JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods

Himanshu Paliwal and Michael R. Shirts*

Department of Chemical Engineering, University of Virginia, Charlottesville, Virginia 22904-4741, United States

Supporting Information

ABSTRACT: There is a significant need for improved tools to validate thermophysical quantities computed via molecular simulation. In this paper we present the initial version of a benchmark set of testing methods for calculating free energies of molecular transformation in solution. This set is based on molecular changes common to many molecular design problems, such as insertion and deletion of atomic sites and changing atomic partial charges. We use this benchmark set to compare the statistical efficiency, reliability, and quality of uncertainty estimates for a number of published free energy methods, including thermodynamic integration, free energy perturbation, the Bennett acceptance ratio (BAR) and its multistate equivalent MBAR. We identify MBAR as the consistently best performing method, though other methods are frequently comparable in reliability and accuracy in many cases. We demonstrate that assumptions of Gaussian distributed errors in free energies are usually valid for most methods studied. We demonstrate that bootstrap error estimation is a robust and useful technique for estimating statistical variance for all free energy methods studied. This benchmark set is provided in a number of different file formats with the hope of becoming a useful and general tool for method comparisons.

1. INTRODUCTION

Simulation and theory communities have developed substantial interest in using free energy calculations for molecular design problems. Specifically, free energy calculations can guide experimental screening techniques for measuring biological interaction energies and offer the potential of a faster and cheaper way to get thermodynamic information over large chemical spaces in a variety of molecular contexts.[1] For example, drug design[2] requires prediction of binding affinities, tautomers, protonation states, membrane permeabilities, and solubilities, all of which involve free energy calculations.[3−6] Similarly, free energy calculations could become useful tools in material design problems ranging from improved protein selectivity and stability on chromatographic surfaces[7] to tailoring metal organic frameworks[8] for applications, such as gas storage and separation. Such potential uses extend to the design of new nanomaterials, such as therapeutic dendrimers, heteropolymers, and hyperbranched polymers for molecular recognition, imaging, sensing/signaling, and controlled payload delivery.[9−13] However, substantial roadblocks to routine use of molecular simulations as a complement to experiment include confusion over suitability of methods for different molecular problems and the lack of rigorous, validated understanding about the reliability of free energy calculations and other observables estimated using statistical methods.

Other computational fields have successfully benchmarked and tested computational methods to improve the reliability and thus the utility of simulations. The field of computational fluid dynamics has also grappled with issues of reliability and standardization. During the late 90s, substantial research efforts in the field of computational fluid dynamics (CFD) were focused on establishing validation benchmarks to improve the reliability of CFD simulations in various design applications.[14−16] This research helped to bring down costs, increase data fidelity, and reduce design cycle time in the early development phases of new airplanes,[17] Formula 1 cars,[18] treatment and diagnosis of cardiovascular diseases,[19,20] and off-shore oil rigs.[21] Similar validation benchmarks were developed for simulations in the nuclear industry, improving the reliability in nuclear reactor safety, underground nuclear waste storage, and nuclear weapon safety.[22] Quantum chemists maintain validation databases for comparison between experimental methods and different QM methods.[23] For molecular simulation to play a similar role in molecular engineering design,[24] benchmarks and validation sets must be established. In this paper we aim to provide one set of tools for improved standardization of molecular simulations through the first version of a benchmark set for free energy calculations for molecular transformations.

There are a large number of free energy methods available,[25−31] which by early 2011 have been cited collectively over 4600 times, with 20% of those citations in the last 18 months.[32] However, the simulation field lacks consensus in choosing a method most appropriate for a given molecular design situation. At least three fundamental issues contribute to this confusion.

First, there is a lack of standard test cases for rigorous comparisons between different free energy methods. Studies of new methods and method comparisons frequently use relatively simple model systems, such as a one- or two-dimensional analytically solvable potential energy function,[28,33] solvation of Lennard-Jones sphere,[28,34] alchemical changes between small molecules,[31,35] or simplified solvation models.[36] These test cases may not capture all of the issues encountered in actual molecular changes. Alternatively, papers comparing methods may use complicated

Received: June 11, 2011
Published: October 18, 2011

biophysical systems, such as protein−ligand binding or pK_a determination that are hard to converge, and therefore make it difficult to accurately gauge true gains in efficiency.[37−40] Both of these extremes put limits on our ability to decide whether a given method will be useful in actual molecular design scenarios.

Second, computing thermophysical properties by molecular simulation involves stochastic sampling of molecular configurations, and all comparisons must deal with the fact that repeated independent measurements have associated statistical error; unlike in quantum mechanics, comparisons must be done on a statistical basis, and it will never be possible to converge most calculations to arbitrary levels of precision in reasonable computational time.[41]

Finally, direct comparisons between methods can be difficult because of the differences between simulation code bases. Free energy calculation capabilities are recent additions to most large-scale molecular simulation codes, and most codes support usually only a small subset of available free energy methods.

As a step toward helping solve these problems, we propose the first version of a molecular test set comprising realistic systems undergoing challenging molecular transformations. We then use this test set to test the efficiency and reliability of different methods for estimating free energy differences of molecular transformations from simulation. Although the molecular design applications listed in the introduction seem very different, there are features common to all free energy calculations required for these applications. All involve determining the preference of a molecule to partition between two environments and can be calculated by way of a difference in the free energies of molecular transformation between these two environments. For example, we might wish to design a solute preferentially solvated by a protein when compared to solvent (pure water) or a different complex medium (another protein), as in the case of drug design. Alternately we might design a solvent which preferentially solvates a given solute in a mixture; for example, designing ionic liquids[42] for sequestering CO_2. These molecular transformations primarily involve either growing or deleting atoms, changing the size or dispersion interaction between atoms, or altering partial charge on mutation sites. Any benchmark test set must include examples of these transformations which are simultaneously challenging enough to push new methods and yet possible to evaluate with sufficiently high precision that we can reach meaningful comparisons about different methods in a reasonable amount of computer time.

The most important features of any property estimation method to understand are the statistical errors inherent in the method, including both statistical bias and statistical uncertainty, and the reliability of the method's estimate of the property of interest. Without knowledge of such features of the methods, we cannot sufficiently trust our calculations or compare two different calculations for validation purposes.

Studies commissioned by US science funding agencies on future directions for simulation based engineering and science have emphasized the fundamental need for improved uncertainty verification and validation.[43,44] Almost all estimators of statistical quantities, like methods for calculating free energies and ensemble averages have some bias, a systematic deviation from the true answer that would be obtained with perfect sampling. Additionally, computing a given observable from independently selected samples gives different estimates of the observable; this variation is the statistical uncertainty of the estimate. Most free energy methods also include estimates of this statistical uncertainty. However, these uncertainty estimates are themselves statistical quantities, with variation from sample to sample, and must be validated.

A few valuable studies have compared[28,31,33,45−49] multiple free energy methods but not necessarily in a systematic way. We use our proposed benchmark set to directly compare the estimated uncertainties with the sample uncertainty. We also compute the change in the mean square error as a function of the number of intermediate states and the number of samples to capture both bias and uncertainty. We also test whether the distribution of free energy estimators is indeed Gaussian, a condition usually assumed when using statistical uncertainty estimates to calculate error. Finally, we also evaluate the bootstrap method[50] as a tool for estimating statistical uncertainty, as this method can be easily implemented for all of free energy algorithms described in this paper, and indeed generally for most statistical estimates of observables.

In this paper, we first explain our proposed benchmark test set for molecular transformations and the rationale behind the molecular choices. Next, we use this set to test and compare the accuracy, precision, and reliability of 10 free energy methods. We then present a summary of the comparison of the methods, with much of the data presented as Supporting Information because of its length, and finally present our recommendations for methods for performing free energy calculations.

2. TEST SET

The systems in this benchmark set are designed to represent "alchemical" changes, or changes of molecular identity, common to most molecular design applications. Alchemical transformations frequently require the deletion or introduction of atoms and large changes in the partial charges. Changes in torsional, angle, or dihedral parameters usually result in smaller changes in phase space, as do small changes in dispersion strength atomic radius, or charge. We therefore focus on atomic introduction/deletion and large changes in partial charge.

2.1. Minimal Test System: OPLS[51] UA Methane in TIP3P Water (MS). The solvation of a Lennard-Jones (LJ) sphere, representing methane, is perhaps the simplest free energy test case that can be truly defined as molecularly realistic. There are no bonds, angles, or torsions terms nor are there solute/solvent charge terms. This system represents a minimal test of whether the free energy method is at all valid or applicable for molecular systems. We examine the transformation of coupling the sphere into water, which corresponds to the solvation free energy of this molecule.

2.2. Charge Mutation System: Dipole Inversion in TIP3P Water (DI). We use an OPLS-UA ethane molecule with the addition of +1/−1 charges on the two atomic centers. This setup avoids computing free energies of ions directly, as changing the total charge of a system with periodic boundary conditions is not always handled exactly in many codes, and requires numerous complicated corrections.[52] This test measures whether a method can handle large water rearrangements around charges and the large energy differences involved in changing large partial charges. The system is a null transform; the free energy change is zero as the final state is identical to the initial state by symmetry.

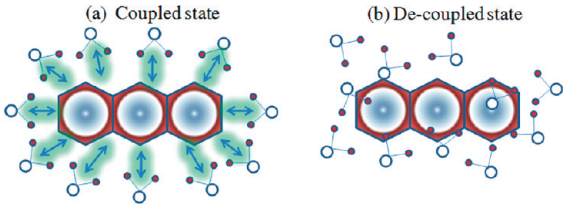

Figure 1. (a) In the coupled state or solvated state both intermolecular and intramolecular interactions for anthracene are turned on. (b) In the decoupled state or vacuum state the intermolecular interactions with water molecules are turned off.

2.3. Large Molecule Mutation System: Absolute Hydration Free Energy of UA Anthracene in TIP3P Water. In our third test, we compute the solvation of anthracene via the decoupling of the intermolecular interactions from water. This system tests whether the method can handle introduction or deletion of multiple atomic sites efficiently. Importantly, there are no internal ligand degrees of freedom to complicate the analysis. Force field parameters are taken from Pitera and Van Gunsteren.[53] Originally, we chose a null transformation of anthracene to anthracene via a benzene intermediate, but the simpler solvation problem was eventually chosen because of the difficulty of supporting such multiphase transformations in other codes and the difficulty of interpreting the statistics of multiple transformations; a key requirement of the benchmark set is simplicity of use. For this test set, we have used a decoupling scheme, turning off the interactions between the solute and the solvent but keeping the intramolecular interactions in the solute turned on as shown in Figure 1. The free energy change of this transformation corresponds to the desolvation free energy of anthracene, but we report the solvation free energy for ease of interpretation.

We note that although these are simplified systems, the dipole inversion and anthracene simulations are by no means toy models. The dipole moment of the dipole inversion test case is 7.20 Debye, significantly greater than the dipole moment of most small molecules, and includes a +1 to −1 charge difference on each atom. The anthracene solvation test set involves the disappearance of three aromatic rings with a total of 14 heavy atoms, which is on the high end of most molecular transformations. There are of course a number of other possible test molecules that could be examined; we will reserve discussion of future extensions of this test set for the discussion.

3. FREE ENERGY METHODS AND ERROR PROPAGATION

Using this benchmark set, we evaluated a total of 10 free energy methods, chosen specifically because all can be computed from the same set of simulation samples. In the following presentations, we assume that the simulations are performed in the isothermal−isobaric ensemble, and thus the Gibbs free energy ΔG is the quantity of interest; the Helmholtz free energy can also be computed if the simulations are performed in the canonical ensemble. U indicates the generalized potential energy, which in the case of the isobaric−isothermal ensemble is actually $U + PV$, β is $1/k_BT$, and λ is a coupling parameter connecting the initial and final states in a user-chosen manner. Brackets indicate an ensemble average over the appropriate ensemble. These methods were chosen to represent a diversity of the most commonly used methods; several of them have alternate variants, and we have not tried to capture all possible variants. The purpose of this paper is to propose a molecular transformation test set and demonstrate its utility in method comparison, not necessarily a test of all possible free energy methods. All of the 10 methods were chosen because they can be computed from the same equilibrium samples, rather than computed independently. We have chosen not to examine expanded ensemble methods[54] or nonequilibrium methods, such as Wang−Landau[55] and Jarzynski's[56] relationship-based methods, they would require independent simulations for each method as well as each having additional parameters that would have to be chosen for optimality. However, by distributing the starting configurations and parameters in a number of formats, performing comparisons between other methods and the methods presented in this paper will hopefully become significantly easier for other research groups.

3.1. Thermodynamic Integration[24] Using a Trapezoid Rule (TI) and a Cubic Spline[26] (TI3). For TI, we compute the ensemble average of the derivative of potential energy function with respect to a coupling parameter λ for a system, i.e., $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ at all λ values and the corresponding variances $\sigma^2{}_i$ of the $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ distributions:

$$\sigma_i^2 = \langle\langle x^2\rangle - \langle x\rangle^2\rangle, \quad \text{where } x = \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_i}\rangle \tag{1}$$

The $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ values at different intermediates are interpolated and then integrated to get an overall free energy change:

$$\Delta G_{10} = G(\lambda = 1) - G(\lambda = 0) = \int_{\lambda=0}^{\lambda=1} \langle(\partial U(\lambda)/\partial\lambda)_{\lambda}\rangle \mathrm{d}\lambda \tag{2}$$

For TI we have used a linear interpolation, which leads to the standard trapezoid rule to integrate the total free energy. For TI3, the $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ vs λ_i curve is fit piecewise to a natural cubic spline and then integrated analytically using the coefficients of the cubic equation (see Appendix for the derivation). Both the trapezoidal and the cubic spline integration can be expressed in the form of weighted sum of individual $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$:

$$\Delta G_{10} = \sum_{i=1}^{K} W_i \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_i}\rangle \tag{3}$$

Here the W_i's are the respective weights corresponding to each state and K is the total number of intermediate states. The variance $\sigma^2{}_{10}$ of this estimate of free energy can be calculated by the following variance propagation formula:

$$\sigma_{10}^2 = \sum_{i=1}^{K} W_i^2 \sigma_i^2 \tag{4}$$

Occasionally, some researchers have computed the variance of the free energy over each interval, i to i+1 individually, and then propagated these results into the total variance. This is incorrect, since the variance of each interval is correlated to the variance of the neighboring intervals. For example, the free energy difference between states 1 and 2 and between states 2 and 3 both contain statistical information from state 2. It is important to propagate the uncertainty directly using eq 4 to avoid potential errors.

A number of alternative TI schemes have been proposed.[26,57,58] However, many of these schemes require some knowledge of the magnitude of the statistical uncertainty for optimality. Other schemes use nonlinear fits to two different functional forms separately

describing LJ and Coulomb contributions to the free energy.[47] Such schemes are not particularly flexible and introduce integration bias that is difficult to quantify. By using cubic splines, we can obtain a higher order formula independent of functional form of $dU/d\lambda$, while propagating error using the same formalism as is used in standard TI (eq 4). For specific applications, different TI weighting schemes may be more appropriate to the particular curvature encountered, and these application-specific methods might easily be better for their intended application than a general-purpose spline, such as the one evaluated here. Because of this large range of slight variants of the weights, we will not attempt to classify all possible TI methods in this study but include one higher order algorithm to examine some method beyond the simplest trapezoidal case.

