SUPPORTING INFORMATION

I. VELOCITY INITIALIZATION IN HYBRID MONTE CARLO

The discussion of Monte Carlo (MC) techniques presented herein closely follows recent work by Allen and Quigley [1]. In standard MC sampling schemes, detailed balance is satisfied by accepting trial moves using the criterion

$$P_{\text{acc}}^{\text{old}\to\text{new}} = \min\left(1, \frac{\varrho(\boldsymbol{r}^{\text{new}})\alpha^{\text{new}\to\text{old}}}{\varrho(\boldsymbol{r}^{\text{old}})\alpha^{\text{old}\to\text{new}}}\right) \tag{1}$$

where $\boldsymbol{r}$ is the complete set of coordinates $\{\boldsymbol{r}_i\}_{i=1}^N$ for a N molecule system, $\varrho(\boldsymbol{r})$ is the stationary distribution, and $\alpha^{\text{old}\to\text{new}}$ is the probability of randomly selecting a trial move. If configurations are sampled, for example, from the canonical distribution $\varrho(\boldsymbol{r}) \propto e^{-\beta U(\boldsymbol{r})}$,

$$P_{\text{acc}}^{\text{old}\to\text{new}} = \min\left(1, e^{-\beta\Delta U}\frac{\alpha^{\text{new}\to\text{old}}}{\alpha^{\text{old}\to\text{new}}}\right), \tag{2}$$

where $\beta = (k_B T)^{-1}$ and $\Delta U = U^{\text{new}} - U^{\text{old}}$ is the change in potential energy.

In hybrid MC (HMC), short microcanonical molecular dynamics (MD) trajectories propagated with a time-reversible and volume-preserving integrator are used as trial moves to perform canonical sampling [2]. Each trajectory is initialized using a complete set of velocities $\boldsymbol{v}$. For rigid bodies, $\boldsymbol{v} = \{\boldsymbol{v}_i^{\text{cm}}, \hat{\boldsymbol{\omega}}_i\}_{i=1}^N$ where $\boldsymbol{v}_i^{\text{cm}}$ is the center of mass velocity and $\hat{\boldsymbol{\omega}}_i$ is the angular velocity in the body-centered frame in which the inertia tensor is a diagonal matrix $\hat{\boldsymbol{I}}_i$. To correct Limmer and Chandler's (LC) code [4, 5], we implemented a routine to draw velocities at random from the Maxwell-Boltzmann (MB) distribution. Components of $\boldsymbol{v}_i^{\text{cm}}$ and $\hat{\boldsymbol{\omega}}_i$ are initialized using

$$v_{i,j}^{\text{cm}} = (\beta m_i)^{-1/2}\,\mathcal{G}_{i,j} \tag{3a}$$

$$\hat{\omega}_{i,j} = \left(\beta\hat{I}_{i,jj}\right)^{-1/2}\mathcal{G}'_{i,j} \tag{3b}$$

for $j = x, y, z$, where $\mathcal{G}_{i,j}$ and $\mathcal{G}'_{i,}$ are Gaussian random numbers with unit variance and zero mean and m_i is the molecular mass. The resulting velocities obey MB statistics

$$P_{\text{MB}}(\boldsymbol{v}^{\text{cm}}, \hat{\boldsymbol{\omega}}) = P_{\text{MB}}(\boldsymbol{v}^{\text{cm}}) \times P_{\text{MB}}(\hat{\boldsymbol{\omega}}) \tag{4}$$

$$* = \left[\prod_{i=1}^{N}\left(\frac{\beta m_i}{2\pi}\right)^{3/2} e^{-\frac{\beta}{2}m_i(\boldsymbol{v}_i^{\text{cm}}\cdot\boldsymbol{v}_i^{\text{cm}})}\right] \times$$

$$\left[\prod_{i=1}^{N}\left(\frac{\beta}{2\pi}\right)^{3/2}(\hat{I}_{i,xx}\hat{I}_{i,yy}\hat{I}_{i,zz})^{1/2}e^{-\frac{\beta}{2}(\hat{\boldsymbol{\omega}}_i^T\hat{\boldsymbol{I}}_i\hat{\boldsymbol{\omega}}_i)}\right].$$

Simplification using the definition of kinetic energy $K(\boldsymbol{v}^{\text{cm}}, \hat{\boldsymbol{\omega}}) \equiv \frac{1}{2}\sum_i[m_i(\boldsymbol{v}_i^{\text{cm}}\cdot\boldsymbol{v}_i^{\text{cm}}) + \hat{\boldsymbol{\omega}}_i^T\hat{\boldsymbol{I}}_i\hat{\boldsymbol{\omega}}_i]$ leads to

$$P_{\text{MB}}(\boldsymbol{v}^{\text{cm}}, \hat{\boldsymbol{\omega}}) = Ce^{-\beta K(\boldsymbol{v}^{\text{cm}}, \hat{\boldsymbol{\omega}})} \tag{5}$$

where C is a temperature-dependent constant. The ratio of probability densities for selecting the forward and reverse moves can therefore be expressed in terms of $\Delta K = K^{\text{new}} - K^{\text{old}}$:

$$\frac{\alpha^{\text{new}\to\text{old}}}{\alpha^{\text{old}\to\text{new}}} = \frac{P_{\text{MB}}(\boldsymbol{v}^{\text{cm,new}}, \hat{\boldsymbol{\omega}}^{\text{new}})}{P_{\text{MB}}(\boldsymbol{v}^{\text{cm,old}}, \hat{\boldsymbol{\omega}}^{\text{old}})} = e^{-\beta\Delta K}. \tag{6}$$

Substitution into Eq. 2 yields the standard HMC acceptance criterion

$$P_{\text{acc}}^{\text{old}\to\text{new}} = \min\left(1, e^{-\beta\Delta U}e^{-\beta\Delta K}\right). \tag{7}$$

As shown in Ref. 3, initial velocities may be drawn from other distributions besides Eq. 4, but the acceptance criterion (Eq. 7) should be modified accordingly to satisfy detailed balance.

