid black line (right) presents the “true” distribution produced with a ten times longer simulation (200 ns).

in Figure 5 (right) are the distributions of the distance  $d$  between the c.o.m. of the toxin and the c.o.m. of the bilayer produced by GSHMC, HMC and MD simulations using a time step of 30 fs; they also confirm the better convergence of the GSHMC results to the “true” distribution.

## Villin

As in the toxin case, we first inspected the acceptance rates for positions and momenta in GSHMC simulations with different integrators and found that the e-MAIA method worked as expected, i.e. provided the best position acceptance rates (Figure 6, left) and maintained the target momenta acceptance rate of 90 % (Figure 6, right) for all choices of time steps.

In contrast to the coarse-grained toxin benchmark, a quantitative analysis of the MAIA’s contribution to the GSHMC performance gain is not feasible with the atomistic villin benchmark. This is because such an analysis would require a long, computationally demanding series of simulations, for a range of time steps, integrators and sampling methods. It is however possible to find evidence of the positive impact of MAIA on the sampling efficiency of GSHMC by using comparatively short simulation runs of 5 ns and metrics directly related

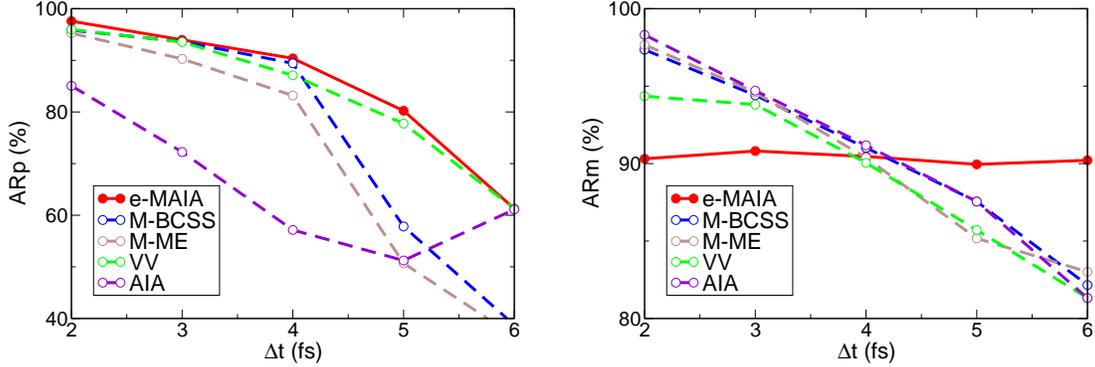


Figure 6: Villin. Acceptance rates for positions (left) and momenta (right) observed in GSHMC simulations when using M-BCSS, M-ME, VV, AIA (all dashed lines) and e-MAIA (solid line). e-MAIA maintains the target  $AR_p$  of 90 % for each value of  $\Delta t$  (right).

to the quality of sampling.

One of such metrics is the Radius of Gyration (RG), which provides an estimation of the compactness of the desired structure, and is computed as

$$\text{RG} = \left( \frac{\sum_{i=1}^n \|r_i\|^2 M_i}{\sum_{i=1}^n M_i} \right)^{1/2},$$

where  $n$  is the number of atoms in the structure,  $r_i$  the distance between atom  $i$  and the center of mass of the structure, and  $M_i$  the mass of atom  $i$ . As in Ref.,<sup>37</sup> we considered the experimental value of 0.94 nm as a target value and investigated the level of convergence to this value in short simulations when using different time steps, numerical integrators and simulation methods.

Another metrics used in this study relates to the positional root-mean-squared deviation (RMSD). The RMSD of a group of atoms in a molecule with respect to a reference structure can be calculated as

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^n \delta_i^2},$$

where  $\delta_i$  is the distance between the positions of atom  $i$  in the two structures being compared.

Following the ideas from Ref.,<sup>37</sup> we calculated the maximal RMSD of the  $\alpha$ -carbon between any two visited structures in each simulation in order to judge the level of exploration of conformational space during the simulation.

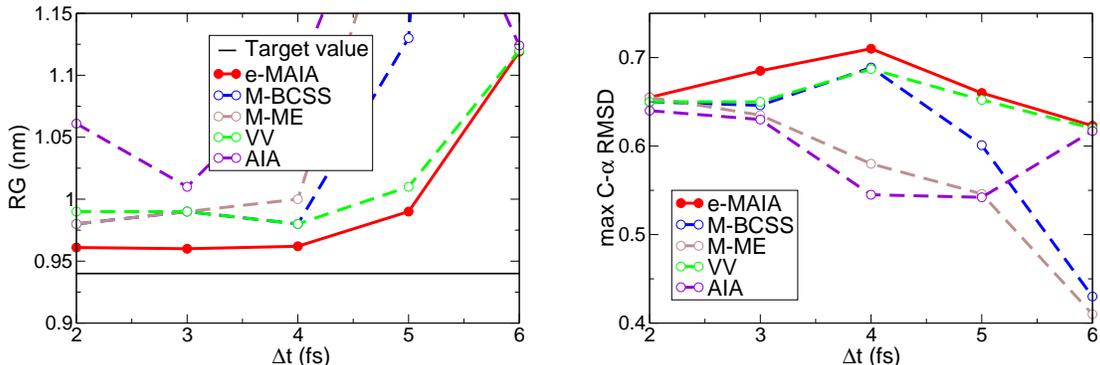


Figure 7: Villin. Sampling efficiency of GSHMC combined with the integrators used in Figure 6: radius of gyration (left) and maximum RMSD of the  $\alpha$ -carbon of the protein (right). The black solid line (left) represents the target experimental value of 0.94 nm.