3.2. Exponential Averaging (EXP) in Two Forms: DEXP and IEXP[27,28]. In exponential averaging schemes, the free energy change ΔG_{ij} is calculated using the exponential average of the difference of the potential energies ΔU_{ij} between two states i and j over one of the ensembles. The free energy difference as a function of potential energy difference ΔU_{ij} and N samples is then

$$\Delta G_{ij} = -\frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{n=1}^{N}\exp[-\beta\Delta U(x_n)_{ij}]\right) \quad (5)$$

This averaging is performed using samples from state i to compute potential energy differences ΔU_{ij} from state i to state j. The free energy of the reverse process can be computed using samples from state j and computing potential energy differences to state i. Since the labels themselves are arbitrary, to remove ambiguity in the direction we will describe such computations as being either "deletion" or "insertion". We will call ΔU_{ij} taken in the direction of decreasing entropy as an "insertion" step and ΔU_{ij} taken in the direction of increasing entropy as a "deletion" step, as inspired by Wu et al.[59] Hence the free energy method with ΔU_{ij} stepping in the direction of increasing entropy in eq 5 is labeled as deletion exponential averaging (DEXP), and the free energy method with steps of ΔU_{ij} in the direction of decreasing entropy in eq 5 is labeled as insertion exponential averaging (IEXP). In both cases, the variance σ^2_{ij} between two adjacent intermediate states can be estimated using standard point estimation theory as

$$\sigma^2_{ij} = \frac{1}{N}\left(\frac{\sigma_x}{\langle x\rangle}\right)^2, \quad x = \exp[-\beta\Delta U(x_n)_{ij}] \quad (6)$$

In both the exponential averaging methods the overall free energy change ΔG_{10} is the sum of intermediate free energy changes ΔG_{ij}, and so the variance $\sigma^2{}_{10}$ is simply the sum of the associated variances $\sigma^2{}_{ij}$. In some cases, the changes from the i state to the $i-1$ state and $i+1$ states might both be deletion or insertion cases; in this case, all the sampling performed at i and the two estimates of the free energy difference will not be statistically independent. Complicated molecular changes will frequently involve both addition and subtractions of accessible phase space and thus will fall somewhere in between these two general schemes.

Any free energy change involving inherent directionality, such as exponential averaging, requires careful definitions to ensure that the direction of entropy change remains constant throughout the process. Otherwise, we cannot interpret the entire transformation as a deletion or an insertion process. Methane and anthracene solvations involve moving molecules from vapor to liquid phase, resulting in a decrease in total entropy. However, in the dipole inversion case, we have a symmetric transformation, and thus deletion and insertion happen within a single process. In dipole inversion, going from a very large magnitude dipole to a small apolar intermediate, we have an increase of entropy as the water around the particle becomes less structured, and thus we use the term deletion. From the intermediate uncharged intermediate state to the reversed −/+ dipolar state during the second half of the inversion, we have the reverse process, and we use the term insertion consistent with the entropy direction definition. To use the terminology IEXP, Gaussian estimate with insertion (GINS), DEXP, and Gaussian estimate with deletion (GDEL) pathways, we therefore need to combine mixed halves of what would typically be called the forward and reverse pathways as illustrated in Figure 2. Although these particular sums are nonzero, they provide a consistent definition of the statistical variance of insertion and deletion. The statistical variance for symmetric transformations will simply be the average of the variance for the deletion and insertion processes.

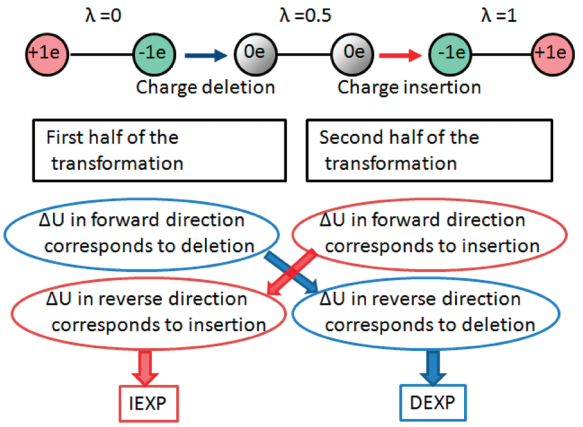

Figure 2. Free energy differences of transitions in the direction of increasing and decreasing entropy should be added separately to get the overall free energy for a dipole inversion.

3.4. Bennett Acceptance Ratio (BAR)[29]. The Bennett acceptance ratio uses samples of the potential energy in both i to j and j to i directions to obtain a provably minimum variance estimate of the free energy difference. Calculation of the free energy change between any two intermediate states through BAR requires self-consistent solution of the two equations:

$$\Delta G_{ij} = \frac{1}{\beta}\ln\left(\frac{\sum_{k=1}^{N_j}\frac{1}{1+\exp[-\beta(\Delta U^j_k - C)]}}{\sum_{l=1}^{N_i}\frac{1}{1+\exp[-\beta(\Delta U^i_l - C)]}}\right) + C - \frac{1}{\beta}\ln\left(\frac{N_j}{N_i}\right) \quad (7)$$

$$C = \Delta G_{ij} + \frac{1}{\beta}\ln\left(\frac{N_j}{N_i}\right) \quad (8)$$

The first equation is true for any constant C, but when eqs 7 and 8 are solved self-consistently the ΔG_{ij} will have minimized variance. There exists a large number of ways to solve the equations self-consistently, and a complete discussion of the

best methods is beyond the scope of this paper. The variance σ^2_{ij} in ΔG_{ij} for any C can be estimated as

$$\sigma_{ij}^2 = \frac{1}{\beta^2 N_i}\left[\frac{\langle f^2(x)\rangle_i}{\langle f(x)\rangle_i^2} - 1\right] + \frac{1}{\beta^2 N_j}\left[\frac{\langle f^2(x)\rangle_j}{\langle f(x)\rangle_j^2} - 1\right] \quad (9)$$

where $f(x)$ is the Fermi function $1/(1+x)$ and $x = \exp[\beta(\Delta U - C)]$. The total free energy change is the sum over changes between consecutive intermediate states. Typically, the variance in the full free energy is computed by assuming independent error and summing the variance for consecutive intermediate states. However, the assumption that the errors add independently is not correct, since the free energy difference from $i-1$ to i and from i to $i+1$ states both depend on the potential energy at i, so their variances are not independent. There is thus no general formula to obtain a statistically unbiased estimate of an entire transformation computed by a series of BAR calculations between neighboring states.

3.5. Unoptimized Bennett Acceptance Ratio (UBAR). Equation 9 in Section 3.4 is valid for any initial estimate of the free energy, though choices of C not given by the implicit equation (eq 9) will not have minimum variance. If we make the choice of $C = \beta^{-1}\ln(N_j/N_i)$, we no longer need to self-consistently solve equations. This can avoid saving, reading, and reprocessing all of the data, potentially saving significant disk space or memory, at a cost of decreased statistical efficiency and increased bias. We can instead accumulate the averages in eqs 7−9 as the simulation progresses. If each intermediate free energy is relatively near zero, then this free energy estimate will be close to optimal. This estimator is directly equivalent to the minimum variance version of transition state Monte Carlo, where Barker acceptance probability[60] is used.[61]

3.6. Range-Based Bennett Acceptance Ratio (RBAR). If we keep track of the ensemble averages in eqs 11 and 12 for a range of trial values of C, we will obtain a number of estimates of the best estimate free energy.[61] Of these free energy estimates, the one that corresponds most closely to the input value of the free energy in the formula for C in eq 8 will be the least biased and will have minimum variance. By choosing this particular value of the free energy from the range of values, we are essentially precalculating the self-consistent solution. To apply this method, a range of starting values of C is chosen. This trial C is fed as an initial guess, and C is calculated using a single iteration of eq 11, with corresponding ΔG and σ then calculated. Accumulated averages are maintained for each choice of C. A decent estimate of the range of C (and therefore ΔG) is therefore a requirement for using this method. In some cases, it may end up being more costly than BAR, as accumulated averages must be maintained for a certain number of trial free energy values, instead of simply performing 5−10 self-consistent iterations. However, the advantage of what we will call in this paper RBAR is that data from each simulation step does not need to be retained for postprocessing, as is required with BAR, and only the accumulated averages need to be retained.

3.7. Multistate Bennett Acceptance Ratio (MBAR)[30]. MBAR is a method to find the free energies of K states simultaneously by minimizing the $K \times K$ matrix of variances of the free energy differences of these K states simultaneously. The derivation of MBAR is a straightforward if mathematically difficult extension of the derivation BAR to more than two states considered simultaneously. This can be a significant improvement over BAR, which minimize variances of the free energy differences for two states at a time. For MBAR, the equation:

$$G_i = -\frac{1}{\beta}\ln \sum_{k=1}^{K}\sum_{n=1}^{N_k} \frac{\exp[-\beta U_i(x_{kn})]}{\sum_{k'=1}^{K} N_{k'}\exp[\beta G_{k'} - \beta U_{k'}(x_{kn})]} \quad (11)$$

is solved self-consistently for each G_i. $\Delta G_{ij} = G(\lambda_j) - G(\lambda_i)$ gives the free energy change between two states i and j. The statistical variance of ΔG_{ij}, σ^2_{ij}, is calculated using eqs 8 and 12 in the paper by Shirts and Chodera.[30]

Importantly, the popular weighted histogram analysis method (WHAM)[46] for computing free energies, based on the multiple histogram algorithms of Ferrenberg and Swendsen,[62,63] can be seen as a histogram approximation to this equation. If instead of computing sums of the samples, we bin the energies U_i into a histogram for each of the intermediate states, then the MBAR equations become equivalent to WHAM equations. Similarly, if one reduces the histogram width to zero, one arrives at the MBAR equations,[64] though this derivation does not allow one to calculate an error estimate. Thus, by testing MBAR, we are also testing WHAM in the limit of sufficiently narrow histograms.

3.8. Gaussian Estimate of Exponential Averaging in Two Forms: GDEL and GINS[31]. If $\sigma^2_{\Delta U} = \langle \Delta U^2\rangle - \langle \Delta U\rangle^2$ is finite and we approximate the ΔU_{ij} distribution as a Gaussian, the free energy can be expressed as a sum of moments of the probability distribution of energy differences[36] by

$$\Delta G_{ij} = \langle \Delta U\rangle_{ij} - \frac{\beta}{2}\sigma^2_{\Delta U_{ij}} \quad (12)$$

The variance over N samples of this free energy difference is approximated in the limit of no higher moments by

$$\sigma_{ij}^2 = \frac{\sigma^2_{\Delta U_{ij}}}{N} + \frac{\beta^2\sigma^4_{\Delta U_{ij}}}{2(N-1)} \quad (13)$$

If the distribution ΔU_{ij} is close to Gaussian, then this estimation method can minimize the statistical effect of rare events, resulting in a more efficient and substantially simpler estimate method. To remove ambiguities with respect to direction of the process, we use the same convention of deletion and insertion as described for exponential averaging. In eq 12, when ΔU_{ij} is in the direction of increasing entropy, we refer to this as GDEL, and if we use ΔU_{ij} in the direction decreasing entropy, we refer to this estimate as GINS. Summing the free energy changes between intermediates again gives the total free energy changes. Total variance is calculated assuming independent sampling at each state, which is not an approximation here, as each calculation depends on samples from only one state. The total free energy is calculated by summing over the free energy changes between neighboring states.

4. SYSTEM PREPARATION, SIMULATION PARAMETERS AND STATISTICAL TESTS

4.1. System Preparation and Simulation Parameters. Topologies for united atom (UA) methane, dipole inversion, anthracene solvation test systems were created by a combination of automated tools (Dundee PRODRG,[65] OpenEye libraries[66]) and manual editing and are available on the Alchemistry.org Web site, http://www.alchemistry.org. Starting configurations were

generated using GROMACS 4.0.4. The automated topologies were solvated using GROMACS genbox, and these solvated systems were minimized with the low-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)[67] minimization method, followed by steepest descent minimization. All systems were then equilibrated at constant volume at 300 K for 100 ps, using Langevin dynamics with a time step of 0.002 fs.[68,69] All hydrogen-containing bonds were constrained using the SHAKE algorithm to a relative tolerance of 10^{-12}. The systems were then equilibrated at constant pressure at 1 atm using a Parrinello–Rahman barostat[44] and a Nose–Hoover thermostat[45] for 100 ps. A coupling time constant of 5 ps was used for both thermostat and barostat. A switching function was used for both particle mesh Ewald (PME) and van der Waals potentials. The PME switch started at 0.88 nm with a coulomb cutoff distance of 0.9 nm for electrostatics. Other PME parameters were: Fourier spacing of 0.12 nm, fourth-order B-spline interpolation, and a Ewald tolerance of 10^{-8}. A van der Waals switch at 0.8 nm and cutoff distance of 0.9 nm were used. A long-range van der Waals dispersion correction was used for both energy and pressure.

4.2. λ Values and Spacing between Intermediate States for Free Energy Calculations. In order to examine the change in bias, statistical error, and mean square error as a function of the spacing between coupling parameter λ values, we choose two sets of λ states for each model: a full λ set, and a sparse λ set.

4.2.1. Full λ Set. Initial simulations (5 ns long, including 0.5 ns equilibration) were performed with 21 equally spaced λ values to guide the selection of the λ values for the main study. The free energy analysis was done using TI. Intermediate states were chosen so that each window contributed equally to the total error such that the uncertainty $\delta(\Delta G_{i,i+1})$ vs λ_i curve for TI was flat, specifically ensuring that the maximum variance among all windows was no larger than the twice of the variance among all windows. The λ values were chosen such that each λ window contributed 0.027 ± 0.006 kJ/mol to the total uncertainty for methane solvation, 0.047 ± 0.005 kJ/mol to the total uncertainty for dipole inversion, and 0.029 ± 0.011 kJ/mol to the total uncertainty for anthracene solvation. For UA methane solvation (MS), 8 intermediates were selected: λ = [0.0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0], where λ = 0 denotes fully interacting UA methane in water and λ = 1 denotes "ghost" UA methane, where there are no interactions with the solvent. For consistency of sign between test cases, the reported results are in terms of the reverse process, the solvation of methane. For dipole inversion (DI), we include 11 intermediate states: λ = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. Here λ = 0 denotes a starting +/− configuration of the dipole; λ = 1 denotes the reversed configuration of dipole, i.e., −/+, with λ = 0.5 a state with zero partial charges. For anthracene solvation (AS), the full set contains 15 total states, with λ = [0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 1.0]; λ = 0 denotes fully interacting UA anthracene in water; λ = 1 denotes the vacuum-state anthracene with no interactions with water. Again this corresponds to a desolvation process, and in tables and charts, we report the free energy of hydration, which is simply the reverse process and so includes a sign reversal. For RBAR, this spacing means that the largest free energy between intervals is approximately 35 kJ/mol, meaning we must use range of −40 to 40 kJ/mol, and we choose increments of 1 kJ/mol.

