Assessment of Nonequilibrium Free Energy Methods

Benjamin P. Cossins,[†] Sebastien Foucher,[†] Colin M. Edge,[‡] and Jonathan W. Essex*,[†]

School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K., and GlaxoSmithKline, New Frontiers Science Park, Coldharbour Road, Harlow, CM19 5AD, U.K.

Received: April 23, 2008; Revised Manuscript Received: December 19, 2008

One of the factors preventing the general application of free energy methods in rational drug design remains the lack of sufficient computational resources. Many nonequilibrium (NE) free energy methods, however, are easily made embarrassingly parallel in comparison to equilibrium methods and may be conveniently run on desktop computers using distributed computing software. In recent years, there has been a proliferation of NE methods, but the general applicability of these approaches has not been determined. In this study, a subset including only those NE methods which are easily parallelised were considered for examination, with a view to their application to the prediction of protein−ligand binding affinities. A number of test systems were examined, including harmonic oscillator (HO) systems and the calculation of relative free energies of hydration of water−methane. The latter system uses identical potentials to the protein ligand case and is therefore an appropriate model system on which methods may be tested. As well as investigating existing protocols, a replica exchange NE approach was developed, which was found to offer advantages over conventional methods. It was found that Rosenbluth-based approaches to optimizing the NE work values used in NE free energy estimates were not consistent in the improvements in accuracy achieved and that, given their computational cost, the simple approach of taking each work value in an unbiased way is to be preferred. Of the two free energy estimators examined, Bennett's acceptance ratio was the most consistent and is, therefore, to be preferred over the Jarzynski estimator. The recommended protocols may be run very efficiently within a distributed computing environment and are of similar accuracy and precision to equilibrium free energy methods.

1. Introduction

In recent years, free energy calculations have become more applicable to rational drug design due to the availability of high performance computing.[1−5] However, these examples are still very system specific and generally limited to groups of ligands which have a common structure (congeneric series of ligands). While approaches are being developed to simulate more diverse ligands,[6−8] the computational cost of these equilibrium free energy calculations remains a handicap.

There have been many studies focused on nonequilibrium (NE) free energy methods most of which are related to the Jarzynski equality (JAR).[9−11] Some of these NE methods have the potential for large-scale embarrassing parallelization of calculations.

NE methods such as that described by Jarzynski employ many independent processes to give an estimate of the free energy difference. These independent processes can be calculated on completely independent processors. This allows for Jarzynski-like NE calculations to be run on large inhomogeneous distributed computers made up of partially used desktops over a wide physical area. Thus, these methods may be able to more easily and cheaply offer large amounts of simulation sampling than established equilibrium methods such as thermodynamic integration (TI) and thus provide more widely applicable calculations with more precise results.

Here, we have investigated many of these NE free energy methods and present work on a subset which we found to offer possible advantages over current equilibrium methods and which are easily applicable to large scale parallelization. We use various combinations of NE methods in comparison with established equilibrium TI approaches on harmonic oscillator (HO) and methane−water test systems. This study is designed to find the most efficient and reliable combination of easily parallelizable NE methods and to assess their utility for use in protein−ligand binding energy calculations. The combination of NE methods found to perform best in this study is compared elsewhere to replica exchange thermodynamic integration (RETI) simulations with two sets of congeneric inhibitors for the enzymes neuraminidase and cyclooxygenase-2.[12]

2. Free Energy Methods

Here, we describe the free energy methods used in this study. The "Equilibrium Methods" section describes the established equilibrium free energy methods which are in general, current use. The "Nonequilibrium Works" section describes the basis of nonequilibrium (NE) free energy methods and how measurements are taken. The "Nonequilibrium Estimators" section discusses NE estimators; these methods take NE measurements and calculate an equilibrium free energy. Hence, a particular NE method can use only one estimator. In this study, results for JAR and Bennett's acceptance ratio (BAR) estimators are presented, although in the process of this work other estimators such as those discussed by Hummer[13] have been investigated but were not included due to perceived problems not discussed here. The "Nonequilibrium Sampling Methods" section discusses NE sampling methods; these are methods which are used in the production of the NE measurements subsequently used by the NE estimators. The "Nonequilibrium Bias Detection"

* Electronic mail: J.W.Essex@soton.ac.uk.
[†] University of Southampton.
[‡] GlaxoSmithKline.

10.1021/jp803532z CCC: $40.75
Published on Web 03/26/2009

section describes methods which attempt to detect error in the form of systematic bias in NE calculations. Many methods which attempt to calculate and detect NE bias were investigated, but only a most useful subset is displayed here. Finally, the "Combining NE Methods" section describes how statistical errors were calculated in this study.

2.1. Equilibrium Methods. ***Free Energy Perturbation.*** It has long been known that the free energy difference (ΔF) between two systems A and B can be found through Zwanzig's equation,

$$\Delta F = F_{\mathrm{B}} - F_{\mathrm{A}} = -1/\beta \ln\langle \exp\{-\beta \Delta U_{\mathrm{AB}}(q)\}\rangle_{\mathrm{A}} \quad (1)$$

where $\langle ...\rangle_{\mathrm{A}}$ denotes an ensemble average over system A, ΔU_{AB}, the difference in the potential energies of A and B for the present set of coordinates, (q), and $\beta = 1/kT$.[14] The computational implementation of Zwanzig's equation is called free energy perturbation (FEP). FEP in its simplest form entails running a simulation of one system (A or B) which at each simulation step adds to the average in eq 1.

Often a series of intermediate systems between A and B are used to ensure good phase space overlap. This is achieved by coupling the differences between systems A and B to a simulation parameter, λ, where $\lambda = 0$ gives system A and $\lambda = 1$ gives system B.

Thermodynamic Integration. Thermodynamic integration (TI) is a well-established rigorous free energy method and is well represented in many texts.[15,16] TI is based on the coupling parameter λ, described above for FEP. Simulations are run at values of λ which allow good phase space overlap from systems A to B. The property accumulated by each simulation is the free energy gradient $(\partial F/\partial\lambda)_\lambda$. ΔF from A to B is then found by integrating over the measured gradients:

$$\Delta F = \int_0^1 \left(\frac{\partial F}{\partial \lambda}\right)_\lambda \mathrm{d}\lambda \quad (2)$$

The free energy gradients can be approximated numerically by the finite difference as in eq 3. TI which uses a finite difference approximation is called finite difference thermodynamic integration (FDTI)[17] and will be used in this study over other forms of TI due to its simplicity.

$$\left(\frac{\partial F}{\partial \lambda}\right)_\lambda = \left(\frac{\Delta F}{\Delta \lambda}\right)_\lambda \quad (3)$$

ΔF in eq 3 can be found using the Zwanzig equation[14] and potential values at λ and $\lambda + \Delta\lambda$. The size of the $\Delta\lambda$ increment made to find a gradient measurement, in FDTI, must be small such that the exact gradient at the correct point is obtained.

Replica Exchange TI. Replica exchange thermodynamic integration (RETI) is a development of TI which incorporates Hamiltonian replica exchange moves between adjacent λ simulations (λ moves).[18,19] λ moves are made periodically and in such a way that each configuration is exchanged with a neighbor. In order that λ moves adhere to detailed balance, they are accepted or rejected with the equivalent of two metropolis tests: one for each configuration introduced to a new simulation. Thus, moves are accepted if,

$$\exp\{\beta[U_{\mathrm{B}}(j) - U_{\mathrm{B}}(i) - U_{\mathrm{A}}(j) + U_{\mathrm{A}}(i)]\} \geq \mathrm{rand}(0,1) \quad (4)$$

is true, where i and j are configurations being exchanged and A and B are the Hamiltonians of the replicas exchanging.

RETI increases sampling especially of the solvent by providing the possibility of ensembles making large jumps in phase space. Also, as simulations are able to move freely across λ, configurations which are more favorable to a particular area of λ may migrate there.

2.2. Nonequilibrium Works. The NE free energy methods discussed here utilize a basic rule of thermodynamics,

$$\Delta F = W_\infty \quad (5)$$

Equation 5 states that over the course of a reversible, isothermal process linking two equilibrium states, the work (W) performed on the system is equal to the free energy difference between the two states. For a process linking two states, also known as a switch, to be truly reversible, in principle, it must be infinitely long. For this reason, switches cannot be truly reversible. The accuracy of eq 5 relies on how close the simulated switch is to the reversible limit.

