Finite-sample bias in free energy bridge estimators

Cite as: J. Chem. Phys. 151, 034105 (2019); doi: 10.1063/1.5097384

Submitted: 25 March 2019 • Accepted: 19 June 2019 •

Published Online: 15 July 2019

Brian K. Radak^{a)}

AFFILIATIONS

Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801-2325, USA

^{a)}Electronic mail: bradak@ks.uiuc.edu

ABSTRACT

In practical free energy estimation, the bias is often neglected once it has been shown to vanish in the large-sample limit. Yet finite-sample bias always exists and ought to be considered in any rigorous study. This work develops a metric for bias in a broad class of free energy “bridge estimators” (e.g., Bennett’s method). The framework complements existing variance estimation methods and provides a means for comparing systematic and statistical errors. Examples show that, contrary to what is often assumed, the bias can be quite substantial when the sample size is modest.

Published under license by AIP Publishing. <https://doi.org/10.1063/1.5097384>

I. INTRODUCTION

Accurate and precise free energy calculations are a central aim of statistical physics and molecular simulations.^{1–3} Indeed, numerous methods have been developed for this purpose,^{4–16} although all of them roughly fall into the two categories of thermodynamic integration (TI)⁴ and free energy perturbation (FEP)⁵ (including analogs based on nonequilibrium processes^{17,18}). Much effort has been exerted in evaluating the strengths and weaknesses of these approaches,^{19–25} and the main concern is the trade-off between errors due to systematic bias and statistical variance.

It was noted by Shirts and Chodera¹⁴ that nearly all methods in the FEP category can be viewed within the “bridge estimator” framework from the statistics literature. Consequently, most FEP-like estimators are now rigorously known to be unbiased in the large-sample limit.²⁶ The framework also provides a general path for developing analytic estimates of the variance.^{14,16} By contrast, there is rarely, if ever, an *a priori* guarantee that TI is asymptotically unbiased. Accordingly, the general practice for free energy calculations of both kinds is simply to estimate sample variances and then assume that statistical noise exceeds any and all bias. However, just because an estimator is asymptotically unbiased does not mean that a finite sample does not contain bias nor does it strictly guarantee that

statistical error dominates.²⁷ Therefore, it would be of significant use to have a practical quantitative metric for bias.

This work develops a simple protocol for estimating bias in free energy calculations based on bridge estimators. This includes the widely used Bennett acceptance ratio (BAR) method⁸ as well as its close cousin simple overlap sampling (SOS).¹³ Ultimately, the approach generalizes an ansatz introduced by Lu *et al.* to analyze conventional FEP calculations.^{13,28–30} In addition to practical computations, the framework here permits a stronger understanding of why methods like BAR are so effective and offers new avenues to think about the relationships between different sources of error.

II. THEORY AND BACKGROUND

Following the standard development of free energy calculations,^{1–3} the relative free energy Δf (in reduced units, absorbing factors of $k_B T$) between two ensembles (or “states”) is being sought. As is customary in FEP, Δf is defined in terms of a transition from some “forward” ensemble to a “backward” ensemble—the subscripts “*f*” and “*b*” will be used as appropriate. Data collection in each ensemble proceeds by gathering the reduced energy difference Δu between states (defined as the reduced energy in the backward ensemble less

that in the forward ensemble). Sampling from the two ensembles thus consists in drawing values of Δu from the probability distribution functions (PDFs) p_f and p_b . Ensemble averages (indicated by angle brackets $\langle \cdot \rangle$) can be computed from these PDFs in the usual way.

In a seminal work, Bennett⁸ derived exact and finite-sample relationships for Δf computed from two ensembles (see also the work of Shirts *et al.*¹⁸). Most notably, he showed that Δf can be written in terms of an arbitrary function α which is now identified as a bridge estimator as follows:^{14,26}

$$e^{\varepsilon - \Delta f} = \frac{\langle \alpha(\Delta u - \varepsilon) \rangle_f}{\langle e^{\Delta u - \varepsilon} \alpha(\Delta u - \varepsilon) \rangle_b}. \quad (1)$$

Here, ε is an arbitrary energy shift parameter. A related perspective was articulated in the overlap sampling framework explored by Lu, Singh, and Kofke.¹³

