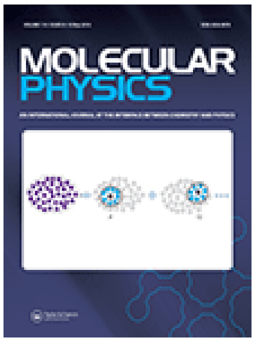


Taylor & Francis
Taylor & Francis Group

Molecular Physics
An International Journal at the Interface Between Chemistry and Physics

ISSN: 0026-8976 (Print) 1362-3028 (Online) Journal homepage: http://www.tandfonline.com/loi/tmph20

Convergence of single-step free energy perturbation

Stefan Boresch & H. Lee Woodcock

To cite this article: Stefan Boresch & H. Lee Woodcock (2016): Convergence of single-step free energy perturbation, Molecular Physics, DOI: 10.1080/00268976.2016.1269960

To link to this article: http://dx.doi.org/10.1080/00268976.2016.1269960

MOLECULAR PHYSICS, 2016
http://dx.doi.org/10.1080/00268976.2016.1269960

SPECIAL ISSUE IN HONOUR OF JOHANN FISCHER

Convergence of single-step free energy perturbation

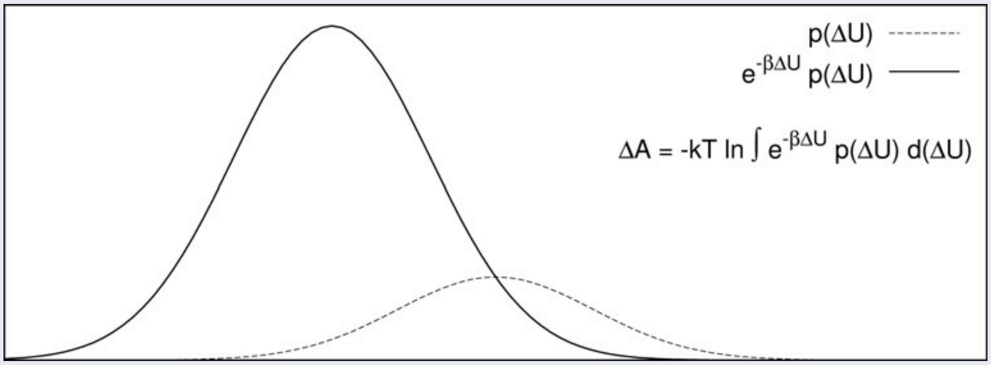
Stefan Boresch [a] and H. Lee Woodcock [b]

[a] University of Vienna, Faculty of Chemistry, Department of Computational Biological Chemistry, Vienna, Austria; [b] Department of Chemistry, University of South Florida, Tampa, FL, USA

ABSTRACT
The convergence of free energy perturbation (Zwanzig's equation) and its non-equilibrium extension (Jarzynski's equation) is herein investigated from a practical point of view. We focus on cases where neither intermediate steps nor two-sided methods (Bennett, Crooks) can be used to compute the free energy difference of interest. Using model data sampled from several probability densities, as well as comparing results of actual free energy simulations with reference values, we find, in agreement with existing theoretical work, that systematic errors are strongly correlated with the variance in the distribution of energy differences / non-equilibrium work values. The bias metric Π introduced by Wu and Kofke (J. Chem. Phys. 121, 8742 (2004)) is found to be a useful test for the presence of bias, i.e. systematic error in the results. By contrast, use of the second-order cumulant approximation to approximate the full Zwanzig or Jarzynski equation leads to poorer results in almost all cases.

ARTICLE HISTORY
Received 11 August 2016
Accepted 1 December 2016

KEYWORDS
Free energy perturbation; Zwanzig's equation; Jarzynski's equation; cumulant expansion; QM/MM free energy simulation; indirect cycle



1. Introduction

Free energy perturbation (FEP) is one of the standard methods used to compute free energy differences from Monte Carlo or molecular dynamics (MD) simulations; such simulations are often referred to as free energy simulations (FES). FEP is also referred to as 'thermodynamic perturbation', 'exponential formula' or, after one of its first users, 'Zwanzig's equation'. While the well-known 1954 article by Zwanzig [1] is commonly considered the canonical reference for FEP, Jorgensen and Thomas have pointed out that the method is indeed much older than that [2]. FEP was the method used in the first calculation of a free energy difference in water from computer simulations [3]. Further, FEP is closely related, if not in essence identical, to Widom's test particle method [4]. Over the past years Bennett's acceptance ratio (BAR) method [5] has in many cases replaced FEP as one of the pillars of FES. Nevertheless, in addition to its historical importance, FEP continues to be an important theoretical foundation of FES.

In fact, in some cases FEP remains the only practically feasible method. Many free energy simulations utilising a mixed quantum chemical molecular mechanical (QM/MM) Hamiltonian rely on what has been called the 'indirect approach' [6]. Rather than computing the free energy difference $\Delta A^{\mathrm{QM/MM}}_{\mathrm{A}\rightarrow\mathrm{B}}$ between two states A and B at the QM/MM level of theory, one computes the corresponding free energy difference $\Delta A^{\mathrm{MM}}_{\mathrm{A}\rightarrow\mathrm{B}}$ at a lower level of theory, e.g. MM. This auxiliary result is then refined by the free energy differences of describing states A and B at the MM and QM/MM levels of theory. Since the free energy is a state function,

CONTACT Stefan Boresch ✉ stefan@mdy.univie.ac.at; H. Lee Woodcock ✉ hlw@mail.usf.edu
Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/00268976.2016.1269960.

the identity $\Delta A^{\mathrm{QM/MM}}_{\mathrm{A}\rightarrow\mathrm{B}} = -\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}_{\mathrm{A}} + \Delta A^{\mathrm{MM}}_{\mathrm{A}\rightarrow\mathrm{B}} + \Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}_{\mathrm{B}}$ holds, where, e.g. $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}_{\mathrm{A}}$ is the free energy difference between the MM and the QM/MM description of state A.

The primary rationale for the indirect approach is the reduction of computational cost; even today, conducting FES by large-scale QM/MM MD is an extremely expensive computational undertaking. By the same rationale, it is often desirable to compute free energy differences of the type $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}_{\mathrm{A}}$ and $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}_{\mathrm{B}}$ from MM simulations alone. To do so, configurations saved during such MM simulations are re-evaluated at the MM and QM/MM levels of theory; the energy differences $U^{\mathrm{QM/MM}} - U^{\mathrm{MM}}$ are the raw input for FEP to compute $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$. In this way, indirect QM/MM FES avoid any MD (or Monte Carlo) simulations at the high (QM/MM) level of theory. In such applications, there is no alternative to FEP. For example, the use of BAR would require a simulation at the other endpoint, which, for the computation of $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$, implies a simulation at the QM/MM level of theory. Thermodynamic integration [7] would be even worse since, in addition to the calculations at the MM and QM/MM end points, converged simulations at intermediate states between MM and QM/MM would be required.

The big unknown, of course, is whether results for $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ obtained from FEP are converged and, therefore, trustworthy in the indirect cycle. This is all the more relevant since for the reasons just outlined, in indirect QM/MM FES it is not practical to utilise intermediate states. When there are doubts about the convergence of a force field-based FES, it is always possible to insert intermediate states between the endpoints 0 and 1, i.e. $\Delta A_{0\rightarrow 1} = \sum_{i=1}^{n-1} \Delta A_{\lambda_i\rightarrow\lambda i+1}$, with λ_1 = state 0 and λ_n = state 1. Thus, checking whether a computed value for $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ is indeed converged is crucial for the correctness of an indirect QM/MM FES. Recent work by multiple groups, e.g. [8–10], confirmed that convergence of free energy differences of the type $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ obtained by FEP is highly questionable.

In 1997, Jarzynski showed that FEP is just a special case of a more general non-equilibrium theorem [11]. In the so-called Jarzynski equation (JAR), the non-equilibrium work $W_{0\rightarrow 1}$ required for forcing the system from state 0 to 1 formally replaces the energy difference $\Delta U_{0\rightarrow 1}$ in FEP. The energy difference can be viewed as the special case of the non-equilibrium work when switching from 0 to 1 instantaneously. We recently explored the use of Jarzynski's equation instead of FEP for the computation of $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ in indirect QM/MM FES [12]. Our data showed that even for rather small systems (less than 30 atoms treated by QM), the systematic error resulting from the use of FEP was non-negligible.

Thus, in the context of indirect QM/MM FES it would be highly useful if one could check whether results for $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ obtained by FEP or JAR were reliable. Obviously, FEP remains the computationally cheapest method; thus, if its convergence can be demonstrated for a particular case, it remains the method of choice. While JAR will work where FEP fails, its computational cost is significantly higher. Therefore, in actual applications tools indicating how many switches to carry out and/or what switching lengths are sufficient are required. At the core of FEP/JAR is the calculation of a nonlinear average, and several detailed mathematical analyses show that systematic errors arise when attempting to estimate such an average from a finite amount of data, e.g. [13,14]. In fact, a large body of work is concerned with the convergence of FEP and JAR, as well as sources of systematic errors in such calculations. An early work of Wood and co-workers [15] already observed that the leading term of this error, referred to as *sample size hysteresis* by the authors, is proportional to the variance of the energy differences $\sigma^2_{\Delta U_i}$. Occasionally, it is recommended that $\sigma_{\Delta U_i}$ be kept below 1–2 $k_{\mathrm{B}}T$, e.g. [14–16]. This is, however, a very small value, i.e. a very stringent criterion even for regular free energy simulations, let alone when attempting to compute $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$. Zuckerman and co-workers [13,17] derived an error estimate for FEP/JAR in terms of the first moments of $\exp(-\beta\Delta U_i)$ or $\exp(-\beta W_i)$; their work encompasses the result by Wood et al. as a special case. Since JAR is also heavily used to evaluate results of non-equilibrium work experiments, such as single molecule pulling experiments [18], the convergence of JAR was investigated in this context as well [19,20]. Recently, Daura et al. [21] applied and tested the corrections suggested in Ref. [19] in a simulation context. Probably the fullest analysis was carried out by Kofke and co-workers in a series of studies, e.g. [22–28].