In Figure 7 we plot, as functions of the time step, the radii of gyration and maximal RMSDs of the  $\alpha$ -carbon calculated from the data collected in GSHMC simulations using e-MAIA, M-BCSS, M-ME, VV and AIA integrators. Clearly, the simulations with e-MAIA (red solid line) produced the best approximations to the experimental data (left plot), the highest values of maximal RMSD (right plot) (implying better sampling) and the smallest performance degradation at the longest time steps.

The comparison of the results obtained using MAIA and e-MAIA in GSHMC simulations of villin confirmed the trends observed earlier in the toxin tests. Both methods achieved almost the same position acceptance rates, whereas the momenta acceptance rates were significantly higher in the simulations with e-MAIA (Figure 8, left). The latter was possible due to the automatic tuning of the parameter  $\varphi$  provided by e-MAIA for maintaining the target  $AR_p = 90\%$  (Figure 8, right); its positive effect can be noticed in Figure 8, center.

Figure 9 compares the radii of gyration (left) and maximal RMSDs of the  $\alpha$ -carbon (right) obtained from the simulations of villin using three different sampling methods, GSHMC,

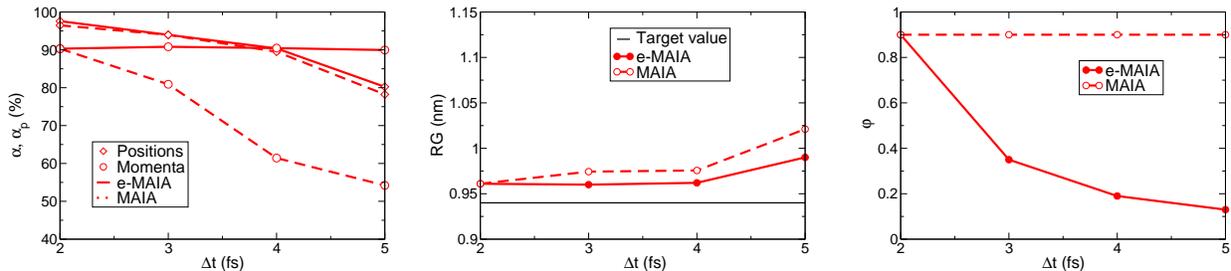


Figure 8: Villin. e-MAIA (solid) vs. MAIA (dashed). Acceptance rates for positions and momenta (left), radii of gyration (center) and the angle  $\varphi$  found by e-MAIA as a function of the time step (right) observed in GSHMC simulations. The angle  $\varphi$  used in MAIA was 0.9 and the target  $AR_p$  for e-MAIA was 90 %.

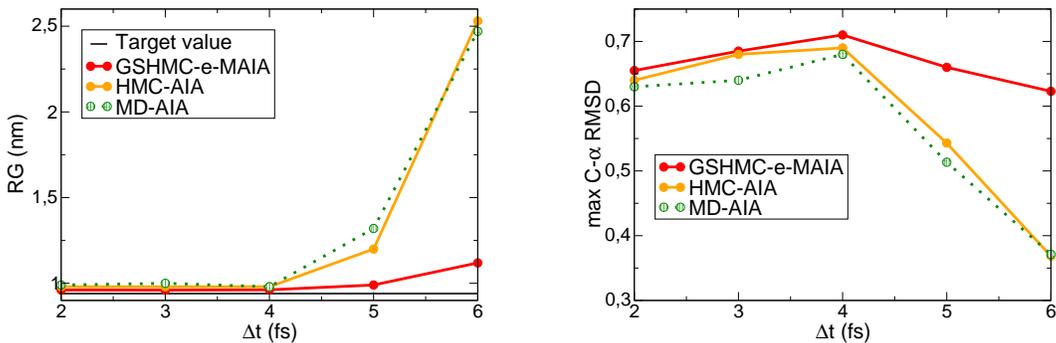


Figure 9: Villin. Sampling efficiency: GSHMC (e-MAIA) vs. HMC (AIA) vs. MD (AIA). The best integrator for each sampling method was employed. Sampling efficiency was measured through the radius of gyration (left) and the maximum RMSD of the  $\alpha$ -carbon of the protein (right). The black solid line (left) represents the target experimental value of 0.94 nm.

HMC and MD. As in the toxin case, the best performing integrator was used for each sampler, i.e. e-MAIA was selected for GSHMC and AIA was combined with HMC and MD. For both metrics, GSHMC demonstrated the best results over the range of time steps. Its advantage over HMC and MD is most visible at longer time steps, when both HMC and MD loose accuracy and sampling efficiency.

Additionally, we have generated Ramachandran plots considering all residues of the protein except for glycine. In Figure 10 the Ramachandran plots, obtained for the largest time step  $\Delta t = 6$  fs, are presented as two-dimensional joint distributions of  $\varphi$  and  $\Psi$  angles.