Bennett derived Eq. (1) by invoking detailed balance and interpreting α as the acceptance ratio for a special kind of *post hoc* Monte Carlo move (hence, the name BAR). A similar view is also at the core of Crooks' theorems pertaining to nonequilibrium work,^{31,32} and it is worth noting that the results here hold equally well in the case that the Δu are work values. Bennett's particular choice, $\alpha(x) = \exp(-x/2) \text{sech}(x/2)$, belongs to a class of functions with the property $\exp(x)\alpha(x) = \alpha(-x)$ such that the denominator in Eq. (1) becomes $\langle \alpha(\varepsilon - \Delta u) \rangle_b$. This equality holds for essentially all estimators in common use and will be used hereafter for simplicity but is otherwise unnecessary.

In the context of a finite-sample estimate of Eq. (1) with (effective) sample sizes N_f and N_b , Bennett also showed that there is a unique, self-consistent value $\varepsilon = \hat{\Delta f} - \ln(N_f/N_b)$ that, along with his choice for α , minimizes the asymptotic variance of the free energy estimate $\hat{\Delta f}$. Using a similar line of reasoning, Lu, Singh, and Kofke¹³ suggested that the even simpler choice $\alpha(x) = \exp(-x/2)$ with $\varepsilon = 0$ (which they called SOS) is nearly as optimal as BAR, at least when Δf is small. Yet another choice briefly considered by Bennett is the standard Metropolis criterion $\alpha(x) = \min[1, \exp(-x)]$. All of these choices are everywhere positive and monotonically decrease to zero. These characteristics are hereafter assumed.

A. Bias in FEP calculations

Long after Bennett's work, Lu and Kofke^{28,29} presented a systematic error analysis for conventional FEP calculations. Standard FEP eschews the two-ensemble approach by making two separate choices: $\alpha_f(x) = \exp(-x)$ ("forward FEP") and $\alpha_b(x) = 1$ ("backward FEP")—each of these eliminates one of the ensemble averages. In a variation on the previous identity, these choices satisfy the relation $\exp(x)\alpha_f(x) = \alpha_b(-x)$. Unfortunately, both types of FEP have rather disappointing properties.^{27–29} In particular, the forward estimate $\hat{\Delta f}_f$ tends to overestimate Δf (i.e., the bias is positive) while the backward estimate $\hat{\Delta f}_b$ tends to underestimate it (i.e., the bias is negative).

Lu and Kofke²⁸ demonstrated the bias in FEP by first noting that even very large samples tend to leave the ensembles undersampled beyond some threshold value. That is, samples from the forward ensemble neglect values below some cutoff Δu_f

while backward samples neglect values above another cutoff $\Delta u_b > \Delta u_f$. Using this threshold ansatz, it can be shown that the biases are²⁸

$$\hat{\Delta f}_f - \Delta f = -\ln[1 - P_b(\Delta u_f)] \approx +P_b(\Delta u_f), \quad (2a)$$

$$\hat{\Delta f}_b - \Delta f = +\ln[1 - \bar{P}_f(\Delta u_b)] \approx -\bar{P}_f(\Delta u_b), \quad (2b)$$

where P_b is the cumulative distribution function (CDF) associated with p_b and $\bar{P}_f(x) \equiv 1 - P_f(x)$ is the complementary CDF associated with p_f ,

$$P_b(\Delta u_f) = \int_{-\infty}^{\Delta u_f} d\Delta u p_b(\Delta u), \quad (3a)$$

$$\bar{P}_f(\Delta u_b) = \int_{\Delta u_b}^{\infty} d\Delta u p_f(\Delta u). \quad (3b)$$

Examples of these PDFs and sample threshold values are shown in the top half of Fig. 1. Taylor expansions for small values of the bias (or CDFs) make the signs more explicit.