4.2.2. Sparse λ Set. For methane solvation, we chose only three λ states [0, 0.5, 1]. For dipole inversion, the sparse set was generated by picking every alternate λ along with λ = 0.5 (which represents zero net charge) [0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0], reducing the number of states from 11 to 7. For the AS test set, every third λ was chosen to create the sparse λ set [0.0, 0.2, 0.5, 0.7, 0.85, 1.0], reducing the number of states from 15 to 6. Note that we did not need to run the separate simulations for the sparse states; we merely select for analysis only a subset of the states from the full λ set. For RBAR, this spacing means that the largest free energy between intervals is approximately 66 kJ/mol, meaning we must use a range of −70 to 70 kJ/mol, again with increments of 1 kJ/mol. In order to also test the bias with respect to number of states, for each model we conduct a set of 5 ns simulations at 51 equally spaced states, with a spacing 0.02 between 2 neighboring λ states.

4.3. Generating an Ensemble of Uncorrelated Configurations. The most important use of the benchmark test set in this paper is to compare estimates of the statistical error of different estimators of the free energy with direct sample error obtained by repeating the experiment N times. For this purpose, we started by generating 100 uncorrelated starting configurations. Using the λ = 0 state from the 5 ns test runs, we used the GROMACS program g_analyze to compute the autocorrelation time of the potential energy, kinetic energy, total energy, Coulomb interactions, and derivative $dU_{pot}/d\lambda$ for all three systems. We used block averaging[70] using the GROMACS g_analyze program to compute the autocorrelation times. The autocorrelation time of potential energy was chosen since it was the longest correlation time of the observables listed here for all molecules. The autocorrelation times of potential energy for UA methane solvation, dipole inversion, anthracene solvation were 25, 30, and 25 ps, respectively.

To generate initial configurations, each of the three test systems were run for 20 ns using a prerelease version of GROMACS 4.5 also used for subsequent free energy configurations. We then selected configurations separated by 2 ns as our uncorrelated starting points. The 2 ns spacing is more than 50 times longer than the 30 ps autocorrelation time of the potential energy. From each of these 10 parent configurations, 20 ns simulation were run with different random seeds to generate a new Maxwell–Boltzmann velocity distribution, and configurations were again selected every 2 ns, giving a total of 100 uncorrelated starting configurations for each system. The velocity Verlet integrator was used with a Nose–Hoover thermostat and the Martyna–Tuckerman–Tobias–Klein (MTTK)[71] barostat was used to control temperature and pressure, respectively, with other parameters set to the defaults discussed above.

The coordinates for simulations at the first state, at λ = 0, were selected from one of the 100 uncorrelated starting configurations. The starting coordinates for each subsequent intermediate state simulation were generated by running consecutive short 10 ps equilibration runs from the ending configuration of the previous λ state. After this initial equilibration round, 5 ns of NPT simulation were then performed for each initial configuration and each separate λ state. Data from the first 500 ps were discarded as equilibration. The remaining 4.5 ns of equilibrium data for each model at each intermediate state were used for all subsequent calculations.

4.4. Statistical Tests. *4.4.1. Quantifying Accuracy and Precision in Uncertainty Estimate of an Estimator.* We estimate the statistical uncertainty for each free energy estimator in three ways. First, we compute the sample standard deviation from ΔG's computed from the series of 100 uncorrelated simulation runs described above. Second, we compute the analytical estimates of error corresponding to each of the

methods. Finally, we use the bootstrap estimator for the standard deviation.[50]

4.4.1.1. Sample Standard Deviation. To compute the sample standard deviation, we take the simulations started from the 100 initial configurations and compute free energy differences from each simulation to obtain a distribution of free energy differences. We then directly compute the sample standard deviation corresponding to each individual estimator from

$$\sigma(\Delta G) = \sqrt{\frac{\sum_{i=1}^{N} (\langle \Delta G \rangle - \Delta G_i)^2}{N-1}} \quad (14)$$

where N = 100 and $\langle \Delta G \rangle$ is the mean over the 100 values of ΔG_i's. Crucially, the standard deviation computed from a finite-sized sample is itself a statistical quantity and must therefore have an associated uncertainty. Rigorously, in order to compute the sample standard deviation of the uncertainty $\delta(\sigma(\Delta G))$, we would need to repeat our 100 simulation experiment 100 times. Instead, we have used the bootstrap method (described in the bootstrap estimate section) to estimate $\delta(\sigma(\Delta G))$; as will be seen later, the bootstrap method is an effective way to compute estimates from independent free energy calculations. From this exercise we finally get $\langle \Delta G \rangle$ and $\langle \sigma(\Delta G) \rangle \pm \delta(\sigma(\Delta G))$; $\langle \Delta G \rangle$ here indicates not an ensemble average but the average over 100 repetitions.

4.4.1.2. Analytical Estimate. Each free energy estimator has an associated uncertainty estimator as discussed in previous sections, namely the square root of the estimated variance of the total free energy. From 100 uncorrelated starting configurations, we will obtain not only 100 ΔG's, but 100 error estimates from each method's analytical uncertainty estimate. We denote the average and standard deviation of these estimated uncertainties over all 100 independent runs as $\langle \delta(\Delta G) \rangle \pm \delta(\delta(\Delta G))$ and call these the analytical uncertainty estimate and the standard deviation of the analytical uncertainty estimate.

4.4.1.3. Bootstrap Estimate. For each of the 100 independent free energy calculations, we also calculate a bootstrap error estimate, a well-known and robust technique in the statistical literature.[50] The bootstrap error is constructed as follows: From each set of potential energy differences or $dU/d\lambda$ values, we generate N bootstrap sets from the original set of molecular simulation data. To generate a bootstrap set, we first subsample the data using an estimate of the autocorrelation time to obtain N statistically uncorrelated values. For each bootstrap set, we draw N samples with replacement from the original set of uncorrelated measurements. For example, if our set was the integers {3,6,8,9}, then a bootstrap set would consist in randomly selecting each of the four numbers for times; {3,3,8,9}, {9,6,6,3}, and {8,8,8,8} would all be valid sets, though clearly the last one would be the rarest. This subsampling process is then repeated many times; in this particular study, we draw 200 bootstrap sets. Standard rules of thumb suggest using 50−200 bootstrapped sets to get robust estimates of uncertainty,[50] though anecdotally some users say they get more consistent results using 1000 or more bootstrap sets. For each of these 200 bootstrap sets, we compute the free energy and the uncertainties using the estimators as if they were the original data set. This gives us 200 ΔG's, one for each of the 200 bootstrapped sets. The average of these 200 bootstrapped ΔG_{bs}'s gives $\langle \Delta G \rangle_{bs}$. The bootstrap estimate of the error, $\delta(\Delta G)_{bs}$ is the sample standard deviation of the 200 $\langle \Delta G \rangle_{bs}$ values for each of the initial configurations. The average $\langle \delta(\Delta G)_{bs} \rangle$ over all 100 initial configurations is the bootstrap error estimate. The statistical uncertainty of this bootstrap uncertainty estimate is estimated by computing the sample standard deviation over the 100 $\delta(\Delta G)_{bs}$ and is denoted by δ ($\delta(\Delta G)_{bs}$).

The analytical estimate of a free energy estimator by definition should agree with the sample standard deviation. If it does not, then any estimate of error using the analytical estimate will be unreliable. For example, the analytic estimate of the uncertainty of EXP diverges from the true estimate well before the error in EXP itself.[28] If the statistics are well-behaved, the bootstrap estimate of the statistical uncertainty should also agree with the sample estimate of the uncertainty. The smaller the difference between the direct sample standard deviation error estimate $\langle \sigma(\Delta G) \rangle$ and the analytical error $\langle \delta(\Delta G) \rangle$ or bootstrap estimates, the better we know the method's variability without having to run multiple trials. Additionally, if we can show that the bootstrap estimate agrees with the sample estimate of the uncertainty, then bootstrap error can substitute for sample uncertainty estimates even when the analytical estimate fails.

4.5. Quantifying Bias of Free Energy Estimates. The average estimate from a statistically biased estimator, even if repeated many times, will still deviate from the true estimate by the bias. There are typically two types of bias in the free energy estimators considered here. Asymptotically unbiased estimators have bias with finite number of samples but the bias decreases to zero in the limit of large numbers of samples. An example is the naïve estimator of the variance in the average of a set of numbers, which is $\mathrm{var}(a) = N^{-1} \Sigma[(\langle A \rangle - A_i)^2]$. This estimate can be shown to always be slightly too large, by an amount proportional to the number of samples, and is thus asymptotically biased. If N is replaced by $N - 1$, however, the estimator becomes unbiased for any number of samples. In the limit of very large N, the bias of an asymptotically biased estimator will be effectively zero, but a given estimator might require a very large number of samples to reach this point. Thermodynamic integration does not have asymptotic bias, because at each intermediate state, the simple average of $dU/d\lambda$ is unbiased for any numbers of samples. Exponential averaging, BAR, and MBAR are only asymptotically unbiased, though the bias of exponential averaging is usually significantly higher than that of BAR and MBAR[28] and careful design of the pathway can minimize this large bias to some extent.[72] However, unlike the simple case of the estimator for the variance of the simple mean of samples, there exist no known unbiased versions of these free energy estimators.

Bias also occurs due to using a limited number of intermediate states because of lack of either phase space overlap between intermediates, as occurs in acceptance ratio methods and exponential averaging, or because of numerical integration error in TI methods. Even for a large benchmark set like the present study, computational expense and storage limits make it very difficult to approach the number of sample limit and the number of intermediate states limit simultaneously. Therefore, we attempt to estimate the two contributions to bias independently. For asymptotic bias, we compare the results from combining a fixed amount of data in either one large data set with averaging the results over or a series of shorter sets of the same data. For bias as a function of number of intermediate states, we vary the number of intermediates with fixed total length of simulation to investigate bias as a function of number of intermediate states.

4.5.1. Bias Due to Number of Samples. Data from all 100 5 ns runs is "stitched" into a single large data set analyzed simultaneously.

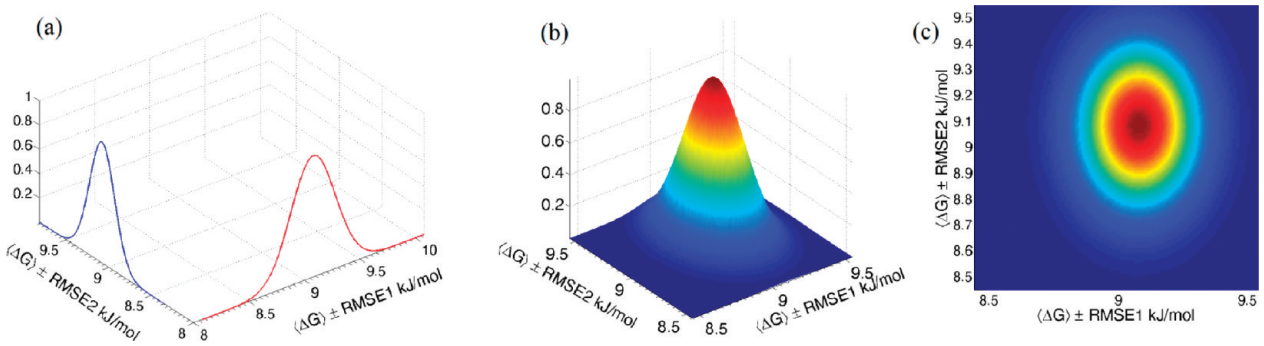

Figure 3. (a) Gaussians are plotted on mutually perpendicular axes. Both have the $\langle \Delta G \rangle$ calculated using a method (here TI for methane solvation) as mean and RMSE1 and RMSE2 as their standard deviations. (b) These are fused to generate a bivariate Gaussian plot. (c) Top view of the bivariate Gaussian plot .

The data corresponding to first 0.5 ns equilibrations were not included while estimating free energies, resulting in 450 ns total simulation data. The difference between free energy estimates computed with 450 ns of data $(\Delta G)_{450}$, and the same data used to compute an average over 100 4.5 ns trajectories $\langle \Delta G \rangle$ is what we report the bias due to number of samples.

4.5.2. Bias Due to Number of Intermediate States. We also ran a set of simulations with 51 λ states for each model, as discussed above. We can estimate the bias due to the number of intermediate states by comparing the differences between free energy estimates computed using 51 λ states $(\Delta G)_{51}$, ΔG estimated using the full set of λ states, and ΔG estimated using the sparse λ states. Besides having low and consistent error estimates, ideal free energy estimation methods should show little or no bias in these tests. In many cases, we are limited by the statistical uncertainty in determining the bias with high accuracy, since it is computationally too demanding to generate 500 ns of simulation for all 51 states for all test sets. In these cases, we can only determine if the bias is statistically insignificant with respect to the statistical uncertainty. Asymptotic bias scales with $1/N$ while statistical uncertainty scales with $N^{-1/2}$, so statistical variance is usually dominant and will always be dominant if sufficient samples are collected.

4.6. Quantifying Reliability of a Free Energy Estimator. Root mean square error (RMSE) is the square root of the sum of squared differences between the sample and the true answer. Alternatively, we can write it as the sum of the variance estimate (σ^2) and the square of the bias. Mean square error is therefore the most general overall measure of the reliability of any statistically estimated observable. Calculating a true RMSE, of course, requires collecting an infinite number of samples at an infinite number of intermediate states to obtain the true answer. Since it is computationally impossible to reach this limit and computationally expensive to approach it, we use the free energy estimate ΔG_{450} from the 450 ns run to approximate the unbiased limit of free energies given a fixed set of intermediates and the free energy estimate ΔG_{51} from 51 λ state simulations as the unbiased limit of the free energy estimate with respect to large numbers of intermediate states, given a fixed amount of sampling. From the two different biases generated from these two reference states, we obtain two different estimates of mean square error. Neither of these is a true RMSE because we lack the true reference answer. However, these estimates of the RMSE's capture the combined effect of the statistical error and the two different sources of bias.

