

Statistically optimal continuous free energy surfaces from biased simulations and multistate reweighting

Michael R. Shirts^{*,†} and Andrew L. Ferguson^{*,‡}

[†]*Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, CO 80305*

[‡]*Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637*

E-mail: michael.shirts@colorado.edu; andrewferguson@uchicago.edu

Abstract

Free energies as a function of a selected set of collective variables are commonly computed in molecular simulation, and of significant value in understanding and engineering molecular behaviors. These free energy ‘surfaces’ are most commonly estimated using variants of histogramming techniques, but such approaches obscure two important facets of these functions. First, the empirical observations along the collective variable are defined by an ensemble of discrete observations and the coarsening of these observations into a histogram bins incurs unnecessary loss of information. Second, the free energy surface is itself almost always a continuous function, and its representation by a histogram introduces inherent approximations due to the discretization. In this study, we relate the observed discrete observations to the inferred underlying continuous probability distribution over the collective variables and derive histogram-free techniques for estimating free energy surfaces. We reformulate

free energy surface estimation as minimization of a Kullback-Leibler divergence between a continuous trial function and the discrete empirical distribution and show this is equivalent to likelihood maximization of a trial function given a set of sampled data. We then present a fully Bayesian treatment of this formalism, which enables the incorporation of powerful Bayesian tools such as the inclusion of regularizing priors, uncertainty quantification, and model selection techniques. We demonstrate this new formalism in the analysis of umbrella sampling simulations for the χ torsion of a valine sidechain in the L99A mutant of T4 lysozyme with benzene bound in the cavity.

1 Introduction

The free energy as a function of a selected set of collective variable is an important observable that is ubiquitous in molecular simulation studies. Applications of free energy profiles include determining the kinetics of a reaction using the free energy along the reaction path,¹⁻³ understanding the behavior of collective interactions such as hydrophobicity,⁴⁻⁶ elucidating transport mechanisms through molecular pores,⁷⁻¹¹ and the parameterization of low-dimensional (generalized) Langevin or Fokker-Planck equations as effective reduced models of the system dynamics.¹²⁻¹⁶ This free energy function is frequently called the “free energy profile”, “free energy surface” or the “potential of mean force,” there can be subtle differences between these quantities in certain situations, which we briefly explain later in this article. We will use the terms “free energy surface” and “free energy surfaces,” and the abbreviation “FES” for both the singular and the plural in this article, to emphasize that the theory holds in more than a single dimension, though we will use the term “free energy profile” interchangeably with “free energy surface” when the collective variable has only a single dimension.

FES are typically estimated from unbiased or biased molecular simulation trajectories using a variant of histogramming techniques, most commonly a type of multiple histogram reweighting technique such as the weighted histogram analysis technique (WHAM).¹⁷

However, the process of histogramming in order to obtain the FES obscures two important issues. First, the true distribution of observations along the desired collective variable or variables in the infinite limit is virtually never actually a histogram but rather a continuous function, so the process of histogramming inherently introduces unnecessary discretization errors. Second, what we actually observe when we perform a simulation is neither a histogram, nor a continuous function, but a discrete set of delta functions, at the observed values of the collective variables. Approximating the “true” FES attained in the limit of infinite sampling of the discrete observations as a histogram inherently entails a loss of information. Although these errors can be and usually are minimized with careful choice of histogram bin width and sufficient sampling, we can resolve these problems with improved approaches to estimate a continuous FES along collective variables directly from the discrete set of empirical observations collected in the simulations, without introducing the approximation and information loss that histogramming incurs.

We are certainly not the first to observe the disadvantages of histogramming approaches. A number of recent studies have proposed histogram-free methodologies to estimate FES. Westerlund et al.¹⁸ have presented an approach that builds FES based on Gaussian mixture models, outperforming histogramming, k-nearest neighbors (kNN) and kernel density estimators (KDE). Schofield¹⁹ presented an adaptive parameterization scheme for a variety of different possible continuous functions for FES. Lee and co-workers^{20,21} presented a variational approach (variational free energy profile, or vFEP) to minimize likelihoods of observations from trial continuous free energy surfaces. Stecher et al.²² have discussed reconstructing free energy surfaces from umbrella sampling using Gaussian process regression that comes inherently equipped with uncertainty estimates. Schneider et al.²³ discuss fitting higher-dimensional FES using artificial neural networks. The umbrella integration method of Käster and Thiel^{24–26} constructs the FES by numerical integration of a weighted average of the derivative of the free energy with respect to the order parameter. Meng and Roux presented a multivariate linear regression framework

to link the biased probability densities of individual umbrella windows to yield a global free energy surface in the desired collective variables, though it uses histograms for some of the intermediate steps.²⁷ Basner and Jarzynski presented an approach to calculate a smoothly varying correction term to a trial continuous potential of mean force.²⁸

The present work shares particular similarities with the vFEP approach of Lee and co-workers^{20,21} and the adaptive parameterization approach of Schofield,¹⁹ but builds upon and goes beyond these works in two main aspects. First, as we detail in our mathematical development, we use the multistate Bennett acceptance ratio (MBAR) approach to furnish the provably minimum variance estimators of the free energy differences required to align independent biased sampling run, and then use these values to compute the maximum likelihood estimate of the unbiased FES. Second, we show how this approach can easily be placed in a fully Bayesian framework that enables transparent incorporation of Bayesian priors, Bayesian uncertainty quantification, and Bayesian model selection.

The calculation of FES parameterized by a small number of collective variables is largely motivated by the “curse of dimensionality”. Molecular systems are intrinsically exceedingly high-dimensional (with numbers of degrees of freedom in the tens or hundreds of thousands), which makes study of the system properties in the full configuration space of limited use in understanding and controlling molecular behaviors. Instead, system microstates are frequently projected into a handful of collective variables motivated by the physics of the problem at hand, and FES are then constructed over this reduced dimensional space to further analyze.

There are a number of ways to estimate FES in these collective variables. One could in theory run a simulation and estimate simply calculate the probability of visiting a representative set of the collective variables using histograms, a kernel density approximation, or averaging the mean force. However, free energy barriers in collective variable space exceeding several $k_B T$ in height—where k_B is Boltzmann’s constant and T is temperature—are crossed with exponentially small probability in standard (unbiased) simulations, re-

sulting in non-ergodic kinetic trapping and the inability to sample transition states and mechanisms.

A number of methods have been proposed to overcome this problem. They typically involve introducing some form of bias of the underlying free energy landscape to enhance sampling of low probability (high free energy) regions and accelerate transitions between high probability (low free energy) metastable states. For example, one can sample rare values of the collective coordinate by constraining a simulation along the collective variable. One can then compute the average value of the force along the collective variable, and properly (though this is nontrivial) integrating along the collective variable to obtain the free energy.^{29–32} The relationship between the mean force and the FES is why the FES in one dimension is also referred to as the “potential of mean force”.

However, perhaps the most popular and straightforward way to perform biased sampling and FES estimation is to run an ensemble of K independent simulations, each of which biases the collective variable using a—usually, but not necessarily, harmonic—biasing potential. Each biasing potential forces the simulation to spend the majority of its time visiting locations with specific ranges of the collective variables consistent with the biases. Assuming sampling orthogonal to the collective variables is sufficiently fast, good sampling of the thermally-relevant domain of the collective variable can be achieved by tiling collective variable space sufficiently densely with biasing potentials such that neighboring biased simulations sample overlapping configuration spaces. The unbiased FES can then be determined using a range of mathematical approaches based in importance sampling.^{17,33–35} Provided the collective variables employed are “good” in the sense that they adequately separate out the relevant metastable states, this methodology, which goes by the name umbrella sampling,³⁶ is a very straightforward and popular approach that works in as many dimensions as one can adequately cover the space with biasing potentials with sufficient configurational overlap. Assuming the potential only depends on the difference in collective variable from the restraint point, then the unbiased FES can

be estimated by *post hoc* analysis of the collective variable at each frame of each biased simulation trajectory without requiring records of the total energies, forces, or any other information from the simulation.¹⁷

In this paper, we establish a mathematical framework to relate a discrete observed empirical distribution determined in a set of biased simulations to the unknown and typically continuous “true” free energy surface in the collective variables one would expect in the limit of infinite sampling. We present a Bayesian treatment of this formalism to enable the incorporation of regularizing priors, uncertainty quantification, and model selection techniques. We demonstrate our approach in the analysis of umbrella sampling simulations for the χ torsion of a valine sidechain in lysozyme L99A with benzene bound in the cavity. The focus of the paper is to present analysis methodology, and so we assume that the data collected from biased simulations is sufficient to provide robust estimates of the FES using reasonable methods. As such, it is our goal to calculate the best estimate of the FES given a set of sampled data from biased simulations, where appropriate definitions of “best” are explored within this paper.

Although we do not do so here, we observe that it is possible to use current best estimates of the FES to adaptively direct additional rounds of sampling, thereby iteratively improving and refine the FES. Such adaptive methods include metadynamics,^{37–39} adiabatic free energy dynamics,⁴⁰ temperature accelerated dynamics,⁴¹ temperature accelerated molecular dynamics⁴² / driven adiabatic free energy dynamics,⁴³ adaptive biasing force approaches,³² variationally enhanced sampling,⁴⁴ and conformational flooding.⁴⁵ This class of method has both significant advantages, such as optimally directing computational effort towards under-sampled regions of collective variable space and efficiently reducing uncertainties in the FES, and significant additional challenges, such as under-sampling slow degrees of motion, and the problems of analyzing simulations that are history-dependent and thus only asymptotically approach equilibrium sampling. For the purposes of this paper we will therefore consider only equilibrium sampling as the way

to generate biased sampling trajectories for the purposes of FES estimation. However, the approach we present is extensible to any collective variable biasing enhanced sampling technique that generates equilibrium samples, and is independent of the type of shape of biasing potential, as long as the potential is not time-dependent. One could not use the current approach with the time-dependent biases in a convergence phase of metadynamics, as it would create uncontrolled biases in the result.

Importantly, we also note that our approach is also applicable to data generated with temperature, restraint, or Hamiltonian exchange,^{46–50} or expanded ensemble.^{51,52} The only requirement on the data is that samples are collected at equilibrium with respect to a time-independent (i.e., stationary) probability distribution, and the biased samples cover the range of interest of the collective variable.

2 Theory: FES estimation from biased sampling data

First, we must be precise about what is being calculated when we calculate a free energy surface. There are two different free energies as a function of collective variable that one could calculate. Hartmann et al. referred to them as as free energies of the “conditional” and “constrained” ensembles, or alternately the “geometric” and “thermodynamic” free energies. The differences between these two definitions involve differential volumes around the surface created by the collective variable constraint. “thermodynamic FES” is defined as

$$F(\vec{\xi}) = -\ln \int_{R_n} e^{-u(\vec{x})} \delta(\Phi(\vec{x}) - \vec{\xi}) d\vec{x} \quad (1)$$

Where we express energies and free energies in terms of reduced quantities, such that $u(\vec{x}) = (k_B T)^{-1} U(\vec{x})$, and similarly for all free energies. The “geometric FES”, in contrast,

is defined as:

$$F(\vec{\xi}) = -\ln \int_{\Sigma(\vec{\xi})} e^{-u(\vec{x})} d\Omega \quad (2)$$

Where $\Sigma(\vec{\xi})$ is the surface of constant $\vec{\xi}$, and $d\Omega$ is the phase space volume of this surface, and thus is the logarithm of probability density of the surface $\Sigma(\vec{\xi})$. This second quantity has also been termed the Riemannian effective potential.^{53,54} Several papers have laid out the very subtle differences in these two definitions,^{29,53} with an examination of the coarea formula being perhaps the clearest way to clarify the relationship.²⁹ The derivatives of both quantities can still be related to the mean force along the collective variable, with proper corrections for changes of variables which are beyond the scope of this summary.^{29,30}

Fortunately, these two free energy surfaces are easily related by transforming the reduced energy $u(\vec{x}) \rightarrow u(\vec{x}) \pm \ln |J_\Phi(\vec{x})|$, where J_Φ is the Jacobian of function $\Phi(\vec{x})$ that maps \vec{x} to $\vec{\xi}$, evaluated at \vec{x} .²⁹ The positive sign takes the thermodynamic energy surface to the geometric one, and the negative in the reverse direction. A non-rigorous argument for this correction, with some abuse of notation, is to note that $\int f(\vec{x}) \delta(\Phi(\vec{x}) - \vec{\xi}) d\vec{x} = \int f(\vec{x}) |J_\Phi|^{-1} \delta(\Phi(\vec{x}) - \vec{\xi}) d\vec{\xi}$, where we switch from integrating the delta function over a volume element of \vec{x} to volume elements of $\vec{\xi}$ because of the presence of Φ in the δ function.

The choice of which free energy surface to use is not always clear. The “geometric” quantity may be more useful for determining transition barriers and it is invariant to the choice of functional form in the constraint,²⁹ but the proper choice is beyond the scope of this article. We simply note that once one decides which quantity to calculate, one can replace $u(\vec{x})$ with a reduced potential with the desired Jacobian correction, and all the steps we present in this paper follow in either case. For more details on the effects of choosing coordinate systems and restraint functional forms, we recommend investigating the following citations.^{29–31,53}

After clarifying what it is we wish to calculate, we focus on how to calculate the free

energy surface. For clarity of exposition, in the present work we will assume the usual case that the biased simulation data are collected at a single temperature and this temperature is the one at which we wish to estimate the unbiased FES. However, the approach we outline here can be generalized to work with simulations in which the biased simulations are carried out at various temperatures^{35,52,55,56} or Hamiltonians,^{57,58} performed with multiple simulations of each biasing function that are each carried out with different temperatures or modified Hamiltonians, or even performed without biasing potentials, and we lay out some preliminary equations for these approaches in the Appendix.

Consider K umbrella sampling simulations with different biasing potentials tiling a collective variable space and enforcing good sampling of all thermally-relevant system configurations with desired values of the collective variable. Typically, the collective variable is 1–3 dimensional, but the formalism holds for arbitrary dimensionality provided the space can be sufficiently densely sampled and sufficient overlaps achieved between neighboring biased distributions.

The reduced potentials $u_{B,k}$ of these states are written in terms of the original potential $u(\vec{x})$ as:

$$u_{B,k}(\vec{x}) = u(\vec{x}) + b_k(\Phi(\vec{x}) - \vec{\xi}_{0,k}) \quad (3)$$

where the subscript k indexes the biased simulation, the subscript B reminds us that the potential is biased, and $b_k(\vec{\xi})$ is a user-defined biasing potential as a function of the collective variables $\vec{\xi}$ in which the umbrella sampling was performed. Most commonly, a harmonic potential is used, though the theory presented here supports *any* functional form of the bias function of the collective variables.

The value of the collective variables corresponding to a particular system configuration \vec{x} is defined by a low-dimensional mapping $\Phi(\vec{x}) = \vec{\xi}$, and the restraint point of the biasing potential in the collective variables is defined by $\vec{\xi}_{0,k}$. The biasing potentials are then chosen so that the set of all simulations with biasing potentials give roughly equal sampling across the relevant range of $\vec{\xi}$ and neighboring biased simulations share overlap

in configurational space.