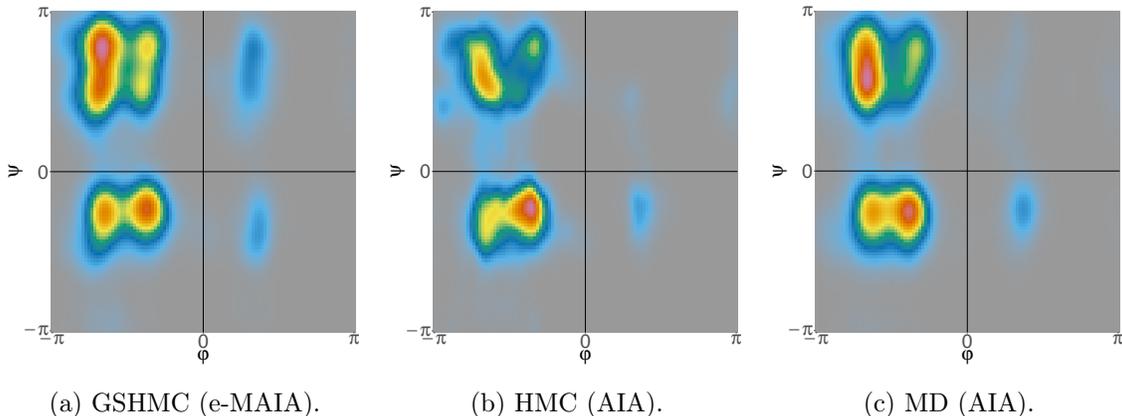


Figure 10: Villin. Sampling efficiency: GSHMC (e-MAIA) vs. HMC (AIA) vs. MD (AIA). Ramachandran plots for all residues of the protein except for glycine with  $\varphi$  torsion on the horizontal axis and  $\Psi$  on the vertical axis. The best integrator for each sampling method was employed. The time step was 6 fs, the largest in these tests.

Figure 10 confirms the advantages of GSHMC over other tested methods. Indeed, GSHMC combined with e-MAIA is the only method capable of sampling all regions including the less populated basins in the  $\varphi, \Psi > 0$  region, which were out of reach for HMC and MD sampling.

Obviously, a deep atomistic study of the villin folding requires significantly longer runs than those presented here, as well as the incorporation of additional sampling techniques, such as, for example, parallel tempering, to the simulations. The latter can be implemented in a similar way, with similar cost for all three methodologies considered in our study. However, the simulations will clearly be more efficient if the underlying sampling method provides higher sampling efficiency, which is the case for GSHMC with e-MAIA.

## Conclusions

We have proposed an adaptive approach for enhancing the accuracy and sampling efficiency of modified Hamiltonian Monte Carlo (MHMC) methods, with the aim of making them strong competitors of popular techniques such as molecular dynamics (MD) and hybrid Monte Carlo (HMC). Given a system to be simulated and a user-chosen time step, the new method, which we call Modified Adaptive Integration Approach or MAIA, identifies the

two-stage numerical integrator which, when used in the Hamiltonian dynamics step of an MHMC method, provides the best conservation of the relevant modified Hamiltonian and thus the highest acceptance rate of the proposed trajectories. An enhanced variant of MAIA, e-MAIA, tailored to Generalized Shadow Hybrid Monte Carlo (GSHMC) methods, additionally supplies a value of the parameter  $\varphi$  that, for the problem under consideration, keeps the momentum acceptance at a user-desired level. The MAIA algorithm has been implemented, with no computational overhead during simulations, in MultiHMC-GROMACS, a modified version of the popular software package GROMACS. The effect of the use of MAIA on the sampling efficiency of GSHMC has been demonstrated by using constrained atomistic and unconstrained coarse-grained benchmarks, and compared with the performance of other suitable integration schemes, including the popular velocity Verlet integrator. The tests revealed that the replacement in GSHMC of any fixed, two-stage integrator with e-MAIA leads systematically to improvements in sampling efficiency of up to an order of magnitude. The performance comparison of GSHMC, HMC and MD combined with their best choices of numerical integrators (e-MAIA, AIA, AIA respectively) confirmed the efficiency and robustness of the GSHMC-MAIA combination, whose advantages are especially noticeable when using the longest possible simulation time steps. For such cases, GSHMC, while maintaining good accuracy in simulation, provided a sampling efficiency (as measured with IACF) up to 30 times higher than the efficiency that may be achieved with MD.

## Acknowledgements

The authors would like to thank the financial support from MTM2013-46553-C3-1-P funded by MINECO (Spain). They have also received funding from the Project of the Spanish Ministry of Economy and Competitiveness with reference MTM2016-76329-R (AEI/FEDER, EU). This work has been possible thanks to the support of the computing infrastructure of the i2BASQUE academic network, and the technical and human support provided by

IZO-SGI SGIker of UPV/EHU and European funding (ERDF and ESF). E.A. and T.R. thank for support Basque Government - ELKARTEK Programme, grant KK-2016/0002. M.F.P. would like to thank the Spanish Ministry of Economy and Competitiveness for funding through the fellowship BES-2014-068640. J.M.S.S. has been additionally supported by project MTM2016-77660-P(AEI/FEDER, UE) funded by MINECO (Spain). This research is also 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) McLachlan, R. I. On the Numerical Integration of Ordinary Differential Equations by Symmetric Composition Methods. *SIAM J. Scient. Comp.* **1995**, *16*, 151–168.
- (2) Blanes, S.; Casas, F.; Sanz-Serna, J. M. Numerical Integrators for the Hybrid Monte Carlo Method. *SIAM J. Scient. Comp.* **2014**, *36*, A1556–A1580.
- (3) Chao, W.-L.; Solomon, J.; Michels, D. L.; Sha, F. Exponential Integration for Hamiltonian Monte Carlo. International Conference on Machine Learning – ICML 2015. 2015; pp 1142–1151.
- (4) Campos, C. M.; Sanz-Serna, J. M. Palindromic 3-stage splitting integrators, a roadmap. *J. Comput. Phys.* **2017**, *346*, 340–355.
- (5) Fernández-Pendás, M.; Akhmatskaya, E.; Sanz-Serna, J. M. Adaptive multi-stage integrators for optimal energy conservation in molecular simulations. *J. Comput. Phys.* **2016**, *327*, 434–449.
- (6) Izaguirre, J. A.; Hampton, S. S. Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *J. Comput. Phys.* **2004**, *200*, 581–604.