B. Bias in bridge sampling calculations

The same threshold ansatz used for FEP can also be applied to Eq. (1). A value of Δf computed with such thresholds will not be the exact value but rather some particular (biased) estimate $\hat{\Delta f}$,³³

$$e^{\varepsilon - \hat{\Delta f}} = \frac{\int_{\Delta u_f}^{\infty} d\Delta u \alpha(\Delta u - \varepsilon) p_f(\Delta u)}{\int_{-\infty}^{\Delta u_b} d\Delta u \alpha(\varepsilon - \Delta u) p_b(\Delta u)}. \quad (4)$$

FIG. 1. Viewed graphically, the threshold ansatz implies that the bias is proportional to the area under the distributions beyond the sample threshold. In this example, the threshold values Δu_f and Δu_b are observed from short simulations while the "exact" distributions are observed from long simulations. The FEP biases are related to integrals of p_f and p_b , whereas the bridge estimate biases depend on integrals of the modified PDFs defined by Eq. (5).

These new bounds of integration are subsumed by those in the exact expression and can be separated out leaving integrals over the tail regions. A series of equalities (see [Appendix A](#)) can be used to transform the integrals into CDFs analogous to those from the FEP result [Eq. (3)] but instead with two new PDFs

$$p_b^\alpha(\Delta u; \varepsilon) \equiv \frac{\alpha(\varepsilon - \Delta u)p_b(\Delta u)}{\int_{-\infty}^{\Delta u_b} d\Delta u' \alpha(\varepsilon - \Delta u')p_b(\Delta u')}, \quad (5a)$$

$$p_f^\alpha(\Delta u; \varepsilon) \equiv \frac{\alpha(\Delta u - \varepsilon)p_f(\Delta u)}{\int_{\Delta u_f}^{\infty} d\Delta u' \alpha(\Delta u' - \varepsilon)p_f(\Delta u')}. \quad (5b)$$

The shapes of these modified PDFs for BAR and SOS are shown in the bottom frames of [Fig. 1](#). Interestingly, in the case of BAR, ε is adjusted so that the finite sample estimates for both of the denominators are equal and identical to the overlap integral noted by Bennett⁸ [Eq. (11) therein].

Solving Eq. (4) for the sample bias yields

$$\Delta \hat{f} - \Delta f = \ln \left[\frac{1 - \overline{P}_f^\alpha(\Delta u_b; \varepsilon)}{1 - \overline{P}_b^\alpha(\Delta u_f; \varepsilon)} \right] \approx P_b^\alpha(\Delta u_f; \varepsilon) - \overline{P}_f^\alpha(\Delta u_b; \varepsilon), \quad (6)$$

which is the main result of this work. It is worth reiterating that this is valid for *any* choice of α and ε , including those choices that correspond with the well-known BAR and SOS estimators.³⁴ Taking the expectation of both sides shows that using an empirical estimate for $P_b^\alpha(\Delta u_f; \varepsilon)$ and $\overline{P}_f^\alpha(\Delta u_b; \varepsilon)$ yields an estimate of the bias that goes to zero no faster than $\sim 1/N$ (see [Appendix B](#)). Details on how to construct and evaluate estimators for finite data sets are given in [Sec. III B](#).

Equation (6) has some instructive properties. For example, the numerator and denominator are effectively the biases due to sampling from only one ensemble. In certain limits, the same bracketing behavior derived for FEP is encountered

$$\lim_{\Delta u_b \rightarrow +\infty} \Delta \hat{f} - \Delta f \approx +P_b^\alpha(\Delta u_f; \varepsilon), \quad (7a)$$

$$\lim_{\Delta u_f \rightarrow -\infty} \Delta \hat{f} - \Delta f \approx -\overline{P}_f^\alpha(\Delta u_b; \varepsilon). \quad (7b)$$

It is also readily apparent that the right-hand side of Eq. (6) vanishes not only when both $\overline{P}_f^\alpha(\Delta u_b; \varepsilon)$ and $P_b^\alpha(\Delta u_f; \varepsilon)$ are zero but also whenever $\overline{P}_f^\alpha(\Delta u_b; \varepsilon) = P_b^\alpha(\Delta u_f; \varepsilon)$ —the two bias contributions expressly cancel. Separately estimating the two quantities can thus be used to establish conservative bounds on the net bias.