Velocities $(\boldsymbol{v}_i^{\text{cm}}, \hat{\boldsymbol{\omega}}_i)$ are the natural choice for describing the kinematics of rigid bodies. LC's HMC code [4, 5], however, uses routines from LAMMPS [6] to generate trial MD trajectories, which are accepted according to the standard HMC criterion (Eq. 7). Although the rigid body integrator employed in their sampling scheme (LAMMPS' *fix nve/rigid molecule* [7]) uses velocities $(\boldsymbol{v}_i^{\text{cm}}, \hat{\boldsymbol{\omega}}_i)$ internally, it computes initial values of these quantities from site velocities $\boldsymbol{v}_{i,k}$ in the space frame (frame defined by the simulation cell) prior to performing time integration. For N rigid water molecules with M interaction sites each, kinetic energy is evaluated from $\boldsymbol{v}_{i,k}$ using $K = \frac{1}{2}\sum_{i,k} m_{i,k}(\boldsymbol{v}_{i,k}\cdot\boldsymbol{v}_{i,k})$ for $i \in N, k \in M$, where $m_{i,k}$ are the site masses. Hence LC's code requires passing a set of space-frame site velocities $\boldsymbol{v}_{i,k}$ directly to LAMMPS. As we explain below in Section II, $\boldsymbol{v}_{i,k}$ are assigned incorrectly at the beginning of each MD trial trajectory in the LC code. We corrected this issue by first drawing initial velocities $(\boldsymbol{v}_i^{\text{cm},0}, \hat{\boldsymbol{\omega}}_i^0)$ from the MB distribution. The initial space-frame site velocities passed to LAMMPS are then obtained using the standard transformation $\boldsymbol{v}_{i,k}^0 = \mathcal{T}(\boldsymbol{v}_i^{\text{cm},0}, \hat{\boldsymbol{\omega}}_i^0)$ described in Section III. This procedure ensures that (i) K fluctuations obey canonical statistics, as required to satisfy detailed balance; (ii) K is *equipartitioned* among translational and rotational molecular degrees of freedom; and (iii) $\boldsymbol{v}_{i,k}^0$ satisfy the *rigidity constraint* $\left|\tilde{\boldsymbol{v}}_{i,k}^{\parallel}\right| = (\boldsymbol{v}_{i,k} - \boldsymbol{v}_i^{\text{cm}})\cdot\tilde{\boldsymbol{r}}_{i,k}\,|\tilde{\boldsymbol{r}}_{i,k}|^{-1} = 0$, where $\tilde{\boldsymbol{r}}_{i,k}$ is the vector connecting the k^{th} interaction site to the molecule's center of mass. The rigidity constraint ensures that each molecule translates as a rigid body and that the velocity components $\tilde{\boldsymbol{v}}_{i,k} = (\boldsymbol{v}_{i,k} - \boldsymbol{v}_i^{\text{cm}})$ give rise to purely rotational motions about the center of mass.

II. LIMMER AND CHANDLER'S VELOCITY INITIALIZATION ROUTINE

The sampling errors produced by LC's HMC algorithm occur because the initial space-frame site velocities passed to LAMMPS violate the rigidity constraint. Instead of drawing random velocities from the MB distribution, LC's code uses saved space-frame site velocities

$\boldsymbol{v}^{\mathrm{f}}_{i,k}$ from the end of the last accepted MD trial trajectory to generate $\boldsymbol{v}^{0}_{i,k}$. Integration with MD ensures that $\boldsymbol{v}^{\mathrm{f}}_{i,k}$ satisfy the rigidity constraint. New velocities $\boldsymbol{v}'_{i,k}$ are generated from $\boldsymbol{v}^{\mathrm{f}}_{i,k}$, however, by performing n_{swaps} velocity swaps, in which space-frame site velocities $\boldsymbol{v}_{i,k}$ are exchanged between randomly selected pairs of molecules. This procedure is inappropriate for rigid molecules because the swapped pairs in general have different space-frame orientations (*i.e.*, different $\tilde{\boldsymbol{r}}_{i,k}$). Hence $\boldsymbol{v}'_{i,k}$ will violate the rigidity constraint. An informal proof of this fact is provided in Section V.

The transformation $(\boldsymbol{v}^{\mathrm{cm}\prime}_i, \hat{\boldsymbol{\omega}}'_i) = \mathcal{T}^{-1}(\boldsymbol{v}'_{i,k})$ applied during initialization of LAMMPS' MD integrator will remove the nonzero velocity components $\tilde{\boldsymbol{v}}^{\parallel\prime}_{i,k}$. To compensate for the associated decrease in K, LC's code invokes internal LAMMPS routines to generate new velocities $\boldsymbol{v}''_{i,k} = \mathcal{T}(\mathcal{T}^{-1}(\boldsymbol{v}'_{i,k}))$. Velocities $\boldsymbol{v}^{0}_{i,k}$ are then obtained using $\boldsymbol{v}^{0}_{i,k} = \alpha \boldsymbol{v}''_{i,k}$, where α is a random scaling factor that ensures K fluctuations obey canonical statistics [8].