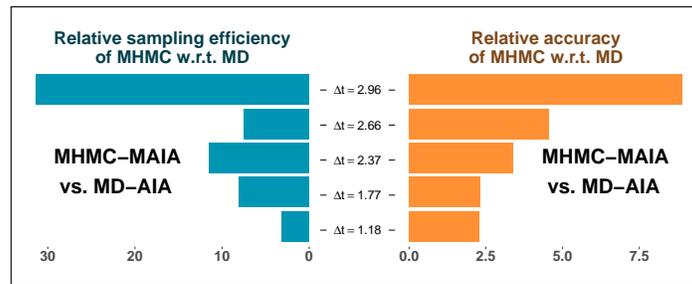
- (7) Sweet, C. R.; Hampton, S. S.; Skeel, R. D.; Izaguirre, J. A. A separable shadow Hamiltonian hybrid Monte Carlo method. *J. Chem. Phys.* **2009**, *131*, 174106.
- (8) Akhmatskaya, E.; Reich, S. GSHMC: An efficient method for molecular simulation. *J. Comput. Phys.* **2008**, *227*, 4934–4954.
- (9) Escribano, B.; Akhmatskaya, E.; Reich, S.; Azpiroz, J. M. Multiple-time-stepping generalized hybrid Monte Carlo methods. *J. Comput. Phys.* **2015**, *280*, 1–20.
- (10) Akhmatskaya, E.; Reich, S. Meso-GSHMC: A stochastic algorithm for meso-scale constant temperature simulations. *Procedia Computer Science* **2011**, *4*, 1353–1362.
- (11) Radivojević, T. Enhancing Sampling in Computational Statistics Using Modified Hamiltonians. Ph.D. thesis, University of the Basque Country (UPV/EHU), Bilbao, 2016.
- (12) Akhmatskaya, E.; Reich, S. New Hybrid Monte Carlo Methods for Efficient Sampling: from Physics to Biology and Statistics (Selected Papers of the Joint International Conference of Supercomputing in Nuclear Applications and Monte Carlo : SNA + MC 2010). *Progress in nuclear science and technology* **2011**, *2*, 447–462.
- (13) Wee, C. L.; Sansom, M. S. P.; Reich, S.; Akhmatskaya, E. 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* **2008**, *112*, 5710–5717.
- (14) Escribano, B.; Lozano, A.; Radivojević, T.; Fernández-Pendás, M.; Carrasco, J.; Akhmatskaya, E. Enhancing sampling in atomistic simulations of solid-state materials for batteries: a focus on olivine  $\{\text{NaFePO}\}_4$ . *Theor. Chem. Acc.* **2017**, *136*, 43.

- (15) Sanz-Serna, J. M.; Calvo, M. P. *Numerical Hamiltonian problems*, 1st ed.; Applied Mathematics and Mathematical Computation 7; Chapman & Hall: London, 1994.
- (16) Leimkuhler, B. J.; Reich, S. *Simulating Hamiltonian dynamics*; Cambridge monographs on applied and computational mathematics; Cambridge Univ.: Cambridge, 2004.
- (17) Hairer, E.; Lubich, C.; Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations; 2nd ed.*; Springer: Dordrecht, 2006.
- (18) Beskos, A.; Pillai, N.; Roberts, G.; Sanz-Serna, J.-M.; Stuart, A. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli* **2013**, *19*, 1501–1534.
- (19) Akhmatskaya, E.; Reich, S. In *New Algorithms for Macromolecular Simulation*; Leimkuhler, B., Chipot, C., Elber, R., Laaksonen, A., Mark, A., Schlick, T., Schütte, C., Skeel, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp 141–153.
- (20) Horowitz, A. M. A generalized guided Monte Carlo algorithm. *Phys. Lett. B* **1991**, *268*, 247–252.
- (21) Kennedy, A.; Pendleton, B. Cost of the generalised hybrid Monte Carlo algorithm for free field theory. *Nucl. Phys. B* **2001**, *607*, 456–510.
- (22) Schlick, T.; Mandziuk, M.; Skeel, R. D.; Srinivas, K. Nonlinear Resonance Artifacts in Molecular Dynamics Simulations. *J. Comput. Phys.* **1998**, *140*, 1–29.
- (23) Mazur, A. K. Common Molecular Dynamics Algorithms Revisited: Accuracy and Optimal Time Steps of Störmer–Leapfrog Integrators. *J. Comput. Phys.* **1997**, *136*, 354–365.
- (24) Akhmatskaya, E.; Bou-Rabee, N.; Reich, S. A comparison of generalized hybrid Monte Carlo methods with and without momentum flip. *J. Comput. Phys.* **2009**, *228*, 2256–2265.