We note two features of our description of umbrella sampling that are germane to our subsequent mathematical developments. First, we do not use the term “windows” as is frequently done when discussing umbrella sampling, as this word possesses significant ambiguity. “Window” could refer to either a specific interval of values of the collective variable $\vec{\xi}$, or it could refer one of the k simulations run with biasing potential b_k . These two concepts are related in that simulations with a biasing potential generally sample values in a relatively restricted volume around $\vec{\xi}_{0,k}$, but they are certainly not the same thing. A biased simulation can, in principle, yield any value of $\vec{\xi}$ (although values far from any of the bias minima are highly unlikely) so the simulation results are not strictly within any finite “window” of $\vec{\xi}$ if run for long enough.

Second, we do not make the problematic assumption that the free energy of biasing a particular simulation is equal to the value of the FES at the restraint point $\vec{\xi}_{0,k}$ of the k th biasing potential. This approximation is often called the “stiff spring” approximation,⁵⁹ as it assumes the collective variable sampling remains very close to the equilibrium position $\vec{\xi}_{0,k}$ of the bias. But the value of the free energy of biasing is a weighted average over all configurations visited by the biasing potential, and so this approximation deteriorates with increasingly weak biasing potentials. Because one has to include biasing potentials of finite width to sufficiently sample the entire volume of $\vec{\xi}$ of interest, there is always a tradeoff between the strength and number of biasing potentials used: fewer biasing potentials require weaker biases, and weaker biases result in less accurate approximations to the free energy at $\vec{\xi}_{0,k}$ under the “stiff spring” approximation. An analysis of this approximation (in the non-equilibrium pulling case) can be found in,⁶⁰ but the approach presented in the present work completely avoids this particular problem.

We also note that the problem of approximating the FES using free energy of the biasing potential is exacerbated by histogramming—as is done in WHAM—which introduces *additional* bias into the free energy calculation itself through binning of the energies as well

as the free energies.⁶¹ Any sort of averaging of the FES in each bin can be problematic because it tends to artificially lower barriers, which are frequently some of the most critical features of the FES that we wish to accurately resolve.

Given data from biased simulations, we seek the statistically optimal estimate of the FES over the collective variables $F(\vec{\xi})$. This distribution contains exactly the same information content and is essentially interchangeable with the unbiased probability distribution $P(\vec{\xi})$. These two distributions are simply related through the logarithm:

$$P(\vec{\xi}) \propto e^{-\beta F(\vec{\xi})} \quad (4)$$

where the constant of proportionality is the integral over the collective variable n -dimensional volume. We will work with whichever of the pair is most natural for the discussion at hand. The relationship above is one of proportionality because $P(\vec{\xi})$ is unnormalized. It can be turned into a proper probability density by dividing by the integral over $\vec{\xi}$ of $e^{-\beta F(\vec{\xi})}$, which will give units of length^{-d} , where d is the dimension of $\vec{\xi}$.

It is typically the case in molecular simulation that we work with relative, rather than absolute, free energies, in which case $F(\vec{\xi})$ is only defined up to an arbitrary additive constant. In this case, our estimate of the unbiased probability distribution $P(\vec{\xi})$ is only defined up to an arbitrary multiplicative constant, but this can be set by enforcing normalization.

When we perform a simulation, we obtain an observed, *empirical* probability distribution, given a set of samples $\{\vec{x}_n\}_{n=1}^N$ distributed over the space of our collective variables $\vec{\xi}$, with probability density in the collective coordinates $\vec{\xi}$:

$$p_E(\vec{\xi}|\{\vec{x}_n\}) = \sum_{n=1}^N W(\vec{x}_n) \delta(\Phi(\vec{x}_n) - \vec{\xi}) \quad (5)$$

Where $W(\vec{x}_n)$ are weights associated with each sample.

$P_E(\vec{\xi})$ is the most precise description of our sampled probability density that we can

express after a simulation, because it only involves non-zero probability where we actually have measurements, and has zero probability at values of $\vec{\xi}$ that are not observed. If we only perform a single, unbiased simulation on a continuous space, then $W(\vec{x}_n) = 1/N$ for every sample, where N is the number of samples, since—in continuous space with arbitrarily high resolution of system configurations and collective variable mapping—each observation occurs only once. However, as we describe in the next section, if we have K biased simulations, we can incorporate data from all $\sum_{k=1}^K N_k = N$ points gathered over all of the K states to better estimate $P_E(\vec{\xi})$.³⁴

2.1 MBAR and the empirical FES

The multistate Bennett acceptance ratio (MBAR) is the statistically optimal approach to estimate the reduced free energies $f_k = \int e^{-u_k(\vec{x})} d\vec{x}$, from $\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$ observations at K thermodynamic state points.³⁴ These K thermodynamic states are defined by the reduced potentials $\{u_1, u_2, \dots, u_K\}$, and we assume that the $\{\vec{x}_n\}_{n=1}^N$ are distributed according to the Boltzmann distribution corresponding to the reduced potential of the state they are collected from. With these assumptions, the MBAR estimate for the reduced free energy differences between these K states is:³⁴

$$e^{-\hat{f}_i} = \sum_{n=1}^N \frac{e^{-u_i(\vec{x}_n)}}{\sum_{k=1}^K N_k e^{\hat{f}_k - u_k(\vec{x}_n)}} \quad (6)$$

where N_k is the number of samples taken from state K . This system of equations must be solved self-consistently for the estimated reduced free energies \hat{f}_i . Since the reduced free energies are typically only defined up to an additive constant, we usually choose to pin exactly one of the estimated free energies \hat{f}_i equal to any constant value we choose and the rest follow as relative free energy differences. We note that MBAR may be considered a binless estimator of free energy differences that can be derived from WHAM in the limit of zero-width bins.^{34,62,63}

After we have solved for these \hat{f}_i , then we can calculate the weight W_i of sample \vec{x}_n in any state i as:^{34,63}

$$W_i(\vec{x}_n) = \frac{e^{\hat{f}_i - u_i(\vec{x}_n)}}{\sum_{k=1}^K N_k e^{\hat{f}_k - u_k(\vec{x}_n)}} \quad (7)$$

The weight $W_i(\vec{x}_n)$ of sample \vec{x}_n at thermodynamic state point i represents the contribution to the average of an observable A in state i under a reweighting from the *mixture distribution*, consisting of all samples collected from all K state points, to the state i .³³

The probability of each sample in the mixture distribution is $p(\vec{x}_n) = \sum_{i=1}^K \frac{N_i}{N} p_i(\vec{x}_n) = \sum_{i=1}^K \frac{N_i}{N} e^{\hat{f}_i - u_i(\vec{x}_n)}$ —in other words, simply the average of all of the individual p_i probability distributions weighted by the number of samples N_i drawn from each of the K states.³³ It can be easily checked from eq. 7 that the $W_i(\vec{x}_n)$ are normalized such that:³⁴

$$\sum_{i=1}^K N_i W_i(\vec{x}_n) = 1 \quad (8)$$

and also from eq. 6 and eq. 7 that:³⁴

$$\sum_{n=1}^N W_i(\vec{x}_n) = 1 \quad (9)$$

The expectation value of the observable A estimated over all samples at all state points may then be written as:

$$\langle A \rangle_i = \sum_{n=1}^N W_i(\vec{x}_n) A(\vec{x}_n) \quad (10)$$

as discussed in eqs. 9 and 15 of the original MBAR paper.³⁴ We denote the weight of sample \vec{x} as obtained via MBAR in the *unbiased* state as $W(\vec{x}_n)$, and in each of the $k = 1 \dots K$ *biased* states as $W_k(\vec{x}_n)$.

By eq. 4, the exponential of minus the free energy surface F_i in state i is a probability density. By combining eq. 4 and eq. 10 under the particular choice for the observable

$A(\vec{x}_n) = \delta(\Phi(\vec{x}_n) - \vec{\xi})$, we have within the MBAR framework that:

$$P(\vec{\xi}) = \langle \delta(\Phi(\vec{x}_n) - \vec{\xi}) \rangle_i = \sum_{n=1}^N W_i(\vec{x}_n) \delta(\Phi(\vec{x}_n) - \vec{\xi}) \quad (11)$$

where $\Phi(\vec{x})$ maps from the full coordinate space to the lower dimensional collective variable space of interest.

Eq. 11 makes clear that the MBAR estimate of the probability density as a function of $\vec{\xi}$ is a weighted sum of delta functions at the observed points. (Technically, it's a distribution, not a function, since it is a sum of delta functions, which are themselves distributions, but this formal distinction doesn't affect any of the development in this paper.) It is instructive to compare this to the empirical distribution function when collecting samples from a single state where $W_i(\vec{x}_n) = 1/N$:

$$P(\vec{\xi}) = \frac{1}{N} \sum_{n=1}^N \delta(\Phi(\vec{x}_n) - \vec{\xi}) \quad (12)$$

from which it can be seen that the empirical distribution $P_E(\vec{\xi}|\{\vec{x}_n\})$ generated using MBAR in eq. 5 is a *weighted* empirical distribution function using data from all states.

The representation of the empirical probability distribution function $P_E(\vec{\xi}|\{\vec{x}_n\})$ of delta functions has both advantages and disadvantages. Estimating expectation values of observables that are a function of $\vec{\xi}$ becomes simply a weighted sum over all observations

$$\langle A \rangle_i = \int A(\vec{x}) P(\vec{x}) d\vec{x} = \sum_{n=1}^N W_i(\vec{x}_n) A(\vec{x}_n). \quad (13)$$

However, it is very complicated to interpret or visualize this delta function representation. Neither can we work with this empirical representation in logarithmic form $F(\vec{\xi}) = -\ln P(\vec{\xi})$ because the logarithm of a sum of delta functions isn't defined, so only the exponential form has a well-defined mathematical meaning. We have implicitly placed the $F(\vec{\xi})$ in reduced form so that it is a pure number. We will maintain this convention

throughout the remainder of this paper. To change into real energy units we simply multiply through by $k_B T$ so that $F_{\text{units}} = (k_B T)F$.

To reiterate, expectations of quantities of interest can be computed by eq. 13 without recourse to $F(\vec{\xi})$ directly, but representing $F(\vec{\xi})$ as a continuous function is valuable for interpretation and understanding of the underlying molecular FES. If we have a continuous probability density, we can then define $F(\vec{\xi}) = -\ln P(\vec{\xi})$ up to an arbitrary normalization constant of dimension length^d required to make the argument of the logarithm unitless. We will use $F(\vec{\xi})$ to refer to the unbiased FES and $F_k(\vec{\xi})$ to the biased free energy FES obtained from each of the $k = 1 \dots K$ biased states.

Developing statistically optimal representations of $F(\vec{\xi})$ that can be visualized and exploited to understand and engineer molecular behaviors is the key motivator of the remainder of this work.

2.2 Representations of $F(\vec{\xi})$ as a continuous function

In most cases, to visualize either a $P(\vec{\xi})$ or $F(\vec{\xi})$, or to use them in some other type of mathematical modeling, we need to choose how to represent them as continuous functions. Additionally, in the infinite sampling limit for molecular systems, they generally *should* be continuous functions due to the inherent continuity of the distribution supported by non-pathological choices of $\vec{\xi}$. We now proceed to describe a number of possible choices for continuous representations of $F(\vec{\xi})$. Most of the mathematical machinery that we develop can, in principle, be deployed in arbitrarily high dimensionalities of $\vec{\xi}$, although the capacity to achieve sufficient sampling will always present an issue. We note at appropriate junctures in the text any special considerations that may arise when generalizing to high-dimensional parameterizations.

1. Represent the FES at specific locations $\vec{\xi}_0$ as the free energy of imposing each of the biasing restraints centered at $\vec{\xi}_0$. Assuming we have well-localized biasing potentials, then the free energy difference between the biased simulation and the unbiased simula-

tion can be estimated as the free energy to restrain the simulation by each of the biasing functions. As described above, this method entails significant drawbacks in overestimating valleys and underestimating peaks, and in a lack of resolution between umbrella centers. We do not pursue this further.

2. Create a histogram out of the empirical distribution. This was the default choice made in the `pymbar` package’s `computePMF` function, which has occasionally been erroneously called the MBAR estimate of the free energy surface in the literature. As we have shown, the use of MBAR is completely independent of the determination of the FES, although it can be *used* in various algorithms to estimate the FES.

We can calculate the expectation of the binning function $I_i(\vec{\xi}_i, \delta, \vec{x}) = 1$ if $\Phi(\vec{x}) > (\vec{\xi}_i - \delta/2)$ and $\Phi(\vec{x}) < (\vec{\xi}_i + \delta/2)$ and $I_i(\vec{\xi}_i, \delta, \vec{x}) = 0$ otherwise, where the $\vec{\xi}_i$ are the centers of the histogram bins and with some abuse of notation δ denotes the multidimensional bin widths, which—for clarity of exposition—we select to be equal in all dimensions. The binning function is used to essentially assign a fractional count to each bin according to the value of $W(\vec{x}_n)$ for \vec{x}_n within the bin. The free energy surface with J total indicator functions:

$$F(\vec{\xi}) = -\ln \sum_{i=1}^J \sum_{n=1}^N W(\vec{x}_n) I_i(\vec{\xi}_i, \delta, \vec{x}_n) \quad (14)$$

where the second sum, as discussed above, is over all N samples collected from all biased simulations. Since we are calculating a log expectation of a function, MBAR gives a straightforward estimate for the error in the uncertainties, as outlined in the original MBAR paper.³⁴ If the bin widths chosen adaptively with the number of samples, the uncertainty becomes more complicated, since a different data set would have a different set of bin widths. If we wished, we could fit this histogram to a smooth function, using a least square fitting method, choosing the function to balance variance and bias. However it is better to avoid any histogramming steps altogether due to the inherent and potentially uncontrolled bias that they introduce. This is especially true with multidimensional histograms, where the curse of dimensionality causes the number of bins required, and thus

the number of samples for equal resolution, to scale exponentially with dimensionality. We emphasize that with sufficient data and attention to histogram bin size, these errors can be minimized, and thus the majority of the free energy surfaces in the literature obtained by histograms are sufficiently accurate for the purposes of their studies.

When WHAM is employed to perform the FES estimation,¹⁷ the histograms used to compute the free energies are the same as the ones used to calculate the FES, which has a tendency to smooth out the FES.⁶¹ With MBAR, one can choose exactly how wide to make the histograms, since the histograms can be of any width that one chooses to best represent the underlying data, and are not constrained by the choice of separation in $\vec{\xi}$ between biasing functions $b_k(\vec{\xi})$.³⁴

3. Employ a kernel density approximation. We can replace each delta function in the empirical FES with a smooth function with weight centered at each sample and scaled by the weight. The most common choice is an isotropic Gaussian kernel $K(\vec{\xi}_i, \delta, \vec{\xi}) = (2\pi\delta^2)^{-1/2} e^{-\frac{(\vec{\xi}-\vec{\xi}_i)^2}{2\delta^2}}$, where δ now plays the role of the kernel bandwidth, but anisotropic Gaussians, “top hat,” and triangle functions are also frequently used. We observe that histogramming can be considered a form of kernel density estimation using indicator functions, with the center of the mass the preassigned bin center rather than the location of the sample. The bandwidth δ can be calculated in a number of ways, although the optimal choice is frequently not obvious.^{64–67} For example, the maximum likelihood approach with the empirical distribution shrinks δ to zero, so other approaches must be used. The FES in the kernel density approximation then becomes:

$$F(\vec{\xi}) = -\ln \sum_{n=1}^N W(\vec{x}_n) K(\Phi(\vec{x}_n), \delta, \vec{\xi}) \quad (15)$$

though to make this well-defined, one should check that the kernels result in probability being defined for all values of $\vec{\xi}$ of interest.

