

# A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations

Tai-Sung Lee,<sup>\*,†</sup> Brian K. Radak,<sup>†,‡</sup> Anna Pabis,<sup>§</sup> and Darrin M. York<sup>\*,†</sup>

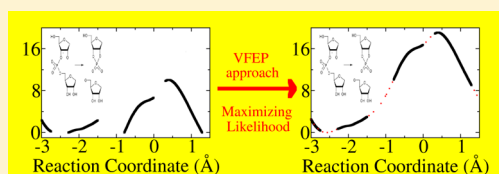
<sup>†</sup>BioMaPS Institute for Quantitative Biology and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States

<sup>‡</sup>Department of Chemistry, University of Minnesota, Minneapolis, Minnesota 55455, United States

<sup>§</sup>Institute of Applied Radiation Chemistry, Faculty of Chemistry, Lodz University of Technology, Zeromskiego 116, 90-924 Lodz, Poland

## S Supporting Information

**ABSTRACT:** A novel variational method for construction of free energy profiles from molecular simulation data is presented. The variational free energy profile (VFEP) method uses the maximum likelihood principle applied to the global free energy profile based on the entire set of simulation data (e.g., from multiple biased simulations) that spans the free energy surface. The new method addresses common obstacles in two major problems usually observed in traditional methods for estimating free energy surfaces: the need for overlap in the reweighting procedure and the problem of data representation. Test cases demonstrate that VFEP outperforms other methods in terms of the amount and sparsity of the data needed to construct the overall free energy profiles. For typical chemical reactions, only  $\sim 5$  windows and  $\sim 20$ – $35$  independent data points per window are sufficient to obtain an overall qualitatively correct free energy profile with sampling errors an order of magnitude smaller than the free energy barrier. The proposed approach thus provides a feasible mechanism to quickly construct the global free energy profile and identify free energy barriers and basins in free energy simulations via a robust, variational procedure that determines an analytic representation of the free energy profile without the requirement of numerically unstable histograms or binning procedures. It can serve as a new framework for biased simulations and is suitable to be used together with other methods to tackle the free energy estimation problem.



## INTRODUCTION

Free energy simulations provide a wealth of insights into complex biomolecular problems. However, the robust calculation of free energies, and in particular free energy surfaces, remains a challenging problem for which much work has been, and continues to be, devoted.<sup>1</sup> One of the primary challenges involves the need to properly sample the necessary degrees of freedom from which a free energy profile can be derived. Strategies to solve this problem are many-fold, and some of the most widespread include multistage/stratified sampling,<sup>2</sup> statically<sup>3–5</sup> and adaptively<sup>6–8</sup> biased sampling, self-guided dynamics,<sup>9</sup> and constrained dynamics,<sup>10,11</sup> as well as multi-canonical<sup>12,13</sup> and replica exchange<sup>14</sup> algorithms. In addition, a number of simulation protocols based on nonequilibrium sampling<sup>15–18</sup> have also been recently proposed as well as hybrid algorithms.<sup>19,20</sup>

One of the most widely used methods for determining free energy surfaces for chemical reactions, where often there are geometric coordinates that are known to be aligned with the overall reaction coordinate, is the “umbrella sampling”<sup>24</sup> technique, which combines stratification with equilibrium, statically biased sampling. Umbrella sampling is particularly amenable to parallel execution, especially in high performance distributed environments,<sup>21,22</sup> as well as extension or combination with replica exchange<sup>23,24</sup> and alchemical

simulation techniques.<sup>25</sup> There are two key difficulties in umbrella sampling methods that remain serious challenges: the problems of “data re-weighting” and “data representation.” Data reweighting refers to the fact that differently biased simulations can only yield accurate information about unbiased simulations after application of a corrective statistical weight. Data representation describes the problem of giving a functional form (either parametric or nonparametric, numerical or analytical) to the target expectation or statistics.

**The Need of Overlap in Data Reweighting.** The data reweighting problem has long been known in the field of molecular simulation and is, in principle, exactly solved by the free energy perturbation (FEP)/Zwanzig relation and the related expression for arbitrary mechanical observables.<sup>4,26,27</sup> However, naive implementation of the FEP estimator is not optimal when considering more than one sample set (see ref 28 for a recent survey). Contemporary methods include the Bennett acceptance ratio,<sup>29</sup> weighted histogram analysis method (WHAM),<sup>30</sup> and multistate Bennett acceptance ratio (MBAR).<sup>31,32</sup> All of these methods are essentially equivalent in their statistical underpinning and rely on the overlap between states (windows) to perform the reweighting but can vary in

Received: August 8, 2012

Published: December 12, 2012

practical applications where sampling is incomplete and, as a result, improved methods continue to be developed.<sup>33–37</sup> The Umbrella Integration (UI) approach of Kästner and Thiel<sup>38–40</sup> assumes a Gaussian model for the unweighted probability density in each umbrella window, from which the analytic derivatives are integrated in order to recover the global probability density, and hence no explicit reweighting is necessary. In fact, UI evades the need of overlap in data reweighting by assuming continuous first derivatives of the free energy profile between windows, even though the usage of a Gaussian model for the unweighted probability density is not ideal in many cases.

**Data Representation.** The data representation problem is particularly important when one is interested in studying mechanisms whereby chemical transformations occur along a minimum free energy pathway. Perhaps the simplest method of data representation is to use a histogram estimate of the probability density.<sup>25,30,36</sup> However, this approach is not numerically stable when data are sparse or sampling does not overlap. Alternatively, one could assume a parametric fit to the biased density in each simulation<sup>41</sup> or apply a more robust kernel density estimator.<sup>8,42</sup> Other methods that address the data representation problem have also been proposed. Maragakis et al. suggested a maximum likelihood approach utilizing a Gaussian-mixture umbrella sampling (GAMUS) model for the global probability density based on the reweighted data<sup>43,44</sup> in order to provide an adaptive bias in umbrella sampling simulations. Basner and Jarzynski proposed a binless estimator based on the optimal correction to an arbitrary reference distribution.<sup>45</sup> Again, UI<sup>38–40</sup> uses Gaussian models for the unweighted probability densities and has also recently been extended to higher order densities (i.e., skewed Gaussians).<sup>46</sup> The result of these assumptions is a significant reduction in the number of data points in each simulation needed to obtain a converged result. This is because parametric estimators converge much more quickly than nonparametric estimators, such as histograms, but often at the expense of increased bias. For example, the approximations/assumptions in UI require near-quadratic (or near quartic) behavior of the local free energy surface. Such behavior can be artificially imposed by using strong harmonic biasing potentials, but this often leads to low overlap between windows and the same kind of failures associated with sparsely populated histogram estimators.<sup>47</sup>

In the present work, we introduce a new variational method for robust determination of free energy profiles (VFEP) from molecular simulation data. The method uses a maximum likelihood principle applied to the global free energy profile and addresses common obstacles: the need for overlap in the data reweighting and the representation problem. In the following sections, the formalism is derived, as well as formulas for the estimation of statistical errors. The method is then applied to a number of numerical simulations, using two general, parametric frameworks based on Akima cubic splines and Floater–Hormann rational function interpolation. The results are compared with those derived from WHAM and MBAR (different reweighting protocols with a histogram density estimate) as well as the UI method. For the test cases examined here, the VFEP method provides extremely robust performance relative to the other methods, particularly in the case of limited or poorly overlapping sampling, and hence appears to be a promising method for robust and rapid

estimation of analytic free energy profiles from molecular simulation data.

## THEORY

Here, we briefly describe the maximum likelihood method utilized in the present work, beginning with a clarification of what the difference is between the terms “probability” and “likelihood” used in this context. In statistical modeling, *probability* refers to the possible outcome of data and is usually modeled by a fixed functional form and a variable set of parameters. On the other hand, *likelihood* refers to how likely a given model can describe a set of observed outcome data.<sup>48</sup> Hence,

- **Probability:**  $p(\{x_n\}|\{\theta_m\})$  is the probability model, defined by a fixed functional form and variable set of parameters  $\{\theta_m\}$ , that returns the probability of observing the data set  $\{x_n\}$ ; i.e., for a given set of model parameters  $\{\theta_m\}$ ,  $p(\{x_n\}|\{\theta_m\})$  predicts the outcome for the set of data  $\{x_n\}$ :  $\{\theta_m\} \rightarrow \{x_n\}$ .
- **Likelihood:**  $\mathcal{L}(\{\theta_m\}|\{x_n\})$  is the likelihood that the observed data set  $\{x_n\}$  was generated by the probability distribution model defined by the set of parameters  $\{\theta_m\}$ ; i.e.,  $\mathcal{L}(\{\theta_m\}|\{x_n\})$ , for a given set of observed data  $\{x_n\}$ , provides an assessment of the goodness of the model parameters:  $\{x_n\} \rightarrow \{\theta_m\}$ .

The maximum likelihood method, or maximum likelihood estimation (MLE),<sup>49,50</sup> is the procedure of finding the optimal set of parameters that maximize the likelihood of the model probability distribution function to represent a given set of observed data.