- (25) Escribano, B.; Akhmatskaya, E.; Mujika, J. I. Combining stochastic and deterministic approaches within high efficiency molecular simulations. *Central European J. Math.* **2013**, *11*, 787–799.
- (26) Fernández-Pendás, M.; Escribano, B.; Radivojević, T.; Akhmatskaya, E. Constant pressure hybrid Monte Carlo simulations in GROMACS. *J. Mol. Model.* **2014**, *20*, 2487.
- (27) Berendsen, H.; van der Spoel, D.; van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **1995**, *91*, 43–56.
- (28) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theo. Comp.* **2008**, *4*, 435–447.
- (29) Duane, S.; Kennedy, A.; Pendleton, B. J.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216–222.
- (30) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
- (31) Oh, S.; Hori, Y. Development of Golden Section Search Driven Particle Swarm Optimization and its Application. 2006 SICE-ICASE International Joint Conference. 2006; pp 2868–2873.
- (32) Jung, H. J.; Lee, J. Y.; Kim, S. H.; Eu, Y.-J.; Shin, S. Y.; Milesco, M.; Swartz, K. J.; Kim, J. I. Solution Structure and Lipid Membrane Partitioning of VSTx1, an Inhibitor of the KvAP Potassium Channel,. *Biochemistry* **2005**, *44*, 6015–6023.
- (33) Bazari, W. L.; Matsudaira, P.; Wallek, M.; Smeal, T.; Jakes, R.; Ahmed, Y. Villin sequence and peptide map identify six homologous domains. *Proceedings of the National Academy of Sciences* **1988**, *85*, 4986–4990.

- (34) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR structure of the 35-residue villin headpiece subdomain. *Nature Structural & Molecular Biology* **1997**, *4*, 180–184.
- (35) Wallace, E. J.; Sansom, M. S. P. Carbon Nanotube/Detergent Interactions via Coarse-Grained Molecular Dynamics. *Nano Lett.* **2007**, *7*, 1923–1928.
- (36) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. Coarse Grained Protein-Lipid Model with Application to Lipoprotein Particles. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.
- (37) van der Spoel, D.; Lindahl, E. Brute-Force. *J. Phys. Chem. B* **2003**, *107*, 11178–11187.
- (38) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: New York, NY, USA, 1989.
- (39) Radivojević, T.; Akhmatkaya, E. Mix & Match Hamiltonian Monte Carlo. *ArXiv e-print 1706.04032* **2017**,

# Graphical TOC Entry



## SUPPORTING INFORMATION

### Adaptive splitting integrators for enhancing sampling efficiency of modified Hamiltonian Monte Carlo methods in molecular simulation

Elena Akhmatskaya<sup>\*1,2</sup>, Mario Fernández-Pendás<sup>1</sup>, Tijana Radivojević<sup>1</sup>, and J.M. Sanz-Serna<sup>3</sup>

<sup>1</sup>BCAM - Basque Center for Applied Mathematics, Alameda de Mazarredo 14, E-48009 Bilbao, Spain

<sup>2</sup>Ikerbasque, Basque Foundation for Science, María Díaz de Haro 3, E-48013 Bilbao, Spain

<sup>3</sup>Departamento de Matemáticas, Universidad Carlos III de Madrid, Avenida de la Universidad 30, E-28911 Leganés (Madrid), Spain

## Contents

<b>1 Derivation of the function <math>\rho</math> used in MAIA and e-MAIA algorithms</b>	<b>S1</b>
<b>2 Flowchart of MAIA and e-MAIA algorithms</b>	<b>S5</b>
<b>3 Validation of the chosen simulation length for the villin benchmark</b>	<b>S6</b>

## 1 Derivation of the function $\rho$ used in MAIA and e-MAIA algorithms

In order to derive integrators with optimal conservation properties, we adopt a strategy similar to the one proposed in ref 1, namely to find the parameters of integrators that minimize the expected value of the energy error. In the present study, the energy error resulting from numerical integration is in terms of the modified Hamiltonian and the expected value is taken with respect to the modified density

$$\tilde{\pi}(\mathbf{q}, \mathbf{p}) \propto \exp\left(-\beta\tilde{H}^{[k]}(\mathbf{q}, \mathbf{p})\right).$$

As in the case when considering the error in the true Hamiltonian,<sup>1</sup> one may prove that the expected error in the modified Hamiltonian  $\mathbb{E}[\Delta\tilde{H}^{[4]}]$  is also positive. Our objective is, therefore, to find a function  $\rho(h, b)$  that upperbounds  $\mathbb{E}[\Delta\tilde{H}^{[4]}]$ , i.e.,

$$0 \leq \mathbb{E}[\Delta\tilde{H}^{[4]}] \leq \frac{1}{\beta}\rho(h, b),$$

where  $b$  is the parameter of the two-stage integrators family

$$\psi_{\Delta t} = \phi_{b\Delta t}^B \circ \phi_{\Delta t/2}^A \circ \phi_{(1-2b)\Delta t}^B \circ \phi_{\Delta t/2}^A \circ \phi_{b\Delta t}^B. \quad (\text{S1})$$

---

\*E-mail: akhmatskaya@bcamath.org

Such family is defined in terms of solution flows of the equations of motion<sup>1</sup>

$$\phi_t^A(\mathbf{q}, \mathbf{p}) = (\mathbf{q}, \mathbf{p} - t\nabla U(\mathbf{q})) \quad (\text{S2})$$

and

$$\phi_t^B(\mathbf{q}, \mathbf{p}) = (\mathbf{q} + tM^{-1}\mathbf{p}, \mathbf{p}). \quad (\text{S3})$$