Unlike the measurement of $\Delta U_{\mathrm{AB}}(q)$ used in a FEP calculation (eq 1), NE methods do not use the same system configuration for systems A and B. Instead, λ is incremented n_{inc} times (where n_{inc} is a constant for the calculation) from 0 to 1, with simulation sampling allowed between increments, and the work performed as a consequence of each λ increment is summed to give W, as follows,

$$W = \sum_{i=1}^{n_{\mathrm{inc}}} U(q_{t_i})_{\lambda_{i+1}} - U(q_{t_i})_{\lambda_i} \quad (6)$$

Estimation of ΔF through equating ΔF with W using eq 6 (where (q_{t_i}) denotes the set of coordinates at the time of λ increment i)[20] invariably produces a systematic error due to the nonequilibrium nature of the perturbation process. The simulation lags behind the changing potential; this is often referred to as Hamiltonian lag.[21] Hamiltonian lag contributes positively to W such that, $\Delta F = W - W_{\mathrm{dis}}$. This contribution is called the dissipated work (W_{dis}); its average is positive and is associated with the increase of entropy during an irreversible process.[22]

2.3. Nonequilibrium Estimators. ***Jarzynski Equality.*** The Jarzynski equality[22] is given in eq 7:

$$\Delta F = -1/\beta \ln\langle \exp\{-\beta W\}\rangle \quad (7)$$

In practice, the number of switches needed to produce an accurate ΔF estimate with eq 7 varies and depends on the nature of the distribution of work values produced by the NE calculation. When close to equilibrium, all W distributions will be Gaussian, as is the W distribution represented by the solid line in Figure 1. The exponential average of the Jarzynski equality can be written as a integral over the distribution of work values,

$$\langle \exp(-\beta W)\rangle = \int \mathrm{d}W\, p(W) \exp(-\beta W) \quad (8)$$

The integrand distribution of eq 8 labeled $p(W)\mathrm{e}^{-\beta W}$ in Figure 1 is the weight of a particular work value in the exponential average $\mathrm{e}^{-\beta W}$ and is peaked to the left of the average work. The work values which contribute most to the right-hand side of eq 7 are those in the peak of $p(W)\mathrm{e}^{-\beta W}$ and in the far left-hand tail

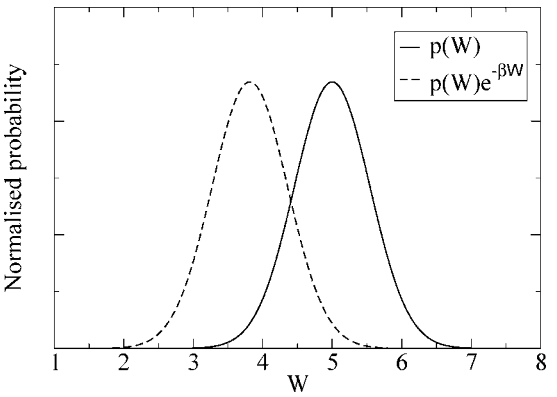

Figure 1. Depiction of a Gaussian distribution of NE work values.

of the work distribution $p(W)$. These important work values are thus produced by rare switches. We can increase the prevalence of these rare, important switches by minimizing the $\bar{W}_{\text{dis}}$ defined as $\bar{W}_{\text{dis}} = \bar{W} - \Delta F$, where $\bar{W}$ is the peak of distribution $p(W)$ in Figure 1.

Bennett's Acceptance Ratio. An interesting development of this switching methodology is the use of Bennett's acceptance ratio (BAR)[23] with switches in both forward (A → B) and backward (B → A) directions used to calculate ΔF.[24,25] Here ΔF is estimated: starting with the lower bound estimate of ΔF, found through the average work, the ΔF estimate is then increased slowly and iteratively until ΔF satisfies,

$$\sum_{i=1}^{n_{\text{F}}}\left(1 + \frac{n_{\text{F}}}{n_{\text{R}}}\exp\{\beta(W_i - \Delta F)\}\right)^{-1} - \sum_{j=1}^{n_{\text{R}}}\left(1 + \frac{n_{\text{F}}}{n_{\text{R}}}\exp\{-\beta(W_j - \Delta F)\}\right)^{-1} = 0 \tag{9}$$

where n_{F} and n_{R} are the numbers of switches in the forward and backward directions and W_i and W_j are work values in the forward and backward directions, respectively. BAR produces the ΔF estimate with the lowest variance for a given set of forward and backward work values and thus may be the most efficient nonequilibrium work estimator. However, in some cases, a JAR estimate using work values from one direction has been shown to be more efficient.[26]

2.4. Nonequilibrium Sampling Methods. ***Rosenbluth NE Methods.*** Another interesting set of methods, developed by Wu and Kofke and explored in this study, use Rosenbluth-like weighting of the size and position of increments to improve the sampling of the rare, important switches.[10] Depending on the perturbation, often particular areas of a switch are more or less prone to producing large work values. When a high work is probable, it is possible that the size of the work can be lowered through the use of favorable system configurations or very small λ increments. Using very small λ increments throughout a JAR calculation may be seen as inefficient, and favorable system configurations cannot be relied upon as they must be chosen at random from the Boltzmann distribution.

Three Rosenbluth NE methods were described by Wu and Kofke: λ bias, which biases the size of each λ increment to minimize the work performed; configuration bias, which biases the configuration the system takes up when undergoing a λ increment; and a hybrid bias, which does both. Rosenbluth methods implemented here were λ bias, described as "method b" by Wu and Kofke[10] (*λ bias* in figures in this study), configuration bias being "method d" (*Confbias-d* in figures in this study), and hybrid bias being "method b-c" (*Hybridbias* in figures in this study).

With λ bias, the λ value to which the system is incremented is found from a continuum weighted by the potential. This continuum is structured in such a way that λ cannot go backward with $\lambda_i \in [\lambda_{i-1}, a_i]$ where a_i is a predefined set of incremental constants, i.e.

$$0 < a_1 \leq a_2 \cdots \leq a_{n-1} = 1 \tag{10}$$

Thus, for each λ increment, the new λ (λ_i) is found from a continuum between the current λ (λ_{i-1}) and the next a value (a_i). This allows a variation in the maximum that λ_i can take.

The selection of λ_i is made according to the potential of the current configuration ($U_{\lambda_i}(q_{i-1})$) with probability density

$$P_{q_{i-1}}(\lambda_i) = \frac{1}{R_i(q_{i-1};\lambda_{i-1})}e^{-\beta\alpha_{\text{K}}U_{\lambda_i}(q_{i-1})} \tag{11}$$

where $R_i(q_{i-1};\lambda_{i-1})$ is the Rosenbluth weight, which ensures a normalized probability:

$$R_i(q_{i-1};\lambda_{i-1}) = \int_{\lambda_{i-1}}^{a_i} e^{-\beta\alpha_{\text{K}}U_{\lambda}(q_{i-1})}\,d\lambda \tag{12}$$

Here α_{K} is a predefined constant designed to control the influence of the potential of the present configuration ($U_{\lambda_i}(q_{i-1})$) on the weighting of the probability density $P_{q_{i-1}}(\lambda_i)$. As there are a predefined number of λ increments (n), the final λ increment must be made such that $\lambda_n = 1$ and hence cannot be weighted.

Each λ bias switch will be slightly different depending on the configurations that the system takes up for each λ increment. Because the probability density applied to find each new λ value in each switch is different, the resulting work values do not give an average which relates directly to the free energy difference. The work performed must be modified to account for the specific weighting of each λ bias switch using eq 13. Wu and Kofke also demonstrated that this definition of the work was consistent with the Jarzynski equality (eq 7) by defining ΔF in terms of these λ values chosen from a distribution.[10]

$$\beta W(\lambda_{i-1} \rightarrow \lambda_i) = \begin{cases} \beta(1-\alpha_{\text{K}})U_{\lambda_i}(q_{i-1}) - \beta U_{\lambda_{i-1}}(q_{i-1}) - \ln[R_i(q_{i-1};\lambda_{i-1})] & \text{if } 1 \leq i < n \\ \beta U_{\lambda_n}(q_{n-1}) - \beta U_{\lambda_{n-1}}(q_{n-1}) & \text{if } i = n \end{cases} \tag{13}$$

There are various options in defining the parameters α_{K} and a_i. The α_{K} parameter should in essence be varied depending on the general size of the system potential. If the potential is very large, it would bias λ increments to be very small, causing a large final λ increment to complete a switch and inevitably large total work values. The α_{K} parameter should be less than one so that it reduces the effect of the potential on the probability density $P_{q_{i-1}}(\lambda_i)$. Wu and Kofke use,

$$\alpha_{\text{K}} = 1/N_{\lambda} \tag{14}$$

where N_λ is the "number of atoms or particles involved in the difference between the A and B systems". Equation 14 was found from direct investigation of the use of α_K values from 0 to 1 on four quite different test systems, each of 10 HOs ($N_\lambda = 10$). A minimum of inaccuracy was found in each of the four cases which corresponds to using α_K defined by eq 14. Although eq 14 is the best definition of α_K for these HO systems, it is possible that a quite different definition may be required for large biosystems with thousands of atoms, of which only a very small fraction are perturbed in changing from systems A to B.

The obvious choice of a_i is to have each a value equal to 1. This would allow each new λ to be any value from the present value to 1. In the case that the potential does not react to a large increase in λ, this would allow the switch to proceed toward $\lambda = 1$ very quickly which would be desirable as it would avoid the situation of have a large forced final λ increment. However, Wu and Kofke found the definition,

$$a_i = i/(n-1) \tag{15}$$

for a_i preferable to the above for their test systems. Equation 15 gives the process an upper bound for each λ increment and prevents initial λ increments from being too large.

Configuration bias sampling uses the idea of Rosenbluth sampling to bias the use of configurations which are used to increment λ to produce switches with a lower W_{dis}. The structure of a standard switch has a predefined set of uniformly spaced points where λ increments are performed. This means that the system configuration present at each of these λ increment points is used whether or not these configurations allow the low work values preferred in producing a low $\bar{W}_{\mathrm{dis}}$.

