JCTC Journal of Chemical Theory and Computation

Article

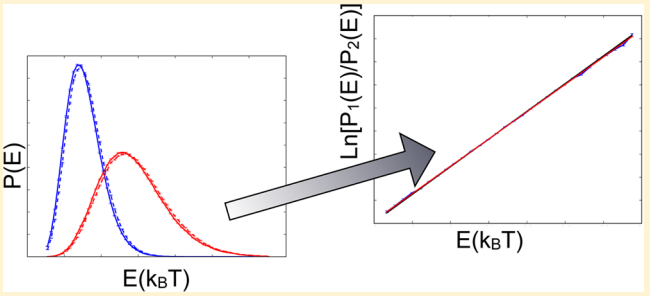
pubs.acs.org/JCTC

Simple Quantitative Tests to Validate Sampling from Thermodynamic Ensembles

Michael R. Shirts*

Department of Chemical Engineering, University of Virginia, Charlottesville, Virginia 22904, United States

ABSTRACT: It is often difficult to quantitatively determine if a new molecular simulation algorithm or software properly implements sampling of the desired thermodynamic ensemble. We present some simple statistical analysis procedures to allow sensitive determination of whether the desired thermodynamic ensemble is properly sampled. These procedures use paired simulations to cancel out system dependent densities of state and directly test the extent to which the Boltzmann distribution associated with the ensemble (usually canonical, isobaric−isothermal, or grand canonical) is satisfied. We demonstrate the utility of these tests for model systems and for molecular dynamics simulations in a range of situations and describe an implementation of the tests designed for end users.



1. INTRODUCTION

Molecular simulations, including both molecular dynamics (MD) and Monte Carlo (MC) techniques, are powerful tools used to study the properties of complex molecular systems. When used to specifically study thermodynamics of such systems, rather than dynamics, the primary goal of molecular simulation is to generate uncorrelated samples from the appropriate ensemble as efficiently as possible. These simulation data can then be used to compute thermodynamic properties of interest. Simulations of several different ensembles may be required to compute some thermodynamic properties, such as free energy differences between states. An ever-expanding number of techniques have been proposed to perform more and more sophisticated sampling from complex molecular systems using both MD and MC, and new software tools are continually being introduced in order to implement these algorithms and to take advantage of advances in hardware architecture and programming languages.

However, it is extremely easy to make subtle errors in both the theoretical development and the computer implementation of these advanced sampling algorithms. Such errors can occur because of numerical errors in the underlying energy functions, theoretical errors in the proposed algorithm, approximations that are too extreme, and the programming bugs that are inevitable when managing more and more complicated code bases.

There are a number of reasons it is difficult to validate a given implementation of an algorithm for the proper thermodynamic behavior. First, we lack analytical results for virtually all complex molecular systems, and analytically soluble toy problems may not have all of the features that more complicated systems of actual research interest may possess. Additionally, molecular simulations generate statistical samples from the probability distribution of the system. Most observables therefore require significant simulation time to reduce statistical noise to a level sufficiently low to allow conclusive identification of small but potentially significant violations of the sampled ensembles.

There are of course some aspects of molecular distributions that can and should always be checked directly. For example, in an NVE ensemble the total energy should be conserved with statistically zero drift. For symplectic integrators with NVE simulations, the RMS error will scale with the square of the step size. For an NVT ensemble when the potential energy is independent of particle momenta (which is true with the rare exception of systems with magnetic forces), then the kinetic energy will follow the Maxwell−Boltzmann distribution, and consistency with this distribution can be estimated using standard statistical methods. NVT simulations must also have an average kinetic energy corresponding to the desired temperature, and NPT simulations must have the proper average instantaneous pressure computed from the virial and kinetic energy. We would encourage all simulation developers to include these automated tests in their code.

However, there are no standard tests for proper distribution for the potential energy, which greatly complicates Monte Carlo simulations, or for total energy of an arbitrary simulation system. Additionally, there are many possible distributions which have the correct average temperature or pressure but do not satisfy the proper Boltzmann probability distributions for our specific ensemble of interest.

It is therefore worthwhile to have physically rigorous strategies and tools for assessing whether a simulation method is indeed generating samples from the desired distribution in its entirety. Such general strategies could help to better answer vital questions such as, “is this thermostat/barostat correct?”, “how much does a very long time step affect my energy

Received: August 3, 2012
Published: December 19, 2012

   dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

distribution?", and of course "have I *finally* got that bug out of my code now?"

2. THEORY

Thermodynamic ensembles all have similar probability distributions with respect to macroscopic intensive parameters and microstates, e.g.:

$$P(\vec{x}|\beta) \propto \exp(-\beta H(\vec{p}, \vec{q})) \quad \text{canonical} \tag{1}$$

$$P(\vec{x}, V|\beta, P) \propto \exp(-\beta H(\vec{p}, \vec{q}) - \beta PV)$$

isobaric-isothermal (2)

$$P(\vec{x}, \vec{N}|\beta, \vec{\mu}) \propto \exp(-\beta H(\vec{p}, \vec{q}) + \sum_{\text{species}} \beta \mu_i N_i)$$

grand canonical (3)

where $P(a|b)$ indicates the probability of a microstate determined by variable or variables a given a macroscopic parameter or parameters b. Specifically, all have the exponential form $\exp(-u(\vec{x}))$ where $\vec{x} = (\vec{p},\vec{q},V,\vec{N})$ is the microstate and $u(\vec{x})$ is a reduced energy term whose form depends on the ensemble.

This reduced energy term is a generalized function of two types of variables. The first type of variable consists of the degrees of freedom determining the microstates of each ensemble, including the positions and velocities of the atoms, but also potentially including the volume of the system V and the number of particles of each of i species in the system N_i. The second type of variable consists of those determining the ensemble of the physical system, including the temperature T, the pressure P, the chemical potentials μ_i, and the specific functional form of the Hamiltonian $H(\vec{p},\vec{q})$. These equations, along with the requirement that all microstates with the same value for the generalized energy term have the same probability, completely define the thermodynamic ensemble. A general test should therefore check as directly as possible that the samples we collect are fully consistent with eqs 1−3. For simplicity, we will perform an initial derivation of such a test using the canonical ensemble and then generalize the derivation to other ensembles.

The probability density of observing a specific energy in the canonical ensemble (eq 1) can be written in terms of the density of states $\Omega(E) = \exp(S(N,V,E)/k_B)$ as

$$P(E|\beta) = Q(\beta)^{-1}\Omega(E)\exp(-\beta E) \tag{4}$$

where S is the entropy, $\beta = (k_B T)^{-1}$, k_B is Boltzmann's constant, and $Q(\beta) = \int \Omega(E)\exp(-\beta E)\,dE$ is the canonical partition function, related to the Helmholtz free energy A by $A = -\beta^{-1}\ln Q$. Q is a function of β, but not E, whereas Ω is a function of E, but importantly, not β. Note that at this point, E is specifically the total energy, though we will examine kinetic and potential energies separately later on.

Without specific knowledge of what the density of states $\Omega(E)$ is for a particular molecular system, no quantity of samples from a single state can identify if the energies indeed have the proper distribution. However, if we take the ratio of the probability distributions of two simulations performed at different temperatures, hence with two different values of β, but with otherwise identical parameters, the unknown density of states cancels, leaving

$$\frac{P(E|\beta_2)}{P(E|\beta_1)} = \frac{\frac{\exp(-\beta_2 E)}{Q(\beta_2)}}{\frac{\exp(-\beta_1 E)}{Q(\beta_1)}} = \exp([\beta_2 A_2 - \beta_1 A_1] - [\beta_2 - \beta_1]E) \tag{5}$$

If we take the logarithm of this ratio, we obtain:

$$\ln\frac{P(E|\beta_2)}{P(E|\beta_1)} = [\beta_2 A_2 - \beta_1 A_1] - [\beta_2 - \beta_1]E \tag{6}$$

which is of the linear form $\alpha_0 + \alpha_1 E$. Note that linear coefficient $\alpha_1 = -(\beta_2 - \beta_1)$ is independent of the (unknown in general) Helmholtz free energies A_2 and A_1.

This relationship forms the basis of the ensemble validation techniques we present in this paper. Similar formulas can be derived for any of the standard thermodynamic ensembles with probability distributions of the form $e^{-u(\vec{x})}$ as long as the reduced energy term is linear in conjugate parameters. Nonexponential probability distributions are certainly possible to generate in simulations, but are much less standard, and so we will not deal directly with them in this study. The same general techniques will work if the probability of a given microstate depends only on the energy of the microstate. We will call agreement of a simulation with its target distribution as described by eq 6 and its analogs for other ensembles *ensemble consistency*.

These tests are particularly simple because they require only the energy, volume (for isobaric−isothermal simulations), and particle number (for grand canonical simulations) data for a pair of simulations to be saved, and because they are general for all molecular systems that have an associated Boltzmann probability; new varieties of the test need not be generated for each new system encountered.

There are a number of ways to check if the distribution of samples from a given pair of NVT simulations satisfies these equations. The most straightforward way starts with binning the energies E from both simulations. If the distributions are sufficiently close together to have statistically well-defined probabilities at overlapping values of E and we have sufficient data, we can fit the ratio of the histogram probabilities to a line in this overlap region. If the slope deviates from $-(\beta_2 - \beta_1)$ by a statistically significant amount, then the data necessarily deviate from a canonical distribution. However, deciding quantitatively what constitutes "statistically significant" can be challenging and will be further explored in this paper.

This test of consistency with eq 6 is a *necessary* test for an algorithm that is consistent with the canonical ensemble; if the slope of the probability ratio deviates from the true line, the data cannot be consistent with the ensemble. However, the test is not necessarily a *sufficient* test of simulation quality as it does not include any direct test of ergodicity. Specifically, it says nothing about whether states with the same energy are sampled with equal probability as is required by statistical mechanics. It also does not say anything about whether there are states that are not sampled. We could have sampling consistent with the desired ensemble but trapped in only a small portion of the allowed phase space of a system.

In general, additional tests of convergence or ergodicity are required before the system can be assumed to be sampled correctly. For example, for molecular dynamics, one could examine the kinetic energy of different partitions of the degrees of freedom as can be used to diagnose such problems as the

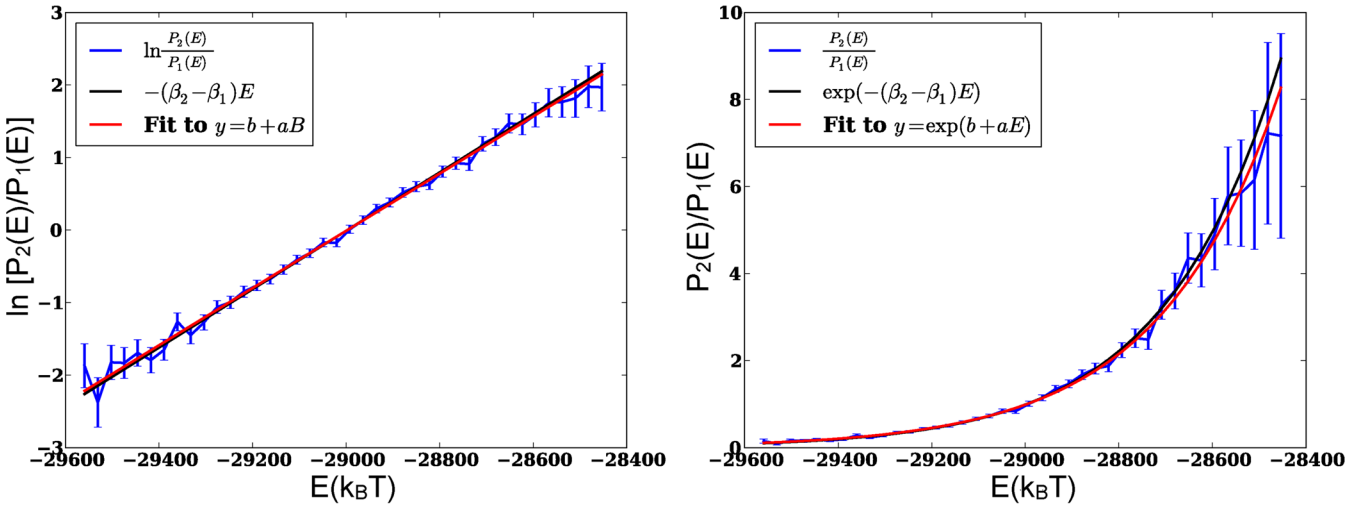


Figure 1. Ensemble validation of water simulations. Validation of the energy distribution of 900 TIP3P water molecules simulated in the NVT ensemble using the Nosé−Hoover algorithm.[2] The predicted value and the actual data for both linear (a) and nonlinear (b) fits quantitatively agree.

"flying ice cube," occurring in some poorly configured simulations when the center of mass degrees of freedom are decoupled from other degrees of freedom.[1] However, for testing algorithms or code, simple systems that are both sufficiently complicated and general can usually be found which will behave ergodically within a reasonable amount of simulation time. Therefore, in the rest of this paper, we will analyze systems which are clearly sampled ergodically and which have converged ensemble averages of interest, so we will not require any additional tests of ergodicity or convergence.

Having analyzed the potential problems with such ensemble validation analysis, we next explore possible methods to quantify deviation from the canonical ensemble using data collected from pairs of simulations.

2.1. Visual Inspection. We can divide the common energy range of the two simulations into bins (perhaps 20−40, depending on the amount of data, numbers chosen solely from experience through trial and error). Bins need not be equally spaced, though this simplifies the analysis considerably by removing the need to correct the probability densities for differing widths of bins. It also greatly simplifies the analysis to select bin divisions that are aligned between the two data sets. Bins can be chosen to exclude a few points on the top and the bottom of each distribution to avoid a small sample error and zero densities at the extremes. $P_1(E)$ and $P_2(E)$ in each bin can then be estimated directly from the histograms. We can compute the ratio of these histograms at each value of the energy at the centers of the bins and plot either the ratio or, more cleanly, the logarithm of this ratio, as shown in Figure 1. If this logarithm ratio is linear, we have a system that for all *qualitative* purposes obeys the proper equilibrium distribution.