MLE begins with the definition of the likelihood function of the sample data. The likelihood function of a set of data is the probability of obtaining that particular set of data, given the probability distribution model function defined by a chosen functional form along with a set of trial model parameters. Here, we consider the probability,  $p(x)$ , of observing a molecular system at a particular value of a single generalized coordinate  $x$  (the extension to multiple dimensions is straightforward). This probability is given by

$$p(x) = \frac{e^{-F(x)}}{\int e^{-F(x')} dx'} \quad (1)$$

where  $F(x) \equiv \mathcal{F}(x)/(k_B T)$  is the unitless scaled free energy profile,  $\mathcal{F}(x)$  is the free energy profile,  $k_B$  is the Boltzmann constant, and  $T$  is the absolute temperature. Consider now a parametric model for the scaled free energy profile  $F(x|\{\theta_m\})$  where  $\{\theta_m\}$  is the set of parameters. The probability distribution model,  $p(x|\{\theta_m\})$ , also contains the set of parameters, due to its relation to  $F(x|\{\theta_m\})$ . Now considering the probability,  $p(\{x_n\}|\{\theta_m\})$ , of a sampled data set  $\{x_n\}$ , if the sampling data points are independent of each other, then

$$p(\{x_n\}|\{\theta_m\}) = p(x_1, x_2, \dots, x_N|\{\theta_m\}) \\ = p(x_1|\{\theta_m\}) \cdot p(x_2|\{\theta_m\}) \cdots p(x_N|\{\theta_m\}) \quad (2)$$

The likelihood  $\mathcal{L}$  of the trial free energy profile  $F\{\theta_m\}$  with the given observed data set  $\{x_n\}$  is

$$\mathcal{L}(F\{\theta_m\}|x_1, \dots, x_N) = \mathcal{L}(\{\theta_m\}|x_1, \dots, x_N) = \prod_{i=1}^N p(x_i|\{\theta_m\}) \quad (3)$$

In the present work, instead of dealing with individual windows, we attempt to find the optimal solution of the above equation by defining a global function  $F(x)$  with a set of defined parameters  $\{\theta_m\}$ . It is practical to use the logarithm of the likelihood function, called the log-likelihood  $\hat{l}$ :

$$\hat{l}(\{\theta_m\}|x_1, \dots, x_N) = \frac{1}{N} \ln \mathcal{L} = \frac{1}{N} \sum_{n=1}^N \ln p(x_n|\{\theta_m\}) \quad (4)$$

Since the likelihood is always positive and the logarithmic function is monotonic, there is no loss of generality in formulating a variational principle based on the log-likelihood, which offers some advantages in terms of numerical stability and is conventional in the literature. Hereafter, we use the term “likelihood” generically to refer to both the likelihood or the log-likelihood and will reference specific equations when the mathematical distinction is necessary. The MLE method estimates  $\{\theta_m\}$  by finding the values of  $\{\theta_m\}$  that maximize  $\hat{l}$ :

$$\begin{aligned} \hat{l}(\{\theta_m^*\}|x_1, \dots, x_N) &= \arg \max_{\{\theta_m\} \in \Theta} \hat{l}(\{\theta_m\}|x_1, \dots, x_N) \\ &= \arg \max_{\{\theta_m\} \in \Theta} \frac{1}{N} \sum_{n=1}^N \ln p(x_n|\{\theta_m\}) \end{aligned} \quad (5)$$

where  $\Theta$  defines the space that  $\{\theta_m\}$  can span. If a biasing potential  $W^\alpha(x)$  is applied in the  $\alpha$ th window in a set of umbrella sampling simulations, the probability of finding the system with a certain coordinate value  $x$  is

$$p^\alpha(x) = \frac{1}{Z^\alpha} \exp\{-[F(x) + W^\alpha(x)]\},$$

where  $Z^\alpha = \int_{-\infty}^{\infty} \exp\{-[F(x) + W^\alpha(x)]\} dx$  (6)

Suppose that for the simulation of the  $\alpha$ th window, there are  $N^\alpha$  points observed with coordinate values  $\{x_i^\alpha\}$ . Since they are observed points, the probability of each point is equal with value  $1/N^\alpha$ . The likelihood of the whole system with an overall free energy profile  $F(x)$  can be expressed as the combination of the likelihood of individual windows obtained from eq 4 and eq 6 as

$$\begin{aligned} \hat{l}(F) &\equiv \sum_{\alpha}^{\text{windows}} c^\alpha \hat{l}^\alpha(\{\theta_m\}|\{x_n^\alpha\}) \\ &= - \sum_{\alpha}^{\text{windows}} c^\alpha \left\{ \ln Z^\alpha + \frac{1}{N^\alpha} \sum_i^{\text{data points}} [F(x_i^\alpha) + W^\alpha(x_i^\alpha)] \right\} \end{aligned} \quad (7)$$

where  $\{c^\alpha\}$  are the combination weights defining the relative contribution of likelihood from different windows when combining the local likelihood into a global likelihood. When assuming all windows contribute equally, the  $c^\alpha$  can simply be set to be equal, i.e.,  $c^\alpha = 1$ . It can also be shown that, in the exact sampling limit, the global optimal  $F$  is also the optimal  $F$  for each individual window; i.e., the choice of  $\{c^\alpha\}$  does not affect the resulting optimal  $F(x)$  (see Supporting Information). In practice, for finite sampling, we observe that the overall result is largely insensitive to the choice of  $c^\alpha$ , and for the present work, we choose  $c^\alpha = 1$  for all windows (also see

Supporting Information). In the above equation for the global likelihood function, we have used  $F$  as the argument to emphasize that optimization of the likelihood function is with respect to the free energy profile  $F$  (by varying the  $\{\theta_m\}$  parameters).

There remains the task of finding the  $F$  that maximizes  $\hat{l}(F)$ . Note that in the above equation, the term  $W^\alpha(x_i^\alpha)$  is constant and does not need to be evaluated if the goal is to maximize the likelihood. Also, the term  $-\ln Z^\alpha$  is equivalent to the relative free energies (or free energy shifts) between windows in other reweighting schemes. In the present VFEP approach, the “reweighting” procedure is implicitly accomplished through the normalization against the global trial function  $F$ .

An alternate strategy is to model  $F(x)$  locally in the region of each window,  $F^\alpha(x)$ , and construct the global  $F(x)$  using the  $F^\alpha(x)$  with the observed data density as weighting. The only variable parameters in this approach are the relative free energy shifts between every window  $\{f^\alpha\}$  (the reference free energy being arbitrary) that establish the relative weights for each window. Thus, the global  $F(x)$  is defined by the parameter set  $\{f^\alpha\}$  and a set of fixed local free energy profiles  $F^\alpha(x)$ . Applying the MLE procedure to  $F(x)$  with respect to the parameter  $\{f^\alpha\}$  leads to the WHAM and the MBAR equations.<sup>31–33,51</sup> Note that within such a context, MBAR is also a parametric procedure where the relative free energy shifts of windows are the MLE parameters and local free energy profiles are predefined in data fitting procedures, whereas the proposed VFEP uses MLE parameters to construct the detailed overall free energy profiles. In summary, the WHAM and MBAR formula are equivalent to the MLE results when the global free energy profile is constructed from the local free energy profiles and the relative free energies are used as the parameters to optimize the likelihood.

In the present work, instead of dealing with individual windows, we attempt to find the optimal solution of eq 7 by defining a global function  $F(x)$  with a set of defined parameters  $\{\theta_m\}$  (i.e.,  $F(x) \equiv F(x|\{\theta_m\})$ ). The procedure is as follows:

1. Choose a trial function  $F(x)$  with a initial parameter set  $\{\theta_m\}$ .
2. Evaluate the likelihood  $\hat{l}(F)$  of the trial function  $F(x)$  according to eq 7.
3. Vary the parameter set  $\{\theta_m\}$  until the maximum of  $\hat{l}(F)$  is reached.
4. The trial  $F(x)$  with the maximal  $\hat{l}(F)$  is the desired overall free energy profile.

Two types of analytic functions were selected to model the overall free energy profile: a cubic spline function<sup>52</sup> and a rational interpolation function.<sup>53</sup> Both were originally designed for interpolation usage. Nevertheless, one could treat the interpolation input data as the variable parameters; for example, a cubic spline function needs to have the  $\{x_i, y_i\}$  data nodes defined in order to build the desired cubic spline interpolation, where  $x_i$  is the independent variable and  $y_i$  is the corresponding observed function value. In this work, we select fixed  $x_i$  and treat  $y_i$  as the MLE parameters to be optimized. For example, a cubic spline function defined by  $\{x_i, y_i\}$  will be the trial free energy function in eq 7, and the optimal free energy profile is reached through changing  $\{y_i\}$ . This is equivalent to assuming that the free energy profile varies slower than a cubic polynomial between windows or that the first and second derivatives of free energy profile are continuous between windows.



## RESULTS

A C++ program was built to test the proposed method. Two interpolation subroutines in the AlgLib (v3.5, <http://www.alglib.net>) package were used: The Akima spline algorithm<sup>52</sup> was employed for cubic spline interpolation, and the Floater–Hormann<sup>53</sup> algorithm for rational interpolation. Both the number of spline function nodes and the number of the rational interpolation poles are set to 2 times the number of windows minus one. There is one node located at the average data position of each window and one node located at the average position of two nodes of two adjacent windows. The results of WHAM were calculated by the program from Grossfield<sup>54</sup> (v2.0.4, <http://membrane.urmc.rochester.edu/content/wham>). The results of MBAR were calculated using the pymbar library of Shirts and Chodera<sup>32</sup> (v2.0b, <http://simtk.org/home/pymbar>). The UI algorithm<sup>38–40</sup> was implemented as part of the VFEP program.

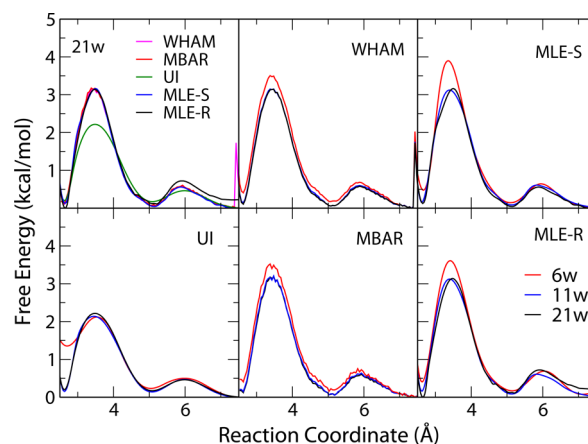
In order to cover a wide range of common free energy profile problems, tests were performed with a benchmark molecular dynamics simulation of a  $\text{Na}^+:\text{Cl}^-$  pair in a water box, two combined quantum mechanical/molecular mechanical (QM/MM) simulations of chemical reactions, and the C–C–C–C torsion rotation of butane. These test cases represent nonbonding interactions, chemical reactions, and conformational transitions. The results for these systems are listed/described in the subsequent sections.