We define the two estimates of mean square errors as MSE1 and MSE2:

$$\begin{aligned} \mathrm{MSE1}_i &= \sigma_i^2 + \mathrm{bias1}_i^2, \quad \text{where bias1}_i \\ &= (\Delta G)_{450} - (\Delta G_{\mathrm{est}})_i \text{ and } 1 \le i \le 100 \end{aligned} \tag{19}$$

$$\begin{aligned} \mathrm{MSE2}_i &= \sigma_i^2 + \mathrm{bias2}_i^2, \quad \text{where bias2}_i \\ &= (\Delta G)_{51} - (\Delta G_{\mathrm{est}})_i \text{ and } 1 \le i \le 100 \end{aligned} \tag{20}$$

RMSE1 and RMSE2 are defined as the averages over the square root of $\mathrm{MSE1}_i$ and $\mathrm{MSE2}_i$, respectively. The errors in RMSE1 and RMSE2, δ(RMSE1) and δ(RMSE2), respectively, are the standard deviations over the 100 $\mathrm{RMSE1}_i$ and 100 $\mathrm{RMSE2}_i$. With the two quantities RMSE1 and RMSE2, we can examine qualitative information about the reliability of the methods using all the information from our experiments. We plot a bivariate Gaussian with variances equal to RMSE1 and RMSE2 on mutually perpendicular axes, as shown in Figure 3a, with the analytical average of the free energy estimate $\langle \Delta G \rangle$ estimated by the method as the mean. An example bivariate Gaussian RMSE plots is shown in Figure 3b. Figure 3c shows the top view of this Gaussian. The overall spread of the rings in Figure 3c is a measure of overall reliability. A poor estimator has large and/or unequal spreads in horizontal and vertical directions, yielding a large circle or an ellipse. A good estimator has small and equal spread in both the horizontal and vertical direction. Vertical spread indicates that bias due to number of intermediate states dominates the uncertainty estimate, while horizontal spread indicates bias due to number of samples dominates the uncertainty estimate.

4.7. Validating the Gaussian Distributions of the Free Energy Differences. When we express uncertainty in the free energy estimate in terms of a single number, the variance or the statistical uncertainty, rather than as a distribution, we are implicitly assuming that the errors are well described by a normal distribution, whose spread is equal to the statistical variance. Most analytical variance methods use propagation estimates that are only rigorously true in the Gaussian limit. As long as the variances are bounded (not infinite), the central limit theorem ensures that with enough samples, variances will indeed converge to the Gaussian limit. However, for finite number of samples, this assumption must be tested, not simply assumed, or else we run the risk of underestimating the chance of large deviations ("black swans") from the average value.

To test whether the shape of the free energy distribution is Gaussian or not, we plot histograms of the distribution of free

Table 1. Statistical Uncertainty Calculated Using Three Different Approaches (Analytical $\langle\delta(\Delta G)\rangle$, Sample Standard Deviation $\sigma(\Delta G)$, and Bootstrap $\langle\delta(\Delta G)_{bs}\rangle$) for UA Methane Solvation Using the Full λ Set[a]

method	$\langle\Delta G\rangle$	$\langle\delta(\Delta G)\rangle \pm \delta(\delta(\Delta G))$	$\sigma(\Delta G) \pm \delta(\sigma(\Delta G))$	$\langle\delta(\Delta G)_{bs}\rangle \pm \delta(\delta(\Delta G)_{bs})$	% dev $\pm$ δ (% dev)
TI	9.081	0.107 ± 0.002	0.115 ± 0.009	0.106 ± 0.006	−7.1 ± 7.9
TI3	9.008	0.111 ± 0.003	0.119 ± 0.012	0.110 ± 0.007	−6.9 ± 9.5
DEXP	8.984	0.320 ± 0.255	0.556 ± 0.122	0.319 ± 0.267	−42.5 ± 47.5
IEXP	8.928	0.104 ± 0.002	0.110 ± 0.011	0.103 ± 0.006	−5.8 ± 9.6
UBAR	8.936	0.079 ± 0.002	0.106 ± 0.010	0.098 ± 0.005	−25.3 ± 7.5
BAR	8.933	0.075 ± 0.001	0.109 ± 0.010	0.099 ± 0.006	−31.3 ± 6.3
RBAR	8.937	0.075 ± 0.001	0.109 ± 0.010	0.099 ± 0.006	−31.0 ± 6.7
MBAR	8.929	0.095 ± 0.002	0.106 ± 0.010	0.094 ± 0.005	−9.9 ± 8.6
GDEL	7.042	0.093 ± 0.002	0.136 ± 0.011	0.121 ± 0.008	−31.7 ± 5.8
GINS	1.097	0.253 ± 0.008	0.399 ± 0.032	0.400 ± 0.169	−36.5 ± 5.5

[a] All quantities are in kJ/mol. $\langle\Delta G\rangle$ is not the ensemble average but the average over 100 repetitions.

Table 2. Statistical Uncertainty Calculated Using Three Different Approaches (Analytical $\langle\delta(\Delta G)\rangle$, Sample Standard Deviation $\sigma(\Delta G)$, and Bootstrap $\langle\delta(\Delta G)_{bs}\rangle$) for UA Methane Solvation Using Sparse λ Set[a]

method	$\langle\Delta G\rangle$	$\langle\delta(\Delta G)\rangle \pm \delta(\delta(\Delta G))$	$\sigma(\Delta G) \pm \delta(\sigma(\Delta G))$	$\langle\delta(\Delta G)_{bs}\rangle \pm \delta(\delta(\Delta G)_{bs})$	% dev $\pm$ δ (% dev)
TI	2.545	0.175 ± 0.007	0.177 ± 0.014	0.175 ± 0.011	−1.5 ± 8.7
TI3	3.792	0.214 ± 0.009	0.217 ± 0.017	0.214 ± 0.014	−1.5 ± 8.7
DEXP	5.631	1.343 ± 0.524	3.179 ± 0.679	1.628 ± 1.186	−57.8 ± 18.8
IEXP	9.091	0.666 ± 0.101	0.638 ± 0.042	0.683 ± 0.108	4.4 ± 17.3
UBAR	8.954	0.344 ± 0.018	0.358 ± 0.024	0.351 ± 0.024	−3.9 ± 8.2
BAR	8.926	0.225 ± 0.006	0.263 ± 0.014	0.232 ± 0.014	−14.4 ± 4.9
RBAR	8.927	0.226 ± 0.005	0.260 ± 0.016	0.233 ± 0.014	−13.2 ± 5.6
MBAR	8.928	0.232 ± 0.006	0.262 ± 0.015	0.232 ± 0.014	−11.4 ± 5.4
GDEL	−3.833	0.112 ± 0.004	0.189 ± 0.014	0.200 ± 0.016	−40.6 ± 4.9
GINS	-1.68×10^{32}	$3 \times 10^{3}0 \pm 34 \times 10^{3}0$	$13 \times 10^{32} \pm 10 \times 10^{32}$	$1 \times 10^{32} \pm 15 \times 10^{32}$	−99.7 ± 2.7

[a] All quantities are in kJ/mol.

Table 3. Free Energy Estimates and Corresponding Uncertainty Estimates in the Large Number of Samples (450 ns) and Large Number of Intermediate States (51 λ states) for UA Methane Solvation[a]

method	$((\Delta G) \pm \delta(\Delta G))_{450,full}$	$((\Delta G) \pm \delta(\Delta G))_{450,sp}$	$((\Delta G) \pm \delta(\Delta G))_{51,full}$
TI	9.085 ± 0.010	2.541 ± 0.017	8.920 ± 0.041
TI3	9.016 ± 0.010	3.786 ± 0.021	8.923 ± 0.041
DEXP	9.083 ± 0.083	12.657 ± 4.018	8.921 ± 0.043
IEXP	8.932 ± 0.010	8.986 ± 0.074	8.928 ± 0.040
UBAR	8.939 ± 0.010	8.930 ± 0.035	8.922 ± 0.040
BAR	8.939 ± 0.010	8.920 ± 0.023	8.921 ± 0.040
RBAR	8.940 ± 0.010	8.923 ± 0.024	8.922 ± 0.040
MBAR	8.936 ± 0.009	8.921 ± 0.023	8.924 ± 0.036
GDEL	7.048 ± 0.012	−3.837 ± 0.020	8.847 ± 0.043
GINS	1.101 ± 0.041	$-1.5 \times 10^{32} \pm 3.2 \times 10^{29}$	8.841 ± 0.040

[a] Bootstrap estimates are reported as they are better than analytical estimates. All quantities are in kJ/mol. The subscript (450, full) denotes the free energy estimate for 450 ns and full λ set and (450, sp) denotes the same for sparse λ set.

energies from each method against Gaussian distributions. For each Gaussian, we use the average free energy estimate as the mean and the analytical uncertainty estimate as the standard deviation of the Gaussian for each method. To validate our analysis from the visual comparison of the curves we have calculated the *p*-value for the Shapiro−Wilk[73] test, which are used to accept or reject the null hypothesis that the ensemble of 100 free energies are drawn from a normal distribution.

5. RESULTS AND DISCUSSION

Results are presented in Tables 1−5 as well as in Figures 4−15. Tables 1 and 2 contain results of the uncertainty analysis for full and sparse λ sets only for methane solvation. Tables 3−5 contain the free energy estimates corresponding to 450 ns and 51 λ states runs, bias analysis, reliability estimates of free energy, and uncertainty predictions for UA methane solvation for full and sparse λ sets. Figures 4−7 provide comparison of the accuracy

Table 4. Bias Estimates Due to Number of Samples and Number of λ States for Full and Sparse λ Sets for UA Methane Solvation[a]

method	$(\text{bias1} \pm \delta(\text{bias1}))_{450,\text{full}}$	$(\text{bias2} \pm \delta(\text{bias2}))_{51,\text{full}}$	$(\text{bias1} \pm \delta(\text{bias1}))_{450,\text{sp}}$	$(\text{bias2} \pm \delta(\text{bias2}))_{51,\text{sp}}$
TI	-0.005 ± 0.015	0.160 ± 0.042	0.005 ± 0.024	-6.374 ± 0.045
TI3	-0.005 ± 0.015	0.088 ± 0.042	0.006 ± 0.030	-5.131 ± 0.046
DEXP	-0.100 ± 0.092	0.062 ± 0.059	-7.044 ± 4.021	-3.308 ± 0.150
IEXP	-0.002 ± 0.014	0.002 ± 0.041	0.097 ± 0.100	0.155 ± 0.078
UBAR	-0.003 ± 0.013	0.014 ± 0.041	0.024 ± 0.049	0.032 ± 0.053
BAR	-0.003 ± 0.012	0.015 ± 0.041	0.009 ± 0.032	0.008 ± 0.046
RBAR	-0.005 ± 0.012	0.013 ± 0.041	0.004 ± 0.033	0.005 ± 0.046
MBAR	-0.007 ± 0.013	0.005 ± 0.037	0.007 ± 0.033	0.004 ± 0.043
GDEL	-0.003 ± 0.015	-1.802 ± 0.044	0.001 ± 0.023	-12.683 ± 0.044
GINS	-0.001 ± 0.048	-7.741 ± 0.047	$-1.4 \times 10^{31} \pm 3.4\text{+E30}$	$-1.7 \times 10^{32} \pm 3.4 \times 10^{3}0$

[a] All quantities are in kJ/mol. The subscript (450, full) denotes the bias estimate for 450 ns and full λ set and (450, sp) denotes the same for sparse λ set. The subscript (51, full) denotes the bias estimate for 51 λ set and full λ set and (51, sp) denotes the same for sparse λ set. Bias1 refers to bias due to number of samples. Bias2 refers to bias due to intermediate states.

Table 5. RMSEs and Statistical Uncertainties in RMSEs for UA Methane Solvation[a]

method	$((\text{RMSE1}) \pm \delta(\text{RMSE1}))_{\text{full}}$	$((\text{RMSE2}) \pm \delta(\text{RMSE2}))_{\text{full}}$	$((\text{RMSE1}) \pm \delta(\text{RMSE1}))_{\text{sp}}$	$((\text{RMSE2}) \pm \delta(\text{RMSE2}))_{\text{sp}}$
TI	0.148 ± 0.054	0.208 ± 0.084	0.237 ± 0.076	6.377 ± 0.178
TI3	0.153 ± 0.059	0.172 ± 0.069	0.290 ± 0.093	5.135 ± 0.218
DEXP	0.531 ± 0.474	0.490 ± 0.512	7.596 ± 2.138	4.478 ± 1.837
IEXP	0.142 ± 0.051	0.142 ± 0.052	0.897 ± 0.271	0.904 ± 0.273
UBAR	0.121 ± 0.054	0.121 ± 0.055	0.473 ± 0.147	0.473 ± 0.148
BAR	0.120 ± 0.055	0.120 ± 0.056	0.330 ± 0.103	0.330 ± 0.103
RBAR	0.120 ± 0.055	0.121 ± 0.056	0.331 ± 0.102	0.331 ± 0.102
MBAR	0.132 ± 0.052	0.132 ± 0.052	0.334 ± 0.102	0.334 ± 0.102
GDEL	0.152 ± 0.065	1.807 ± 0.137	0.201 ± 0.090	12.682 ± 0.190
GINS	0.427 ± 0.205	7.748 ± 0.402	$3.1 \times 10^{32} \pm 1.5 \times 10^{33}$	$1.5 \times 10^{32} \pm 1.5 \times 10^{33}$

[a] All quantities are in kJ/mol. The subscripts (full) and (sp) denotes the free energy estimates for the full λ set and sparse λ set, respectively. RMSE1 uses bias due to number of samples and sample standard deviation uncertainty estimate. RMSE2 uses bias due to number of intermediate states and sample standard deviation uncertainty estimate. Free Energy estimates are largest for GINS and GDEL in sparse λ set, while the acceptance ratio methods consistently show low RMSEs and the corresponding uncertainty in the RMSEs.

and precision in free energy and the uncertainty predictions for all 10 methods for the three test cases. Bivariate Gaussian plots presenting the analysis of reliabilities of free energy methods are presented in Figures 8−10. Figures 11 and 12 compare the actual distribution of the free energies (computed using 100 5 ns simulations with full and sparse λ sets) with Gaussians using the corresponding variance and the mean estimates of the free energy methods, along with the Gaussians of 450 ns trajectory solution and the 51 λ state simulation set for comparison. In some plots, we omit GDEL and GINS, as their errors are significantly larger than the scale of errors of the other methods in the plots. Figures 13 and 14 contain the Shapiro−Wilk test comparison for all methods for methane solvation for full and sparse λ sets. Figure 15 contains a free energy convergence comparison for all methods. Tables and figures for dipole inversion and anthracene hydration free energy test cases are presented in the Supporting Information.

5.1. Validation of Uncertainty Estimates. The first column in Table 1 is the free energy change calculated as an average over the 100 repetitions from uncorrelated configurations. The next three columns are the average of the analytical estimate of uncertainty over all repetitions $\langle\delta(\Delta G)\rangle$, the sample standard deviation $\sigma(\Delta G)$, and the average bootstrap estimate of the uncertainty over all repetitions $\langle\delta(\Delta G)_{\text{bs}}\rangle$. The last column gives the percent deviation of the analytical estimate of uncertainty from the sample standard deviation along with the propagated error. Standard deviations $\{\delta(\delta(\Delta G)),\ \delta(\sigma(\Delta G)),\ \delta(\delta(\Delta G)_{\text{bs}}\}$ for the error estimate distributions are also reported. Importantly, the error estimates are generally predicted very consistently with the exception of the error estimates for DEXP, with a relative error of usually between 5% and 10%. Thus, the error estimates obtained by almost all methods can generally be relied on to be consistent between different repetitions. This does not necessarily guarantee that they will accurately predict the variation in the free energies obtained with different data sets but at least means the error estimates will be consistent between data sets.