4. Identify a parameterized continuous probability distribution that best represents

the empirical distribution. The fundamental difficulty with this approach is that there is no unambiguous “best” continuous distribution that stands independent of any other assumptions beyond those made so far. Specifically, the closest parameter-independent continuous function to a set of δ functions, for any reasonable definitions of close, are continuous functions that are essentially indistinguishable from the δ functions themselves. It is necessary, therefore, to instead impose some constraints upon the family of continuous functions that represent our understanding of the empirical distribution as a discrete finite-data sampling of what should be a smooth and continuous distribution in the limit of infinite samples. This is an extremely flexible and generic point-of-view which allows for a variety of ways to represent the function with minimal bias and which naturally admits Bayesian formulations. The examination of this fourth perspective is our focus for the remainder of the paper. We now proceed to present a number of possible “best” choices for the representation for this continuous function along with proposed quantitative definitions of “best”.

2.3 Kullback-Leibler divergence as a measure of distance

Before we start examining mathematical forms of the trial FES, we need to decide how we will evaluate how “close” a (continuous) trial function $P_T(\vec{\xi}|\vec{\theta})$ of some arbitrary parameters $\vec{\theta}$ is to the empirical distribution $P_E(\vec{\xi}|\{\vec{x}_n\})$. For the purposes of the present mathematical development we will leave the form of $P_T(\vec{\xi}|\vec{\theta})$ abstract, but it can be useful to consider that a number of possible parameterizations for the trial function are possible, including linear interpolants, cubic splines, or piecewise cubic Hermite interpolating polynomial (PCHIP) interpolations. For non-pathological continuous representations of $P_T(\vec{\xi}|\vec{\theta})$, the corresponding FES is simply $F(\vec{\xi}|\vec{\theta}) = -\ln P_T(\vec{\xi}|\vec{\theta})$.

One logical definition of “closeness” is the Kullback-Leibler (KL) divergence from the empirical distribution in the state of interest (the one without any biasing distribution) to our trial distribution $P_T(\vec{\xi}|\vec{\theta})$, over the volume Γ of collective variables. The Kullback-

Leibler divergence from Q to P , denoted $D_{\text{KL}}(P||Q)$, can be interpreted as a measure of the information lost when Q is used to approximate P , and is defined as:

$$D_{\text{KL}}(P||Q) = \int_{\Gamma} P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{Q(\mathbf{x})} d\mathbf{x} \quad (16)$$

In later usage, we will generally omit the explicit reference to the volume Γ over the collective variable space. We will develop several different formulations of the KL divergence that each consist of a weighted sum of the function evaluated at each sampled point, and the integral of the simulation over all the entire FES (or sum of several integrals). We present them here and then later report the results of numerical tests to demonstrate their performance.

C.1. Unbiased state Kullback-Leibler divergence. The KL divergence from $P_T(\vec{\xi}|\vec{\theta})$ to $P_E(\vec{\xi}|\{\vec{x}_n\})$ is:

$$\begin{aligned} D_{\text{KL}}(\vec{\theta}) &= \int P_E(\vec{\xi}|\{\vec{x}_n\}) \ln \frac{P_E(\vec{\xi}|\{\vec{x}_n\})}{P_T(\vec{\xi}|\vec{\theta})} d\vec{\xi} \\ &= \int \left[P_E(\vec{\xi}|\{\vec{x}_n\}) \ln P_E(\vec{\xi}|\{\vec{x}_n\}) \right. \\ &\quad \left. - P_E(\vec{\xi}|\{\vec{x}_n\}) \ln P_T(\vec{\xi}|\vec{\theta}) \right] d\vec{\xi} \end{aligned} \quad (17)$$

The first term in the integral is somewhat problematic, in that it has a factor of $\ln P_E(\vec{\xi}|\{\vec{x}_n\})$, which is not well-defined for delta functions. Even taking Gaussian approximations for the delta functions and allowing them to shrink to zero-width fails to yield a well-defined value since the entire integral $\int P_E(\vec{\xi}) \ln P_E(\vec{\xi})$ is unbounded in the positive direction as the width of the δ function goes to zero. Fortunately, whatever the value may be, it is independent of the parameters $\vec{\theta}$. Accordingly, we may neglect the first term in our minimization with respect to $\vec{\theta}$ and focus only on minimization of the second term. For the purposes of functional optimization we will—with some abuse of terminology—use $D_{\text{KL}}(\vec{\theta})$ to stand for the second, $\vec{\theta}$ -dependent term, with the dropping of the first parameter-

independent term understood.

Using eq. 4, the normalized trial probability distribution can be equivalently expressed in terms of a trial free energy surface $F_T(\vec{\xi}|\vec{\theta})$:

$$P_T(\vec{\xi}|\vec{\theta}) = \frac{e^{-F_T(\vec{\xi}|\vec{\theta})}}{\int_{\Gamma} e^{-F_T(\vec{\xi}'|\vec{\theta})} d\vec{\xi}'} \quad (18)$$

If we set $W(\vec{x}) = W_{\text{unbiased}}(\vec{x})$ to be the weighting function for our unbiased reduced potential energy $u(\vec{x})$, and seek the trial free energy surface in the unbiased state $F_T(\vec{\xi}|\vec{\theta}) = F(\vec{\xi}|\vec{\theta})$, the function to be minimized reduces to:

$$\begin{aligned} D_{\text{KL}}(\vec{\theta}) &= \int -P_E(\vec{\xi}|\{\vec{x}_n\}) \ln P_T(\vec{\xi}|\vec{\theta}) d\vec{\xi} \\ &= \int P_E(\vec{\xi}|\{\vec{x}_n\}) F(\vec{\xi}|\vec{\theta}) d\vec{\xi} + \int P_E(\vec{\xi}) \ln \int e^{-F(\vec{\xi}'|\vec{\theta})} d\vec{\xi}' d\vec{\xi} \\ &= \int P_E(\vec{\xi}|\{\vec{x}_n\}) F(\vec{\xi}|\vec{\theta}) d\vec{\xi} + \ln \int e^{-F(\vec{\xi}'|\vec{\theta})} d\vec{\xi}' \\ &= \sum_{n=1}^N W(\vec{x}_n) F(\vec{\xi}_n|\vec{\theta}) + \ln \int e^{-F(\vec{\xi}'|\vec{\theta})} d\vec{\xi}' \end{aligned} \quad (19)$$

Between the 2nd and 3rd steps we can integrate out the $P_E(\vec{\xi}|\{\vec{x}_n\})$ term as $P_E(\vec{\xi}|\{\vec{x}_n\})$ is normalized, is independent of the dummy variable $\vec{\xi}'$, and $\vec{\xi}_n = \Phi(\vec{x}_n)$, and between the 3rd and 4th steps we employ eq. 13 to estimate the expectation value over the data. Minimization of eq. 19 presents a prescription to adjust $\vec{\theta}$ to find the free energy surface $F(\vec{\xi}_n|\vec{\theta})$ which is the logarithm of the closest distribution to the empirical delta function distribution calculated from MBAR.

Before proceeding to do so, it is instructive to make several observations about eq. 19.

- The biasing functions do not appear explicitly *anywhere* in eq. 19. The biases appear only implicitly through the weights associated with samples from biased states. One may therefore also carry out any other type of accelerated sampling, in addition to, or instead of biasing functions of the collective variable, as long as these simulations

have a time-independent potential (they cannot involve adaptive biasing), and are included in the K states for which MBAR reweighting is carried out and the weights $W(\vec{x}_n)$ are determined; the sum is over *all* points, collected in whatever simulation is used.

- The contribution $F(\vec{\theta}) = -\ln \int e^{-F(\vec{\xi}|\vec{\theta})} d\vec{\xi}$ is independent of the samples, and thus penalizes free energy surfaces that are simply low everywhere.
- Low free energy regions of the FES contribute more to the integral $F(\vec{\theta}) = -\ln \int e^{-F(\vec{\xi}|\vec{\theta})} d\vec{\xi}$ than high free energy regions. Accordingly, we should expect better estimates at the low values of F (high probability states), but may sacrifice accuracy at large values of F (low probability states).

C.2. Summed biased state Kullback-Leibler divergence. We can measure closeness to the KL divergence in a slightly different way, and try to find a single function that minimizes the sum of KL divergences from the K empirical distribution functions observed at each biased sample state to the trial function with the biased potential added. The motivation for this ansatz is that it will force the trial function close to the free energy surface force in all regions the biased simulations have high density and therefore good sampling. When summing over the K different biased simulations, we elect to weight the KL divergence proportional to the number of samples N_k from that state. The motivation for this choice is that states with few samples should contribute less information than states with many. We will see that this assumption leads to particularly simple results.

Under these choices we define the sample-weighted sum of Kullback-Leibler diver-

gences and function to be minimized as:

$$\begin{aligned}
\sum_{k=1}^K N_k D_{\text{KL}}(\vec{\theta}) &= \sum_{k=1}^K N_k \left(\sum_{n=1}^N W_k(\vec{x}_n) F_k(\vec{\xi}_n | \vec{\theta}) \right. \\
&\quad \left. + \ln \int e^{-F_k(\vec{\xi} | \vec{\theta})} d\vec{\xi} \right) \\
&= \sum_{k=1}^K N_k \left(\sum_{n=1}^N W_k(\vec{x}_n) \left(F(\vec{\xi}_n | \vec{\theta}) + b_k(\vec{\xi}_n) \right) \right. \\
&\quad \left. + \ln \int e^{-F(\vec{\xi} | \vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi} \right) \\
&= \sum_{n=1}^N \left(\sum_{k=1}^K N_k W_k(\vec{x}_n) \right) F(\vec{\xi}_n | \vec{\theta}) \\
&\quad + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi} | \vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi} \\
&= \sum_{n=1}^N F(\vec{\xi}_n | \vec{\theta}) \\
&\quad + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi} | \vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi}
\end{aligned} \tag{20}$$

where $F_k(\vec{\xi})$ is the free energy surface of the k th biased state, $F(\vec{\xi}_n)$ and $F_k(\vec{\xi}_n)$ are the values of F and F_k at $\Phi(\vec{x}_n) = \vec{\xi}_n$, $b_k(\vec{\xi}_n)$ is the value of the biasing potential associated with biased simulation k at $\Phi(\vec{x}_n) = \vec{\xi}_n$, and $F_k(\vec{\xi} | \vec{\theta}) = F(\vec{\xi} | \vec{\theta}) + b_k(\vec{\xi})$. We note that in moving from the second to third line we dropped the term $\sum_{k=1}^K \left(\sum_{n=1}^N W_k(\vec{x}_n) b_k(\vec{\xi}_n) \right)$ because it is independent of the $\vec{\theta}$, and thus does not affect the minimization, and in moving from the third to fourth line we appeal to the normalization condition for $W_k(\vec{x}_n)$ in eq. 8. The latter operation eliminates the weights from each individual state, leaving as the first term in our final expression an unweighted sum over the trial functions at the empirical data points. The second term is a weighted sum over an integral over the trial functions and biasing potentials and contains significant contributions only where the biasing potential is low. Large biasing potentials result in small contributions and essentially free variations of the trial function. However, as long as the trial function has significant weight

in one of the biasing functions, then it will be constrained over that region of space. In our numerical tests discussed below, it appears that eq. 20 gives additional accuracy in the densely sampled regions by sacrificing accuracy in the sparsely sampled regions, but provides superior global fits compared to those achieved by minimization of eq. 19.

It is possible in many cases to include simulations performed with other accelerated sampling methods in addition to biasing in the collective variable, but unlike in the case of the results are more complicated. We provide a preliminary analysis in the Appendix, but do not further analyze these combinations in this paper.

C.3. Summed sampled biased state Kullback-Leibler divergence. The final alternative we consider is to sum the KL divergences from the K empirical distribution functions with the biased potential added as we do in the preceding section, but only using the N_k actual samples from each biased state. In this case, each weight will be simply $1/N_k$, as each of the N_k samples will be equally weighted. We will continue to weight each state by the number of samples N_k collected from the state, as states with more samples contribute proportionally more information to the KL divergence. Following a similar development to that which led to eq. 20 and again dropping terms that are not dependent on $\vec{\theta}$ yields

the expression to be minimized as:

$$\begin{aligned}
\sum_{k=1}^K N_k D_{\text{KL}}(\vec{\theta}) &= \sum_{k=1}^K N_k \left(\sum_{n=1}^{N_k} \frac{1}{N_k} F_k(\vec{\xi}_n | \vec{\theta}) \right. \\
&\quad \left. + \ln \int e^{-F_k(\vec{\xi}^T | \vec{\theta})} d\vec{\xi}^T \right) \\
&= \sum_{k=1}^K N_k \left(\sum_{n=1}^{N_k} \frac{1}{N_k} \left(F(\vec{\xi}_n | \vec{\theta}) + b_k(\vec{\xi}_n) \right) \right. \\
&\quad \left. + \ln \int e^{-F(\vec{\xi}^T | \vec{\theta}) - b_k(\vec{\xi}^T)} d\vec{\xi}^T \right) \\
&= \sum_{k=1}^K \sum_{n=1}^{N_k} F(\vec{\xi}_n | \vec{\theta}) \\
&\quad + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}^T | \vec{\theta}) - b_k(\vec{\xi}^T)} d\vec{\xi}^T \\
&= \sum_{n=1}^N F(\vec{\xi}_n | \vec{\theta}) \\
&\quad + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}^T | \vec{\theta}) - b_k(\vec{\xi}^T)} d\vec{\xi}^T
\end{aligned} \tag{21}$$

Somewhat surprisingly, this result is exactly the same as eq. 20. This emerges due to the normalization condition for $W_k(\vec{x}_n)$ defined by eq. 8. Accordingly, whether we sum the contribution to the KL divergence of each sample over all states using the MBAR weights, or simply sum the contribution of each sample to its biased state, we will be minimizing the same function, provided we weight by the number of samples N_k from each distribution.

We could, in principle, also choose to sum over the K KL divergences without weighting each biased distribution by N_k . Doing so and following the steps leading to eq. 21

yields the expression:

$$\begin{aligned} \sum_{k=1}^K D_{\text{KL}}(\vec{\theta}) &= \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} F(\vec{\xi}_n | \vec{\theta}) \\ &\quad + \sum_{k=1}^K \ln \int e^{-F(\vec{\xi}' | \vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \end{aligned} \quad (22)$$

which is both less mathematically elegant and less intuitively satisfying than eq. 21 since simulations conducted at a state point with small N_k contribute equally to those with large N_k . Likewise, if we follow the logic of eq. 20 but employing equal weightings, we end up with a similarly unsatisfying result:

$$\begin{aligned} \sum_{k=1}^K D_{\text{KL}}(\vec{\theta}) &= \sum_{n=1}^N \left(\sum_{k=1}^K W_k(\vec{x}_n) \right) F(\vec{\xi}_n | \vec{\theta}) \\ &\quad + \sum_{k=1}^K \ln \int e^{-F(\vec{\xi}' | \vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \end{aligned} \quad (23)$$

which is not only more complicated than eq. 20, but also differs (as numerical tests confirm) from eq. 22 unless all N_k are equal, in which case $\sum_{k=1}^K W_k(\vec{x}_n) = K/N = 1/N_k$, and equality is restored. Due to these mathematically and intuitively unsatisfying features, we will not pursue eq. 22 and eq. 23 further.