**$\text{Na}^+:\text{Cl}^-$  Pair.** A  $\text{Na}^+:\text{Cl}^-$  pair was put in a TIP3P water box<sup>55</sup> ( $20 \text{ \AA} \times 20 \text{ \AA} \times 20 \text{ \AA}$ ) with the CHARMM27 force field.<sup>56</sup> The distance between  $\text{Na}^+$  and  $\text{Cl}^-$ , defined as the relevant coordinate, was scanned from 2.4 to 7.4  $\text{\AA}$  with 21 windows separated by 0.25  $\text{\AA}$ . A biasing potential of either 5 or 100  $\text{kcal/mol/\AA}^2$  was applied to each window. The NAMD package (v 2.7)<sup>57</sup> was used, and simulations were performed under periodic boundary conditions in the NpT ensemble at 300 K and 1 atm (NAMD uses a modified Nosé–Hoover method<sup>58,59</sup> in which Langevin dynamics is used to control fluctuations in the barostat). Each window was simulated for 1 ns of equilibration and 1 ns of data collection (10 000 data points per window).

**Weak Biasing Potential.** In the first set of simulations, a biasing potential of 5  $\text{kcal/mol/\AA}^2$  was applied to every umbrella sampling window, which is relatively weak, affording considerable overlap between windows. This allows fewer windows to be required to construct the overall profile than if a larger umbrella potential were used. However, in the case of weak umbrella biasing, one would expect a quadratic approximation of the local (biased) free energy profile within any given window not to be ideal.

The results with the weak biasing potential of 5  $\text{kcal/mol/\AA}^2$  are shown in Figure 1. The upper left panel shows the results from all methods with 21 windows. Other panels show the results from different methods with different numbers of windows (11 and 6). While all other methods converge with 21 windows (with statistic errors less than 0.05  $\text{kcal/mol}$ , see Table 1) and give similar results with 11 or even six windows, UI, using a quadratic approximation, delivers a quantitatively incorrect free energy profile.

**Strong Biasing Potential.** In the second set of simulations, a relatively strong biasing potential with a strength of 100  $\text{kcal/mol/\AA}^2$  was applied to every window. Contrary to the weak potential set of simulations, one would expect that a quadratic approximation of the local free energy profile would perform



**Figure 1.** The free energy profiles calculated with different methods for the  $\text{Na}^+:\text{Cl}^-$  pair from a 21-window umbrella sampling simulation with a weak biasing potential of 5  $\text{kcal/mol/\AA}^2$ . The upper left panel shows the results from all methods with 21 windows. Other panels show the results from individual methods with different numbers of windows: six (red), 11 (blue), and 21 (black) windows. While all methods converge with 21 windows and give similar results with 11 or even six windows, UI yields an incorrect free energy profile, as expected.

well, but the requirement of the numbers of windows will increase since the overlap between windows will be diminished.

The results with the strong biasing potential of 100  $\text{kcal/mol/\AA}^2$  are shown in Figure 2. The upper left panel again shows the results from all methods with 21 windows, and other panels show the results from individual methods with different numbers of windows (11 and 6). All methods, including UI, converge with 21 windows and give similar results. With 11 windows, however, WHAM and MBAR fail to produce correct results, while with six windows, WHAM, UI, and MBAR all fail to converge due to the lack of sufficient overlap between windows. On the other hand, the VFEP approach, with both the spline function (MLE-S) and rational interpolation function (MLE-R), gives very good results for 11 windows compared to the 21 window results and gives qualitatively correct results with only six windows.

**Reduced Data Set.** In the case of a weak biasing potential, WHAM gives good results with only six windows. One would expect, however, that many data points would be necessary to model individual windows well. Figure 3 shows the results with six windows from WHAM and the proposed VFEP methods using the weak biasing potential of 5  $\text{kcal/mol/\AA}^2$ , the same as the above results (Figure 1), but the data points are stripped out when performing analysis. The upper panel and middle panel show the VFEP results with spline function (MLE-S) and rational interpolation function (MLE-R), respectively, while the WHAM results are shown in the bottom panel. WHAM fails to converge with 100 data points or less per window, and MBAR gives similar results, both due to insufficient data points in the histograms. Hence only WHAM results are shown. VFEP still delivers qualitatively correct results with only 20 data points per window where the statistical error (by bootstrapping) is less than 1  $\text{kcal/mol}$  (see Table 2).

**QM/MM Phosphoryl Transfer Reactions.** The phosphate 2'-O-transesterification reaction for two model compounds were simulated by QM/MM umbrella sampling using the AMBER12 simulation package<sup>60</sup> (Figure 4). The first model, 2-(hydroxypropyl)-4-nitrophenyl phosphate (HpPNP), contains

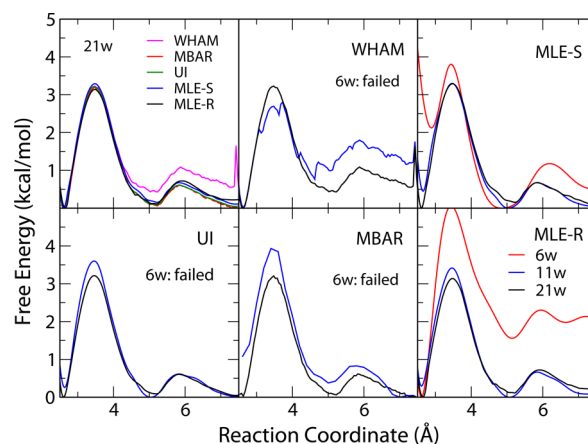
**Table 1.** Estimated Errors for the Free Energy Profile of the  $\text{Na}^+:\text{Cl}^-$  System from VFEP and MBAR<sup>a</sup>

window ( $\alpha$ )	$\bar{x}^\alpha$	$\Delta\hat{f}^\alpha(s,g)$	$\Delta\hat{f}^\alpha(s,m)$	stat error ( $N = 50$ )	stat error ( $N = 100$ )	MBAR
1	2.629	0.022	0.060	0.018	0.018	0.043
2	2.655	0.040	0.015	0.018	0.018	0.042
3	2.686	0.125	-0.001	0.018	0.018	0.049
4	2.754	0.421	-0.034	0.018	0.018	0.070
5	3.113	0.555	-0.071	0.018	0.015	0.086
6	3.945	0.304	-0.034	0.018	0.010	0.073
7	4.275	0.160	-0.113	0.010	0.010	0.052
8	4.460	0.016	0.013	0.010	0.010	0.042
9	4.638	0.002	-0.025	0.010	0.010	0.040
10	4.797	0.002	0.007	0.009	0.009	0.040
11	4.964	0.002	0.012	0.008	0.008	0.039
12	5.127	0.003	-0.034	0.008	0.008	0.038
13	5.334	0.016	-0.008	0.009	0.008	0.039
14	5.588	0.015	-0.006	0.009	0.009	0.041
15	5.863	0.006	0.001	0.010	0.010	0.041
16	6.188	0.006	-0.012	0.010	0.011	0.038
17	6.500	0.002	-0.007	0.012	0.012	0.037
18	6.719	0.002	-0.012	0.013	0.014	0.036
19	6.955	0.000	0.010	0.015	0.015	0.035
20	7.194	0.000	-0.014	0.016	0.017	0.035
21	7.413	0.000	-0.004	0.017	0.018	0.000
RMS		0.172	0.036	0.013	0.013	0.048

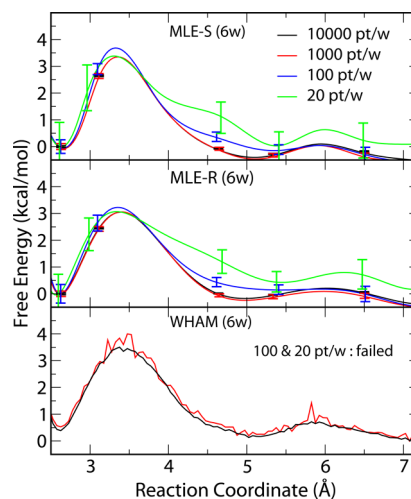
<sup>a</sup>The numbers here are derived from a 21-window umbrella sampling simulation on a  $\text{Na}^+:\text{Cl}^-$  pair in a TIP3P water box with a biasing potential of 5 kcal/mol/Å<sup>2</sup> (see the Results section, also Figure 1).  $\bar{x}^\alpha$  is the average of the sampled coordinates of the  $\alpha$ th window.  $\Delta\hat{f}^\alpha(s,m)$  and  $\Delta\hat{f}^\alpha(s,g)$  are likelihood errors defined in eqs 8 and 9, respectively. The “stat error” is the statistical error estimated by performing bootstrap error analysis on the free energy shift term,  $-\ln Z^\alpha$ , with the same calculations performed on 50 or 100 randomly chosen data sets. The numbers reported are the standard deviations of  $-\ln Z^\alpha$  from different sets of data. MBAR errors are from the MBAR output. The last row (RMS) is the root-mean-square values of the corresponding column. All values are in units of  $k_B T$ , except for  $\bar{x}^\alpha$ , which is in units of Å.

an enhanced leaving group and is therefore expected to have a free energy profile with a significantly different shape. The second is an abasic RNA dinucleotide which has been studied previously in our group. Both sets of simulations used the AM1/d-PhoT QM/MM Hamiltonian,<sup>61</sup> which has been verified and demonstrated to be able to reproduce high-level DFT results within chemical accuracy in describing phosphate chemistry by our group<sup>62–65,87</sup> and others,<sup>66,67</sup> under periodic boundary conditions using QM/MM Ewald summations as implemented in AMBER12.<sup>68</sup> The QM region was defined as the entire solute. The reaction coordinate is defined as the difference between the nucleophile to phosphorus distance ( $r_1$ ) and the phosphorus to leaving group distance ( $r_2$ ). For umbrella sampling simulations, a harmonic biasing potential was applied to this reaction coordinate,  $r_1 - r_2$ .

