MDPI

Article

Quantifying configuration-sampling error in Langevin simulations of complex molecular systems

Josh Fass [1,6], David A. Sivak [2], Gavin E. Crooks [3], Kyle A. Beauchamp [4], Ben Leimkuhler [5], John D. Chodera [6*]

[1] Tri-Institutional PhD Program in Computational Biology & Medicine, New York, NY 10065; josh.fass@choderalab.org
[2] Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; dsivak@sfu.ca, http://www.sfu.ca/physics/sivakgroup.html
[3] Rigetti Computing; gec@threeplusone.com
[4] Counsyl, South San Francisco, CA 94080; kyleabeauchamp@gmail.com
[5] School of Mathematics and Maxwell Institute of Mathematical Sciences, James Clerk Maxwell Building, Kings Buildings, University of Edinburgh, Edinburgh, EH9 3JZ, UK; B.Leimkuhler@ed.ac.uk, http://kac.maths.ed.ac.uk/~bl
[6] Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065; john.chodera@choderalab.org, http://choderalab.org
[*] Correspondence: john.chodera@choderalab.org

Version February 16, 2018 submitted to Entropy

Abstract: While Langevin integrators are widely popular in the study of equilibrium properties of complex systems, it is challenging to estimate the the timestep-induced discretization error: the degree to which the sampled phase space or configuration space probability density departs from the desired target density due to the use of a finite integration timestep. In [1], Sivak *et al.* introduced a convenient approach to quantifying the a natural measure of distribution error between the sampled density and the target equilibrium density, the KL divergence, in *phase space*, but did not specifically address the issue of *configuration-space properties*, which are much more commonly of interest in molecular simulations. Here, we introduce a variant of this near-equilibrium estimator capable of measuring the error in the configuration-space marginal density, validating it against a complex but exact nested Monte Carlo estimator to show that it reproduces the KL divergence with high fidelity. To illustrate its utility, we employ this new near-equilibrium estimator to assess a claim that a recently proposed Langevin integrator introduces extremely small configuration-space density errors up to the stability limit at no extra computational expense. Finally, we show how this approach to quantifying sampling error can be applied to a wide variety of stochastic integrators by following a straightforward procedure to compute the appropriate shadow work, and describe how it can be extended to quantify the error in arbitrary marginal or conditional distributions of interest.

Keywords: Langevin dynamics; Langevin integrators; KL divergence; nonequilibrium free energy; molecular dynamics integrators; integrator error; sampling error; BAOAB; VRORV; OBABO; OVRVO; VVVR; Bussi-Parrinello; shadow work; integrator error

---

Contents

1 Introduction **2**

2 Results **9**

2.1 Common Langevin integrator splittings induce comparable timestep-dependent error in phase space . . . 9
2.2 A simple modification to the near-equilibrium estimator can compute KL divergence in configuration space . . . 11
2.3 Common Langevin integrators induce substantially different timestep-dependent configuration space error . . . 11
2.4 An exact but expensive estimator of KL divergence validates the near-equilibrium estimate 11

3 Discussion **13**
3.1 The near-equilibrium estimator is a fast and useful way to measure integrator error . . . 13
3.2 Future directions . . . 15

4 Detailed methods **15**
4.1 One-dimensional model system: Double-well . . . 16
4.2 Model molecular mechanics system: A harmonically restrained water cluster . . . 16
4.3 Caching equilibrium samples . . . 16
4.4 Computing shadow work for Strang splittings . . . 17
4.5 Computation of shadow work for OVRVO . . . 19
4.6 Near-equilibrium estimators for KL divergence . . . 20
4.7 Variance-controlled adaptive estimator for KL divergence . . . 20

A Relation to GHMC acceptance rates **21**

B Statistics of shadow work distributions **23**

References **25**

1. Introduction

Langevin integrators are widely employed to simulate the equilibrium thermodynamics and kinetics of microscopic molecular systems. The origin of these methods lies in the Langevin equation [2], which describes the behavior of condensed phase systems subject to random weak collisions with fictitious bath particles at thermal equilibrium,

$$d\begin{bmatrix}\mathbf{x}\\ \mathbf{v}\end{bmatrix} = \begin{bmatrix}\mathbf{v}\\ 0\end{bmatrix}\mathrm{d}t + \begin{bmatrix}0\\ -M^{-1}\nabla U(\mathbf{x})\end{bmatrix}\mathrm{d}t + \begin{bmatrix}0\\ -\gamma\,\mathbf{v}\,\mathrm{d}t + \sigma M^{-1/2}\,dW\end{bmatrix}. \tag{1}$$

Here, $\mathbf{x}$ and $\mathbf{v}$ denote the positions and velocities of all particles in the system in Cartesian coordinates, t is time, M the diagonal mass matrix, and $U(\mathbf{x})$ is the potential energy. The constant $\sigma^2 \equiv (2k_BT\gamma)$ quantifies the rate of heat exchange with the bath, with k_BT denoting the thermal energy, γ the collision rate (with dimensions of inverse time), and $W(t)$ a standard multidimensional Wiener process [3,4].

On a digital computer, integration of (1) requires discretizing these equations in time to produce a *finite timestep Langevin integrator* capable of generating a temporally discrete dynamical trajectory from which equilibrium or dynamical properties can be estimated [5]. A wide variety of schemes have been proposed for this discretization, which we refer to herein as "Langevin integrators" [6–15].

For example, the popular integrator of Bussi and Parrinello [12] (which we will denote OVRVO for reasons that will become clear shortly), also known as velocity Verlet with velocity randomization

(VVVR) [16] due to its use of a velocity Verlet integrator core (substeps VRV below) [17], consists of the discrete update equations:

$$\begin{aligned}
\mathbf{v}_{t+1/4} &= a_2 \mathbf{v}_t + \sqrt{1-a_2^2}\,(\beta M)^{-1/2}\,\xi_{t+1/4} && \mathsf{O} \\
\mathbf{v}_{t+1/2} &= \mathbf{v}_{t+1/4} - \frac{\Delta t}{2} M^{-1} \nabla \mathcal{U}(\mathbf{x}_t) && \mathsf{V} \\
\mathbf{x}_{t+1} &= \mathbf{x}_t + \Delta t\, \mathbf{v}_{t+1/2} && \mathsf{R} \\
\mathbf{v}_{t+3/4} &= \mathbf{v}_{t+1/2} - \frac{\Delta t}{2} M^{-1} \nabla \mathcal{U}(\mathbf{x}_{t+1}) && \mathsf{V} \\
\mathbf{v}_{t+1} &= a_2 \mathbf{v}_{t+3/4} + \sqrt{1-a_2^2}\,(\beta M)^{-1/2}\,\xi_{t+3/4} && \mathsf{O}
\end{aligned} \tag{2}$$

where $a_k = e^{-\gamma(\Delta t/k)}$ is a constant that depends on timestep Δt and collision rate γ, and k denotes the number of O steps appearing in the integrator. The intermediate timestep subscripts ($t+1/4$, $t+1/2$, $t+3/4$) denote variables used solely for convenience (without physical meaning), and the ζ denote unit normal random variates.

An alternative splitting championed by Leimkuhler and Matthews termed BAOAB [14,18] (which we will refer to as VRORV) consists of a different set of discrete update equations:

$$\begin{aligned}
\mathbf{v}_{t+1/4} &= \mathbf{v}_t - \frac{\Delta t}{2} M^{-1} \nabla \mathcal{U}(\mathbf{x}_t) && \mathsf{V} \\
\mathbf{x}_{t+1/2} &= \mathbf{x}_t + \frac{\Delta t}{2} \mathbf{v}_{t+1/4} && \mathsf{R} \\
\mathbf{v}_{t+3/4} &= a_1 \mathbf{v}_{t+1/4} + \sqrt{1-a_1^2}\,(\beta M)^{-1/2}\,\xi_{t+1/2} && \mathsf{O} \\
\mathbf{x}_{t+1} &= \mathbf{x}_{t+1/2} + \frac{\Delta t}{2} \mathbf{v}_{t+3/4} && \mathsf{R} \\
\mathbf{v}_{t+1} &= \mathbf{v}_{t+3/4} - \frac{\Delta t}{2} M^{-1} \nabla \mathcal{U}(\mathbf{x}_{t+1}) && \mathsf{V}
\end{aligned} \tag{3}$$

While both discrete time integration schemes reduce to the same stochastic differential equations in the limit that $\Delta t \to 0$, they can behave quite differently for finite timesteps ($\Delta t > 0$), especially for timesteps of practical interest for atomistic molecular simulation.

Langevin integrators introduce sampling bias that grows with the size of the timestep

Langevin integrators intend to sample from the equilibrium density,

$$\pi(\mathbf{x},\mathbf{v}) \propto e^{-h(\mathbf{x},\mathbf{v})} \propto e^{-u(\mathbf{x})} e^{-t(\mathbf{v})}, \tag{4}$$

where β is the inverse temperature, $h(\mathbf{x},\mathbf{v})$ is a separable reduced (unitless) Hamiltonian $h(\mathbf{x},\mathbf{v}) = \beta H(\mathbf{x},\mathbf{v}) = u(\mathbf{x}) + t(\mathbf{v})$, $u(\mathbf{x}) \equiv \beta U(\mathbf{x})$ is the reduced potential [19], and $t(\mathbf{v}) \equiv \beta T(\mathbf{v})$ is the reduced kinetic energy. If only configuration-space properties are of interest, as is often the case in molecular simulations, Langevin integrators aim to sample from the marginal density in configuration space, $\pi_{\mathbf{x}}$,

$$\pi_{\mathbf{x}}(\mathbf{x}) \quad = \quad \int d\mathbf{v}\, \pi(\mathbf{x},\mathbf{v}) \propto e^{-u(\mathbf{x})}, \tag{5}$$

if the velocities are ignored while the positions are used to estimate expectations or other configuration-dependent properties.

The continuous-time Langevin equations of motion (1) possess the target equilibrium density π (4) as their stationary density, suggesting that, at least in principle, temporal averages along Langevin trajectories can be used to approximate averages with respect to the equilibrium distribution. However, numerical simulations with a finite timestep $\Delta t > 0$ will generally sample a slightly different

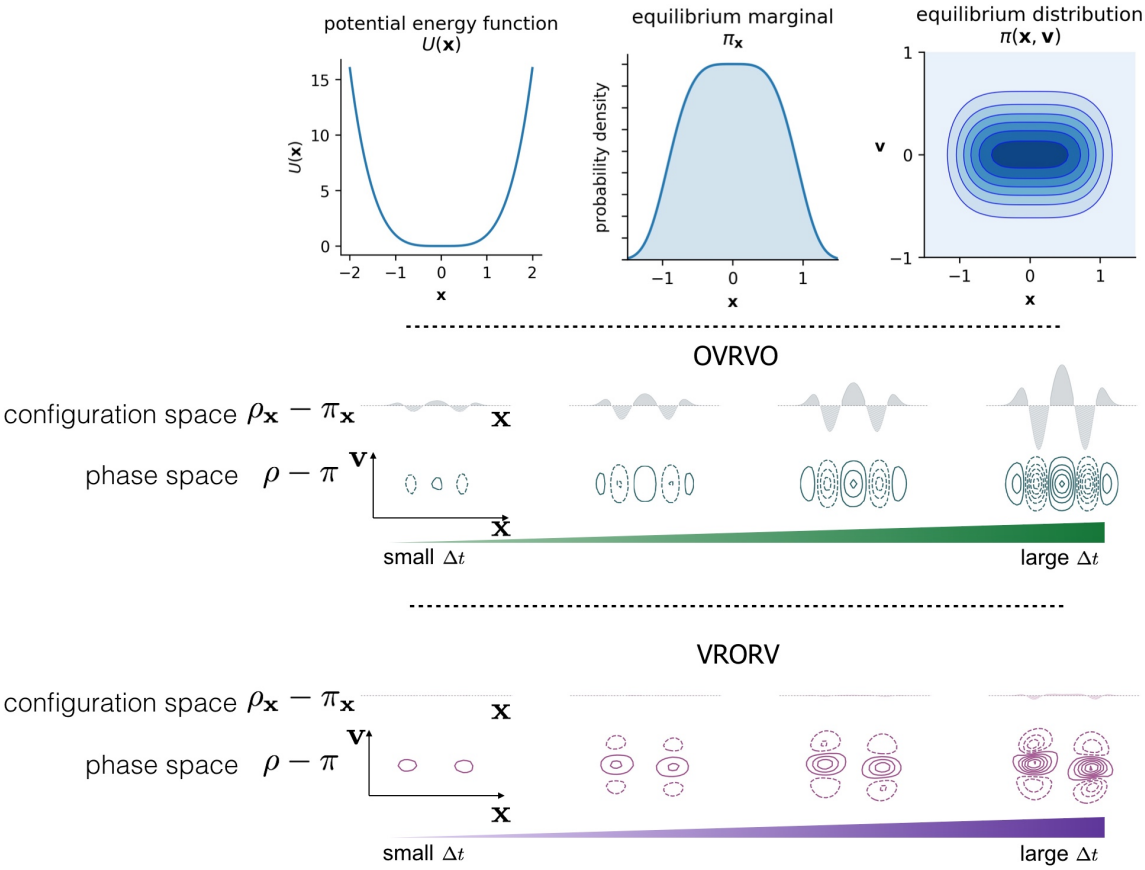

Figure 1. Different Langevin integrators can induce dramatically different errors in sampled phase space and marginal configuration space densities. For a simple 1D system with the quartic potential $U(\mathbf{x}) = \mathbf{x}^4$, the error in sampled phase space density ρ and its marginal density $\rho_{\mathbf{x}}$ grows as a function of timestep Δt. However, different Langevin integrators (OVRVO and VRORV shown here) derived from symmetric Strang splittings can lead to drastically different error structures in phase space, which can induce fortuitous cancellation of error in in the marginal distribution under certain circumstances (VRORV). In the **top row**, we illustrate the definition of the 1D system (**left**: the potential energy function, $U(\mathbf{x}) = \mathbf{x}^4$; **middle**: the equilibrium marginal density over configuration space, $\pi_{\mathbf{x}}(\mathbf{x}) \propto e^{-\beta U(\mathbf{x})}$; **right**: the equilibrium joint distribution over phase space $\pi(\mathbf{x}, \mathbf{v})$. In the **middle row**, we illustrate the *increasing discrepancy between the sampled distribution ρ and the equilibrium distribution π*, for both the full phase space and the marginal configuration space, as a function of timestep Δt, for a particular choice of Langevin integrator OVRVO (2). In the **bottom row**, we illustrate the timestep-dependent error for another integrator VRORV (3).