The insights from these theoretical analyses seem to have been mostly ignored in practical applications of indirect QM/MM FES. While it is being recognised that FEP may not result in converged free energy differences $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ [8–10], until very recently there was a lack of alternatives. Curiously, one 'remedy' often used in applications is not even considered in most of the theoretical studies. The FEP/JAR expression can be expanded in terms of cumulants [29–31]. Provided the distribution of energy differences, ΔU, or non-equilibrium work values, W, is Gaussian, only the first two terms in the cumulant expansion, which are essentially functions of the arithmetic mean and the variance of the energy differences/work values, are needed. An example for the use of the first two cumulants instead of the full FEP/JAR equation in the context of QM/MM FES is, e.g. Ref. [32]. Clearly, mean and variance converge much more quickly than the exponential average at the core of FEP/JAR. However, it is not clear whether the distributions of energy values/work values are indeed Gaussian, and the error associated with applying the second-order cumulant approximation to the non-Gaussian case is unknown; cf., e.g. Ref. [16]. At the same time, it should be pointed out that many of the theoretical analysis also rely, at least in part, on the assumption that the distribution of energy differences/non-equilibrium work values is Gaussian.

This manuscript attempts to bridge the gap between the theoretical analyses of FEP/JAR and their use in QM/MM FES to compute $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$. Our recent exploration of the use of JAR instead of FEP [12] provides us with a body of data which can be used for this purpose, in particular since we have additional reference data available for several of these cases. We proceed in a twofold manner. First, we try to determine empirical guidelines from the available data-sets, e.g. correlating the deviations between obtained and reference results for the free energy difference of a given data-set with standard deviation of potential energy differences (FEP) or non-equilibrium work (JAR). Particular attention is paid to the performance of the second- order cumulant approximation compared to the full FEP/JAR equation, because of its frequent use in applications. The question of FEP convergence is explored further by means of model data, which were obtained by drawing random values from several probability densities. Second, among the various corrections and criteria resulting from the theoretical analyses, we chose the bias metric Π introduced by Kofke and co-workers [25,26] since all data needed for its computation are available from a routine statistical analysis of the raw data.

2. Theory

2.1. Background

The free energy difference between two states 0 and 1 when using FEP is given by

$$\Delta A_{0\rightarrow 1} = -k_{\mathrm{B}}T \ln \left\langle \exp\left[\frac{-\Delta U_{0\rightarrow 1}}{k_{\mathrm{B}}T}\right]\right\rangle_0 \qquad (1)$$

The symbols in Equation (1) have the usual meaning: k_{B} is Boltzmann's constant, T is the temperature and the angular brackets denote an ensemble average, i.e. averaging over all configurations sampled during the underlying MD or MC simulation. The energy difference $\Delta U_{0\rightarrow 1} = U_1 - U_0$ is obtained by re-evaluating the configurations sampled at state 0 — hence the subscript $<>_0$ — with the potential energy functions for state 0 and 1. In the context of computing $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ states 0 and 1 correspond to the MM and QM/MM descriptions of interactions for the system of interest.

JAR is obtained formally from Equation (1) by replacing $\Delta U_{0\rightarrow 1}$ by the non-equilibrium work $W_{0\rightarrow 1}$ required to force the system from state 0 to 1, i.e.

$$\Delta A_{0\rightarrow 1} = -k_{\mathrm{B}}T \ln \left\langle \exp\left[\frac{-W_{0\rightarrow 1}}{k_{\mathrm{B}}T}\right]\right\rangle_0 \qquad (2)$$

The subscript $\langle\rangle_0$ in Equation (2) indicates that the switching simulations were started from configurations (coordinates and velocities) sampled in equilibrium simulations at state 0 under conditions corresponding to the canonical ensemble. For a detailed description of non-equilibrium work methods with emphasis on practical details we refer the reader to Ref. [16].

2.2. Numerical considerations

Naive attempts to evaluate FEP/JAR may result in numerical instabilities. If the absolute values of energy differences ΔU_i are large, which is typically the case concerning energy differences between MM and QM representations of a system ($\Delta U_{0\rightarrow 1} = \Delta U_{\mathrm{MM}\rightarrow\mathrm{QM/MM}} = U^{\mathrm{QM/MM}} - U^{\mathrm{MM}}$), inserting them directly into the exponential function may result in numerical over- or underflow, even in double or quadruple precision floating point arithmetic. The identity

$$\begin{aligned}\Delta A &= -\frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp(-\beta\Delta U_i)\right)\\ &= -\frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp(-\beta(\overline{\Delta U}+\delta U_i))\right)\\ &= \overline{\Delta U} - \frac{1}{\beta}\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp(-\beta\delta U_i)\right) \qquad (3)\end{aligned}$$

avoids this pitfall. Here, $\delta U_i = \Delta U_i - \overline{\Delta U}$ and $\overline{\Delta U} = \frac{1}{N}\sum_{i=1}^{N}\Delta U_i$, and $\beta = 1/k_{\mathrm{B}}T$ as usual. In the following, it is assumed that any numerical over/underflow issues were taken care of by use of Equation (3). In the case of a very broad distribution, or a long tail, numerical issues might arise despite use of Equation (3). To guard against such worst cases, we have implemented our analysis routines based on Equation (3) in double, as well as in quadruple precision floating point arithmetic and check at least once for each data-set with the higher precision version. So far, double and quadruple precision arithmetic always led to identical results. A more general solution to numerical issues of this kind was described by Berg [33].

2.3. Cumulant expansion

Occasionally, it is suggested to avoid direct use of Zwanzig's or Jarzynski's equation and instead use cumulant expansions;[29–32], i.e.

$$\Delta A = \sum_{k=1}^{\infty} C_k \frac{(-\beta)^{k-1}}{k!} = \overline{\Delta U} - \beta\sigma^2_{\Delta U_i}/2 + \sum_{k=3}^{\infty} C_k \frac{(-\beta)^{k-1}}{k!} \tag{4}$$

On the right-hand side of Equation (4) the first two terms of the expansion, which are proportional to the mean $\overline{\Delta U}$ and the variance $\sigma^2_{\Delta U_i}$ of the energy differences ΔU_i, are given explicitly. In fact, as mentioned, if the distribution of the ΔU_i is Gaussian, the cumulants of third and higher order all vanish [34], i.e. in this case $\Delta A = \overline{\Delta U} - \beta\sigma^2_{\Delta U_i}/2$. Thus, if ΔU_i are distributed normally or approximately normally, then computing ΔA from solely $\overline{\Delta U}$ and $\sigma^2_{\Delta U_i}$ may well be advantageous compared to the full Zwanzig equation. Further, as one sees from a comparison of Equations (3) and (4), any numerical issues are avoided automatically. However, as pointed out, e.g. by Hummer and Dellago [16], in the case of marked deviations from a Gaussian distribution, the convergence properties of a cumulant expansion are not known.

2.4. Sources of error – the bias measure Π

The practical question we are concerned with is the error resulting during evaluations of the exponential averages in Equations (1) and (2) with finite amount of data. Our primary motivation is the computation of $\Delta A^{\mathrm{MM}\to\mathrm{QM/MM}}$, where circumventing convergence problems through insertion of intermediate states is not practical. Having to compute $\Delta A_{0\to1}$ in a single step/switch implies that the magnitude of fluctuations in $\Delta U_{0\to1}$ ($W_{0\to1}$) is not under our control.

A starting point of most theoretical analyses is to rewrite Equation (1) in terms of the probability density of $\Delta U_{0\to1}$ [35], i.e.

$$\Delta A_{0\to1} = -k_{\mathrm{B}}T \ln \int_{-\infty}^{+\infty} \exp(-\Delta U_{0\to1}/k_{\mathrm{B}}T) \times p_0(\Delta U_{0\to1})\, d(\Delta U_{0\to1}) \tag{5}$$

Obviously, JAR can be rewritten analogously. Here $p_0(\Delta U_{0\to1})$ denotes the probability density of the energy difference $\Delta U_{0\to1}$, with the subscript 0 indicating that sampling occurred at state 0. Through the change of variables, the FEP (JAR) equation is formally reduced to a one-dimensional integral. As an aside, we note that, in principle, one has to ascertain the convergence of the improper integral with limits $\pm\infty$ in Equation (5). If $p_0(\Delta U_{0\to1})$ is Gaussian, this is certainly the case; however, if $p_0(\Delta U_{0\to1})$ obeyed, for example, a Student's t-distribution, the integral in Equation (5), would diverge.