Initial violation of the rigidity constraint by LC's velocity swapping scheme removes rotational kinetic energy from the system. Kinetic energy is subsequently added back by rescaling $\boldsymbol{v}_{i,k}$, but this does not equipartition K. Consequently, the initial velocities $(\boldsymbol{v}^{\mathrm{cm},0}_i, \hat{\boldsymbol{\omega}}^{0}_i)$ calculated from $\boldsymbol{v}^{0}_{i,k}$ by LAMMPS' initialization routines do not obey MB statistics at the target sampling temperature (Fig. 1). Although the distributions of $\boldsymbol{v}^{\mathrm{cm},0}_i$ and $\hat{\boldsymbol{\omega}}^{0}_i$ extracted from LC's code at 300 K, for example, are roughly Gaussian, their variances are consistent with MB distributions at ca. 597 K and 3 K, respectively. This violation of equipartition is repeated at the beginning of each MD trial move and leads to sampling errors (Fig. 1).

III. CALCULATING SPACE-FRAME SITE VELOCITIES

Space-frame site velocities $\boldsymbol{v}_{i,k}$ for use with LAMMPS' MD integrator are obtained via the transformation $\boldsymbol{v}_{i,k} = \mathcal{T}(\boldsymbol{v}^{\mathrm{cm}}_i, \hat{\boldsymbol{\omega}}_i)$. In practice, this transformation is performed numerically for each molecule $i \in N$ using the following procedure:

(i) Calculate the residual positions vector $\tilde{\boldsymbol{r}}_{i,k} = \boldsymbol{r}_{i,k} - \boldsymbol{r}^{\mathrm{cm}}_i$ for each of the molecule's M sites. Vectors $\boldsymbol{r}_{i,k}$ and $\boldsymbol{r}^{\mathrm{cm}}_i$ are the space-frame Cartesian coordinates of the k^{th} site and the molecule's center of mass, respectively.

(ii) Compute the inertia tensor $I_{i,\{l,m\}} = \sum_{k=1}^{M} m_{i,k}[|\tilde{\boldsymbol{r}}_{i,k}|^2 \delta_{l,m} - \tilde{r}_{i,k,l}\tilde{r}_{i,k,m}]$, where $m_{i,k}$ is the mass of the k^{th} site. Indices l and m run over the three components of the Cartesian position vector $\tilde{\boldsymbol{r}}_{i,k}$.

(iii) Diagonalize the inertia tensor $\boldsymbol{I}_i = \boldsymbol{Q}_i^T \hat{\boldsymbol{I}}_i \boldsymbol{Q}_i$. The components of the resulting diagonal matrix $\hat{\boldsymbol{I}}_i$ are the principal moments of inertia, and the columns in the orthonormal rotation matrix $\boldsymbol{Q}_i$ are the principal axes.

(iv) Rotate $\hat{\boldsymbol{\omega}}_i$ into the space frame: $\boldsymbol{\omega}_i = \boldsymbol{Q}_i^T \hat{\boldsymbol{\omega}}_i$

(v) Calculate $\boldsymbol{v}_{i,k} = \boldsymbol{v}^{\mathrm{cm}}_i + \boldsymbol{\omega}_i \times \tilde{\boldsymbol{r}}_{i,k}$

IV. CONVERTING SITE VELOCITIES TO ANGULAR VELOCITIES

The inverse transformation $(\boldsymbol{v}^{\mathrm{cm}}_i, \hat{\boldsymbol{\omega}}_i) = \mathcal{T}^{-1}(\boldsymbol{v}_{i,k})$ is applied by computing space-frame angular velocities $\boldsymbol{\omega}_i$ and then rotating $\boldsymbol{\omega}_i$ into the body frame. Velocities $\boldsymbol{\omega}_i$ can be obtained by solving $\tilde{\boldsymbol{v}}_{i,k} = \boldsymbol{\omega}_i \times \tilde{\boldsymbol{r}}_{i,k}$ for all $k \in M$, where $\tilde{\boldsymbol{r}}_{i,k} = \boldsymbol{r}_{i,k} - \boldsymbol{r}^{\mathrm{cm}}_i$ and $\tilde{\boldsymbol{v}}_{i,k} = \boldsymbol{v}_{i,k} - \boldsymbol{v}^{\mathrm{cm}}_i$. This systems of equations is overdetermined because $\boldsymbol{\omega}_i$ can be uniquely evaluated using only two site velocities. Due to the existence of floating point errors in the data obtained from molecular simulation, however, it is often desirable to compute $\boldsymbol{\omega}_i$ from all of the site velocities using a least squares approach:

$$\mathcal{L}(\boldsymbol{\omega}_i) = \sum_{k=1}^{M} \left|\tilde{\boldsymbol{v}}_{i,k} - \boldsymbol{\omega}_i \times \tilde{\boldsymbol{r}}_{i,k}\right|^2 . \tag{8}$$

Setting $\partial\mathcal{L}/\partial\omega_{i,l} = 0$ for each of the $l = 1, 2, 3$ vector components results in the linear system:

$$\left[\sum_{k=1}^{M} (\tilde{\boldsymbol{r}}^T_{i,k}\tilde{\boldsymbol{r}}_{i,k}\boldsymbol{I}_3 - \tilde{\boldsymbol{r}}_{i,k}\tilde{\boldsymbol{r}}^T_{i,k})\right] \boldsymbol{\omega}_i = \sum_{k=1}^{M} \tilde{\boldsymbol{r}}_{i,k} \times \tilde{\boldsymbol{v}}_{i,k} \tag{9a}$$

$$\boldsymbol{S}_i \boldsymbol{\omega}_i = \boldsymbol{u}_i \tag{9b}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix. Hence space-frame angular velocities may be computed by solving the linear system, $\boldsymbol{\omega}_i = \boldsymbol{S}_i^{-1}\boldsymbol{u}_i$. Transformation to the body frame is achieved via $\hat{\boldsymbol{\omega}}_i = \boldsymbol{Q}_i\boldsymbol{\omega}_i$, where $\boldsymbol{Q}_i$ is the orthonormal rotation matrix obtained by diagonalizing the molecule's inertia tensor (see Section III).