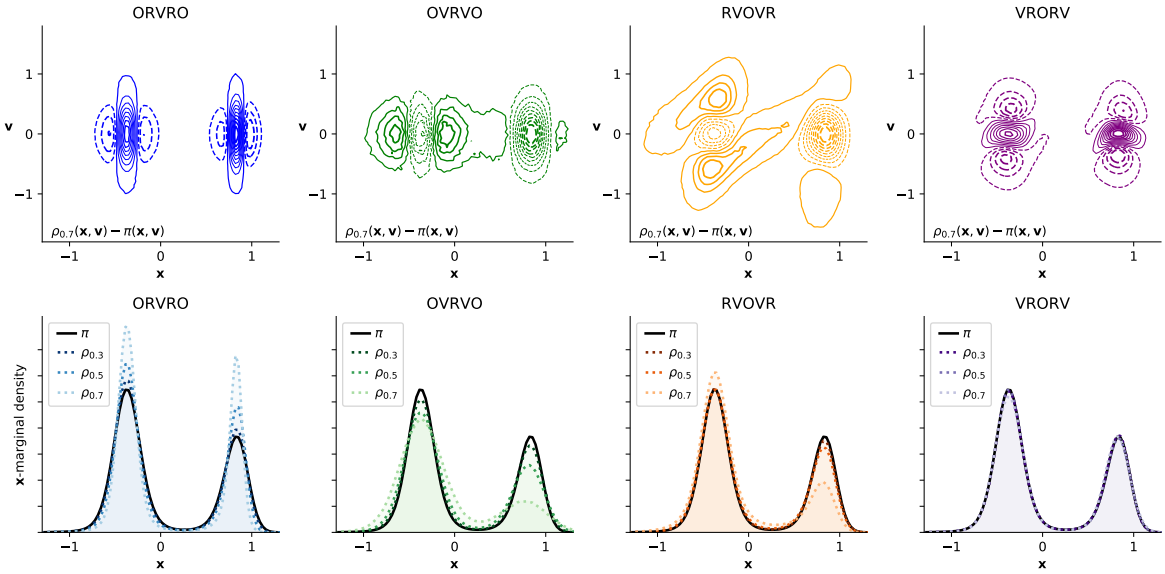

Figure 2. Different numerical integrators introduce different error structure in phase space, illustrated in a double-well system. Here, we illustrate the timestep-dependent discretization error introduced by four integrators on a 1D double-well potential $[U(\mathbf{x}) \equiv \mathbf{x}^6 + 2\cos(5(\mathbf{x}+1))]$.
The **top row** of 2D contour plots illustrates the difference between the phase-space density $\rho(\mathbf{x},\mathbf{v})$ sampled at the maximum timestep considered ($\Delta t = 0.7$, close to the stability limit) and the equilibrium density $\pi(\mathbf{x},\mathbf{v})$; solid lines indicate positive contours, while dashed lines indicate negative contours. The **bottom row** of 1D density plots shows timestep-dependent perturbation in the sampled marginal distribution in configuration space, $\rho_{\mathbf{x}}$, with the equilibrium distribution $\pi_{\mathbf{x}}$ is depicted as a solid black line. The sampled marginal distributions $\rho_{\mathbf{x}}$ are shown for increasingly large timesteps, denoted $\rho_{\Delta t}$, depicted by increasingly light dotted lines, for $\Delta t = 0.3, 0.6, \ldots, 0.7$ (arbitrary units). Inspecting the contour plots suggests that some integrator splittings (especially VRORV) induce error that fortuitously "cancels out" when the density is marginalized by integrating over $\mathbf{v}$, while the error in other integrator splittings (ORVRO, OVRVO) constructively sums to amplify the error in configuration space.

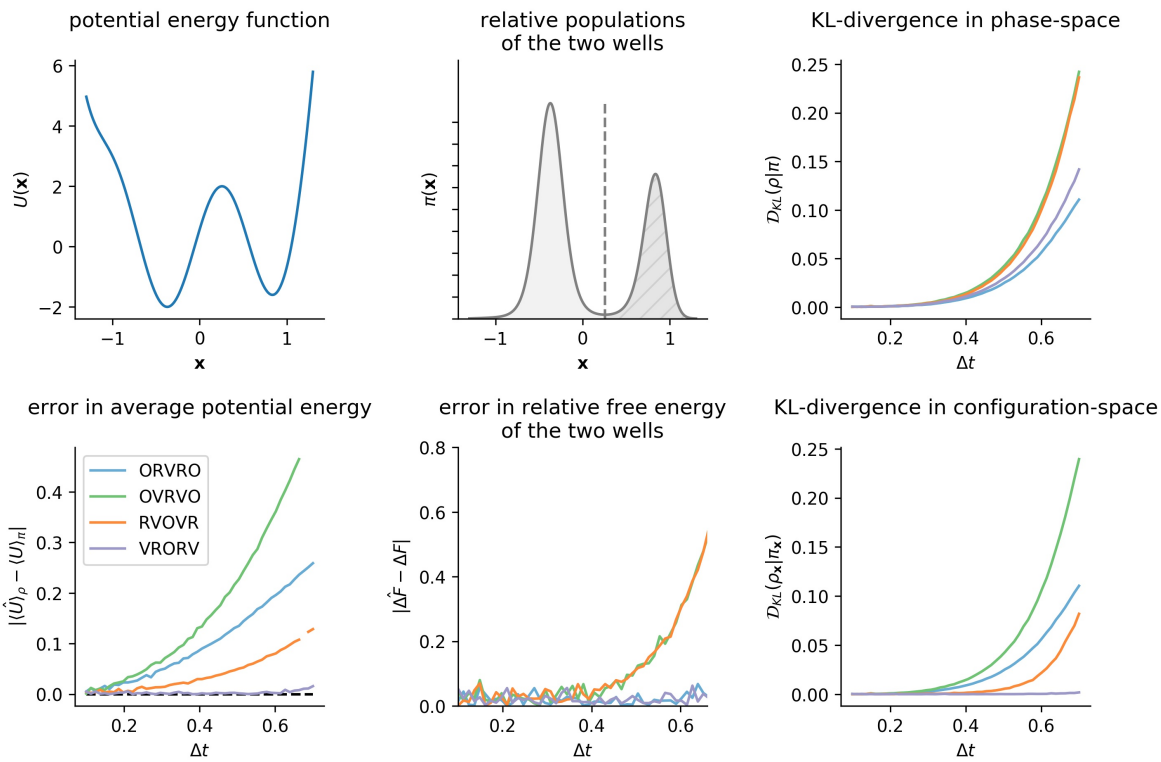

Figure 3. KL divergence is a natural measure of sampling error, though system-specific observables display different sensitivities to sampling error. Even for the simple double-well potential considered in Figure 2, configuration-space properties display different sensitivities to sampling error, motivating the use of a "universal" error measure, such as the KL divergence. The **top left** panel illustrates the double-well potential energy function from Figure 2, and the **top center** panel shows the resulting marginal equilibrium density, $\pi_{\mathbf{x}}$, at $\beta = 1$. The **bottom left** panel shows, as a function of Δt, growth in the magnitude of the error in average potential energy, $|\langle U\rangle_\rho - \langle U\rangle_\pi|$, which has been used previously as a sensitive measure of sampling error [14]. The **bottom center** panel shows the error in the apparent free energy difference between the two wells as a function of Δt. Note that the timestep-dependent behavior of these two observables imply different rankings of integrator fidelity that may mislead one into believing error in *all* observables remains low with increasing timestep. However, as is clear here, just because an integrator introduces low timestep-dependent error in one observable doesn't mean that the method will introduce low error in another observable: for example, OVRVO preserves the well populations as accurately as VRORV, but introduces the largest errors in the average potential energy. The **right column** summarizes the growth in timestep-dependent error, as measured by the KL divergence. While all four integrators introduce comparable levels of Δt-dependent error in the phase-space distribution, they induce dramatically different magnitudes of error in the configuration-space marginal.

distribution, which we will denote by $\rho(\mathbf{x}, \mathbf{v})$, which implicitly depends on timestep Δt. The discrepancy between the distributions ρ and π will grow with Δt at some rate (*e.g.* $\mathcal{O}(\Delta t^2)$ or $\mathcal{O}(\Delta t^4)$), until numerical simulation becomes unstable altogether. Note that this phenomenon is completely separate from numerical issues introduced by finite-precision arithmetic on digital computers, which introduces roundoff error in mathematical operations; here, we presume that computations can be carried out to arbitrary precision, and analyze only the effects of time-discretization.

In Figure 1, we illustrate a few key behaviors of this stepsize-dependent sampling bias in a simple quartic 1D system. Note that: (1) the numerically sampled distribution deviates from the target distribution, (2) this deviation increases with timestep Δt, and (3) the deviation in phase space $(\mathbf{x}, \mathbf{v})$ may be different than the deviation in configuration space only $(\mathbf{x})$.

Many discrete-time integrators can be constructed from the continuous Langevin equations of motion

There are several possible ways to discretize Langevin dynamics in such a way that we recover the exact Langevin equation in the limit $\Delta t \to 0$, and thus recover the desired equilibrium distribution, but as we shall soon see, these methods can have drastically different behaviors for finite timesteps ($\Delta t > 0$). A particularly flexible way to construct numerical integrators for Langevin dynamics is operator splitting, where the Langevin system is commonly split into three components, here labeled V, R, and O because they deal with increments to positions (R), velocities (V), or an Ornstein-Uhlenbeck like process (O)[1],

$$d\begin{bmatrix}\mathbf{x}\\ \mathbf{v}\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{v}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathsf{R}} + \underbrace{\begin{bmatrix}0\\ -M^{-1}\nabla \mathrm{U}(\mathbf{x})\end{bmatrix}\mathrm{d}t}_{\mathsf{V}} + \underbrace{\begin{bmatrix}0\\ -\gamma\,\mathbf{v}\,\mathrm{d}t + \sqrt{2\gamma}\,(\beta M)^{-1/2}dW\end{bmatrix}}_{\mathsf{O}} \tag{6}$$

where each component can be solved "exactly" for a small time increment[2].

The Strang symmetric operator splitting approach considers splitting the overall propagator $e^{\mathcal{L}\Delta t}$ into a product of propagators for the individual R, V, and O components, defining a family of splittings which we will index by strings indicating the order of appearance of these individual propagators. For example, we use OVRVO to refer to the propagator splitting,

$$e^{\mathcal{L}\Delta t} \approx e^{\mathcal{L}_{\mathsf{OVRVO}}\Delta t} \quad = \quad e^{\mathcal{L}_{\mathsf{O}}\frac{\Delta t}{2}} e^{\mathcal{L}_{\mathsf{V}}\frac{\Delta t}{2}} e^{\mathcal{L}_{\mathsf{R}}\Delta t} e^{\mathcal{L}_{\mathsf{V}}\frac{\Delta t}{2}} e^{\mathcal{L}_{\mathsf{O}}\frac{\Delta t}{2}}. \tag{7}$$

While equality between the true propagator and the Strang splitting is only achieved in the limit that $\Delta t \to 0$, the great advantage of this approach is that each component propagator, $e^{\mathcal{L}_{\mathsf{R}}\Delta t/n_{\mathsf{R}}}$, $e^{\mathcal{L}_{\mathsf{V}}\Delta t/n_{\mathsf{V}}}$,

---

1 These components are also often called A, B, and O in the literature [4,14].
2 The R and O components can be integrated exactly for any time increment h, but V can be integrated only to first order.

and $e^{\mathcal{L}_{\mathsf{O}}\Delta t/n_{\mathsf{O}}}$, has a corresponding finite-time update equation that can be used to implement the action of that operator on a digital computer:

$$\mathsf{R} \;:\; e^{\mathcal{L}_{\mathsf{R}}\Delta t/n_{\mathsf{R}}} \;:\; \begin{bmatrix}\Delta\mathbf{x}\\ \Delta\mathbf{v}\end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{v}\\ 0\end{bmatrix}}_{\mathsf{R}} \frac{\Delta t}{n_{\mathsf{R}}} \tag{8}$$

$$\mathsf{V} \;:\; e^{\mathcal{L}_{\mathsf{V}}\Delta t/n_{\mathsf{V}}} \;:\; \begin{bmatrix}\Delta\mathbf{x}\\ \Delta\mathbf{v}\end{bmatrix} = + \underbrace{\begin{bmatrix}0\\ -M^{-1}\nabla \mathrm{U}(\mathbf{x})\end{bmatrix}}_{\mathsf{V}} \frac{\Delta t}{n_{\mathsf{V}}} \tag{9}$$

$$\mathsf{O} \;:\; e^{\mathcal{L}_{\mathsf{O}}\Delta t/n_{\mathsf{O}}} \;:\; \begin{bmatrix}\Delta\mathbf{x}\\ \Delta\mathbf{v}\end{bmatrix} = + \underbrace{\begin{bmatrix}0\\ (a_k-1)\mathbf{v} + \sqrt{1-a_k^2}\,(\beta M)^{-1/2}\,\xi\end{bmatrix}}_{\mathsf{O}} \tag{10}$$

where $a_k = e^{-\gamma(\Delta t/n_{\mathsf{O}})}$ and $\xi \sim \mathcal{N}(0,1)$ is a standard normal random variate drawn for each degree of freedom for each O step. The integers n_{R}, n_{V}, and n_{O} denote the number of times each operator appears in the propagator splitting, which is equal to the number of times each corresponding update equation is used in one integrator cycle.