We consider the one-dimensional harmonic oscillator with potential  $U(q) = (k/2)q^2$  ( $k > 0$  a constant) and mass  $M$ , whose equations of motion are

$$\frac{dq}{dt} = \frac{p}{M}, \quad \frac{dp}{dt} = -kq. \quad (\text{S4})$$

Using a linear change of variables  $\bar{q} = \sqrt{k}q, \bar{p} = p/\sqrt{M}$  and denoting the non-dimensional time step as  $h = \omega\Delta t$ , where  $\omega = \sqrt{k/M}$ , lead to the dynamics

$$\frac{d\bar{q}}{dt} = \omega\bar{p}, \quad \frac{d\bar{p}}{dt} = -\omega\bar{q}, \quad (\text{S5})$$

with Hamiltonian

$$H(\bar{q}, \bar{p}) = \frac{1}{2}\bar{p}^2 + \frac{1}{2}\bar{q}^2$$

and the fourth order modified Hamiltonian for integrators of the family eq S1

$$\tilde{H}^{[4]}(\bar{q}, \bar{p}) = \frac{1}{2}\bar{p}^2 + \frac{1}{2}\bar{q}^2 + h^2\lambda\bar{p}^2 + h^2\mu\bar{q}^2. \quad (\text{S6})$$

The numerical integration of the system eq S5 in the new variables  $\bar{q}, \bar{p}$  is equivalent to the application of the above change of variables to the numerical solution  $q, p$  obtained by integration of the system eq S4.

In order to find the error in the modified Hamiltonian after  $L$  integration steps of the dynamics eq S5 with a time step  $h$ , i.e.,

$$\Delta\tilde{H}^{[4]} = \tilde{H}^{[4]}(\Psi_{h,L}(\bar{q}, \bar{p})) - \tilde{H}^{[4]}(\bar{q}, \bar{p}), \quad (\text{S7})$$

we first find the numerical solution to the dynamics eq S5 for a single time step  $(\bar{q}_{n+1}, \bar{p}_{n+1}) = \psi_h(\bar{q}_n, \bar{p}_n)$ . In matrix form this is given by

$$\begin{bmatrix} \bar{q}_{n+1} \\ \bar{p}_{n+1} \end{bmatrix} = \tilde{M}_h \begin{bmatrix} \bar{q}_n \\ \bar{p}_n \end{bmatrix}, \quad \tilde{M}_h = \begin{bmatrix} A_h & B_h \\ C_h & A_h \end{bmatrix},$$

where the coefficients  $A_h, B_h, C_h$  depend on the integrator. After  $L$  integration steps the state of the system  $(\bar{q}_L, \bar{p}_L) = \Psi_{h,L}(\bar{q}, \bar{p})$  is given by

$$\begin{bmatrix} \bar{q}_L \\ \bar{p}_L \end{bmatrix} = \underbrace{\tilde{M}_h \cdots \tilde{M}_h}_{L \text{ times}} \begin{bmatrix} \bar{q} \\ \bar{p} \end{bmatrix} = \tilde{M}_h^L \begin{bmatrix} \bar{q} \\ \bar{p} \end{bmatrix}. \quad (\text{S8})$$

For the two-stage family of integrators the matrix  $\tilde{M}_h$  can be calculated as

$$\tilde{M}_h = B(b) \cdot A\left(\frac{1}{2}\right) \cdot B(1-2b) \cdot A\left(\frac{1}{2}\right) \cdot B(b),$$

where

$$A(a) = \begin{bmatrix} 1 & ah \\ 0 & 1 \end{bmatrix}, \quad B(b) = \begin{bmatrix} 1 & 0 \\ -bh & 1 \end{bmatrix}$$

correspond to the flows  $\phi_h^A$  and  $\phi_h^B$ , respectively (eqs S2 and S3). The resulting entries of  $\tilde{M}_h$  are

$$\begin{aligned} A_h &= \frac{h^4}{4}b(1-2b) - \frac{h^2}{2} + 1 \\ B_h &= -\frac{h^3}{4}(1-2b) + h \\ C_h &= -\frac{h^5}{4}b^2(1-2b) + h^3b(1-b) - h. \end{aligned} \tag{S9}$$

It is well known that if time step  $h$  is such that  $|A_h| < 1$  the integration is stable.<sup>2,3</sup> In that case one may define variables

$$\zeta_h := \arccos A_h, \quad \chi_h := B_h / \sin \zeta_h,$$

for which the one-step and  $L$ -steps integration matrices  $\tilde{M}_h$  and  $\tilde{M}_h^L$ , respectively, are

$$\tilde{M}_h = \begin{bmatrix} \cos(\zeta_h) & \chi_h \sin(\zeta_h) \\ -\chi_h^{-1} \sin(\zeta_h) & \cos(\zeta_h) \end{bmatrix}$$

and

$$\tilde{M}_h^L = \begin{bmatrix} \cos(L\zeta_h) & \chi_h \sin(L\zeta_h) \\ -\chi_h^{-1} \sin(L\zeta_h) & \cos(L\zeta_h) \end{bmatrix}. \tag{S10}$$

In order to calculate the expected value of the error eq S7 we follow ideas from the proof of Proposition 3 in ref 1 and denote

$$\begin{aligned} c &= \cos(L\zeta_h), \\ s &= \sin(L\zeta_h), \\ S_1 &= 1 + 2h^2\mu, \\ S_2 &= 1 + 2h^2\lambda. \end{aligned}$$