While we are primarily interested in one-sided approaches, i.e. the 0→1, MM to QM/MM direction, the potential errors arising in FEP/JAR can only be fully understood by also considering the backward direction 1→0 [14,25,26]. Obviously, Equations (1), (2) and (5) can also be formulated in the 1→0 direction, and it is insightful to plot the histograms/densities p of forward (0→1) and backward (1→0) energy differences/work values. We illustrate this in Figure 1 with data for a real system, blocked serine in water (cf. supporting information (SI)). Data for the energy differences, i.e. FEP are shown on the left, work values from (relatively short, $\tau_{\mathrm{switch}} = 0.5$ ps) switching simulations on the right. The thicker, solid line is used for the forward distribution since this is the only data one would have available in most applications; the distribution of the negative backward energy differences/work values is shown as a thinner, dashed line.

Pohorille et al. [14] summarised the requirement to compute $\Delta A_{0\to1}$ from microstates sampled at state 0 as follows: '...ΔA can be estimated reliably only if microstates representative of state 1 have been sampled from state 0'. Looking at the left-hand side of Figure 1, one sees that this condition is clearly not met for FEP; the forward distribution of energy differences (thick solid line) barely touches the distribution of (negative) backward energy differences (thin dashed line). By contrast, in the JAR case (right-hand side of Figure 1) the thick solid line representing the forward work values extends beyond the peak area of the distribution of backward work values. Thus, FEP is expected to fail for this case (left hand side of Figure 1), whereas JAR is likely to work, i.e. lead to a result that is free from major systematic error.

Figure 1 can also serve to illustrate the starting point of the more detailed analysis by Kofke and co-workers, e.g. [25,26]. In Equation (5) the integration is carried out from $-\infty$ to $+\infty$. However, looking at Figure 1 one sees that the actual energy differences sampled are restricted to a finite interval $[\Delta U^{\mathrm{min}}_{0\to1}, \Delta U^{\mathrm{max}}_{0\to1}]$. Because of the nature of the integrand, the upper limit poses no problem, but the lower limit does. Because of limited sampling, the exact expression Equation (5) is approximated by

$$\Delta A_{0\to1} \approx -k_{\mathrm{B}}T \ln \int_{\Delta U^{\mathrm{min}}_{0\to1}}^{+\infty} \exp(-\Delta U_{0\to1}/k_{\mathrm{B}}T) \times p_0(\Delta U_{0\to1})\, d(\Delta U_{0\to1}) \tag{6}$$

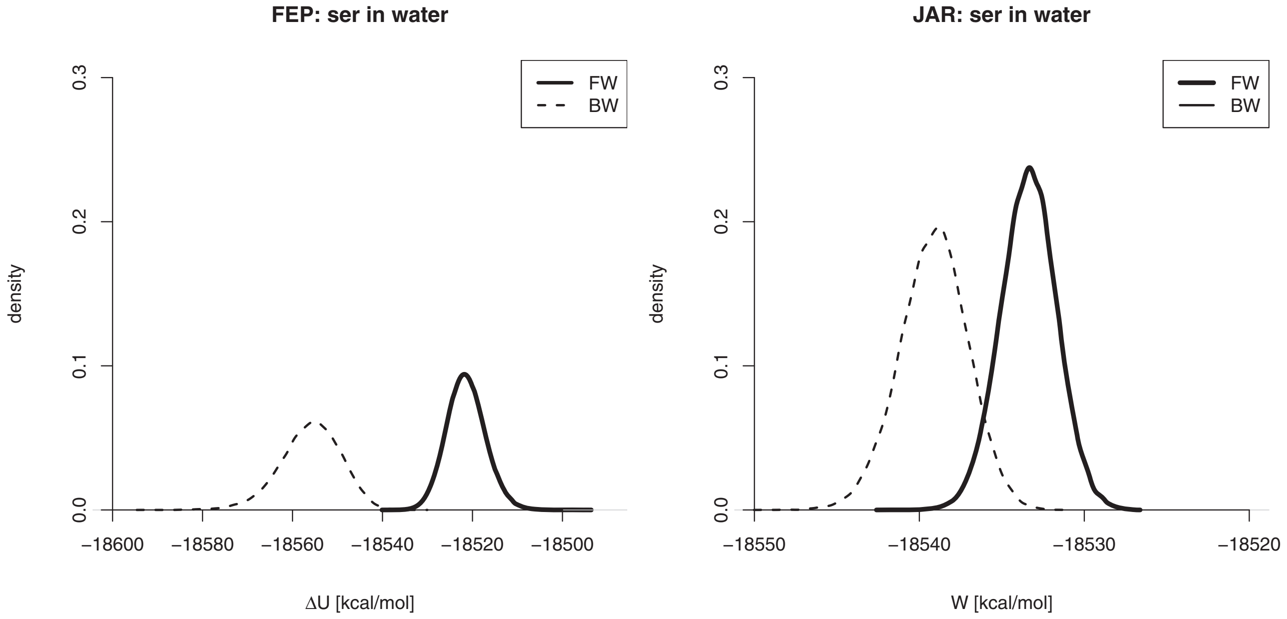


Figure 1. Forward (FW)/backward (BW) distributions of ΔU (FEP) and W (JAR) for blocked serine in water.

In order for the integral Equation (5) to converge, $p(\Delta U_{0\to1})$ has to approach zero for $\Delta U_{0\to1}\to-\infty$ faster than the competing term $\exp(-\beta\Delta U_{0\to1})$; therefore, in principle, the lower bound $-\infty$ can be replaced by some suitable, finite lower bound. By considering the forward and backward direction for FEP or JAR, Kofke and co-workers showed that a systematic error arises if the finite lower bound $\Delta U^{\min}_{0\to1}$ in Equation (6) is greater than the negative tail of the distribution of $p(-\Delta U_{1\to0})$ (FEP) or $p(-W_{1\to0})$ (JAR) for the corresponding backward process. As one sees in the left-hand side of Figure 1, $\Delta U^{\min}_{0\to1} \approx -18,540$ kcal/mol, whereas the most negative value of the backward distribution is extending down to almost $-18,600$ kcal/mol. Thus, all of this range, which was not sampled in the simulation of state 0, contributes to systematic error when attempting to compute $\Delta A_{0\to1}$ by FEP. Clearly, the situation is much better for JAR (right-hand side of Figure 1). Here the forward distribution extends slightly below $-18,540$ kcal/mol, and the most negative value of the backward distribution of work values is at $\approx -18,550$ kcal/mol. While there will be some error, it is likely to be significantly lower than in the FEP case.

The argumentation just sketched was carried through in full detail in Ref. [26], leading to criteria concerning the degree of overlap between forward and backward distributions required for free energy differences that can be considered free from systematic error (Equation (4) of [26]). In addition, Wu and Kofke also pursued a second approach, based on ideas from information theory. They showed that entropies – in the information theory sense – correspond to the dissipated work in a non-equilibrium process, i.e.

$$\begin{aligned} s_{0\to1} &= \beta(\langle W_{0\to1}\rangle_0 - \Delta A)\\ s_{1\to0} &= \beta(\langle -W_{1\to0}\rangle_1 + \Delta A) \end{aligned} \tag{7}$$

These are then used to define bias measures Π

$$\begin{aligned} \Pi_{0\to1} &= \sqrt{\frac{s_{0\to1}}{s_{1\to0}}\mathbf{W}_{\mathrm{L}}\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2s_{0\to1}}\\ \Pi_{1\to0} &= \sqrt{\frac{s_{1\to0}}{s_{0\to1}}\mathbf{W}_{\mathrm{L}}\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2s_{1\to0}} \end{aligned} \tag{8}$$

Here, $\mathbf{W}_{\mathrm{L}}$ is the Lambert W function, defined as the solution for w in $x = we^w$, and M is the number of data points, i.e. the number of energy differences or work values. Based on heuristics, Wu and Kofke argue that at least $\Pi > 0$ is required for free energy differences that can be expected to be free from systematic error.

The result just outlined implies two conditions to hold, a Gaussian distribution of work values, and a phase space subset relation between the 0→1 and 1→0 directions (see, e.g. Figure 1a in [26]). Our real world data fulfill neither criterion; even by visual inspection of Figure 1 one can discern that distributions of both energy differences and work values differ slightly from the Gaussian case. Also, the subset relation is not fulfilled; instead, there is, at best, a partial overlap relation. Thus, applying Equation (8) is strictly speaking wrong. However, attempting to use it as criterion for potential bias (error) still seems justified, in particular when contrasted with the widespread practice of either ignoring potential error completely, or approximating the FEP/JAR equation by the second-order cumulant approximation, even though the underlying distributions are rarely Gaussian.

Provided one has forward/backward energy differences or work values, criterion Equation (8) is straightforward to apply since all data are typically already computed. However, when attempting to compute $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$, the backward data will typically not be available, which seems to limit the utility of Equation (8). The derivations in [26] are a generalisation of a more special case derived a year earlier [25]. From a theoretical point of view, the drawback of this special case is that it assumes that the amount of dissipated work is identical for the forward and the backward direction. Thus, this simplified model does not reflect that the errors for the forward and backward direction may be different. An extreme example would be particle insertion/deletion, where only the former will work in practice, whereas the latter will lead to erroneous results. On the other hand, as illustrated in Figure 1, in our case the forward and backward distributions are relatively similar in terms of the width of the distributions. Since we are applying Equation (8) in a non-rigorous fashion, we can reasonably assume that the forward and backwards distributions have an identical spread, i.e. $s_{0\rightarrow 1} = s_{1\rightarrow 0}$. Applying this assumption yields the fundamental criterion applied through this study (cf. Ref. [25]):

$$\Pi = \sqrt{\mathbf{W}_{\mathrm{L}}\left[\frac{1}{2\pi}(M-1)^2\right]} - \sqrt{2\beta(\langle W\rangle_0 - \Delta A)} \quad (9)$$

Wu and Kofke suggest for the looser measure Equation (9) $\Pi > 0.5$ to have some confidence that the corresponding free energy difference ΔA be free from systematic error [25]. We stress that applying criterion Equation (9) to work values from relatively short switching simulations, let alone energy differences in FEP, violates assumptions made in its derivation. However, the overall goal of this study is to provide ways to estimate the reliability of free energy differences by FEP and JAR. Since we have available independent reference free energy differences, which we consider trustworthy, we can test empirically how the Π criterion performs when applied outside its strict area of validity. Also, as shown in the more general derivation [26], requiring $\Pi > 0.5$ provides some margin of error.