V. SWAPPING SPACE-FRAME SITE VELOCITIES VIOLATES THE RIGIDITY CONSTRAINT

Let A and B be identical molecules with residual site positions $\tilde{\boldsymbol{r}}_{A,k} = \boldsymbol{r}_{A,k} - \boldsymbol{r}^{\mathrm{cm}}_B$ and $\tilde{\boldsymbol{r}}_{B,k} = \boldsymbol{r}_{B,k} - \boldsymbol{r}^{\mathrm{cm}}_B$, respectively. Suppose that molecule B has been created by rotating molecule A using the orthogonal matrix $\boldsymbol{R}$, *i.e.*, $\tilde{\boldsymbol{r}}_{B,k} = \boldsymbol{R}\tilde{\boldsymbol{r}}_{A,k}$ for all $k \in M$. Let $\boldsymbol{v}_{A,k}$ be site velocities for molecule A, that satisfy the rigidity constraint in the sense that there exists a $\boldsymbol{\omega}_A \in \mathbb{R}^3$ such that $\tilde{\boldsymbol{v}}_{A,k} = \boldsymbol{\omega}_A \times \tilde{\boldsymbol{r}}_{A,k}$ or all $k \in M$. Now suppose we assign the velocities $\boldsymbol{v}_{A,k}$ to the sites of molecule B by setting $\boldsymbol{v}_{B,k} = \boldsymbol{v}_{A,k}$. Obviously, $\boldsymbol{v}^{cm}_B = \boldsymbol{v}^{cm}_A$ and $\tilde{\boldsymbol{v}}_{B,k} = \tilde{\boldsymbol{v}}_{A,k}$ after the assignment is made. The new site velocities of molecule B $\boldsymbol{v}_{B,k}$, however, will not necessarily satisfy

the rigidity constraints, as $\tilde{\boldsymbol{v}}_{i,k} \perp \tilde{\boldsymbol{r}}_{i,k}$ does not necessarily imply that $\tilde{\boldsymbol{v}}_{i,k} \perp \boldsymbol{R}\tilde{\boldsymbol{r}}_{i,k}$. In fact, for a random $\boldsymbol{R}$ uniformly chosen from $O(3)$, the rigidity constraint will be violated with probability one. Hence exchanging space-frame site velocities between molecules with different space-frame orientations will generally lead to violations of the rigidity constraint.

Because Eq. 9 is a nonsingular linear system, it will always yield a solution. The residuals $\mathcal{L}(\boldsymbol{\omega}_i)$ will only vanish, however, if $\tilde{\boldsymbol{v}}_{i,k} \perp \tilde{\boldsymbol{r}}_{i,k}$. For $\tilde{\boldsymbol{v}}_{B,k}$ that do not satisfy $\tilde{\boldsymbol{v}}_{B,k} \perp \tilde{\boldsymbol{r}}_{B,k}$, the solution $\boldsymbol{\omega}_B = \boldsymbol{S}_B^{-1}\boldsymbol{u}_B$ will be invalid and yield $\mathcal{L}(\boldsymbol{\omega}_B) \neq 0$. It follows that $|\boldsymbol{\omega}_B| < |\boldsymbol{\omega}_A|$. Exchanging site velocities between molecules will therefore not preserve the MB distribution of $\boldsymbol{\omega}_i$ because it does not preserve the magnitudes of the angular velocities.

VI. CONSISTENCY CHECKS

The sampling errors generated by LC's HMC code can be unambiguously detected using a number of rigorous, quantitative consistency checks. To perform these checks, we used LC's HMC code to simulate a system of 512 ST2a water molecules (ST2 with conducting Ewald boundary conditions) at a fixed density of 0.94 g cm^{-3}. Interactions between molecules were truncated at 0.90 nm in the simulations, and the Ewald tolerance was set to 1×10^{-5}. A similar protocol was employed in our equation of state calculations for ST2b (Fig. 1) and free energy calculations for ST2a (Fig. 2). A 0.75 nm cutoff was used in those simulations, however, because they were performed for smaller systems containing 216 molecules. Time series data from the simulations were used to estimate the statistical inefficiency $g \equiv 1+2\tau_U$ for sampling the potential energy, where $\tau_U \equiv \int_0^\infty C_U(t)dt$ is the mean correlation time computed from the normalized auto-correlation function $C_U(t) = (\langle U(t)U(0)\rangle - \langle U\rangle^2)/(\langle U^2\rangle - \langle U\rangle^2)$. The values of g were used to estimate the number of independent samples generated from each simulation for the purpose of calculating statistical uncertainties. Following equilibration for $\sim 10^3 g$, statistics for the analysis were collected over a production period $> 5\times10^3 g$ in duration.

A. HMC normalization condition

Standard HMC schemes use a time-reversible and volume-preserving integrator to propagate trial MD trajectories. Invoking the volume-preserving property, it is straightforward to show that an HMC sampling algorithm should obey the normalization condition [10–12]:

$$\langle e^{-\beta\Delta H}\rangle = 1, \tag{10}$$

where $\Delta H = \Delta U + \Delta K$. As pointed out by an anonymous referee, this condition serves as a built-in consistency check for HMC schemes that is useful for detecting systematic sampling errors.