2.4 Likelihood as a measure of distance

As an alternative to the Kullback-Leibler divergence, we can measure distances using likelihoods. Specifically, we can take our trial probability distribution $P_T(\vec{\xi} | \vec{\theta})$ and compute the *likelihood* of one of our N observations by evaluating the P_T associated with that observation. The observations taken together comprise our data D . Assuming the samples are independent and identically distributed (i.i.d.) observations, then we can calculate the total likelihood as the product of the individual likelihoods. The trial probability distribution as a function of θ that maximizes this likelihood will be the one closest to the

empirical distribution. In a similar manner to the KL divergence, we may construct this distribution in a number of ways. We shall show that the two choices we propose contain the same information as the KL divergence expressions, but offer greater interpretability and amenability to a Bayesian treatment.

D.1. Product over unbiased state likelihoods. Perhaps the simplest choice is to consider the joint likelihood of each weighted sample in the unbiased state. In this case, since we can consider each sample to be observed according to its weight $W(\vec{x}_n)N$ (the expected number of counts at \vec{x}_n given the empirical distribution), then the overall likelihood as a function of $\vec{\theta}$ is:

$$\ell(\vec{\theta}|\{\vec{x}_n\}) = \prod_{n=1}^N P_T(\vec{\xi}_n|\vec{\theta})^{W(\vec{x}_n)N} \quad (24)$$

and the log likelihood is:

$$\begin{aligned} \ln \ell(\vec{\theta}|\{\vec{x}_n\}) &= \sum_{n=1}^N NW(\vec{x}_n) \ln P_T(\vec{\xi}_n|\vec{\theta}) \\ &= \sum_{n=1}^N NW(\vec{x}_n) \left(-F(\vec{\xi}_n|\vec{\theta}) - \ln \int e^{-F(\vec{\xi}'|\vec{\theta})} d\vec{\xi}' \right) \\ &= -N \sum_{n=1}^N W(\vec{x}_n) F(\vec{\xi}_n|\vec{\theta}) - N \ln \int e^{-F(\vec{\xi}'|\vec{\theta})} d\vec{\xi}' \end{aligned}$$

In going from the second to the third line, we employ normalization condition in eq. 9. As expected,⁶⁸ we quickly verify that eq. 25 is identical to eq. 19 up to a factor of $(-N)$, so maximizing this log likelihood is the same as minimizing the unbiased state KL divergence.

D.2. Product over biased state likelihoods. We could also calculate the overall likelihood as the product of the likelihoods of the individual samples in each of the biased

simulations:

$$\ell(\vec{\theta}|\{\vec{x}_n\}) = \prod_{k=1}^K \prod_{n=1}^{N_k} P_T(\vec{\xi}_n|k, \vec{\theta}) \quad (25)$$

where we have denoted the probability distribution resulting from the trial FES plus the k th bias as $P_T(\vec{\xi}_n|k, \vec{\theta})$. The corresponding log likelihood is:

$$\begin{aligned} \ln \ell(\vec{\theta}|\{\vec{x}_n\}) &= \sum_{k=1}^K \sum_{n=1}^{N_k} \ln P_T(\vec{\xi}_n|k, \vec{\theta}) \\ &= \sum_{k=1}^K \sum_{n=1}^{N_k} \left(-F(\vec{\xi}_n|\vec{\theta}) - b_k(\vec{\xi}_n) \right. \\ &\quad \left. - \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \right) \\ &= \sum_{k=1}^K \left(\sum_{n=1}^{N_k} -F(\vec{\xi}_n|\vec{\theta}) \right. \\ &\quad \left. - N_k \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \right) \\ &= - \sum_{n=1}^N F(\vec{\xi}_n|\vec{\theta}) \\ &\quad - \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \end{aligned} \quad (26)$$

where in going from the second to third line we drop the $b_k(\vec{\xi}_n)$ term as independent of $\vec{\theta}$ and therefore irrelevant to the maximization. Eq. 26 is identical to eq. 20 up to a minus sign, so maximizing the product of biased state likelihoods is equivalent to minimizing the summed biased KL divergence.

D.3. Weighted product over biased state likelihoods. We could try to construct a likelihood that was consistent with the KL divergence in eq. 22 by constructing a sum of KL divergences over each state weighted by the reciprocal of the number of samples in each state:

$$\ell(\vec{\theta}|\{\vec{x}_n\}) = \prod_{k=1}^K \prod_{n=1}^{N_k} P_T(\vec{\xi}_n|k, \vec{\theta})^{\frac{1}{N_k}}, \quad (27)$$

for which the corresponding log likelihood is:

$$\begin{aligned}
\ln \ell(\vec{\theta}|\{\vec{x}_n\}) &= \sum_{k=1}^K \sum_{n=1}^{N_k} \frac{1}{N_k} \ln P_T(\vec{\xi}_n|k, \vec{\theta}) \\
&= \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} \left(-F(\vec{\xi}_n|\vec{\theta}) - b_k(\vec{\xi}_n) \right. \\
&\quad \left. - \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \right) \\
&= \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} -F(\vec{\xi}_n|\vec{\theta}) \\
&\quad - \sum_{k=1}^K \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \\
&= - \sum_{k=1}^K \frac{1}{N_k} \sum_{n=1}^{N_k} F(\vec{\xi}_n|\vec{\theta}) \\
&\quad - \sum_{k=1}^K \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \tag{28}
\end{aligned}$$

Eq. 28 is identical to eq. 22 up to a minus sign, and so maximizing the former is equivalent to minimizing the latter. However, as discussed above, there appears to be no real theoretical or practical justification reason to weight samples in the manner expressed in eq. 27 and for this reason we do not advocate the use of this formulation.

2.5 Least squares as a measure of distance

Finally, we could choose to adopt a functional form, and then perform a least squares fit to the empirical distribution or to the empirical FES in order to define a distance between the distributions. Although seemingly quite a natural and straightforward approach, it does not give rise to easily interpretable or implementable expressions. Accordingly, we defer an analysis of the least squares approach to the Appendix and do not pursue this further.

2.6 How does vFEP fit into this framework?

We now examine the correspondence of our development with the variational free energy profile (vFEP) approach developed by Lee and co-workers.^{20,21}

We first note the potential ambiguity within vFEP regarding the definition of the term “window”, which as described before, could refer to either a biasing potential, the data collected from a simulation run with that biasing potential, or a region of collective variable space within which a biased simulation has high probability density are related, but not equivalent, concepts. In the present comparison with vFEP, we will assume “window” as used in the vFEP definition refers to a biasing potential plus the data collected during simulations with that biasing potential. Under this definition of “window”, samples in the window are not included or excluded based on the associated values of $\vec{\xi}$, only on the basis of biased simulation from which they were collected.

Using the original vFEP notation, $Z^a = \int e^{-F_{i,a}(\theta,x)} dx$ is the partition function of biased simulation a and $F_{i,a}(\theta, x) = F_i(\theta, x) + W_a(x)$ is the biased trial partition function determined by parameters θ and collective variable x , where $W_a(x)$ is the biasing potential, and vectors in x and θ are implicit. Since $W_a(x)$ is not a function of θ and does not affect the minimization, the log likelihood to be maximized with respect to the parameters θ of the trial function F is:

$$\begin{aligned} \ln \ell(\theta) &= \sum_a \left[-\ln Z^a - \frac{1}{N_a} \sum_{i=1}^N F_{i,a}(\theta, x_a) \right] \\ &= \sum_a \left[-\frac{1}{N_a} \sum_{i=1}^N F_i(\theta, x_a) - \ln \int_{\Gamma_a} e^{-F_{i,a}(\theta,x)} dx \right] \end{aligned} \quad (29)$$

To proceed, we must make two assumptions: (i) the substitution of k as a label for biasing potential rather than a as the label of “windows”, (ii) the recognition that \int_{Γ_a} should be either the same or approximately the same as \int_{Γ} , since samples from biased potential

will be mostly constrained to subset of Γ , but can in principle appear anywhere in Γ . In this case, we can translate vFEP into the terminology of the present paper. The window a becomes the biased simulation k , N_a becomes N_k , x becomes $\vec{\xi}$, vectors are noted explicitly, and we obtain:

$$\ln \ell(\vec{\theta}|\{\vec{x}\}) = \sum_{k=1}^K \left[-\frac{1}{N_k} \sum_{i=1}^{N_k} F(\vec{\xi}_i|\vec{\theta}) - \ln \int e^{-F(\vec{\xi}'|\vec{\theta}) - b_k(\vec{\xi}')} d\vec{\xi}' \right] \quad (30)$$

This expression is identical to eq. 28 and, up to a minus sign, eq. 22. Accordingly, when viewed through the lens of the development presented in this paper—and with the previously mentioned assumptions about the definitions of windows and range of integrals—vFEP would correspond to a particular choice of biased state weighting within a Kullback-Leibler divergence (eq. 22) or likelihood formulation (eq. 28). As discussed above, this weighting of all simulations equally is problematic, since it puts equal weight on simulations regardless of how many samples they have. If the direct sum over biasing potentials is changed to one weighted by N_k , then it becomes eq. 26, which both easier to work with and better justified, with umbrellas with larger numbers of samples having more weight.

3 A Bayesian framework for FES estimation

Equipped with the prescriptions to calculate likelihood of observations under the different assumptions detailed in Section 2.4, we can switch to a Bayesian framework to find distributions possessing desirable features of an analytical form, continuity, and smoothness that is most consistent with our understanding of $F(\vec{\xi})$. We note that our use of a likelihood formulation, which was shown to be fully consistent with the KL divergence framework, is crucial in opening the door to a Bayesian formulation.

At the first step in this framework, we take a candidate trial distribution $P_T(\vec{\xi}|\vec{\theta})$ and

optimize its parameters $\vec{\theta}$ to form the maximum *a posteriori* probability (MAP) estimate of $P_T(\vec{\xi}|\vec{\theta})$. This estimate maximizes the Bayes posterior probability of the trial distribution, rather than simply the likelihood, given the collected (biased) samples and MBAR estimates of the relative free energy differences $\Delta f_{ij} = f_j - f_i$ between biased states.

As we introduce our Bayesian formulation, we note that the free energies emerging from the MBAR equations have no free parameters; they are the only estimated normalizing constants satisfying the self-consistent equations in eq. 6. It is possible to employ a Bayesian approach to free energy estimation by sampling of either the density of states⁶⁹ or weights of each sample in the unbiased state,⁷⁰ allowing one to incorporate additional priors about the simulations in addition to priors on the shape of the free energy surface. However, since the free energy is defined completely by the Boltzmann distribution, and since the MBAR equations provide the lowest variance importance sampling estimator and are asymptotically unbiased, then in the absence of other information about the system, it is the simplest and least biased approach to employ MBAR estimates for $\{f_i\}$.

A difference from previous efforts is that we cast our approach within a Bayesian framework that enables transparent incorporation of Bayesian priors, Bayesian uncertainty quantification, and Bayesian model selection about the functional form of the potential of mean force. Although we do not do so here, this formalism also sets the stage for adaptive sampling, in which regions of the probability distribution containing the most uncertainty are identified for additional biased sampling to optimally direct computational resources. This is similar in spirit to, but would go beyond, the adaptive approach of Schofield, which presents an elegant means to alter the analytical representation of the unbiased probability distribution to minimize uncertainty,¹⁹ to actually guiding the collection of additional data to optimally reduce uncertainty in the estimated distribution. We note that we follow a fairly standard Bayesian approach that can be found in many textbooks and other resources; one excellent presentation of Bayesian techniques in data analysis in general is offered by Ref. 71. We also note that one of the authors has previ-

ously presented a fully Bayesian treatment of WHAM in Ref. 35 that goes into more detail about the Bayesian aspects of parameter optimization as it applies to potentials of mean force.

Given the set of biased samples $\{\vec{x}_n\}$ and their collective variable mappings $\{\vec{\xi}_n\} = \{\Phi(\vec{x}_n)\}$ and the associated weights in the (unbiased) thermodynamic state calculated by MBAR $W(\vec{x}_n)$ (eq. 7), we apply Bayes' theorem⁷¹ to construct an expression for the posterior probability of the parameters $\vec{\theta}$ given the data $\{\vec{x}_n\}$, obtaining:

$$\mathcal{P}(\vec{\theta}|\{\vec{x}_n\}) = \frac{\mathcal{P}(\{\vec{x}_n\}|\vec{\theta})\mathcal{P}(\vec{\theta})}{P(\{\vec{x}_n\})} \quad (31)$$

where $\mathcal{P}(\vec{\theta}|\{\vec{x}_n\})$ is the *posterior probability* of the parameters $\vec{\theta}$ given the sampled data, $\mathcal{P}(\{\vec{x}_n\}|\vec{\theta}) = \ell(\vec{\theta}|\{\vec{x}_n\})$ is the previously-defined *likelihood* specifying the probability of the collected samples given the particular choice of parameters, $\mathcal{P}(\vec{\theta})$ is the *prior probability* of the parameters before any data have been collected, and $P(\{\vec{x}_n\}) = \int \mathcal{P}(\{\vec{x}_n\}|\vec{\theta})\mathcal{P}(\vec{\theta})d\vec{\theta}$ is the probability of observing the samples that we did (the *evidence*), serves to normalize the posterior, and contains no dependence on the parameters $\vec{\theta}$. Importantly, the prior enables us to transparently encode any prior beliefs or knowledge about the system into our analysis that can serve to regularize and stabilize our estimation.

The MAP estimate of the parameters follows from maximization of the log posterior:

$$\begin{aligned} \vec{\theta}^{\text{MAP}}(\{\vec{x}_n\}) &= \underset{\vec{\theta}}{\text{argmax}} \ln \mathcal{P}(\vec{\theta}|\{\vec{x}_n\}) \\ &= \underset{\vec{\theta}}{\text{argmax}} \left(\ln \mathcal{P}(\{\vec{x}_n\}|\vec{\theta}) + \ln \mathcal{P}(\vec{\theta}) \right) \\ &= \underset{\vec{\theta}}{\text{argmax}} \left(\ln \ell(\vec{\theta}|\{\vec{x}_n\}) + \ln \mathcal{P}(\vec{\theta}) \right) \end{aligned} \quad (32)$$

Exploiting our previous observation that maximizing a log likelihood is the same as minimizing the corresponding KL divergence from an empirical distribution,⁶⁸ we can equivalently view maximization of the Bayes posterior (eq. 32) from a frequentist perspective

as minimization of the Kullback-Leibler divergence or maximization of the log likelihood subject to regularization by the logarithm of the Bayes prior.