We find that the analytic error estimate of most methods underestimates the true sample standard deviation, most often only slightly, but occasionally more significantly. For some of the methods, this deviation is always within either one or two standard deviations and thus likely within the statistical noise. Examining the data in Tables 1 and S2 and S3 in the Supporting Information for TI and TI3, analytical and sample estimates of uncertainty differ by 1 or less standard deviation for all but the anthracene solvation with the full λ set, where it differs by approximately 2 standard deviations and is likely therefore noise.

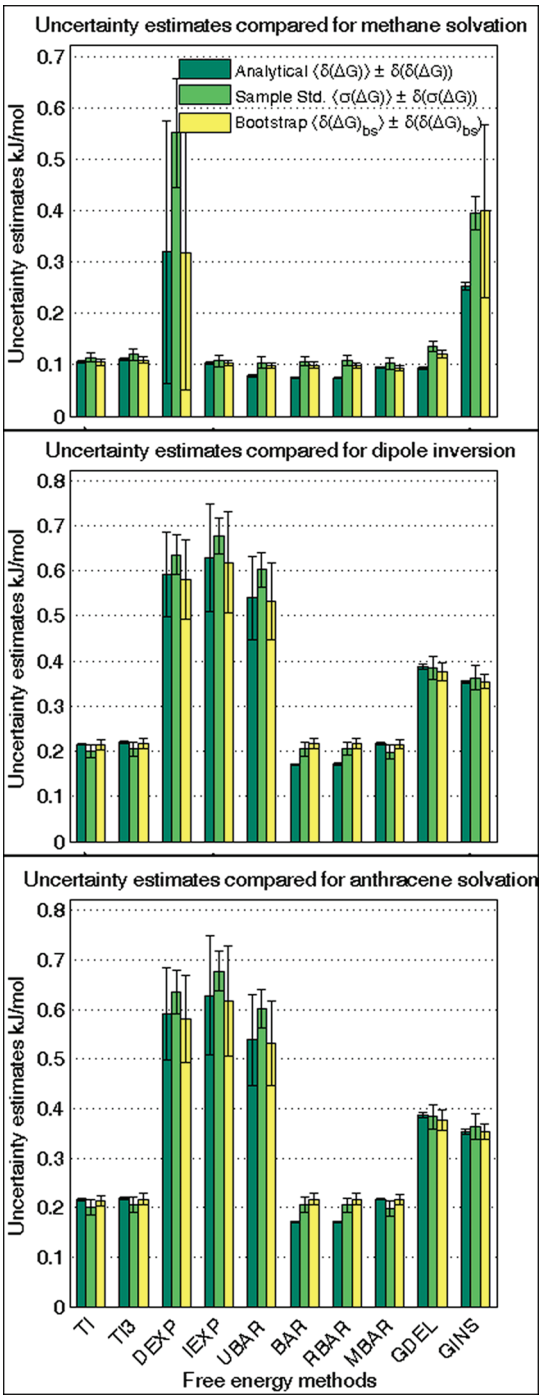

Figure 4. Uncertainty estimates (sample standard deviation, analytical, bootstrap) are plotted for all the methods in three different test cases for full λ set. Consistent free energy methods have bars equal in height, and the most precise methods have the shortest bars.

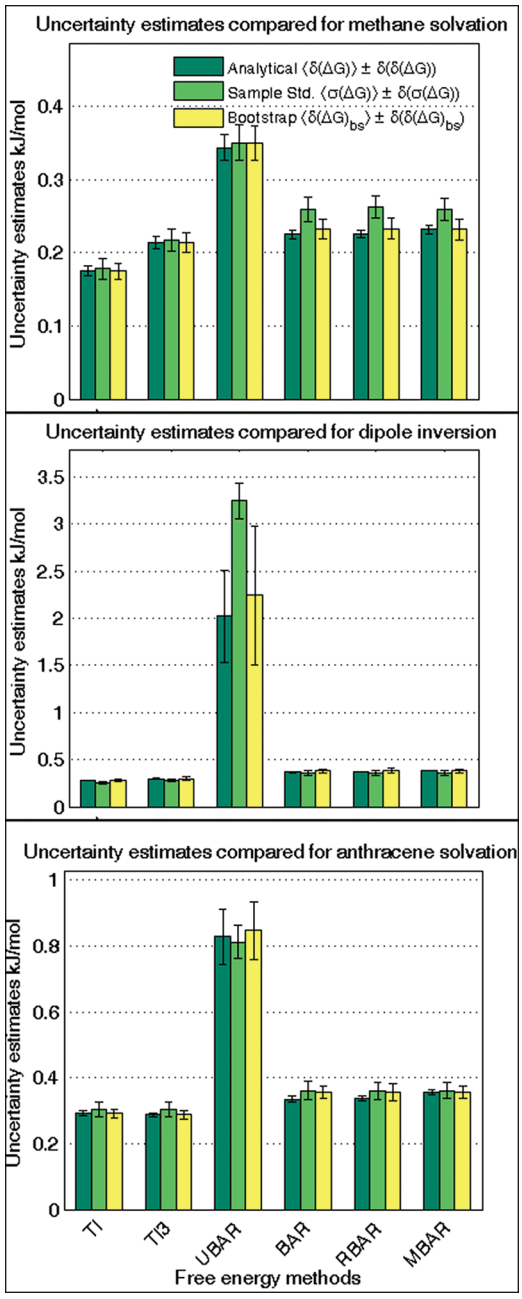

Figure 5. Uncertainty estimates (sample standard deviation, analytical, bootstrap) are plotted for six methods in three different test cases for sparse λ set. Four (DEXP, IEXP, GINS, and GDEL) are not shown because they did not converge properly for any of the three uncertainty estimates.

For IEXP and DEXP, analytical and sample estimates also differ by 1 or in some cases 2 standard deviations. However, analytical and sample uncertainty estimates for BAR, RBAR, and UBAR have differences of up to 5 standard deviations, which indicates that the BAR uncertainty estimates can be significantly inaccurate for multiple intermediate states. In contrast, analytical uncertainty estimates with MBAR have less than 1 standard deviation from sample uncertainty estimates. Quantitatively, as seen in column six of Table 1, the percentage deviation of the analytical uncertainty estimate from sample uncertainty for TI is $-7 \pm 8\%$ and for MBAR is $-10 \pm 8\%$, with the negative sign indicating a negative deviation of the analytical estimate. In both cases, this appears to be within the noise. For BAR, the deviation from the true uncertainty is $-31 \pm 5\%$, which is clearly statistically significant, with similar results for RBAR and UBAR. As with BAR and its variants, analytical and sample estimates GDEL and GINS are off by $30 \pm 5\%$. Over all methods, MBAR and TI analytical error estimates deviate least from the sample estimate.

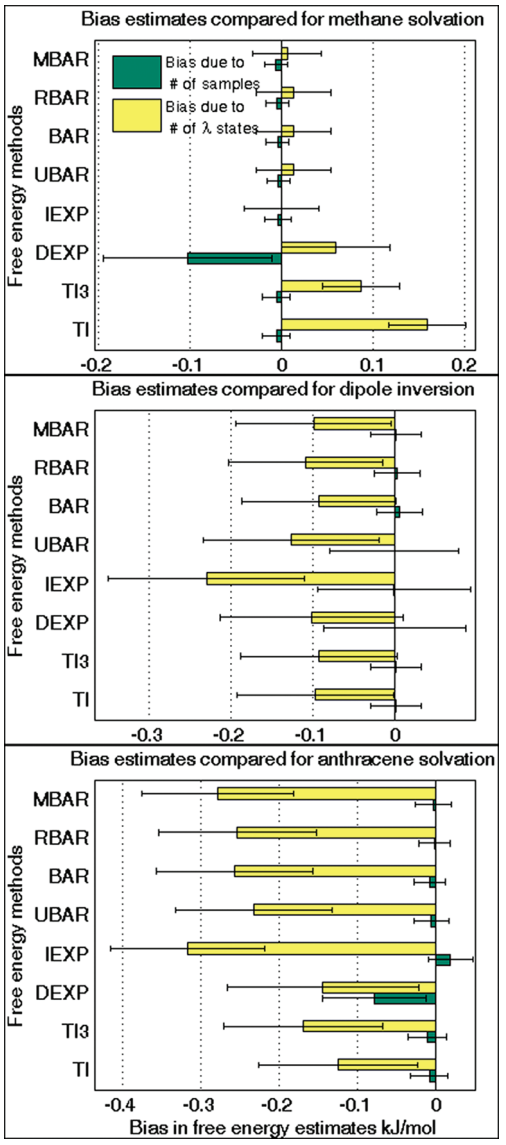

Figure 6. Bias plots for different test cases for the full λ set. DEXP, TI, and TI3 show numerically significant bias for methane, while all methods show moderate bias with respect to number of states for anthracene, with TI and TI3 showing possibly less bias.

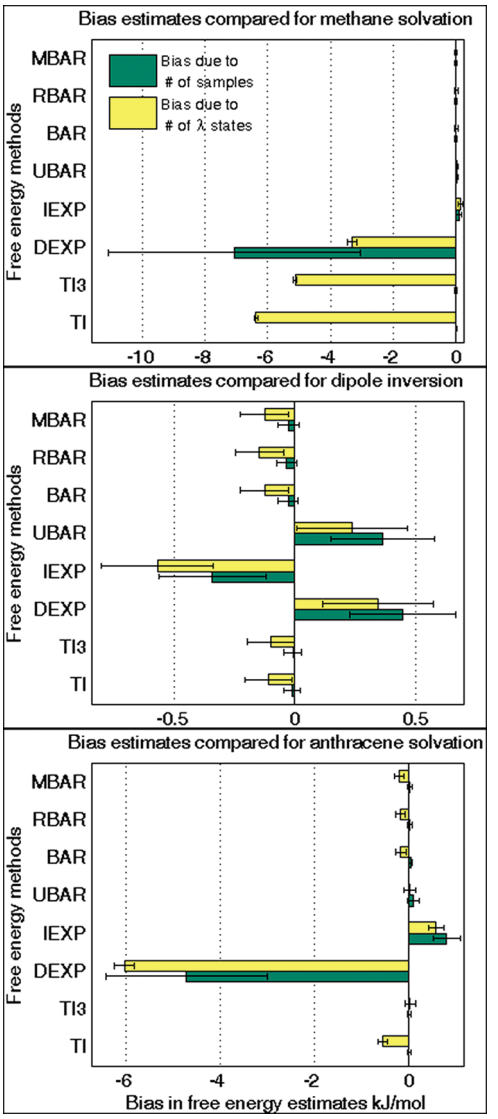

Figure 7. Bias plots for different test cases or sparse λ set. DEXP and IEXP show large biases both due to number of samples and due to number of intermediate states.

Bootstrap uncertainty is a robust alternative to the sample standard deviation for all methods. Bootstrap estimates of the uncertainty (in column five of Table 1) are very close to sample uncertainty (in column four of Table 1). Specifically in the case of BAR, RBAR, UBAR, GDEL, and GINS bootstrap estimates of uncertainty are statistically equivalent to the sample standard error estimates, unlike their analytical counterparts. In cases where the analytical estimate does not accurately predict the sample standard deviation, such as with BAR, RBAR, and UBAR, the bootstrap method clearly provides a useful estimate of the error in the free energy without a need for performing repeated sampling.

Figure 4 visually demonstrates the efficiency of free energy methods and the consistency of uncertainty estimators. Short bars indicate precise free energy estimates. Equal height bars indicate that the analytical and bootstrap uncertainties are consistent with the sample standard deviation. For the full λ set, MBAR, TI, and TI3 predict free energies with the highest precision and with the most reliable error estimate. BAR, RBAR, and UBAR analytical estimates have large deviations from the standard deviation estimate, but the bootstrap estimate closely matches sample standard uncertainty estimate. IEXP and DEXP have the largest deviations of the analytical error estimate particularly for large transformations like dipole inversion and anthracene solvation. GINS and GDEL have the high uncertainty estimates and therefore give imprecise estimates of free energy.

For the sparse λ set, as shown in Tables 2 and S2 and S4 in the Supporting Information and Figure 5, TI and TI3 still show the lowest percent deviation from sample standard deviation. However, the free energy estimate of methane solvation is off by 6.5 kJ/mol for TI and 5 kJ/mol for TI3. GDEL and GINS have clearly unconverged free energy and uncertainty estimates; the free energy estimate of methane solvation for GDEL is off by 12 kJ/mol and for GINS is off by $\sim 10^{32}$ kJ/mol! This clearly indicates the failure of the Gaussian approximation of ΔU for most molecular transformation problems. The free energy

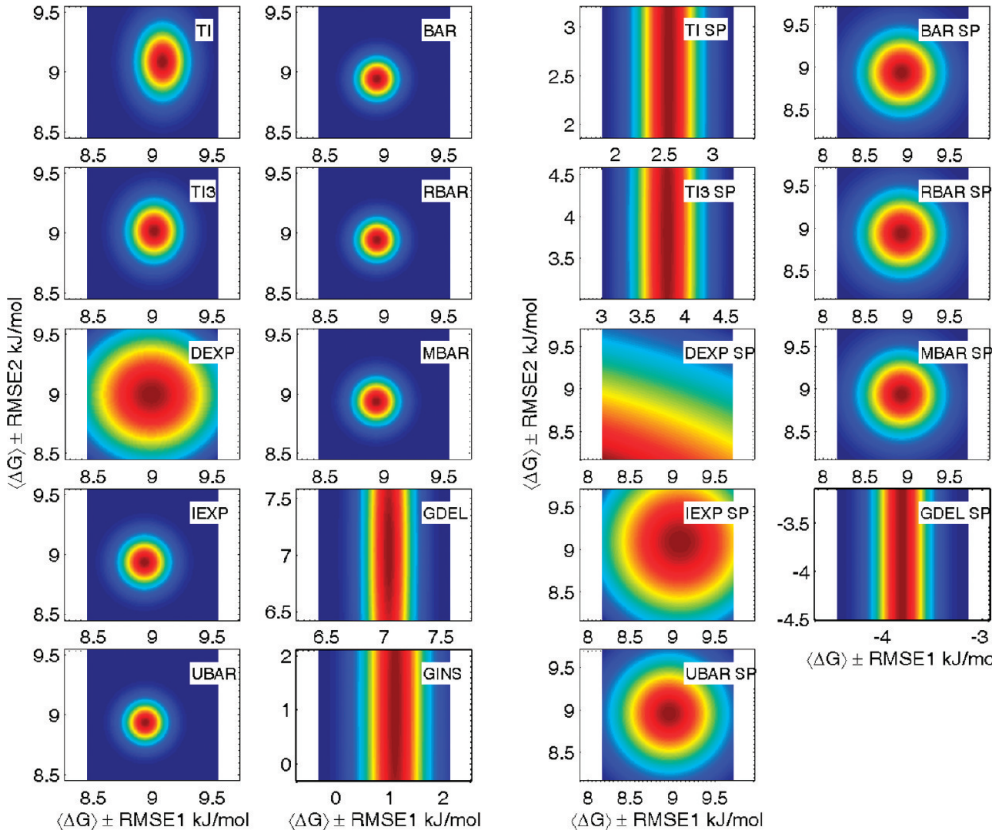

Figure 8. Bivariate Gaussian plots for UA methane solvation. TI and TI3 tend to have high bias for sparse λ sets. GDEL and GINS are not at all reliable and are not shown, and MBAR and BAR are accurate but the precision in BAR is misleading as it underestimates uncertainty, as shown in Figures 4 and 5. SP after the method name indicates sparse λ set.

estimate of methane solvation for DEXP differs from the converged answer by 3 kJ/mol, and its uncertainty estimate is 5 times larger than the largest estimated uncertainty shown in Figure 5. IEXP differs from the converged answer by only 0.2 kJ/mol, but its uncertainty estimate is twice the largest plotted uncertainty. MBAR again has the lowest and most consistent uncertainty estimates. However, unlike with the full λ set, BAR and RBAR have uncertainty estimates which are as accurate as MBAR to within statistical noise. This appears to be because as the energy differences between the i and $i+1$ and i and $i-1$ get larger, their correlations between the estimators computed from these energies decrease.