**HpPNP.** HpPNP was solvated in a box of TIP4P-Ew water<sup>69</sup> at 300 K using the NVT ensemble with an Andersen thermostat.<sup>70</sup> Twenty-five short (100 ps) umbrella sampling simulations were performed with a biasing potential strength of 60 kcal/mol/Å<sup>2</sup>. The QM/MM free energy profile results for HpPNP are shown in Figure 5. Similar to Figure 1, The upper left panel shows the results from all methods with 25 windows. Other panels show the results from individual methods with



**Figure 2.** The free energy profiles calculated with different methods for the  $\text{Na}^+:\text{Cl}^-$  pair from a 21-window umbrella sampling simulation with a strong biasing potential of 100 kcal/mol Å<sup>2</sup>. The upper left panel shows the results from all methods with 21 windows. Other panels show the results from individual methods with different numbers of windows: six (red), 11 (blue), and 21 (black) windows. All methods, including UI, converge with 21 windows and give similar results. With 11 windows, however, MBAR fails to produce correct results, while with six windows, WHAM, UI, and MBAR all fail to converge due to the lack of sufficient overlap between windows. On the other hand, the VFEP approach, both with spline (MLE-S) and rational interpolation (MLE-R) functions, gives very good results for 11 windows compared to the 21 window results and gives qualitatively correct results with only six windows.



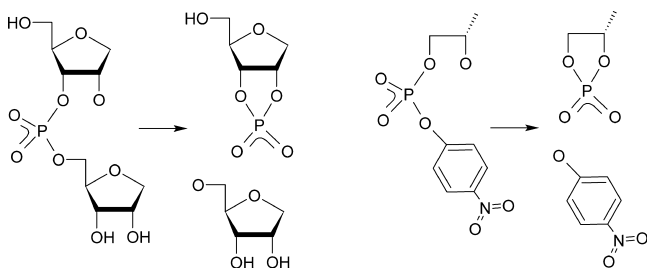
**Figure 3.** The free energy profiles calculated with WHAM and VFEP for the  $\text{Na}^+:\text{Cl}^-$  pair from a six-window umbrella sampling simulation with a weak biasing potential of 5 kcal/mol Å<sup>2</sup>. The data points are reduced at different levels: 10 000 pt/w (black), 1000 pt/w (red), 100 pt/w (blue), and 20 pt/w (green). The error bars are bootstrap errors calculated from 100 random data sets with the same size. The upper panel and middle panel show the VFEP results with spline (MLE-S) and rational interpolation (MLE-R) functions, respectively, while the WHAM results are shown in the bottom panel. WHAM fails to converge with 100 or fewer data points per window, while VFEP still delivers qualitatively correct results with only 10 data points per window.

different numbers of windows (15 and five). While all methods converge with 25 windows, only the VFEP method, both with a spline function (MLE-S) and a rational interpolation function (MLE-R), still gives good results for five windows.

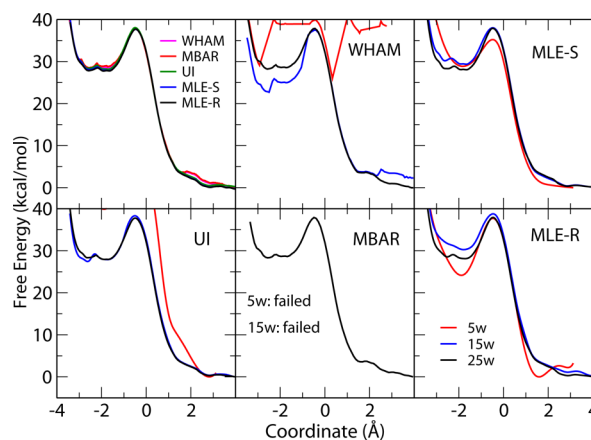
**Table 2.** Estimated Bootstrap Errors (50 and 100 Calculations with Random Data Sets) of Free Energy Shifts Calculated by the VFEP Method<sup>a</sup>

	window	N = 50		N = 100	
		average	SD	average	SD
10000 pt/w	1	0.758	0.032	0.760	0.033
	2	2.139	0.026	2.141	0.026
	3	-0.153	0.017	-0.152	0.018
	4	-0.776	0.014	-0.777	0.017
	5	-0.720	0.028	-0.722	0.026
	6	-1.249	0.033	-1.250	0.032
RMS			0.026		0.026
1000 pt/w	1	0.763	0.118	0.761	0.119
	2	2.146	0.101	2.142	0.091
	3	-0.153	0.057	-0.151	0.053
	4	-0.763	0.066	-0.771	0.065
	5	-0.722	0.093	-0.721	0.089
	6	-1.270	0.111	-1.260	0.105
RMS			0.094		0.088
100 pt/w	1	0.752	0.219	0.715	0.254
	2	2.135	0.189	2.087	0.206
	3	-0.149	0.185	-0.138	0.175
	4	-0.770	0.214	-0.761	0.181
	5	-0.685	0.236	-0.664	0.246
	6	-1.283	0.299	-1.239	0.301
RMS			0.227		0.232
20 pt/w	1	0.679	0.819	0.633	0.902
	2	2.028	0.734	1.965	0.855
	3	-0.118	0.474	-0.097	0.586
	4	-0.772	0.483	-0.728	0.54
	5	-0.654	0.652	-0.647	0.653
	6	-1.162	0.775	-1.126	0.781
RMS			0.670		0.732

<sup>a</sup>The system is the Na<sup>+</sup>:Cl<sup>-</sup> system with six windows. The numbers here are derived from a six-window umbrella sampling simulation on a Na<sup>+</sup>:Cl<sup>-</sup> pair in a TIP3P water box with a biasing potential of 5 kcal/mol/Å<sup>2</sup> (see the Results section, also Figure 3). The results are estimated by performing bootstrap type error analysis on the free energy shift term,  $-\ln Z^a$ , with the same calculations performed on 50 or 100 randomly chosen data sets. "SD" is the standard deviation, while "RMS" is the root-mean-square value of the corresponding column. Results from different numbers of data points used in a window (10 000 pt/w, 1000 pt/w, 100 pt/w, and 20 pt/w) are shown. All values are in units of  $k_B T$ .

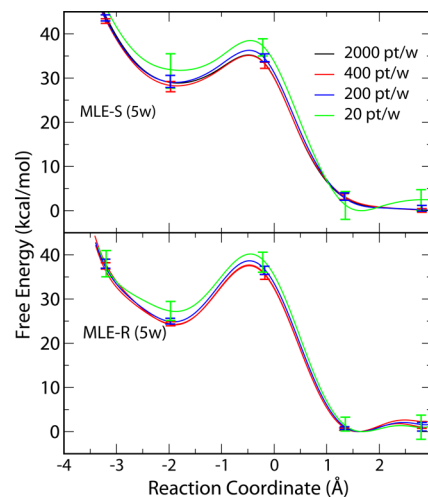


**Figure 4.** Reaction schemes for QM/MM phosphoryl transfer reactions of an abasic RNA dinucleotide and 2-(hydroxypropyl)-4-nitrophenyl phosphate (HpPNP), a model compound with an enhanced leaving group.



**Figure 5.** The QM/MM free energy profile results for HpPNP. Similar to Figure 1, the upper left panel shows the results from all methods with 25 windows. Other panels show the results from individual methods with different numbers of windows: five (red), 15 (blue), and 25 (black) windows. For the case of 15 and five windows, MBAR fails due to a lack of overlap between windows when 75 bins are used (no data in certain bins). While all methods converge with 25 windows, only the VFEP method, with spline (MLE-S) and rational interpolation (MLE-R) functions, still gives good results for five windows.

**Reduced Data Set.** The QM/MM free energy profile results for HpPNP with reduced numbers of data points are shown in Figure 6. Similar to Figure 3, VFEP, both with a spline function



**Figure 6.** The QM/MM free energy profile results for HpPNP with reduced numbers of data points (2000 pt/w (black), 400 pt/w (red), 200 pt/w (blue), and 20 pt/w (green)). The error bars are bootstrap errors calculated from 100 random data sets with the same size. VFEP, with both spline (MLE-S) and rational interpolation (MLE-R) functions, still delivers qualitatively correct results with only 20 data points in each of five windows. Note that all other methods fail with only five windows and hence cannot be compared here.

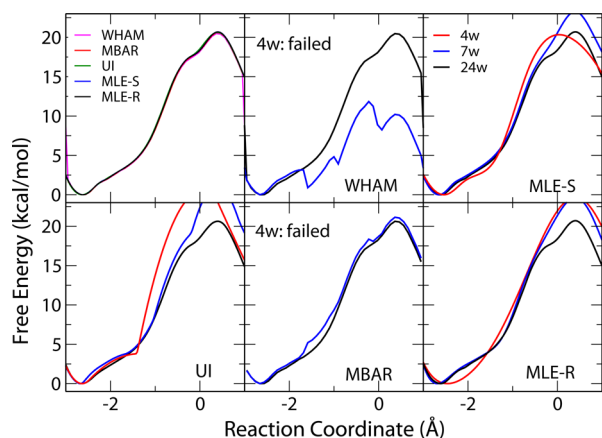
(MLE-S) and a rational interpolation function (MLE-R), still delivers qualitatively correct results with only 20 data points per window (with five windows) where the bootstrapping errors are around 3 kcal/mol.