At the point a λ increment is performed, configuration bias selects a system configuration from a subset of those configurations the simulation has taken up since the previous λ increment. The choice of configuration is biased to one which produces a low work value when the λ increment is performed. Thus, the switch produced via configuration bias may be more important to the exponential average in eq 7.

The configuration to be used in a λ increment is selected from a set of m taken at uniform intervals from the simulation since the previous λ increment according to,

$$P_{\lambda_i}(q_{i-1}) = \frac{1}{R_i(\lambda_i)} \exp\{-\beta f[U_{\lambda_i}(q_{i-1})]\} \tag{16}$$

In eq 16, the Rosenbluth weight is defined as

$$R_i(\lambda_i) = \sum_{j=1}^{m} \exp\{-\beta f[U_{\lambda_i}(q_{i-1,j})]\} \tag{17}$$

The term $f[U_{\lambda_i}(q_{i-1,j})]$ is a function of the potential for which Wu and Kofke list two possible options: configuration bias-c, whereas with λ bias, $f[U_{\lambda_i}(q_{i-1,j})] = \alpha_K U_{\lambda_i}(q_{i-1})$ and α_K has the same definition as for λ bias (eq 14) and configuration bias-d where $f[U_{\lambda_i}(q_{i-1,j})] = U_{\lambda_i}(q_{i-1}) - U_{\lambda_{i-1}}(q_{i-1})$ and is simply the work incurred in performing a λ increment.

Again the definition of the work performed on each switch must be modified to account for the differences in internal structure using eq 18. With configuration bias-d, equation 18 reduces to only the term containing the Rosenbluth weight.

$$\beta W(\lambda_{i-1} \rightarrow \lambda_i) = \beta U_{\lambda_i}(q_{i-1}) - \beta U_{\lambda_{i-1}}(q_{i-1}) - \beta f[U_{\lambda_i}(q_{i-1,j})] - \ln[R_i(\lambda_i)/m] \tag{18}$$

As recognized by Wu and Kofke, the present λ bias algorithm has limitations for systems with small or no phase space overlap.[10] If a λ bias switch has barriers to sampling after the initial stages, λ increments will be small and the final forced increment will incur large amounts of work. Wu and Kofke[10] attempt to alleviate this problem to some degree through a hybrid of both λ and configuration bias.

Hybrid bias is organized so that first a number of system configurations are generated and one chosen in a way biased by the subsequent λ increment. The size of the subsequent λ increment is then chosen as discussed above with λ bias. The hybrid bias probability density for selecting the configuration with which to perform a λ increment is

$$P'_{\lambda_{i-1}}(q_{i-1}) = \frac{R_i(q_{i-1}; \lambda_{i-1})}{R'_i(\lambda_i)} \tag{19}$$

In eq 19, $R_i(q_{i-1}; \lambda_{i-1})$ is the same term as found in the λ bias method described above (eq 12) and the Rosenbluth weight $R'_i(\lambda_i)$ is defined as

$$R'_i(\lambda_i) = \sum_{j=1}^{m} R_i(q_{i-1,j}; \lambda_{i-1}) \tag{20}$$

Then the modified definition of the work for hybrid bias is found using eq 21.

$$\beta W(\lambda_{i-1} \rightarrow \lambda_i) = \begin{cases} \beta(1-\alpha_K)U_{\lambda_i}(q_{i-1}) - \beta U_{\lambda_{i-1}}(q_{i-1}) - \ln\{[R'_i(\lambda_{i-1})/mI_i(\lambda_{i-1})]\} & \text{if } 1 \le i < n \\ \beta U_{\lambda_n}(q_{n-1}) - \beta U_{\lambda_{n-1}}(q_{n-1}) - \beta f[U_{\lambda_n}(q_{n-1})] - \ln\{R_n(\lambda_n)/m\} & \text{if } i = n \end{cases} \tag{21}$$

As suggested by Wu and Kofke, we only consider the use of hybrid bias with the configuration bias-c definition of $f[U_{\lambda_n}(q_{n-1})]$. This is because the α_K parameter was shown to be important to λ bias, and hybrid bias mainly uses the parameters of λ bias.

Division of λ Coordinate. As with FEP, for NE calculations, it may be optimum to divide the λ coordinate into a number of intervals (i.e., 0−0.1, 0.1−0.2,..., 0.9−1.0) which are evaluated independently, as follows,

$$\Delta F = \sum_{\lambda=0}^{1} -\frac{1}{\beta} \ln\langle \exp\{-\beta W_\lambda\}\rangle \tag{22}$$

something suggested in a number of studies.[26,27] Strangely, this potential optimization is often not explored in studies of chemical systems.[28,29]

As well as possibly optimizing the calculation, dividing a switch into many smaller switches allows more parallelization. Splitting the calculation into 10 smaller calculations allows the production of the starting configurations to be parallelized as well. Any possible optimization found through division of the λ coordinate will be investigated in this study with, for example, JAR-BY10 denoting a JAR calculation using 10 independent calculations across λ.

Replica Exchange NE. It is not possible to make the λ swap moves discussed for RETI between nonequilibrium switches, as apart from the initial ones, the configurations are not part of an equilibrium ensemble. Thus, combining RE and NE can only be achieved by performing λ swap moves between the equilibrium seed simulations to give more diverse starting configurations for switches. λ swap moves would again be accepted on the basis of equation 4.

When attempting a λ swap move it is important that there is a large amount of phase space overlap between the two ensembles involved. Attempting λ swaps between equilibrium seed simulations at $\lambda = 0$ and $\lambda = 1$ would result in an extremely low acceptance rate for all but the smallest perturbations. By using a protocol with many equilibrium seed simulations across the λ coordinate (as in eq 22), we can increase the phase space overlap of those equilibrium seed simulations adjacent in λ. The use of λ swap moves in the generation of NE starting configurations should in theory reduce error associated with incomplete sampling of large systems.

RENE methods will be investigated in this study and compared to original NE and RETI to discern if they have any potential in protein−ligand free energy calculations.

2.5. Nonequilibrium Bias Detection. Owing to the possible presence of nonsymmetric bias in JAR calculations,[11] in the case that the BAR estimator is not as accurate as JAR, it may become necessary to make a choice between JAR estimates in the forward and backward directions. Recent studies have found a relation between the $\bar{W}_{\mathrm{dis}}$ and the probable efficiency of JAR estimates in forward and backward directions.[11,30] These studies suggest that the JAR direction with the largest $\bar{W}_{\mathrm{dis}}$ should be more efficient. In this study the sensitivity of this relation will be investigated and called the $\bar{W}_{\mathrm{dis}}$ measure with $\bar{W}^{\mathrm{F}}_{\mathrm{dis}}$ denoting the forward direction and $\bar{W}^{\mathrm{B}}_{\mathrm{dis}}$ denoting the backward direction.

A method of predicting an unbiased JAR estimate has recently been developed by Wu and Kofke[10] (Kofke bias measure). The Kofke bias measure is defined such that both switching directions are explored and one is chosen when Π is above 0 (or to be safe 0.5) as follows,

$$\Pi_{\mathrm{A}\rightarrow\mathrm{B}} = \sqrt{\frac{\bar{W}^{\mathrm{F}}_{\mathrm{dis}}}{\bar{W}^{\mathrm{B}}_{\mathrm{dis}}} W_{\mathrm{L}}\left[\frac{1}{2\pi}(N-1)^2\right]} - \sqrt{2\bar{W}^{\mathrm{F}}_{\mathrm{dis}}} \tag{23}$$

$$\Pi_{\mathrm{B}\rightarrow\mathrm{A}} = \sqrt{\frac{\bar{W}^{\mathrm{B}}_{\mathrm{dis}}}{\bar{W}^{\mathrm{F}}_{\mathrm{dis}}} W_{\mathrm{L}}\left[\frac{1}{2\pi}(N-1)^2\right]} - \sqrt{2\bar{W}^{\mathrm{B}}_{\mathrm{dis}}} \tag{24}$$

Here, $W_{\mathrm{L}}(x)$ is the Lambert W function, defined as the solution for w in $x = w \exp w$.

2.6. Statistical Errors. Most estimates of standard errors use the variance of averages due to K independent blocks of measurements (block variance method). The number of blocks used for each type of estimate will be stated and is chosen to try and ensure block independence.

The block variance methods discussed here can only give an idea of the variance in the data. If a simulation is unable to overcome barriers in the energy surface of a system and only samples from a subset of phase space, these methods will not give a good estimate of the possible range of results. The best way to gauge the possible range of results is to independently repeat calculations a number of times. This will be explored in this study by comparison of standard errors to the actual range of a number of repeats of a calculation.

TABLE 1: Possible Combination of NE Methods Discussed and Tested in This Study[a]

estimator	Rosenbluth	λ div	RE	bias
JAR	λ bias	1	no	no
	Confbias			
	Hybridbias			
BAR	orig JAR	10	yes	Kofke

[a] The Rosenbluth column refers to the Rosenbluth NE sampling methods or to the absence of a Rosenbluth method, original JAR. The column named "λ div" refers to the use of 1 or 10 divisions of the λ coordinate when producing switches. The RE column refers to the use or otherwise of replica exchange methods. The Bias column refers to the use or otherwise of the NE bias correction method "Kofke bias" described here. One method from each column may be chosen in a NE protocol in this study.