To use eq. 32 we need to adopt a form for the likelihood $\ell(\vec{\theta}|\{\vec{x}_n\})$ and prior $\mathcal{P}(\vec{\theta})$. The development in Section 2.4 suggests we adopt eq. 24 or 25 as candidates for the likelihood, where we explicitly assumed samples to be i.i.d. distributed. If the samples cannot be treated as i.i.d., then the counts N or N_k should be corrected by an inefficiency factor reflecting the presence of correlations in the sampling procedure.^{72,73} The simplest and most common choice for the prior is a uniform prior $\mathcal{P}(\vec{\theta}) = 1$. With no dependence on the model parameters $\vec{\theta}$, it drops out of the maximization in eq. 32 and the MAP estimate $\vec{\theta}^{\text{MAP}}$ becomes coincident with the maximum likelihood (ML) estimate $\vec{\theta}^{\text{ML}}$:

$$\vec{\theta}^{\text{ML}}(\{\vec{x}_n\}) = \underset{\vec{\theta}}{\text{argmax}} \ln \ell(\vec{\theta}|\{\vec{x}_n\}). \quad (33)$$

In principle, arbitrary priors are admissible—even improper priors that do not have a finite integral—provided the posterior is proper (i.e., integrates to unity).⁷⁴ In a Bayesian sense, we use the prior to encode prior knowledge or belief about the character of the probability distribution (such as smoothness of the splines). In the frequentist sense, the prior serves to regularize the probability estimate, providing bias-variance trade-off and compensating for sparse data. In a practical sense, the appropriate prior to adopt depends on the form of the model selected $P_T(\vec{\xi}|\vec{\theta})$, the size and quality of the simulation data, and the degree of prior belief or understanding of the system. Adopting the likelihood eq. 24 the maximization in eq. 32 can be expressed as:

$$\begin{aligned} \vec{\theta}^{\text{MAP}}(\{\vec{x}_n\}) &= \underset{\vec{\theta}}{\text{argmax}} \left[-N \sum_{n=1}^N W(\vec{x}_n) F(\vec{\xi}_n|\vec{\theta}) - N \ln \int e^{-F(\vec{\xi}|\vec{\theta})} d\vec{\xi} + \ln \mathcal{P}(\vec{\theta}) \right] \\ &= \underset{\vec{\theta}}{\text{argmin}} \left[N \sum_{n=1}^N W(\vec{x}_n) F(\vec{\xi}_n|\vec{\theta}) + N \ln \int e^{-F(\vec{\xi}|\vec{\theta})} d\vec{\xi} - \ln \mathcal{P}(\vec{\theta}) \right] \\ &= \underset{\vec{\theta}}{\text{argmin}} \left[N \sum_{n=1}^N W(\vec{x}_n) F(\vec{\xi}_n|\vec{\theta}) - \ln \mathcal{P}(\vec{\theta}) \right] \quad \text{s.t.} \quad \int_{\Gamma} e^{-F(\vec{\xi}_n|\vec{\theta})} d\vec{\xi} = 1, \end{aligned} \quad (34)$$

where in going from line 2 to 3 we have appealed to the identity $P(\vec{\xi}|\vec{\theta}) = e^{-F(\vec{\xi}|\vec{\theta})}$ (eq. 4) and asserted that this distribution must be normalized.

Adopting the product of likelihoods eq. 25 the maximization in eq. 32 becomes:

$$\begin{aligned}\vec{\theta}^{MAP}(\{\vec{x}_n\}) &= \vec{\theta}^{\text{argmax}} \left[-\sum_{n=1}^N F(\vec{\xi}_n|\vec{\theta}) - \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}|\vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi} + \ln \mathcal{P}(\vec{\theta}) \right] \\ &= \vec{\theta}^{\text{argmin}} \left[\sum_{n=1}^N F(\vec{\xi}_n|\vec{\theta}) + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}|\vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi} - \ln \mathcal{P}(\vec{\theta}) \right] \\ &= \vec{\theta}^{\text{argmin}} \left[\sum_{n=1}^N F(\vec{\xi}_n|\vec{\theta}) - \ln \mathcal{P}(\vec{\theta}) \right] \quad \text{s.t.} \quad \int_{\Gamma} e^{-F(\vec{\xi}|\vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi} = 1 \quad \forall k. \quad (35)\end{aligned}$$

There are thus two approaches to find the MAP or ML estimate: an unconstrained minimization enforcing the normalization implicitly (second-to-last lines in eq. 34 and 35), and a constrained minimization enforcing the normalization explicitly (last lines in eq. 34 and 35). The constrained minimization versions of the above expressions can be solved using the method of Lagrange multipliers or through any other constrained optimization method such as the interior point method or sequential quadratic programming (SQP). The relative efficiency of the two approaches will depend on the details of software methods available as well as the particular forms of the biases and $F(\vec{\xi}_n|\vec{\theta})$.

4 Model selection

The Akaike information criterion (AIC) or Bayesian information criterion (BIC) provide a principled means to discriminate between different possible choices for the Bayes prior and the trial probability distribution, The AIC is defined as:⁷⁵

$$AIC = 2k - 2 \ln \ell(\vec{\theta}|\{\vec{x}_n\}), \quad (36)$$

where k is the number of estimated parameters in the model. The BIC is defined as:⁷⁶

$$BIC = k \ln N - 2 \ln \ell(\vec{\theta} | \{\vec{x}_n\}), \quad (37)$$

where N is the number of data points. If we compute $\vec{\theta} = \vec{\theta}^{\text{MAP}}$ for a number of model choices i , we can use these parameter estimates to compute the set of AIC or BIC values $\{a_i\}$ for the candidate models. The model with the lowest a_i is the single model that is best supported by the data.

A more sophisticated approach to model selection defines the smallest of the $\{a_i\}$ as a_{\min} , then assigns the relative likelihood of model i as $r_i = e^{-\Delta_i/2} = e^{-(a_i - a_{\min})/2}$. The model weights follow from the normalized r_i and provide the likelihood of model i :¹⁹

$$\omega_i = \frac{r_i}{\sum_k r_k} = \frac{e^{-\Delta_i/2}}{\sum_k e^{-\Delta_k/2}}. \quad (38)$$

Adopting a threshold $q = 0.05$ (for example), the $\{r_i\}$ can be used to discard models from consideration and/or determine that there is insufficient evidence to choose one model over the other. The $\{\omega_i\}$ may also be used as weighting factors with which to construct a multi-model composed from the weighted sum of the predictions of each candidate model.

5 Bayesian uncertainty quantification

The $\vec{\theta} = \vec{\theta}^{\text{MAP}}$ estimate represents the single best point estimate of the parameters of the trial distribution $P_T(\xi | \vec{\theta})$ given the data $\{\vec{x}_n\}$ and the prior $\mathcal{P}(\vec{\theta})$. Uncertainties around these point estimates may be approximated by analytical error expectations or through bootstrap estimation.⁷⁷ A fully Bayesian uncertainty estimate is defined by the distribution of $\vec{\theta}$ dictated by the Bayes posterior.³⁵ Empirical samples of $\vec{\theta}$ from the Bayes posterior may be generated using the Metropolis-Hastings algorithm. This Markov Chain Monte-

Carlo (MCMC) approach generates a sequence of parameter realizations that converges to the stationary distribution of the Bayes posterior.⁷⁸ Under this approach we propose trial moves in $\vec{\theta}$ that are accepted or rejected according to the Metropolis-Hastings acceptance criterion:^{78,79}

$$\begin{aligned}\alpha(\vec{\theta}^\nu|\vec{\theta}^\mu) &= \min \left[\frac{\mathcal{P}(\vec{\theta}^\nu|\{\vec{x}_n\}) \cdot q(\vec{\theta}^\mu|\vec{\theta}^\nu)}{\mathcal{P}(\vec{\theta}^\mu|\{\vec{x}_n\}) \cdot q(\vec{\theta}^\nu|\vec{\theta}^\mu)}, 1 \right] \\ &= \min \left[\frac{\mathcal{P}(\{\vec{x}_n\}|\vec{\theta}^\nu) \cdot \mathcal{P}(\vec{\theta}^\nu) \cdot q(\vec{\theta}^\mu|\vec{\theta}^\nu)}{\mathcal{P}(\{\vec{x}_n\}|\vec{\theta}^\mu) \cdot \mathcal{P}(\vec{\theta}^\mu) \cdot q(\vec{\theta}^\nu|\vec{\theta}^\mu)}, 1 \right] \\ &= \min \left[\frac{\ell(\vec{\theta}^\nu|\{\vec{x}_n\}) \cdot \mathcal{P}(\vec{\theta}^\nu) \cdot q(\vec{\theta}^\mu|\vec{\theta}^\nu)}{\ell(\vec{\theta}^\mu|\{\vec{x}_n\}) \cdot \mathcal{P}(\vec{\theta}^\mu) \cdot q(\vec{\theta}^\nu|\vec{\theta}^\mu)}, 1 \right]\end{aligned}\tag{39}$$

where $\alpha(\vec{\theta}^\nu|\vec{\theta}^\mu)$ is the probability of accepting a trial move from parameter set $\vec{\theta}^\mu$ to parameter set $\vec{\theta}^\nu$, and $q(\vec{\theta}^\nu|\vec{\theta}^\mu)$ is the probability of proposing this trial move. We have invoked Bayes' Theorem (eq. 31) in going from the first line to the second, and observe that (importantly) the evidence has canceled top and bottom. In going from the second line to the third, we employed the identity $\mathcal{P}(\{\vec{x}_n\}|\vec{\theta}) = \ell(\vec{\theta}|\{\vec{x}_n\})$. In the event that symmetric trial move proposal probabilities are adopted such that $q(\vec{\theta}^\nu|\vec{\theta}^\mu) = q(\vec{\theta}^\mu|\vec{\theta}^\nu)$, the Metropolis-Hastings acceptance criterion reduces to the Metropolis criterion:^{78,80}

$$\alpha(\vec{\theta}^\nu|\vec{\theta}^\mu) = \min \left[\frac{\ell(\vec{\theta}^\nu|\{\vec{x}_n\}) \cdot \mathcal{P}(\vec{\theta}^\nu)}{\ell(\vec{\theta}^\mu|\{\vec{x}_n\}) \cdot \mathcal{P}(\vec{\theta}^\mu)}, 1 \right]\tag{40}$$

We initialize the Markov chain from $\vec{\theta}^{\text{MAP}}$ corresponding to the maximum of the Bayes posterior $\mathcal{P}(\vec{\theta}|\{\vec{x}_n\})$ and propose trial moves that maintain normalization $\int_{\Gamma} \mathcal{P}(\vec{\xi}|\vec{\theta}) d\vec{\xi} = 1$. By monitoring $\mathcal{L}(\vec{\theta}|\{\vec{x}_n\}) = \ln \left(\mathcal{P}(\{\vec{x}_n\}|\vec{\theta}) \mathcal{P}(\vec{\theta}) \right) = \ln \ell(\vec{\theta}|\{\vec{x}_n\}) + \ln \mathcal{P}(\vec{\theta})$ —which is proportional to the Bayes posterior up to an additive constant with no $\vec{\theta}$ dependence (eq. 31)—we can determine that the Markov chain has converged when $\mathcal{L}(\vec{\theta}|\{\vec{x}_n\})$ plateaus to fluctuate around a stable mean. At this point we may harvest realizations of $\vec{\theta}$ distributed according to the Bayes posterior. Using these parameter realizations, we can construct realizations

of $P_T(\vec{\xi}|\vec{\theta})$ to quantify the uncertainties in this estimated distribution.

6 Example: Umbrella sampling of protein sidechain torsion within binding cavity

As an illustrative example, we consider the application of our mathematical framework to compute a 1D FES from an umbrella sampling simulation. Code implementing these methods can be found publicly available in the `pymbar4` branch of `pymbar` (located at <http://github.com/choderalab/pymbar>), in the script `examples/umbrella-sampling/umbrella-sampling-advanced-fes.py`. The data is from an umbrella sampling simulation for the χ torsion of a valine sidechain in lysozyme L99A with benzene bound in the cavity⁸¹ (fig. 1).

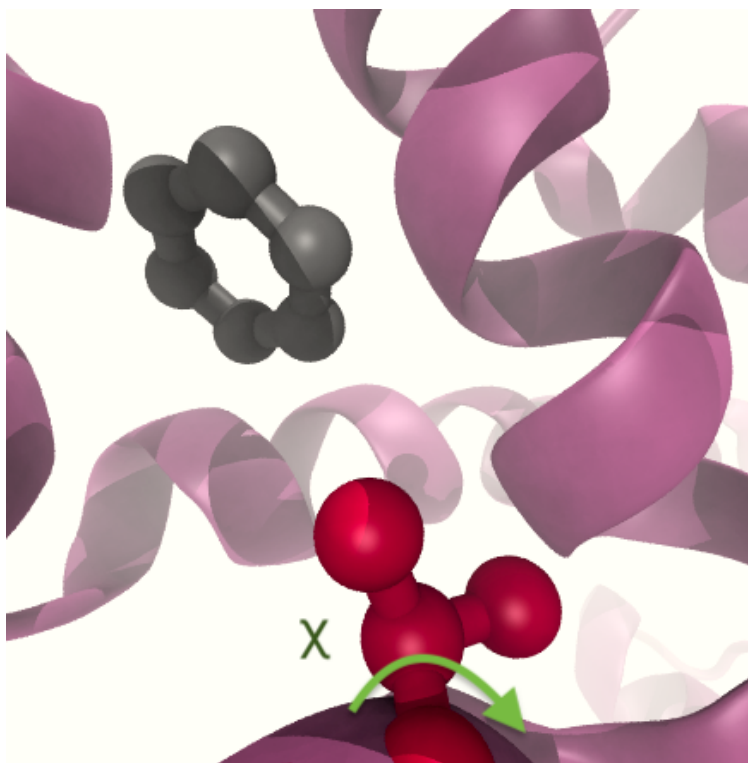


Figure 1: χ torsion angle in Lys111 of L99a T4 lysozyme around which the free energy surface is calculated using umbrella sampling.

We analyze data from 26 biased simulations employing umbrella potentials at a range of dihedral values with harmonic biasing constants of between 100 and 400 kJ/mol/nm². A 100 ps simulation was carried out under each umbrella potential with angles and energies saved every 0.2 ps for a total of 500 samples at each state. The data was analyzed for correlations, and approximately every other data point is taken (exact frequency varying with state) for a total of 7446 data points, ranging from 42 to 410 points per umbrella.

We examine the histogram approach (with 30 bins, a number chosen to be visually clear—the number of bins can be chosen completely independently of the number of umbrella simulations run), and the kernel density approximation with a Gaussian kernel, with bandwidth parameter half of the bin size, in this case, $\frac{1}{2} \times 360/30 = 6$ degrees. We also look at parameterized splines as our representation; in this case, using B-splines with varying numbers of knots placed uniformly, using cubic splines in this example; the theory is independent of these particular choices of spline.

We note that one could use splines to fit to either the FES $F(\vec{\xi}|\vec{\theta})$ or the probability distribution $P(\vec{\xi}|\vec{\theta})$. However, we find that it becomes difficult to satisfy the non-negativity condition of $P(\vec{\xi}|\vec{\theta})$ when using standard spline implementations, and that large changes in FES propagate exponentially to the probability distribution making it challenging to fit stably and robustly. For numerical stability, we therefore recommend using splines to approximate $F(\vec{\xi}|\vec{\theta})$ rather than $P(\vec{\xi}|\vec{\theta})$.

We examine the parameterized spline representations emerging from the optimizations defined by the expressions in eq. 34—corresponding to the unbiased state likelihood in eq. 24, log likelihood in eq. 25, and KL divergence in eq. 19—and eq. 35—corresponding to the product of biased states likelihood in eq. 25, log likelihood in eq. 26, and KL divergence in eq. 20. We will refer to the first as the “unbiased state likelihood”, and the second as the “biased states likelihood,” as it combines samples from all biased states.