Substituting eqs S6, S10 and S8 into eq S7 leads to

$$\begin{aligned} 2\Delta\tilde{H}^{[4]} &= S_1 (c\bar{q} + \chi_h s\bar{p})^2 + S_2 \left( -\frac{1}{\chi_h} s\bar{q} + c\bar{p} \right)^2 - S_1 \bar{q}^2 - S_2 \bar{p}^2 \\ &= s^2 \left( \frac{1}{\chi_h^2} S_2 - S_1 \right) \bar{q}^2 + s^2 \left( \chi_h^2 S_1 - S_2 \right) \bar{p}^2 + 2sc \left( S_1 \chi_h - S_2 \frac{1}{\chi_h} \right) \bar{q}\bar{p}. \end{aligned}$$

Since the expectations are taken with respect to the modified density  $\tilde{\pi}$ ,

$$\mathbb{E}[\bar{q}^2] = \frac{1}{\beta S_1}, \quad \mathbb{E}[\bar{p}^2] = \frac{1}{\beta S_2}, \quad \mathbb{E}[\bar{q}\bar{p}] = 0,$$

it follows that

$$2\mathbb{E}[\Delta\tilde{H}^{[4]}] = \frac{1}{\beta} s^2 \left( \frac{1}{\chi_h^2} \frac{S_2}{S_1} + \chi_h^2 \frac{S_1}{S_2} - 2 \right).$$

The last expression can be simplified by defining

$$\tilde{\chi}_h^2 := \chi_h^2 \frac{S_1}{S_2} = \chi_h^2 \mathcal{S},$$

to obtain

$$\mathbb{E}[\Delta\tilde{H}^{[4]}] = \frac{1}{\beta} s^2 \rho(h, b),$$

where

$$\rho(h, b) = \frac{1}{2} \left( \tilde{\chi}_h - \frac{1}{\tilde{\chi}_h} \right)^2 = \frac{(SB_h + C_h)^2}{2S(1 - A_h^2)} \quad (\text{S11})$$

and

$$S = \frac{1 + 2h^2\mu}{1 + 2h^2\lambda}.$$

We note that the conditions for stable integration and positivity of  $\rho(h, b)$  are that  $|A_h| < 1$  and  $S > 0$ . For the two-stage integrators and the fourth order modified Hamiltonian this is equivalent to the following conditions

$$\begin{aligned} h &< \sqrt{12/(1 - 6b)} && \text{for } b < \frac{1}{6}, \\ h &> \sqrt{12/(1 - 6b)} && \text{for } b > \frac{1}{6}, \\ 0 &< h &< \min \{ \sqrt{2/b}, \sqrt{1/(1 - 2b)} \}, \end{aligned}$$

which are always satisfied for  $b \in (0, \frac{1}{2})$ .

Finally, by substituting eq S9 into eq S11 we obtain the expression

$$\rho(h, b) = \frac{h^8 \left( b(12 + 4b(6b - 5) + b(1 + 4b(3b - 2))h^2) - 2 \right)^2}{4(2 - bh^2)(4 + (2b - 1)h^2)(2 + b(2b - 1)h^2)(12 + (6b - 1)h^2)(6 + (1 + 6(b - 1)b)h^2)}, \quad (\text{S12})$$

which bounds the expected error in the modified Hamiltonian. This function is then used within an optimization routine to find the value  $b$  that provides the optimal conservation of the modified Hamiltonian for a specific system.