Qualitatively, if the actual slope of the log energy ratios is below the expected slope, it means that the low β (high temperature) simulation samples that particular energy less than it should, while if it is above the true line, it means that portion of the distribution is oversampled. A consistently higher observed slope therefore means that the distribution is narrower than it should be, and a lower observed slope means that the distribution is wider than it should be.

2.2. Quantitative Fitting. The relationships presented so far are not entirely novel; visual inspection of probability ratios of paired temperature replica exchange simulations has been used previously to check that neighboring replicas have the proper distributions relative to each other.[3,4] However, there has not been an effort to use this relationship as a general test to quantitatively analyze simulations for goodness-of-fit to the putative ensemble distributions.

2.2.1. Linear Fitting. To make this ensemble test quantitative, we estimated the error in the occupancy probability p_k of each bin i as $\delta p_k = (p_k(1 - p_k)/n)^{1/2}$ (a standard result for the binomial distribution) and propagate the error in the individual bin probabilities into the ratio $P_2(E)/P_1(E)$ (a process detailed in Appendix B). If the true slope lies consistently outside of the error estimates, then it is very likely the simulation is not correctly sampling the desired ensemble. Calculation of the histogram uncertainties also allows us to perform weighted linear and nonlinear least-squares fitting (details also in Appendix B). This allows us to include the effect of small sample error at the extremes of the distribution in our fitting. We can use standard error propagation methods to propagate the error in the histogram occupancy ratios into the error in the linear parameters.

2.2.2. Nonlinear Fitting. It is well-known that linearizing a mathematical relationship in order to perform linear least-squares can introduce bias in the estimation of the parameters. It is therefore often preferable to minimize the direct sum of the residuals $S_r(\vec{\alpha}) = \sum_i (y_i - f(\vec{\alpha},x_i))^2$, which is a nonlinear function of $\vec{a}$, and then propagate the error in the histogram bins into the uncertainties of the components of $\vec{\alpha}$. In this particular problem, we want to determine the two parameter fit that minimizes the sum of residuals $S_r(\alpha_0, \alpha_1)$ for the function

$$S_r(\alpha_0, \alpha_1) = \sum_i \left[\frac{P_1(E_i)}{P_2(E_i)} - \exp(\alpha_0 + \alpha_1 E_i) \right]^2 \quad (7)$$

2.2.3. Maximum Likelihood Estimates. Any results from either the linear or nonlinear case may be affected by the choice of histogram binning we use. In theory, we can vary the number of histogram bins to ensure that the answers are not dependent

on the number of bins. However, we can completely eliminate the histogram dependence as well as include the data at the tails rather than truncate them by using a maximum likelihood approach. A maximum likelihood approach allows us to predict the most likely parameters for a given statistical model from the data that have been observed.

Previously, we used such a maximum likelihood approach to compute the free energy difference between forward and reverse work distributions between two thermodynamic states at the same temperature,[5] which is equivalent to computing the value of $\ln Q_1/Q_2$ with fixed β. In that case, we use information from two unnormalized distributions to find a single unknown, the ratio of normalizing constants. In the present case, however, we have *two* parameters in the distribution which we must fit, $\alpha_0 = \ln Q_1/Q_2$ and $\alpha_1 = -(\beta_2 - \beta_1)$. Applying a maximum likelihood approach along the lines described in the paper[5] leads to log likelihood equations:

$$\ln L(\alpha|\text{data}) = \sum_{i=1}^{N_1} \ln f(-\alpha_0 - \alpha_1 E_i) + \sum_{j=1}^{N_2} \ln f(\alpha_0 + \alpha_1 E_j)$$

where $f(x)$ is the Fermi function $f(x) = [1 + \exp(-x)]^{-1}$, and where the first sum is over energies sampled at temperature T_1 and the second sum is over the energies sampled at T_2. The most likely parameters are the ones which maximize the likelihood function in eq 8. This particular function can be shown to have no minima and only one maximum, so it will always converge. The change from one unknown to two unknowns does not affect the procedure to calculate the value but does affect the uncertainty. Details of these differences are discussed in Appendix C.

Equation 8 can be solved by any of the standard techniques for multidimensional optimization as it is everywhere concave. There is one minor technicality; clearly, the variance can be minimized to zero by setting $\alpha_0 = \alpha_1 = 0$, which is not physically consistent with the data. There is therefore an additional constraint we must first identify to find a unique minimum.

In performing this likelihood maximization, we note that although there are four parameters explicitly stated, A_1, A_2, β_1, and β_2, only two of them are actually free parameters. Examining eq 6, we can express the relationship to the physical quantities as $\alpha_0 = \beta_2 A_2 - \beta_1 A_1$ and $\alpha_1 = -(\beta_2 - \beta_1)$. We also note that eq 6 does not allow us to test for β_1 and β_2 directly but instead is only a function of the difference $\beta_2 - \beta_1$, so we must actually treat this as one variable corresponding to a single degree of freedom. A simple choice is to treat $\beta_1 + \beta_2$ as a constant in what amounts to a choice of the energy scale. We can therefore set $\beta_{\text{ave}} = 1/2(\beta_{1,\text{user}} + \beta_{2,\text{user}})$, the user specified temperatures. A_1 and A_2 are the free energies of the system, so there is no physical meaning to their absolute value, only their difference. Without a loss of generality, we set $A_1 + A_0 = 0$ and treat $\Delta A = A_2 - A_1$ as our second independent variable. These two choices allow us to then solve for unique values of α_0 and α_1, rewriting $\alpha_0 + \alpha_1 E = \beta_{\text{ave}}(A_2 - A_1) - (\beta_2 - \beta_1)E = \beta_{\text{ave}}\Delta A - \Delta\beta E$, an expression that explicitly only has two free parameters.

One down side of using a maximum likelihood analysis is that it does not give a graphical representation; it is histogram independent, and so we do not have a histogram that we can plot! A linear fit should therefore be performed in conjunction with maximum likelihood analysis to quickly visualize the data as a sanity check.

2.3. Error Estimates. Once we have an estimate of the slope $\beta_2 - \beta_1$, we must ask if the slope deviates from the true by a statistically significant amount or if the difference is more likely due to random chance. For this, we can turn to error estimation techniques to find a statistically robust approximation for the error in $\beta_2 - \beta_1$ and to determine if any deviations from the true value are most likely a result of statistical noise or actual errors in the simulation.

For weighted linear least-squares, weighted nonlinear least-squares, and multiparameter maximum likelihood logistic regression, the analytic asymptotic error estimators for the covariance matrix of fitting parameters are all well-known statistical results:

$$\text{linear} \quad \text{cov}(\vec{\alpha}) = (X^T W X)^{-1} \tag{8}$$

$$\text{nonlinear} \quad \text{cov}(\vec{\alpha}) = (J^T W^{-1} J)^{-1} \tag{9}$$

$$\text{maximum likelihood} \quad \text{cov}(\vec{\alpha}) = (H(\ln L)_\alpha)^{-1} \tag{10}$$

In all equations, $\vec{\alpha}$ is the vector of parameters we are estimating. In eq 8, X is the $(M + 1) \times N$ matrix with the first column all ones, and the second through the $(M + 1)$th column, the values of the N observations of the M observables. In eq 9, J is the Jacobian of the model with respect to the vector of parameters, evaluated at N observations and the values of the parameters minimizing the nonlinear fit. In eq 10, $H(\ln L)$ is the Hessian of log likelihood with respect to the parameters, and W is a weight matrix consisting of the variances of the values of each data point estimated from the histograms. We explore these expressions more completely in Appendices B and C.

We can also use bootstrap sampling of the original distribution data to generate error estimates, which has proven to be a reliable error estimation method for free energy calculations.[6] Although more computationally intensive, the total burden is relatively low. For example, it takes only 20 min on a single core of a 2.7 GHz Intel i7 processor to perform 200 bootstrap samples, even with 600 000 energy evaluations from each simulation.

Once we have generated error estimates for our estimates of the parameters, we can ask the underlying statistical question of whether deviations from the true result are likely caused by statistical error or by errors in the underlying data. In most cases, we will have collected enough samples that the deviation from the fit should be distributed normally. In this case, we can simply compute the standard deviation of the fit parameters and ask how many standard deviations the calculated slope $\beta_2 - \beta_1$ is from the user specified slope. If this difference is consistently more than 2−3 σ away from the true value in repeated tests, it indicates that there are likely errors with the simulations as the two distributions do not have the relationship that they would have if they obeyed a canonical distribution. More sophisticated statistical tests are possible that do not assume normality, but the straightforward normal assumption appears to work fairly well to diagnose problems for all cases presented here. It is important to note that the number of standard deviations a result is estimated away from the expected value is not necessarily a measure of the size of the error. Instead, it is a measure of how certain we are of the error, as we may be measuring either a very small error with extremely high numerical precision or a large error with lower precision.

2.4. Choosing the Parameter Gap. We note that the relationship in eq 6 is true for any choice of the temperatures β_1 and β_2. However, if β_1 and β_2 are very far apart, then the two

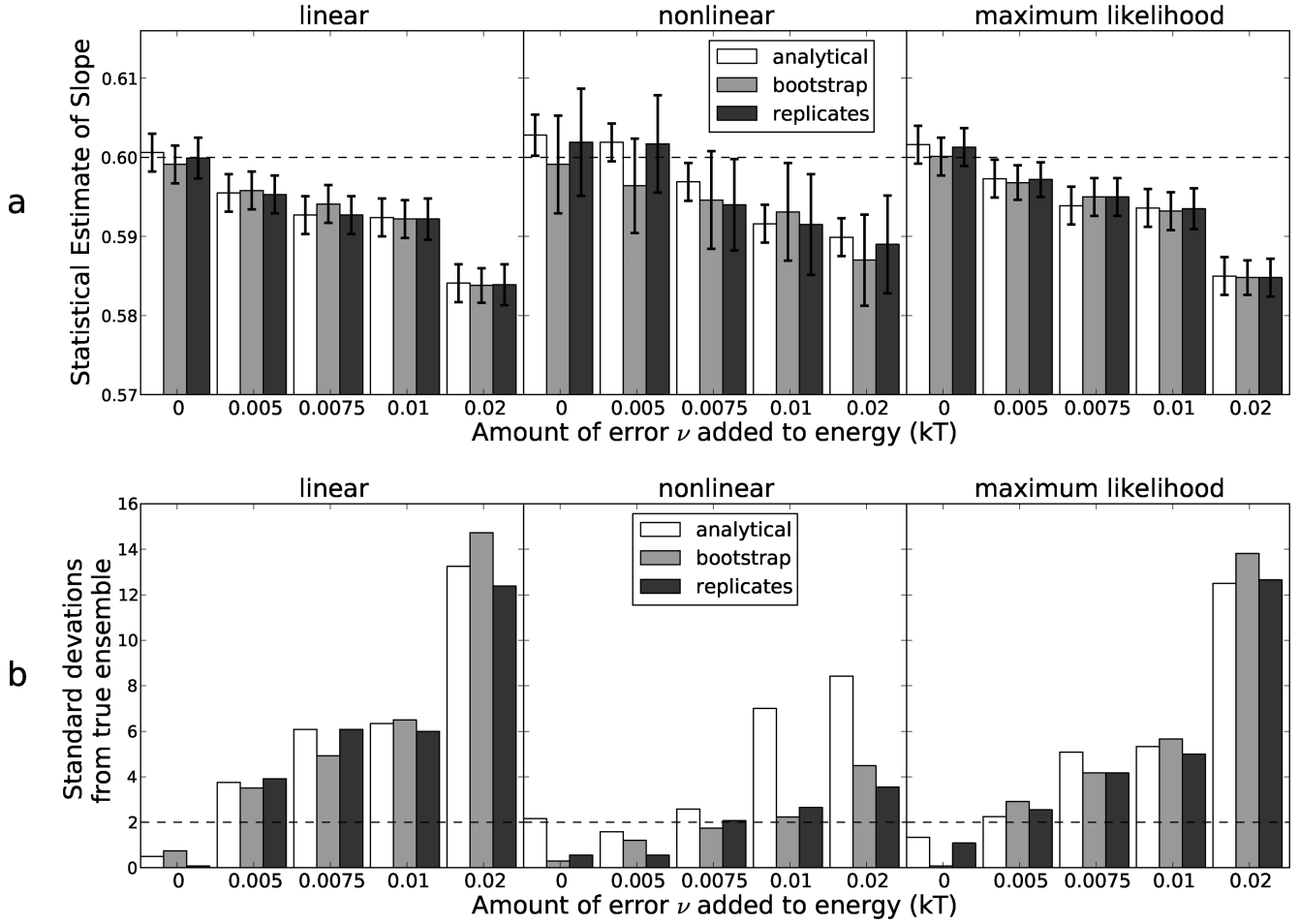


Figure 2. All fitting methods are sensitive determinants of noise in the energy. Upper figure (a) shows the estimated values of $\beta_2 - \beta_1$ (dotted line is true value of 0.6) for the three fitting methods (linear, nonlinear, and maximum likelihood), as a function of the added noise parameter ν. Errors and values are calculated using analytical error estimates (white), bootstrap sampling (light gray), or by repeating the calculations 200 times (dark gray). All three methods for estimating error are consistent except for the nonlinear analytical estimate. All methods have high senstivity to error, though nonlinear fitting is less sensitive. Lower figure (b) shows for the same data the number of standard deviations (dotted line is 95% confidence level) that $\beta_2 - \beta_1$ is from the true value for each fitting method using the three error estimate methods. The dotted line is 95% confidence level.

probability distributions $P(E|\beta_1)$ and $P(E|\beta_2)$ will not be well determined over any range of E in any simulation of reasonable length. If, on the other hand, $\beta_1 = \beta_2$, no information can be obtained because the simulations will be statistically identical. If the two simulations are not statistically identical, there are deeper problems to worry about than if the simulations are ensemble consistent!

Coming in from these two limits, if β_1 and β_2 are moderately far apart, small-sample noise from the extremes of the distribution will make it difficult to determine the deviations from $\beta_2 - \beta_1$. If β_1 and β_2 are too close together, even the relatively small statistical noise at the centers of the distributions will swamp out the information contained in the very slight difference between the user-specified temperature gap and the simulation's actual value for $\beta_2 - \beta_1$. There should therefore be some ideal range of temperature gaps giving the most statistically clear information about deviations from ensemble consistency. We will examine specific choices of this gap for different systems in this study.