By chaining these operations in the order specified by the splitting string, we can unroll these operations into the sequence of mathematical updates needed to implement one cycle of the integrator for a total time Δt. For VRORV, for example, translating the splitting string into the appropriate sequence of update equations in (8)–(10) produces the equations for one complete integrator timestep:

$$\begin{aligned}
\mathbf{v}_{t+1/4} &= \mathbf{v}_t - \frac{\Delta t}{2} M^{-1}\nabla \mathrm{U}(\mathbf{x}_t)\\
\mathbf{x}_{t+1/2} &= \mathbf{x}_t + \frac{\Delta t}{2}\mathbf{v}_{t+1/4}\\
\mathbf{v}_{t+3/4} &= a_1\,\mathbf{v}_{t+1/4} + \sqrt{1-a_1^2}\,(\beta M)^{-1/2}\,\xi_{t+1/2}\\
\mathbf{x}_{t+1} &= \mathbf{x}_{t+1/2} + \frac{\Delta t}{2}\mathbf{v}_{t+3/4}\\
\mathbf{v}_{t+1} &= \mathbf{v}_{t+3/4} - \frac{\Delta t}{2} M^{-1}\nabla \mathrm{U}(\mathbf{x}_{t+1})
\end{aligned} \tag{11}$$

Different integrators may approximate different properties with varying fidelity

It has been proposed that some integrators of Langevin dynamics (particularly the VRORV aka "BAOAB" integrator of Leimkuhler and Matthews) preserve the configuration distribution with significantly greater fidelity than other equal-cost integration algorithms [14,18,20]. However, as the formal arguments for this "superconvergence" property rely on a high-friction limit, it is unclear how large the friction coefficient needs to be in practice for the argument to apply. Formal descriptions of the error are typically generic, in that they do not provide guidance on precisely which Δt introduces a tolerable amount of bias for a particular system, and they do not provide a way to predict how other choices, such as mass matrix modifications (*e.g.*, hydrogen mass repartitioning) [21–24]), will affect the error for a system of interest.

Until now, assessing whether specific integrators sample the true equilibrium density with greater fidelity than others in specific settings has relied on computing low-dimensional marginal distributions of manually selected observables perceived to be sensitive to configuration-space sampling errors [25], such as radial distribution functions, marginal distributions of internal coordinates [22], or the configurational temperature [14,26,27]. While it is clear that some observables are more sensitive to errors in configuration space density than others (Figure 2), and the error in the observables of interest is paramount for a particular application, this highlights the risk of using the error in a single

physical property as a surrogate for judging integrator quality, as the error in other observables of interest may be large despite small error in the test observable.

To evaluate numerical Langevin integrators, there would be great utility in a *computable, universal measure* of the bias they introduce in specific concrete settings, such that low error in this measure ensures low error in all observables of interest. There is currently no computable measure of the total configuration-sampling bias for concrete choices of integrator parameters and target system.

KL divergence is a natural measure of sampling bias

Controlling the magnitude of the integrator-induced sampling bias is crucial when computing quantitative predictions from simulation. However, because ρ does not have a closed-form, easily computable expression[3], it is difficult to quantify the error introduced by a given choice of timestep or integrator. We will show how, for a particularly useful measure of error, we can circumvent this problem and develop a simple, effective approach to measuring error in complex molecular systems.

An ideal measure of the discrepancy between the sampled distribution $\rho_{\mathbf{x}}$ and the equilibrium distribution $\pi_{\mathbf{x}}$ should be "universal" in the sense that driving that measure to zero implies that error in any expectation also goes to zero. It should also be defined for all densities, and not rely on a system-specific choice of observables. One such measure is the Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\mathrm{KL}}(p\|q) \equiv \int d\mathbf{x}\, p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) . \tag{12}$$

The KL divergence is defined and non-negative for any pair of distributions on the same support, and $\mathcal{D}_{\mathrm{KL}}(p\|q) = 0$ if and only if $p = q$.

A near-equilibrium estimator can measure timestep-dependent KL divergence for molecular systems

In [1], Sivak and colleagues demonstrated how to approximate the KL divergence of the sampled distribution ρ from the target distribution π *over the full phase-space distribution*, $\mathcal{D}_{\mathrm{KL}}(\rho\|\pi)$, in terms of a work-like quantity—the *shadow work*—that is readily computable for a large family of Langevin integrators (Figure 4a). This estimator depends only on the ability to draw samples from π and to measure a suitable work-like quantity. This method was applied in [1] to measure the phase space sampling bias introduced by a particular Langevin integrator (OVRVO) on periodic boxes of TIP3P water [29].

Since the velocity marginal is usually not of interest, and since some integrators are thought to preserve the configuration marginal of the target distribution with higher fidelity than the phase space joint distribution, we sought to generalize the technique to estimate the KL divergence of the sampled configuration-space marginal, $\mathcal{D}_{\mathrm{KL}}(\rho_{\mathbf{x}}\|\pi_{\mathbf{x}})$. Below, we show how a simple modification of the estimator described in [1] can achieve this goal, and illustrate how this provides a useful tool for measuring the integrator-induced error in configuration-space densities for real molecular systems.

2. Results

2.1. Common Langevin integrator splittings induce comparable timestep-dependent error in phase space

We first applied the original near-equilibrium method of Sivak *et al.* [1] to measure the timestep-dependent phase-space error introduced by four common Langevin integrators on a molecular mechanics model system (a cluster of 20 TIP3P waters [29] in a harmonic restraining potential), for a range of timesteps Δt between 0.1 and 8 femtoseconds (0.1 fs, 0.5 fs, 1.0 fs, . . . , 7.5 fs,

[3] The concept of a *shadow Hamiltonian* has been used to embed this density in a canonical density context, but the shadow Hamiltonian cannot be directly compute, though some approaches to approximate it via expansion (generally requiring higher-order derivatives than gradients) have been proposed – see [28], Chapter 3 of [4], and references therein.

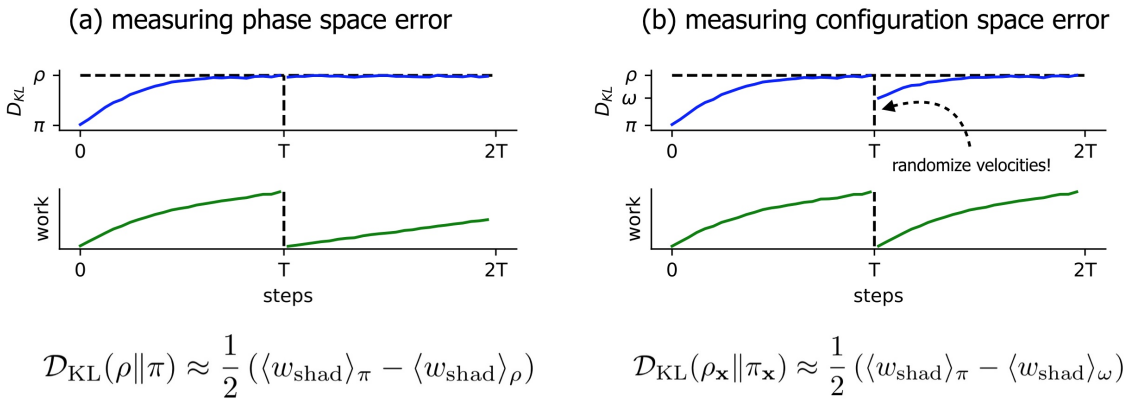

Figure 4. A simple nonequilibrium protocol allows measurement of the KL divergence in phase and configuration space close to equilibrium. Simple nonequilibrium protocols can be used in complex molecular systems to rapidly estimate—utilizing the Crooks fluctuation theorem—the KL divergence of sampled Langevin densities from equilibrium. In both panels, the x-axis is the number of steps taken so far in the length-$2T$ protocol, and $\langle w_{\mathrm{shad}}\rangle_\pi$ indicates the average (reduced unitless) shadow work accumulated over T steps of Langevin dynamics, initialized from equilibrium $((\mathbf{x}_0, \mathbf{v}_0) \sim \pi)$. **(a)** The original scheme described in Sivak *et al.* [1] to measure the KL divergence between the sampled phase space density ρ and the equilibrium phase space density π. $\langle w_{\mathrm{shad}}\rangle_\rho$ is the average shadow work accumulated over T steps of Langevin dynamics, initialized from the integrator's steady state $((\mathbf{x}_0, \mathbf{v}_0) \sim \rho)$. **(b)** The modified scheme introduced here to measure the KL divergence in the *configuration-space* marginal density between the marginal sampled configuration-space density $\rho_{\mathbf{x}}$ and marginal equilibrium density $\pi_{\mathbf{x}}$. $\langle w_{\mathrm{shad}}\rangle_\omega$ is the average shadow work accumulated over T steps of Langevin dynamics, where the initial configuration is drawn from the integrator's steady state, and the initial velocities are drawn from equilibrium $(\mathbf{x}_0 \sim \rho_{\mathbf{x}}, \mathbf{v}_0 \sim \pi(\mathbf{v}|\mathbf{x}_0))$. We denote this distribution $\omega(\mathbf{x}, \mathbf{v}) \equiv \rho_{\mathbf{x}}(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})$. The **top row** schematically illustrates "distance from equilibrium", with y-axis ticks for $\mathcal{D}_{\mathrm{KL}}(\pi\|\pi) = 0$, $\mathcal{D}_{\mathrm{KL}}(\omega\|\pi) \leq \mathcal{D}_{\mathrm{KL}}(\rho\|\pi)$. The **bottom row** illustrates the average work (here, just shadow work) accumulated throughout each protocol.

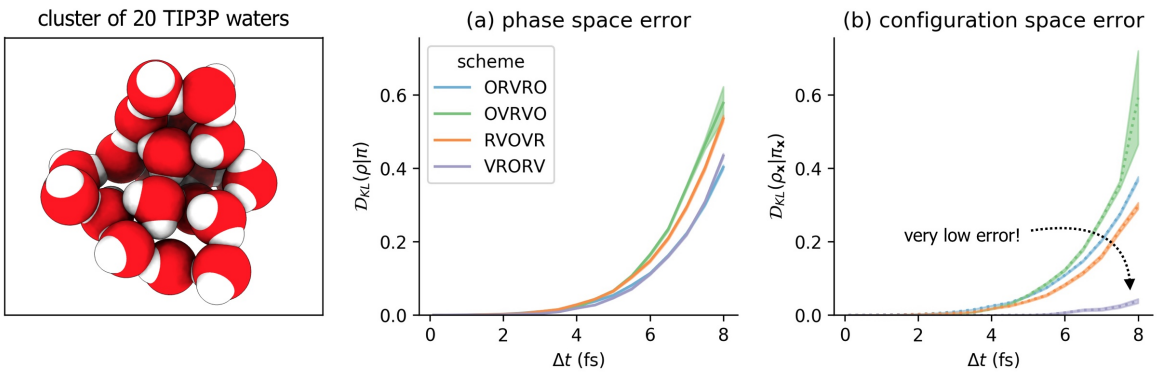

Figure 5. Using the near-equilibrium approximation, some numerical methods introduce far less configuration-space bias in molecular mechanics models than others. The results here are reported for a small cluster of rigid TIP3P waters, described in more detail in the Detailed Methods section, and illustrated **in the leftmost panel.** On the x-axis is the timestep Δt, measured in femtoseconds (fs). On the y-axis is the estimated KL divergence $\mathcal{D}_{\mathrm{KL}}$. **(a)** The error over the joint distribution on $\mathcal{D}_{\mathrm{KL}}(\rho\|\pi)$. **(b)** The error over the configuration-space marginal $\mathcal{D}_{\mathrm{KL}}(\rho_{\mathbf{x}}\|\pi_{\mathbf{x}})$. Each colored curve corresponds to a numerical scheme for Langevin dynamics. The shaded region is the mean $\pm$ 95% confidence interval.

8.0 fs). All four integrator splitting schemes introduced comparable error in phase space (although we could resolve statistically significant differences among the schemes), as illustrated in Figure 5a. This may be unsurprising, as none of these integrator schemes are known to provide a significant reduction in acceptance rates—which depend in some manner on the induced phase space sampling error—for Metropolized versions of these integrators (Figure A1a; see Appendix A for more details about the relationship between Metropolized acceptance rates and KL divergence). Consistent with the results of [1], the phase-space error appears to scale approximately as $\mathcal{O}(\Delta t^4)$.