## 2 Flowchart of MAIA and e-MAIA algorithms

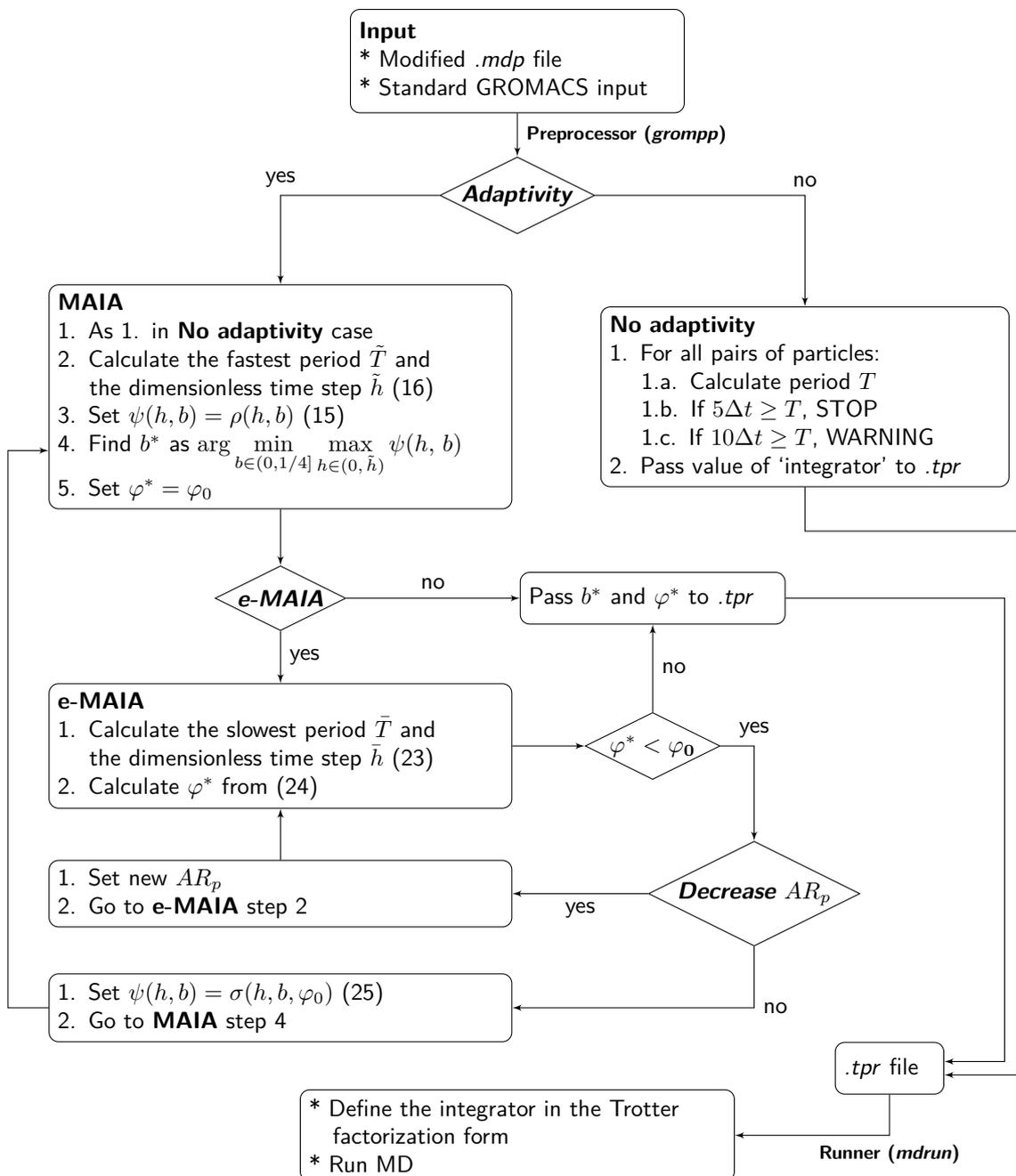


Figure S1: Flowchart of the Modified Adaptive Integration Approach (MAIA) and the extended MAIA (e-MAIA) as implemented in MultiHMC-GROMACS. The references (15)-(16) and (23)-(25) correspond to the equations in the paper.

### 3 Validation of the chosen simulation length for the villin benchmark

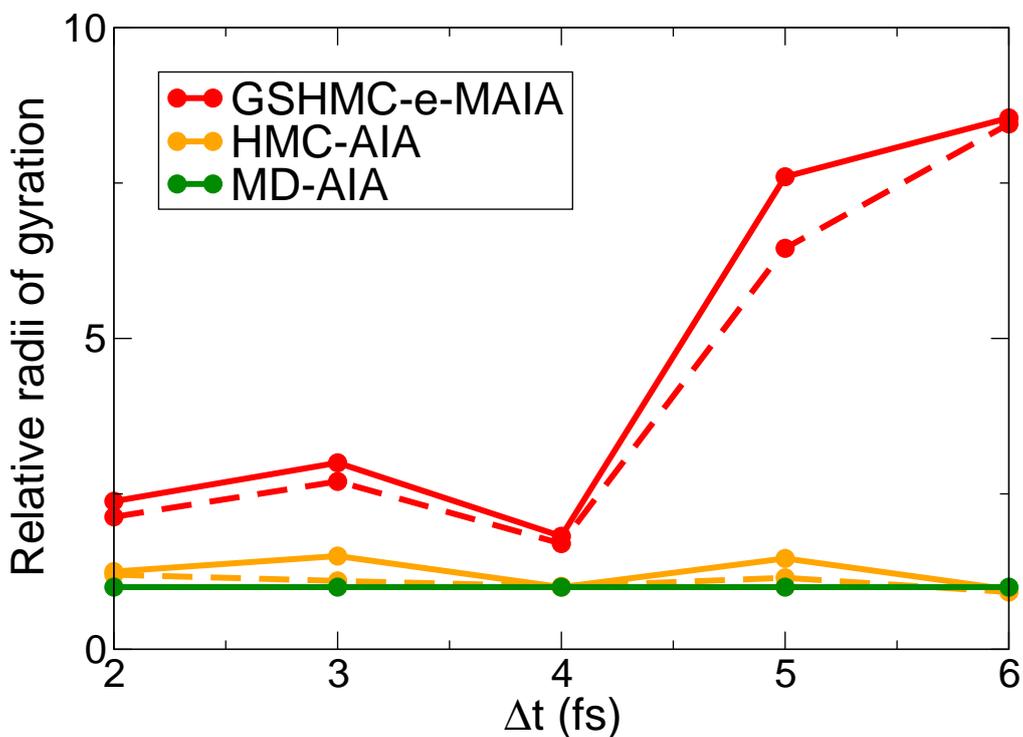


Figure S2: Evolution with time of the relative radii of gyration (RG) observed for each simulation method with respect to the RG found in MD simulations. The dashed lines represent the RG at half of the simulation time (2.5 ns) whereas the solid lines are used for the full simulations of 5 ns. The efficiency of GSHMC with e-MAIA, relative to MD, expressed in terms of radii of gyration, increases with simulation time. This suggests that simulations longer than those presented in this study are at least as favorable to the proposed algorithms as we claim; they may be even more favorable as the simulations become longer.

### References

- [1] Blanes, S.; Casas, F.; Sanz-Serna, J. M. Numerical Integrators for the Hybrid Monte Carlo Method. *SIAM J. Scient. Comp.* **2014**, *36*, A1556–A1580.
- [2] Sanz-Serna, J. M.; Calvo, M. P. *Numerical Hamiltonian problems*, 1st ed.; Applied Mathematics and Mathematical Computation 7; Chapman & Hall: London, 1994.
- [3] Hairer, E.; Lubich, C.; Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations; 2nd ed.*; Springer: Dordrecht, 2006.