3. Methods

As was already outlined, our primary focus here is to explore under what circumstances FEP/JAR leads to reliable results, in particular in situations when the fluctuations of $\Delta U/W$ cannot easily be kept low. In addition, we wanted to investigate the applicability of the second-order cumulant expansion, as well as the sensitivity of the bias measure Π.

Table 1. Overview of various model distributions used.

N1	Normal distribution $\mu = 0, \sigma = 0.5$	
N2	Normal distribution $\mu = 0, \sigma = 1$	
N3	Normal distribution $\mu = 0, \sigma = 2$	
N4	Normal distribution $\mu = 0, \sigma = 3$	
G1	Gumbel distribution[a] $\mu = 0, \beta = 1$	$\sigma = 1.28$[c]
G2	Gumbel distribution[a] $\mu = 0, \beta = 4$	$\sigma = 5.13$[c]
T10′	'Truncated' Student's *t*-distribution with 10 degrees of freedom; only values $-20 \leq x \leq +20$ are considered	$\sigma = 1.12$[c]
B	Beta distribution[b] $p_{\mathrm{Beta}}(x/5, 15, 4)$	$\sigma = 0.46$[c]

[a] $p_{\mathrm{Gumbel}}(x, \mu, \beta) = 1/\beta \exp\left[-\left(z + e^{-z}\right)\right], z = (x - \mu)/\beta$.

[b] $p_{\mathrm{Beta}}(x, a, b) = x^{a-1}(1-x)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b}, x \in (0, 1)$.

[c] Theoretical standard deviation for chosen parameters of the probability density.

Our starting point is the FEP/JAR master equation as rewritten in Equation (5). Given, as in our first set of tests, an assumed, analytic probability density $p(\Delta U_{0\rightarrow 1})$ (*vide infra*), Equation (5) can be integrated numerically. We employed the QUADPACK numerical quadrature routines for this purpose [36]. We further used the R language [37] to sample ten million (10,000,000) random numbers from several probability densities. These random numbers were considered to be energy differences $\Delta U_{0\rightarrow 1}$ as would be obtained in free energy simulations, and the corresponding 'free energy difference' was computed using both the full FEP equation (Equation (1)), as well as the cumulant expansion Equation (4) truncated after the second term. In addition, the bias measure Π was computed (Equation (9)). Units are, of course, arbitrary, but could be assumed to be kcal/mol.

The probability distributions considered are summarised in Table 1; some are also plotted in Figure 2. They consisted of four normal distributions labelled N1–N4 with $\mu = 0$, but varying σ, two Gumbel distributions (G1,G2) [38], a 'truncated' Student's *t*-distribution (T10'), as well as a Beta distribution (B) [38]. The Gumbel or Extreme Value Type I distribution was chosen since it is an asymmetric distribution which, nevertheless, does not deviate too dramatically from a normal distribution. Further, through the scale parameter β it can be made narrower or broader at will. As already pointed out, Equation (5) diverges for the Student's *t*-distribution. Therefore, we 'truncated' the Student's *t*-distribution with ten (10) degrees of freedom by somewhat arbitrarily defining $p_{t,10} = 0$ for $|x| > 20$. It should be noted that the probability for x to be outside this range is extremely small; e.g. for the 10,000,000 random samples not a single value lay outside this region. Based on visual inspection alone, T10' would be very difficult to discern from a standard normal

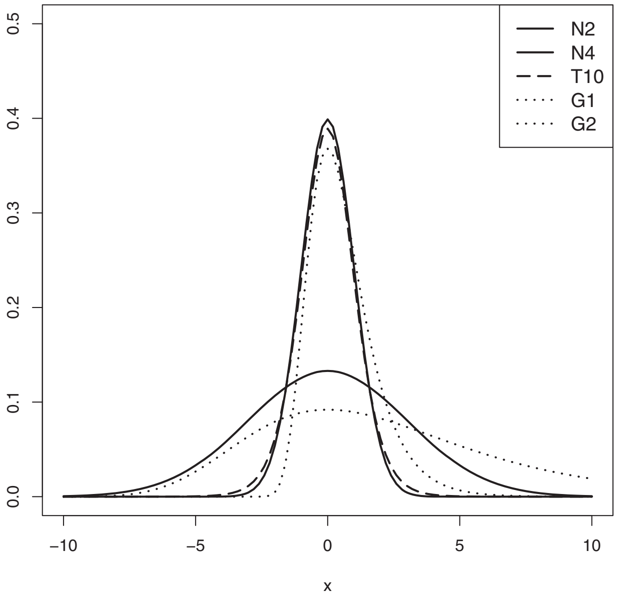


Figure 2. Probability density of some of the model distributions used. See Table 1 for further details.

distribution (i.e. N2). The Beta distribution B, $p_{\mathrm{Beta}}(x/5, 15, 4)$, is somewhat of an outlier. First, it is strictly bounded within $0 < x < 5$; within this range it consists of a single sharp narrow peak, which, nevertheless, is asymmetric. Thus, while T10' probes for differences between a true normal distribution and a distribution that appears very similar to a normal distribution, B is a test case for an asymmetric distribution, i.e. clearly deviating from the Gaussian case, yet narrow and strictly bounded. Finally, we stress that our choice of densities was motivated by considerations of shape and width, and does in no way indicate an assumption that true distributions of energy differences $p(\Delta U_{0\to1})$ obey any of these laws.

A second set of data was accumulated during our ongoing work exploring the use of JAR instead of FEP to compute $\Delta A^{\mathrm{MM}\to\mathrm{QM/MM}}$ [12]. It consists of all data generated for Refs. [12] and [39]. Specifically, the systems include small- to medium- sized organic molecules in the gas phase (methanol, methanethiol, ethane, methyl formate, *n*-butane, *bis*-2-chloroethylether, phenyl-trifluoroethyl-ether, 1-octanol, triacetyl glycerol, as well the blocked amino acids alanine and serine, viz. *N*-acetyl-alanine-methylamide and *N*-acetyl-serine-methylamide). For the two blocked amino acids, data for various force fields, as well as various parameters for the semi-empirical QM method are available. Furthermore, we have results for ethane, methanol and the two blocked amino acids in aqueous solution, with the solute being described either by MM or a semi-empirical QM method; see SI for details. Solvent waters were always described classically. The protocols are identical or at least very similar to the ones described in Ref. [12]. For each of the systems, we computed $\Delta A^{\mathrm{MM}\to\mathrm{QM/MM}}$ based on data generated with various methods/protocols, i.e. energy differences $\Delta U_{0\to1}$ for FEP and non-equilibrium work values $W_{0\to1}$ obtained from switching simulations of varying switching lengths. Whenever the convergence of the result was in doubt, we also carried out calculations at the high level of theory (a semi-empirical QM method, cf. Ref. [12] and SI) to obtain backward energy differences $\Delta U_{1\to0}$ and work values for backward switches $W_{1\to0}$. Thus, we could use Bennett's acceptance ratio method [5] as well as Crook's equation [40] in order to compute reference results.

In the present work, the focus is not on demonstrating the superiority of JAR compared to FEP when computing $\Delta A^{\mathrm{MM}\to\mathrm{QM/MM}}$, and the detailed numbers are of minor importance. The relevance of the data lies in the fact that the results were obtained with several protocols and methods. Thus, the results obtained from the most extensive protocols – either JAR with the longest switching protocol, or the Bennett/Crooks results – can be viewed as reference results, and used to gauge the accuracy of the other methods/less expensive protocols, as well as, to compare to results obtained with the second order cumulant approximation. Further, we can check whether the bias measure Π detects biased results.

Aside from some of the reference results, all free energy differences were obtained with FEP/JAR. The quantities of interest in the present context are (i) the variance $\sigma^2_{\Delta U_i}$ of the energy differences/work values, (ii) the deviation $\delta\Delta A_{\mathrm{REF}}$ of the result obtained for a particular method/protocol from the reference result for this system, and (iii) $\delta\Delta A_{C2}$, the difference between the result obtained with the second-order cumulant approximation (first two terms in Equation (4)) and the FEP/JAR result (Equations (1) and (2)), respectively, for that particular data-set. In addition to metrics (i)–(iii), we also probed for sample-size hysteresis (iv), cf. Ref. [15]. To monitor sample-size hysteresis, free energy differences were computed using all available data with Equations (1) or (2). Then, the raw data (100,000 or 200,000 energy differences, 10,000 or 20,000 work values) were divided into 10 blocks and the free energy differences were computed for each of the blocks. The difference between ΔA obtained for the full data and the arithmetic mean of the 10 block results is labelled $\delta\Delta A_{\mathrm{BLOCK}}$, and we consider deviations $\delta\Delta A_{\mathrm{BLOCK}} > 0.5$ kcal/mol as indication of sample size hysteresis. Finally, (v) Π was computed according to Equation (9). Data for these five metrics is listed in Table 1 of SI.