Bootstrap and analytical estimates of the error in the sparse λ set are slightly lower than the sample standard deviation for acceptance ratio methods for methane solvation, though not for the other two molecules. For this sparse set of λ states, MBAR does not provide the same clear advantage over BAR in estimating the uncertainty as with the full λ set. We hypothesize that this advantage may only exist when the overlap between states is non-negligible. However, even our full λ set uses relatively aggressive spacing compared to typical free energy calculations. Therefore, unless we are certain we are in a low overlap regime, MBAR will be preferred to BAR.

5.2. Analysis of Bias. The statistical uncertainty is the most important measure to quantify in order to understand the reliability of the free energy estimate but understanding systematic bias including both bias due to finite sample size and numerical integration is also important. Table 3 shows the free energy and the uncertainty estimates predicted by different methods for UA methane solvation for large number of samples (450 ns trajectory) and large number of intermediate states (51 λ states). Tables 4 and 5 include estimates of both types of bias in free energy estimates, with respect to number of samples and with respect to number of intermediate states, and the corresponding RMSE estimates for UA methane solvation. For UA methane solvation, MBAR, BAR, RBAR, UBAR, TI, and TI3 have very low bias in free energy estimates with respect to number of samples. The acceptance ratio methods, MBAR, BAR, RBAR, and UBAR, have biases with respect to number of states within the statistical noise. TI and TI3 show larger bias in free energy estimates with respect to number of states than the other methods. DEXP and IEXP show large biases in free energy estimates both with respect to number of samples and states. GDEL and GINS show the largest bias with respect to number of states. All methods show a larger bias with respect to the number of λ states compared to the bias with respect to number of samples. However, this may be an artifact of the lower precision of bias determination as a function of the number of states.

Figures 6 and 7 show the biases for different methods for all three test cases. When error bars are larger than the bias bars, for example, for MBAR in methane solvation, comparisons for accuracy between free energy methods become difficult. DEXP and IEXP have statistically significant bias for all the cases except in methane solvation. For UA methane solvation sparse set, TI unsurprisingly has the largest bias due to number of states, as only three states were used.

Overall, MBAR, BAR, UBAR, and RBAR have consistently less bias compared to other methods and hence are more accurate for

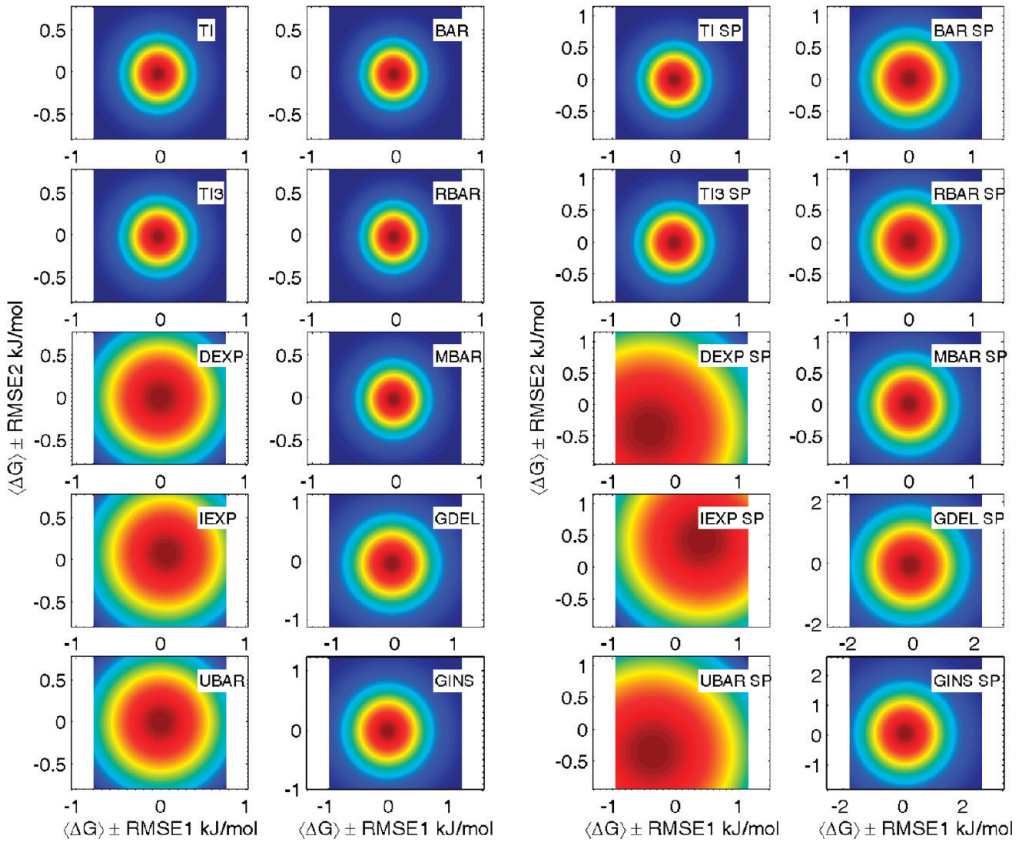

Figure 9. Bivariate Gaussian plots for dipole inversion. TI is reliable for this molecule, and DEXP, IEXP, GDEL, and GINS again are the least accurate and precise.

estimating free energy. TI and TI3 have more bias, which becomes substantial for sparse sets. IEXP and DEXP and especially GDEL and GINS have even larger bias. For dipole inversion, DEXP and IEXP show the largest bias both with respect to results from large number of samples and large number of intermediate states. All other methods for dipole inversion, including GINS and GDEL, show almost equal biases within statistical error limits of zero. For anthracene hydration free energies BAR, UBAR, RBAR, and MBAR show moderate biases for the full λ set compared to TI and TI3, but these results are again likely to be noise (see Table S9 in the Supporting Information). DEXP and IEXP as usual show high biases.

5.3. Overall Reliability. The knowledge of bias and uncertainty can now be put together to analyze the reliability of a method in estimating the free energy, which in this case we will define as consistently lowest RMSE. The bivariate Gaussian plots (Figures 8−10) show the reliability of a method, by visualizing the RMSE using both bias estimates for each test case. For each figure, the first two columns contain bivariate RMSE plots for full λ state runs and the last two show the results of the sparse λ state runs.

Figure 8, with data for UA methane solvation, shows that MBAR, RBAR, BAR, and UBAR have small and equal spreads in both horizontal and vertical directions indicating that these are reliable estimates of free energy both for sparse as well as full λ sets. TI and TI3 give reliable estimates of free energy only in full λ state runs but are dominated by bias due to number of intermediate states with the sparse set. IEXP has lower RMSE compared to TI and TI3. GDEL and GINS are unreliable in both the full and sparse λ sets.

In Figure 9 for dipole inversion, free energy estimates from TI and TI3 are comparable in reliability with MBAR, BAR, RBAR, and UBAR. GDEL and GINS also give fairly accurate estimates of free energy for dipole inversion given the poor performance in other systems. This can be explained in the light of the work done by Hummer, Pratt, and Garcia on free energy of ionic hydration,[74] in which they found that the electrostatic potential energy distribution follows Gaussian behavior. Thus even for large changes in charges, Gaussian methods may still be a viable and useful method.

The anthracene solvation test set is a harder problem, and no method is as accurate as with the other two molecular cases, as seen in Figure 10. TI, TI3, IEXP, BAR, RBAR, UBAR, and MBAR perform equally well within noise for full λ set. In the sparse λ set, however, IEXP and UBAR become significantly worse than the other methods with TI being slightly worse. Improved performance of TI relative to acceptance ratio methods for the anthracene test set is because the sparse λ set for anthracene solvation case (four states between end points) is not as aggressive as the methane solvation case (only one state between end points). GDEL, GINS, and DEXP are unreliable estimators of anthracene hydration free energy for both the full and sparse λ sets, with both low accuracy and precision in their predicted free energy and uncertainty estimates.

5.4. Testing the Gaussian Distribution of Free Energies. Asymptotic error estimate methods assume normal distribution of error, as does the use of a standard deviation to describe the error distribution. It is important to test if this assumption is actually valid. Figure 11 demonstrates graphically the distribution

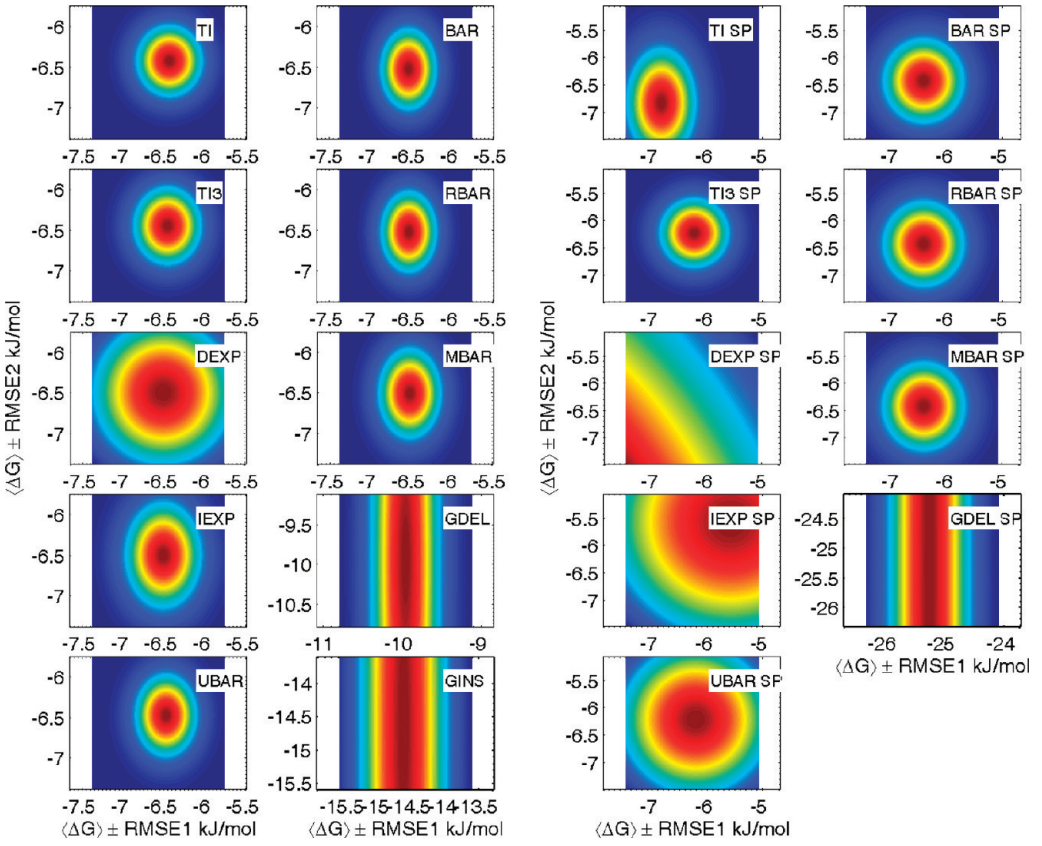

Figure 10. Bivariate Gaussian plots for anthracene solvation free energy. The effect of bias due to the number of intermediate states is evident in almost all methods as all spreads are elliptical in vertical direction. TI3 and MBAR both appear the most reliable, especially for the full λ set.

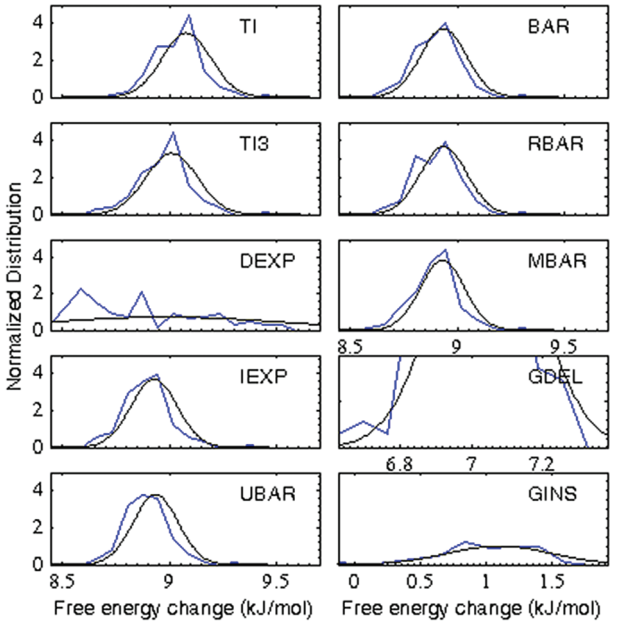

Figure 11. Subplots compare the distribution of estimated free energies from independent 100 repetitions (in blue) and Gaussian with the mean $\langle \Delta G \rangle$ and standard deviation $\langle \delta(\Delta G) \rangle$ from 100 independent repetitions (in black).