2.2. A simple modification to the near-equilibrium estimator can compute KL divergence in configuration space

The method in Sivak *et al.* [1] approximates the KL divergence between the nonequilibrium steady state density ρ and equilibrium density π in terms of two work averages: the work $\langle w_{\text{shad}} \rangle_\pi$ required to drive from equilibrium π into the steady state ρ, and the steady-state work $\langle w_{\text{shad}} \rangle_\rho$ expended over the same length of time, but starting in ρ (Figure 5a).

We can construct an analogous procedure to estimate the KL divergence of arbitrary marginals of ρ and π, provided we can also sample from an appropriate conditional distribution of π. For example, we are interested in estimating the KL divergence between configuration-space marginals $\rho_{\mathbf{x}}$ and $\pi_{\mathbf{x}}$,

$$\rho_{\mathbf{x}}(\mathbf{x}) \equiv \int d\mathbf{v}\, \rho(\mathbf{x}, \mathbf{v}) \tag{13}$$

$$\pi_{\mathbf{x}}(\mathbf{x}) \equiv \int d\mathbf{v}\, \pi(\mathbf{x}, \mathbf{v})\,. \tag{14}$$

To compute these using the machinery of Sivak *et al.* [1], we replace the second work average $\langle w_{\text{shad}} \rangle_\rho$ with $\langle w_{\text{shad}} \rangle_\omega$, where the expectation of the shadow work is now computed over a modified density constructed from the nonequilibrium steady-state configuration density but with Maxwell-Boltzmann velocity density,

$$\omega(\mathbf{x}, \mathbf{v}) \equiv \rho(\mathbf{x})\pi(\mathbf{v}|\mathbf{x}). \tag{15}$$

Practically, this corresponds to drawing samples from the nonequilibrium steady-state ρ and replacing the velocities $\mathbf{v}$ with an i.i.d. sample from the Maxwell-Boltzmann distribution $\pi(\mathbf{v}|\mathbf{x})$. The modified procedure depicted schematically in Figure 4b.

2.3. Common Langevin integrators induce substantially different timestep-dependent configuration space error

Figure 5b shows that the measured KL divergence between the configuration-space marginals $\rho_{\mathbf{x}}$ and $\pi_{\mathbf{x}}$ can be drastically different among the four integrator schemes, and in some cases grow much more slowly than the associated phase-space sampling error (Fig. 5a). In particular, for VRORV, the error in the $\mathbf{x}$-marginal is very nearly zero for the entire range of feasible timesteps, and it can be run at $\Delta t \approx 6$ fs while introducing the same amount of configuration error as other methods at $\Delta t \approx 2$ fs. This is consistent with prior findings [4,14], which showed the VRORV scheme introduces very little error in the average potential energy and multiple other system-specific observables.

We also note that for the OVRVO scheme, $\mathcal{D}_{\text{KL}}(\rho_{\mathbf{x}} \| \pi_{\mathbf{x}}) \approx \mathcal{D}_{\text{KL}}(\rho \| \pi)$ over the range of measured timesteps (Fig. 5), consistent with Sivak et al.'s prior findings that estimates of $\mathcal{D}_{\text{KL}}(\rho \| \pi)$ tracked well with several measures of configuration-sampling error.

2.4. An exact but expensive estimator of KL divergence validates the near-equilibrium estimate

The accuracy of the near-equilibrium approximation is largely unexplored. While the near-equilibrium approximation introduced by Sivak *et al.* is computationally and statistically appealing, it is important to validate the accuracy of the approximation over the practical timestep Δt range of relevance to molecular simulation. In particular, it is unknown whether the near-equilibrium approximation produces an over-estimate or under-estimate of the KL divergence, or how accurate the approximation is for high-dimensional systems. Further, it is unknown whether any bias introduced by the approximation is uniform across different numerical methods for Langevin dynamics.

How well does the near-equilibrium estimator approximate the true KL divergence of relevant timestep ranges? The task of validating the near-equilibrium approximation is numerically challenging, since we are unaware of exact estimators for $\mathcal{D}_{\mathrm{KL}}(\rho \| \pi)$ that remain tractable in high dimensions. [4] In the case of simple fluids, approximate methods are available that express the KL divergence in terms of a series of N-body correlations (as in [31]), typically truncating to 2-body correlation functions (*i.e.* comparing the radial distribution functions). However, in general we don't know the effect of truncating the expansion, since the successive terms in the series don't necessarily have decreasing magnitude.

Thus, we provide an asymptotically exact reference method. First, we will derive an exact expression for the KL divergence between ρ and π in terms of quantities that we can measure, then discuss practical challenges that arise when using this expression, and under what conditions it becomes impractical. We start by writing the KL divergence as an expectation w.r.t. ρ (17), since we can't evaluate $\rho(\mathbf{x}, \mathbf{v})$ pointwise, but we can draw samples $(\mathbf{x}, \mathbf{v}) \sim \rho$.

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}[\rho \| \pi] &= \int d\mathbf{x}\, d\mathbf{v}\, \rho(\mathbf{x}, \mathbf{v}) \ln \left[\frac{\rho(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})} \right] && (16) \\
&= \left\langle \ln \left[\frac{\rho(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})} \right] \right\rangle_{\rho} && (17) \\
&= \left\langle \ln \left[\frac{\pi(\mathbf{x}, \mathbf{v}) \langle e^{-w} \rangle_{\mathbf{x}, \mathbf{v}; \tilde{\Lambda}}}{\pi(\mathbf{x}, \mathbf{v})} \right] \right\rangle_{\rho} && (18) \\
&= \left\langle \ln \langle e^{-w} \rangle_{\mathbf{x}, \mathbf{v}; \Lambda} \right\rangle_{\rho} && (19) \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \left(\ln \sum_{j=1}^{M} \frac{1}{M} e^{-w_{ij}} \right) && (20)
\end{aligned}$$

We note that the inner ratio of nonequilibrium steady-state to equilibrium densities, $\rho(\mathbf{x}, \mathbf{v}) / \pi(\mathbf{x}, \mathbf{v})$, can be expressed in terms of $\langle e^{-w} \rangle_{\mathbf{x}, \mathbf{v}; \tilde{\Lambda}}$, the average of exponentiated nonequilibrium work measured under the application of the time-reversed protocol $\tilde{\Lambda}$ starting from $(\mathbf{x}, \mathbf{v})$ (18). Λ denotes the protocol used to generate ρ from a sample from π—in this case, T applications of the Langevin integrator step kernel; $\tilde{\Lambda}$ denotes the time-reverse of this protocol. Since the protocol we apply to generate ρ from π is time-symmetric for integrators derived from symmetric Strang splittings, we can substitute $\Lambda = \tilde{\Lambda}$[5]. In the final step, we substitute a simple Monte Carlo estimator of that average, in terms of work samples w_{ij}, where w_{ij} is the jth reduced (unitless) work measurement collected from initial condition i. Here, N is the number of initial conditions sampled (*i.e.* the number of "outer-loop" samples), and M is the number of work samples (*i.e.*, the number of "inner-loop" samples) collected at each initial condition $(\mathbf{x}_i, \mathbf{v}_i) \sim \rho$.

The required work values w_{ij} can be easily computed from simulations. To sample an initial condition i from ρ, we simply run the Langevin integrator of interest for a sufficient number of steps to sample a new uncorrelated configuration from the nonequilibrium steady-state sampled by the integrator. To compute the work accumulated from a given starting condition, we use the notion of *shadow work* [32]. For numerical methods constructed from symmetric Strang splittings involving the R, V, and O operations described above, we simply need to compute sum of the total energy changes during the deterministic substeps (*i.e.*, the potential energy change during deterministic updates of the position variables, and the kinetic energy change during deterministic updates of the momentum

[4] Discretization or density estimation are infeasible, due to curse of dimensionality. There are direct estimators of the KL-divergence based on Euclidean nearest-neighbor distances that perform well in some high-dimensional settings (*e.g.* [30]), but Euclidean distance is an unsuitable metric on molecular configurations.

[5] Note that applying this methodology to non-symmetric integrators (where the integrator and its time-reverse are not identical) would require modifications to this scheme, as well as the manner in which shadow work is computed.

variables). For convenience, we use reduced (unitless) energies and work values throughout, where factors of $k_B T$ has been removed, without loss of generality. See Detailed Methods (Section 4.4) for a detailed description on how shadow work can be computed from this family of Langevin integrators in general.

Like the near-equilibrium scheme, this nested scheme can be modified analogously to measure the configuration-space error in isolation, by initializing instead from the distribution $\omega(\mathbf{x}, \mathbf{v}) \equiv \rho_{\mathbf{x}}(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})$, allowing us to compute $\mathcal{D}_{\mathrm{KL}}[\omega|\pi]$, a quantity that is identical to $\mathcal{D}_{\mathrm{KL}}[\rho_{\mathbf{x}}\|\pi_{\mathbf{x}}]$:

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}[\omega|\pi] &= \int d\mathbf{x}\, d\mathbf{v}\, \omega(\mathbf{x}, \mathbf{v}) \ln \left[\frac{\omega(\mathbf{x}, \mathbf{v})}{\pi(\mathbf{x}, \mathbf{v})}\right] && (21)\\
&= \int d\mathbf{x}\, d\mathbf{v}\, \rho_x(\mathbf{x})\pi(\mathbf{v}|\mathbf{x}) \ln \left[\frac{\rho_x(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})}{\pi_{\mathbf{x}}(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})}\right] && (22)\\
&= \int d\mathbf{x}\, \rho_x(\mathbf{x}) \left[\int d\mathbf{v}\, \pi(\mathbf{v}|\mathbf{x})\right] \ln \left[\frac{\rho_x(\mathbf{x})}{\pi_{\mathbf{x}}(\mathbf{x})}\right] && (23)\\
&= \int d\mathbf{x}\, \rho_x(\mathbf{x}) \ln \left[\frac{\rho_x(\mathbf{x})}{\pi_{\mathbf{x}}(\mathbf{x})}\right] && (24)\\
&= \mathcal{D}_{\mathrm{KL}}[\rho_{\mathbf{x}}\|\pi_{\mathbf{x}}] && (25)
\end{aligned}$$

Specifically, to measure the full KL divergence, we sample initial conditions from the Langevin integrator's steady state: $(\mathbf{x}_i, \mathbf{v}_i) \sim \rho$. To measure configuration-space-only KL divergence, we draw initial configuration from the integrator's steady state, and velocities from equilibrium: $\mathbf{x}_i \sim \rho_{\mathbf{x}}$, $\mathbf{v}_i \sim \pi(\mathbf{v}_i|\mathbf{x}_i)$. Note that, for constrained systems, $\pi(\mathbf{v}|\mathbf{x})$ is not independent of $\mathbf{x}$, and care must be taken to eliminate velocity components along constrained degrees of freedom before measuring the contribution of the integrator substep to the shadow work (see Detailed Methods).

We note that the nested plug-in Monte Carlo estimator of the KL divergence is asymptotically exact only when both N (the number of "outer-loop" samples) and M (the number of "inner-loop" samples) go to infinity. In practice, we use a simple adaptive scheme (described in detail in Section 4.7) that draws inner- and outer-loop samples until uncertainty thresholds are met.

We compared the estimates of the KL divergence on the molecular mechanics system introduced in Figure 5, and confirmed that the two estimates are roughly consistent for all four integrator schemes, over a range of feasible timesteps (Figure 6). At large $\mathcal{D}_{\mathrm{KL}}$, the near-equilibrium scheme deviates from the nested scheme. However, the nested scheme is likely to be biased in the direction of *under-estimating* the actual $\mathcal{D}_{\mathrm{KL}}$ in this region, and the magnitude of this bias increases with the variance of the shadow work distribution. We conclude that the near-equilibrium approximation is empirically reliable for measuring integrator bias on molecular mechanics models for practical timesteps.

3. Discussion

3.1. The near-equilibrium estimator is a fast and useful way to measure integrator error

Here, we introduced and validated a work-based estimator of the KL divergence over the configuration-space marginal sampled by Langevin dynamics. We demonstrated that we could use this estimator to measure differences between the timestep-dependent configuration-sampling error introduced by each member of a small class of Langevin integrators on molecular mechanics models. Work-based estimators are especially attractive for biomolecular systems, since expectations over work distributions are often tractable when other approaches are not.