2.5. Sampling from the Canonical Ensemble with a Harmonic Oscillator. To study these ensemble validity tests in practice, we first examine a toy model, sampling from a D-dimensional harmonic oscillator, where we can generate uncorrelated samples directly from the analytical Boltzmann distribution. We then use this model to demonstrate the use of this method to identify simulation errors.

For a D-dimensional harmonic oscillator with an equal spring constant K in each dimension and equilibrium location $x_{i,0}$ in each direction, the total potential energy of the system is $E = 1/2K\sum_{i=1}^{D}(x_i - x_{i,0})^2$. The partition function for this model is $Q(\beta) = (2\pi/\beta K)^{D/2}$, meaning the free energy is $A(\beta) = -(D/2\beta)\ln[(2\pi)/(\beta K)]$, and the probability of a given configuration $\vec{x}$ is

$$P(\vec{x}|\beta) = \left(\frac{\beta K}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta K}{2}\sum_{i}^{D}|x_i - x_{i,0}|^2\right)$$

For this exercise, we set $x_{i,0} = 0$ for all i for simplicity, and choose $D = 20$. We specifically do not choose $D = 1$, because it can give results that may not be typical for other choices of dimensions. For $D = 1$, the density of states $\sigma(E)$ is constant for this choice of E, i.e., $P(E) \propto \exp(-\beta E)$ for all spring constants. Unlike most physical densities of states, in this case $E = 0$ has nonzero probability for all temperatures, which means samples from all temperatures have nonnegligible overlap. Harmonic oscillators with $D \gg 1$ have $\Omega(E) = 0$ at $E = 0$ and then rapidly

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

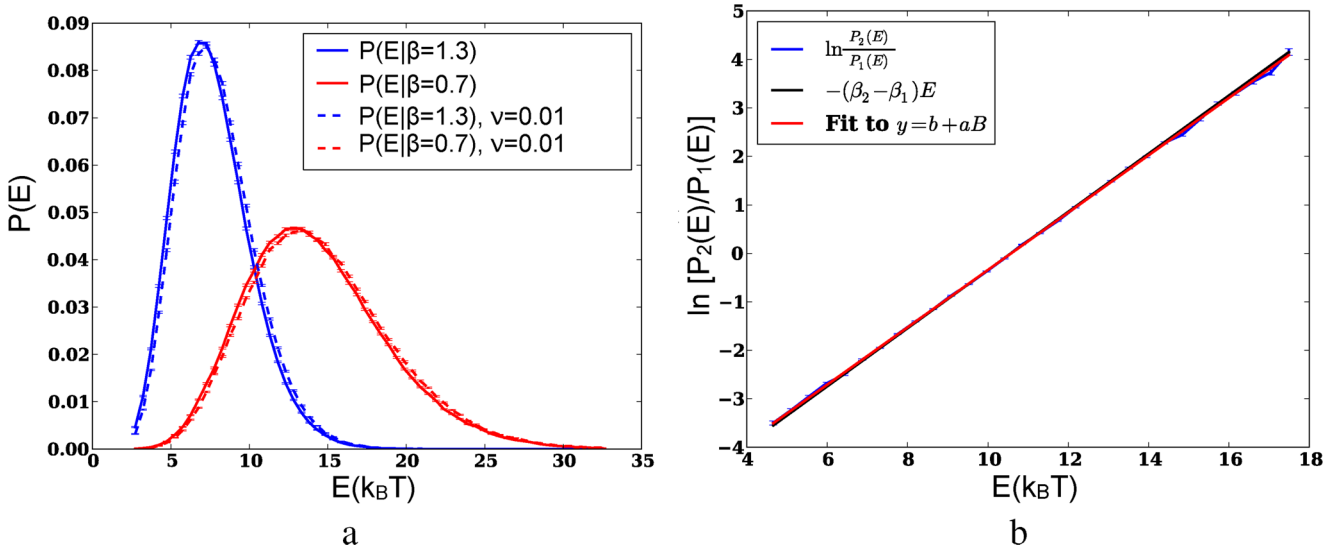


Figure 3. Model energy distributions and discrimination of error. A small amount of noise (0.06% of the average energy) is added to each sample. Such differences affect the distribution minimally (a). When the linear graph of the log ratio probabilities is inspected visually (b), it can be difficult to distinguish errors in the sampling, but fitting quantitatively to the distribution reveals that the deviation in the distribution from analytical results is 5−7 standard deviations (depending on the fitting method used) from the expected value. The system is the same as that used for Figure 2 with error scale $\nu = 0.01$.

increasing as E increases, much more characteristic of more typical physical systems.

2.5.1. Testing for Ensemble Validity with a Toy System with Simulation Noise. Using the toy system introduced in section 2.5, we generate samples with $K = 1$ and $\beta = 1.3$ and 0.7 (the specific choice of temperature gap is explained later) and record the energies of these samples. After generating the samples and recording the energies, we add random noise $\delta E = \nu|N(0,1)|$, where $N(0,1)$ is a Gaussian random variate with mean zero and standard deviation 1, and ν is some small positive constant. The addition of random noise allows us to test the ability of the algorithm to identify simple errors in the energy distributions. In each case, we carry out 200 independent repetitions of this procedure, each time with 500 000 samples from each of the distributions at the two different temperatures. This particular type of error means that the data are generated with the correct probability, but their energies are recorded incorrectly. This pattern might not be typical of actual errors observed in molecular simulations but serves as a useful starting point for characterizing the sensitivity of this procedure. The results are shown as a function of noise in Figure 2, with 0.6 the exact result for $\beta_2 - \beta_1$. We examine the linear, nonlinear, and maximum likelihood fits, with the error calculated by the analytical estimates, sample standard deviations over 200 repetitions, and bootstrap sampling using 200 bootstrap samples from the first of the 200 repetitions.

In all cases as seen in Figure 2, bootstrap sampling closely matches the standard sample error from 200 independent samples, suggesting that bootstrap error estimation is likely to be as effective as independent sampling to identify ensemble errors, as was also observed in previous free energy calculations.[6] Additionally, the analytical error estimates for linear and maximum likelihood fitting closely match the sample standard deviation. This is particularly encouraging because it means that single pairs of simulations are enough to calculate error estimates robustly.

Nonlinear fitting is somewhat less useful, as the nonlinear analytical error estimates appear to noticeably underestimate the actual error, as determined by the sample standard deviation over 200 repetitions. The statistical error in nonlinear fitting is larger than the error in the linear and maximum likelihood estimates, possibly because of a magnified effect of small sample errors. However, all fitting forms (linear, nonlinear, maximum likelihood) in combination with all estimators of the error (analytic, independent replicas, and bootstrap sampling) are relatively sensitive determinants of noise in the energy. Deviations of more than 3σ occur consistently for ν as low as 0.0075, or less than 1% of k_BT, demonstrating that these errors have become statistically significant. Even with $\nu = 0.01$, where the slope is between 5 and 7 standard deviations from the true slope, the visual difference between estimates becomes virtually unnoticeable, for both the actual distributions and the ensemble validation fit, as seen in Figure 3. The ability to sensitively identify errors that cannot be directly visualized demonstrates the utility of this quantitative approach. Overall, it appears that maximum likelihood error estimates are the best method to use, as discretization errors due to poor histogram choice will not matter. However, linear fitting also appears robust, at least for this system.

The ensemble validation relationship is true for all choices of β_1 and β_2, but as discussed, for finite numbers of samples, there are problems with choices of $\beta_1 - \beta_2$ that are either too large or two small. For large slopes, a small sample error in the tails dominates; for small slopes, the small magnitude of the slope becomes difficult to distinguish even at the moderate levels of statistical error occurring near the peaks of the energy distributions. In Figure 2, we use a fixed difference in temperatures. Can we identify an optimal temperature difference to detect error? For this exercise, we select a fixed low level of random error ($\nu = 0.01$) and vary the slope $\beta_2 - \beta_1$ with the average $1/2(\beta_1 + \beta_2)$ fixed at 1, using the analytical estimate from the maximum likelihood parameter estimation and again using 500 000 samples from each distribution.

For fixed noise in the energy function, we see in Table 1 the number of standard deviations from the true slope to the

Table 1. Optimizing Temperature Spacing to Improve Error Detection in the Distribution[a]

β_2	β_1	$\beta_2 - \beta_1$	estimated $\beta_2 - \beta_1$	σ deviation
1.05	0.95	0.1	0.0993 ± 0.0006	1.1
1.10	0.90	0.2	0.1981 ± 0.0007	2.7
1.15	0.85	0.3	0.2970 ± 0.0008	3.9
1.20	0.80	0.4	0.3960 ± 0.0009	4.7
1.25	0.75	0.5	0.4948 ± 0.0010	5.2
1.30	0.70	0.6	0.5936 ± 0.0012	5.4
1.40	0.60	0.8	0.7913 ± 0.0017	5.1
1.50	0.50	1.0	0.9907 ± 0.0027	3.5
1.60	0.40	1.2	1.1930 ± 0.0047	1.5
1.70	0.30	1.4	1.3916 ± 0.0100	0.8

[a]Deviation from the correct slope of the log ratio of the energy distributions as a function of increasing distance between the two distributions, as measured by the magnitude of $\beta_2 - \beta_1$, with fixed noise. The ability to discriminate the error in $\beta_2 - \beta_1$ reaches an optimum at intermediate separation of distributions.

observed slope as a function of the energy gap. The ability to discriminate the error in $\beta_2 - \beta_1$ is lower for both very large and very small temperature gaps, though there is a relatively broad range near the middle where the sensitivity of the test, measured in the number of standard deviations the measured slope is from the true slope, is relatively constant.

Examining the energy distributions at the maximum error discrimination point ($\beta_1 = 0.6$, $\beta_2 = 1.3$), we find that the difference between the centers of the distributions ($14.3k_BT - 7.7k_BT = 6.6k_BT$) is approximately equal to the sum of the standard deviations of the distributions ($4.5k_BT + 2.4k_BT = 6.9k_BT$). This suggest (though does not prove) a general rule-of-thumb that we can maximize the ability to identify errors by using temperatures separated by approximately the sum of the standard deviations of the distributions. The precise value of the difference will not matter particularly in most cases as long as we are somewhat near the optimum. With less data, we might err on the side of using a slightly smaller gap to guarantee good overlap in the distributions.

This rule is simply intended as a guideline, as some sources of error might show up preferentially in the tails and thus require larger temperature gaps to observe but provides useful starting criteria. One example of a physical system which violates this rule is a 1-D harmonic oscillator, which has a constant density of states $\Omega(E)$. Although the statistical error does indeed increase with decreasing overlap, the slope increases faster, and thus sensitivity to statistical error always increases with increasing temperature gap. With fixed error, the sensitivity with noise magnitude $\nu = 0.02$ increases from less than one standard deviation for $\beta_2 - \beta_1 = 0.1$ to over five standard deviations for $\beta_2 - \beta_1 = 1.8$. However, this case is atypical, because the density of states is a constant with the maximum probability always at $E = 0$, so that even when the temperatures are very different there is still nonnegligible overlap in the distributions.

To apply this rule, we still need to estimate the standard distributions of the two distributions. If we assume the variance in energy (and therefore the value of the heat capacity) does not change very much over the relative narrow range of temperature spacings used to perform ensemble validation, then the distributions will also be the same. We can estimate the width σ of the distribution given a known heat capacity C_V. Specifically, $\sigma_E = T(C_V k_B)^{1/2}$, so that for a temperature gap to result in a difference in the centers of the energy distribution of $2\sigma_E$, we must have $2\sigma_E = (\partial E/\partial T)\Delta T = C_V \Delta T$, which reduces to $\Delta T/T = (2k_B/C_V)^{1/2}$. Alternatively, in many cases it may make the most sense to run a short simulation at the "center" temperature $1/2(T_1 + T_2)$ to estimate the variance, and we can use the equivalent relationship $\Delta T/T = 2k_BT/\sigma_E$ to identify a reasonable temperature gap $\beta_2 - \beta_1$ for simulations of a specific system.

2.6. Isobaric–Isothermal Ensembles. Our discussion up to this point has been restricted to NVT systems. However, the same principles can also be applied to check the validity of simulations run at constant temperature and pressure, and of simulations run at constant temperature and chemical potential. We will analyze isobaric–isothermal simulations extensively in this section. We will not examine grand canonical simulations in this paper, though we do include the derivations in Appendix A.

There are at least three useful ways we can analyze NPT simulations for validity. First, let us assume that we have two simulations run at the same pressure but different temperatures. Then, the microstate probabilities are

$$P(\vec{x}, V|\beta, P) = \Delta(\beta, P)^{-1} \exp(-\beta E(\vec{x}) - \beta PV) \tag{11}$$

where $\Delta(\beta,P)$ is the isothermal–isobaric partition function. We then integrate out configurations with fixed instantaneous enthalpy $H = E(\vec{x}) + PV$, where $\vec{x}$ here is shorthand for both position and momentum variables, not the entire microstate specification. We then have

$$\begin{aligned} P(H|\beta, P) &= \frac{\beta P}{h^{3N}} \int_V \int_{\vec{x}} \delta[E(\vec{x}) + PV - H]\Delta(\beta, P)^{-1} \\ &\quad \exp(-\beta(E(\vec{x}) + PV)\, d\vec{x}\, dV \\ &= \frac{\beta P}{h^{3N}} \Omega'(H, P)\Delta(\beta, P)^{-1} \exp(-\beta H) \end{aligned}$$

where $\Omega'(H,P)$ is a density of states counting the number of states with a given value of $H = E + PV$ and is explicitly a function of P, but not β. The prefactor of βP comes from the requirement to cancel the units in the integral, ignoring factors of N relating to the distinguishability of particles, which will cancel in the ratio of distributions in all cases. Because both simulations have the same pressure, we arrive directly at a new ensemble validation relationship:

$$\frac{P(H|\beta_2, P)}{P(H|\beta_1, P)} = \frac{\beta_1 \Delta(\beta_1, P)}{\beta_2 \Delta(\beta_2, P)} \exp(-[\beta_2 - \beta_1]H) \tag{12}$$

$$\ln\left(\frac{P(H|\beta_2, P)}{P(H|\beta_1, P)}\right) = \ln(\beta_1/\beta_2) + [\beta_2 G_2 - \beta_1 G_1] - [\beta_2 - \beta_1]H \tag{13}$$

The exact same ensemble validation statistical tests can therefore be applied with H in place of E and the Gibbs free energy (or free enthalpy) G plus a small correction factor in the place of A.