The rigid body integrator in LAMMPS used by LC's HMC code (*fix nve/rigid molecule*) is reversible and volume preserving [7]. Hence, in the absence of implementation errors, LC's HMC scheme should satisfy the normalization condition, independent of the choice of sampling parameters such as the MD integration time step δt, the number of MD steps n_{steps}, or the number of velocity swaps n_{swaps} . Contrary to this expectation, however, the average in Eq. 10 becomes significantly larger as δt and n_{swaps} increase (Fig. S1A). For all parameter choices, the violations are statistically significant and range from 10^1 to $10^2\sigma$, where σ is the estimated uncertainty (Fig. S1B). The spurious sensitivity to parameter choice and statistically significant violations of the normalization condition, a rigorous statistical mechanical relationship, demonstrate that LC's original HMC code produces systematic sampling errors.

For comparison, we have also performed a similar analysis using the version of LC's HMC code that we corrected to draw initial velocities from the MB distribution at the target sampling temperature. As expected, within statistical uncertainty, the corrected sampling scheme satisfies the normalization condition independent of parameter choice (Fig. S1). Hence this simple yet rigorous consistency check illustrates that significant sampling errors are introduced by LC's inappropriate velocity initialization algorithm. This observation is consistent with the conclusions drawn from the equation of state (Fig. 1) and free energy (Fig. 2) calculations performed using $\delta t = 3-4$ fs and $n_{\text{steps}} = 10-15$.

Equation 10 implies that the average HMC acceptance probability will follow $\langle P_{\text{acc}}\rangle = \text{erfc}\left(\frac{1}{2}\sqrt{\beta\langle\Delta H\rangle}\right)$ [10–12] as $\delta t \to 0$ and $N \to \infty$. Although these limits are never reached in simulation, this expression for $\langle P_{\text{acc}}\rangle$ has been shown to be remarkably accurate in practice [11]. The values of $\langle P_{\text{acc}}\rangle$ computed with LC's HMC code, however, deviate significantly from this expected relation (Fig. S2). Further, by Jensen's inequality, Eq. 10 requires $\langle\Delta H\rangle \geq 0$ [10], which is clearly violated by LC's HMC sampling algorithm when $\langle P_{\text{acc}}\rangle \gtrsim 0.6$. Conversely, both of these expectations are satisfied when sampling is performed with the version of LC's code that we corrected (Fig. S2).

B. Ensemble consistency

LC's HMC algorithm fails consistency checks based on equation of state comparisons (Fig. 1) and the normalization condition in Eq. 10 (Fig. S1) because of sampling errors introduced by their inappropriate velocity initialization scheme. In the absence of algorithm-specific consistency checks or a properly functioning reference code, sampling errors can be detected through fluctuation analysis [13]. If canonical sampling is performed for a system at two different temperatures $\beta_1 = (k_B T_1)^{-1}$ and $\beta_2 = (k_B T_2)^{-1}$, for example, the log ratio of the poten-

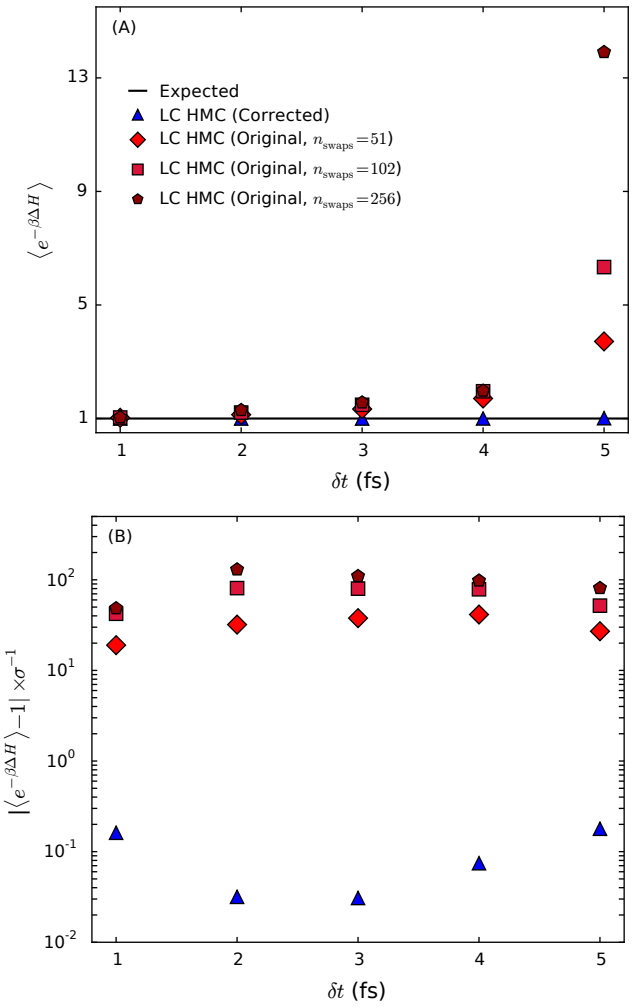

FIG. S1. Tests of the HMC normalization condition (Eq. 10) performed with canonical ensemble simulations of 512 ST2a molecules [5] (ST2 with conducting Ewald boundary conditions) at 310 K and a fixed density of 0.94 g cm^{-3}. The expected value of 1 should be obtained independent of the chosen HMC sampling parameters. (A) The values of $\langle e^{-\beta\Delta H}\rangle$ computed with LC's HMC code deviate from 1 and become larger as δt and n_{swaps} increase. (B) For all parameter choices, the violations are statistically significant and range from 10^1 to $10^2\sigma$, where σ is the estimated uncertainty. By contrast, when sampling is performed with the version of LC's HMC code that we corrected as described in Section I, the normalization condition is satisfied within statistical uncertainty (*i.e.*, deviations are $\ll \sigma$). Values of $n_{\text{steps}} = 40$, 30, 15, 10, and 20 were used for the simulations performed with $\delta t = 1$, 2, 3, 4 and 5 fs, respectively. For comparison, the velocity auto-correlation time for ST2 is $\sim$ 200 fs under these conditions. Symbol sizes in (A) are larger than the estimated statistical uncertainties.