Reliable estimates of KL divergence using the work-based estimator considered here require knowledge of the time to reach nonequilibrium steady state. This near-equilibrium approach requires the user select a trajectory length T sufficiently large to reach the nonequilibrium steady state, or else the KL divergence estimate could be substantially biased. Opposing the choice of large T is the variance of the estimate, since the contribution of the steady-state work to the variance of the estimate

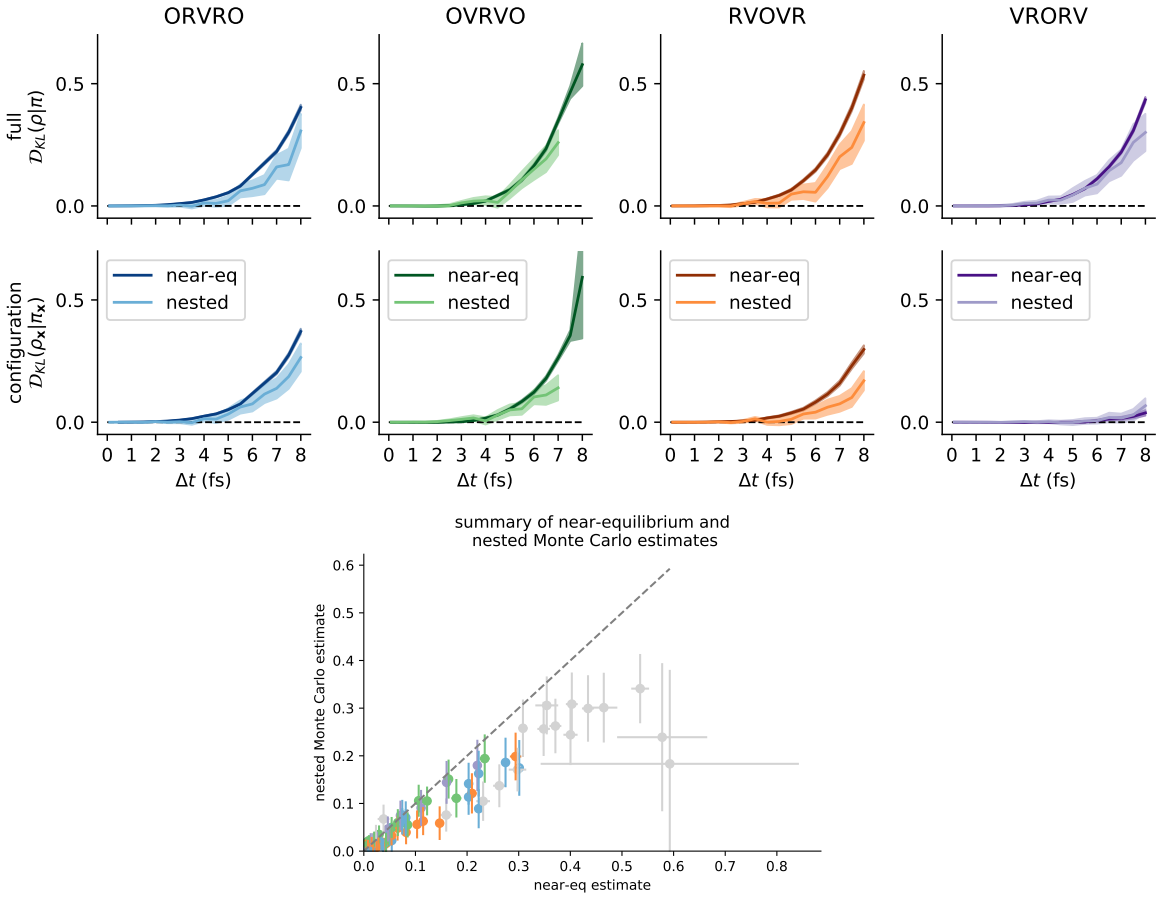

Figure 6. The near-equilibrium estimator is consistent with the nested Monte Carlo estimator for a practical range of Δt. In the **top row**, we validate near-equilibrium estimates of the KL divergence on the full state space $(\mathbf{x}, \mathbf{v})$. In the **bottom row**, we validate near-equilibrium estimates of the KL divergence on configuration space $(\mathbf{x})$ alone. Each column corresponds to a numerical method for Langevin dynamics. The **darker band** in each plot corresponds to the near-equilibrium estimate $\pm$ 95% confidence intervals from asymptotic uncertainty estimate (details in section 4.6). The **lighter band** corresponds to the nested Monte Carlo estimate $\pm$ 95% confidence intervals from bootstrapping (details in section 4.7).

In the **second panel**, we summarize these results by plotting all near-equilibrium estimates vs. all nested Monte Carlo estimates. The colored dots and bars correspond to the means $\pm$ uncertainties used in the earlier panels. The dashed diagonal line shows parity. Grey error dots and error bars correspond to conditions where the nested Monte Carlo estimate reached the computational budget (5×10^4 inner-loop samples) but failed to reach the inner-loop uncertainty threshold, and is thus more biased.

grows as T. Taken together, this suggests the smallest time T that produces unbiased estimates is optimal. In our calculations, it was sufficient to use a protocol that was twice the average collision time, but in general this choice should be validated for the system under study. One way to do this, for example, is to perform the same computation for T and $2T$ and ensure estimates are concordant to within statistical error.

Generating equilibrium samples from $\pi(\mathbf{x}, \mathbf{v})$ can be difficult for large, complex molecular systems. The near-equilibrium method requires access to a large number of independent samples from the equilibrium distribution of the system under study. In this work, we used extra-chance HMC to construct a large cache of independent equilibrium samples, amortizing the cost of equilibrium sampling across the many integrator variants (described in Section 4.3). In cases where we would like to compare a large number of integrator variants on the same system, this can be an acceptable cost, but in other cases, it may be prohibitive. It is unknown whether we can relax this requirement in practice, and use only samples in "local equilibrium," since integrator error may be dominated by local features of the energy landscape. An alternative would be that, if the primary contributions to the dissipation process that drive integration errors arise from high-frequency motions, adding a weak restraint to parts of the system under study (such as a biological macromolecule) may also allow rapid assessment of integrator schemes in a region of configuration space where estimates can be easily converged. We have not yet tested this hypothesis, but if it is possible to relax the requirement of i.i.d. equilibrium samples and retain accurate estimates, then the method will be much cheaper to apply in difficult settings.

3.2. Future directions

The validation of the near-equilibrium estimate makes it possible to apply the technique to a systematic comparison of sampling bias in integrator variants and biomolecular systems. Although we considered only four Langevin integrators here, this approach can be applied to any stochastic integrator for which the relative path action can be computed (see [33] for examples of how to compute the relative action for stochastic integrators not necessarily derived from operator splitting).

Independently, the work-based estimate for $\ln(\rho(\mathbf{x})/\pi(\mathbf{x}))$ we used in the expensive reference method 2.4 could be useful for other analyses. For example, an estimate of $\ln(\rho(\mathbf{x})/\pi(\mathbf{x}))$ could be used to interpret what features of $\mathbf{x}$ are most distorted by integrator bias, *e.g.* by checking which features of $\mathbf{x}$ are most predictive of extreme values of $\ln(\rho(\mathbf{x})/\pi(\mathbf{x}))$.

We also note that nothing about the derivation is specific to the partition between configuration degrees of freedom and velocities. We could also use this method to measure the KL divergence over any subset S of the state variables $\mathbf{z} = (\mathbf{x}, \mathbf{v})$, provided we can sample from the conditional distribution for the complementary subset S' of the state variables: $\pi(\mathbf{z}_{S'}|\mathbf{z}_S)$. To measure KL divergence over the configuration variables, we need only sample from the conditional distribution of velocities given positions, which is typically tractable. Provided that the required conditional distribution is tractable, this method could also prove useful in contexts other than measuring integrator error.

4. Detailed methods

All code used in this paper, along with a manifest of all conda-installable prerequsites and version numbers needed to run the code, is available at https://github.com/choderalab/integrator-benchmark under the permissive OSI-approved MIT license.

A byproduct of this work is a flexible implementation of Langevin integrators derived from operator splitting for the GPU-accelerated OpenMM molecular simulation framework [34], also available under the MIT license in the `openmmtools` library: https://github.com/choderalab/openmmtools. This implementation allows the user to specify a Langevin integrator using a splitting string (like OVRVO) and can automatically compute shadow work.

4.1. One-dimensional model system: Double-well

For illustration and to have a model system where the exact $\mathcal{D}_{\mathrm{KL}}$ was readily computable using histograms, we constructed and analyzed a double-well model in 1D. The potential energy function of this model is $U(x) \equiv x^6 + 2\cos(5(x+1))$, illustrated in Figure 3. We implemented the four Langevin schemes under study using Numba 0.35.0 [35] for use with 1D toy models. We used a temperature of $\beta = 1$, a collision rate of $\gamma = 10$, and a mass of $m = 10$. For these conditions, we found a maximum stable timestep of approximately $\Delta t = 0.7$. Histogram-based estimates of the configuration-space density and phase-space density used 100 bins per dimension, where the bin edges were set by bounding box of a trial run at the maximum Δt. The equilibrium density of each bin was computed using numerical quadrature (using the trapezoidal rule, `numpy.trapz`). The KL divergence between a given ρ and π was then computed using `scipy.stats.entropy` on the histogram representation.

4.2. Model molecular mechanics system: A harmonically restrained water cluster

As noted, the exact Monte Carlo method involves exponential work averages, resulting in a statistical inefficiency that grows rapidly with both the size of the system and the distance from equilibrium. Since we are interested in identifying whether the near-equilibrium approximation breaks down over the timestep Δt range of interest to molecular simulations, it is important to be able to compute a reliable estimate of $\mathcal{D}_{\mathrm{KL}}$ far from equilibrium. Thus, we aim to select the smallest system we think will be representative of the geometry of molecular mechanics models generally in order to allow the exact estimate to be computable with reasonable computing resources.

To compare the proposed method with a reference estimator, we needed to select a test system which met the following criteria:

1. The test system must have **interactions typical of solvated molecular mechanics models**, so that we would have some justification for generalizing from the results. This rules out 1D systems, for example, and prompted us to search for systems that weren't alanine dipeptide in vacuum.
2. The test system must have **sufficiently few degrees of freedom that the nested Monte Carlo estimator remains feasible**. Because the nested estimator requires converging many exponential averages, the cost of achieving a fixed level of precision grows dramatically with the standard deviation of the steady-state shadow work distribution. The width of this distribution is extensive in system size. Empirically, this ruled out using the first water box we had tried (with approximately 500 rigid TIP3P waters [29], with 3000 degrees of freedom). Practically, there was also a limit to how small it is possible to make a water system with periodic boundary conditions in OpenMM (about 100 waters, or 600 degrees of freedom), which was also infeasible.
3. The test system must have **enough disordered degrees of freedom that the behavior of work averages is typical of larger systems**. This was motivated by our observation that it was paradoxically much easier to converge estimates for large disordered systems than it was to converge estimates for the 1D toy system.

To construct a test system that met all of those criteria, we used a `WaterCluster` test system, which comprises 20 rigid TIP3P waters weakly confined in a central harmonic restraining potential with force constant $K = 1$ kJ/mol/nm^2 applied to all atoms. This test system is available in version 0.14.0 of the `openmmtools` package [36]. Simulations were performed in double-precision using the `Reference` platform in OpenMM 7.2 [37] to minimize the potential for introducing significant round-off error due to finite floating point precision.

4.3. Caching equilibrium samples

To enable this study, we attempted to amortize the cost of collecting i.i.d. samples from each test system's equilibrium distribution π and various integrator-and-Δt-specific distributions ρ. Since there are many different ρ distributions and all are relatively small perturbations of π, we invest initial effort

into sampling π exhaustively, and then we draw samples from each integrator-specific ρ by running the integrator of interest from initial conditions $(\mathbf{x}_0, \mathbf{v}_0) \sim \pi$.

For each test system, we pre-computed a large collection of $K = 1000$ equilibrium samples

$$\pi_{\mathbf{x}}(\mathbf{x}) \approx \pi^{\text{cache}}(\mathbf{x}) \equiv \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}^{(k)}) \tag{26}$$

using Extra-Chance Hamiltonian Monte Carlo (XC-HMC) [38,39], implemented as a `CustomIntegrator` in OpenMM [37]. In brief, XC-HMC is a strategy to reduce the adverse effects of momentum-flipping on sampling autocorrelation times in HMC. HMC uses leapfrog integration of Hamiltonian dynamics as a proposal mechanism, and accepts or rejects each proposal according to a Metropolis criterion. Whenever the Metropolis test fails, the proposal is rejected and the momentum must be reversed; this is necessary to maintain detailed balance [3,40], but can lead to extremely large autocorrelation times when the acceptance rate is not sufficiently close to 100%, see [39] for empirical examples and further discussion. In "extra-chance" HMC, rather than immediately flipping the momentum whenever the Metropolis criterion fails, the proposal trajectory is instead extended to generate a new proposal, and another (suitably modified) Metropolis criterion is checked. For suitable choices of parameters (length of trajectory proposal, timestep, number of extra chances, length of "extra-chance" trajectories), this strategy can virtually eliminate the effect of momentum-flipping, at the cost of increasing the average length of proposal trajectories. Our initial experiments (on larger systems than reported here) suggested that the cost of collecting uncorrelated samples using Generalized HMC without "extra-chances" was prohibitive, since we needed to make the timestep extremely small (around 0.1–0.25 fs) to keep the acceptance rate sufficiently near 100% that the effect of momentum-flipping was acceptable.

Specifically, we equilibrated for approximately 1 ns (10^5 XC-HMC iterations, 10 steps per XC-HMC proposal trajectory, 15 extra-chance trajectories per iteration, 1 fs per timestep) from an energy-minimized starting structure. We then saved one sample $\mathbf{x}^{(i)}$ every 10^4 XC-HMC iterations afterwards.

To draw an i.i.d. sample from $\pi(\mathbf{x}, \mathbf{v})$, we draw a sample $\mathbf{x}$ uniformly from π^{cache}, and then sample $\mathbf{v}$ from the equilibrium distribution of velocities conditioned on $\mathbf{x}$, $\mathbf{v} \sim \pi(\mathbf{x}|\mathbf{v})$. (In the presence of holonomic constraints, the velocity distribution is not independent of the configuration distribution. For example, if bond lengths involving hydrogen atoms are constrained, the velocity of a hydrogen minus the velocity of the bonded heavy atom cannot have any component parallel to the bond.)