2.7. Combining NE Methods. Table 1 helps explain the possible combinations of the NE methods described here. In finding an NE protocol, we may pick one estimator, an NE sampling method, and possibly an NE bias correction method.

NE methods investigated but not presented here were discounted for the following reasons. The cumulant expansion and symmetric NE estimators discussed by Hummer[27] were not included as they are only accurate for strictly Gaussian W distributions (unpublished results). The extrapolation methods of Zuckerman et al.[31,32] were not included due to perceived increased levels of statistical uncertainty and possible inaccuracy when applied to test systems with less Guassian-like work distributions (unpublished). The NE path sampling methods of Sun[33] and Ytreburg and Zuckerman[34] may offer advantages for very demanding systems. However, the method of propagating switches for these methods is to derive a starting configuration for a new switch from the previous switch. It is therefore not possible to run NE switches independently for these methods. Hence, NE path sampling methods were not included in this study, as they cannot be easily applied to large-scale embarrassingly parallel computers.

3. Simulation Methods

3.1. Harmonic Oscillator Systems. The HO systems used in this study were originally proposed by Wu and Kofke in their study of Rosenbluth NE sampling.[10] Systems A and B both have the same number of oscillating particles, N, with differing Hamiltonians,

$$H_{\mathrm{A}} = \sum_{i=1}^{N} \omega_{\mathrm{A}} x_i^2 \tag{25}$$

$$H_{\mathrm{B}} = \sum_{i=1}^{N} \omega_{\mathrm{B}} (x_i - x_0)^2 \tag{26}$$

where x_i is the reference coordinate for particle i and ω_{A} and ω_{B} are the force constants of the two systems which control the size of oscillations the particles will undergo. ω_{A} and ω_{B} control the size of phase space each system explores. x_0 displaces the reference position of particles in system B which together with ω_{A} and ω_{B} gives control over the amount of phase space overlap between the systems.

TABLE 2: Parameters of Four Test HO Systems Used to Test Our NE Implementations[a]

case	N	ω_B	ω_A	x_0	$\beta\Delta F$
A	10	500	1	0	31.07
B	10	20	1	0	14.98
C	10	20	1	1	14.98
D	10	5	1	3	8.05

[a] The terms ω_A, ω_B, and x_0 are in arbitrary units.

This HO model is very simple and many of its properties can be calculated analytically, including ΔF which can be found using eq 27. This is very useful as the result from our protocol can be compared to the right answer rather than an answer found using an exhaustive free energy protocol.

$$\Delta F = \frac{1}{2}Nk_BT\ln\left(\frac{\omega_B}{\omega_A}\right) \tag{27}$$

Each of the cases described in Table 2 model difficulties encountered with free energy calculations on more complex, chemically relevant systems. Each of cases A−D represents a different relationship between the accessible phase-spaces of systems A and B.[10] For cases A and B, system B is a subsystem of system A, and for cases C and D, system B is offset such that systems A and B have differing levels of phase-space overlap.

There are differences between the way these HO systems were sampled for this study and that of Wu and Kofke.[11] Only two of the most important differences are discussed here. Other differences were seen as inconsequential to verifying the Rosenbluth NE methods.

The HO model parameter ω_A in Table 2 may be different from that used by Wu and Kofke. This may mean the energetic barriers encountered for systems C and D will be bigger than those used by Wu and Kofke.

Wu and Kofke describe their coupling of the Hamiltonian as a linear scaling of the reference and target Hamiltonians (eq 28). This method uses two separate systems and in the context of a molecular system this is labeled dual topology.[8] The coupling used in the present study is achieved by scaling the parameters of a single system in a single topology arrangement as in eq 29.

$$H_\lambda(q) = H_A(q) + \lambda[H_B(q) - H_A(q)] \tag{28}$$

$$H_\lambda(q) = \sum_{i=1}^{N}((1-\lambda)\omega_A)(x_i)^2 + \sum_{i=1}^{N}(\lambda\omega_B)(x_i - \lambda x_0)^2 \tag{29}$$

H in eqs 28 and 29 is the Hamiltonian, and q is the present system configuration. It is unclear whether either dual or single topology calculations have an advantage for these HO cases (dual/single problem). In theory it is possible that dual topology sampling has an advantage for cases C and D as the most important regions of phase space do not shift from their previous position as λ is changed, as they do for single topology sampling. However, the dual topology method may have higher levels of sampling error compared to single topology as the system does not truly mutate as λ changes and systems A and B could be in different areas of phase space. When applied to large complex systems (such as protein−ligand systems), the high levels of sampling noise generally found with dual topology methods can be a large disadvantage.[8] Therefore, single topology methods have been used throughout this study as the main focus is toward calculating ligand binding affinities. The issues involved in this comparison of dual and single topology HO calculations are not clear and their clarification is not within the scope of this study.

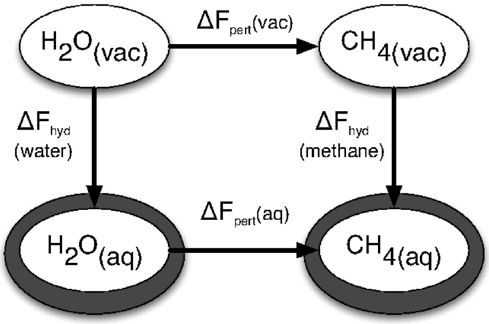

Figure 2. Thermodynamic cycle used to calculate the relative hydration free energy of water and methane.

3.2. Methane in Water. The relative hydration free energy of water and methane can be calculated through the thermodynamic cycle in Figure 2. Thus, perturbations or switches must be performed from water to methane in vacuum and water environments. The relative free energy difference ($\Delta\Delta F_{hyd}$) is found by taking the ΔF of the water leg from the ΔF of the vacuum leg, as follows,

$$\Delta\Delta F_{hyd} = \Delta F_{pert}(\text{vac}) - \Delta F_{pert}(\text{aq}) \tag{30}$$

$$= \Delta F_{hyd}(\text{water}) - \Delta F_{hyd}(\text{methane}) \tag{31}$$

The experimental free energy of hydration of methane is unfavorable at 2.00 kcal·mol^{-1} at 298 K and 1 atm, while the hydration of water has a favorable ΔF of −6.31 kcal·mol^{-1}.[35] This gives an experimental water to methane relative hydration free energy of 8.31 kcal·mol^{-1}.

This calculation has been performed in the literature by Woods et al. using many free energy methods in a direct comparison.[18] It is convenient to use the same system setup to give a quick comparison of NE methods with the equilibrium methods investigated by Woods et al. The water to methane model consists of a TIP4P water molecule switching to an OPLS united atom methane molecule.[36,37] The oxygen atom of the TIP4P water is switched to the OPLS methane while the TIP4P hydrogens and extra "M" atom are switched to dummy atoms. The bond lengths of the TIP4P hydrogens are reduced to 0.2 Å as they become dummy atoms to help smooth the switch and to pull them inside the influence of the methane molecule. The water−methane resides in an periodic, orthorhombic box of 1679 TIP4P molecules.

Both the TIP4P water and the OLPS methane are rigid-molecule models. Consequently, for this calculation, the vacuum leg of the calculation from Figure 2 will give a ΔF of zero and can be discounted. The calculation now consists of a single perturbation and the free energy difference is now due to the solvent rearrangement only. This is useful as a free energy method can be assessed solely on its ability to evaluate solvent rearrangment and solute−solvent interactions.

Neutron diffraction results for a methane system show peaks for both hydrogen−methane and oxygen−methane radial dis-

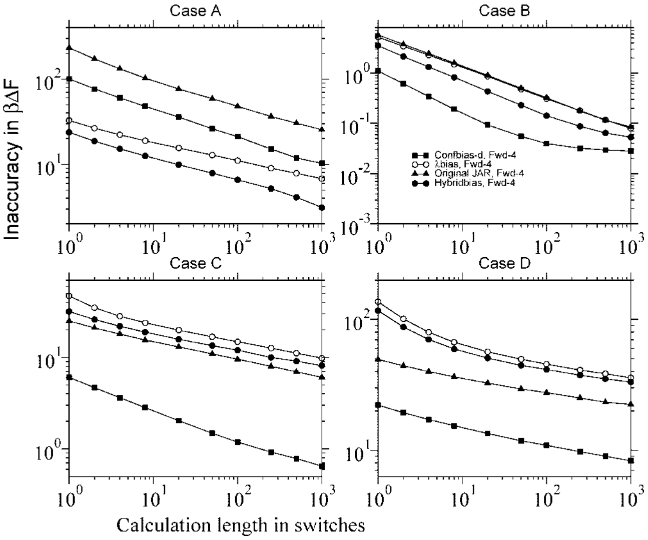

Figure 3. HO results for cases A−D showing inaccuracy in $\beta\Delta F$ against numbers of trajectories where a trajectory has 10 λ increments and 1000 MC trials per λ increment. For configuration bias, 10 configuration samples were used per λ increment. Here, for example, "Confbias-d, Fwd-4" denotes a configuration-d bias calculation using switching protocol 4 from Table 3, in the forward direction only.

tribution functions (RDFs) at 3.5 Å.[38] As the hydrogen peak is broader than the oxygen peak it is thought that the waters orientate themselves with hydrogens toward the methane. The results for pure water show a sharp oxygen−oxygen RDF peak at 2.8 Å with a second shell at 4.5 Å and a third at 6.8 Å.[18] This change in structure is likely to cause JAR calculations to display some level of Hamiltonian lag whereas TI equilibrium calculations would not. Also, as the ΔF is mainly due to solvent rearrangement, it is likely that this solvated water to methane perturbation will display a smooth potential energy change. This will allow TI calculations to integrate over the free energy gradient with a good level of accuracy.