Interestingly, the extremely common occurrence $\Delta u_f \leq \langle \Delta u \rangle_b$ leads to $P_b^\alpha(\Delta u_f; \varepsilon) \leq P_b(\Delta u_f)$ (see [Appendix C](#)). This implies that *any* choice of α (that meets the criteria so far) leads to a positive contribution to the bias that is very often less than the analogous value from forward FEP. Similarly, $\Delta u_b \geq \langle \Delta u \rangle_f$ leads to $\overline{P}_f^\alpha(\Delta u_b; \varepsilon) \leq \overline{P}_f(\Delta u_b)$. Nonetheless, because of bias cancellation, these special occurrences are unnecessary for the overall bias to be less than that from FEP.

Finally, Eq. (6) is quite different from other bias expressions in the literature. It rigorously utilizes data from both ensembles and makes no assumptions as to the underlying form of the PDFs.²⁰ Although there is some resemblance, it does not have the same bias

cancellation properties as a simple average of the FEP estimates [i.e., half the sum of Eqs. (2a) and (2b)]. Indeed, such an estimate is not effective because the two components often do not decrease with sample size at the same rate¹³ (more clever weighting likely will not help^{21,22}). With regard to BAR, Shirts and Pande²¹ derived an expression for the large-sample bias that is proportional to one-half the variance times a constant containing multiple derivatives of the log-likelihood function evaluated at Δf . This is informative but lacks a practical form for estimation. Finally, Lu, Singh, and Kofke¹³ actually did propose an approximate expression for the bias that can be made to look very similar to Eq. (6). However, that expression utilized an equality involving the product $p_f(\Delta u)p_b(\Delta u)$ and thus requires some additional manipulations before it can be employed as an estimator.

III. METHODS

A. Example calculations

Some bias calculations are presented here for simple, but realistic, multistage free energy calculations using BAR and FEP. The data come from alchemical simulations of 15 neutral amino acid side-chain analogs designed to compute their absolute solvation free energies. BAR and FEP were applied to successive pairs of alchemical simulations, and the net free energy, variance, and bias across all stages of the transformation were appropriately summed. Three independent trials were performed. The “observed” bias (a proxy for the exact expected bias) was calculated as the difference in free energies computed from a subsample and the average from the three complete data sets (a best guess for Δf). The “estimated” biases were evaluated with either Eq. (2) or (6) (see below for details). For each sample size, the results represent the mean of equal sized, equally spaced, and nonoverlapping blocks (e.g., 100 samples would yield the average of 6 blocks of size 15 with 10 samples discarded between blocks). Additional data were generated for a single-step ethane-to-ethane null transformation so that Δf is known to be exactly zero. This was repeated 500 times when extremely detailed statistics were needed.

All simulations were performed with NAMD 2.13,³⁵ and analysis was done with in-house Python tools that replicate and extend the functionality of ParseFEP.³⁶ The exact details of the simulations will be published elsewhere, and the data shown here are only intended to represent a diverse set of PDFs as might be found, for example, in ligand binding applications.

B. Practical estimation of the bias

When evaluating FEP bias, it is common practice to approximate the CDFs by first making a histogram estimate of the PDF and then summing over a subset of the bins. However, this approach can be sensitive to the choice of bin width. Worse yet, because the forward and backward PDFs tend to have very different shapes, it is difficult (or impossible) to choose a bin width that is simultaneously optimal for both. A simple alternative is to use the empirical CDFs by counting the number of data points below a particular value. This can be thought of as the average of a Heaviside function H [$H(0) = 1$], and thus, all of the necessary terms in Eqs. (2) and (6) can be evaluated directly from sample means. An estimate

of the forward and backward bias components in Eq. (6) can thus be computed as

$$\hat{P}_f^\alpha(\Delta u_b; \varepsilon) = \frac{\sum_{n=1}^{N_f} \alpha(\Delta u_n - \varepsilon) H(\Delta u_n - \Delta u_b)}{\sum_{n=1}^{N_f} \alpha(\Delta u_n - \varepsilon)}, \quad (8a)$$

$$\hat{P}_b^\alpha(\Delta u_f; \varepsilon) = \frac{\sum_{n=1}^{N_b} \alpha(\varepsilon - \Delta u_n) H(\Delta u_f - \Delta u_n)}{\sum_{n=1}^{N_b} \alpha(\varepsilon - \Delta u_n)}, \quad (8b)$$

where the sums run over the forward and backward samples, respectively. For BAR, one first estimates $\Delta \hat{f}$ and then takes $\varepsilon = \Delta \hat{f} - \ln \frac{N_f}{N_b}$, while for SOS $\varepsilon = 0$ and the form of α also changes.