**Abasic Dinucleotide.** Mimicking the experimental conditions of Harris et al.<sup>71</sup> for a UpG dinucleotide, the system was solvated in a rhombic dodecahedron of TIP3P water<sup>55</sup> with sodium chloride<sup>72</sup> under physiological conditions (310 K) in



the NVT ensemble with an Andersen thermostat.<sup>70</sup> Data from 24 long (1.75 ns each) umbrella sampling simulations were used.

The QM/MM free energy profile results for the abasic dinucleotide are shown in Figure 7. Similar to Figure 5, the



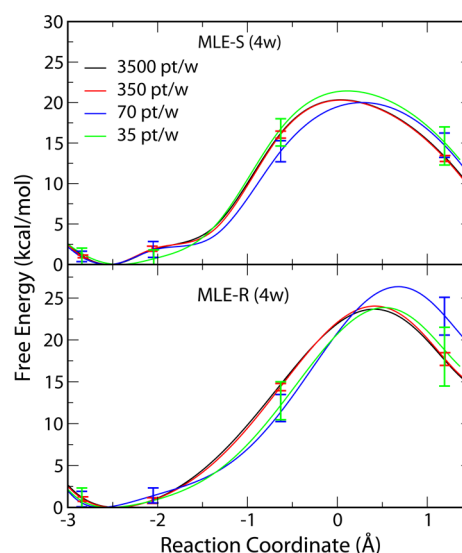
**Figure 7.** The QM/MM free energy profile results for an abasic RNA dinucleotide (Figure 4). Similar to Figure 5, the upper left panel shows the results from all methods with 24 windows. Other panels show the results from individual methods with different numbers of windows: four (red), seven (blue), and 24 (black) windows. While all methods converge with 24 windows, both WHAM and MBAR fail to converge with four windows. UI and VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still gives good result for four windows.

upper left panel shows the results from all methods with 24 windows. Other panels show the results from individual methods with different numbers of windows (24, seven, and four windows). While all methods converge with 24 windows, both WHAM and MBAR fail with four windows. The UI and VFEP methods, both with a spline function (MLE-S) and a rational interpolation function (MLE-R), still give good results for four windows. When they succeed, all of the methods produce a free energy barrier comparable to the experimental value of 19.9 kcal/mol, as inferred from the rate constant extrapolated to “infinite” pH<sup>71</sup> and transition state theory.

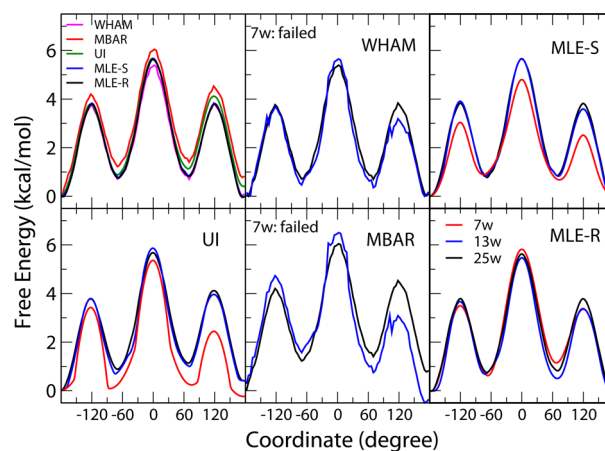
**Reduced Data Set.** The QM/MM free energy profile results for the abasic dinucleotide with reduced numbers of data points are shown in Figure 8. VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still delivers qualitatively correct results with only 35 data points in each of four windows. However, the quantitative inaccuracy is readily apparent in the bootstrapping errors around 3 kcal/mol.

**Torsion Rotation of Butane.** A butane molecule was modeled using the AMBER ff99 force field in a generalized Born solvent at 300 K using Langevin dynamics as implemented in the AMBER12 simulation package.<sup>60</sup> The umbrella sampling simulations were performed by applying harmonic restraints on the C–C–C–C torsion with a force constant of 32.83 kcal/mol/rad<sup>2</sup> (0.02 kcal/mol/degree<sup>2</sup>). The equilibrium position of the torsion angle ran from –180 to +180° in increments of 15, resulting in 25 windows. Each window was simulated for 0.5 ns of equilibration and 1 ns of data collection (10 000 data points per window).

The free energy profile results for the C–C–C–C torsion of butane are shown in Figure 9. Similar to Figure 1, The upper left panel shows the results from all methods with 25 windows.



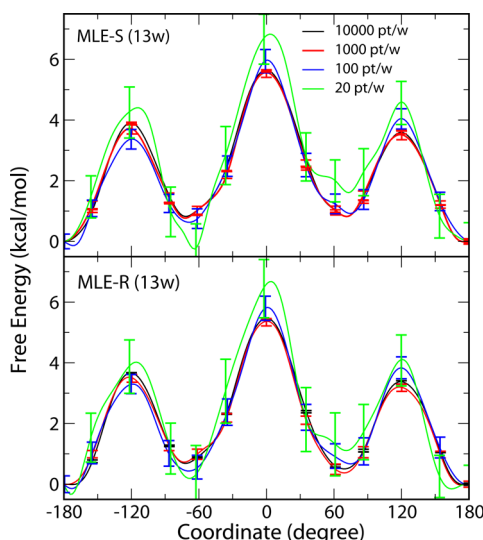
**Figure 8.** The QM/MM free energy profile results for an abasic RNA dinucleotide with reduced numbers of data points: 3500 pt/w (black), 350 pt/w (red), 70 pt/w (blue), and 35 pt/w (green). The error bars are bootstrap errors calculated from 100 random data sets with the same size. VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still delivers qualitatively correct results with only seven data points in each of four windows.



**Figure 9.** The free energy profile of C–C rotation of butane. Similar to Figure 5, the upper left panel shows the results from all methods with 25 windows (15° spacing). Other panels show the results from individual methods with different numbers of windows: seven (red), 13 (blue), and 25 (black) windows. While all methods converge with 25 windows (MBAR and UI have some deviation due to lack of periodic constraint), both WHAM and MBAR fail to converge with seven windows. UI and VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still give good results for seven windows.

Other panels show the results from individual methods with different numbers of windows (13 and 7). While all methods converge with 25 windows, only the VFEP method, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still gives good results for seven windows.

**Reduced Data Set.** The free energy profile results for butane with reduced numbers of data points are shown in Figure 10. Similar to Figure 3, VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still delivers qualitatively correct results with only 20 data points in each



**Figure 10.** The free energy profile of C–C rotation of butane with reduced numbers of data points: 10 000 pt/w (black), 1000 pt/w (red), 100 pt/w (blue), and 20 pt/w (green). The error bars are bootstrap errors calculated from 100 random data sets with the same size. VFEP, both with spline (MLE-S) and rational interpolation (MLE-R) functions, still delivers qualitatively correct results with only 20 data points in each of 13 windows.

of 13 windows. Statistical errors from bootstrapping are around 1 kcal/mol.

**Error Analysis. Likelihood Error.** The likelihood of a set of trial probability  $\{p(x_i)\}$  with given observed probability set  $\{p_{\text{obs}}(x_i)\}$  can be written as

$$\hat{l} = \sum_i p_{\text{obs}}(x_i) \ln p(x_i)$$

Assuming that the trial probability is a Boltzmann distribution due to the trial effective potential  $F_{\text{eff}}$  and the observed data points are unbiased, then the corresponding observed likelihood function is

$$\hat{l}(F_{\text{eff}}) = \sum_i \frac{1}{N} \ln \left\{ \frac{1}{Z} e^{-F_{\text{eff}}(x_i)} \right\}$$

where the normalization factor  $Z$  is defined as  $Z \equiv \int e^{-F_{\text{eff}}(x)} dx$ . For the  $\alpha$ th umbrella sampling simulation window, the trial effective potential is the combination of the trial free energy profile  $F(x)$  and the added biasing potential  $W^\alpha(x)$ . Hence

$$\begin{aligned} \hat{l}_s^\alpha(F_{\text{eff}}) &= \hat{l}_s^\alpha(F) = -\ln Z^\alpha - \frac{1}{N} \sum_i [F(x_i^\alpha) + W^\alpha(x_i^\alpha)] \\ &= -\ln Z^\alpha - \langle F(x) + W^\alpha(x) \rangle_{\text{sample}} \end{aligned}$$

Note that the above equation is exactly the same as eq 7.  $\hat{l}_s^\alpha$  can be expressed as a functional of either  $F_{\text{eff}}$  or  $F$  since they only differ by a known function  $W^\alpha$ . The subscript “s” denotes that the likelihood is calculated based on the sampling data, and  $\langle \dots \rangle_{\text{sample}}$  indicates that the average is calculated using the observed sample probability distribution.  $\hat{l}_s^\alpha(F)$  is the functional to be optimized in the present work as described in the Theory section. If the trial free energy profile is the true system free energy profile and the sampling is exact and infinite, then the “ideal” likelihood is now

$$\begin{aligned} \hat{l}_m^\alpha(F) &= -\ln Z^\alpha - \int e^{-[F(x) + W^\alpha(x)]} \{F(x) + W^\alpha(x)\} dx \\ &= -\ln Z^\alpha - \langle F(x) + W^\alpha(x) \rangle_{\text{model}} \end{aligned}$$

The subscript “m” denotes that the likelihood is calculated based on the modeled free energy profile function, and  $\langle \dots \rangle_{\text{model}}$  indicates that the average is calculated using the modeled probability distribution. In the present work, since  $\hat{l}_s^\alpha(F)$  is the functional to be optimized and  $\hat{l}_m^\alpha(F)$  is the “ideal” target likelihood, the difference between them, denoted as  $\Delta \hat{l}^\alpha(s, m)$ , can be viewed as the limit that the optimization process can reach, or equivalently, the lower bound error of the proposed method. That is

$$\begin{aligned} \Delta \hat{l}^\alpha(s, m) &\equiv \hat{l}_s^\alpha(F) - \hat{l}_m^\alpha(F) \\ &= \langle F(x) + W^\alpha(x) \rangle_{\text{model}} \\ &\quad - \langle F(x) + W^\alpha(x) \rangle_{\text{sample}} \end{aligned} \quad (8)$$

Apparently,  $\Delta \hat{l}^\alpha(s, m)$  is just the difference in the expectation values computed with the effective potentials from the sampling data and from the optimized free energy profile.