Water−methane simulation parameters used by Woods et al.[18] are used in the present calculations. The present study used the ProtoMS 2.1 Monte Carlo (MC) simulation software[39] rather than MCPRO 1.5[40] used by Woods et al. ProtoMS 2.1 is unable to force volume moves every 10 375 MC trials as is possible with MCPRO. Instead, volume moves are made with a probability relative to solute and solvent moves. The MC trial probability ratios used for water−methane simulation in the present study are volume 2/solute 13/solvent 20 800. This is very close to the MC trial probabilities used by Woods et al.

All NE water−methane simulations reported here were performed on an inhomogeneous distributed computing cluster at the University of Southampton, consisting of approximately 1500 processors, running under the Condor software.[41] The water−methane calculations discussed here used 100 condor processors each and are thus more highly parallelised than the similar equilibrium simulations of Woods et al.[18]

4. Results

4.1. Rosenbluth Validation. The aim of this validation is to show that the Rosenbluth methods implemented here are able to reproduce the efficiency trends which are demonstrated by Wu and Kofke.[10] To achieve this validation we make a direct comparison of our results with those of Wu and Kofke in Figure 3.

Figure 3 shows the inaccuracy of ΔF estimates (i.e., the difference between the estimate and the analytical result) as the numbers of switches used per estimate is increased, using the different Rosenbluth NE methods. Only the JAR estimator is used in the comparison in Figure 3. The calculation presented in Figure 3 used switches of 10 λ increments with 1000 MC trials between each λ increment. This switching protocol is protocol 4 in Table 3 (as indicated by Fwd-4 in the legend of Figure 3). Table 3 describes the different switches used in this study, the column labeled "total MC trials" displays the total MC trials each NE calculation uses.

TABLE 3: HO Test NE Switching Protocols

protocol	λ inc	MC trials/λ inc	total MC trials
1	10	100	1×10^7
2	50	20	1×10^7
3	200	5	1×10^7
4	10	1000	1×10^8
5	50	200	1×10^8
6	200	50	1×10^8

For each of cases A−D, forward switches will produce more efficient JAR calculations than backward switches as the important phase space of system B is more easily accessible from A than the other way around.[11] Thus, here backward JAR estimates are not considered.

Another issue to be aware of when studying Figures 3, 4, and 5 is the method of data point averaging used. Wu and Kofke describe "additional outer repetitions" made to better characterize the inaccuracy in the free energy difference in their estimates. These outer repetitions are used to reduce the statistical uncertainty by averaging each JAR estimate over many individual calculations. Wu and Kofke used 10 000 outer repetitions. Our work has used 100 000 NE switches for each analysis. Thus, for each case in Figure 3 100 000 NE switches were made. Hence, the data point for one switch per NE estimate was averaged over 100 000 estimates and the data point for 1000 switches per NE estimate was averaged over 100 estimates.

Case B is the least demanding of cases A−D, shown in Figure 3. All the Rosenbluth methods can calculate a free energy very

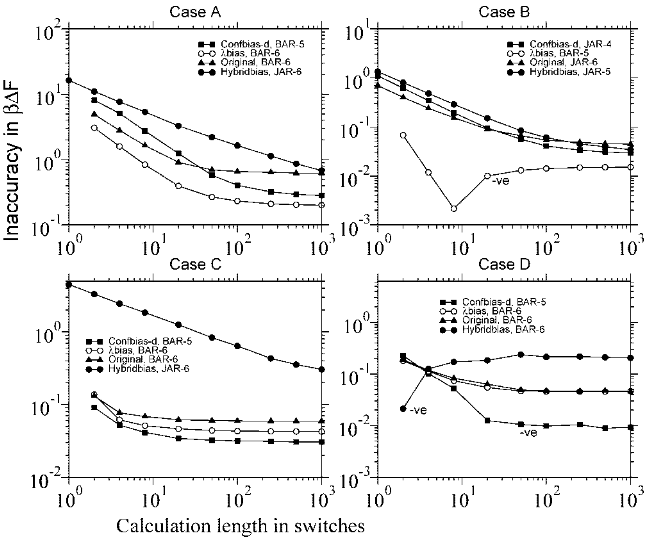

Figure 4. HO results for cases A−D using BAR and JAR with any of the 1×10^8 switching protocols from Table 3, showing inaccuracy in $\beta\Delta F$ against numbers of switches. For each switch sampling method (configuration bias, original NE, etc.), only the calculation with the lowest total inaccuracy from 100, 1000 switch estimates is plotted. The switching protocol plotted for each switch sampling method is listed in the legend. For example, λ bias, BAR-6 denotes λ bias with the BAR estimator and switching protocol 6 from Table 3. In the legend, JAR always denotes JAR estimates using forward switches only.

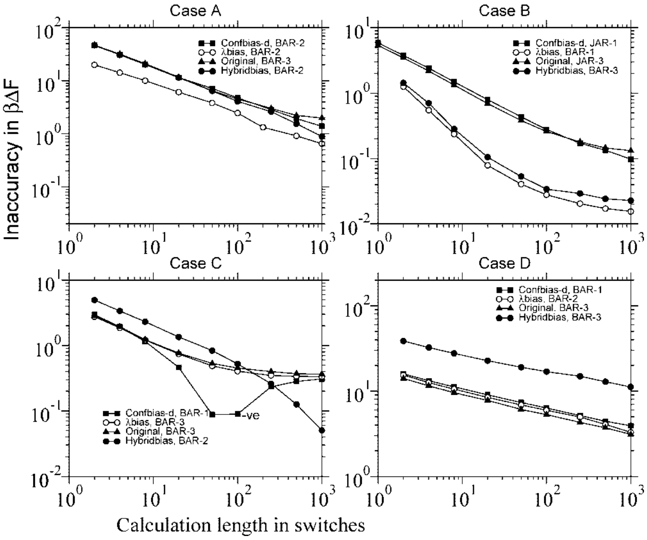

Figure 5. HO results for cases A−D using BAR and JAR with any of the 1 × 10^7 switching protocols from Table 3, showing inaccuracy in $\beta\Delta F$ against numbers of switches. For each switch sampling method (configuration bias, original NE, etc.), only the calculation with the lowest total inaccuracy from 100, 1000 switch estimates is plotted. The switching protocol plotted for each switch sampling method is listed in the legend. For example, λ bias, BAR-6 denotes λ bias with the BAR estimator and switching protocol 6 from Table 3. In the legend, JAR always denotes JAR estimates using forward switches only.

close to the analytical value. Our results, for case B, reproduce those seen by Wu and Kofke very well.[10]

Case A is a more demanding free energy difference to evaluate than case B. Consequently, original JAR can only estimate ΔF to within 20−30 units. Hybrid bias and λ bias JAR give a performance improvement of almost an order of magnitude. Configuration bias-d JAR also shows improved efficiency over original JAR. This result is in line with that seen by Wu and Kofke.

Case C is similar to case B but provides an extra barrier to sampling due to the 1 unit displacement of system B. Our results for case C show significant difference from those of Wu and Kofke. Original JAR displays similar performance when only one switch is used for each ΔF estimate. However as the numbers of switches is increased, our results do not have the same improvement in accuracy with our data point at 10^3 switches being at around 6 inaccuracy units, compared to that of Wu and Kofke at around 2. Wu and Kofke show hybrid bias and λ bias to perform better than original JAR in this case whereas our results show them to perform slightly worse. This suggests that there is some difference between the calculations, as discussed earlier. Case D has very similarly sized areas of important phase space for systems A and B but with larger displacement barriers to sampling than case C. Our results for this case show the same trends as Wu and Kofke but with increased inaccuracy for all methods except configurational bias-d.

For cases A and B, we have quantitatively reproduced the results of Wu and Kofke. The differences found for cases C and D may be rationalized through the differences in protocol from the study of Wu and Kofke. While the difference in protocol may have harmed the comparison with the study of Wu and Kofke, the protocols to be used in this study have been validated and the possible improved performance of these NE methods demonstrated.

4.2. Best NE Combination for HO Cases A−D. Here, we compare a range of combinations of NE methods which are all easily parallelizable. Table 3 describes six switching protocols, which have been used with the Rosenbluth NE methods described by Wu and Kofke[10] and the JAR and BAR estimators.