We can also look at the probability of the volume alone by integrating out the energy E at fixed volume:

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909–926

$$P(V|\beta, P_1) = \frac{\beta_1 P_1}{h^{3N}}\Delta(\beta, P_1)^{-1}Q(\beta, V)\exp(-\beta P_1 V)$$
$$P(V|\beta, P_2) = \frac{\beta_2 P_2}{h^{3N}}\Delta(\beta, P_2)^{-1}Q(\beta, V)\exp(-\beta P_2 V)$$
$$\frac{P(V|\beta, P_2)}{P(V|\beta, P_1)} = \frac{P_1\Delta(\beta, P_1)}{P_2\Delta(\beta, P_2)}\exp(-\beta[P_2 - P_1]V) \quad (14)$$

$$\ln\frac{P(V|\beta, P_2)}{P(V|\beta, P_1)} = \ln(P_1/P_2) + [\beta(G_2 - G_1)] - [\beta(P_2 - P_1)V] \quad (15)$$

We can then use the same techniques already described with $\Delta(\beta,P_1)$ in the place of $Q(\beta,V)$, P_1 and P_2 in the place of β_1 and β_2, and βV in the place of E.

Finally, we can treat the joint probability distributions with both V and E varying independently:

$$P(V, E|\beta_1, P_1) = \frac{\beta_1 P_1}{h^{3N}}\Omega(V, E)\Delta(\beta_1, P_1)^{-1}\exp(-\beta_1 E - \beta_1 P_1 V)$$
$$P(V, E|\beta_2, P_2) = \frac{\beta_2 P_2}{h^{3N}}\Omega(V, E)\Delta(\beta_2, P_2)^{-1}\exp(-\beta_2 E - \beta_2 P_2 V)$$
$$\frac{P(V, E|\beta_2, P_2)}{P(V, E|\beta_1, P_1)} = \frac{\beta_1 P_1\Delta(\beta_1, P_1)}{\beta_2 P_2\Delta(\beta_2, P_2)}\exp([\beta_2 - \beta_1]E + [\beta_2 P_2 - \beta_1 P_1]V) \quad (16)$$

$$\ln\frac{P(V, E|\beta_2, P_2)}{P(V, E|\beta_1, P_1)} = \ln(\beta_1 P_1/\beta_2 P_2) + [\beta_2 G_2 - \beta_1 G_1] - [\beta_2 - \beta_1]E - [\beta_2 P_2 - \beta_1 P_1]V \quad (17)$$

We can apply most of the same methods described previously with slight modifications for the additional dimensions. For example, when fitting the log ratio of the distributions, we must now perform a multilinear fit in V and E. Multiple variable nonlinear fitting can also be employed. However, in both cases, we can quickly run into numerical problems because of the difficulty of populating multidimensional histograms with a limited number of samples, making discretization error worse. The maximum likelihood method, which already appears to be the most reliable method for estimating single variables, does not require any histograms and thus is free from discretization error in any dimension. In examining joint variation in E and V in this study, we therefore focus on only the maximum likelihood method.

For maximum likelihood maximization, we again need to clarify what the free variables are in order to fix the form of the probability distribution. The first two are $\Delta G = G_2 - G_1$, setting $G_1 + G_2 = 0$, and $\Delta\beta = \beta_2 - \beta_1$, setting $(\beta_1 + \beta_2)/2 = \beta_{ave} = \text{const}$, as before. By analogy, we set $(P_1 + P_2)/2 = P_{ave}$, with the variable $\Delta P = P_1 - P_2$. Both β_{ave} and P_{ave} are then set at the averages of the applied β and P of the two simulations. We then find that

$$(\beta_2 P_2 - \beta_1 P_1) = \frac{1}{2}((\Delta\beta)(P_2 + P_1) + (\beta_2 + \beta_1)(\Delta P)) = (\Delta\beta)P_{ave} + \beta_{ave}(\Delta P) \quad (18)$$

The explicit maximum likelihood equations for enthalpy, volume, and joint energy and volume are then

$$\ln\frac{P(H|\beta_2, P)}{P(H|\beta_1, P)} = \beta_{ave}(\Delta G) - (\Delta\beta)H \quad (19)$$
$$\ln\frac{P(V|\beta, P_2)}{P(V|\beta, P_1)} = \beta(\Delta G - (\Delta P)V) \quad (20)$$
$$\ln\frac{P(E, V|\beta_2, P_2)}{P(E, V|\beta_1, P_1)} = \beta_{ave}(\Delta G) - (\Delta\beta)E - (\Delta\beta)P_{ave}V - (\Delta P)\beta_{ave}V = \beta_{ave}(\Delta G) - (\Delta\beta)(E + P_{ave}V) - \beta_{ave}(\Delta P)V \quad (21)$$

omitting the unchanged prefactors involving logarithms of the ratios of the known intensive variables β_1, β_2, P_1, and P_2. In general, we can ignore this term because we usually do not care about the exact value of the free energy difference ΔG between the paired simulations and so therefore do not need to break the constant term down into its components.

2.7. Sampling from the Isobaric−Isothermal Ensemble for a Toy Problem. To better understand how to validate the volume ensemble, we examine a toy model sampling from a modified harmonic oscillator potential. In this case, the harmonic spring constant is increased by decreasing the system volume in order to add a PV work term to the system. We set the harmonic force constant $K = (a/V)^2$, and for simplicity set $x_0 = 0$. This means that $\Delta(P, \beta) = \int_V Q(V, \beta)\exp(-\beta V)\,dV$, which gives

$$\Delta(P, \beta) = (\beta P)^{-2}\sqrt{\frac{2\pi}{a^2\beta}}$$
$$P(x, V|\beta, P) = a(\beta P)^2\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{\beta a^2 x^2}{2V^2} - \beta PV\right)$$

We use the Gibbs sampler[7] to generate configurations from the joint distribution $P(x,V)$ in eq 22 by alternating sampling in $P(E|V)$ and $P(V|E)$. To sample randomly from $P(x|V)$, we observe that x will always be distributed as a Gaussian, with standard deviation $\sigma = (K/\beta)^{1/2} = (V/a)\beta^{-1/2}$. To perform conditional sampling in the system volume dimension, we must sample according to the conditional distribution $P(V|x_i) \propto \exp(-\beta(a^2x_i^2/2V^2) - \beta PV)$. This is not a typical continuous probability family, so there is no simple formula for generating samples from this distribution. However, we note that the distribution is strictly less than $M\exp(-\beta PV)$, where M is the ratio of the normalizing constant for the exponential distribution and the normalizing constant for the exponential plus the harmonic term. We can then sample V from the exponential distribution $\exp(-\beta PV)$ and perform rejection sampling to sample from the strictly smaller desired distribution $P(V|x)$. Initially, it appears that the smaller the difference between the two distributions (i.e., the smaller $-\beta a^2x_i^2/2V^2$) is, the more efficient the sampling will be. However, because x_i is

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

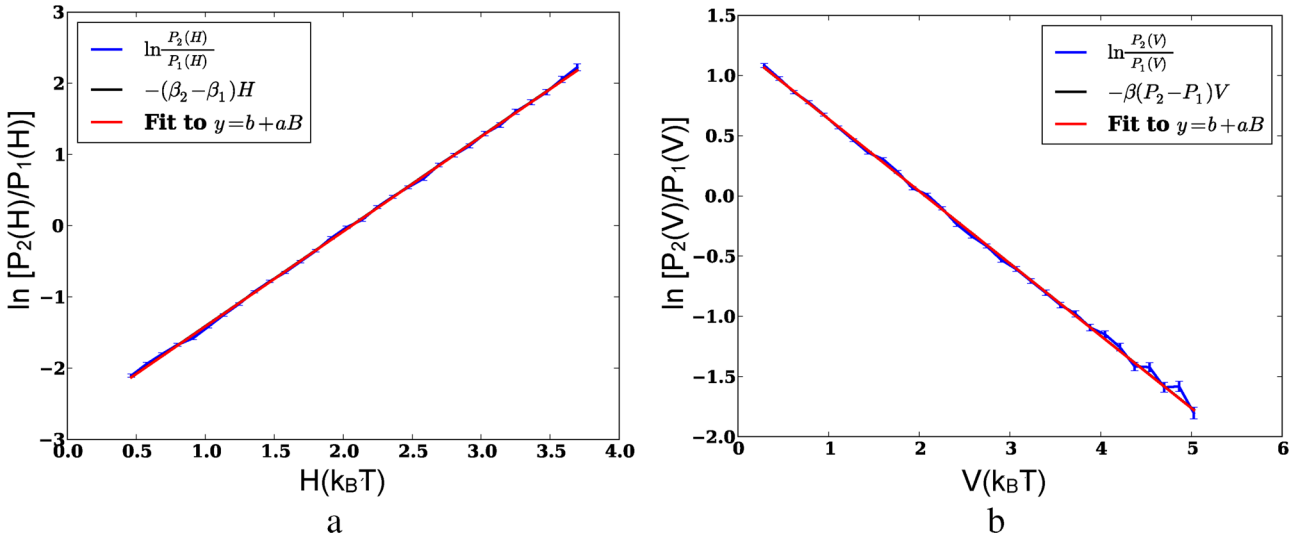


Figure 4. Validation of distributions for harmonic oscillators with pressure. We can accurately validate the isothermal−isobaric distributions of enthalpy (a) and volume (b) for our harmonic oscillator toy problem with a volume-dependent spring constant.

generated from a Gaussian distribution, $\langle x^2 \rangle = \beta V/a^2$, then the average efficiency reduction factor becomes $\exp(-\beta/2)$, independent of P or a, so the acceptance ratio is only significantly affected by the temperature.

2.7.1. NPT Model System Results. For all tests, we generate 250 000 samples from each of the paired distributions. To examine the enthalpy, we pick $\beta_1 = 2/3$, $\beta_2 = 2$, and $P_1 = P_2 = 1$, and using the maximum likelihood method we estimate $\beta_2 - \beta_1 = 1.3341 \pm 0.0040$, only 0.2 standard deviations from the true answer of 4/3 (see Figure 4a for the linear plot). To validate the volume sampling, we pick $\beta_1 = \beta_2 = 1.0$ and $P_1 = 1.3$ and $P_2 = 0.7$ and find that $\beta(P_2 - P_1) = -0.6013 \pm 0.0025$, 0.53 standard deviations from the true answer of −0.6 (see Figure 4b for the linear plot). Finally, when we examine the joint variation of energy and volume, we use $\beta_1 = 0.6$, $\beta_2 = 0.8$, $P_1 = 0.8$, and $P_2 = 1.2$, which gives us 0.20035 ± 0.00318 for the slope $(\beta_2 - \beta_1)$ and -0.48129 ± 0.00185 for the slope $\beta_2 P_2 - \beta_1 P_1$, which are 0.1 and 0.7 standard deviations from the true answers 0.2 and −0.48, respectively. We see that indeed these equations properly capture entropy and volume distributions.

2.7.2. Picking Intervals for Enthalpy and Volume Tests. In the NPT case with differing temperatures and constant pressure, the instantaneous enthalpy $E + PV$ takes the place of the energy, and a two standard deviation temperature gap will mean choosing temperatures separated by $(2k_B/C_P)^{1/2}$, instead of $(2k_B/C_V)^{1/2}$. In the case of an NPT simulation performed with constant temperature and at differing pressures, we want $2\sigma_V = \Delta P(\partial V/\partial P)_T$. We can use the distribution of volume fluctuations to find that $2\sigma_V = (\partial V/\partial P)_T = \sigma_V^2/k_B T$. We therefore must have that $|\Delta P| = 2k_B T/\sigma_V$, or in terms of the physical measurable isothermal compressibility $\kappa_T = -1/V(\partial V/\partial P)_T$, $\Delta P = (2k_B T/V\kappa_T)^{1/2}$. Again, this is a guideline, not a strict rule; short simulations at the simulation average can also be useful to identify the spread of the distributions, as the answer must only be in the right range. For joint distributions, the analysis is more complicated, but it seems reasonable to use 2σ in both directions, perhaps erring on the low side to ensure sufficient samples.