**Error Due to Gaussian Distribution Approximation.** The same concept can be applied to Gaussian distributions, as many approaches use Gaussian distributions to model the probability distribution for individual windows. The likelihood of a perfect Gaussian probability distribution is

$$\hat{l}_g^\alpha = \sum_i \frac{1}{N^\alpha} \ln \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i^\alpha - \bar{x}^\alpha)^2 / 2\sigma^2} \right\}$$

where  $\bar{x}^\alpha$  is the average of the sample data  $\{x_i^\alpha\}$  and  $\sigma^2$  is the unbiased variance defined as  $\sigma^2 = \sum_i (x_i^\alpha - \bar{x}^\alpha)^2 / (N^\alpha - 1)$ . The likelihood can be expressed analytically as

$$\hat{l}_g^\alpha(F) = -\ln(\sqrt{2\pi}\sigma) - \frac{N^\alpha - 1}{2N^\alpha}$$

The difference  $\Delta \hat{l}^\alpha(s, g)$ , defined as

$$\begin{aligned} \Delta \hat{l}^\alpha(s, g) &\equiv \hat{l}_s^\alpha(F) - \hat{l}_g^\alpha \\ &= -\ln Z^\alpha - \langle F(x) + W^\alpha(x) \rangle_{\text{sample}} \\ &\quad + \ln(\sqrt{2\pi}\sigma) + \frac{N^\alpha - 1}{2N^\alpha} \end{aligned} \quad (9)$$

can be viewed as the likelihood of the sampling data set of the  $\alpha$ th window being Gaussian distributed.

**Sampling Error.** As already mentioned in the Results section, simple bootstrapping methods<sup>73</sup> were utilized to estimate the statistical sampling errors in the present work. The error of a target observable is estimated by calculating the standard deviation between randomly chosen data sets with the same data size.

**Optimum of the Trial Free Energy.** For the entire set of umbrella sampling simulations, the likelihood is eq 7



Table 3. Estimated Free Energy Shifts of the Na<sup>+</sup>:Cl<sup>-</sup> System from VFEP, MBAR, and WHAM<sup>a</sup>

window ( $\alpha$ )	$\bar{x}^\alpha$	$\Delta f_{\text{VFEP}}^\alpha$	$\Delta f_{\text{MBAR}}^\alpha$	$\Delta f_{\text{WHAM}}^\alpha$	$\Delta \Delta f_{\text{VFEP/MBAR}}^\alpha (x10^3)$	$\Delta \Delta f_{\text{VFEP/WHAM}}^\alpha (x10^3)$
1	2.629					
2	2.655	-0.241	-0.240	-0.240	-0.840	-0.900
3	2.686	-0.021	-0.017	-0.017	-3.340	-3.350
4	2.754	0.629	0.633	0.633	-4.060	-4.030
5	3.113	1.476	1.477	1.479	-1.140	-3.010
6	3.945	1.512	1.511	1.519	0.790	-7.150
7	4.275	0.694	0.694	0.705	-0.410	-11.070
8	4.460	-0.084	-0.083	-0.071	-1.100	-13.410
9	4.638	-0.662	-0.660	-0.646	-1.520	-15.450
10	4.797	-1.048	-1.047	-1.031	-1.840	-17.370
11	4.964	-1.264	-1.262	-1.245	-2.050	-19.250
12	5.127	-1.317	-1.315	-1.297	-2.120	-20.870
13	5.334	-1.225	-1.223	-1.204	-2.040	-21.840
14	5.588	-1.056	-1.054	-1.033	-1.860	-22.610
15	5.863	-0.938	-0.937	-0.914	-1.720	-24.320
16	6.188	-0.970	-0.968	-0.943	-1.770	-26.700
17	6.500	-1.114	-1.112	-1.083	-1.930	-30.880
18	6.719	-1.285	-1.283	-1.235	-2.110	-49.640
19	6.955	-1.431	-1.429	-1.308	-2.130	-122.670
20	7.194	-1.532	-1.530	-1.206	-2.050	-325.970
21	7.413	-1.588	-1.586	-0.832	-2.500	-756.610
RMS					1.990	182.73

<sup>a</sup>The numbers here are derived from the same set of simulation as in Table 1.  $\bar{x}^\alpha$  is the average of the sampled coordinates of the  $\alpha$ th window. For the VFEP approach, the  $\Delta f_{\text{VFEP}}^\alpha$ 's are calculated from  $-\ln Z^\alpha$ , and the values relative to the first window are listed.  $\Delta f_{\text{MBAR}}^\alpha$  and  $\Delta f_{\text{WHAM}}^\alpha$  values were taken directly from the MBAR and WHAM output, respectively. The last two columns are differences (multiplied by  $10^3$ ) between these three types of  $\Delta f$ 's. The last row (RMS) is the root-mean-square values of the corresponding column. All values are in units of  $k_B T$ , except for  $\bar{x}^\alpha$ , which is in units of Å.

$$\begin{aligned}\hat{l}(F) &\equiv \sum_{\alpha}^{\text{windows}} c^{\alpha} \hat{l}^{\alpha}(F) \\ &= - \sum_{\alpha}^{\text{windows}} c^{\alpha} \left\{ \ln Z^{\alpha} + \frac{1}{N^{\alpha}} \sum_i^{\text{datapoints}} [F(x_i^{\alpha}) \right. \\ &\quad \left. + W^{\alpha}(x_i^{\alpha})] \right\}\end{aligned}$$

The variation of  $\hat{l}(F)$ ,  $\Delta \hat{l}(F)$ , due to a variation of  $F$ ,  $\Delta F$ , can be expressed as

$$\Delta \hat{l}(F) = \int \frac{\delta \hat{l}(F(x))}{\delta F(x)} \Delta F(x) dx \quad (10)$$

where  $\delta \hat{l}(F(x))/\delta F(x)$  is the functional derivative of  $\hat{l}(F(x))$  with respect to  $F(x)$ . Explicitly taking the functional derivative on eq 7, we get

$$\begin{aligned}\frac{\delta \hat{l}(F(x))}{\delta F(x)} &= \sum_{\alpha}^{\text{windows}} c^{\alpha} \left\{ \frac{1}{Z^{\alpha}} e^{-\{u(x) + w^{\alpha}(x)\}} \right. \\ &\quad \left. - \frac{1}{N^{\alpha}} \sum_i^{\text{datapoints}} \delta(x - x_i^{\alpha}) \right\}\end{aligned} \quad (11)$$

where  $\delta(x - x_i^{\alpha})$  is the Dirac delta function. Assuming  $c^{\alpha} = 1$  for all  $\alpha$ 's and plugging the above equations into eq 10, the likelihood variation becomes (note that  $F$  can be chosen relative to an arbitrary constant; hence,  $\Delta F$  can simply be replaced by  $F$ ):

$$\begin{aligned}\Delta \hat{l}(F) &= \int \frac{\delta \hat{l}(F)}{\delta F} \Delta F dx \\ &= \sum_{\alpha}^{\text{windows}} \{ \langle F(x) \rangle_{\text{model}}^{\alpha} - \langle F(x) \rangle_{\text{samplepoints}}^{\alpha} \}\end{aligned} \quad (12)$$

At the optimal  $F$ ,  $\delta \hat{l}(F(x))/\delta F(x)$  is zero at all  $x$ , thus  $\sum_{\alpha}^{\text{windows}} \{ \langle F(x) \rangle_{\text{model}}^{\alpha} - \langle F(x) \rangle_{\text{samplepoints}}^{\alpha} \} = 0$  (or  $\Delta \hat{l}(F) = 0$ ). Hence  $\Delta \hat{l}(F) = 0$  can be a simple criterion of judging optimal  $F$ . The above derivation is not limited to the VFEP method. Any free energy profile should hold this criterion if the optimization is based on the entire system likelihood. In all simulations reported in this paper, the magnitude of  $\Delta \hat{l}(F)$  is  $3.0 \times 10^{-5}$  or less, which indicates that the optimal (in terms of likelihood)  $F$  is reached in all of our simulations.

**Free Energy Shifts.** While in our VFEP method there is no explicit reweighting procedure involved, the term  $-\ln Z^{\alpha}$  is the relative free energy shifts,  $\Delta f^{\alpha}$ , defined in MBAR or WHAM. In VFEP, they are obtained implicitly through global optimization of the free energy profile, while in the MBAR and WHAM approaches they are calculated as the results of the reweighting procedure. Calculated  $\Delta f^{\alpha}$ 's from VFEP, MBAR, and WHAM are listed in Table 3, for the Na:Cl system with 21 windows and with a biasing potential of 5 kcal/mol/Å<sup>2</sup> (Figure 1). The relative free energy shift  $\Delta f_{\text{MLE}}^{\alpha}$  from VFEP is similar to  $\Delta f_{\text{MBAR}}^{\alpha}$  from MBAR (RMS = 0.00199), which suggests that VFEP is able to implicitly reweight windows just as MBAR. The larger differences between  $\Delta f_{\text{MLE}}^{\alpha}$  and  $\Delta f_{\text{WHAM}}^{\alpha}$  (RMS = 0.18273) may suggest that the number of data points is still not sufficient from the WHAM approach, especially for  $x > 6.7$ . For VFEP, cubic

spline functions are used for the above error analysis. Using rational interpolation functions gives virtually identical results.