To draw an i.i.d. sample from ρ, we draw an i.i.d. sample from π and then simulate Langevin dynamics for a large number of steps. We tested using 1000 steps, $(1/\gamma)/\Delta t$ steps, and $(2/\gamma)/\Delta t$ steps.

4.4. Computing shadow work for Strang splittings

Here, we demonstrate how to compute the appropriate shadow work for a given discrete-timestep Langevin integration scheme. While there has been a great deal of confusion in the literature about how nonequilibrium work should be computed [41], fortunately, there is an unambiguous mechanical (if tedious) approach to the computation of the appropriate work-like quantity.

Assemble the sequence of all steps and substeps of an integrator cycle into a trajectory Z. For example, for OVRVO, we have

$$Z \equiv \{z_0 \xrightarrow{\mathsf{O}} z_1 \xrightarrow{\mathsf{V}} z_2 \xrightarrow{\mathsf{R}} z_3 \xrightarrow{\mathsf{V}} z_4 \xrightarrow{\mathsf{O}} z_5\} \tag{27}$$

where $z_n \equiv (\mathbf{x}_n, \mathbf{v}_n)$ are phase space points. Let $\tilde{Z}$ denote the time-reversal of all substeps of Z with all velocities negated,

$$\tilde{Z} \equiv \{\tilde{z}_5 \xrightarrow{\mathsf{O}} \tilde{z}_4 \xrightarrow{\mathsf{V}} \tilde{z}_3 \xrightarrow{\mathsf{R}} \tilde{z}_2 \xrightarrow{\mathsf{V}} \tilde{z}_1 \xrightarrow{\mathsf{O}} \tilde{z}_0\} \tag{28}$$

where $\tilde{z}_n = (\mathbf{x}_n, -\mathbf{v}_n)$ denotes negating the velocity of phase space point z_n.

To compute the reduced, unitless shadow work $w[Z]$, we use the definition of work that satisfies the Crooks fluctuation theorem for starting with a sample from the target equilibrium density $\pi(z_0)$ and taking one integrator cycle step (eqn 4 of [1]):

$$w[Z] = \ln \frac{P[Z|\Lambda]}{P[\tilde{Z}|\tilde{\Lambda}]} + \Delta f_{\mathrm{eq}} \tag{29}$$

where the target equilibrium density $\pi(z) = e^{f-h(z)}$, where f is a log normalizing constant (dimensionless free energy), $h(z) \equiv u(\mathbf{x}) + t(\mathbf{v})$ is the reduced Hamiltonian, $u(\mathbf{x})$ the reduced potential, and $t(\mathbf{v})$ the reduced kinetic energy [19]. The quantity $P[\tilde{Z}|\tilde{\Lambda}]$ denotes the probability of generating the time-reversed trajectory $\tilde{Z}$ by starting with $\tilde{z}_5$ drawn from the target equilibrium density $\pi(\tilde{z}_5)$ (and *not* the nonequilibrium steady state) and applying the reverse sequence of integrator operations $\tilde{\Lambda}$, which is identical to the forward sequence of integrator operations Λ because the integrators we consider here are symmetric. Since the Hamiltonian is time-independent, the free energy change $\Delta f_{\mathrm{eq}} = 0$, and this simplifies to

$$\begin{aligned} w[Z] &= \ln \frac{P[Z|\Lambda]}{P[\tilde{Z}|\tilde{\Lambda}]} = \ln \frac{\pi(z_0)}{\pi(\tilde{z}_5)} \frac{P[Z|z_0]}{P[\tilde{Z}|\tilde{z}_5]} && (30) \\ &= \ln \frac{e^{f-h(z_0)}}{e^{f-h(\tilde{z}_5)}} + \ln \frac{P[Z|z_0]}{P[\tilde{Z}|\tilde{z}_5]} && (31) \\ &= \Delta h[Z] + \ln \frac{P[Z|z_0]}{P[\tilde{Z}|\tilde{z}_5]} && (32) \end{aligned}$$

Computation of the shadow work then proceeds by simple mechanical algebra by computing the log ratio of conditional path probabilities in the last term.

For the family of integrators considered here (symmetric Strang splittings of the propagator, composed of R, V, and O steps), the shadow work has an especially simple form:

$$w[Z] = \Delta h[Z] - \Delta q[Z] \tag{33}$$

where Δh is the total change in reduced Hamiltonian, and Δq is the total change in reduced heat across each of the O substeps. We note that accumulation of the shadow work during integration requires no extra force evaluations, and simply requires knowledge of the potential energy at the beginning and end of the integrator cycle as well as the changes in kinetic energy for each O substep.

For OVRVO, this is

$$\begin{aligned} \Delta h[Z] &\equiv [u(\mathbf{x}_5) + t(\mathbf{v}_5)] - [u(\mathbf{x}_0) + t(\mathbf{v}_0)] && (34) \\ \Delta q[Z] &\equiv [t(\mathbf{v}_1) - t(\mathbf{v}_0)] + [t(\mathbf{v}_5) - t(\mathbf{v}_4)] && (35) \end{aligned}$$

We illustrate how to arrive at this result in detail for OVRVO below.

4.5. Computation of shadow work for OVRVO

For OVRVO, we can represent the forward and reverse of a single integrator cycle diagrammatically as

$$Z \equiv \{z_0 \xrightarrow{\mathsf{O}} z_1 \xrightarrow{\mathsf{V}} z_2 \xrightarrow{\mathsf{R}} z_3 \xrightarrow{\mathsf{V}} z_4 \xrightarrow{\mathsf{O}} z_5\} \tag{36}$$

$$\tilde{Z} \equiv \{\tilde{z}_5 \xrightarrow{\mathsf{O}} \tilde{z}_4 \xrightarrow{\mathsf{V}} \tilde{z}_3 \xrightarrow{\mathsf{R}} \tilde{z}_2 \xrightarrow{\mathsf{V}} \tilde{z}_1 \xrightarrow{\mathsf{O}} \tilde{z}_0\} \tag{37}$$

where $\tilde{z}_n = (\mathbf{x}_n, -\mathbf{v}_n)$ denotes negating the velocity of phase space point z_n.

To compute the conditional path probability $P[Z|z_0]$, we write a transition probability density kernel for each substep:

$$P[Z|z_0] = K_{\mathsf{O}}(z_0, z_1)\, K_{\mathsf{V}}(z_1, z_2)\, K_{\mathsf{R}}(z_2, z_3)\, K_{\mathsf{V}}(z_3, z_4)\, K_{\mathsf{O}}(z_4, z_5) \tag{38}$$

We can write the log ratio of conditional path probabilities as

$$\ln \frac{P[Z|z_0]}{P[\tilde{Z}|\tilde{z}_5]} = \ln \frac{K_{\mathsf{O}}(z_0, z_1)}{K_{\mathsf{O}}(\tilde{z}_1, \tilde{z}_0)} \frac{K_{\mathsf{V}}(z_1, z_2)}{K_{\mathsf{V}}(\tilde{z}_2, \tilde{z}_1)} \frac{K_{\mathsf{R}}(z_2, z_3)}{K_{\mathsf{R}}(\tilde{z}_3, \tilde{z}_2)} \frac{K_{\mathsf{V}}(z_3, z_4)}{K_{\mathsf{V}}(\tilde{z}_4, \tilde{z}_3)} \frac{K_{\mathsf{O}}(z_4, z_5)}{K_{\mathsf{O}}(\tilde{z}_5, \tilde{z}_4)} \tag{39}$$

The probability kernels K_{V} and K_{R} are both deterministic, so as long as we are considering a trajectory Z and its time-reverse $\tilde{Z}$ generated by a symmetric integrator splitting, the ratios involving these kernels are unity.

To compute the ratios involving K_{O} kernels, we note that $K_{\mathsf{O}}(z_0, z_1)$ perturbs the velocity according to the update equation

$$\mathbf{v}_1 = a_2 \mathbf{v}_0 + \sqrt{1 - a_2^2}(\beta M)^{-1/2}\, \xi \tag{40}$$

where ξ is a random variate drawn from the unit normal density, which allows us to solve for the random variate required to propagate from z_0 to z_1,

$$\xi = (1 - a_2^2)^{-1/2}(\beta M)^{+1/2}(\mathbf{v}_1 - a_2 \mathbf{v}_0) \tag{41}$$

$$\xi' = (1 - a_2^2)^{-1/2}(\beta M)^{+1/2}(\tilde{\mathbf{v}}_0 - a_2 \tilde{\mathbf{v}}_1) \tag{42}$$

where the probability density is given by

$$p(\xi) \propto e^{-\frac{1}{2}|\xi|^2}. \tag{43}$$

We can then rewrite the log ratio of O kernels as

$$\ln \frac{K_{\mathsf{O}}(z_0, z_1)}{K_{\mathsf{O}}(\tilde{z}_1, \tilde{z}_0)} = \ln \frac{p(\xi)}{p(\xi')} \tag{44}$$

$$= -\frac{\beta M}{2(1 - a_2)^2} \left[\left(|\mathbf{v}_1|^2 - 2a_2 \mathbf{v}_1 \cdot \mathbf{v}_0 + a_2^2 |\mathbf{v}_0|^2 \right) - \left(|\mathbf{v}_0|^2 - 2a_2 \mathbf{v}_0 \cdot \mathbf{v}_1 + a_2^2 |\mathbf{v}_1|^2 \right) \right] \tag{45}$$

$$= -\frac{\beta M}{2(1 - a_2)^2} (1 - a_2)^2 \left(|\mathbf{v}_1|^2 - |\mathbf{v}_0|^2 \right) \tag{46}$$

$$= -[t(\mathbf{v}_1) - t(\mathbf{v}_0)] \tag{47}$$

Combining this with (33), this provides the overall work as

$$w[Z] = [u(\mathbf{x}_5) - u(\mathbf{x}_0)] + [t(\mathbf{v}_5) - t(\mathbf{v}_0)] - [t(\mathbf{v}_1) - t(\mathbf{v}_0)] - [t(\mathbf{v}_5) - t(\mathbf{v}_4)] \tag{48}$$

4.6. Near-equilibrium estimators for KL divergence

We seek the KL divergence between the stationary distribution produced by a given Langevin integrator, and the desired equilibrium distribution. Sivak and Crooks [42] derived a near-equilibrium estimator, in terms of work averages, for the KL divergence between an arbitrary distribution ρ and the equilibrium distribution π. Sivak, Chodera, and Crooks [1] demonstrated how to apply this estimator to the case where ρ is the biased stationary distribution of a single numerical Langevin integrator (OVRVO). For ρ, the stationary distribution resulting from repeated application of the Langevin integrator, the KL divergence is approximately given by,

$$\mathcal{D}_{\mathrm{KL}}(\rho\|\pi) \approx \frac{1}{2}\left(\langle w\rangle_{\pi;\Lambda} - \langle w\rangle_{\pi,\Lambda;\Lambda}\right) \tag{49}$$

$$= \frac{1}{2}\left(\langle w\rangle_{\pi;\Lambda} - \langle w\rangle_{\rho;\Lambda}\right) \tag{50}$$

$$= \frac{1}{2}\left(\langle w_{\mathrm{shad}}\rangle_{\pi} - \langle w_{\mathrm{shad}}\rangle_{\rho}\right) \ . \tag{51}$$

Here $\langle\cdots\rangle_{p;\Lambda}$ indicates an average over the dynamical ensemble produced by initialization in microstates sampled from density p and subsequent driving by protocol Λ. Note that $\langle w_{\mathrm{shad}}\rangle_{\pi,\Lambda;\Lambda}$ represents an expectation computed with respect to an initial distribution (π,Λ) prepared by first sampling over π followed by the application of protocol Λ, subsequently measuring the shadow work over the application of the protocol Λ again; this is distinct from $\langle w_{\mathrm{shad}}\rangle_{\pi;\Lambda}$, which denotes the expectation where the initial sample is selected from π and the shadow work is measured during the execution of protocol Λ.

In this study, we are especially interested in the configuration-space marginal distribution $\rho_{\mathbf{x}}$, so we introduce a distribution $\omega(\mathbf{x},\mathbf{v}) \equiv \rho_{\mathbf{x}}(\mathbf{x})\pi(\mathbf{v}|\mathbf{x})$ that differs from π only in its $\mathbf{x}$-marginal (so that $\mathcal{D}_{\mathrm{KL}}(\omega\|\pi) = \mathcal{D}_{\mathrm{KL}}(\rho_{\mathbf{x}}\|\pi_{\mathbf{x}})$) and compute the near-equilibrium approximation for the KL divergence between ω and π:

$$\mathcal{D}_{\mathrm{KL}}(\rho_{\mathbf{x}}\|\pi_{\mathbf{x}}) = \mathcal{D}_{\mathrm{KL}}(\omega\|\pi) \approx \frac{1}{2}\left(\langle w_{\mathrm{shad}}\rangle_{\pi} - \langle w_{\mathrm{shad}}\rangle_{\omega}\right) \tag{52}$$

4.7. Variance-controlled adaptive estimator for KL divergence

As noted in Section 2.4, the nested Monte Carlo estimator we use as an expensive, but in principle asymptotically exact, estimate of the KL divergence requires converging a separate exponential average $\ln\langle e^{-w}\rangle\mathbf{x},\mathbf{v};\Lambda$ for every sample $(\mathbf{x},\mathbf{v}) \sim \rho$ or ω. It is obviously impossible to compute this exactly; any practical approach employing finite computational resources can only estimate this quantity to some finite statistical precision, and even then, the logarithm will induce some bias in the computed estimate for finite sample sizes. Here, we take the approach of setting a sensible target statistical error for this inner estimate, arbitrarily selecting 0.01, since we would like to resolve features in $\mathcal{D}_{\mathrm{KL}}$ larger than this magnitude. Notably, the difficulty in achieving this threshold increases exponentially as the width of the sampled distribution $p(w)$ increases with increasing timestep Δt.