3. MOLECULAR SYSTEMS

3.1. Kinetic Energy and Potential Energy Independently Obey the Ensemble Validation Equation. In most molecular systems (for example, ones without applied magnetic fields), the potential energy of the system can be assumed to be independent of the velocities and masses of the particles. Thus, the potential and kinetic energy are separable, and we can write

$$\begin{aligned} P(E_{pot} + E_{kin}|\beta) &= Q_{kin}(\beta)^{-1} Q_{pot}(\beta)^{-1} \Omega(E_{pot}) \Omega(E_{kin}) \\ &\quad \exp(-\beta E_{pot}) \exp(-\beta E_{kin}) \\ &= [Q_{kin}(\beta)^{-1} \Omega(E_{kin}) \exp(-\beta E_{kin})] \\ &\quad [Q_{pot}(\beta)^{-1} \Omega(E_{pot}) \exp(-\beta E_{pot})] \\ &= P(E_{pot}|\beta) P(E_{kin}|\beta) \end{aligned}$$

The separability of the density of states occurs again because the momenta can be sampled independently of the coordinates. The ensemble validation algorithm is therefore valid for the kinetic and potential energies independently as well, so that

$$\frac{P(E_{kin}|\beta_2)}{P(E_{kin}|\beta_1)} = \frac{Q_{kin}(\beta_2)}{Q_{kin}(\beta_1)} \exp(-[\beta_2 - \beta_1]E_{kin}) \tag{22}$$

$$\frac{P(E_{pot}|\beta_2)}{P(E_{pot}|\beta_1)} = \frac{Q_{pot}(\beta_2)}{Q_{pot}(\beta_1)} \exp(-[\beta_2 - \beta_1]E_{pot}) \tag{23}$$

In the case of kinetic energy, Q_{kin} is simply $\Pi_{i=1}^{N} \int_{-\infty}^{\infty} \exp(-\beta p_i^2/m_i)\, dp_i = \Pi_{i=1}^{N} (m_i/\pi\beta)^{3/2}$, meaning the probability ratio is

$$\frac{P(E_{kin}|\beta_2)}{P(E_{kin}|\beta_1)} = \left(\frac{\beta_2}{\beta_1}\right)^{3N/2} \exp([\beta_1 - \beta_2]E_{kin}) \tag{24}$$

$$= \left(\frac{\beta_{ave} + \Delta\beta}{\beta_{ave} - \Delta\beta}\right)^{3N/2} \exp(-\Delta\beta E_{kin}) \tag{25}$$

which is now in terms of the single free parameter $\Delta\beta = \beta_2 - \beta_1$ rather than two parameters. Note that this is true for both

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

identical and nonidentical particles, since the mass terms will cancel out for all i. In the case of kinetic energy, we can obtain a distribution for each distribution alone, because the kinetic energy is simply the sum of $3N$ random normal variables with standard deviations $(m_i)^{-1/2}p_i$ and is thus a χ^2 distribution with $3N$ (minus any center of mass variables removed from the simulation) degrees of freedom (DOF). For more than 60 DOF, corresponding to about 20 particles, the χ^2 distribution is essentially indistinguishable from a normal distribution with the mean equal to the sum of the means of the individual distributions, which in this case is simply the average kinetic energy. By equipartition, the total kinetic energy will simply be $3N/2\beta$. The standard deviation can be computed by noting that the $\sigma^2 = k_B T^2 C_V$ and that the heat capacity due to the kinetic energy is the ideal gas heat capacity, $3Nk_B/2$. Thus, $\sigma^2 = 3N/2\beta^2$, and

$$P(E_{kin}) = \frac{\beta}{\sqrt{3N\pi}} \exp\left(-\frac{\left(\beta E_{kin} - \frac{3N}{2}\right)^2}{3N}\right)$$

to high accuracy for any number of molecules typical in molecular simulations. In the above formulas, $3N$ should be replaced by the correct number of DOF if constraints are implemented or if any center of mass degrees of freedom are removed. Standard methods for testing the normality of distributions with known means and standard deviations can be used, such as inspecting Q−Q plots or the Anderson−Darling test.[8] If the number of degrees of freedom is not available, as may be the case when one is analyzing data provided by someone else, then this can be estimated from the average of the kinetic energy by equipartition as $\langle E_{kin}\rangle = (k_B T/2)(\#DOF)$. If the kinetic energy is not equal to this value, then the reported temperature will not even be correct, which should be noticed from simpler outputs of the simulation before running any other more sophisticated analysis like the procedures described in this paper.

The kinetic energy distribution, in addition to following the ensemble validation formula, can therefore be checked directly as well with essentially no overhead, though this does not seem to be common practice in molecular simulation validation. The potential energy formula can be used to either validate the potential energies separately or can be used for Monte Carlo simulations, where only potential energies are defined. It is also possible to perform this separation in terms of ideal gas and canonical partition functions, but it does not change the results, as the volume is constant.

To obtain separability of kinetic and potential energies in an NPT ensemble, we start by writing the isobaric−isothermal partition function in terms of kinetic and potential energy portions of the canonical partition functions and note that the kinetic energy part is independent of the volume.

$$\Delta(\beta, P) = \beta P \int_V Q_{kin}(\beta) Q_{pot}(\beta, V) \exp(-PV)\, dV$$

$$\Delta(\beta, P) = Q_{kin}\beta P \int_V Q_{pot}(\beta, V) \exp(-PV)\, dV$$

$$\Delta(\beta, P) = Q_{kin}(\beta)\Delta_{pot}(\beta, V)$$

Then in terms of probabilities, we have

$$P(E, V|\beta, P) = Q_{kin}(\beta)^{-1}\Delta_{pot}(\beta, V)^{-1} \exp(-\beta E_{kin} - \beta E_{pot} - \beta PV)$$

$$P(E_{kin}|\beta, P) = Q_{kin}(\beta)^{-1} \exp(-\beta E_{kin})$$

$$P(E_{pot}, V|\beta, P) = \Delta_{pot}(\beta, V) \exp(-\beta E_{pot} - \beta PV)$$

This separation again makes it possible to validate NPT Monte Carlo simulations by removing the kinetic energy.

3.2. Molecular Dynamics of Lennard-Jones Spheres. We next illustrate the utility of the ensemble validation formula for molecular simulations. For this study, we used a simulation of 300 Lennard-Jones particles using a beta version of the Gromacs 4.6 simulation code compiled in double precision. We used the Rowley, Nicholson, and Parsonage argon parameters for Lennard-Jones spheres (σ = 0.3405 nm, ε = 119.8 K, k_B = 0.996072 kJ/mol)[9] and simulated at $\rho = 0.85\rho_c$ (where ρ_c is the Lennard-Jones critical pressure), meaning the box is of length 3.5328256 nm and $T = 0.85T_c$ = 135.0226. Velocity Verlet integration was used, with the exception of the Gromacs stochastic integration method, which is only defined for the leapfrog Verlet algorithm. The linear center of mass momentum was removed every step, and a long-range homogeneous dispersion correction was applied to the energy. Unless otherwise specified, a Lennard-Jones switch between 0.8 and 0.9 nm was used, including a homogeneous long-range correction with a neighborlist at 1.0 nm, a neighborlist update frequency of 5 step, and a time step of 8 fs. Temperature coupling algorithms were carried out with a coupling constant of τ_T = 1.0 ps. A total of 62.5 million MD steps were simulated for all simulations, equivalent to 500 ns with an 8 fs time step, with the last 490 ns used for analysis. Unless otherwise specified, the low and high temperatures are T = 132.915 and T = 137.138, respectively, chosen to be approximately 0.7 times the estimated ideal σ gap from the rule of thumb, using $C_V \approx$ 8.5 kJ K^{-1} mol^{-1} from a preliminary simulation of the system.

3.3. Molecular Example: Validating Temperature Control Algorithms. Using this Lennard-Jones system, we first examine temperature control algorithms implemented in Gromacs: Bussi−Parrinello,[10] with stochastic scaling of the target temperature, Andersen temperature control,[11] a variant of Andersen temperature control with the velocity of all atoms randomized at some regular interval τ_t, Nosé-Hoover,[2] stochastic dynamics, and Berendsen velocity scaling.[12] All of these temperature control algorithms are proven in theory to give the correct canonical distribution in the limit of long time scales[13] with the exception of the Berendsen temperature algorithm, which is known to give an incorrect, overly narrow kinetic energy distribution.[14−16] We examine the deviations of the total, potential, and kinetic energies, using analytic errors from the maximum likelihood fits. In this analysis, we will often use the ΔP and ΔT (from the maximum likelihood expressions) to describe the deviations from the true distribution to make them more intuitive. We can calculate ΔT from $\Delta\beta$ by assuming an average $\beta_{eve} = 1/2(\beta_1 + \beta_2)$ and calculating $T_2 = k_B^{-1}(\beta_{eve} + \Delta\beta/2)^{-1}$ and $T_1 = k_B^{-1}(\beta_{eve} - \Delta\beta/2)^{-1}$. In all molecular simulations, we also compute the correlation times τ of the energy observables, using the timeseries module of the pymbar code distribution[17] and subsample the data with frequency $2\tau + 1$ to obtain uncorrelated samples. We find that for the kinetic energies alone, the correlation times are actually artificially short when using the algorithm in the timeseries module, which only

Table 2. Ensemble Validation of Different Temperature Control Algorithms[a]

	true ΔT = 4.223					
	total		potential		kinetic	
thermostat	estimated ΔT	σ deviation	estimated ΔT	σ deviation	estimated ΔT	σ deviation
none (NVE)	N/A (constant)		4.388 ± 0.115	1.4	3.048 ± 0.112	10.5
Berendsen	9.369 ± 0.122	42.2	4.606 ± 0.086	4.5	29.034 ± 0.364	68.3
stochastic	4.172 ± 0.066	0.8	4.098 ± 0.081	1.6	4.251 ± 0.091	0.3
Nosé–Hoover	4.197 ± 0.067	0.4	4.220 ± 0.082	0.03	4.186 ± 0.090	0.4
Andersen	4.212 ± 0.066	0.2	4.226 ± 0.081	0.03	4.226 ± 0.090	0.03
Andersen (Massive)	4.188 ± 0.079	0.4	4.176 ± 0.097	0.5	4.217 ± 0.107	0.06
Bussi–Parrinello	4.167 ± 0.066	0.8	4.272 ± 0.082	0.6	4.155 ± 0.089	0.8

[a]All studied thermostats are consistent with a canonical ensemble, with the exception of the Berendsen thermostat, with deviations from the true slope generally 1σ or less. The true slope is 0.027865 $k_B T^{-1}$, equivalent to ΔT = 4.223. All errors are computed using the maximum likelihood method with the analytical error estimate. NVE simulations also deviate from the canonical ensemble, though the potential energy distributions do not statistically deviate.

Table 3. Effect of Step Size on Ensemble Consistency[a]

			true ΔT = 4.223 K					
	true T_{low} = 132.915 K		total		potential		kinetic	
Δt (fs)	T_{low}(K)	σ deviation	estimated ΔT	σ deviation	estimated ΔT	σ deviation	estimated ΔT	σ deviation
8	132.924 ± 0.040	0.2	4.230 ± 0.047	0.2	4.237 ± 0.058	0.2	4.186 ± 0.063	0.6
16	132.933 ± 0.028	0.7	4.183 ± 0.032	1.2	4.253 ± 0.040	0.8	4.106 ± 0.043	2.7
24	132.933 ± 0.023	0.8	4.058 ± 0.026	6.4	4.140 ± 0.032	2.6	4.023 ± 0.035	5.8
32	132.905 ± 0.020	0.5	3.967 ± 0.030	8.6	4.199 ± 0.028	0.9	4.054 ± 0.022	7.6
40	132.948 ± 0.019	1.8	3.988 ± 0.020	11.6	4.178 ± 0.026	1.7	3.877 ± 0.027	12.9
40 (E_{kin} ave)	132.917 ± 0.018	0.1	4.266 ± 0.021	2.6	4.275 ± 0.026	2.0	4.296 ± 0.029	2.6

[a]Total and kinetic energy gradually deviate from the true ensemble as step size increases, becoming statistically noticeably near, but not at the instability point. Potential energy distributions deviate less significantly from a canonical distribution than the kinetic energy distributions. The average half step kinetic energy estimator using the leapfrog verlet integration algorithm deviates less from the true distribution.

integrates out to the first crossing of the x axis. We therefore in this study use the correlation times for the potential energies, which are equal to or longer than the correlation times of the kinetic or total energies. Subsampling more frequently than required only affects the results by decreasing the statistical accuracy due to collecting to few uncorrelated measurements, which for a validation test is not as large a problem as significantly undersampling the statistical error, which results in using correlated data. For the thermostat comparison, we use the subsampling frequency of 40 ps, which is the maximum among all methods, except for the Andersen massive variant, for which we use 60 ps.

This comparison is presented in Table 2, with all estimates and errors computed using maximum likelihood methods described in this paper. We see that all temperature control methods appear to be consistent with a canonical ensemble, with deviations from the true slope generally 1σ or less, with the exception of the Berendsen temperature control method. NVE kinetic energy distributions deviate from the canonical ensemble, though interestingly, potential energy distributions do not deviate from the correct distribution to a statistically noticeable level. In all cases where there are deviations of the kinetic energy, the distributions of the potential energies are closer to the true distribution than the kinetic energy or total energy distributions are; as noted, for NVE, the potential energy distribution is statistically indistinguishable from the NVT potential energy distribution.

3.3.1. Molecular Example: The Effect of Large Step Size. It is well-known that step sizes that are too large can lead to rapid heating of an NVE molecular dynamics simulation as the integration deviates from the conserved energy trajectory. This deviation was one of the initial motivations leading to the development of thermostats. However, using a thermostat to bleed out the extra thermal energy created by violations of the conservation of energy effectively creates a steady state system. The system has heat being both pumped in by numerical integration error and pumped out by the thermostat, with the kinetic energy having the desired average. However, this steady state process does not necessarily have the correct Boltzmann probability distribution.

There has been relatively little investigation of the effect of step size on the ensemble itself when temperature control is applied,[18] especially for atomistic simulations. Here, we examine step sizes from 8 to 40 fs. In the Gromacs code, a step size of 48 fs with Lennard-Jones argon cause segmentation faults within just a few nanoseconds and therefore represents the upper limit of stability with a thermostat coupling constant with τ_T = 1 ps. In these units, the reduced time is $\sigma(M/\varepsilon)^{1/2}$ = 0.1245 ps, so the stability limit is about 0.386 reduced time units.

However, being below the limit of stability does not necessarily mean that the ensemble is correctly reproduced. To analyze the distributions generated by long step sizes, we use the Bussi–Parrinello thermostat algorithm and step sizes ranging from 8 to 40 fs (Table 3). Uncorrelated potential energy samples were 20 ps apart as determined by the timeseries module, consistent over all steps sizes to within 10%. Uncertainties in effective temperature are determined directly from the subsampled kinetic energies, rather than using

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909–926

Table 4. Effect of Abrupt Cutoff on Ensemble Validation[a]

				true ΔT = 4.223					
T_{low} = 132.915				total		potential		kinetic	
r_c (LJ σ)	E_{cons} (gained kJ/ns)	estimated T_{low} (K)	σ deviation	estimated ΔT	σ deviation	estimated ΔT	σ deviation	estimated ΔT	σ deviation
2	9400	133.952 ± 0.045	23.0	4.102 ± 0.058	2.1	4.018 ± 0.084	2.4	4.122 ± 0.070	1.4
2.5	1140	133.043 ± 0.045	2.9	4.206 ± 0.052	0.3	4.177 ± 0.059	0.8	4.176 ± 0.071	0.7
3	239	132.941 ± 0.045	0.6	4.232 ± 0.051	0.2	4.291 ± 0.065	1.1	4.213 ± 0.071	0.1
3.5	104	132.930 ± 0.045	0.3	4.226 ± 0.050	0.1	4.302 ± 0.058	1.4	4.135 ± 0.070	1.2
4	78	132.929 ± 0.045	0.3	4.302 ± 0.050	1.6	4.307 ± 0.057	1.6	4.192 ± 0.071	0.4

[a]Distributions are surprisingly ensemble consistent for most values of abrupt cutoff for Lennard-Jones spheres, with only the shortest cutoff distances (less than 3 LJ σ) showing statistically clear violations.