Bennett's acceptance ratio method and Crook's equation can also be approximated by a cumulant

Table 2. Comparison of results for random data for various distributions.

	Num.Int.[a]	FEP[b]	C2[c]	Π[d]
N1	− 0.21	− 0.21	− 0.21	4.37
N2	− 0.82	− 0.84	− 0.84	3.53
N3	− 3.35	− 3.35	− 3.35	1.85
N4	− 7.55	− 7.71	− 7.55	0.12
G1	− 0.25	− 0.25	− 0.80	3.54
G2	− 4.74	− 4.73	− 19.76	0.35
T10′	− 7.20	− 5.24	− 1.05	1.02
B	3.74	3.74	3.77	4.38

[a] Numerical integration of Equation (5).
[b] Result obtained from inserting the 10 million random data into the FEP master equation (Equation (1)).
[c] Result obtained from the first two terms of the cumulant expansion in Equation (4).
[d] Equation (9).

expansion [31]:

$$\Delta A \approx \frac{1}{2}\left(\Delta U_{0\to1} - \Delta U_{1\to0}\right) - \frac{1}{12k_BT}\left(\sigma^2_{\Delta U_{0\to1}} - \sigma^2_{\Delta U_{1\to0}}\right),$$

where for Crook's equation ΔU is replaced by W. The first term in Equation (10) is just the two-sided linear response approximation (LRA), i.e. the arithmetic mean of forward and backward energy differences. Warshel and co-workers occasionally employed LRA to compute $\Delta A^{\mathrm{MM}\to\mathrm{QM/MM}}$ [41]. For the cases where we have data available, we also check the agreement between using Equation (10) and the full Bennett's/Crook's equation.

4. Results

4.1. Analytical model data

Results for hypothetical data corresponding to probability distributions of Table 1 are shown in Table 2. We compare results obtained from the numerical integration of Equation (5) (column Num.Int.) with the full FEP equation, i.e. inserting the 10 million random data into Equation (1), as well as the second-order cumulant approximation (column C2). In most cases the FEP and Num.Int. results agree well. There is a small deviation for the normal distribution with the largest σ (N4). The FEP results for the two Gumbel distributions, even the very broad G2, as well as the Beta distribution (B) are also in excellent agreement with the reference result. By contrast, for the 'truncated' Student's t-distribution with 10 degrees of freedom (T10') the error is large, almost 2 units. On the one hand, T10' is a problematic case since the full integral from $-\infty$ to $+\infty$ diverges. Thus, the choice of bounds ±20 is arbitrary, and the value of the integral varies considerably for narrower or wider bounds. On the other hand, though most likely not relevant for practical FEP/JAR calculations, the example drives home the point about large, systematic deviations resulting from limited sampling. The most negative value in the T10' random sample consisting of ten million data points was −14.82, considerably more positive than the arbitrary lower bound of −20.

Table 3. Effect of using subsets of data for FEP and C2.

	N3[a]			N4[b]		
Data points	FEP[c]	C2[d]	Π[g]	FEP[c]	C2[d]	Π[g]
10,000,000	− 3.35	− 3.36	1.85	− 7.71	− 7.55	0.12
1,000,000	− 3.25	− 3.35	1.46	− 7.14	− 7.53	− 0.13
100,000	− 3.38	− 3.35	0.91	− 7.26	− 7.51	− 0.66
10,000	− 3.12	− 3.39	0.50	− 6.55	− 7.42	− 0.96
1000	− 2.93	− 3.52	0.00	− 5.33	− 7.04	− 1.11
	G1[e]			G2[f]		
10,000,000	− 0.25	− 0.80	3.54	− 4.73	− 19.74	0.35
1,000,000	− 0.25	− 0.80	3.10	− 4.71	− 19.86	− 0.10
100,000	− 0.25	− 0.80	2.62	− 4.79	− 19.81	− 0.61
10,000	− 0.27	− 0.80	2.06	− 4.75	− 19.75	− 1.13
1000	− 0.24	− 0.82	1.46	− 4.92	− 18.54	− 1.82

[a] Reference result by numerical integration of Equation (5) for N3 is −3.35.
[b] Reference result: −7.55.
[c] Result obtained from inserting the 10M random data into the FEP master equation (Equation (1)).
[d] Result obtained from the first two terms of the cumulant expansion in Equation (4).
[e] Reference result: −0.25.
[f] Reference result: −4.74.
[g] Equation (9).

With the second-order cumulant approximation (column C2), all normal distributions, including N4, agree perfectly with the reference result. This is not surprising since in this case the approximation becomes exact [14]. In fact, for N4 the C2 result is superior to the FEP result. In addition, the C2 result for B is in very good agreement with the reference result. For the remaining cases, G1, G2 and T10', the second-order cumulant approximation fails. While the result for G1 (−0.80 vs. −0.25) might still be considered acceptable, the G2 and T10' results are completely wrong. Table 2 also contains the bias measure Π. For N4, where FEP starts to fail, the value $\Pi = 0.12$ is much lower than, e.g. for N1–N3 with $\Pi \gg 1$. For G2 $\Pi = 0.35$ is relatively low, failing the 'safe' threshold 0.5 of Ref. [25], although the 'free energy difference' obtained by FEP is correct. For T10' the Π measure of 1.02 erroneously does not indicate problems, but as just discussed this model distribution is to some degree questionable.

In Table 3 we examine the influence of using fewer data points on the FEP and C2 results for distributions N3, N4, G1 and G2. Starting from the full set of 10,000,000 values, we diluted the data by factors of 10 until only 1000 values were used. For the two relatively broad normal distributions N3 and N4 we see that the FEP result gradually deteriorates; in particular the results obtained

Table 4. Correlation matrix between five characteristics of using FEP/JAR obtained from 38 raw data.

	$\sigma^2_{\Delta U_i}$ [a]	$\delta\Delta A_{\text{BLOCK}}$ [b]	$\delta\Delta A_{\text{REF}}$ [c]	$\delta\Delta A_{\text{C2}}$ [d]	Π [e]
$\sigma^2_{\Delta U_i}$	1.00	0.65	0.79	−0.91	− 0.91
$\delta\Delta A_{\text{BLOCK}}$		1.00	0.39	−0.37	− 0.70
$\delta\Delta A_{\text{REF}}$			1.00	−0.71	− 0.70
$\delta\Delta A_{\text{C2}}$				1.00	0.69
Π					1.00

[a] Variance $\sigma^2_{\Delta U_i}$ of the $\Delta U/W$ raw data entering the FEP (Equation (1)) and JAR (Equation (2)) master equations.
[b] Estimate of sample size hysteresis, $\Delta A - 1/10\sum_{i=1}^{10}\Delta A_i$, where ΔA is the FEP/JAR result obtained for all available data, and ΔA_i denotes a FEP/JAR free energy difference obtained for 1/10 of the full data, i.e. after blocking the full data into 10 subsets (blocks); cf. Ref. [15].
[c] Difference between the FEP/JAR result (using all available data) and a reference result obtained with a more elaborate protocol.
[d] Difference between the result obtained with the second-order cumulant approximation (C2, first two terms in Equation (4)) and the FEP/JAR result of Equations (1) and (2), respectively.
[e] Equation (9).

with just 1000 and 10,000 data points are noticeably in error. For N3, this is reflected by decreasing Π values: whereas the FEP result of -3.38 obtained with 100,000 data points ($\Pi = 0.91$) is still in good agreement with the reference value, the results obtained with 10,000 ($\Pi = 0.5$) and 1000 ($\Pi = 0.0$) data points, respectively, deviate noticeably. For N4 even the full data set leads to a low Π value (cf. above); for all reduced data sets Π is negative. The C2 results, on the other hand, remain almost in perfect agreement with the reference results for as few as 10,000 data points; only when using just 1000 data values, one can discern deviations. For the two Gumbel distributions, the C2 result was wrong even for the full set of data; obviously, the results do not improve when reducing the number of data used in evaluation. These wrong results, however, appear to be well converged even when using a subset of the data points. The FEP results are also very little affected when reducing the number of data points; only the G2 result with just 1000 data points deviates discernibly from the reference value. The good FEP results for G2 are slightly surprising, as the distribution is very broad. Further, Π for all reduced G2 data sets is negative, indicating that there is the possibility of systematic errors in the result. In fact, the good agreement of the FEP results with the reference value may be slightly fortuitous since the longer tail of the Gumbel distributions is towards the more positive values which have negligible weight in the exponential average needed for FEP.