Efficient optimization of these expressions requires calculating the gradient and potentially the Hessians. The use of B-splines, which construct the spline in terms of local

basis function, makes this calculation relatively efficient, as detailed in the Appendix. For simplicity, we elect to use a uniform distribution of spline knot locations over the domain, but these could be adaptively situated by optimizing their locations to maximize the MAP as proposed by Schofield.¹⁹

For the Bayes prior, where we compute the full posterior, rather than just the likelihood, we adopt a unnormalized Gaussian prior on the difference between successive spline knot values:

$$\mathcal{P}(\vec{\theta}) = \prod_{c=1}^{C-1} e^{-\alpha(\theta_c - \theta_{c+1})^2} \quad (41)$$

where α is a hyperparameter that controls the degree of smoothing regularization imposed upon the trial distribution. Selecting $\alpha = 0$ corresponds to a uniform prior that drops out of the maximization and $\vec{\theta}^{\text{MAP}} = \vec{\theta}^{\text{ML}}$. Selecting $\alpha > 0$ favors smoother splines with less variation from knot to knot. We examine the effect of priors governed by choice of α , where $\alpha = k/n$, where n is the number of spline knots, for some constant k . Uncertainties are estimated by MCMC sampling of the Bayes posterior using the Metropolis-Hastings algorithm and acceptance criteria (eq. 39).

The time limiting factor, both for optimizations and MCMC sampling of the posterior, is the numerical quadrature of the integral $\int P_T(\vec{\xi}|\vec{\theta})d\vec{\xi}$. For the log likelihoods from the unweighted state (eq. 25), the integral enforcing the normalization of P_T is only carried out over the unbiased trial function, whereas for approaches considering all states (eq. 26), the integral is carried out over all K trial functions with biases and is thus roughly K times slower.

The AIC and BIC allow us to select the number of spline knots best supported by the data. We plot in fig. 2 the AIC (eq. 36) and BIC (eq. 37) for the unbiased state likelihood and biased states likelihood choices. In the unbiased state case, the AIC exhibits a local minimum at 16 knots and a global minimum at 26, whereas the BIC—which penalizes

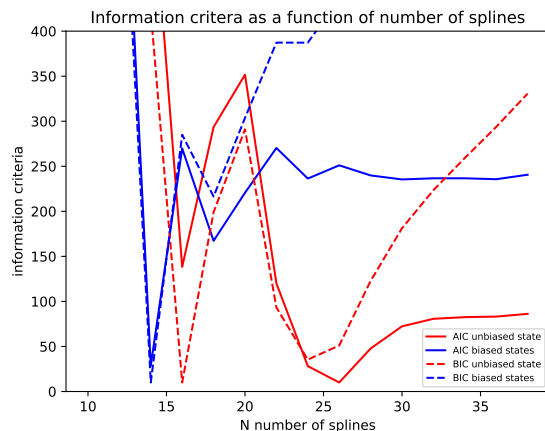


Figure 2: AIC (solid) and BIC (dotted) for splines maximizing FES likelihoods for the unbiased state estimator (red, eq. 25) and biased states estimator (blue, eq. 26) as a function of the number of spline knots, referenced from the minimum of each method. Although the curves are noisy, and occasionally nonmonotonic, they provide a useful guide towards choosing optimal numbers of parameters for models, as can be seen by comparison to Fig. 3.

excessive parameters more strongly than the AIC—possesses a local minimum at 24 knots and a global minimum at 16. In the biased states case, the AIC and BIC both exhibit clear global minima at 14 knots.

We can see how the behavior of FES changes as a function of the number of knots and how the AIC and BIC help select optimal knot numbers in fig. 3. In this figure, we plot maximum likelihood FES under the unbiased state likelihood (eq. 34, in fig. 3a) and biased states likelihood (eq. 35, in fig. 3b) as a function of the number of spline knots, along with the histogram estimate equipped with uncertainties generated from error propagation from the weights via MBAR.³⁴ As expected, higher numbers of knots provide improved fitting, but overfitting becomes clear for larger numbers of knots, especially in the case of fits using the unbiased state likelihood. However, model complexities corresponding to AIC/BIC minima fit the data relatively well in both cases. We note that the unbiased state FES fits in fig. 3a, even for the 10-knot spline, are tightly grouped at the various FES minima, but they vary significantly at the maxima, as there are less constraints on the maxima than the minima using this approach. In contrast, all fits with sufficient functional flex-

ibility (more than 10 spline knots) using the biased states approach agree relatively well across the entire range of the FES (fig. 3b), even with as few as 14 spline knots, the value corresponding to the minimum of both AIC and BIC for the biased states likelihood.

Adding bootstrapped uncertainty estimates to the FES help better show the relationship between the methods and their strengths and weaknesses. We present in Fig. 4 a comparison of the histogram (with 30 bins), kernel density approximation (with Gaussian kernels with σ of 6°), unbiased state likelihood and biased states likelihood splines employing the AIC/BIC optimal number of knots, and vFEP (using with the same number of splines as the biased states likelihood case). Uncertainties all estimates are estimated from an ensemble of 40 bootstrap samples from each of the umbrellas. All methods give relatively similar results, which is to be expected with a well-sampled system and careful selection of parameters. In particular, the FES calculated using vFEP (subject to the assumptions discussed earlier in the text) is close to the biased states likelihood approximation. This result is expected because the two approaches coincide in the limit of equal numbers of uncorrelated samples per state.

In fig. 5 we demonstrate the utility of fully Bayesian uncertainty quantification. Uncertainties in the MAP splines are computed from 50,000 (for biased states posteriors, which is slower) and 200,000 (for unbiased state posteriors) steps of MCMC sampling from the Bayes posterior. Uncertainties represent the 95% confidence intervals at each spline knot. In both cases, we show results for 10, 20, and 30 splines for two different Gaussian priors (eq. 41): (i) $\alpha = 0.1/n$ in fig. 5a and fig. 5c, where n is the number of spline knots, and (ii) $\alpha = 1/n$ in fig. 5b and fig. 5d. We recall that larger values of α impose a stronger influence of the smoothing prior and are expected to result in smoother posterior distributions. The choice of $\alpha = 0.1/n$ produces very minor differences between the ML and MAP curves (fig. 5a and 5c), whereas $\alpha = 1/n$ results in a visibly apparent difference between the two curves (fig. 5b and 5d). We see that under the biased states formulation (figs. 5c and 5d), uncertainties are relatively low and constant across the full range of the FES, whereas in

the unbiased state formulation (figs. 5a and 5b), the uncertainties are largest at the high free energy regions where the likelihood function is least constrained (cf. eq. 34). Under the unbiased state formulation, the stronger smoothing prior with $\alpha = 1/n$ (fig. 5b) is valuable in reducing the size of the confidence intervals at the peaks of the FES (note the larger y-axis range in fig. 5a required to accommodate the large uncertainty envelopes). We note that due to the significant freedom in the 30-knot splines, MCMC sampling of the probability nearly diverges in fig. 5a with $\alpha = 0.1/n$. In contrast, the biased states formulation provides more constraints across the entire FES (cf. eq. 35), and the MCMC error bounds are smaller over the entire range of the FES for both choices of α (fig. 5c and 5d).

7 Conclusions

In this paper, we have presented a Bayesian formalism to compute free energy surfaces from the empirical distributions generated by biased sampling. Within this formalism, we avoid any arbitrary choice of histogram in either the definition of the FES or the calculation of the weights, and provide clear and explicit criteria to decide which continuous free energy surfaces are most consistent with the biased sampling data. The choice and optimization of the representation of the continuous FES is completely decoupled from the choice of biasing functions and calculation of the relative free energies between the biased simulations. Biasing functions of the collective variables can be chosen, with freedom of the biasing functional form, to give appropriate sampling along the collective variables of interest, and the samples and their associated Boltzmann weights are used to construct the FES. The Bayesian formalism allows us to choose the FES that is sufficiently close to the empirical distribution of the samples we have collected, and explicitly include any prior information that we include by our choice of representation of our FES functional form. Our development also clearly demonstrates the equivalence of the likelihood-based

Bayesian formulation and Kullback-Leibler-based frequentist formulation.

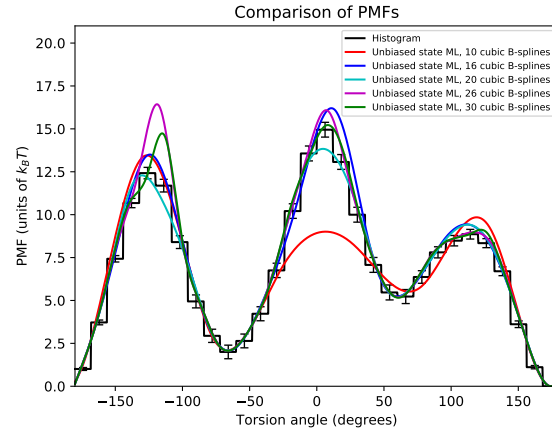
We find that the maximum likelihood calculated only from the unbiased state (eqs. 25 and 19) has a tendency to underestimate the free energy barriers in the collective variable. The product of likelihoods from all the unweighted samples collected from each biased state, weighted by the number of samples collected from each biased state (eqs. 26 and 20), has much better overall performance over the entire FES range. Surprisingly, this likelihood is exactly equal to the likelihood generated from the product over all states of the reweighted contribution of *all* samples to each biased state, again weighted by the number of samples collected from each state (cf. eqs. 20 and 21).

We can then take these likelihoods and directly incorporate them into a Bayesian inference framework. Priors on the parameters of the FES can then be chosen using whatever criteria is most appropriate; in this study we considered a Gaussian prior enforcing smoothness, but the selection can be made based on any user-defined criteria, such as tethering free energies to particular values or enforcing similarity to previously estimated distributions. We can then use MCMC sampling of the posterior of the FES curves to perform uncertainty quantification for arbitrary choices of prior.

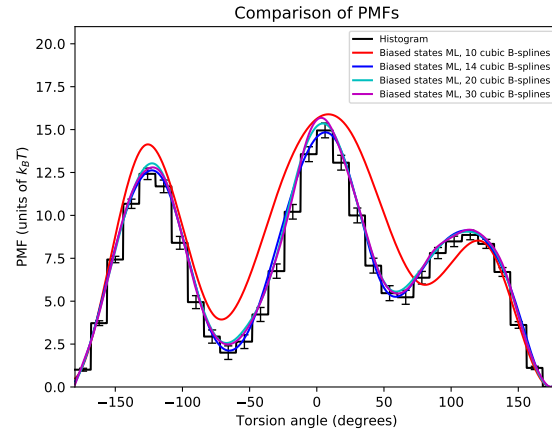
We demonstrate our approach in an application to calculation of the FES for the leucine rotation in the L99A mutant of T4 lysozyme. The unbiased state likelihood has some clear failures in that it insufficiently constrains the FES at the highest points. This failure shows up in multiple ways. When computing bootstrap uncertainties, the unbiased states approach has very high uncertainty in the barriers. With MCMC sampling, the issues become even clearer, with significant fluctuation in the parameters at the barriers unless a relatively severe prior is imposed. The biased states likelihood, however, behaves much more stably, with a well-constrained FES over the entire range, even under weak priors.

Code implementing this approach is distributed in `pymbar`, where the previous free energy surface functionality, using histograms to represent the FES, is replaced with a more comprehensive module implementing the formalism presented in this paper.

The Bayesian approach we present here approach is directly extensible to multidimensional free energy surfaces. However, the numerical details of performing the fitting may be challenging in some cases. Both the optimization processes and the MCMC require successive quadrature of the integrals $\int P_T(\vec{\xi}|\vec{\theta})d\vec{\xi}$, which in all but the simplest cases cannot be carried out analytically. The authors of vFEP have already noted this challenge²⁰ in even two dimensions with splines. This approach may also be extensible to other methods that construct biasing functions and FES adaptively, though the equations presented above will require modification if the sampling is not strictly stationary.



(a)



(b)

Figure 3: Splines maximizing the (a) unbiased state likelihood (eq. 25 or eq. 34 with uniform prior) and (b) biased states likelihood (eq. 26 or eq. 35 with uniform prior) as a function of the number of spline knots, with a histogram (black) as a reference. Knot numbers identified as optimal by both AIC and BIC appear to be good fits compared to other numbers of splines that under- or overfit the curve .

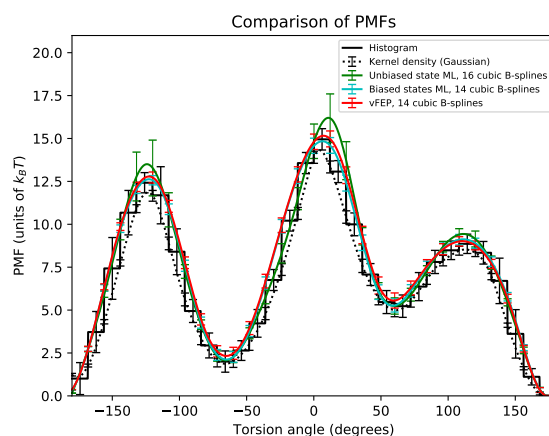


Figure 4: Comparison of methods including with bootstrap uncertainty estimates. The number of splines employed in each method was selected according to the AIC / BIC analysis in fig. 2. The same number of spline knots is used for vFEP as for the biased states estimator. The histogram employs 30 bins and Gaussian kernels with $\sigma = 6^\circ$. Uncertainties are estimated by bootstrap resampling with $n = 40$. We observe that error bars are significantly greater at the barriers for the FES maximizing the likelihood in eq. 25 than maximizing the likelihood in eq. 26, which has very low uncertainty throughout the entire range of values. Histogram uncertainties are moderately large over the entire range.