tial energy distributions should vary linearly with U:

$$\ln\left[\frac{P(U|\beta_2)}{P(U|\beta_1)}\right] = -(\beta_2 - \beta_1)U + C, \qquad (11)$$

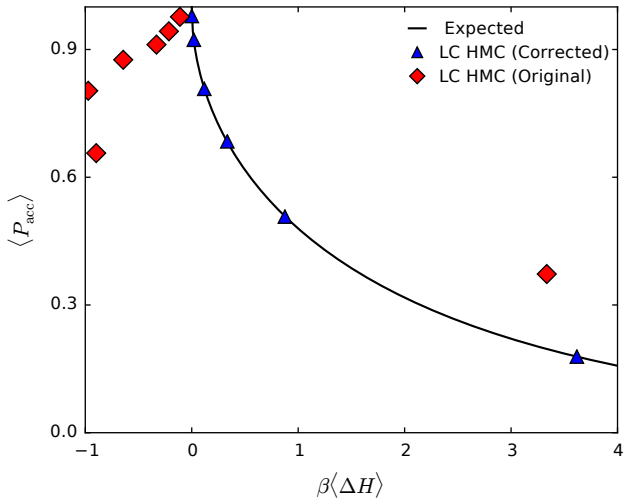

FIG. S2. Variation of the average HMC acceptance probability $\langle P_{\text{acc}}\rangle$ with $\langle\Delta H\rangle$. Simulations performed with LC's original code significantly deviate from the expected relation $\langle P_{\text{acc}}\rangle = \text{erfc}\left(\frac{1}{2}\sqrt{\beta\langle\Delta H\rangle}\right)$ and do not satisfy the constraint $\langle\Delta H\rangle \geq 0$ implied by Eq. 10. By contrast, simulations performed with the corrected version of LC's HMC code satisfy both expectations. The simulations were performed at 310 K using a system of 512 ST2a molecules [5] (ST2 with conducting Ewald boundary conditions) at a fixed density of 0.94 g cm^{-3}. Each data point represents an independent simulation with parameters $\delta t = 1-10$ fs and $n_{\text{steps}} = 10-30$ chosen to achieve the reported average acceptance rate. Statistical uncertainties are smaller than the symbol sizes.

where C is an additive constant that depends on β_1 and β_2. This relationship provides a straightforward route to test that sampled potential energy fluctuations obey canonical statistics. To implement this consistency check, sampling is performed at two nearby temperatures, and the log ratio of the the resulting potential energy distributions is fit to a straight line. According to Eq. 11, the slope of the fit line should be equal to $-(\beta_2 - \beta_1)$ within statistical uncertainty if canonical fluctuations are correctly sampled.

Free energy calculations are notoriously sensitive to sampling errors. Hence *qualitative* tests, such as visual inspection of histogram data to check for consistency with Eq. 11, may be insufficient to detect systematic sampling errors that can significantly influence these types of calculations. To address this issue, Shirts [13] has developed the Python software package *checkensemble* [14] for performing *quantitive* consistency tests based on Eq. 11 and analogous relationships that can be used to validate schemes designed to sample the isothermal-isobaric and grand canonical ensembles. The *checkensemble* package provides routines for rigorously estimating statistical uncertainties in fit parameters based on analytical error expressions and bootstrapping methods [13]. Additionally, *checkensemble* implements maximum likelihood methods for parameter estimation. Rather

than fitting to histograms, maximum likelihood methods make use of time series data directly and thus avoid potential issues related to improper choice of bin width [13].

We have performed quantitative tests using the *checkensemble* package to scrutinize LC's original code and determine whether the potential energy fluctuations sampled by their HMC scheme are consistent with canonical statistics (Fig. S3; Table S1). Simulations with LC's HMC code were performed at 310 and 320 K using the system and sampling protocols described at the beginning of Section VI. Following the recommendations outlined by Shirts [13], the temperature gap $\Delta T^{\text{true}} = 10$ K was chosen such that the peaks of the potential energy distributions were separated by approximately twice the standard deviation at 310 K. The results from this rigorous analysis show that the potential energy fluctuations generated by LC's original HMC code are not quantitatively consistent with canonical statistics at the specified sampling temperatures (Fig. S3A). The effective temperature gap $\Delta T^{\text{obs.}} \pm \sigma = 8.721 \pm 0.067$ extracted from the slope of the fit line to the histogram data deviates significantly from the expected value of 10 K. The absolute magnitude of the deviation is $|\Delta T^{\text{true}} - \Delta T^{\text{obs.}}| = 1.279$ K or equivalently 19.08 σ, which indicates that significant sampling errors are likely present. By contrast, an effective temperature gap of $\Delta T^{\text{obs.}} \pm \sigma = 10.103 \pm 0.147$ is obtained when sampling is performed using the version of LC's HMC code that we corrected to draw initial velocities from the MB distribution at the target sampling temperature (Fig. S3B). This value corresponds to a deviation of only 0.70 σ from the true temperature gap, demonstrating that the corrected code samples the expected equilibrium distribution.