Table 5. Molecular Validation of Ideal Gap Guidelines[a]

			kinetic		potential		total	
$\Delta T/T$	$\beta_2 - \beta_1$	$n \times gap_{opt}$	est. slope	σ deviation	est. slope	σ deviation	est. slope	σ deviation
0.0156	0.01393	0.4	0.01302 ± 0.00023	4.0	0.01413 ± 0.00021	0.9	0.013518 ± 0.00016	2.6
0.0313	0.02787	0.7	0.02598 ± 0.00025	7.7	0.02585 ± 0.00018	1.4	0.026433 ± 0.00018	7.8
0.0469	0.04182	1.1	0.03892 ± 0.00027	10.7	0.04152 ± 0.00027	1.1	0.039990 ± 0.00023	8.1
0.0626	0.05578	1.4	0.05190 ± 0.00031	12.6	0.05526 ± 0.00031	1.8	0.053343 ± 0.00029	8.6
0.0938	0.08378	2.1	0.07791 ± 0.00042	14.2	0.08290 ± 0.00045	2.0	0.079531 ± 0.00049	8.6
0.1251	0.11189	2.8	0.10464 ± 0.00059	12.2	0.11150 ± 0.00071	1.6	0.107768 ± 0.0010	4.2
0.1877	0.16867	4.2	0.1522 ± 0.0016	10.1	0.1654 ± 0.0022	1.5	0.165805 ± 0.0057	0.5
0.2502	0.22645	5.6	0.2099 ± 0.0037	4.5	0.220 ± 0.010	0.6	0.249 ± 0.072	0.3

[a]We test the temperature gap for maximum discrimination rule with Lennard-Jones argon with time step Δt = 32 fs. Maximum discrimination of error in the ensemble consistency for the different energy terms occurs between 1 and 2 times the estimated gap rule (column 3).

Gromacs g_energy output, in order to have a more consistent treatment of uncertainties between different observables.

In Table 3, we note that total and kinetic energy gradually deviate from the true ensemble with the deviation becoming extremely large near the instability point. For this particular system, average temperatures determined by averages of the kinetic energy from a simulation (shown for the lower temperature simulation in Table 3) are not as useful in distinguishing systems that are being forced back to the desired average kinetic energy using the thermostat.

Interestingly, potential energy distributions deviate much less significantly from the canonical distribution than the kinetic energy distributions to the extent that this deviation is not statistically significant. This may relate to the fact that the standard estimator of the kinetic energy in the velocity Verlet algorithm, the sum of the squared full step velocities times the masses, is not as accurate as the estimator of the kinetic energy of the leapfrog Verlet algorithm, which uses the averaged half-step kinetic energies. Although deviations increase with the square of the step size in both cases, the full step kinetic energies deviate more quickly.[19] We note that it appears to be the choice of kinetic energy estimator, not the integration method per se, that makes a difference, since the two methods give identical NVE trajectories up to numerical precision. The hypothesis that the choice of kinetic energy estimator may make a difference was confirmed by performing the same 40 fs time step simulation with the leapfrog Verlet integrator and the Bussi–Parrinello algorithm, resulting in significantly better kinetic energy distribution without statistically altering the potential energy distributions. We note that in this case, although the deviation is statistically very clear, it is not necessarily that large. Even for the 40 fs step kinetic energy, the fitted temperature difference is only off 10%, which is about 0.4 K, which will not make a difference for most applications. We also note that simulations of different molecular systems with different potential functions may have different deviations from ensemble consistency as a function of the distance from the time step stability limit.

3.3.2. Example: Examining the Effect of Cutoffs on Ensemble Consistency. An abrupt cutoff of a radial potential function creates a discontinuity in the force, resulting in steadily increasing temperature in an NVE simulation. This temperature rise can, as in the case of large time step, again be disguised by adding a thermostat, creating a steady state system that does not necessarily obey the canonical distribution. We can examine the effect of this truncated potential on the NVT ensemble using our ensemble consistency tests. We run the same Lennard-Jones argon system with abrupt cutoffs at r_c = 2.0σ, 2.5σ, 3.0σ, 3.5σ, and 4.0σ, where σ here is the Lennard-Jones size, not the standard deviation. Because of quirks in the way Gromacs handles abrupt cutoffs, we create an abrupt cutoff using a potential switch over a distance of 10^{-9} nm, which on the integration time scale effectively becomes an abrupt cutoff. We can measure how much such a simulation violates conservation of energy by monitoring the average increase in the conserved quantity per unit time. In these simulations, we use the Bussi–Parrinello thermostat, with τ_T = 1.0 ps, approximately 120 times the time step, with T_{low} = 132.915 and T_{high} = 137.138. We can measure the magnitude of energy drift by monitoring the change in the conserved quantity over time, which varies from 9.40 × 10^3 kJ mol^{-1} ns^{-1} for r_c = 2.0σ to 78 kJ $mol^{-1}ns^{-1}$ for r_c = 4.0σ. Times between uncorrelated samples, as determined by potential energy differences, were no larger than 25 ps for all systems, so we use this sampling time frequency for all three quantities.

We see in Table 4 that the distributions are surprisingly ensemble consistent for most values of abrupt cutoff for Lennard-Jones spheres despite the fact that the simulation is

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909–926

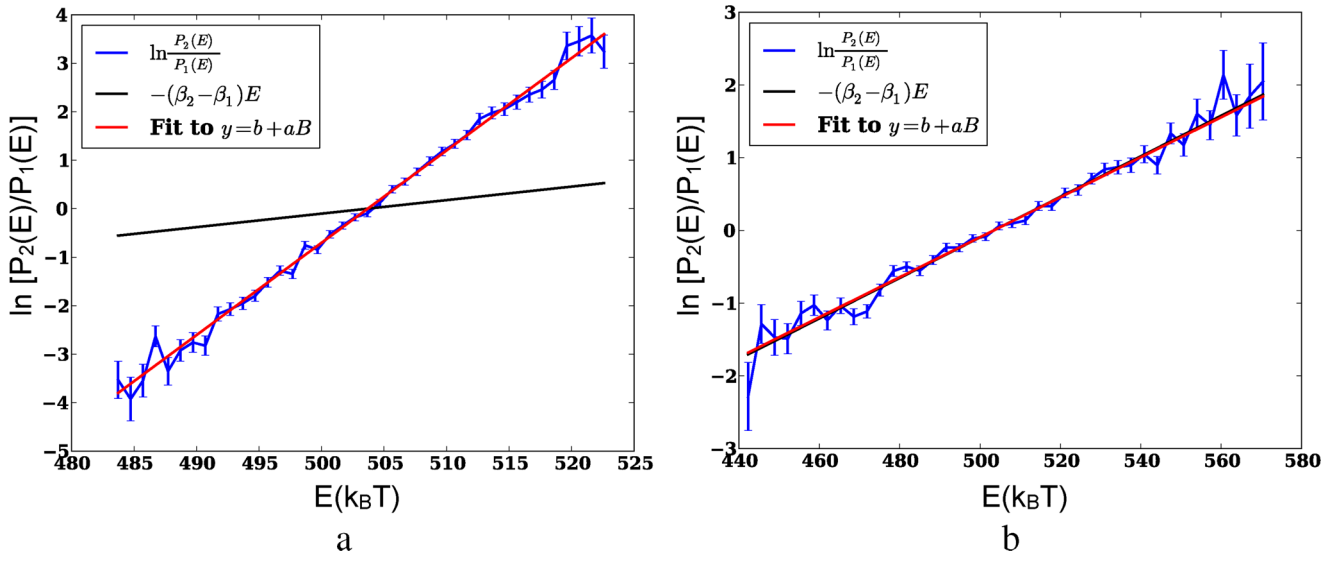


Figure 5. Differences in validation of Berendsen and Nosé–Hoover thermostats. (a) Berendsen temperature control produces simulations deviating greatly from the true distribution; in this case, the slope $\beta_2 - \beta_1$ of the kinetic energy log ratio is 7 times higher than it should be, 68 standard deviations away from the true value. (b) The Nosé–Hoover thermostat, like most others examined here, gives a slope statistically indistinguishable from the proper slope for the kinetic energy portion of the canonical ensemble.

Table 6. Ensemble Validation of Pressure Control Algorithms[a]

	enthalpy		volume		joint energy and volume			
barostat	ΔT	σ deviation	ΔP	σ deviation	ΔT	σ deviation	ΔP	σ deviation
Berendsen	4.176 ± 0.121	19.8	79.5 ± 4.4	17.1	0.69 ± 0.14	7.6	−318.661 ± 7.322	43.9
Parrinello–Rahman	7.022 ± 0.033	3.5	114.58 ± 0.57	9.5	7.168 ± 0.036	0.8	110.971 ± 0.529	7.4
MTTK	7.105 ± 0.029	1.2	115.51 ± 0.50	9.0	7.152 ± 0.031	0.5	111.312 ± 0.457	7.8

[a]Tests of enthalpy distribution, volume distribution, and joint energy and volume distributions. The Berendsen barostat fails badly in all three tests. The other two barostats give correct enthalpy distributions but have small (ΔP off by 5 bar or ≈5%) but statistically clear (7–9σ) errors in the volume distributions. The correct ΔT = 7.138 (1.784 for Berendsen) and the correct ΔP = 120 bar.

gaining more than 200 kJ/mol/ns with a 3 σ cutoff. We note that in this case, the deviation from desired temperature as calculated from average kinetic energy is fairly clear (23 standard deviations for a 2σ cutoff!), and therefore this measure appears to be better at distinguishing violations from the correct distribution than the ensemble consistency check. This contrasts with the case of varying step size, where the ensemble consistency check was more sensitive than the deviation from the correct average kinetic energy. Clearly, multiple validation methods should always be performed.

3.3.3. Validating the Gap Selection Criteria for Molecular Systems. Finally, we attempt to validate our rule of thumb for the ideal temperature gap with molecular systems, since it was derived for a simplified model system. We test the ability to detect error using the same Lennard-Jones argon system with time step Δt = 32 fs using velocity Verlet (see Table 5), as for higher temperatures, a time step of Δt = 40 fs can crash in simulations extending for hundreds of nanoseconds. Measuring the heat capacity as 8.5 kJ mol^{-1} K^{-1} at 135 K leads to a standard deviation of 36 kJ/mol and an estimated ideal temperature gap of 6 K between the means of the two total energy distributions. We see (Figure 5) that we are most sensitive to error in the total energy between 1 and 2 times the estimated gap, meaning that our analytical guidelines were close, but that a slightly larger gap might sometimes be more effective in identifying errors. We note that the kinetic energy standard deviation at 135 K (24 kJ/mol) is only about 2/3 of the total energy standard deviation, but since the total heat capacity (8.5 kJ K^{-1} mol^{-1}) is more than twice as large as the ideal gas heat capacity (3.72 kJ K^{-1} mol^{-1} for this size of system), the kinetic energy distributions have closer mean values than the total energy distributions. Thus, the range of peak discrimination for kinetic energy still falls in the 1 to 2 times the "twice the central standard deviation" rule of thumb when using the distribution of kinetic energies. For molecular systems, the ideal gap might therefore be better estimated using a temperature gap 1.5 to 2 times the estimated gap range. However, a relatively wide range of values allows for discriminating a lack of ensemble validity if sufficient data are collected.

3.4. Examining Pressure Control Algorithms. There are currently three pressure control algorithms implemented in Gromacs: Berendsen,[12] Parrinello–Rahman,[20,21] and the Martyna–Tuckerman–Tobias–Klein (MTTK) algorithm.[22,23] The first two are defined using the leapfrog integrator in Gromacs, and the first and last are defined using the velocity Verlet integrator. We next examine the same small argon system for fluctuations of enthalpy and volume, and the joint fluctuation of volume and energy. A velocity Verlet integrator was used except for Parrinello–Rahman, with Δt = 8 fs. We set the pressure coupling τ_p to 5 ps in all cases and use P = 90 bar and T = 125 K as the average pressure and temperature, resulting in a system well below the critical point. When testing volume fluctuations or joint energy and volume fluctuations, a low pressure of 30 bar and a high pressure of 150 bar were used (ΔP = 120 bar), except for Berendsen pressure control, where

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909–926

Table 7. Ensemble Validation of Different Temperature Control Algorithms with Water[a]

	total		potential		kinetic	
thermostat	slope	σ deviation	slope	σ deviation	slope	σ deviation
Berendsen	51.6 ± 1.1	44.2	7.20 ± 0.12	34.7	4.86 ± 0.12	15.8
stochastic	2.998 ± 0.059	0.04	2.944 ± 0.069	0.8	3.032 ± 0.090	0.4
Nosé−Hoover	2.921 ± 0.058	1.4	2.953 ± 0.068	0.7	2.837 ± 0.089	1.8
Andersen	3.028 ± 0.083	0.4	3.114 ± 0.098	1.2	2.870 ± 0.126	1.0
Andersen (Massive)	3.086 ± 0.083	1.0	3.048 ± 0.097	0.5	3.136 ± 0.127	1.0
Bussi−Parrinello	2.955 ± 0.058	0.8	2.956 ± 0.068	0.6	3.021 ± 0.090	0.2

[a] ΔT = 3 K corresponding to a inverse temperature slope of 0.004023 $(k_B T)^{-1}$. Results are consistent with those performed with argon, with all temperature control algorithms ensemble consistent except for Berendsen.

low and high pressures of 88 and 92 bar (ΔP = 4 bar) were used. A lower range is needed for the Berendsen weak coupling algorithm as the volume distributions are far smaller than is correct for the distribution (already demonstrating a problem). When testing enthalpy fluctuations or joint energy and volume fluctuations, a low temperature of 121.431 K and a high temperature of 128.569 K ($\Delta\beta$ = 0.054987 $k_B T^{-1}$, ΔT = 7.138 K) were used, generated using an estimated C_P of 10.2 kJ/mol from short initial simulations for this system using the estimated gap formula. For Berendsen thermostat simulations, a temperature range of 124.108 to 125.892 K was used ($\Delta\beta$ = 0.013736, equivalent to ΔT = 1.784), as again the overlap between the distributions is very poor for wider parameter differences. Nosé−Hoover temperature control with τ_T = 1 ps was used for both Parrinello−Rahman and MTTK algorithms. The ΔP for joint energy and volume comparisons is smaller because the simulations are run at different temperatures and is equal to ΔP = 114.861 bar for Parrinello−Rahman and MTTK and 2.715 bar for Berendsen.