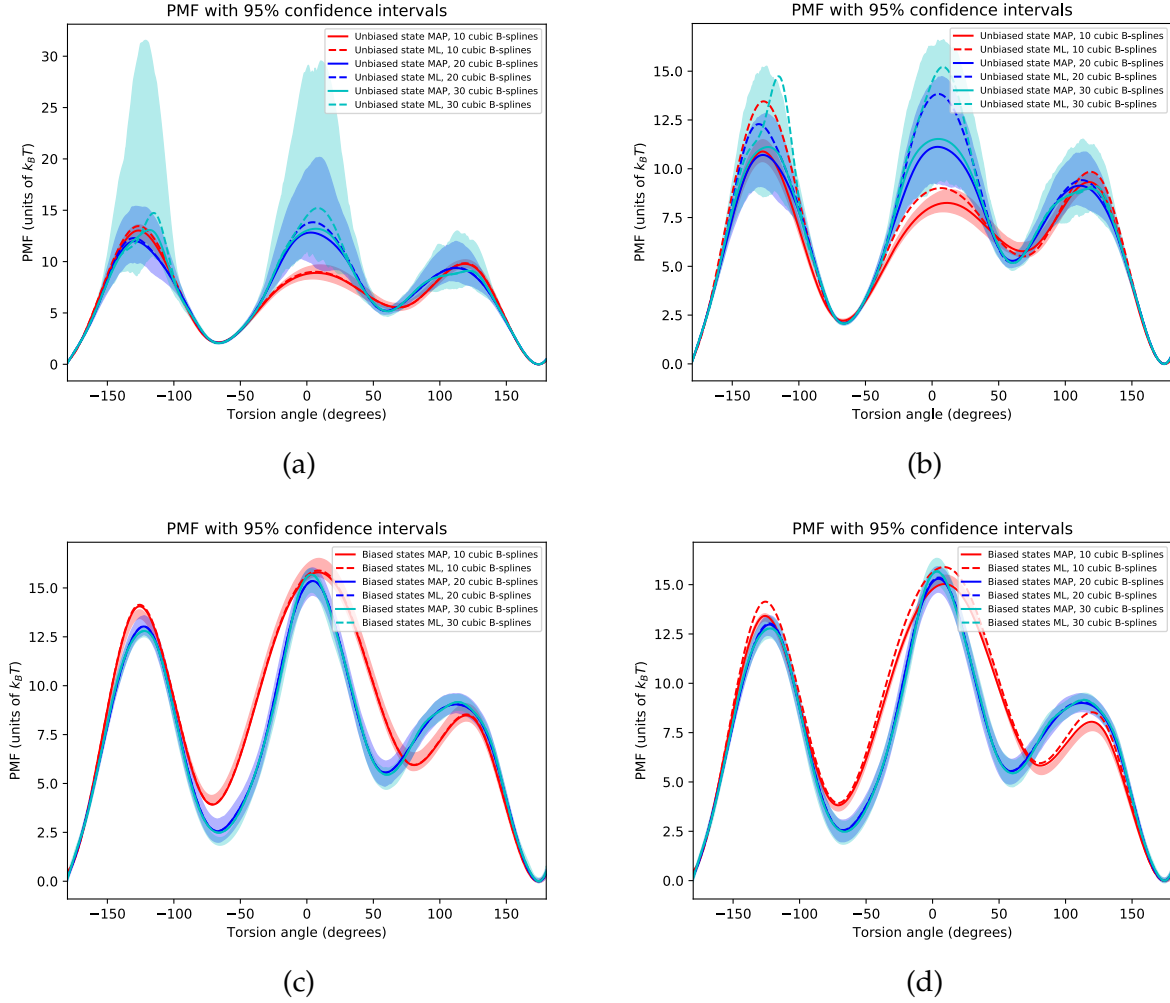


Figure 5: Comparison of MAP estimates as a function of the number of spline knots with uncertainty estimates and a Gaussian prior (eq. 41) with (a, c) $\alpha = 0.1/n$ and (b, d) $\alpha = 1/n$, where n is the number of spline knots. We illustrate the MAP distributions for (a,b) the unbiased state likelihood (eq. 34) and (b,d) biased states likelihood (eq. 35). The shading represents the 95% confidence intervals in the MAP estimate evaluated at each spline knot by MCMC sampling of the posterior, and the dashed line represents the ML solution. The MAP and ML solutions are coincident for $\alpha = 0$. The choice $\alpha = 0.1/n$ results in only minor differences between the ML and MAP solutions, whereas $\alpha = 1/n$ results in a visible difference between the two curves. In the biased states formulation (figs. 5c and 5d), the uncertainties are approximately constant across the range of the FES, whereas under the unbiased state formulation (figs. 5a and 5b), the uncertainty is largest at the high free energy regions where the likelihood function is least constrained.

Appendix

7.1 Least squares functional fitting

One possibility briefly mentioned in the main text is to minimize a least squares fit of our trial function to the empirical distribution by writing the function to be minimized as

$$\begin{aligned}
S(\vec{\theta}) &= \int \left(P_E(\vec{\xi}|\{\vec{x}_n\}) - e^{-F(\vec{\xi}|\vec{\theta})} \right)^2 d\vec{\xi} \\
&= \int P_E(\vec{\xi}|\{\vec{x}_n\})^2 - 2P_E(\vec{\xi}|\{\vec{x}_n\})e^{-F(\vec{\xi}|\vec{\theta})} \\
&\quad + e^{-2F(\vec{\xi}|\vec{\theta})} d\vec{\xi} \\
&= -2 \sum_{i=1}^N W(\vec{x}_n) e^{-F(\vec{\xi}_n|\vec{\theta})} \\
&\quad + \int e^{-2F(\vec{\xi}|\vec{\theta})} d\vec{\xi}
\end{aligned}$$

where we neglect the terms independent of $\vec{\theta}$ and employ eq. 13 to estimate the thermal average. However, this integral is problematic as it is strongly biased towards low free energy regions. Large values of F contribute very little to the sum or the log and are therefore largely unconstrained.

One could consider ameliorating this issue by minimizing over the relative error instead of the absolute. Since we can't divide by delta functions, we would have to divide by the trial function:

$$\begin{aligned}
S(\vec{\theta}) &= \int \left(\frac{P_E(\vec{\xi}|\{\vec{x}_n\}) - e^{-F(\vec{\xi}|\vec{\theta})}}{e^{-F(\vec{\xi}|\vec{\theta})}} \right)^2 d\vec{\xi} \\
&= \int \left(P_E(\vec{\xi}|\{\vec{x}_n\})^2 e^{2F(\vec{\xi}|\vec{\theta})} \right. \\
&\quad \left. - 2P_E(\vec{\xi}|\{\vec{x}_n\})e^{F(\vec{\xi}|\vec{\theta})} + 1 \right) d\vec{\xi}
\end{aligned}$$

This integral is, however, even more problematic since squares of integrals of delta functions are not well-defined and the integral over the square of a delta function is infinite. In

the direct least squares approach, we didn't really care, because this undefined function was independent of $\vec{\theta}$ and could be dropped, but in this case we must maintain this term. This seems an insurmountable deficiency and so we choose to abandon this approach.

Finally, we could consider minimizing over the squared log probabilities (i.e the FES), instead of the weights. This is *not* the Kullback-Leibler divergence, but does penalize divergence in the positive as well as the negative direction:

$$\begin{aligned}
S(\vec{\theta}) &= \int P_E(\vec{\xi}|\{\vec{x}_n\}) \left(\ln \left(\frac{P_E(\vec{\xi}|\{\vec{x}_n\})}{P_T(\vec{\xi}|\vec{\theta})} \right) \right)^2 d\vec{\xi} \\
&= \int P_E(\vec{\xi}|\{\vec{x}_n\}) \left(\ln P_E(\vec{\xi}|\{\vec{x}_n\}) - \ln P_T(\vec{\xi}|\vec{\theta}) \right)^2 d\vec{\xi} \\
&= \int P_E(\vec{\xi}|\{\vec{x}_n\}) \left(\ln P_E(\vec{\xi}|\{\vec{x}_n\})^2 \right. \\
&\quad \left. - 2 \ln P_E(\vec{\xi}|\{\vec{x}_n\}) \ln P_T(\vec{\xi}|\vec{\theta}) + \ln P_T(\vec{\xi}|\vec{\theta})^2 \right) d\vec{\xi}
\end{aligned}$$

It appears that square minimizing the log weights isn't really possible, because the logarithm of the empirical distribution of delta functions that occurs in the cross-term is not well defined. However, other least square alternatives to determining similarities of distributions involving the *cumulative distribution* have been previously presented by Schofield.¹⁹

7.2 Using biasing functions in conjunction with other accelerated sampling methods

We remove the requirement that the biasing functions are functions of the collective variable, and simply assume that they are carried out with different reduced potentials.

With these choices, the sample-weighted sum of Kullback-Leibler divergences is

$$\sum_{k=1}^K N_k D_{\text{KL}}(\vec{\theta}) = \sum_{k=1}^K N_k \left(\sum_{n=1}^N W_k(\vec{x}_n) F_k(\vec{\xi}_n|\vec{\theta}) + \ln \int e^{-F_k(\vec{\xi}|\vec{\theta})} d\vec{\xi} \right)$$

We first need to clarify what $\int e^{-F_k(\vec{\xi}'|\vec{\theta})} d\vec{\xi}'$ means if the biasing function is not a function of $\vec{\xi}$. If this is the case, then there appears to be no clear relationship between $F_k(\vec{\xi}|\vec{\theta})$ and $F(\vec{\xi}|\vec{\theta})$, so information about $F_k(\vec{\xi}|\vec{\theta})$ will not help us. So we claim that at present, we must sum over only biased simulations that have energy function of the form of eq. 3, a sum of the $u(\vec{x})$ of interest and a bias function that only depends on $\vec{\xi}$.

However, each of the biased simulations *can* have many (say, M_k) simulations accelerated with other methods associated with it as are of interest, and we can use this information to build our empirical estimate of the K biasing potentials which we do sum over. There are two primary situations we can consider.

First, reweighting is performed only between simulations that share the same biasing function, and they are reweighted to only that simulation with only the bias and no other potential-based acceleration. These K reweighted terms are summed with some K -dependent weights. In this case, there are thus K different sets of weights $W_k^{k'}(\mathbf{x}_n)$, one for each of the K MBAR evaluations for reweighting, where the subscripts denote that the weight is determined for the k simulations with biases alone, and the superscripts denote which set of weights are involved. In this case, N_k corresponds to the total number of samples from all M_k simulations associated with that biasing weight.

However, we don't know what the simplest factor is with which to weight the K biasing simulations. Because the effective number of samples at any of the K biased states will be less than N_k , we replace the weighting N_k with a constant C_k to be

determined later. We then find:

$$\begin{aligned}
\sum_{k=1}^K C_k D_{\text{KL}}(\vec{\theta}) &= \sum_{k=1}^K C_k \left(\sum_{n=1}^{N_k} W_k^{k'}(\vec{x}_n) F_k(\vec{\xi}_n | \vec{\theta}) \right. \\
&\quad \left. + \ln \int e^{-F_k(\vec{\xi} | \vec{\theta})} d\vec{\xi} \right) \\
&= \sum_{k=1}^K \left(C_k \sum_{n=1}^{N_k} W_k^{k'}(\vec{x}_n) F_k(\vec{\xi}_n | \vec{\theta}) \right) \\
&\quad + \sum_{k=1}^K C_k \ln \int e^{-F_k(\vec{\xi} | \vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi}
\end{aligned}$$

Unlike for the derivation of eq. 20, we cannot interchange the order of summation, and so there are no obvious choices for C_k . One could choose an “effective number of samples” for C_k , such as $\sum_{n=1}^{N_k} W_k(\vec{x}_n)$, though it is not clear if this is optimal. However, this is still a usable equation to minimize divergence or as a log-likelihood.

In the second case, we assume that all $M = \sum_{k=1}^K M_k$ simulations are used to calculate the MBAR weights, meaning we must be able to calculate the potential energy of each simulation in all M simulations. In this case, there is a single set of weights $W_k(\vec{x}_n)$. However, the normalization is a bit different than is used in eq. 20. Although there is a single $W_k(\vec{x}_n)$ corresponding to the weights in the k biased potentials, the normalization $\sum_{k=1}^K N_k W_k(\vec{x}_n) = 1$ no longer holds, since we would need to sum over all M states, not just the K different bias potentials. We again use a weighted linear scaling C_k because the “best” weighting is not clear:

$$\begin{aligned}
\sum_{k=1}^K C_k D_{\text{KL}}(\vec{\theta}) &= \sum_{k=1}^K C_k \left(\sum_{n=1}^N W_k(\vec{x}_n) F_k(\vec{\xi}_n | \vec{\theta}) \right. \\
&\quad \left. + \ln \int e^{-F_k(\vec{\xi} | \vec{\theta})} d\vec{\xi} \right) \\
&= \sum_{n=1}^N \left(\sum_{k=1}^K C_k W_k(\vec{x}_n) \right) F(\vec{\xi}_n | \vec{\theta}) \\
&\quad + \sum_{k=1}^K C_k \ln \int e^{-F(\vec{\xi} | \vec{\theta}) - b_k(\vec{\xi})} d\vec{\xi}
\end{aligned}$$

This version is again somewhat more complex, but usable as log-likelihood or a divergence to minimize. One can again choose an “effective” number of samples in the biased state for C_k , such as $C_k = \sum_{n=1}^N W_k(\vec{x}_n)$, though again it is not entirely clear if this is optimal.

7.3 Solving the minimization problem

We briefly describe efficient optimization routines to solve the minimization problems defined in eqs. 34 and 35 in the case of splines. In below, we suppress explicit dependence of F on θ for compactness. We start by examining the minimization of eq. 35:

$$S(\theta) = \sum_{n=1}^N F(\vec{\xi}_n) + \sum_{k=1}^K N_k \ln \int e^{-F(\vec{\xi}') - b_k(\vec{\xi}')} d\vec{\xi}' - \ln \mathcal{P}(\vec{\theta})$$

Various minimization approaches are required to compute the gradient and Hessian of this function with respect to the parameter vector $\vec{\theta}$. For convenience, we define the equilibrium average performed with biasing function k of some observable A that is a function of $\vec{\theta}$ as:

$$\langle A(\vec{\theta}) \rangle_k = \frac{\int A(\vec{\xi}'|\vec{\theta}) e^{-F(\vec{\xi}',\theta) - b_k(\vec{\xi}')} d\vec{\xi}'}{\int e^{-F(\vec{\xi}',\theta) - b_k(\vec{\xi}')} d\vec{\xi}'}$$

The i components of the gradient are then:

$$\nabla S(\theta)_i = \sum_{n=1}^N \frac{\partial F(\vec{\xi})}{\partial \theta_i} + \sum_{k=1}^K N_k \left\langle \frac{\partial F(\vec{\xi}')}{\partial \theta_i} \right\rangle_k - \frac{1}{\mathcal{P}(\theta)} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i}$$

We note that if we have linear basis functions, the first term is independent of $\vec{\theta}$ and can be precomputed, as $\frac{\partial F}{\partial \theta_i}$ is simply the corresponding basis function. Additionally, the integral term will have only limited support for each basis function, so the integrals are relatively easy to carry out, and the calculations scales easily in the number of basis functions.

The ij entries in the Hessian are::

$$\begin{aligned}
\nabla^2 S(\theta)_{ij} = & \sum_{n=1}^N \frac{\partial^2 F(\vec{\xi}_n)}{\partial \theta_i \partial \theta_j} \\
& - \sum_{k=1}^K N_k \left[\left\langle \frac{\partial^2 F(\vec{\xi})}{\partial \theta_i \partial \theta_j} \right\rangle_k - \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle_k \right. \\
& \left. + \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \right\rangle_k \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle_k \right] \\
& - \left[\frac{1}{\mathcal{P}(\theta)} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i \partial \theta_j} - \frac{1}{\mathcal{P}(\theta)^2} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_j} \right]
\end{aligned} \tag{42}$$

If we assume that we have a trial function that is linear in the parameters, then the initial terms involving mixed second derivatives vanish, leaving only:

$$\begin{aligned}
\nabla^2 S(\theta)_{ij} = & \sum_{k=1}^K N_k \left[\left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle_k \right. \\
& \left. - \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \right\rangle_k \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle_k \right] \\
& - \left[\frac{1}{\mathcal{P}(\theta)} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i \partial \theta_j} - \frac{1}{\mathcal{P}(\theta)^2} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_j} \right]
\end{aligned} \tag{43}$$

If the function is linear in the parameters (again, such as splines), this will only be nonzero in areas where basis functions have mutual support, essentially just banded along the diagonal, so are be relatively inexpensive to compute.