**Calculated Errors.** The likelihood error estimator functions mentioned above represent the lower bound of the errors due to the usage of model functions, while the statistical sampling errors can be obtained from the bootstrapping analysis.

Table 1 lists these error estimators,  $\Delta\hat{I}^a(s,g)$  and  $\Delta\hat{I}^a(s,m)$ , the bootstrapping errors of the free energy shifts based on 50 and 100 random data sets, as well as the errors reported from MBAR. The likelihood error estimator  $\Delta\hat{I}^a(s,m)$  for this system is  $0.0356 k_B T$  (RMS value), which suggests that the model functions employed here are adequate. The Gaussian likelihood error estimator  $\Delta\hat{I}^a(s,g)$  is quite small for most of the windows. The exceptions are windows #3 to #7, which suggest that Gaussian approximation may be not ideal between  $x = 2.5$  and  $x = 4.2$ . The corresponding accumulated error from this region is around  $1.6 k_B T$ , or  $1.0$  kcal/mol, which qualitatively agrees with the fact that the converged UI result is off by about  $1.0$  kcal/mol when compared to other methods (Figure 1). The bootstrapping errors of the free energy shifts are  $0.013 k_B T$  (RMS value) for VFEP. The combined errors for VFEP (likelihood errors plus sampling errors) are roughly the same as the reported MBAR errors.

**Reduce Data Set.** Table 2 lists the bootstrapping errors for different sizes of data sets for the Na:Cl system with six windows and with a biasing potential of  $5$  kcal/mol/Å<sup>2</sup> (Figure 3). The calculated average values of free energy shifts of different windows are consistent using different numbers of data points, which indicates the reliability of the calculations. The standard deviations, seen as the sampling errors, are around  $0.03$  kcal/mol for  $10\,000$  pt/w,  $0.1$  kcal/mol for  $1000$  pt/w,  $0.2$  kcal/mol for  $100$  pt/w, and  $0.7$  kcal/mol for  $20$  pt/w.

## ■ DISCUSSIONS

Traditional methods for estimating free energy differences or free energy profiles from umbrella sampling simulations usually consist of two steps. The first step is to model the free energy profile of each window and the second step is to merge/combine the free energy profiles from individual windows into a global free energy profile. As already mentioned earlier in the Introduction, two major types of problems are inevitably associated with these traditional methods: the reweighting (combination) problem and the data representation problem. On the other hand, instead of dealing with individual windows, the proposed VFEP approach finds the global free energy profile that gives an optimal likelihood based on the entire set of simulation data that spans the free energy surface. In other words, VFEP looks for a global free energy profile that every data point is consistent with, while traditional methods look for a global free energy profile that is the best combination of local free energy profiles of individual windows. In this section, we discuss the results presented in the last section in a broader context with regard to the reweighting and data fitting problems and their relation with other methods.

### The Need of Overlap in the Reweighting Procedure.

In traditional methods, it is necessary to have overlap information between sampling windows; otherwise it is impossible to reasonably combine the free energy profiles of individual windows. Consequently, when the number of windows is not adequate and/or individual window sampling regions are too small to overlap with neighboring windows, the reweighting problem becomes intractable. In Figure 2, a strong

biasing potential leads to a small window region, and UI/MBAR/WHAM all fail to converge with six windows. The same situation happens in the two QM/MM cases as well (shown in Figures 5 and 7). Although UI evades the need of overlap in data reweighting by assuming continuous first derivatives of the free energy profile between windows, UI would fail due to numerical instability in some cases. On the other hand, the proposed VFEP approach searches for the optimal global function based on all available data and, through the usage of cubic functions, implicitly assumes continuous first and second derivatives of the free energy profile between windows; hence the lack of overlap between windows is much less severe of a problem. In all test cases, the VFEP approach gives plausible results even with very few windows, although one clearly should not expect quantitatively correct results with such sparse data. Nevertheless, the VFEP delivers a reasonable, rough estimate in these more extreme limits compared to the other methods that have been tested here.

**The Data Representation Problem.** In the traditional methods mentioned, it is desirable for the local free energy profiles of individual windows to be modeled with a stable analytic function. The quadratic approximation used in UI is often a good choice, particularly when strong biasing potentials are used, as shown in Figures 2, 5, and 7. However, this approach also leads to the need for a large number of windows, each of which is strongly localized by a harmonic biasing potential. Conversely, when weak biasing potentials are used, the quadratic approximation will begin to break down as shown in Figure 1 (the UI case). Using histograms, as in the cases of WHAM and (most commonly) MBAR, will avoid this problem but will suffer from the requirement of dense sampling in each bin in order to be numerically stable. As shown in Figure 3, WHAM will fail when the number of data points for a given window is not enough to provide sufficient sampling density. The VFEP method utilizes higher order functions to model the local free energy profiles (third order in the case of cubic spline functions) and performs very well in all test cases. Using analytic functions, VFEP also requires many fewer independent data points as shown in Figures 3, 6, and 8. Note that these reduced data sets are obtained by subsampling the original data and hence represent sparse independent data points. The results could be very different from those obtained using shorter simulation data sets possibly with higher correlations.

On the basis of the test results presented here, the proposed VFEP approach outperforms all listed methods in dealing with the above two major types of problems in estimating free energy when the overlap or the data points are not sufficient. As a result, the following potential advantages of VFEP could significantly advance the current free energy estimation techniques:

**Fast Estimate of Rough Biasing Potentials.** In recent years, much effort has been devoted to the field of adaptive approaches for free energy simulations.<sup>6,74–77</sup> In order to obtain optimal sampling, instead of fixed biasing potentials, the biasing potentials are modified adaptively according to knowledge obtained from the available simulation results. Nevertheless, adaptive approaches require at least some knowledge of the target free energy profile before any sensible modification of the biasing potentials can be made. Due to the two major problems of free energy estimation discussed above, the very first round of estimating the target free energy profile already requires significant computational resources. Test results here suggest that the VFEP is capable of delivering a

qualitatively correct free energy profile with only about five windows and ~20 to 35 independent data points per window for typical chemical reactions. With help from the VFEP approach, one may be able to establish a very quick coarse-grained picture of the free energy profile and apply an adaptive biasing potential approach to build the best biasing potentials for the next iteration of free energy estimation.

**Free Energy Profiles in Multiple Dimensions.** Theoretically it is possible to calculate the free energy profiles in multiple dimensions by slight modification of the WHAM, UI, and MBAR approaches.<sup>40,43,78–80</sup> However, in practice, it is not always feasible to do so since numerous data points are needed in order to construct a multidimensional free energy profile. The GAMUS approach<sup>43,44</sup> uses a global Gaussian fit to reduce the data points needed and can be practically used in multidimensional free energy profiles. However, the authors pointed out that the GAMUS approach was designed to explore free energy basins and is not necessarily appropriate to describe the location and magnitude of barriers along a minimum free energy pathway, possibly due to the limitation of the Gaussian approximation in providing sufficient resolution of the local free energy profiles. Nevertheless, the VFEP approach can easily be extended to multidimensional cases as eq 7 is not limited to the one-dimensional case. VFEP provides a way of constructing free energy profiles in multiple dimensions since it only needs a very small number of data points when only a qualitatively correct free energy profile is needed. As a result, one could be able to identify free energy basins quickly and focus only on important regions instead of performing simulations in all regions. Furthermore, the VFEP approach can be used iteratively with more data points to generate the quantitatively detailed free energy profile when more data are available.

**Analytic Forms of Biasing Potentials.** Another potentially significant advantage of VFEP over other methods is that the resulting free energy profiles are in analytic forms. Hence it would be straightforward to calculate the free energy derivatives with respect to the relevant coordinates. The availability of free energy derivatives will be particularly useful in the multidimensional case, in which the minimal free energy paths between two basins could be easily calculated. Such an approach has already been advocated in conjunction with the UI method.<sup>81</sup> Further, these derivatives would provide biasing forces from a global biasing potential in order to smooth out the free energy landscape for improved sampling such as in metadynamics and adaptive biasing potential methods.<sup>5–8,74–77,82–86</sup>

## CONCLUSION

In the present work, we demonstrate that the two major problems in estimating free energy profiles from umbrella sampling data can be addressed through modeling the overall free energy profile based on the whole set of simulation data. The VFEP method presented here is a variational approach based on the maximum likelihood principle and is demonstrated to generally outperform other methods for a variety of test cases in terms of the number of required windows and data points needed to construct the overall free energy profile. Whereas several other existing methods all converge to the correct free energy profile in the limit that there is sufficiently rich, well-distributed data, the VFEP method is shown to offer clear advantages in delivering stable, analytic free energy profiles under circumstances in which the data are more sparse, as are often encountered in practice. Test cases demonstrate that, for typical chemical reactions, only about five windows and

~20 to 35 data points per window are sufficient to obtain a qualitatively correct coarse-grained free energy profile that can be used to focus sampling in the most relevant regions of the surface, for example, in adaptive asynchronous Hamiltonian replica exchange simulations. The VFEP-modeled free energy profile behaves significantly better than the quadratic function-based approaches, or methods that require significant overlap between windows. Hence, VFEP provides a potentially powerful tool in the arsenal of methods used to attack the problem of free energy estimation from computer simulations of chemical reactions and processes.

## ASSOCIATED CONTENT

### Supporting Information

The choice of the combination weights. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [taisung@biomaps.rutgers.edu](mailto:taisung@biomaps.rutgers.edu); [york@biomaps.rutgers.edu](mailto:york@biomaps.rutgers.edu).