Looking at Table 6, we see that the Parrinello−Rahman and MTTK algorithms reproduce very accurately the correct enthalpy distributions, deviating very little from the correct $\Delta\beta$, with very high statistical confidence. The precision is high partly because the time between uncorrelated samples (in this case, determined from the largest correlation time of either the energy or the volume) is quite short, in the range of 4−6 ps. The volume distributions, however, are somewhat off, with the effective ΔP in both cases near 115 ± 0.6 instead of 120. For most cases, this will be sufficiently accurate to model physical processes (and is far better than the Berendsen results) but might not be sufficiently accurate for very high precision thermodynamic measurements. The 9σ deviation from the true answer is again not necessarily a sign of how bad the simulation is. In this case, because the slope is nearly correct, it is a sign that it is statistically very *likely* the simulation is at least somewhat off rather than simply being very *bad*. Similar patterns are seen in the joint distribution of E and V, where the effective ΔP is still off by about 5 bar (or around 5%). The deviations are similar for both MTTK and Parrinello−Rahman, even though these integration routines are mostly separated in the Gromacs code.

For Berendsen, the results are uniformly bad. In all cases, the deviation from the expected values is significantly higher than with MTTK or Parrinello−Rahman, with the slopes being much further from the true value even though the statistical error is much higher as well. This deviation exists even though the average temperatures and pressures in the Berendsen case were all well within statistical noise. For example, for the joint distribution analysis, the low and high average pressures were indeed 87.996 ± 0.003 and 91.998 ± 0.005 bar and the average temperatures were 125.865 ± 0.015 K and 124.081 ± 0.015, well within the statistical noise. Errors in the fitting parameters are therefore due to unphysically narrow distributions, not the average values themselves. We note one other potential strange problem with Berendsen volume control combined with the Bussi−Parrinello thermostat. The autocorrelation times are much longer than with other simulation variables, on the order of 20 ps for the energies and 110−130 ps for the volumes. It is not clear what exactly is causing such a slow change of these variables when the time constants themselves are much lower—in this case τ_T = 1 ps and τ_P = 5 ps—but perhaps indicates another reason to avoid Berendsen pressure control.

3.5. Water Simulations. We also examine a somewhat more typical system for molecular simulation, a small box of 900 TIP3P water, a size that might be used to compute pure water properties or small molecule solvation free energies. We again use velocity Verlet integration (with the exception of the Gromacs stochastic integration, which is only defined for the leapfrog Verlet algorithm) with linear center of mass momentum removal every step and a long-range homogeneous dispersion correction applied to the energy and virial. We use a Lennard-Jones switch between 0.8 and 0.9 nm with a neighborlist at 1.0 nm and particle mesh Ewald electrostatics with a cutoff of 1.0 nm, PME order 6, and an Ewald cutoff tolerance of 10^{-6}. In all cases, a neighborlist update frequency of 10 steps was used with a time step of 2 fs. SETTLE[24,25] was used to constrain the water bonds and angles, and a total of 10 million steps (20 ns) were simulated, with the last 19 ns used for analysis. Temperature coupling algorithms were carried out with a coupling constant of τ_T = 1.0 ps for the NVT simulations and τ_T = 5.0 and τ_P = 5.0 for the NPT simulations. The low temperature is 298 and 301 K, with $\Delta T/T$ = 0.01 estimated from σ_E in the total energy from a single short simulation using the relationships for the ideal temperature gap.

For the NPT simulations, using a σ_V = 0.25 nm^3 at 1 bar from a short simulation predicts a ΔP of 238 bar using the formula presented here, but to err on the side of having sufficient samples we instead use ΔP = 175 with the low pressure at 1 bar and the high at 351 bar, though we are potentially losing some precision. In the case of Berendsen pressure control, we used ΔT = 1 K and ΔP = 30 bar to ensure overlap because of the narrowed distributions using Berendsen methods. For NVT, the interval between uncorrelated samples is determined from correlation times of the potential energy which is 2 ps for all methods except the Andersen method, where we use 4 ps. For NPT, we use the maximum of the uncorrelated sample intervals between the volume and the energy. Correlation times for MTTK are much smaller, around 0.3−0.4 ps for both energy and volume, whereas for Berendsen the energy and volume uncorrelated sample intervals are both 4

Table 8. Ensemble Validation of Pressure Control Algorithms in Water[a]

	enthalpy		volume		joint energy and volume			
barostat	ΔT	σ deviation	ΔP	σ deviation	ΔT	σ deviation	ΔP	σ deviation
Berendsen	1.03 ± 0.15	0.2	262 ± 25	9.0	1.67 ± 0.21	3.3	250 ± 30	7.4
Parrinello−Rahman	2.65 ± 0.21	1.7	309.3 ± 3.7	11.1	4.09 ± 0.34	3.2	354 ± 19	0.2
MTTK	2.978 ± 0.053	0.4	335.7 ± 3.9	3.7	3.026 ± 0.074	0.4	345.7 ± 4.6	0.6

[a]Tests of enthalpy distribution, volume distribution, and joint energy and volume distributions. For Parrinello−Rahman and MTTK, the true $\Delta T = 3$ and true $\Delta P = 350$, while for Berendsen, they are $\Delta T = 1$ and $\Delta P = 30$ in the joint energy and volume case. The Berendsen barostat performs significantly worse than the other two methods, requiring a much narrower range of variables to get any overlap. The other two barostats give statistically valid enthalpy distributions, with MTTK appearing to have fairly accurate volume distributions and with Parrinello−Rahman having somewhat worse volume behavior.

ps, and for Parrinello−Rahman, the energy and volume intervals are 6 and 0.4 ps, respectively. Thus, the NPT MTTK results are somewhat more precise.

We first examine the NVT results in Table 7. These results are completely in keeping with the argon results before, with all temperature control methods well within statistical error, with the exception of Berendsen, which is again wildly incorrect. These results demonstrate that the utility of ensemble validation is applicable to more typical molecular simulations, with data set sizes that are more typical for a standard testing pipeline.

From the NPT results in Table 8, we see that Parrinello−Rahman and MTTK have reasonable performance in describing the enthalpy distribution. Berendsen in this case is also reasonable, perhaps because the energy contribution dominates for the nearly incompressible water. MTTK has somewhat better results for volume fluctuations than Parrinello−Rahman. It is interesting to speculate on exactly the source of the difference between the volume fluctuation results in the argon and the water examples. In the argon example, both pressure control algorithms had small but statistically noticeable errors that were consistent between the two algorithms. In the water example, MTTK appears to be fairly ensemble consistent, whereas Parrinello−Rahman is slightly worse. Parrinello−Rahman with leapfrog is known to be inexact because the pressure lags by one time step, as the pressure and temperature are not both known at a given time t until after the next half step. This may be more of a problem in the case of water because with a higher compressibility, volume integration is a stiffer equation, requiring more exact solutions. We can tentatively conclude that typical aqueous simulations using MTTK may be more consistent with an NPT ensemble than Parrinello−Rahman, though both are far better than Berendsen temperature control. However, this test is with homogeneous fluids; it is possible that with an inhomogeneous system, such as a lipid bilayer in water, other artifacts might appear, but testing all possible systems and all possible methods is beyond the scope of this study.

4. TOOLS

To make these ensemble consistency checks easier, we have created a set of tools to assist other researchers to more easily measure the ensemble validations. This code is hosted by SimTK, at http://simtk.org/home/checkensemble and includes automatic plotting of linear and nonlinear graphs, as well as linear, nonlinear, and maximum likelihood parameter analysis for NVT, NPT, and μVT ensembles. It also provides validation of kinetic energies versus the Maxwell−Boltzmann distribution. These software tools were used for all analysis in this paper. These tools include example code for parsing Gromacs, CHARMM, Desmond, and flat text output files for ensemble consistency for NVT, NPT, and grand canonical simulations (for single component fluids), including testing enthalpy, volume, joint energy and volume fluctuations, instataneous Helmholtz energy, particle number, and joint energy and particle number. Scripts to regenerate all the analytic harmonic oscillator tests described in this paper are also included in the distribution.

5. CONCLUSIONS

We have shown that for molecular distributions characterized by Boltzmann distributions, which is true for all molecular simulations performed at NVT and NPT or in the grand canonical ensemble, we can easily check for consistency with the intended ensemble regardless of the details of the simulation. This test thus verifies a necessary condition all simulations must satisfy, regardless of the molecular details. We simply require pairs of simulations with differing external parameters such as temperature, pressure, or chemical potential. These paired simulations allow system-dependent properties such as densities of states to cancel out, resulting in a linear relationship between the distribution of extensive quantities such as energy, volume, enthalpy, and number of particles. Importantly, the constant of proportionality in this linear relationship is completely determined by the intensive variables that are set by the user.

Tests of simple model systems show that these relationships are not only qualitatively useful but also, with proper error analysis, can provide quantitative validation of the statistics of the distributions. We have demonstrated the utility of these relationships with simple analytical toy models of harmonic oscillators in both the NVT and NPT ensembles as well as with molecular simulations of argon and water. We see that these ensemble consistency relationships are able to identify thermostats and barostats that are inconsistent with the ensemble as well as identify differences in distributions caused by long time steps or abrupt cutoffs. All tested thermostats except the Berendsen thermostat give statistically good results. Barostats were somewhat more problematic, with MTTK giving the best results and Parrinello−Rahman being acceptable for many uses, while Berendsen pressure control is simply wrong for any calculation where volume fluctuations are important. In all cases, simpler checks such as making sure estimators of quantities like the temperature and pressure calculated from the kinetic energy and the virial do indeed have the correct value are useful as diagnosis tools and may occasionally identify problems that are not easily identified by the ensemble consistency methods tested here.

These relationships between pair distributions are true for all differences in applied external thermodynamic variables.

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

However, there are statistical reasons for choosing specific differences in the parameters. We have shown that for simple potentials both small and large differences in the applied system parameters lead to difficulty in distinguishing systems with errors from systems with the correct distributions. We have also shown that for typical probability distributions, choosing distributions whose means are separated by gaps 2 to 4 times the sum of the standard deviations appears to maximize the ability to discriminate between data that are or are not consistent with the desired ensemble, erring on the shorter side in cases where less data might be available. It is also important not to underestimate the autocorrelation time for the energy variables to be able to accurately use the error estimates, as it may give inaccurately high deviations from the correct distribution. Indeed, in typical simulation cases, the ability to properly estimate correlation times may be the largest source of uncertainty, as all other parts of the calculations are highly robust. We also emphasize that the size of the statistical deviation is a measure of how certain we are of the discrepancy, not necessarily the size of the discrepancy, as with sufficient data, we can statistically identify with a high certainty small deviations that generally do not affect simulation properties significantly. Finally, we note that these are very sensitive necessary tests, but they are not sufficient tests; they cannot guarantee that all states with the same energy are equally sampled, nor can they guarantee that all important regions of phase space are sampled.

We have also developed easy-to-use software tools to easily perform the statistical validation discussed here, requiring only lists of the relevant extensive variables and specification of the intensive applied variables. These tools can be easily incorporated into the workflow for molecular simulation testing, hopefully greatly reducing the difficulty of determining whether a given algorithm or software program is producing the desired thermodynamic ensemble. Future potential improvements of these tools include adapting the tools for grand canonical simulations and translating the relatively unsophisticated accounting of the number of standard deviation errors that are observed into full statistical hypothesis testing.

A. GRAND CANONICAL ENSEMBLE

Although no grand canonical simulations were carried out in this study, all the equations are essentially equivalent in the case of the isobaric−isothermal ensemble with $-\mu$ taking the place of P and N taking the place of V.

$$P(x, N|\beta, \mu) = \Xi(\beta, \mu)^{-1} \exp(-\beta E + \beta\mu N) \tag{26}$$

Examining the probability of N at fixed β and P performed at two different chemical potentials μ_1 and μ_2, we obtain

$$\begin{aligned} P(x, N|\beta, \mu_1) &= \Xi(\beta, \mu_1)^{-1} Q(\beta, N) \exp(\beta_1 \mu N) \\ P(x, N|\beta, \mu_2) &= \Xi(\beta, \mu_2)^{-1} Q(\beta, N) \exp(\beta_2 \mu N) \\ \frac{P(N|\beta, \mu_2)}{P(N|\beta, \mu_1)} &= \frac{\Xi(\beta, \mu_1)}{\Xi(\beta, \mu_2)} \exp([\beta\mu_1 - \beta\mu_2]N) \end{aligned} \tag{27}$$

$$\ln \frac{P(N|\beta, \mu_2)}{P(N|\beta, \mu_1)} = \beta(-[(PV)_2 - (PV)_1] + [\mu_2 - \mu_1]N) \tag{28}$$

We note that in the grand canonical case, N is already discrete, so a histogramming approach introduces no additional approximations as long as the histograms are fine grained down to integers. For samples sizes large enough that larger bins are required for accurate determination of probabilities, the maximum likelihood method will be more accurate.