4.2. Real MM→ QM/MM data

Table 4 provides condensed results for the real data from various calculations attempting to compute $\Delta A^{\text{MM}\rightarrow\text{QM/MM}}$. The raw data used to generate it and some additional technical details can be found in Table 1 of SI. In this study we are looking for criteria to gauge when the result of a FEP/JAR calculation can be expected to be reliable. Shown in Table 4 are the correlation coefficients between the variance $\sigma^2_{\Delta U_i}$ of the energy differences (FEP)/work values (JAR), a measure for sample size hysteresis $\delta\Delta A_{\text{BLOCK}}$, the deviation of the FEP/JAR result from the best available reference result $\delta\Delta A_{\text{REF}}$, the difference between the second cumulant approximation and the full FEP/JAR result $\delta\Delta A_{\text{C2}}$, and the bias measure Π (Equation (9)); cf. Section 3. The strongest (negative) correlations (-0.91) exist between $\sigma^2_{\Delta U_i}$ and $\delta\Delta A_{\text{C2}}$ on the one hand, and $\sigma^2_{\Delta U_i}$ and Π on the other hand. This indicates that the second-order cumulant approximation is performing more poorly as the variance of the raw data (energy differences/work values) increases. Similarly, with increasing $\sigma^2_{\Delta_i}$, Π becomes more negative. The correlation between $\delta\Delta A_{\text{REF}}$ and $\sigma^2_{\Delta U_i}$ is +0.79, followed by correlations of -0.71 and -0.70 between $\delta\Delta A_{\text{REF}}$ and $\delta\Delta A_{\text{C2}}$ and Π, respectively.

Arguably, the most relevant correlations are those involving $\delta\Delta A_{\text{REF}}$. Theoretical considerations [13–15] indicate that the leading term of the error of FEP/JAR with finite amount of data is proportional to $\sigma^2_{\Delta U_i}$. Indeed, the correlation between $\sigma^2_{\Delta U_i}$ and $\delta\Delta A_{\text{REF}}$ (0.79) is fairly strong. The bias measure Π is applied by us to cases which lie outside its strict area of validity; nevertheless, the correlation between $\delta\Delta A_{\text{REF}}$ and Π (-0.70) is quite strong. By contrast, while we find correlation (+0.65) between $\sigma^2_{\Delta U_i}$ and our estimate of sample size hysteresis $\delta\Delta A_{\text{BLOCK}}$, the correlation coefficient between $\delta\Delta A_{\text{REF}}$ and $\delta\Delta A_{\text{BLOCK}}$ is low (+0.39). This weak correlation confirms observations during ongoing work that the absence of sample size hysteresis (as quantified by, e.g. $|\delta\Delta A_{\text{BLOCK}}| \leq 0.5$ kcal/mol) is a necessary, but not a sufficient criterion for excluding the possibility of a systematic error in the result.

The results of this correlation analysis (Table 4) suggest that both $\sigma^2_{\Delta U_i}$ and Π should have some predictive utility concerning deviations from the respective reference result. In Figure 3 we show a scatter plot between the variance $\sigma^2_{\Delta U_i}$ of the energy differences/work values and $|\delta\Delta A_{\text{REF}}|$, i.e. the absolute value of the difference of a particular result from its reference result. The linear regression line for the data is shown as well (solid line, correlation coefficient is 0.83). The dotted, horizontal line indicates a deviation (error) of $k_{\text{B}}T \approx 0.6$ kcal/mol at room temperature. The solid and dotted lines intersect at $\sigma^2_{\Delta U_i} \approx 5.85$. This suggests that for our data standard deviations of ΔU or W which are less than or equal to about 2.4 kcal/mol ($\approx 4k_{\text{B}}T$ at room temperature) result in errors $|\delta\Delta A_{\text{REF}}| \leq 0.6$ kcal/mol. Obviously, this is at best a crude rule of the thumb. In Figure 3 there are

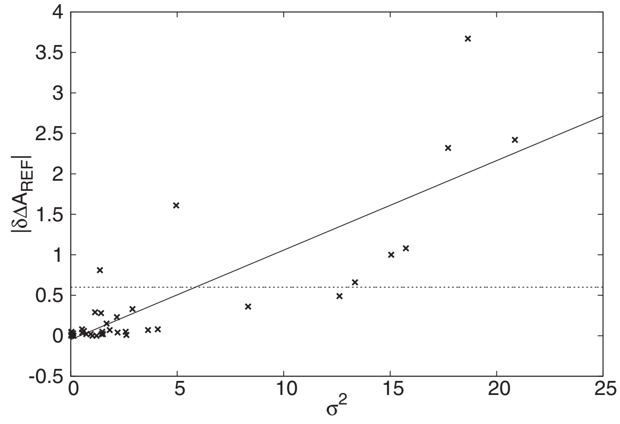


Figure 3. Scatter plot of $|\delta\Delta A_{\text{REF}}|$ vs. variance $\sigma^2_{\Delta U_i}$ of the energy differences/work values. The horizontal dotted line indicates an error of $k_\text{B}T \approx 0.6$ kcal/mol. The solid line shows the linear regression for the data, $|\delta\Delta A_{\text{REF}}| \propto -0.05 + 0.11\sigma^2_{\Delta U_i}$, which has a correlation coefficient of 0.83.

two outliers in the 'allowed' $\sigma^2_{\Delta U_i} < 5.85$ range, for which $|\delta\Delta A_{\text{REF}}| > 0.6$ kcal/mol. In two cases significantly larger values of $\sigma^2_{\Delta U_i}$ still result in an error below $k_\text{B}T$. The two outliers with larger than 'expected' error concern the same system, *bis*-2-chloroethylether, specifically the FEP and the JAR result based on the shortest switching protocol used.

In Figure 3 eight systems have $|\delta\Delta A_{\text{REF}}| >$ 0.6 kcal/mol. The exact numbers can be found in Table 1 in SI, which also contains the bias measure Π. In seven cases, $\Pi \ll 0.5$; this includes the FEP result for *bis*-2-chloroethylether. For the corresponding JAR result, $\Pi = 0.61$, i.e. just above the 0.5 threshold considered safe. Given that most Π values found for the real data are >1 (28 out of 38, see Table 1 in SI), the relatively low value obtained in this case provides at least some warning. There is one more comparatively low Π value (0.7), corresponding to the JAR result for blocked serine in water, i.e. the system depicted on the right in Figure 1. Finally, among the 38 data-sets there is one very low Π value, which does not correspond to a large deviation from the reference result. Nevertheless, the bias measure Π performs better than the ad hoc $\sigma^2_{\Delta U_i}$ criterion. The special case of *bis*-2-chloroethylether and the 'false positive' are discussed further in Section 4.3.

In eight cases the free energy differences $\Delta A^{\text{MM}\rightarrow\text{QM/MM}}$ were not only computed with FEP/JAR, but also with Bennett's acceptance ratio method and Crook's equation. These free energy differences were computed with the full Bennett/Crooks equations, as well as their first- and second-order cumulant approximation Equation (10) [31]. For two of the Bennett and one of the Crooks results, use of LRA (i.e. only the first term of Equation (10)) incurred a sizable error (1.5–3 kcal/mol) compared to the use of the full equation. The largest error using the two-sided second-order cumulant expression was 0.6 kcal/mol, i.e. on the order of $k_\text{B}T$, which should be acceptable in most applications.

4.3. *Additional remarks concerning the Π measure and distributions encountered*

Given that the bias measure Π in most cases works well to detect systematic error, despite being applied in cases violating the assumptions underlying its derivation, we carried out some additional checks to investigate its utility. First, we repeat that we employ the Π criterion in its simplest form [25], which assumes equal width of the forward and backward distributions of work values. As already mentioned in Section 2 and illustrated by the case depicted in Figure 1, we assume any error resulting from this to be low as forward and backward distributions for MM ↔ QM/MM tend to be of comparable width. If this is the case, the entropies $s_{0\rightarrow1}$ and $s_{1\rightarrow0}$ are similar, and any difference between Equations (8) and (9) should become small. Indeed, for the work values for the blocked serine in water (shown on the right in Figure 1) this is the case: $\Pi = 0.71$ when computed with Equation (8), compared to $\Pi = 0.70$ when computed with Equation (9).

In general one does not expect energy differences ΔU_i needed for FEP to obey a normal distribution. Nevertheless, the Π criterion does well at identifying systematically wrong FEP results. Most likely, in these cases the entropy $s = \langle\Delta U\rangle - \Delta A$ becomes so large that the details of the underlying distribution cease to matter, i.e. extremely large numbers of data points would be required to result in $\Pi > 0$, let alone $\Pi > 0.5$. Indeed, as one can see in Figure 1 of SI, most ΔU_i distributions are relatively broad, i.e. s is large.

In the limit of nearly reversible switches, the work values obtained in JAR calculations are expected to be distributed normally [11,19]. In Figure 1 we showed one example of forward/backward distributions of work values; small deviations from the Gaussian case case are clearly discernible. However, there are some systems with much more pronounced deviations (see Figure 1 of SI), including *bis*-2-chloroethylether. Since this is the one case for which the $\Pi \geq 0.5$ threshold failed to detect a systematic bias, and since this molecule is the most difficult case we encountered to date, we find it instructive to present the available data in full detail. The system was part of the model set used in Ref. [39]. There, we observed that convergence was poor even when using JAR, since sampling with the standard MM force field [42] and with the semi-empirical SCC-DFTB method (cf. SI for methodological details) visits rather different regions of conformational

Table 5. Detailed results for *bis*-2-chloroethylether. The MM calculations employed either the CHARMM CGENFF force field (2CLE) [42], or a reparameterised force field obtained with the GAAMP protocol (2CLE-2) [43].