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The project is supported by the CDI type-II grant #1125332 fund from the National Science Foundation (to D.M.Y.). The authors are grateful for financial support provided by the National Institutes of Health (GM62248 to D.M.Y.). Computational resources from The Minnesota Supercomputing Institute for Advanced Computational Research (MSI) were utilized in this work. This research was supported in part by the National Science Foundation through TeraGrid resources provided by Ranger at TACC and Kraken at NICS under grant numbers TG-MCB100054 to T.-S.L. and TG-CHE100072 to D.M.Y.

## REFERENCES

- (1) Pohorille, A.; Chipot, C. *Free Energy Calculations; Springer Series in Chemical Physics*; Springer: Berlin, 2007.
- (2) Valleau, J. P.; Card, D. N. *J. Chem. Phys.* **1972**, *57*, 5457–5462.
- (3) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.
- (4) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (5) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
- (6) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. *J. Chem. Phys.* **2008**, *128*, 144120.
- (7) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.
- (8) Babin, V.; Karpusenka, V.; Moradi, M.; Roland, C.; Sagui, C. *Int. J. Quantum Chem.* **2009**, *109*, 3666.
- (9) Wu, X.; Brooks, B. R. *Adv. Chem. Phys.* **2012**, *150*, 255–326.
- (10) den Otter, W. K. *J. Chem. Phys.* **2000**, *112*, 7283–7292.
- (11) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9183.
- (12) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.
- (13) Nakajima, N.; Nakamura, H.; Kidera, A. *J. Phys. Chem. B* **1997**, *101*, 817–824.
- (14) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042–6051.
- (15) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.
- (16) Crooks, G. E. *J. Stat. Phys.* **1998**, *90*, 1481–1487.
- (17) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 3658–3661.
- (18) Minh, D. D.; Chodera, J. D. *J. Chem. Phys.* **2009**, *131*, 134110–134110.



- (19) Nilmeier, J. E.; Crooks, G. E.; Minh, D. D. L.; Chodera, J. D. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, 108.
- (20) Ballard, A. J.; Jarzynski, C. *J. Chem. Phys.* **2012**, 136, 194101.
- (21) Luckow, A.; Lacinski, L.; Jha, S. *The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*; ACM: New York, 2010; Chapter SAGA BigJob: An Extensible and Interoperable Pilot-Job Abstraction for Distributed Applications and Systems, pp 135–144.
- (22) Luckow, A.; Santcroos, M.; Weidner, O.; Merzky, A.; Maddineni, S.; Jha, S. *Proceedings of the 21st International Symposium on High-Performance Parallel and Distributed Computing, HPDC'12*; ACM: New York, 2012; Chapter Towards a common model for pilot-jobs.
- (23) Jiang, W.; Roux, B. *J. Chem. Theory Comput.* **2010**, 6, 2559–2565.
- (24) Gallicchio, E.; Levy, R. M. *Curr. Opin. Struct. Biol.* **2011**, 21, 161–166.
- (25) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, 135, 40–57.
- (26) Zwanzig, R. W. *J. Chem. Phys.* **1954**, 22, 1420–1426.
- (27) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: New York, 1987.
- (28) Paliwal, H.; Shirts, M. R. *J. Chem. Theory Comput.* **2011**, 7, 4115–4134.
- (29) Bennett, C. H. *J. Comput. Phys.* **1976**, 22, 245–268.
- (30) Kumar, S.; Bouzida, D.; Swendsen, R.; Kollman, P.; Rosenberg, J. *J. Comput. Chem.* **1992**, 13, 1011–1021.
- (31) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, 91, 140601.
- (32) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, 129, 124105.
- (33) Bartels, C. *Chem. Phys. Lett.* **2000**, 331, 446–454.
- (34) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, 122, 144107.
- (35) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, 109, 6722–6731.
- (36) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, 3, 26–41.
- (37) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. *J. Chem. Phys.* **2012**, 136, 144102.
- (38) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, 123, 144104.
- (39) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2006**, 124, 234106.
- (40) Kästner, J. *J. Chem. Phys.* **2009**, 131, 034109.
- (41) Chakravorty, D. K.; Kumarasiri, M.; Soudackov, A. V.; Hammes-Schiffer, S. *J. Chem. Theory Comput.* **2008**, 4, 1974–1980.
- (42) Chodera, J. D.; Swope, W. C.; Noe, F.; Prinz, J.-H.; Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2011**, 134, 244107.
- (43) Maragakis, P.; van der Vaart, A.; Karplus, M. *J. Phys. Chem. B* **2009**, 113, 4664–4673.
- (44) Spiriti, J.; Kamberaj, H.; Van Der Vaart, A. *Int. J. Quantum Chem.* **2012**, 112, 33–43.
- (45) Basner, J. E.; Jarzynski, C. *J. Phys. Chem. B* **2008**, 112, 12722–12729.
- (46) Kästner, J. *J. Chem. Phys.* **2012**, 136, 234102.
- (47) Kästner, J. *WIREs Comput. Mol. Sci.* **2011**, 1, 932–942.
- (48) Edwards, A. *Likelihood*; Cambridge University Press: Cambridge, U. K., 1972.
- (49) Fisher, R. A. *Phil. Trans. R. Soc. London, Ser. A* **1922**, 222, 309–368.
- (50) Aldrich, J. *Stat. Sci.* **1997**, 12, 162–176.
- (51) Maragakis, P.; Spichty, M.; Karplus, M. *Phys. Rev. Lett.* **2006**, 96, 100602.
- (52) Akima, H. *J. ACM* **1970**, 17, 589–602.
- (53) Floater, M. S.; Hormann, K. *Numer. Math.* **2007**, 107, 315–331.
- (54) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.4. <http://membrane.urmc.rochester.edu/content/wham> (accessed Dec. 2012).
- (55) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, 79, 926–935.
- (56) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, 4, 187–217.
- (57) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, 26, 1781–1802.
- (58) Nosé, S.; Klein, M. L. *Mol. Phys.* **1983**, 50, 1055–1076.
- (59) Hoover, W. G.; Ree, F. H. *J. Chem. Phys.* **1967**, 47, 4873–4878.
- (60) Case, D. A. et al. *AMBER 12*; University of California, San Francisco: San Francisco, CA, 2012.
- (61) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, 3, 486–504.
- (62) Nam, K.; Gao, J.; York, D. M. *J. Am. Chem. Soc.* **2008**, 130, 4680–4691.
- (63) Nam, K.; Gao, J.; York, D. *RNA* **2008**, 14, 1501–1507.
- (64) Lee, T.-S.; Giambasu, G. M.; Moser, A.; Nam, K.; Silva-Lopez, C.; Guerra, F.; Nieto-Faza, O.; Giese, T. J.; Gao, J.; York, D. M. *Multi-scale Quantum Models for Biocatalysis*; Springer Verlag: New York, 2009; Vol. 7, Chapter Unraveling the mechanisms of ribozyme catalysis with multi-scale simulations.
- (65) Wong, K.-Y.; Lee, T.-S.; York, D. M. *J. Chem. Theory Comput.* **2011**, 7, 1–3.
- (66) Marcos, E.; Anglada, J. M.; Crehuet, R. *Phys. Chem. Chem. Phys.* **2008**, 10, 2442–2450.
- (67) Lopez-Canut, V.; Marti, S.; Bertran, J.; Moliner, V.; Tunon, I. *J. Phys. Chem. B* **2009**, 113, 7816–7824.
- (68) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, 1, 2–13.
- (69) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, 120, 9665–9678.
- (70) Andersen, H. C. *J. Chem. Phys.* **1980**, 72, 2384–2393.
- (71) Harris, M. E.; Dai, Q.; Gu, H.; Kellerman, D. L.; Piccirilli, J. A.; Anderson, V. E. *J. Am. Chem. Soc.* **2010**, 132, 11613–11621.
- (72) Joung, I. S.; Cheatham, T. E., III. *J. Phys. Chem. B* **2008**, 112, 9020–9041.
- (73) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.
- (74) Lelievre, T.; Rousset, M.; Stoltz, G. *J. Chem. Phys.* **2007**, 126, 134111–134111.
- (75) Li, H.; Min, D.; Liu, Y.; Yang, W. *J. Chem. Phys.* **2007**, 127, 094101–094101.
- (76) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, 105, 20227–20232.
- (77) Zheng, H.; Zhang, Y. *J. Chem. Phys.* **2008**, 128, 204106.
- (78) Roux, B. *Comput. Phys. Commun.* **1995**, 91, 275–282.
- (79) Bartels, C.; Karplus, M. *J. Comput. Chem.* **1997**, 18, 1450–1462.
- (80) Bartels, C.; Schaefer, M. *J. Chem. Phys.* **1999**, 111, 8048–8067.
- (81) Bohner, M. U.; Kästner, J. *J. Chem. Phys.* **2012**, 137, 034105.
- (82) Wu, X.; Wang, S. *J. Phys. Chem. B* **1998**, 102, 7238–7250.
- (83) Lahiri, A.; Nilsson, L.; Laaksonen, A. *J. Chem. Phys.* **2001**, 114, 5993–5999.
- (84) Wang, J.; Gu, Y.; Liu, H. *J. Chem. Phys.* **2006**, 125, 094907–094907.
- (85) Babin, V.; Roland, C.; Darden, T. A.; Sagui, C. *J. Chem. Phys.* **2006**, 125, 204909–204909.
- (86) Hansen, H. S.; Hunenberger, P. H. *J. Comput. Chem.* **2010**, 31, 1–23.
- (87) Radak, B. K.; Harris, M. E.; York, D. M. *J. Phys. Chem. B* **2012**, DOI: 10.1021/jp3084277.