To determine the number of inner-loop samples required to meet this statistical error threshold, we periodically compute an estimate of $\mathrm{stdev}(\ln\widehat{\langle e^{-w}\rangle}_{x;\Lambda})$ from the available samples using the first term of a Taylor expansion. We compare the estimated standard deviation with the user-defined threshold, and continue to draw samples until we meet the threshold or exceed the sample budget of 5×10^4 samples[6]. The scaling of computational effort with Δt is shown in Figure 7.

Choosing the inner-loop threshold is subtle. If the inner-loop threshold is chosen too large, then the resulting estimate will be very biased (the whole procedure only becomes unbiased in the limit that

[6] This limited the maximum CPU time spent collecting an "outer-loop" sample to approximately 2 hours – conditions where this limit was met are colored grey in the lower panel of Figure 6.

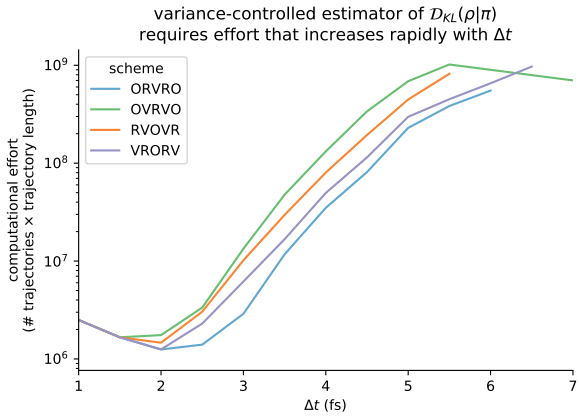

Figure 7. The computational effort required to reach a fixed uncertainty threshold depends sharply on Δt. The total computational effort is determined by the total number of trajectories sampled in that condition (*i.e.* $\sum_{i=1}^{N} M_i$), multiplied by the length of each trajectory. Note that the curves are not monotonic, since the number of required trajectories increases superlinearly with Δt, but the number of timesteps in each trajectory decreases linearly with Δt.

$\sigma_{\text{inner}} \to 0$). If the threshold is chosen too small, then the computational effort becomes prohibitive. Controlling the *precision* of the inner-loop estimates should also be expected to control their *bias*, since the bias of the inner-loop estimates is approximately $\sigma_{\text{inner}}^2/2M$ (see Section II.B, eqn. 8 in [43]), in the direction of *under-estimating* the $\mathcal{D}_{\text{KL}}$.

To compute and report uncertainty in Figure 6, we use bootstrap resampling, rather than Taylor propagation. The data for each condition is a jagged array of sampled work values, where each row represents an "outer-loop sample" (*i.e.*, a different initial condition $(\mathbf{x}, \mathbf{v})$ sampled from ρ (or ω)), and the length of each row is variable, reflecting the number of "inner-loop" samples required to estimate $\ln\left[\frac{\rho(\mathbf{x},\mathbf{v})}{\pi(\mathbf{x},\mathbf{v})}\right]$ (or $\ln\left[\frac{\omega(\mathbf{x},\mathbf{v})}{\pi(\mathbf{x},\mathbf{v})}\right]$) to the desired precision. To generate a single bootstrap sample, we resample first the rows uniformly with replacement, and then, within each row, resample the columns uniformly with replacement. The error bands for the "exact" estimator in Figure 6 are computed from 100 bootstrap samples per condition.

Appendix Relation to GHMC acceptance rates

What is the relationship between the bias introduced by an integrator at steady state, and the acceptance rate of the corresponding Metropolized integrator? Specifically, why not just Metropolize VRORV to guarantee samples are drawn appropriately from the equilibrium target density $\pi(\mathbf{x})$? Following [3,38], we can construct an exact MCMC method that uses one or more steps of Langevin dynamics as a proposal, by using

$$\alpha \equiv \min[1, \exp(-w_{\text{shad}})] \tag{A1}$$

as the acceptance criterion. The resulting method is called "GHMC" (Generalized Hybrid Monte Carlo). A natural question arises: if an integrator introduces low configuration-space error, is the rejection rate of the corresponding GHMC method also low?

To answer this question, we estimated the GHMC acceptance rate at all conditions for which we have estimated steady-state $\mathcal{D}_{\text{KL}}$. Given a collection of equilibrium samples (described in Section 4.3), we can efficiently estimate the acceptance rate of an MCMC proposal by taking the sample average of the acceptance ratio α over proposals originating from equilibrium, $(\mathbf{x}_0, \mathbf{v}_0) \sim \pi$.

We compared the GHMC acceptance rate to the histogram-based $\mathcal{D}_{\text{KL}}$ estimates for the 1D double-well system in Figure A1. There does not appear to be a consistent relationship between

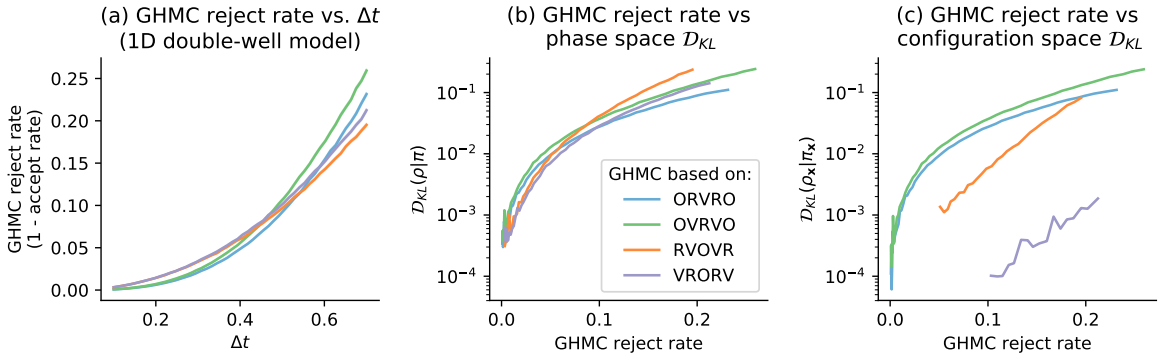

Figure A1. No consistent relationship between the GHMC acceptance rate and the steady-state bias is apparent for a 1D system. Since the GHMC rejection rate grows similarly with Δt across all four schemes, but the configuration space KL divergence does not, the GHMC rejection rate can be overly "conservative" for some splittings. Panel **(a)** shows the growth in the GHMC rejection rate as a function of timestep Δt, for the 1D double-well model considered in Figures 2 and 3. On the x-axis is an evenly spaced grid of 50 timesteps between 0.1 and 0.7. On the y-axis is the estimated rejection rate, which is based on a sample average of the GHMC acceptance criterion. The shaded region is the mean $\pm$ 95% confidence interval. Panel **(b)** compares the GHMC rejection rate vs. the *phase-space* bias at steady state, over the range of timesteps plotted in panel (a). The y-axis is KL divergence between the phase-space histograms, plotted on a log-scale. Panel **(c)** compares the GHMC rejection rate vs. the *configuration-space* bias at steady state, over the range of timesteps plotted in panels (a), (b). The y-axis is the KL divergence between the configuration-space histograms, plotted on a log-scale. Note that in panel (c), we have truncated the leftmost parts of the curves for RVOVR and VRORV rejection rates less than 0.05 and 0.1, respectively, due to noise in histogram estimates of very small $\mathcal{D}_{\mathrm{KL}}(\rho_{\mathbf{x}} \| \pi_{\mathbf{x}})$.

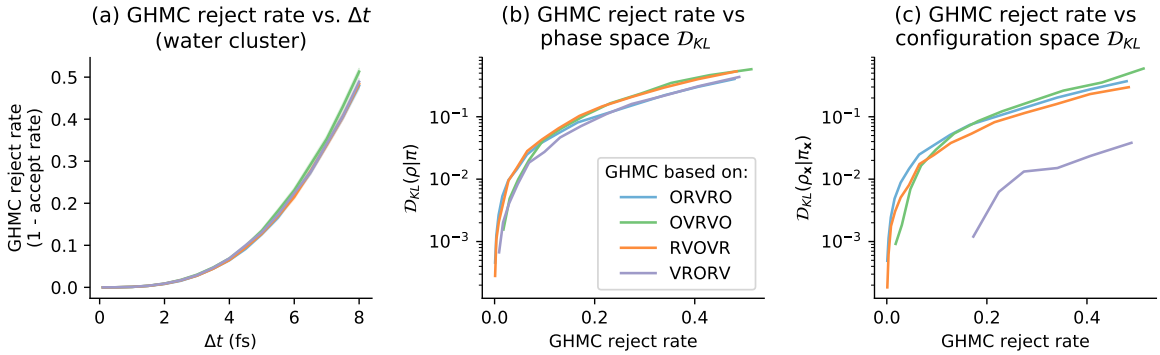

Figure A2. Near-equilibrium measurements recapitulate the relationship between steady-state $\mathcal{D}_{\mathrm{KL}}$ and GHMC acceptance rate for the water cluster test system. Panel **(a)** shows the growth in the GHMC rejection rate (1 minus the acceptance rate) as a function of timestep Δt (in femtoseconds), for the water cluster test system illustrated in Figure 5. On the x-axis are timesteps [0.1 fs, 0.5 fs, 1.0 fs, ...7.5 fs, 8fs]. On the y-axis is the estimated rejection rate, which is based on a sample average of the GHMC acceptance criterion, over 10000 proposals per condition. The shaded region is the mean $\pm$ 95% confidence interval. Panel **(b)** compares the GHMC rejection rate vs. the *phase-space* bias at steady state, over the range of timesteps plotted in panel (a). The y-axis is the KL divergence between the phase-space distributions as measured by the near-equilibrium estimate, plotted on a log-scale. Panel **(c)** compares the GHMC rejection rate vs. the *configuration-space* bias at steady state, over the range of timesteps plotted in panels (a), (b). Note that in panels (b) and (c), we have truncated at $\mathcal{D}_{\mathrm{KL}} \leq 10^{-4}$, due to noise in near-equilibrium estimates of very small $\mathcal{D}_{\mathrm{KL}}$.

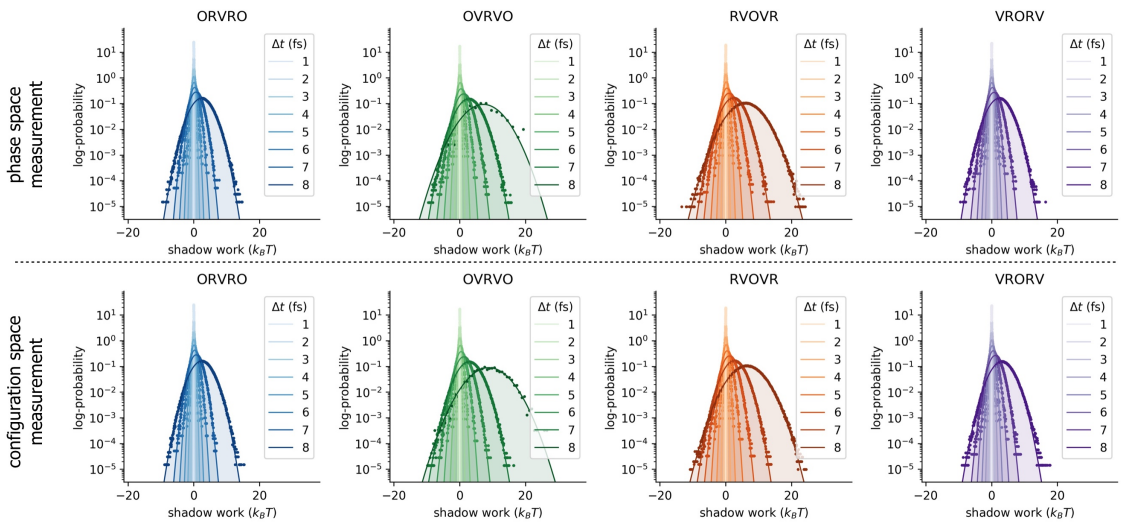

Figure A3. Shadow work distributions for the water cluster are approximately Gaussian for all integrators examined. In all panels, solid lines and shaded regions denote Gaussian fits, while dots denote histogram estimates. The **top row** depicts work distributions where initial conditions are sampled from the nonequilibrium steady-state induced by the corresponding integrator and timestep ($(\mathbf{x}, \mathbf{v}) \sim \rho$); these shadow work values are used to measure phase-space error in the near-equilibrium estimates of $\mathcal{D}_{\mathrm{KL}}$. The **bottom row** depicts work distributions where initial conditions are sampled from the ω ensemble ($\mathbf{x} \sim \rho_{\mathbf{x}}$, $\mathbf{v} \sim \pi(\mathbf{v}|\mathbf{x})$); these work values are used to estimate configuration-space error in the near-equilibrium estimates of $\mathcal{D}_{\mathrm{KL}}$.

$\mathcal{D}_{\mathrm{KL}}$ and acceptance rate across the four schemes. Notably, the GHMC reject rate can be extremely "conservative" for splittings such as VRORV.