	2CLE				2CLE-2			
	Π[a]	$\sigma^2_{\Delta_i}$[b]	$\lvert\delta\Delta A_{C2}\rvert$[c]	$\lvert\delta\Delta A_{REF}\rvert$[d]	Π[a]	$\sigma^2_{\Delta_i}$[b]	$\lvert\delta\Delta A_{C2}\rvert$[c]	$\lvert\delta\Delta A_{REF}\rvert$[d]
FEP	0.09	4.95	1.08	1.63	−0.16	8.94	1.64	1.23
JAR, 0.1 ps	0.61	1.37	1.76	0.82	0.29	3.70	0.44	0.26
JAR, 0.5 ps	0.52	2.04	1.37	0.24	1.28	2.58	0.37	0.05
JAR, 1.0 ps	0.62	1.16	1.08	0.32	1.62	2.06	0.40	0.03

[a] Equation (9).
[b] Variance $\sigma^2_{\Delta U_i}$ of the $\Delta U/W$ raw data entering the FEP (Equation (1)) and JAR (Equation (2)) master equations.
[c] Difference between the result obtained with the second-order cumulant approximation and the full FEP/JAR result.
[d] Difference between the FEP/JAR result and the reference result obtained using Crook's equation [40].

(=dihedral angle) space. Changes of conformation from the region preferred by MM to the preferred region of the semi-empirical method occurred very rarely during the relatively short switching simulations (0.1–1 ps). We, therefore, repeated the MM calculations with a reparameterised force field, obtained using the GAAMP protocol [43], and for longer switching times (0.5 and 1 ps, but not for 0.1 ps) satisfactory results were obtained [39]. Data obtained with the regular force field are labelled 2CLE, those obtained with the reparameterised force field are labelled 2CLE-2, and relevant data are summarised in Table 5. The corresponding distributions of energy differences and work values are shown in Figure 4

In Figure 4 one sees immediately that for 2CLE the work distributions become bimodal as the switching length increases. At the shortest switching length (0.1 ps), the work distribution is noticeably skewed towards negative values; this tail develops into separate peaks at switching lengths of 0.5 and 1 ps. This is also reflected in the variance of the work values. As one sees in Table 5, $\sigma^2_{\Delta U_i}(\text{JAR, } 0.1\,\text{ps}) = 1.37$, whereas $\sigma^2_{\Delta U_i}(\text{JAR, } 0.5\,\text{ps}) = 2.04$. Similarly, Π is actually lower for the 0.5 ps switches than in the case of 0.1 ps switches, and remains at the comparatively low value of 0.62 for the 1 ps switching time. Obviously, when attempting to compute $\Delta A^{\text{MM}\rightarrow\text{QM/MM}}$ for 2CLE, i.e. based on MM simulations employing the regular CHARMM force field, the assumption of a more or less Gaussian distribution of work values is severely violated. With the reparameterised force field (2CLE-2), the situation is improved (see right-hand side of Figure 4). Nevertheless, all work distributions do have a shoulder, i.e. are somewhat bimodal as well; however, in this case the deviation from the normal distribution is located towards more positive values. For 2CLE-2 the Π criterion works: Not surprisingly, the use of FEP leads to a very broad distribution of energy differences, which results in a negative value of Π (see Table 5). For the shortest switching time (0.1 ps), $\Pi = 0.29$, indicating potential problems. The deviation from the reference result is ≈0.3 kcal/mol in this case, which, however, is already smaller than our threshold of $k_\text{B}T$. The two 2CLE-2 results for the longer switching times (0.5 and 1 ps) lead to free energy differences that agree almost perfectly with the reference result (which were obtained in the 2CLE-2 case, and also in the case of 2CLE, using Crook's theorem based on forward/backward switching simulations of 0.5 ps length). In both cases $\Pi \gg 0.5$. In passing we note that for 2CLE-2 our simplistic $\sigma^2_{\Delta U_i} < 5.85$ criterion also works.

Summing up, switching simulations from MM→QM/MM over the sub- or low picosecond range are too short to result in normally distributed work values. Thus, in principle, the assumptions underlying the derivation of the bias measure Π are not fulfilled. Yet, the criterion proved quite sensitive. Visual inspection of the distribution of work values can easily help spot problematic cases. In the 2CLE case the persistent difference between the full FEP/JAR result and the second cumulant approximation, $\delta\Delta A_{\text{C2}}$ could serve as an additional hint. As one sees in Table 5, $\delta\Delta A_{\text{C2}}$ for the 0.1 and 0.5 ps switches is actually larger than that for FEP.

Finally, in the full data-set (Table 1 of SI) there was one 'positive outlier', i.e. the FEP result for phenyl-trifluoroethyl-ether (PTFE, $\Pi = 0.05$, $\delta\Delta A_{\text{REF}} = 0.36$). Looking at the distribution of energy differences and work values (Figure 1 of SI), one sees that $p(\Delta U_i)$ is broad, but extends as far into the negative range as the work values obtained from switching simulations. Thus, the low Π value is explained by the width of the distribution, but there seem to be sufficiently many ΔU_i values in the negative tail to lead to a more or less correct result. Interestingly, the distributions of work values also exhibit

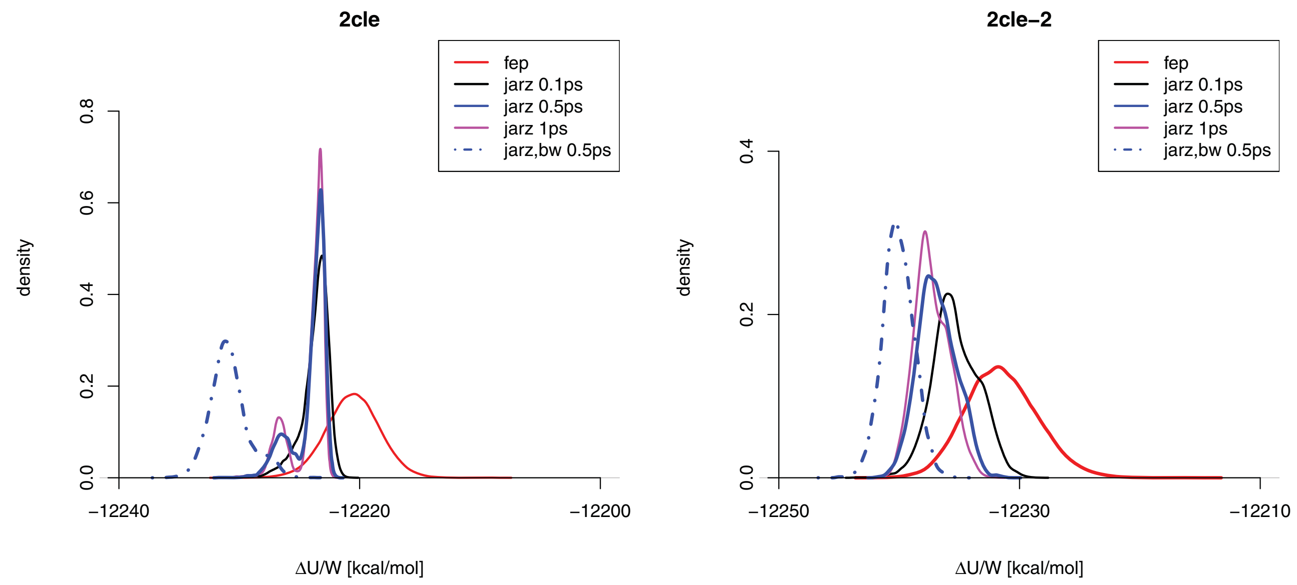


Figure 4. Various forward (FW) and one backward (BW) distributions of ΔU (FEP) and W (JAR) for *bis*-2-chloroethylether (2CLE). Left: MM data generated with the CHARMM CGENFF force field [42]. Right: MM data generated with revised parameters obtained with the GAAMP protocol [43].

shoulders towards positive values, similar to what was found for 2CLE-2 (Figure 4).

5. Conclusions

We have searched for criteria indicating whether a FEP/JAR calculation has converged and can be expected to be free from error, in particular free from systematic bias. Our particular interest is the reliable computation of $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$ in the context of so-called indirect cycles. Because of computational cost, neither two-sided methods (Bennett, Crooks) nor intermediate states are practical options in this case, as they would require extensive simulations at the QM/MM level of theory. A large body of theoretical work is concerned with this question [13,13–17,19–28], though many approaches rely, at least to some degree, on the Gaussian distribution of energy differences/work values. Similarly, in the field of QM/MM FES, the exponential average in Zwanzig's equation is often approximated by the second-order cumulant expansion, although it is valid only in the case of a Gaussian distribution of energy differences. Since this work arose out of practical need, we gave preference to simple, even naive, approaches. In particular, we searched (i) whether a clear correlation between the variance of the distribution of energy differences/work values and systematic error exists, and whether it can be exploited to provide some quantitative guidance. Given the widespread use of the second cumulant approximation, (ii) we attempted to validate it. Based on ease of use, we (iii) picked the bias measure Π introduced by Wu and Kofke [25–27] from the various criteria/error estimates found in the literature. Because backwards distributions are not likely to be available, we restrict ourselves to the assumption that the dissipative work (entropy) is the same in the forward and backward direction [25]. Our findings are pertinent to any situation where a single step FEP/JAR calculation is the only option for completing the indirect thermodynamic cycle, unless, of course, the requirements for one of the stricter theoretical approaches, such as Gaussian distribution of energy differences/work values is fulfilled, in which case this approach or these approaches should be employed.