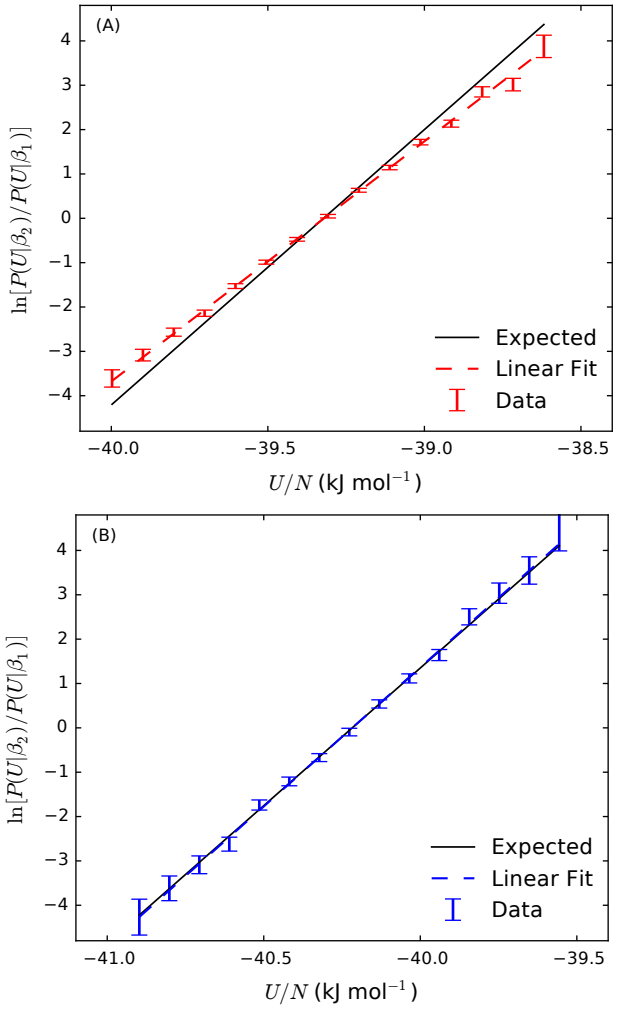

FIG. S3. Ensemble consistency tests performed with 512 ST2a molecules [5] (ST2 with conducting Ewald boundary conditions) at a fixed density of 0.94 g cm^{-3}. Histogram data from HMC simulations at 310 K and 320 K using $\delta t = 3$ fs and $n_{\text{steps}} = 15$ were analyzed with the *checkensemble* [14] package to determine if the sampled fluctuations are consistent with canonical statistics (Eq. 11). (A) Simulations performed with LC's original HMC code using $n_{\text{swaps}} = 51$ exhibit statistically significant deviations (19.08σ) from Eq. 11 (black line), indicating that the sampled fluctuations are inconsistent with the expected canonical distribution. (B) Data from the version of LC's HMC code with the corrected velocity initialization routine are statistically indistinguishable (deviation of 0.70σ) from Eq. 11 (black line).

TABLE S1. Ensemble consistency tests for ST2a performed on data from simulations at 310 and 320 K ($\Delta T^{\text{true}} = 10$ K) using maximum likelihood estimation. Deviations larger than 2-3σ indicate that systematic sampling errors are likely present.

Code	δt (fs)	n_{steps}	n_{swaps}	$\Delta T^{\text{obs.}} \pm \sigma$ (K)[a]	$\lvert\sigma$ Dev.$\rvert$[b]
HMC-c[c]	3	15	–	10.127 ± 0.140	0.907
HMC-o[d]	1	10	51	6.703 ± 0.117	28.2
	1	20	51	7.567 ± 0.107	22.7
	1	40	51	8.874 ± 0.106	10.6
	2	30	51	9.060 ± 0.081	11.6
	3	15	51	8.757 ± 0.066	18.8
	7	3	51	8.423 ± 0.107	14.7
	1	40	102	7.982 ± 0.132	15.3
	2	30	102	8.218 ± 0.076	23.4
	3	15	102	8.349 ± 0.142	11.6
	7	3	102	6.569 ± 0.123	27.9
	1	40	256	7.292 ± 0.149	18.2
	2	30	256	7.764 ± 0.106	21.1
	3	15	256	7.616 ± 0.102	23.4

[a] Maximum likelihood estimate ± analytical uncertainty
[b] Absolute deviation from ΔT^{true} in units of σ
[c] Version of LC's code that we corrected to draw initial velocities from the Maxwell-Boltzmann distribution at the target sampling temperature
[d] LC's original code that generates initial velocities using the inappropriate swapping procedure described in Section II

The quantitative consistency checks reported above use histogram data to perform tests based on Eq. 11 (Fig. S3). Although these tests are convenient for vi-

sualizing the data and linear fits, the results may be sensitive to the choice of histogram bin width. Hence we have also performed a similar analysis using the maximum likelihood estimation approach in *checkensemble*. The results again show that the simulations performed with LC's original HMC code exhibit potential energy fluctuations that are quantitatively inconsistent with the expected equilibrium distribution (Table S1). Signs of significant sampling errors are present independent of the choice of HMC parameters n_{steps}, δt, or n_{swaps}. For comparison, the same consistency check was also performed on data collected from the corrected version of LC's HMC code. The potential energy fluctuations sampled by the corrected HMC algorithm are found to be consistent with canonical statistics (Table S1).