Next, we compared the GHMC rejection rate with the near-equilibrium $\mathcal{D}_{\mathrm{KL}}$ estimates for the water cluster considered in Figure A2. A similar pattern is recapitulated in this molecular mechanics model as in the 1D system—there is not a consistent relationship between configuration-space bias introduced by a Langevin integrator and the rejection rate of its corresponding GHMC method.

This complicates the decision of whether to Metropolize or not. As noted in Section 4.3, incurring even a small rejection rate in GHMC can have a large effect on statistical efficiency, due to the effect of momentum flipping. An open challenge is to construct Metropolis criteria for GHMC that might be less "wasteful" for Langevin splittings that introduce low configuration-space bias.

Appendix Statistics of shadow work distributions

The exact expression for the KL divergence used for validation in Section 2.4 requires estimating the expectation of e^{-w} averaged over $p(w)$. Work distributions for various integrators and timesteps are plotted in Figure A3, and appear to be approximately Gaussian, as can be seen by comparison with Gaussian fits (solid lines). As expected, the width of the work distributions increases with increasing timestep, which can be seen more clearly in Figure A4, which plots the standard deviation of the work distribution as a function of timestep.

While the near-equilibrium estimate will find the difficulty of reaching estimates of a given statistical precision grows linearly with the variance in $p(w)$, the exact estimator must converge expectations of the *exponentiated* shadow work, which becomes exponentially difficult with increasing variance. This effect is illustrated in Figure A5, where we plot e^{-w} along with the Gaussian fits to these work distributions.

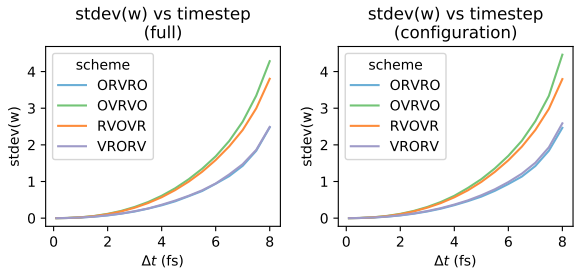

Figure A4. Standard deviation of water cluster work distribution grows with time step. The left panel summarizes the top row of Figure A3 (shadow work distributions for trajectories initialized in π), and the right panel summarizes the bottom row of Figure A3 (trajectories initialized in ω.

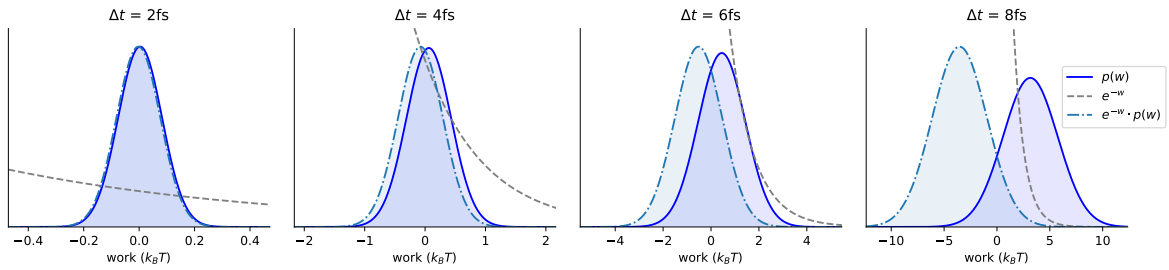

Figure A5. Exponential averages with respect to the shadow work distribution become increasingly difficult with increasing timestep. It becomes increasingly difficult to estimate the expectation of e^{-w} with respect to Gaussian fits to work distributions $p(w)$ for the water cluster, as the timestep Δt increases. The four panels increase in Δt from left to right: note the changing x-axis scales. The **solid line** is the shadow work distribution $p(w)$ measured at each timestep. The **dashed line** is e^{-w}. The **dash-dotted line** is $e^{-w} \cdot p(w)$.

Acknowledgments: The authors gratefully acknowledge members of Ben Leimkuhler's lab (University of Edinburgh) for stimulating discussions on the nature of errors and efficiencies of Langevin integrators; Grant Rotskoff (University of California, Berkeley) for his input on nonequilibrium shadow work measurement schemes; Charlie Matthews and Brian Radak (University of Chicago) for discussions on geodesic integrators; Jason Wagoner (Stony Brook) for discussions about Metropolized Langevin integrators; and members of the Chodera laboratory for input on various aspects of the implementation and theory. We are also grateful to Peter Eastman (Stanford) for implementing the `CustomIntegrator` facility within OpenMM that greatly simplifies the implementation of the integrators considered here. We also thank the *Entropy* editors for their patience. JDC acknowledges support from the Sloan Kettering Institute and NIH grant P30 CA008748 and NIH grant R01 GM121505. JF acknowledges support from NSF grant CHE 1738979. DAS acknowledges support from a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and the Canada Research Chairs program.

1. Sivak, D.A.; Chodera, J.D.; Crooks, G.E. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Physical Review X* **2013**, *3*.
2. Lemons, D.S.; Gythiel, A. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530-533 (1908)]. *American Journal of Physics* **1997**, *65*, 1079–1081.
3. Lelièvre, T.; Stoltz, G.; Rousset, M. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press: London ; Hackensack, N.J, 2010. OCLC: ocn244765923.
4. Leimkuhler, B.; Matthews, C. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*; Springer: Cham, 2015. OCLC: 914391557.
5. Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press, Inc.: Orlando, FL, USA, 2001.
6. Maruyama, G. Continuous Markov Processes and Stochastic Equations. *Rendiconti del Circolo Matematico di Palermo* **1955**, *4*, 48–90.
7. Ermak, D.L.; Yeh, Y. Equilibrium Electrostatic Effects on the Behavior of Polyions in Solution: Polyion-Mobile Ion Interaction. *Chemical Physics Letters* **1974**, *24*, 243–248.
8. Brünger, A.; Brooks, C.L.; Karplus, M. Stochastic Boundary Conditions for Molecular Dynamics Simulations of ST2 Water. *Chemical Physics Letters* **1984**, *105*, 495–500.
9. Pastor, R.W.; Brooks, B.R.; Szabo, A. An Analysis of the Accuracy of Langevin and Molecular Dynamics Algorithms. *Molecular Physics* **1988**, *65*, 1409–1419.
10. Athénes, M. A Path-Sampling Scheme for Computing Thermodynamic Properties of a Many-Body System in a Generalized Ensemble. *The European Physical Journal B* **2004**, *38*, 651–663.
11. Adjanor, G.; Athènes, M.; Calvo, F. Free Energy Landscape from Path-Sampling: Application to the Structural Transition in LJ38. *The European Physical Journal B* **2006**, *53*, 47–60.
12. Bussi, G.; Parrinello, M. Accurate Sampling Using Langevin Dynamics. *Physical Review E* **2007**, *75*, 056707.
13. Izaguirre, J.A.; Sweet, C.R.; Pande, V.S. Multiscale Dynamics of Macromolecules Using Normal Mode Langevin. In *Biocomputing 2010*; World Scientific, 2009; pp. 240–251.
14. Leimkuhler, B.; Matthews, C. Robust and Efficient Configurational Molecular Sampling via Langevin Dynamics. *The Journal of Chemical Physics* **2013**, *138*, 174102.
15. Leimkuhler, B.; Matthews, C. Efficient Molecular Dynamics Using Geodesic Integration and Solvent–solute Splitting. *Proc. R. Soc. A* **2016**, *472*, 20160138.
16. Sivak, D.A.; Chodera, J.D.; Crooks, G.E. Time Step Rescaling Recovers Continuous-Time Dynamical Properties for Discrete-Time Langevin Integration of Nonequilibrium Systems. *The Journal of Physical Chemistry B* **2014**, *118*, 6466–6474.
17. Swope, W.C.; Andersen, H.C.; Berens, P.H.; Wilson, K.R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *The Journal of Chemical Physics* **1982**, *76*, 637–649.
18. Leimkuhler, B.; Matthews, C. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Appl. Math. Res. EXpress* **2012**, [1203.5428].
19. Shirts, M.R.; Chodera, J.D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

20. Leimkuhler, B.; Matthews, C. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*; Springer: Cham, 2015. OCLC: 914391557.
21. Bennett, C.H. Mass Tensor Molecular Dynamics. *Journal of Computational Physics* **1975**, *19*, 267–279.
22. Hopkins, C.W.; Le Grand, S.; Walker, R.C.; Roitberg, A.E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation* **2015**, *11*, 1864–1874.
23. Pomès, R.; McCammon, J.A. Mass and Step Length Optimization for the Calculation of Equilibrium Properties by Molecular Dynamics Simulation. *Chemical Physics Letters* **1990**, *166*, 425–428.
24. Plecháč, P.; Rousset, M. Implicit Mass-Matrix Penalization of Hamiltonian Dynamics with Application to Exact Sampling of Stiff Systems. *Multiscale Modeling & Simulation* **2010**, *8*, 498–539.
25. Grubmüller, H.; Tavan, P. Multiple Time Step Algorithms for Molecular Dynamics Simulations of Proteins: How Good Are They? *Journal of Computational Chemistry* **1998**, *19*, 1534–1552.
26. Butler, B.D.; Ayton, G.; Jepps, O.G.; Evans, D.J. Configurational Temperature: Verification of Monte Carlo Simulations. *The Journal of Chemical Physics* **1998**, *109*, 6519–6522.
27. Leimkuhler, B.; Matthews, C. Efficient Molecular Dynamics Using Geodesic Integration and Solvent–solute Splitting. *Proc. R. Soc. A* **2016**, *472*, 20160138.
28. Sweet, C.R.; Hampton, S.S.; Skeel, R.D.; Izaguirre, J.A. A Separable Shadow Hamiltonian Hybrid Monte Carlo Method. *The Journal of Chemical Physics* **2009**, *131*, 174106.
29. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.
30. Perez-Cruz, F. Kullback-Leibler Divergence Estimation of Continuous Distributions. 2008 IEEE International Symposium on Information Theory, 2008, pp. 1666–1670.
31. Dhabal, D.; Nguyen, A.H.; Singh, M.; Khatua, P.; Molinero, V.; Bandyopadhyay, S.; Chakravarty, C. Excess Entropy and Crystallization in Stillinger-Weber and Lennard-Jones Fluids. *The Journal of Chemical Physics* **2015**, *143*, 164512.
32. Sivak, D.A.; Chodera, J.D.; Crooks, G.E. Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Physical Review X* **2013**, *3*, 001007+.
33. Nilmeier, J.P.; Crooks, G.E.; Minh, D.D.L.; Chodera, J.D. Nonequilibrium Candidate Monte Carlo Is an Efficient Tool for Equilibrium Simulation. *PNAS* **2011**, *108*, E1009–E1018.
34. Eastman, P.; Swails, J.; Chodera, J.D.; McGibbon, R.T.; Zhao, Y.; Beauchamp, K.A.; Wang, L.P.; Simmonett, A.C.; Harrigan, M.P.; Stern, C.D.; Wiewiora, R.P.; Brooks, B.R.; Pande, V.S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017**, *13*, 1–17.
35. Lam, S.K.; Pitrou, A.; Seibert, S. Numba: A LLVM-Based Python JIT Compiler. Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC; ACM: New York, NY, USA, 2015; LLVM '15, pp. 7:1–7:6.
36. Chodera, J.; Rizzi, A.; Naden, L.; Beauchamp, K.; Grinaway, P.; Fass, J.; Rustenburg, B.; Ross, G.A.; Simmonett, A.; Swenson, D.W. Openmmtools: 0.14.0 - Exact Treatment of Alchemical PME Electrostatics, Water Cluster Test System, Optimizations.
37. Eastman, P.; Swails, J.; Chodera, J.D.; McGibbon, R.T.; Zhao, Y.; Beauchamp, K.A.; Wang, L.P.; Simmonett, A.C.; Harrigan, M.P.; Stern, C.D.; Wiewiora, R.P.; Brooks, B.R.; Pande, V.S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Computational Biology* **2017**, *13*, e1005659.
38. Campos, C.M.; Sanz-Serna, J.M. Extra Chance Generalized Hybrid Monte Carlo. *Journal of Computational Physics* **2015**, *281*, 365–374, [1407.8107].
39. Wagoner, J.A.; Pande, V.S. Reducing the Effect of Metropolization on Mixing Times in Molecular Dynamics Simulations. *The Journal of Chemical Physics* **2012**, *137*, 214105, [1209.5944].
40. Akhmatskaya, E.; Bou-Rabee, N.; Reich, S. Erratum to "A Comparison of Generalized Hybrid Monte Carlo Methods with and without Momentum Flip" [J. Comput. Phys. 228 (2009), S. 2256 - 2265] **2009**.
41. Adib, A.B. Comment on "On the Crooks fluctuation theorem and the Jarzynski equality" [J. Chem. Phys. 129, 091101 (2008)]. *The Journal of Chemical Physics* **2009**, *130*, 247101.
42. Sivak, D.A.; Crooks, G.E. Near-Equilibrium Measurements of Nonequilibrium Free Energy. *Physical Review Letters* **2012**, *108*.

43. Shirts, M.R.; Pande, V.S. Comparison of Efficiency and Bias of Free Energies Computed by Exponential Averaging, the Bennett Acceptance Ratio, and Thermodynamic Integration. *The Journal of Chemical Physics* **2005**, *122*, 144107.

© 2018 by the authors. Submitted to *Entropy* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).