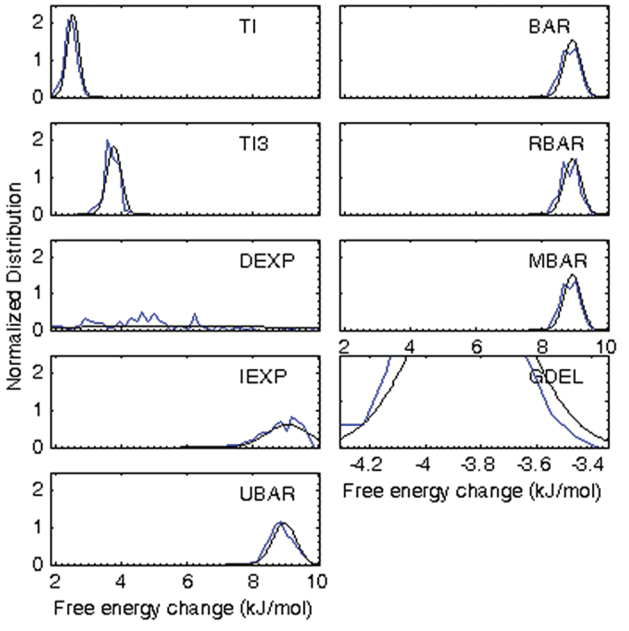

Figure 12. Each subplot compares the distribution of estimated free energies from 100 repetitions (in blue) and the Gaussian with the mean $\langle \Delta G \rangle$ and standard deviation $\langle \delta(\Delta G) \rangle$ from 100 repetitions (in black).

of free energy results for each test case, comparing it with a Gaussian with the mean and the variance of the distribution. The free energy distributions are plotted as histograms with the optimal bin width calculated from Scotts' formula,[75] with optimum bin width $h = 3.5\,\sigma/n^{(1/3)}$, where n is the number of samples

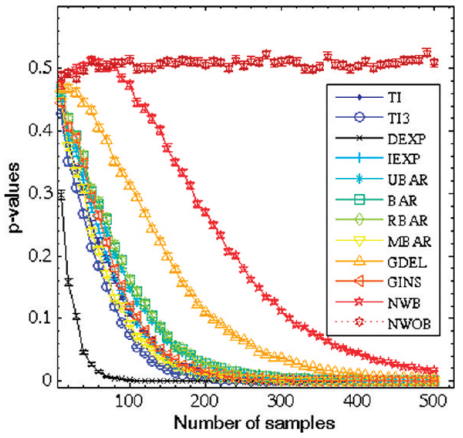

Figure 13. For methane solvation with full λ set, all p-value curves except DEXP are above 0.05 for 100 samples, resulting in the rejection of the null (normal distribution) hypothesis at the 95% confidence level for DEXP for as low as 40 samples.

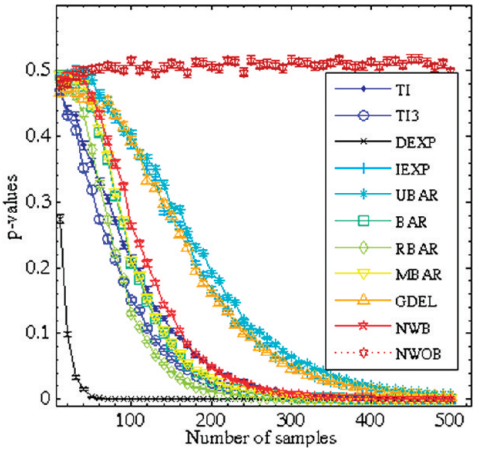

Figure 14. Methane solvation with the sparse λ set appears very similar to the full λ set case, with DEXP failing normality tests at 95% confidence level for as low as 25 samples. GINS is omitted because of a nonphysical (10^{32}) variance.

(i.e., 100), and σ is the standard deviation. For calculating the optimum bin width for the first eight methods, we have used the mean of the standard deviations from the 100 repetitions predicted by MBAR, though the variance is within a factor of 2−5 for all methods and only affects the qualitative visualization of the data. For GDEL and GINS, their own mean of predicted uncertainties is used to calculate the optimum bin width because uncertainties estimated with GDEL and GINS are drastically different from that of MBAR.

In Figure 11, the blue curve is the histogram of free energies, and the black curve is the Gaussian with the mean and the standard deviation estimated by the specified method. The shapes of the blue curve will match the black curve within noise if the distribution of free energies is indeed Gaussian, with the specified variance.

In Figures 11 and 12 and S1−S4 in the Supporting Information, we see that the distributions of free energies from all methods except DEXP well approximate the Gaussian, even when the variances are large. However, in several cases (GDEL and GINS)

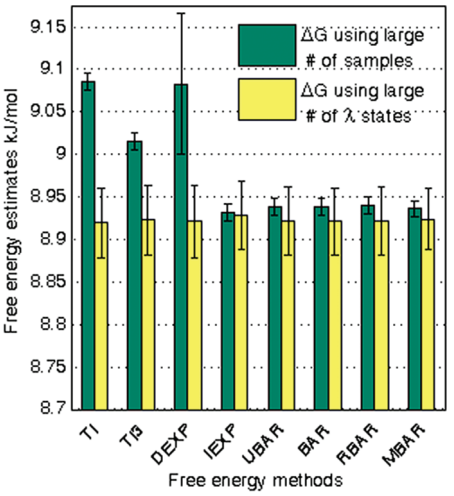

Figure 15. Free energies estimated using large number of intermediate states and a large number of samples converge to a single value for all free energy estimators.

the variances are so large that the comparisons become statistically meaningless. Besides DEXP, which fails in all cases IEXP and UBAR are sufficiently noisy for dipole inversion that the comparison to a Gaussian is problematic. For the sparse λ set IEXP and UBAR fail completely to be Gaussian. Similarly, for anthracene solvation IEXP and UBAR fail to be Gaussian for the sparse λ set. The 51 λ state results are omitted from GDEL and GINS plots as they lie outside the plot axis limits. Interestingly, we find that typically errors are distributed normally with the variance given by the analytical estimate even if the bias in the free energy is noticeable.

This visual analysis is only qualitative, and it is useful to also have quantitative tests of normality. We use the Shapiro−Wilk test for normality with an unknown mean and variance[73] to test the null hypothesis that a sample of free energies came from a normal population. The null hypothesis is rejected if the p-value is less than the specified probability. In this case, we choose a cutoff p-value of 0.05, corresponding to 95% confidence intervals. Importantly, the p-value depends on the number of samples; if a distribution is even the slightest different than a true Gaussian, with enough samples, the null hypothesis will always be rejected. Therefore for a continuous measure of normality, we calculate p-values as a function of number of samples and find the minimum number of samples on average that are required to reject the null hypothesis.

We randomly pick N ($10 < N < 500$) samples 2000 times with replacement from the original ensemble of 100 free energies, creating 2000 bootstrap replicas. For each bootstrap replica, we calculate the p-value. We then calculate the mean of the p-value over all the bootstrap replicas. This mean serves as an estimate of the p-value corresponding to N completely independent samples. Errors in the p-value are computed as the standard deviation over the 2000 bootstrap replicas. We use two separate controls for each method for this experiment. First, we generate N samples 2000 times from the Gaussian distribution with mean and variance from MBAR and calculate the corresponding p-values. Then to test the effect of bootstrapping from a finite distribution, we generate 100 normally distributed free energies, and then from this distribution, we randomly pick N samples 2000 times to calculate bootstrap replicas and again calculate p-values estimated as a

function of the number of samples. The control experiment using samples from a normal distribution with no bootstrap is abbreviated as NWOB, and the one which uses bootstrap is abbreviated as NWB.

Figures 13 and 14 and S5−S8 in the Supporting Information show average p-values as a function of number of samples pulled from the distribution for the 10 methods and two control cases for the three test systems. NWOB has the expected p-value = 0.5 curve, as the distribution is exactly Gaussian. NWB and all 10 methods have a decaying p-value curve with increasing numbers of samples. Decreasing p-value indicates increasing deviation from normality and increasing probability of rejecting the normality hypothesis. For NWB, although samples come from a known normal distribution, the null hypothesis is rejected at 95% confidence level when the number of samples is between 110 and 270 samples depending on the test set, indicating an approximate upper level of normality that can be observed with a finite bootstrapped set.

Examining Figures 13 and 14 and S5−S8 in the Supporting Information, we see that in almost all cases, for all methods other than DEXP, we would generally need more than 90 samples to reject the hypothesis that the samples were normal at a 95% confidence level, which should be sufficient for all general purposes. In all but the dipole inversion case, the null hypothesis of normality is rejected at the 95% confidence level for DEXP as low as 40 samples, which can be seen qualitatively in the distribution graphs in Figures 11 and 12 and S1−S4 in the Supporting Information. These results indicate for all methods except for DEXP, there are essentially no long tails or "black swans", and the error estimates can be assumed to be Gaussians for all practical purposes. We note that variances that are known to be too large might require more samples to demonstrate that they are definitively not normal, but the accurate prediction of the magnitude of the variance is not measured by this statistical test, and the effects show up only in the large (more than 100) sample regime.

5.5. Convergence Properties. The true free energy estimate is not necessarily the experimental value of the free energy change of the process, but instead the infinite sampling limit of the particular choice of molecular model. We see that with a large number of intermediate states, all methods converge to the same value (Figure 15), whereas the 450 ns results with the sparse λ data set vary for different methods, with as usual large deviations seen in GDEL and GINS, indicating significant bias with respect to the overlap between states. Given sufficient sampling, increasing the number of λ states appears to be the best way to obtain asymptotic convergence to the true answer for most molecular transformations that have converged sampling.

5.6. Amount of Time Required for Free Energy Estimation Methods. We chose the calculation of anthracene solvation over 4.5 ns with the full λ set as our test system to compare computational time required by methods to compute free energies, because the system has the largest number of intermediate states with large free energy changes between states. We report the time required to calculate the free energies 201 times (for the original set and for 200 bootstrap sets) to eliminate variability caused by computational overhead in single calculations. The time required to read in data, make bootstrap samples, and perform bookkeeping required by all methods was 249.5 s was subtracted from the total time to yield the analysis time of each method. Time, i.e., time required for generating samples is also not included in the analysis, as it is same for all methods.

Table 6. Analysis Time for Different Free Energy Methods[a]

method	time taken (s)
TI	0.2
TI3	5.8
DEXP	13.5
IEXP	11.5
UBAR	15.3
BAR	93.5
RBAR	1148.2
MBAR	4913.5
GDEL	4.0
GINS	4.3

[a] Times are for 201 repetitions of a 4.5 ns dataset for anthracene solvation with 15 states. Time to generate samples is not included.

Table 7. Summary of All Statistical Tests

Reliability (accuracy plus bias) of free energy estimate (high to low)
MBAR > BAR = RBAR > UBAR > TI3 > TI > IEXP > DEXP > GDEL = GINS
Reliability of uncertainty estimate (high to low)
MBAR > TI3 = TI > BAR = RBAR > UBAR > IEXP > DEXP > GDEL = GINS
Computational cost of analysis of data (high to low)
MBAR > RBAR > BAR > UBAR > IEXP = DEXP > GDEL = GINS > TI3 > TI
Is distribution of estimated free energies Gaussian?
Yes for [MBAR, RBAR, BAR, UBAR, TI, TI3, IEXP, GDEL, GINS].
No for [DEXP]
Is bootstrap better than analytical uncertainty estimate?
Yes for [RBAR, BAR, UBAR, GDEL, GINS].
Both equally good for [MBAR,IEXP,DEXP,TI,TI3]

From Table 6 it is evident that MBAR takes the longest of all methods as it processes information from all the intermediate states to give an estimate of free energy and uncertainty. RBAR is the next most computationally costly. RBAR takes more time compared to BAR because multiple BAR calculations are performed over a range of free energies at each intermediate stage if a large range of possible values for the self-consistent constants are evaluated. UBAR takes less time compared to BAR because only a single iteration is performed at each stage. DEXP and IEXP are similar in cost UBAR. GDEL and GINS take less time compared to DEXP and IEXP. TI3 takes slightly longer compared to TI as it fits a spline of higher degree, but both are much cheaper than any others. However, the total time required even by MBAR is orders of magnitude less than the time required to perform the sampling, so the higher cost of MBAR is not an obstacle in most cases. Note that the time includes bootstrapping of 200 samples, so in the case of MBAR, where analytical variance is sufficient, the single calculation time is only 24 seconds.

6. CONCLUSIONS

In this paper we have proposed the first iteration of a set of test sets which can be used for benchmarking free energy calculation methods for small molecule solvation. We have demonstrated the utility of this test set by comparing 10 equilibrium free energy methods on 3 test cases for molecular solvation, with different spacing between intermediate states. We estimated the uncertainty in three different ways: the sample standard deviation, the

analytical estimate, and the bootstrap estimate as well as the uncertainty in each of these estimates of the uncertainty. We also calculated biases in free energy estimates at the large number of samples limit and large number of intermediate states limit separately. We graphically demonstrated the effect of the variance and two separate types of bias by bivariate Gaussian plots expressing the overall reliability of the methods. We demonstrated that bootstrap sampling accurately predicts the properties of the sample distribution observed from 100 independent simulations for all the free energy methods. We find that all uncertainty estimates for all but the worst performing methods are highly consistent, with relative errors of only 5% to 10%, and thus can be generally expected to be consistent from sample to sample. We also showed that the histogram of free energies from 100 independent simulations has a Gaussian form for TI, TI3, BAR, RBAR, UBAR, MBAR, GDEL, and GINS, but that DEXP and sometimes IEXP deviate from Gaussian distributions.

We have found that MBAR is the most reliable of all free energy estimators, showing consistency in accuracy and precision in both free energy and uncertainty prediction. TI and TI3 are better uncertainty estimators compared to BAR, UBAR, and RBAR, with equal performance to MBAR when sufficient intermediate states are included but are biased with respect to the number of intermediate states if such care is not taken. It is likely that this bias can be reduced with the judicious use of application-specific integration schemes, though the variance will not be reduced, since the uncertainties for the $\partial U/\partial\lambda$ observables do not change in alternative choices of the integration weights. When the $\partial U/\partial\lambda$ vs λ curve has low curvature, such as in dipole inversion, both TI and TI3 are equally reliable. But when the curve is nonlinear, i.e., when LJ spheres grow or disappear, TI3 gives better estimates of free energy than TI. BAR and RBAR have relatively negligible bias, but their uncertainty estimates are frequently underestimated by 25% to 30% when overlap between states is not negligible. UBAR is often as good as BAR and RBAR but can fail with low numbers of intermediate states. IEXP and DEXP are less reliable than TI and acceptance ratio methods and should be avoided if samples can be collected from all intermediates in both the forward or back direction or if the derivative of the Hamiltonian along the pathway can be computed. IEXP does works in some cases, but in general, IEXP and DEXP give poor estimates for uncertainty and free energies. GINS and GDEL do not compare well with the other methods in all the test systems except dipole inversion test case. They only work if there is large number of intermediate states or if the distributions are inherently Gaussian. However, even here they are not as accurate or precise as the other methods.

MBAR is the most expensive, but the amount of time required for analysis is orders of magnitude less than the time required for collecting data. UBAR takes less time compared to RBAR and BAR and should only be considered as a quick and easy (but not so reliable) alternative to BAR and RBAR. RBAR requires some knowledge of the maximum free energy gap but does not require storing all the energy data. BAR, like MBAR, requires storing energy data for later analysis but requires no knowledge about the size of the maximum free energy difference. GDEL and GINS take less time compared to DEXP and IEXP but they heavily sacrifice accuracy and precision for speed in virtually all cases. Finally, TI takes least time to estimate free energies and uncertainties but a little extra computation (fitting cubic splines) in TI3 improves the accuracy in free energy estimate using TI. We summarize these conclusions in Table 7.

Availability of the Benchmark Data Set for Multiple Simulation Packages. The benchmark test set is available for distribution and use at http://www.alchemistry.org. It contains starting configurations and parameter files for all 100 uncorrelated starting configurations in three formats, corresponding to three different molecular dynamics packages, GROMACS (*.gro, *.top, and *.mdp), AMBER (*.inpcrd, *.prmtop, and *.in), and DESMOND (*.cms and *.cfg) as well as detailed instructions for the test set's use. To ensure that the parameter files are correctly constructed, we have calculated the single point energies of the 100 structures using the input files and the corresponding MD package. We invite users of other simulation packages to contact us in order to add energy comparisons to these simulation packages.