We can also treat the joint probability distributions of N and E.

$$\begin{aligned} P(N, E|\beta_1, \mu_1) &= \Omega(N, E)\Xi(\beta_1, \mu_1)^{-1} \\ &\quad \exp(-\beta_1 E + \beta_1\mu_1 N) \\ P(N, E|\beta_2, \mu_2) &= \Omega(N, E)\Xi(\beta_2, \mu_2)^{-1} \\ &\quad \exp(-\beta_2 E + \beta_2\mu_2 N) \\ P(N, E|\beta_2, \mu_2)P(N, E|\beta_1, \mu_1) &= \frac{\Xi(\beta_1, \mu_1)}{\Xi(\beta_2, \mu_2)} \exp(-[\beta_2 - \beta_1]E \\ &\quad + [\beta_2\mu_2 - \beta_1\mu_1]N) \end{aligned} \tag{29}$$

$$\begin{aligned} &\ln \frac{P(N, E|\beta_2, \mu_2)}{P(N, E|\beta_1, \mu_1)} \\ &= -(\beta_2(PV)_2 - \beta_1(PV)_1) - [\beta_2 - \beta_1]E \\ &\quad + [\beta_2\mu_2 - \beta_1\mu_1]N \end{aligned} \tag{30}$$

This approach can easily be generalized to multiple chemical species, especially when using maximum likelihood methods to allow minimization of the resulting multidimensional probability ratios. For example, for an arbitrary number of species $\vec{N}$ with associated chemical potentials $\vec{\mu}$, we have

$$\begin{aligned} P(\vec{N}, E|\beta_1) &= \Omega(E, \vec{N})\Xi_1(\beta_1, \vec{\mu}_1)^{-1} \\ &\quad \exp(-\beta_1 E + \beta_1\vec{\mu}_1\cdot\vec{N}) \\ P(\vec{N}, E|\beta_2) &= \Omega(E, \vec{N})\Xi_2(\beta_2, \vec{\mu}_2)^{-1} \\ &\quad \exp(-\beta_2 E + \beta_2\vec{\mu}_2\cdot\vec{N}) \\ P(\vec{N}, E|\beta_2, \mu_2)P(\vec{N}, E|\beta_1, \mu_1) &= \frac{\Xi_1(\beta_1, \vec{\mu}_1)}{\Xi_2(\beta_2, \vec{\mu}_2)} \exp(-[\beta_2 - \beta_1]E \\ &\quad + [\beta_2\vec{\mu}_2 - \beta_1\vec{\mu}_1]\cdot\vec{N}) \end{aligned} \tag{31}$$

$$\begin{aligned} &\ln \frac{P(\vec{N}, E|\beta_2, \mu_2)}{P(\vec{N}, E|\beta_1, \mu_1)} \\ &= -(\beta_2(PV)_2 - \beta_1(PV)_1) - [\beta_2 - \beta_1]E \\ &\quad + [\beta_2\vec{\mu}_2 - \beta_1\vec{\mu}_1]\cdot\vec{N} \end{aligned} \tag{32}$$

B. WEIGHTED LEAST-SQUARES FITTING TO HISTOGRAM RATIOS

Assume we are collecting data from a continuous, one-dimensional probability distribution in a histogram H with $k = 1...K$ bins. We have N total samples, with $\{n_1, n_2, ..., n_k\}$ observations in each bin, so that $\sum_{k=1}^{K} n_k = N$. The empirical probability of finding an observation in bin k is simply $p_k = n_k/N$. Repeating this experiment will lead to slightly different results for the p_k. The standard estimator of variance of p_k due to this sampling variance is a standard result for the binomial distribution and is equal to $p_k(1 - p_k)/N$.

Given two histograms H_1 and H_2 that have aligned bins with N_1 and N_2 samples each, the ratio of the probabilities of H_2

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

over H_1 will be $r_k = p_{k,1}/p_{k,2}$ for each bin, where $p_{k,1}$ and $p_{k,2}$ are the probabilities in the kth bin for the first and second simulation in the pair. The data in the two histograms are collected independently, so the statistical variance in the logarithm of the ratio $\ln r_k = \ln(p_{k,2}/p_{k,1})$ will be to first order:

$$\begin{aligned}\mathrm{var}(\ln r_k) &= \frac{\mathrm{var}(p_{k,1})}{p_{k,1}^2} + \frac{\mathrm{var}(p_{k,2})}{p_{k,2}^2} \\ &= \frac{1 - p_{k,1}}{N_1 p_{k,1}} + \frac{1 - p_{k,2}}{N_2 p_{k,2}} \\ &= \frac{1}{n_{k,1}} - \frac{1}{N_1} + \frac{1}{n_{k,2}} - \frac{1}{N_2}\end{aligned} \tag{33}$$

The variance in the ratio of the histograms themselves, useful for computing nonlinear estimates of the error, will be

$$\begin{aligned}\frac{\mathrm{var}(r_k)}{r_k^2} &= \frac{\mathrm{var}(p_{k,1})}{p_{k,1}^2} + \frac{\mathrm{var}(p_{k,2})}{p_{k,2}^2} \\ \mathrm{var}(r_k) &= \left(\frac{p_{k,2}}{p_{k,1}}\right)^2 \frac{\mathrm{var}(p_{k,1})}{p_{k,1}^2} + \frac{\mathrm{var}(p_{k,2})}{p_{k,2}^2} \\ \mathrm{var}(r_k) &= \left(\frac{n_{k,2}N_1}{n_{k,1}N^2}\right)^2 \left(\frac{1}{n_{k,1}} - \frac{1}{N_1} + \frac{1}{n_{k,2}} - \frac{1}{N_2}\right)\end{aligned} \tag{34}$$

Define a diagonal weight matrix W, with one over the variance in the ith measurement along the diagonal. If we have a multivariate function F linearly dependent on data vector X as $F = AY$, with A a constant matrix, then the covariance matrix of uncertainties $\mathrm{cov}(F)$ will be equal to $A\,\mathrm{cov}(Y)A^T$. In the case of weighted linear least-squares, $\mathrm{cov}(Y)$ is the matrix of weights W, where $W_{ii} = \sigma^{-2}$, the variance of each histogram ratio point. If α is the vector of parameters and X is the $(M + 1) \times N$ matrix of observables, with the first column all ones and the second through the $(M + 1)$th column, the values of the observations of the M observables, then we will have for α

$$\alpha = (X^T W X)^{-1} X^T W Y$$

Plugging this into the equation for $\mathrm{cov}(\alpha)$ in terms of $\mathrm{cov}(Y)$, some linear algebra leads to a covariance matrix of the parameters $\vec{\alpha}$ of $(X^T W X)^{-1}$. If we have instead a nonlinear least-squares problem, at the minimum, we obtain a similar covariance matrix, except that we replace X with the linear approxiation to the nonlinear system of equations, the Jacobian matrix J defined by $J_{ij} = \partial f(y_i, \vec{\alpha})/\partial \alpha_j$, where f is the nonlinear model, and y_i is the ith data point. This leads to a final equation for the covariance of the parameters:

$$\mathrm{cov}(\alpha) = (J^T W J)^{-1}$$

C. MAXIMUM LIKELIHOOD ESTIMATION AND ANALYTICAL ERROR ESTIMATES

For a general Boltzmann-type probability distribution, the ratio of probabilities must satisfy

$$\frac{P_2(\vec{X})}{P_1(\vec{X})} = \exp(-\vec{\alpha}\cdot\vec{X}) \tag{35}$$

where the $\vec{X}$ are the M sample variables (such as E or V), the $M + 1$ α_j variables are the corresponding conjugate variables specified by the simulation ensemble, and $\vec{\alpha}\cdot\vec{X}_i$ is shorthand for $\alpha_0 + \sum_{j=1}^{M}\alpha_j X_j$ rather than the standard dot product.

We develop the solution by finding maximum likelihood parameters along the lines of the solution presented in ref 5. The ratio in eq 35 can be interpreted as $P(\vec{X}|1)/P(\vec{X}|2)$ where $P(\vec{X}|i)$ is the conditional probability that an observation is from the ith simulation given only the information $\vec{X}$. We would like to compute the likelihood of a given set of α parameters given sets of measurements with the specific simulation i each set comes from known.

Using the rules of conditional probabilities, and the fact that $P(\vec{X}|1) + P(\vec{X}|2) = 1$, we rewrite this probability distribution as follows:

$$\begin{aligned}\frac{P(\vec{X}|2)}{P(\vec{X}|1)} &= \frac{\frac{P(2\,|\,\vec{X})P(\vec{X})}{P(2)}}{\frac{P(1\,|\,\vec{X})P(\vec{X})}{P(1)}} \\ &= \frac{P(2|\vec{X})P(1)}{P(1|\vec{X})P(2)} \\ &= \frac{P(2|\vec{X})}{1 - P(2|\vec{X})}\frac{P(1)}{P(2)}\end{aligned} \tag{36}$$

We note that $P(1)/P(2) = N_1/N_2$, where N_2 and N_1 are the number of samples from the two simulations, respectively. Although either $P(\vec{X}|1)$ or $P(\vec{X}|2)$ can be eliminated, we are left with one independent continuous free energy distribution. Writing either $P(\vec{X}|2)$ or $P(\vec{X}|1)$ in a closed form is system dependent; specifically, it depends on the unknown density of states. We define the constant $M = \ln(N_1/N_2)$ and rewrite eq 35 as

$$\frac{P(2|\vec{X})}{1 - P(2|\vec{X})} = \exp(-M - \alpha_0 - \sum_{j=1}^{M}\alpha_j X_j) \tag{37}$$

Given eq 37, we can rewrite the probability of a single measurement $P(1|\vec{X}_i)$ or $P(2|\vec{X}_i)$ as

$$\begin{aligned}P(1|\vec{X}_i) &= \frac{1}{1 + \exp(M + \vec{\alpha}\cdot\vec{X}_i)} \\ P(2|\vec{X}_i) &= \frac{1}{1 + \exp(-M - \vec{\alpha}\cdot\vec{X}_i)}\end{aligned} \tag{38}$$

The total likelihood of any given observation X_i is the product of all the individual likelihoods, giving

$$\begin{aligned}\ln L(\vec{\alpha}|\mathrm{data}) = &\sum_{i=1}^{N_1} \ln f(-M - \vec{\alpha}\cdot\vec{X}_i) \\ &+ \sum_{i=1}^{N_2} \ln f(M + \vec{\alpha}\cdot\vec{X}_i)\end{aligned}$$

where $f(x) = [1 + \exp(x)]^{-1}$ is the Fermi function. This likelihood equation can be minimized directly or by finding the gradient with respect to the α parameters and solving for $\nabla(\ln L) = 0$ to give the maximum likelihood result. The log likelihood function has a single minimum, and thus there will be only a single root to $\nabla(\ln L) = 0$.

The covariance matrix of each α_j can be written in terms of the Fisher information:

dx.doi.org/10.1021/ct300688p | J. Chem. Theory Comput. 2013, 9, 909−926

$$\mathrm{var}(\alpha_j) = I(\alpha_j)^{-1} = -\left(\frac{\partial^2 \ln L(\alpha)}{\partial \alpha_j^2}\right)^{-1} \quad (39)$$

Note that in ref 5 an additional factor dependent on the number of samples was required to get the correct uncertainty estimates. In that case, we assumed that the simulation was conducted properly, so that β was known and thus had an additional constraint, leaving only a single parameter estimated from the ratio of two distributions, which ends up reducing the uncertainty by this constant factor.[26] In this case, we are solving for two parameters using the data from two distributions, and no implicit constraints are applied. Thus the correction is not required.

AUTHOR INFORMATION

Corresponding Author

*E-mail: michael.shirts@virginia.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author wishes to thank Ed Maginn (Notre Dame University) and Lev Gelb (UT-Dallas) for comments and suggestions; Joe Basconi (University of Virginia), Daniel Sindhikara (Institute for Molecular Sciences, Japan), David Mobley (UC-Irvine), John Chodera (Memorial Sloan-Kettering), and Lev Gelb (UT-Dallas) for careful reading of the manuscript; and SimTK.org for hosting the checkensemble code.

REFERENCES

(1) Harvey, S. C.; Tan, R. K. Z.; Cheatham, T. E. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.* **1998**, *19*, 726−740.

(2) Hoover, W. G. Canonical dyanmics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(3) Rhee, Y. M.; Pande, V. S. Multiplexed Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* **2003**, *84*, 775−786.

(4) Sindhikara, D.; Meng, Y.; Roitberg, A. E. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.* **2008**, *128*, 024103.

(5) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys. Rev. Lett.* **2003**, *91*, 140601.

(6) Paliwal, H.; Shirts, M. R. A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *J. Chem. Theory Comput.* **2011**, *7*, 4115−4134.

(7) Geman, S.; Geman, D. Stochastic Relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal.* **1984**, *6*, 721−741.

(8) Anderson, T. W.; Darling, D. A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Ann. Math. Stat.* **1952**, *23*, 193−212.

(9) Rowley, L. A.; Nicholson, D.; Parsonage, N. Grand Ensemble Monte-Carlo Studies of Physical Adsorption. 1. Results for multilayer adsorption of 12−6 argon in field of a plane homogeneous solid. *Mol. Phys.* **1976**, *31*, 365−387.

(10) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(11) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384−2393.

(12) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3584−3690.

(13) Technically, Nosé−Hoover does not give a completely ergodic trajectory, as there are low energy states that are inaccessible.[27,28] However, for large chaotic systems, the contribution from these states appears to be negligible.

(14) Morishita, T. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.* **2000**, *113*, 2976−2982.

(15) S., L. A.; Balabaev, N. K. On the Berendsen thermostat. *Mol. Simul.* **2002**, *13*, 177−187.

(16) Golo, V. L.; Shaitan, K. V. Dynamic attractor for Berendsen's thermostat and the slow dynamics. *Biofizika* **2002**, *47*, 611−617.

(17) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(18) Pastor, R. W.; Brooks, B. R.; Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* **1988**, *65*, 1409−1419.

(19) Cuendet, M. A.; van Gunsteren, W. F. On the calculation of velocity-dependent properties in molecular dynamics simulations using the leapfrog integration algorithm. *J. Chem. Phys.* **2007**, *127*, 184102.

(20) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182−7190.

(21) Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.* **1983**, *50*, 1055−1076.

(22) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal-isobaric ensemble. *J. Phys. A* **2006**, *39*, 5629−5651.

(23) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **1996**, *87*, 1117−1157.

(24) Andersen, H. C. Rattle: A "Velocity" version of the Shake Algorithm for Molecular Dynamics Calculations. *J. Comput. Phys.* **1983**, *52*, 24−34.

(25) Miyamoto, S.; Kollman, P. A. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.

(26) Anderson, J. A. Separate Sample Logistic Discrimination. *Biometrika* **1972**, *59*, 19−35.

(27) Legoll, F.; Luskin, M.; Moeckel, R. Non-ergodicity of the Nose-Hoover thermostatted harmonic oscillator. *Arch. Rational Mech. Anal.* **2007**, *184*, 449−463.

(28) Legoll, F.; Luskin, M.; Moeckel, R. Non-ergodicity of Nose-Hoover dynamics. *Nonlinearity* **2009**, *22*, 1673−1694.