Figure 4 shows only the results for those protocols which achieve the lowest total inaccuracy in $\beta\Delta F$ from 100 estimates with 1000 switches. Each plot in Figure 4 compares each of the Rosenbluth methods using either JAR or BAR estimators, whichever is the more accurate, for each of cases A−D, using a total of 1 × 10^8 MC trials. Figure 5 shows these results for switching protocols with a total of 1 × 10^7 MC trials. In both Figures 4 and 5, a data point marked with "−ve" marks where inaccuracies have become negative. In all cases, data points plotted after (on the same line with more switches) a −ve have negative values of inaccuracy.

In Figures 4 and 5, BAR data with negative inaccuracy seem to be more inaccurate as we increase the number of switches in each estimate i.e. these calculations converge but to the wrong answer. To investigate this for one protocol we increased the level of averaging such that the data points for 1000 switches were averaged over 700 rather than 100 estimates. This had the effect of bringing the converged inaccuracy closer to zero (data not shown). We presume that the origin of this problem with BAR, Rosenbluth data lies in the very poor reverse work values in these cases. For example, it has been shown that reverse JAR has a standard error going to infinity for $\omega_B > 2\omega_A$.[26] Wu and Kofke also report reverse FEP (closely related exponential free energy method) calculations with case B which have slowly converging negative inaccuracy after 1 × 10^6 switches.[11]

It is clear that the introduction of BAR and/or a choice of switching protocols significantly increases the efficiency and accuracy of calculations for all cases. For case A, Figure 4 shows that although λ bias with BAR and switching protocol 6 (λ bias, BAR-6) is most efficient, configuration bias-d with BAR and switching protocol 5 (confbias-d, BAR-5) has produced estimates with similar accuracy using 1000 switches. It is interesting that hybrid bias is most accurate when used with forward JAR and protocol 6.

For case B, all best performing calculations use JAR except λ bias; all produce very similar results suggesting that for this case and possibly other similar cases forward JAR is more efficient than BAR. It is interesting that configuration bias-d is most accurate with switching protocol 4 while the other methods use protocols with more increments. This may mean that configuration bias performs best with many MC trials between increments. This is probably due to the need to produce independent configurations from which configuration bias can pick a configuration advantageous to a given increment.

For case C, all methods perform best with BAR and switching protocols 5 or 6 except hybrid bias which prefers JAR with protocol 6. Again, all methods produce very similar results for case C apart from hybrid bias, however configuration bias-d and λ bias display a slight increase in accuracy over original NE. Case D has all Rosenbluth methods using BAR and switching protocols 5 or 6 with each seeming to converge to a maximum accuracy when using around 50 switches. Configuration bias-d finds a maximum accuracy slightly lower than original and λ bias and over an order of magnitude more accurate than hybrid bias.

Figure 5 shows results for switching protocols with a total of 1 × 10^7 MC trials, ie. switches an order of magnitude shorter than for Figure 4. Here, trends seen in Figure 4 are generally repeated. Case B apart, no Rosenbluth method is able to offer a significant increase in efficiency or accuracy compared to

TABLE 4: Comparison of Inaccuracy in ΔF Values Estimated with NE for Cases A and D with Switches Split into 10 and Not Split at All

protocol	200-5-BY10	20-5-BY10	200-5
no. switches	2×10^4	2×10^5	2×10^4
case A, fwd	0.20	0.52	1.22
case A, bwd	0.76	2.59	2.56
case A, BAR	0.37	0.42	0.15
case D, fwd	0.21	0.13	15.73
case D, bwd	0.40	0.24	16.61
case D, BAR	0.73	0.7	0.41

original, BAR-3. For case B hybrid and λ bias offer significantly increased accuracy compared to configuration bias-d and original; however, all methods are accurate to within 0.1 units.

These Rosenbluth methods are significantly more computationally complex and time-consuming compared to original NE. This is especially problematic for hybrid bias and λ bias and with switching protocols 3 and 6 these methods are prohibitively slow. Taking this into account with the results of Figures 4 and 5, it seems sensible to use the original sampling method together with BAR as the estimator, with a switching protocol with many increments for these simple harmonic oscillator systems.

4.3. Simple Investigation of NE-BY10. Table 4 shows results from NE calculations for cases A and D comparing calculations which are broken up into 10 smaller calculations (as described in the methods section and denoted by BY10) and calculations with continuous switches from $\lambda = 0$ to $\lambda = 1$. Both JAR and BAR estimators have been used. The labels of each column refer to the switching protocol used, with the first number referring to the number of increments and the second number after the "−" referring to the number of MC trials between each λ increment. All calculations in Table 4, used 2×10^7 MC trials in their switches. Thus, the 20-5-BY10 calculations used 200 000 switches while the others used 20 000 switches, indicated in the row labeled no. switches.

Splitting switches into 10 has improved the performance for both cases A and D. This is especially marked for the JAR estimates. It is interesting to compare the estimates for case A which use the 200-5-BY10 and 20-5-BY10 protocols from Table 4. These protocols give similar results except the backward JAR estimate for 20-5-BY10 is far less accurate and similar to the 200-5 estimate. This suggests that for case A backward JAR estimates, the improvement seen for 200-5-BY10 comes from the effective increase in length of switching from $\lambda = 0$ to $\lambda = 1$ rather than the extra equilibrated starting points for switches across λ. If the same protocols are compared for case D, it is clear that to 20-5-BY10 estimates are much improved compared to the 200-5 estimates. Thus, the extra equilibrated points across λ are more important in this case.

Comparing BAR estimates for BY10 and non-BY10 protocols seems to suggest that BAR estimates do not gain accuracy through the addition of extra equilibrated intermediates across λ.

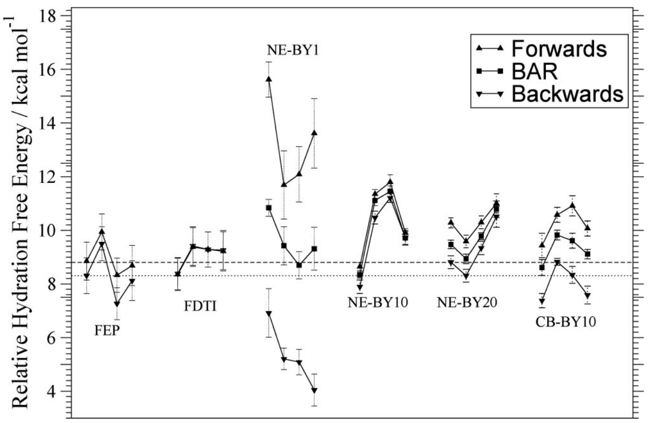

Figure 6. Relative hydration free energy of water and methane, estimated by four repetitions for each non-RE method. Each set of four estimates is linked with a line and labeled with the method abbreviation. These estimates are compared to the experimental value of 8.31 kcal·mol^{-1} which is the dotted line and the recent RETI estimate of 8.8 kcal·mol^{-1} of Woods et al.[42]

4.4. Relative Hydration Free Energy of Water and Methane. ***Nonreplica Exchange Methods.*** Figure 6 shows water−methane results for nonreplica exchange (non-RE) protocols, comparing four repeated calculations (different random number seeds), which all started with the same equilibrated structure and are joined by a line as an aid to the eye. FEP, TI (taken from ref 18), and various NE protocols are used. Of the Rosenbluth NE methods, only configuration bias is presented here owing to the conclusions of the HO system simulations and the extra computational cost of the λ bias and hybrid bias approaches. Triangles pointing up denote JAR estimates where λ is being incremented from 0 (water) to 1 (methane) and triangles pointing down where λ is being decremented from 1 to 0, with squares representing BAR calculations. Also, it should be noted that JAR estimates in this water−methane study use half the number of switches as BAR and TI estimates. The dotted line marks the experimental relative hydration free energy of 8.31 ± 0.5 kcal·mol^{-1}, and the dashed line marks the recently exhaustively calculated relative hydration free energy of 8.8 ± 0.1 kcal·mol^{-1} using RETI.[42]

FEP (212 million MC trials) estimates display a hysteresis of 0.5−1 kcal·mol^{-1}. The problem of choosing a sampling direction may be solved by application of BAR to the FEP results, which would find the optimum estimate due to the variance of the data. However, more of an issue is the spread of FEP estimates which is around 2 kcal·mol^{-1}. This large spread is despite the relatively simple system and the common equilibrated starting structure for each calculation repetition. Each FEP calculation repetition probably samples different areas of configuration phase space giving these different results. This problem can be called the random sampling error, as discussed by Woods et al.[18] FDTI (212 million MC trials) shows very low levels of hysteresis; however, this is obviously a poor measure of the level of possible error as the random sampling error is comparable to that of FEP.