VII. SUMMARY

We have scrutinized the hybrid Monte Carlo (HMC) [2] code used by Limmer and Chandler (LC) to perform free energy calculations for the ST2 model water [4, 5]. We have shown mathematically that LC's velocity initialization scheme based on swapping space-frame site velocities between pairs of water molecules violates the bond constraints of rigid polyatomic models (Sections II and V). Hence the initial velocities generated by LC's code at the beginning of each HMC trial move violate equipartition and do not obey Maxwell-Boltzmann statistics at the target sampling temperature (Fig. 1). Additionally, our extensive numerical tests reveal:

1. LC's HMC code does not predict the correct equation of state for ST2 (Fig. 1); it produces average potential energies that are significantly larger (more positive) than those generated by our publicly available HMC code [15] and the MD code provided to us by Ni and Skinner [16].

2. LC's HMC code generates a free energy surface for ST2 (Fig. 2), similar to that reported in the LC studies [4, 5], not exhibiting two liquid basins under deeply supercooled conditions, in disagreement with calculations for ST2 performed by a number of independent groups [17–23].

3. LC's HMC code does not satisfy rigorous quantitative consistency checks [10–12] based on HMC's normalization condition (Eq. 10; Fig. S1); the violations of the normalization condition show that LC's HMC scheme produces systematic sampling errors.

4. LC's HMC code does not satisfy rigorous quantitative tests [13] for ensemble consistency (Eq. 11; Fig. S3; Table S1); the tests demonstrate that the fluctuations sampled by LC's code are inconsistent with canonical statistics at the target sampling temperature.
5. The version of LC's HMC that we corrected to draw initial velocities from the Maxwell-Boltzmann distribution at the target sampling temperature (a) correctly predicts ST2's equation of state (Fig. 1); (b) predicts free energy surfaces with two liquid basins under deeply supercooled conditions (Fig. 2), in accord with other studies [17–23]; (c) satisfies HMC's rigorous normalization condition (Fig. S1); and (d) samples fluctuations consistent with canonical statistics (Fig. S3; Table S1).

Our numerical tests therefore demonstrate that the simulation code made available to us by LC implements a HMC sampling protocol that is inappropriate for rigid polyatomic water models. The resulting sampling errors significantly distort ST2's equation of state and hence have prevented LC from accurately characterizing ST2's phase behavior in Refs. [4] and [5].

[1] M. Allen and D. Quigley, Mol. Phys. **111**, 3442-3447 (2013).
[2] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216-222 (1987).
[3] J. C. Palmer, A. Haji-Akbari, R. S. Singh, F. Martelli, R. Car, A. Z. Panagiotopoulos and P. G. Debenedetti, arXiv: https://arxiv.org/abs/1712.08278 (2017).
[4] D. T. Limmer and D. Chandler, J. Chem. Phys. **135**, 134503 (2011).
[5] D. T. Limmer and D. Chandler, J. Chem. Phys. **138**, 214504 (2013).
[6] S. Plimpton, J. Comp. Phys. **117**, 1-19 (1995).
[7] T. F. Miller III, M. Eleftheriou, P. Pattnaik, A. Ndirango, and D. Newns, J. Chem. Phys **116**, 8649 (2002).
[8] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys **126**, 014101 (2007).
[9] H. Kamberaj, R. J. Low, and M. P. Neal, J. Chem. Phys **122**, 224114 (2005).
[10] M. Creutz, Phys. Rev. D **38**, 1228-1238 (1988).
[11] B. Mehlig, D. W. Heermann, and B. M. Forest, Phys. Rev. B **45**, 679-685 (1992).
[12] V. Weber, S. Merchant, P. Dixit, and D. Asthagiri, J. Chem. Phys **132**, 204509 (2010).
[13] M. R. Shirts, J. Chem. Theory Comput. **9**, 909-926 (2013).
[14] The *checkensemble* package is available on GitHub: https://github.com/shirtsgroup/checkensemble.
[15] Our code has been publicly available since Sept. 2014 (http://pablonet.princeton.edu/pgd/html/links.html)
[16] Y. Ni and J. L. Skinner, J. Chem. Phys. **145**, 124509 (2016).
[17] Y. Liu, A. Z. Panagiotopoulos and P. G. Debenedetti, J. Chem. Phys. **131**, 104508 (2009).
[18] F. Sciortino, I. Saika-Voivod and P. H. Poole Phys. Chem. Chem. Phys. **13**, 19759-19764 (2011).
[19] Y. Liu, J. C. Palmer, A. Z. Panagiotopoulos and P. G.

Debenedetti, J. Chem. Phys. **137**, 214505 (2012).
[20] P. H. Poole, R.K. Bowles, I. Saika-Voivod and F. Sciortino, J. Chem. Phys. **138**, 034505 (2013).
[21] T. A. Kesselring, E. Lascaris, G. Franzese, S. V. Buldyrev, H. J. Herrmann, and H. E. Stanley, J. Chem. Phys. **138**, 244506 (2013).
[22] J. C. Palmer, F. Martelli, Y. Liu, R. Car, A. Z. Panagiotopoulos and P. G. Debenedetti, Nature **510**, 385-388 (2014).
[23] F. Smallenburg and F. Sciortino, Phys. Rev. Lett. **115**, 015701 (2015).
[24] A. M. Horowitz, Phys. Lett. B **268**, 247-252 (1991).
[25] J. A. Wagoner and V. S. Pande, J. Chem. Phys. **137**, 214105 (2012).