This is intended to be the first version of the benchmark test set. It is not comprehensive and will require further expansion to be more useful to a wider range of researchers. We hope that these data sets will be useful for other researchers to compare other free energy estimators, such as alternative TI schemes, nonequilibrium free energy methods, and other more novel methods.

The intention is for future versions of this benchmark test set to be developed in response to feedback and as resources become available. Future versions of the benchmark will ideally include model molecules with long correlation time internal motion. A model system for the transformation of bonded terms of would also be useful, though these transformations may have sufficiently short correlation times to make the differences in efficiency less relevant. It is also sometimes useful to compute molecular potentials of mean force in place of molecular transformations. Variance is directly related to phase space overlap, so the patterns discovered for alchemical transformations should also hold for construction of PMF. However a test system to validate this hypothesis will be useful. Finally, it would be ideal to eventually test the efficiency of methods to calculate free energies of well-studied ligand-binding systems and increasingly tractable protein−ligand binding systems, such as T4 lysozyme. We look forward to collaborating with other researchers to further develop the benchmark set and will post results and comparisons with other methods and between different codes on the http://alchemistry.org Web site along with the benchmark set data.

APPENDIX: THERMODYNAMIC INTEGRATION USING CUBIC SPLINES

We fit the $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ vs λ_i curve piecewise to a series of cubic polynomials $S_i(\lambda)$:

$$S_i(\lambda) = a_i + b_i(\lambda - \lambda_i) + c_i(\lambda - \lambda_i)^2 + d_i(\lambda - \lambda_i)^3 \quad \forall\ 1 \le i \le K-1 \tag{A1}$$

Here K is the total number of intermediate states, creating $K-1$ intervals for splining. Each spline of a given interval has its own set of coefficients a_i, b_i, c_i, and d_i, which can be computed by standard linear algebra methods using the conditions that define a natural cubic spline.

$$S_i(\lambda_i) = a_i = \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_i}\rangle \quad \forall\ 1 \le i \le K-1 \tag{A2}$$

$$S_i(\lambda_{i+1}) = \langle(\partial U(\lambda)/\partial\lambda)_{\lambda_{i+1}}\rangle \quad \forall\ 1 \le i \le K-1 \tag{A3}$$

$$S_i'(\lambda_{i+1}) = S_{i+1}'(\lambda_{i+1}) \quad \forall\ 1 \le i \le K-2 \tag{A4}$$

$$S_i''(\lambda_{i+1}) = S_{i+1}''(\lambda_{i+1}) \quad \forall\ 1 \le i \le K-2 \tag{A5}$$

$$S_1''(\lambda_1) = 0 \tag{A6}$$

$$S_{K-1}''(\lambda_K) = 0 \tag{A7}$$

When we integrate piece wise over all the intervals, we get the total free energy change:

$$\Delta G = \sum_{i=1}^{K-1} \int_{i}^{i+1} \mathrm{d}\lambda S_i(\lambda) \tag{A8}$$

$$\Delta G = \sum_{i=1}^{K-1} a_i(\lambda_{i+1} - \lambda_i) + \frac{b_i}{2}(\lambda_{i+1} - \lambda_i)^2 + \frac{c_i}{3}(\lambda_{i+1} - \lambda_i)^3 + \frac{d_i}{4}(\lambda_{i+1} - \lambda_i)^4 \tag{A9}$$

We need to write ΔG in the form described in eq 3, as a weighted sum of $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ at each λ, so that we can propagate the error using eq 4. We must solve for the coefficients a_i, b_i, c_i, and d_i in such a way such that they be expressed as a linear weighted sum of individual $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$.

$$\begin{bmatrix} a_1 \\ \vdots \\ a_{K-1} \end{bmatrix} = \begin{bmatrix} A_{1,1} & \cdots & A_{1,K} \\ \vdots & \ddots & \vdots \\ A_{K-1,1} & \cdots & A_{K-1,K} \end{bmatrix} \begin{bmatrix} \langle(\partial U(\lambda)/\partial\lambda)_1\rangle \\ \vdots \\ \langle(\partial U(\lambda)/\partial\lambda)_K\rangle \end{bmatrix} \tag{A10}$$

Here A_{ij} are the weights in a weight matrix for a_i. Similarly b_i, c_i, and d_i are expressed as linear weighted sums of $\langle(\partial U(\lambda)/\partial\lambda)_{\lambda i}\rangle$ with B_{ij}, C_{ij}, and D_{ij} as the weights in the respective $K-1 \times K$ matrices. There exist a unique solution for a, b, c, and d. The A, B, C, and D matrices are all of rank $K-1$ and are invertible.

We can then finally combine these into a single weight matrix. Defining $h_i = \lambda_{i+1} - \lambda_i$, then eq A9 can be written as a linear weighted sum:

$$\Delta G = \sum_{i=1}^{K-1}\sum_{j=1}^{K}\left(h_i A_{ij} + \frac{h_i^2}{2}B_{ij} + \frac{h_i^3}{3}C_{ij} + \frac{h_i^4}{4}D_{ij}\right)(\langle(\partial U(\lambda)/\partial\lambda)_j\rangle) \tag{A11}$$

Equation A11 can be further written as:

$$\Delta G = \sum_{i=1}^{K-1}\sum_{j=1}^{K}(W_{ij})(\langle(\partial U(\lambda)/\partial\lambda)_j\rangle) \tag{A12}$$

Here

$$W_{ij} = h_i A_{ij} + \frac{h_i^2}{2}B_{ij} + \frac{h_i^3}{3}C_{ij} + \frac{h_i^4}{4}D_{ij} \tag{A13}$$

Once we have the weights, we can calculate the overall free energy change using eq A12 and the uncertainty estimate using the following equation.

$$\sigma_{10}^2 = \sum_{i=1}^{K-1}\sum_{j=1}^{K} W_{ij}^2\sigma_j^2 \tag{A14}$$

An implementation of this weighting for TI using GROMACS is included in the examples section of the pymbar distribution at https://simtk.org/home/pymbar.

■ ASSOCIATED CONTENT

Supporting Information. Figures S1−S8 and Tables S1−S10 referred to in the text. It also contains the current version of files described in the section "Availability of the benchmark data set for multiple simulation packages." The most recent version of the benchmark can be found at www.alchemistry.org. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

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

■ ACKNOWLEDGMENT

The authors acknowledge support for the Oak Ridge Associated Universities Ralph E. Powe, Jr. Faculty Enhancement Award. We also would like to acknowledge significant help from James Watney at D. E. Shaw Research for help in running Desmond.

■ REFERENCES

(1) Davies, J. W.; Glick, M.; Jenkins, J. L. *Curr. Opin. Chem. Biol.* **2006**, *10*, 343–351.
(2) Merz, K. M.; Ringe, D.; Reynolds, C. H. *Drug Design: Structure- and Ligand-Based Approaches*; Cambridge University Press: New York, 2010.
(3) Mobley, D.; Dill, K. *Structure* **2009**, *17*, 489–498.
(4) Bai, H.; Lai, L. *Acta Physicochim. Sin.* **2010**, *26*, 1988–1997.
(5) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.
(6) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.
(7) Zamolo, L.; Salvalaglio, M.; Cavallotti, C.; Galarza, B.; Sadler, C.; Williams, S.; Hofer, S.; Horak, J.; Lindner, W. *J. Phys. Chem. B* **2010**, *114*, 9367–9380.
(8) Duren, T.; Bae, Y.; Snurr, R. *Chem. Soc. Rev.* **2009**, *38*, 1237–1247.
(9) Nel, A. E.; Madler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. *Nat. Mater.* **2009**, *8*, 543–557.
(10) Liong, M.; Lu, J.; Kovochich, M.; Xia, T.; Ruehm, S. G.; Nel, A. E.; Tamanoi, F.; Zink, J. I. *ACS Nano* **2008**, *2*, 889–896.
(11) Shi, X.; Wang, S.; Shen, M.; Antwerp, M.; Chen, X.; Li, C.; Petersen, E.; Huang, Q.; Weber, W.; Baker, J. *Biomacromolecules* **2009**, *10*, 1744–1750.
(12) Nishide, H. *Advanced Nanomaterials*; Wiley-VCH: Hoboken, NJ, 2010.
(13) Yuan, H.; Zhang, S. *Appl. Phys. Lett.* **2010**, *96*, 033704.
(14) Fisher, E. H.; Rhodes, N. *Proc. Inst. Mech. Eng., Part C* **1996**, *210*, 91–94.
(15) Roache, P. J. *Annu. Rev. Fluid Mech.* **1997**, *29*, 123–160.
(16) Oberkampf, W. L.; Trucano, T. G. *Prog. Aerosp. Sci.* **2002**, *38*, 209–272.
(17) Johnson, F.; Tinoco, E.; Yu, N. *Comput. Fluids.* **2005**, *34*, 1115–1151.
(18) Kellar, W.; Savill, A.; Dawes, W. *Proc. High Perform. Comput. Networking* **1999**, *1593*, 90–98.
(19) del Álamo, J. C.; Marsden, A. L.; Lasherasa, J. C. *Rev. Esp. Cardiol.* **2009**, *62*, 781–805.
(20) Botar, C. C.; Vasile, T.; Sfrangeu, S.; Clichici, S.; Agachi, P. S.; Badea, R.; Mircea, P.; Cristea, M. V. In *20th European Symposium on Computer Aided Process Engineering*; Elsevier: Waltham, MA, 2010; Vol. 28, pp 205−210.
(21) Sampson, B. *Prof. Eng.* **2007**, *20*, 37–37.
(22) Oberkampf, W.; Trucano, T. *Nucl. Eng. Des.* **2008**, *238*, 716–743.

(23) Johnson, R. D., III CCCBDB Computational Chemistry Comparison and Benchmark Database, *NIST Standard Reference Database Number 101*, release 15b, August 2011. http://cccbdb.nist.gov/.
(24) Maginn, E. J.; Elliott, J. R. *Ind. Eng. Chem. Res.* **2010**, *49*, 3059–3078.
(25) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300.
(26) Bruckner, S.; Boresch, S. *J. Comput. Chem.* **2011**, *32*, 1320–1333.
(27) Zwanzig, R. W. *J. Chem. Phys.* **1955**, *23*, 1915.
(28) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 144107.
(29) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245–268.
(30) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
(31) Ytreberg, F. M.; Swendsen, R. H.; Zuckerman, D. M. *J. Chem. Phys.* **2006**, *125*, 184114.
(32) These numbers were generated by adding the primary citations for each of the listed methods (23−30) and searches for thermodynamic integration (TI) and free energy perturbation in ISI Web of Science, as the original papers for these methods are older than the ISI database.
(33) Ytreberg, F. M.; Zuckerman, D. M. *J. Chem. Phys.* **2004**, *120*, 10876.
(34) Lin, C.; Wood, R. H. *J. Comput. Chem.* **1994**, *15*, 149–154.
(35) Radmer, R.; Kollman, P. *J. Comput. Chem.* **1997**, *18*, 902–919.
(36) Lelievre, T.; Stoltz, G.; Rousset, M. *Free Energy Computations: A Mathematical Perspective*; 1st ed.; Imperial College Press: London, 2010.
(37) Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. *Biopolymers* **2003**, *68*, 16–34.
(38) Meirovitch, H.; Cheluvaraja, S.; White, R. P. *Curr. Protein Pept. Sci.* **2009**, *10*, 229–243.
(39) Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, *6*, 2559–2565.
(40) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227–20232.
(41) Cencek, W.; Szalewicz, K. *Int. J. Quantum Chem.* **2008**, *108*, 2191–2198.
(42) Gurkan, B.; Goodrich, B.; Mindrup, E.; Ficke, L.; Massel, M.; Seo, S.; Senftle, T.; Wu, H.; Glaser, M.; Shah, J.; Maginn, E.; Brennecke, J.; Schneider, W. *J. Phys. Chem. Lett.* **2010**, *1*, 3494–3499.
(43) Oden, J. T.; Belytschko, T.; Fish, J.; Hughes, T. J. R.; Johnson, C.; Keyes, D.; Laub, A.; Petzold, L.; Srolovitz, D.; Yip, S. *A Report of the National Science Foundation Blue Ribbon Panel on Simulation-Based Engineering Science: Revolutionizing Engineering Science through Simulation*; National Science Foundation: Arlington, VA, 2006.
(44) Glotzer, S.; Kim, S.; Cummings, P.; Deshmukh, A.; Head-Gordon, M.; Karniadakis, G.; Petzold, L.; Sagui, C.; Shinozuka, M. *WTEC Panel Report on International Assessment of Research and Development in Simulation-Based Engineering and Science*; World Technology Evaluation Center, Inc.: Baltimore, Maryland, 2009.
(45) Pohorille, A.; Jarzynski, C.; Chipot, C. *J. Phys. Chem. B* **2010**, 1420–1426.
(46) Radmer, R.; Kollman, P. *J. Comput. Chem.* **1997**, *18*, 902–919.
(47) Jorge, M.; Garrido, N. M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. *J. Chem. Theory Comput.* **2010**, *6*, 1018–1027.
(48) Kofke, D. A.; Cummings, P. T. *Mol. Phys.* **1997**, *92*, 973–996.
(49) Hummer, G. *J. Chem. Phys.* **2001**, *114*, 7330.
(50) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; 1st ed.; Chapman and Hall/CRC: Norman, OK, 1994; Vol. 5.
(51) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.
(52) Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856.
(53) Pitera, J.; Van Gunsteren, W. *Mol. Simulat.* **2002**, *28*, 45–65.
(54) Martínez-Veracoechea, F. J.; Escobedo, F. A. *J. Phys. Chem. B.* **2008**, *112*, 8120–8128.
(55) Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.
(56) Jarzynski, C. *Phys. Rev. E* **1997**, *56*, 5018.
(57) Resat, H.; Mezei, M. *J. Chem. Phys.* **1993**, *99*, 6052.
(58) Mark, A. E.; van Helden, S. P.; Smith, P. E.; Janssen, L. H. M.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2011**, *116*, 6293–6302.
(59) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 054103.
(60) Barker, A. *Aust. J. Phys.* **1965**, *18*, 119–134.
(61) Fenwick, M. K.; Escobedo, F. A. *J. Chem. Phys.* **2004**, *120*, 3066.
(62) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.
(63) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635.
(64) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40–57.
(65) Schuttelkopf, A.; van Aalten, D. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 1355–1363.
(66) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.
(67) Xiao, Y.; Li, D. *Math. Meth. Oper. Res.* **2008**, *67*, 443–454.
(68) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, new ed.; Clarendon Press: Oxford, U.K., 1989.
(69) Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512–518.
(70) Hess, B. *J. Chem. Phys.* **2002**, *116*, 209.
(71) Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L. *Mol. Phys.* **1996**, *87*, 1117.
(72) Kofke, D. *Mol. Phys.* **2006**, *104*, 3701–3708.
(73) Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591–611.
(74) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.
(75) Scott, D. W. *Biometrika* **1979**, *66*, 605–610.