In the case of eq. 34, this becomes:

$$\nabla S(\theta)_i = N \sum_{n=1}^N W_n(\vec{x}_n) \frac{\partial F(\vec{\xi})}{\partial \theta_i} - N \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \right\rangle - \frac{1}{\mathcal{P}(\theta)} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i} \tag{44}$$

$$\begin{aligned}
\nabla^2 S(\theta)_{ij} = & N \left(\left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle \right. \\
& \left. - \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_i} \right\rangle \left\langle \frac{\partial F(\vec{\xi})}{\partial \theta_j} \right\rangle \right) \\
& - \left[\frac{1}{\mathcal{P}(\theta)} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i \partial \theta_j} - \frac{1}{\mathcal{P}(\theta)^2} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_i} \frac{\partial \mathcal{P}(\theta)}{\partial \theta_j} \right]
\end{aligned} \tag{45}$$

Where expectations are now over the *unbiased* state rather than any of the K biased simulations.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1841810 (to A.L.F.). The authors thank Mahmoud Moradi, Lucie Delemotte, and Darrin York for useful insights and discussion.

References

- (1) Chandler, D. Statistical Mechanics of Isomerization Dynamics in Liquids and the Transition State Approximation. *J. Chem. Phys.* **1978**, *68*, 2959–2970.
- (2) Northrup, S. H.; Pear, M. R.; Lee, C. Y.; McCammon, J. A.; Karplus, M. Dynamical Theory of Activated Processes in Globular Proteins. *PNAS* **1982**, *79*, 4035–4039.
- (3) Schenter, G. K.; Garrett, B. C.; Truhlar, D. G. Generalized Transition State Theory in Terms of the Potential of Mean Force. *J. Chem. Phys.* **2003**, *119*, 5828–5833.
- (4) San Biagio, P. L.; Bulone, D.; Martorana, V.; Palma-Vittorelli, M. B.; Palma, M. U. Physics and Biophysics of Solvent Induced Forces: Hydrophobic Interactions and Context-Dependent Hydration. *Eur. Biophys. J.* **1998**, *27*, 183–196.

- (5) Sobolewski, E.; Makowski, M.; Czaplewski, C.; Liwo, A.; Odziej, S.; Scheraga, H. A. Potential of Mean Force of Hydrophobic Association: Dependence on Solute Size. *J. Phys. Chem. B* **2007**, *111*, 10765–10774.
- (6) Makowski, M.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. Potential of Mean Force of Association of Large Hydrophobic Particles: Toward the Nanoscale Limit. *J. Phys. Chem. B* **2010**, *114*, 993–1003.
- (7) Hub, J. S.; de Groot, B. L. Mechanism of Selectivity in Aquaporins and Aquaglyceroporins. *PNAS* **2008**, *105*, 1198–1203.
- (8) Hub, J. S.; Winkler, F. K.; Merrick, M.; de Groot, B. L. Potentials of Mean Force and Permeabilities for Carbon Dioxide, Ammonia, and Water Flux across a Rhesus Protein Channel and Lipid Membranes. *J. Am. Chem. Soc.* **2010**, *132*, 13251–13263.
- (9) Allen, T. W.; Andersen, O. S.; Roux, B. Molecular Dynamics Potential of Mean Force Calculations as a Tool for Understanding Ion Permeation and Selectivity in Narrow Channels. *Biophys. Chem.* **2006**, *124*, 251–267.
- (10) Medovoy, D.; Perozo, E.; Roux, B. Multi-Ion Free Energy Landscapes Underscore the Microscopic Mechanism of Ion Selectivity in the KcsA Channel. *BBA-Biomembranes* **2016**, *1858*, 1722–1732.
- (11) Sigg, D. Modeling Ion Channels: Past, Present, and Future. *J. Gen. Physiol.* **2014**, *144*, 7–26.
- (12) Yang, S.; Onuchic, J. N.; Garca, A. E.; Levine, H. Folding Time Predictions from All-Atom Replica Exchange Simulations. *J. Mol. Biol.* **2007**, *372*, 756–763.
- (13) Hummer, G.; Kevrekidis, I. G. Coarse Molecular Dynamics of a Peptide Fragment: Free Energy, Kinetics, and Long-Time Dynamics Computations. *J. Chem. Phys.* **2003**, *118*, 10762–10773.

- (14) Kopelevich, D. I.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. Coarse-Grained Kinetic Computations for Rare Events: Application to Micelle Formation. *J. Chem. Phys.* **2005**, *122*, 044908.
- (15) Rzepiela, A. J.; Schaudinnus, N.; Buchenberg, S.; Hegger, R.; Stock, G. Communication: Microsecond Peptide Dynamics from Nanosecond Trajectories: A Langevin Approach. *J. Chem. Phys.* **2014**, *141*, 241102.
- (16) Chiavazzo, E.; Gear, C. W.; Dsilva, C. J.; Rabin, N.; Kevrekidis, I. G. Reduced Models in Chemical Kinetics via Nonlinear Data-Mining. *Processes* **2014**, *2*, 112–140.
- (17) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
- (18) Westerlund, A. M.; Harpole, T. J.; Blau, C.; Delemotte, L. Inference of Calmodulins Ca²⁺-Dependent Free Energy Landscapes via Gaussian Mixture Model Validation. *J. Chem. Theory Comput.* **2018**, *14*, 63–71.
- (19) Schofield, J. Optimization and Automation of the Construction of Smooth Free Energy Profiles. *J. Phys. Chem. B* **2017**, *121*, 6847–6859.
- (20) Lee, T.-S.; Radak, B. K.; Huang, M.; Wong, K.-Y.; York, D. M. Roadmaps through Free Energy Landscapes Calculated Using the Multidimensional vFEP Approach. *J. Chem. Theory Comput.* **2014**, *10*, 24–34.
- (21) Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M. A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 153–164.
- (22) Stecher, T.; Bernstein, N.; Csnyi, G. Free Energy Surface Reconstruction from Um-

- rella Samples Using Gaussian Process Regression. *J. Chem. Theory Comput.* **2014**, *10*, 4079–4097.
- (23) Schneider, E.; Dai, L.; Topper, R. Q.; Drechsel-Grau, C.; Tuckerman, M. E. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. *Phys. Rev. Lett.* **2017**, *119*, 150601.
- (24) Kstner, J.; Thiel, W. Bridging the Gap between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: Umbrella Integration. *J. Chem. Phys.* **2005**, *123*, 144104.
- (25) Kstner, J. Umbrella Integration in Two or More Reaction Coordinates: The. *J. Chem. Phys.* **2009**, *131*, 034109.
- (26) Kstner, J. Umbrella Integration with Higher-Order Correction Terms. *J. Chem. Phys.* **2012**, *136*, 234102.
- (27) Meng, Y.; Roux, B. Efficient Determination of Free Energy Landscapes in Multiple Dimensions from Biased Umbrella Sampling Simulations Using Linear Regression. *J. Chem. Theory Comput.* **2015**, *11*, 3523–3529.
- (28) Basner, J. E.; Jarzynski, C. Binless Estimation of the Potential of Mean Force. *J. Phys. Chem. B* **2008**, *112*, 12722–12729.
- (29) Hartmann, C.; Latorre, J. C.; Ciccotti, G. On Two Possible Definitions of the Free Energy for Collective Variables. *Eur. Phys. J. Spec. Top.* **2011**, *200*, 73–89.
- (30) den Otter, W. K. Revisiting the Exact Relation between Potential of Mean Force and Free-Energy Profile. *J. Chem. Theory Comput.* **2013**, *9*, 3861–3865.
- (31) Chipot, C.; Kollman, P. A.; Pearlman, D. A. Alternative Approaches to Potential of Mean Force Calculations: Free Energy Perturbation versus Thermodynamic Integra-

- tion. Case Study of Some Representative Nonpolar Interactions. *J. Comput. Chem.* **1996**, *17*, 1112–1131.
- (32) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive Biasing Force Method for Scalar and Vector Free Energy Calculations. *J. Chem. Phys.* **2008**, *128*, 144120.
- (33) Shirts, M. R. Reweighting from the Mixture Distribution as a Better Way to Describe the Multistate Bennett Acceptance Ratio. *Arxiv Prepr.* **2017**, [704.00891.
- (34) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.
- (35) Ferguson, A. L. BayesWHAM: A Bayesian Approach for Free Energy Estimation, Reweighting, and Uncertainty Quantification in the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2017**, *38*, 1583–1605.
- (36) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.
- (37) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *PNAS* **2002**, *99*, 12562–12566.
- (38) Huber, T.; Torda, A. E.; van Gunsteren, W. F. Local Elevation: A Method for Improving the Searching Properties of Molecular Dynamics Simulation. *J. Comput.-Aid. Mol. Des.* **1994**, *8*, 695–708.
- (39) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.
- (40) Rosso, L.; Minry, P.; Zhu, Z.; Tuckerman, M. E. On the Use of the Adiabatic Molecular Dynamics Technique in the Calculation of Free Energy Profiles. *J. Chem. Phys.* **2002**, *116*, 4389–4402.
- (41) Sørensen, M. R.; Voter, A. F. Temperature-Accelerated Dynamics for Simulation of Infrequent Events. *J. Chem. Phys.* **2000**, *112*, 9599–9606.

- (42) Maragliano, L.; Vanden-Eijnden, E. A Temperature Accelerated Method for Sampling Free Energy and Determining Reaction Pathways in Rare Events Simulations. *Chem. Phys. Lett.* **2006**, *426*, 168–175.
- (43) Abrams, J. B.; Tuckerman, M. E. Efficient and Direct Generation of Multidimensional Free Energy Surfaces via Adiabatic Dynamics without Coordinate Transformations. *J. Phys. Chem. B* **2008**, *112*, 15742–15757.
- (44) Valsson, O.; Parrinello, M. Variational Approach to Enhanced Sampling and Free Energy Calculations. *Phys. Rev. Lett.* **2014**, *113*, 090601.
- (45) Grubmiller, H. Predicting Slow Structural Transitions in Macromolecular Systems: Conformational Flooding. *Phys. Rev. E* **1995**, *52*, 2893–2906.
- (46) Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J Chem Phys* **2000**, *113*, 6042.
- (47) Bergonzo, C.; Henriksen, N. M.; Roe, D. R.; Swails, J. M.; Roitberg, A. E.; Cheatham, T. E. Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide. *J. Chem. Theory Comput.* **2014**, *10*, 492–499.
- (48) Li, P.-C.; Miyashita, N.; Im, W.; Ishido, S.; Sugita, Y. Multidimensional Umbrella Sampling and Replica-Exchange Molecular Dynamics Simulations for Structure Prediction of Transmembrane Helix Dimers. *J. Comput. Chem.* **2014**, *35*, 300–308.
- (49) Dickson, A.; Ahlstrom, L. S.; Brooks, C. L. Coupled Folding and Binding with 2D Window-Exchange Umbrella Sampling. *J. Comput. Chem.* **2016**, *37*, 587–594.
- (50) Kstner, J. Umbrella Sampling. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.
- (51) Fenwick, M. K.; Escobedo, F. A. Expanded Ensemble and Replica Exchange Methods for Simulation of Protein-like Systems. *J Chem Phys* **2003**, *119*, 11998.

- (52) Chodera, J. D.; Shirts, M. R. Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing. *J. Chem. Phys.* **2011**, *135*, 194110.
- (53) Fakharzadeh, A.; Moradi, M. Effective Riemannian Diffusion Model for Conformational Dynamics of Biomolecular Systems. *J. Phys. Chem. Lett.* **2016**, *7*, 4980–4987.
- (54) Goolsby, C.; Fakharzadeh, A.; Moradi, M. Thermodynamic and Kinetic Characterization of Protein Conformational Dynamics within a Riemannian Diffusion Formalism. *bioRxiv* **2019**, 707711.
- (55) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (56) Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.
- (57) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* **2002**, *116*, 9058–9067.
- (58) Kwak, W.; Hansmann, U. H. E. Efficient Sampling of Protein Structures by Model Hopping. *Phys. Rev. Lett.* **2005**, *95*, 138102.
- (59) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free Energy Calculation from Steered Molecular Dynamics Simulations Using Jarzynski's Equality. *J. Chem. Phys.* **2003**, *119*, 3559–3566.
- (60) Hummer, G.; Hummer, G. Free Energy Profiles from Single-Molecule Pulling Experiments. *PNAS* **2010**, *107*, 21441–21446.
- (61) Fajer, M.; Swift, R. V.; McCammon, J. A. Using Multistate Free Energy Techniques

- to Improve the Efficiency of Replica Exchange Accelerated Molecular Dynamics. *J. Comput. Chem.* **2009**, 30, 1719–1725.
- (62) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, 136, 144102.
- (63) Bartels, C. Analyzing Biased Monte Carlo and Molecular Dynamics Simulations. *Chem. Phys. Lett.* **2000**, 331, 446–454.
- (64) Park, B.; Turlach, B. *Practical Performance of Several Data Driven Bandwidth Selectors*; 1992.
- (65) Cao, R.; Cuevas, A.; Gonzalez Manteiga, W. A Comparative Study of Several Smoothing Methods in Density Estimation. *Comput. Stat. Data An.* **1994**, 17, 153–176.
- (66) Jones, M. C.; Marron, J. S.; Sheather, S. J. A Brief Survey of Bandwidth Selection for Density Estimation. *J. Am. Stat. Assoc.* **1996**, 91, 401–407.
- (67) Sheather, S. J.; Jones, M. C. A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *J. Roy. Stat. Soc. B Met.* **1991**, 53, 683–690.
- (68) Eguchi, S.; Copas, J. Interpreting KullbackLeibler Divergence with the NeymanPearson Lemma. *J. Multivariate Anal.* **2006**, 97, 2034–2040.
- (69) Habeck, M. Bayesian Estimation of Free Energies From Equilibrium Simulations. *Phys. Rev. Lett.* **2012**, 109.
- (70) Moradi, M.; Enkavi, G.; Tajkhorshid, E. Atomic-Level Characterization of Transport Cycle Thermodynamics in the Glycerol-3-Phosphate:Phosphate Antiporter. *Nat Commun* **2015**, 6, 1–11.
- (71) Sivia, D.; Skilling, J. *Data Analysis: A Bayesian Tutorial*, 2nd ed.; Oxford University Press: Oxford, 2006.

- (72) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. Temperature Weighted Histogram Analysis Method, Replica Exchange, and Transition Paths. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.
- (73) Zhu, F.; Hummer, G. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2012**, *33*, 453–465.
- (74) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. *Bayesian Data Analysis*, 3rd ed.; Chapman and Hall/CRC: Boca Raton, 2013.
- (75) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
- (76) Schwarz, G. Estimating the Dimension of a Model. *Ann. Statist.* **1978**, *6*, 461–464.
- (77) Paliwal, H.; Shirts, M. R. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J. Chem. Theory Comput.* **2011**, *7*, 4115–4134.
- (78) Smith, R. C. *Uncertainty Quantification: Theory, Implementation, and Applications*; SIAM-Society for Industrial and Applied Mathematics: Philadelphia, 2013.
- (79) Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.
- (80) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (81) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. Predicting Absolute Ligand Binding Free Energies to a Simple Model Site. *J. Mol. Biol.* **2007**, *371*, 1118–1134.

Table of Contents Graphic