The results labeled NE-BY1 use a protocol (212 million MC trials) with uninterrupted switches from $\lambda = 0$ to 1. The numbers of NE switches, λ increments and MC trials per λ increment are listed in Table 5. Table 5 also shows the numbers of seed

TABLE 5: Non-RE NE Protocols Used in Figure 6

NE protocol	switches	Δλs	MC trials per Δλ	λ split	seed blocks	seed block size	total MC trials (millions)
NE-BY1	340	1000	500	1	340	10^5	212
NE-BY10	400	1000	375	10	220	10^5	207
NE-BY20	600	500	375	20	315	10^5	209
CB-BY10	400	100	4000	10	220	10^5	207

blocks and the seed block size. Here, a seed block refers to the equilibrium simulation needed to produce a starting configuration for the next switch. Hence, for NE-BY1, 340 seed blocks of 10^5 MC trials each were used to produce 340 switches. However, for NE-BY10, most seed blocks give rise to a forward and backward switch (only those at $\lambda = 0$ and 1 give rise to a single switch) and so only 220 are needed. The λ split referred to in Table 5 is the number of times the λ coordinate is divided with switches linking each division. Two million MC trials were used in an initial equilibration (initial equilibration) of the system, before this configuration was taken as the starting point for each λ division, which were then equilibrated for a further 3 million MC trials (λ equilibration). The total MC trials used in each protocol is easily found through eq 32.

$$\text{MC trials} = (\text{seed blocks} \times \text{seed block size}) + (\text{switches} \times \Delta\lambda\text{s} \times \text{MC trials}) + \lambda\ \text{equilibration} + \text{initial equilibration} \quad (32)$$

The hysteresis of JAR estimates is very large at around 13 kcal·mol^{-1}. However, the BAR estimates from the same switches is of similar quality to FEP and FDTI.

NE-BY10 (207 million MC trials) uses the same switching protocol as NE-BY1. However, as the λ coordinate is split into 10 (Table 5), the use of a NE-BY10 protocol massively lowers the hysteresis of JAR ΔF estimates. BAR estimates due to the NE-BY10 protocol are not improved over BAR estimates with uninterrupted switches, as was observed with our HO systems.

The NE-BY20 (209 million MC trials) protocol produces quite similar estimates to the NE-BY10 protocol (Figure 6).

The CB-BY10 (207 million MC trials) protocol uses the configuration bias-d method to perform switches with small numbers of λ increments and many MC trials between λ increments designed to allow maximum sampling for configuration choices (Table 5). These CB-BY10 estimates display a higher level of hysteresis between forward and backward JAR estimates than the NE-BY10 calculations which suggests a lower level of convergence. CB-BY10 BAR estimates have a much smaller range than the NE-BY10 BAR estimates, although this may be due to random effects.

Of the methods used in this non-RE comparison, FDTI looks to be the method of choice as it is the most accurate and reliable. However, all the NE protocols in Figure 6 produce BAR estimates with comparable quality to FDTI taking into account possible random effects. All statistical errors displayed here are calculated with the block variance method discussed earlier in the methods section, with BY1 protocols using 10 blocks, BY10 using 2 blocks, and BY20 using 4 blocks, to maintain the independence of the blocks.

Replica Exchange Methods. Methods involving λ swap moves display significantly more consistent ΔF estimates (Figure 7), as shown by Woods et al.

RETI ΔF estimates show very low random sampling error and extremely good agreement with experiment.[18] REFEP has slightly larger levels of hysteresis and random sampling error.

The four RETI estimates of Figure 7 were performed with MCPRO.[40] The dashed line marks the more recent RETI estimate performed with ProtoMS 2.1[39] independently of this study.[42] The RETI-ProtoMS 2.1 estimate uses the same protocol as the four RETI estimates in all other ways.

RENE-BY10 (221 million MC trials) estimates in Figure 7 have less random sampling error than the other NE protocols. RENE-BY10 JAR and BAR estimates are very close to the RETI value of 8.8 kcal·mol^{-1} and nearly as consistent as the four RETI estimates while using approximately half the number of MC trials. This suggests that RENE methods can be as consistently accurate as RETI.

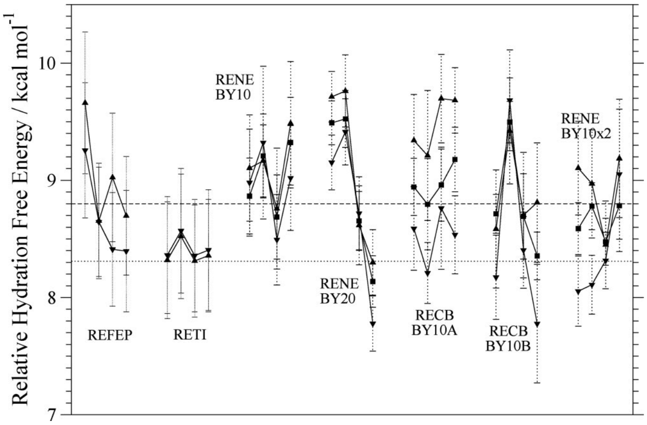

Figure 7. Relative hydration free energy of water and methane, estimated by four repetitions for each RE method. Each set of four estimates is linked with a line and labeled with the method abbreviation. These estimates are compared to the experimental value of 8.31 kcal·mol^{-1} which is the dotted line and the recent RETI estimate of 8.8 kcal·mol^{-1} of Woods et al.[42]

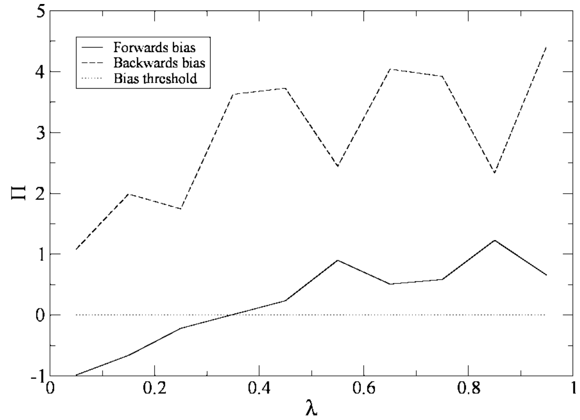

Figure 8. Plot of Kofke bias measure values (Π) for each individual calculation of the first repetition of the RENE-BY10 protocol (Figure 7). In the legend "Forwards bias" and "Backwards bias" denote the Kofke bias measure in the forward and backward direction respectively. Also, "Bias threshold" denotes the point at which Π suggests a converged NE calculation.

RENE-BY20 (213 million MC trials) estimates in Figure 7 display estimates with less random sampling error compared to NE-BY20. However, RENE-BY20 does not seem to offer any improvement compared to RENE-BY10.

The RECB-BY10A (214 million MC trials) and RECB-BY10B (210 million MC trials) protocols use the configuration bias-d method to perform switches with differing numbers of λ increments and MC trials, as shown in Table 6. The estimates from these two protocols seem less consistent than those produced by the RENE-BY10 protocol.

The RENE-BY10 × 2 calculations of Figure 7 have switches which are twice the length of those in the RENE-BY10 calculations. These RENE-BY10 × 2 calculations seem to offer no improvement in accuracy and consistency.

In summary, the use of RE moves to generate the initial structures used in the NE simulations significantly improves the precision of the calculated free energies, presumably by sampling more diverse configurations more efficiently. In general, the BAR estimator is more reliable than JAR, and subdividing the λ coordinate improves performance up to a

TABLE 6: REFG Protocols Used in Figure 7

NE protocol	switches	$\Delta\lambda$s	MC trials per $\Delta\lambda$	λ split	seed blocks	seed block size	total MC trials (millions)
RENE-BY10	600	1000	200	10	330	20^5	221
RENE-BY20	720	500	200	10	378	20^5	213
RECB-BY10A	600	125	1500	20	330	20^5	214
RECB-BY10B	360	250	1500	20	198	20^5	210
RENE-BY10 × 2	600	1000	400	10	330	20^5	341

TABLE 7: $\bar{W}_{\mathrm{dis}}$ Values and RENE-BY10 Estimates of the Relative Hydration Free Energy of Water and Methane (kcal·mol^{-1})[a]

rep	$\bar{W}_{\mathrm{dis}}^{\mathrm{F}}$	$\bar{W}_{\mathrm{dis}}^{\mathrm{B}}$	FJAR	BJAR	BAR
1	1.83	2.17	9.10	8.98	8.86
2	1.97	2.06	9.17	9.32	9.20
3	1.65	1.40	8.76	8.49	8.68
4	1.60	1.36	9.48	9.02	9.56

[a] $\bar{W}_{\mathrm{dis}}^{\mathrm{F}}$ and $\bar{W}_{\mathrm{dis}}^{\mathrm{B}}$ denote the $\bar{W}_{\mathrm{dis}}$ in the forwards and backwards direction, respectively. Also, FJAR and BJAR denote JAR free energy estimates in the forwards and backwards directions, respectively.

point. The use of configuration bias in this context offers little improvement over conventional approaches to generating the NE trajectories.

Detection of NE Calculation Bias. As seen for case B of the harmonic oscillator study and the RENE-BY10 calculations in Figure 7, JAR is sometimes more accurate than BAR. However, as shown in other studies,[11] often one switching direction provides more efficient estimates due to asymmetric bias. The Kofke bias measure (eq 24) and the $\bar{W}_{\mathrm{dis}}$ measure can be used in an attempt to predict the most accurate JAR estimate. However, it is difficult to know when a JAR estimate should be used over a BAR estimate.

NE protocols such as RENE-BY10 have a number of small calculations across λ rather than one calculation relating the A and B systems. There are a number of ways combining the individual ΔF estimates to give the total ΔF across λ; here, we will test two:

Independently Chosen Kofke Bias Measure. Each individual NE calculation can be treated independently and a particular estimator chosen for each, with the result being that different estimators are used for different parts of the complete free energy calculation.