Our analysis relies on the availability of reference data, which can be considered free from systematic error. Both the variance of the distribution of energy differences/work values $\sigma^2_{\Delta_i}$ and the bias measure Π are correlated relatively strongly with the deviation of results from the respective reference value. Based on Figure 3, we arrive at the crude recommendation that the standard deviation σ_{Δ_i} of the energy differences (work values) should be kept below $4k_\mathrm{B}T$. We would like stress the qualifier *crude*, as even in Figure 3 one clearly sees two violations of this rule. On the other hand, at least for the data considered here, the recommendation to keep σ_{Δ_i} below 1–$2k_\mathrm{B}T$ [14–16] appears too restrictive. In this context, several studies recommend plotting cumulative averages of $-k_\mathrm{B}T\ln\langle\exp(-\Delta U_{0\rightarrow1}/k_\mathrm{B}T)\rangle_0$ and watching for 'saw-tooth' patterns, which indicate rare events where the exponential average is dominated by just a few energy differences (work values) [14]. When applying this criterion to our data, most of the 38 free energy differences utilised in Table 4 and Figure 3 would have to be considered as either not or poorly converged. If strict theoretical criteria were applied, e.g. $\sigma_{\Delta_i} < 1.5 k_B T$ and the absence of saw-tooth patterns in plots of cumulative averages, FEP could never be used to compute $\Delta A^{\mathrm{MM}\rightarrow\mathrm{QM/MM}}$.

The bias measure Π was able to detect most problematic results. This is insofar remarkable, as we utilise the criterion outside its area of validity. Overall, our ad hoc $\sigma^2_{\Delta_i}$ criterion and the requirement $\Pi > 0.5$ either agree or complement each other. Failures of the criteria can be rationalised by considering the details of the distribution of energy differences/work values. In the most problematic case, *bis*-2-chloroethylether, the distributions are far from Gaussian (see Figure 4). Clearly, much longer switching simulations would be required; this would be doable for this case, but not in real-world applications.

Use of the second-order cumulant expansion overall led to significantly poorer results. Not surprisingly, the method works well for Gaussian distributions, e.g. the N1–N4 model distributions. In fact, for broad Gaussian distributions the second-order cumulant approximation gives correct results even for a relatively small number of data points. However, it failed, often by a considerable margin, for most non-Gaussian model distributions, and it gave poorer results for actual application data. When computed with the full FEP/JAR expression, six (out of 38) results deviated by ≥1 kcal/mol from the respective reference result. By contrast, when computing free energy differences using the same raw data with the second-order cumulant approximation, 10 out of 38 results deviated by ≥1 kcal/mol from the reference free energy difference. In all six cases where FEP/JAR failed, the second-order cumulant approximation also failed. Thus, for the 38 data points used in this work, there was no single instance where the second-order cumulant approximation gave a more accurate result than obtained with the full FEP/JAR expression.

While we did not find a single, all-encompassing quality criterion, we feel somewhat confident to give some recommendations. First, use the full FEP/JAR equation rather than the second-order cumulant expression. Second, when evaluating the FEP/JAR equation, compute $\sigma^2_{\Delta U_i}$, as well as Π. Out of 38 systems, a single one failed the dual thresholds $\sigma^2_{\Delta U_i} \leq 4 k_B T$ and $\Pi > 0.5$. This single failure, however, did jump out when plotting distributions of energy differences/work values, as this case deviated noticeably from a Gaussian distribution. From the use of JAR in non-equilibrium work measurements [19] it is well known that too fast switching rates result in work values far from the near-reversible regime, and theoretical treatments of this regime are scarce [20]. Thus, ultimately, we need to develop a better understanding of what constitutes near-reversible switching in the context of going from an MM to a QM/MM representation of the system. Recent work [39] contains first pointers in this direction. In addition, as illustrated by the case of 2CLE vs. 2CLE-2, an improved MM representation speeds up convergence; clearly further work in this area is needed.

Notes

1. While we report/utilise some results from backwards calculations, i.e., QM/MM→ MM, in this work, we stress again that this would not be feasible in most real world applications!
2. Obviously, numerical over-/underflow issues must be taken care of either by means of Equation (3) or the techniques suggested by Berg [33].

Acknowledgments

We thank Fiona L. Kearns and Phillip S. Hudson for careful reading and helpful comments on the manuscript. Additionally, we would like to highlight that this material is based upon work supported by the National Science Foundation under CHE-1464946. Further, HLW would like to thank NSF for computational support via their Major Research Instrumentation Program (MRI 1531590) and via XSEDE (MCB150037).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

National Science Foundation [Grant Number CHE-1464946]; Major Research Instrumentation Program [Grant Number MRI 1531590]; XSEDE [Grant Number MCB 150037].

ORCID

Stefan Boresch http://orcid.org/0000-0002-2793-6656

References

[1] R.W. Zwanzig, J. Chem. Phys. **22**(8), 1420 (1954).
[2] W.L. Jorgensen and L.L. Thomas, J. Chem. Theory Comput. **4**(6), 869 (2008).
[3] W.L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**(6), 3050 (1985).
[4] B. Widom, J. Chem. Phys. **39**(11), 2808 (1963).
[5] C.H. Bennett, J. Comput. Phys. **22**(2), 245 (1976).
[6] W. Yang, Q. Cui, D. Min, and H. Li, in *Annual Reports in Computational Chemistry*, edited by Ralph A. Wheeler (Elsevier, Amsterdam, 2010), Vol. 6, pp. 51–62.
[7] J.G. Kirkwood, J. Chem. Phys. **3**(5), 300 (1935).
[8] J. Heimdal and U. Ryde, Phys. Chem. Chem. Phys. **14**(36), 12592 (2012).
[9] G. König, P.S. Hudson, S. Boresch, and H.L. Woodcock, J. Chem. Theory Comput. **10**(4), 1406 (2014).

[10] C. Cave-Ayland, C.K. Skylaris, and J.W. Essex, J. Phys. Chem. B **119**(3), 1017 (2015).
[11] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[12] P.S. Hudson, H.L. Woodcock, and S. Boresch, J. Phys. Chem. Lett. **6**(23), 4850 (2015).
[13] D.M. Zuckerman and T.B. Woolf, Phys. Rev. Lett. **89**, 180602 (2002).
[14] A. Pohorille, C. Jarzynski, and C. Chipot, J. Phys. Chem. B **114**(32), 10235 (2010).
[15] R.H. Wood, W.C.F. Muhlbauer, and P.T. Thompson, J. Phys. Chem. **95**(17), 6670 (1991).
[16] C. Dellago and G. Hummer, Entropy **16**(1), 41 (2013).
[17] F.M. Ytreberg and D.M. Zuckerman, J. Comput. Chem. **25**(14), 1749 (2004).
[18] J. Liphardt, S. Dumont, S.B. Smith, I. Tinoco and C. Bustamante, Science **296**(5574), 1832 (2002).
[19] J. Gore, F. Ritort and C. Bustamante, Proc. Natl. Acad. Sci. USA **100**(22), 12564 (2003).
[20] M. Palassini and F. Ritort, Phys. Rev. Lett. **107**(6), 060601 (2011).
[21] X. Daura, R. Affentranger, and A.E. Mark, Chem. Phys. Chem. **11**(17), 3734 (2010).
[22] D.A. Kofke and P.T. Cummings, Mol. Phys. **92**(6), 973 (1997).
[23] N. Lu and D.A. Kofke, J. Chem. Phys. **114**(17), 7303 (2001).
[24] N. Lu and D.A. Kofke, J. Chem. Phys. **115**(15), 6866 (2001).
[25] D. Wu and D.A. Kofke, J. Chem. Phys. **121**(18), 8742 (2004).
[26] D. Wu and D.A. Kofke, J. Chem. Phys. **123**(5), 054103 (2005).
[27] D. Wu and D.A. Kofke, J. Chem. Phys. **123**(8), 084109 (2005).
[28] D.A. Kofke, Mol. Phys. **104**(22–24), 3701 (2006).
[29] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, New York, 2001).
[30] G. Hummer and A. Szabo, J. Chem. Phys. **105**(5), 2004 (1996).
[31] G. Hummer, J. Chem. Phys. **114**, 7330 (2001).
[32] J. Kästner, H.M. Senn, S. Thiel, N. Otte, and W. Thiel, J. Chem. Theory Comput. **2**(2), 452 (2006).
[33] B.A. Berg, Comput. Phys. Commun. **153**(3), 397 (2003).
[34] J. Marcinkiewicz, Math. Z. **44**, 612 (1939).
[35] M.P. Allen and D.J. Tildesley, *Computer Simulations of Liquids* (Oxford University Press, New York, 1989).
[36] R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber, and D.K. Kahaner, *Quadpack: A Subroutine Package for Automatic Integration* (Springer, Berlin, Heidelberg, 1983).
[37] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2014).
[38] *NIST-International Sematech E-Handbook: NIST Handbook 151. Dataplot : NIST Handbook 148* (National Institute of Standards and Technology, Gaithersburg, 2003).
[39] F.L. Kearns, P.S. Hudson, H.L. Woodcock, and S. Boresch, J. Comput. Chem., in press. DOI:10.1002/jcc.24706
[40] G.E. Crooks, Phys. Rev. E **61**, 2361 (2000).
[41] E. Rosta, M. Klähn, and A. Warshel, J. Phys. Chem. B **110**(6), 2934 (2006).
[42] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A.D. Mackerell, J. Comput. Chem. **31**(4), 671 (2010).
[43] L. Huang and B. Roux, J. Chem. Theory Comput. **9**(8), 3543 (2013).