Totalled $\bar{W}_{dis}$ Measure. The $\bar{W}_{\mathrm{dis}}$ measure can be totalled across the λ coordinate and a direction chosen from these totals, so the same estimator is used for each individual NE calculation across λ. This method may not discriminate the correct estimator when using few, very long switches and calculations are well behaved, as the $\bar{W}_{\mathrm{dis}}$ is likely to be similar in the forward and backward direction; consequently, random fluctuations could have an impact.

Figure 8 shows the Kofke bias measures for each individual calculation of the first RENE-BY10 repetition in Figure 7. For the three NE calculations between 0 and 0.3 of the λ coordinate, the Kofke bias for forward switches measure is negative, predicting that these forward JAR estimates are not converged. The Kofke bias measure suggests that in general the backward JAR estimates are more converged than the forward JAR estimates. These trends are seen in all repetitions of these RENE-BY10 calculations suggesting the initial portion of the λ coordinate is difficult to converge in the forward direction and overall the backward direction may converge faster. It is difficult to interpret these results as all NE estimators seem to give comparable levels of accuracy for RENE-BY10 in Figure 7.

$\bar{W}_{\mathrm{dis}}$ measures totalled across λ for the RENE-BY10 calculations given in this study are in Table 7. Using the totalled $\bar{W}_{\mathrm{dis}}$ measure for the RENE-BY10 repetitions, forward JAR estimates are picked for repetitions 3 and 4 while backward JAR estimates are picked for repetitions 1 and 2. All RENE-BY10 estimates are very close, and it is clear that these methods of predicting the most accurate estimator are not sufficiently sensitive in this case.

5. Conclusions

Over recent years, many promising NE, highly parallelizable free energy methods have been reported. Here, we have investigated possible combinations of these NE methods with a view to their use in calculating protein−ligand binding free energies.

The Rosenbluth methods investigated here, originally presented by Kofke et al. offer advantages for the test protocols originally presented.[10] However, when other protocols and estimators are explored these advantages do not necessarily remain. Original NE switches are comparable to any of the Rosenbluth NE methods as long as the size of λ increments is kept small, and given the computational cost and complexity of the Rosenbluth methods, using unbiased work values is generally to be recommended. It is clear that for HO systems, in general, BAR is the most efficient NE estimator, although when enough MC trials are available JAR may be able to provide a more accurate estimate. Also, JAR calculations seem to be more accurate when the λ coordinate is split into a number of independent calculations but generally offer no improvement over BAR.

Splitting a switch into many independent switches makes sense from a computing point of view as it allows more parallelization. Often the bottleneck for NE calculations is the production of the equilibrium starting configurations. Splitting the complete calculation across λ into 10 smaller calculations allows the production of the starting configurations to be parallelized as well. It is also necessary to take a computing point of view when selecting the length of switches. Many studies of NE methods suggest that the longest switches possible should be used.[26,43,44] This detracts from the main advantage of NE methods which is the possible embarrassing parallelization.

The addition of replica exchange moves to NE calculations for the generation of the initial equilibrium configurations was investigated. For the relative hydration free energy of water and methane, coupled with subdividing the total calculation into 10, this approach improves accuracy and precision. RENE-BY10 is able to produce estimates of similar accuracy and reproducibility to RETI. This protocol has subsequently been applied to protein−ligand systems.[12]

When forward and backward JAR and BAR estimators do not give the same value, it may be difficult to choose the most accurate estimate. When forward and backward JAR are very similar, this may signify a lack of bias in the calculation and that both provide an accurate estimate, although this is not a fool-proof method. To assess whether the more accurate estimate can be determined a priori, two measures were investigated. Neither the Kofke bias nor $\bar{W}_{\mathrm{dis}}$ measures were sufficiently sensitive in the relative hydration free energy of water and methane calculations presented here. However, these methods may give better predictions when estimates are less well converged. It may be simpler to use BAR on all occasions as BAR is generally accurate.

We therefore recommend that original (non-Rosenbluth) work values, coupled with the BAR estimator, and replica exchange moves to generate the equilibrium seed configurations, be used for molecular systems.

If the limiting factor to calculations is not the number of MC trials but instead is the wall clock time, then NE methods may have a large advantage over equilibrium approaches, on account of their ready use in a distributed computing environment. This being the case, and depending on the number of computers used, more MC trials could be used in a RENE-BY10 calculation than for a RETI calculation taking the same amount of time. As RENE-BY10 has been shown to give similar performance to RETI with the same number of MC trials used, with this extra computational advantage RENE methods must be seen as viable alternatives to RETI, and may be used in preference in certain situations when significant distributed computing resources are available.

Acknowledgment. We thank GlaxoSmithKline, the EPSRC, and the University of Southampton for support.

References and Notes

(1) Price, M. P.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 9455–9466.

(2) Oostenbrink, C.; van Gunsteren, W. F. *Prot. Struct. Func. Gen.* **2004**, *54* (2), 237–246.

(3) Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. *Bioorg. Med. Chem. Lett.* **2006**, *16* (3), 663–667.

(4) Kim, J. T.; Hamilton, A. D.; Bailey, C. M.; Domoal, R. A.; Wang, L. G.; Anderson, K. S.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2006**, *128* (48), 15372–15373.

(5) Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Med. Chem.* **2006**, *49*, 7427–7439.

(6) Deng, Y. Q.; Roux, B. *J. Chem. Theory. Comput.* **2006**, *2*, 1255–1273.

(7) Oostenbrink, C.; van Gunsteren, W. F. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6750–6754.

(8) Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Chem. Theory. Comput.* **2007**, *3*, 1645–1655.

(9) Jarzynski, C. *Phys. Rev. E* **1997**, *56*, 5018–5035.

(10) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *122*, 204104.

(11) Wu, D.; Kofke, D. A. *J. Chem. Phys.* **2005**, *123*, 054103.

(12) Cossins, B. P.; Foucher, S.; Edge, C. M.; Essex, J. W. *J. Phys. Chem. B.* **112**, 14985–14992.

(13) Hummer, G. *Mol. Sim.* **2002**, *28*, 81–90.

(14) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

(15) Leach, A. R., *Molecular modelling, principles and applications*, first ed.; Longman: Harlow, U.K., 1996.

(16) Frenkel, D.; Smit, B. *Understanding molecular simulation*; Academic Press: New York, 1996.

(17) Mezei, M. *J. Chem. Phys.* **1987**, *86*, 7084–7088.

(18) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B.* **2003**, *107*, 13703–13710.

(19) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13711–13718.

(20) Hu, H.; Yun, R. H.; Hermans, J. *Mol. Sim.* **2002**, *28*, 67–80.

(21) Pearlman, D. A.; Kollman, P. A. *J. Chem. Phys.* **1989**, *91* (12), 7831–7839.

(22) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

(23) Bennett, C. H. *J. Comp. Phys.* **1976**, *22*, 245.

(24) Crooks, G. E. *Phys. Rev. E* **1999**, *60*, 2721–2726.

(25) Shirts, M. R.; Blair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, *91*, 140601–140604.

(26) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 144107.

(27) Hummer, G. *J. Chem. Phys.* **2001**, *114*, 7330–7337.

(28) Oostenbrink, C.; van Gunsteren, W. F. *Chem. Phys.* **2006**, *323*, 102–108.

(29) Ytreburg, M. F.; Swendsen, R. H.; Zuckerman, D. *Physics* **2006**, *2*, 0602088.

(30) Crooks, G. E.; Jarzynski, C. *Phys. Rev.* **2007**, *75* (2), 021116.

(31) Zuckerman, D. M.; Woolf, T. B. *Chem. Phys. Lett.* **2002**, *351*, 445–453.

(32) Ytreburg, F. M.; Zuckerman, D. *J. Comput. Chem.* **2004**, *25*, 1749–1759.

(33) Sun, S. X. *J. Chem. Phys.* **2003**, *118*, 5769.

(34) Ytreberg, F. M.; Zuckerman, D. M. *J. Chem. Phys.* **2004**, *121*, 50225023) .

(35) Zhou, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117–9133.

(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(37) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

(38) Dejong, P. H. K.; Wilson, J. E.; Neilson, G. W.; Buckingham, A. D. *Mol. Phys.* **1997**, *91*, 99–103.

(39) Woods, C. J.; Michel, J. *ProtoMS2.1*; in-house Monte Carlo software, 2002−2006.

(40) Jorgensen, W. L. *MCPRO*; version 1.5, Yale University: New Haven, CT, 1996.

(41) Litzkow, M.; Livny, M.; Mutka, M. In *Condor — A Hunter of Idle Workstations*; San Jose, California, June 13−17, *Proceedings of the 8th International Conference of Distributed Computing Systems* **1988**, 104−111.

(42) Woods, C. J.; Manby, F. R.; Mulholland, A. J. *J. Chem. Phys.* **2008**, *128*, 014109−8.

(43) Gore, J.; Ritort, F.; Bustamante, C. *Proc. Nat. Acad. Sci. USA.* **2003**, *100*, 12564–12569.

(44) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. Lett.* **2002**, *89*, 180602.

JP803532Z