Finally, it is worth considering the compromises to be made in selecting values for Δu_f and Δu_b . In principle, these can simply be chosen as the minimum and maximum of their respective samples. However, the integrity of the threshold ansatz also hinges on the accuracy of the observed distributions and these may not be reliable for data near the tail regions. This trade-off can be reduced to a single parameter by setting Δu_f equal to the q th percentile value of the forward data set and Δu_b to the $(100 - q)$ th percentile value of the backward data set [$q \in (0, 50)$]. Provided that the value is reasonably small (1–10, say), there does not appear to be a significant relationship between q and the quality of the bias estimate (Appendix D). A value of $q = 1$ is used for all calculations.

IV. RESULTS AND DISCUSSION

A. Bias cancellation and sample size invariance

Example bias calculations for multiple systems and sample sizes confirm the predicted trends. Most notably, for all systems studied here, the separate components of the BAR bias are always smaller than their FEP counterparts with the same sign (Fig. 2, top, dashed lines). This follows from the fact that the observed threshold values are smaller than the mean of the opposite distribution (see, for example, Fig. 1). Only the individual components obviously decay with sample size, and the net bias is instead flat and near zero. Bias cancellation thus appears to be rather efficient and has a fortuitous side effect of the variance minimization procedure in BAR. Both log terms in Eq. (6) vanish slightly slower than $\sim 1/N$ and with opposite signs (Appendix B). The signed bias thus oscillates near zero while the squared bias has the more expected $\sim 1/N^2$ behavior (Fig. 2, fit lines in the bottom panel). This is to be contrasted with the FEP biases which decay even more slowly ($\sim 1/N^{0.6}$ or worse).

With some caveats (discussed later), the estimated squared bias [computed from Eq. (6)] correlates relatively well with the observed squared bias (the squared difference between $\Delta \hat{f}$ and a best guess of Δf). This can be seen in the reasonable overlap of the square and circle data points in Fig. 2. The notable underestimation of some points appears to be an outlier due to the reference value being poorly determined for one particular molecule (i.e., the observed bias calculation itself possessed a high variance). The observed and estimated biases are also seen to vanish with essentially the same sample size dependence (prefactors of 172 ± 9 vs 162 ± 7 , respectively). Similar, but slightly worse, agreement is seen for estimates using SOS (data not shown).

FIG. 2. Bias estimates (top) for Δf computed over a broad set of molecules illustrate the expected trend, with BAR (circles) yielding considerably lower values than FEP (triangles and squares). The variance is generally the dominant contributor to the squared error (bottom). The reported sample size is the number of (presumed independent) data points from each simulation. Trend lines are nonlinear least squares fits to the power law $1/N^\alpha$, $\alpha \leq 1$.

B. Bias/variance trade-off

By construction, the expected variance reaches a minimum (which is dependent on the sample size) when ε is optimally chosen according to Eq. (1). However, this is *not* the same as minimizing the sample variance [see Eq. (10a) from the work of Bennett⁸] or the squared sample bias estimated from Eq. (6); the sample quantities are only minimized on average. Indeed, the variances and squared biases from a given sample may have completely different minima with respect to ε , and neither of these minima need correspond to $\Delta \hat{f} - \ln(N_f/N_b)$. Figure 3 illustrates this with representative data compared against the average of 500 repeated free energy calculations. This may initially appear paradoxical and somewhat counter to (arguably imprecise) statements from the literature. However, it is obviously true from the data here and does indeed make sense in light of the fact that multiple variance expressions for BAR have been suggested.^{14,16} All such expressions are correct in the large-sample limit but likely have different sample size dependences and/or prefactors. Unlike Bennett's variance expression, others may also not be valid for arbitrary ε .

A consequence of the energy shift dependence is that the correlation between systematic and statistical error is effectively random within a given sample. That is, since BAR will only jointly minimize both errors on average, there is no guarantee that either is minimized in practice. Furthermore, the BAR optimum might actually lie in between the sample minima such that improvement in one exacerbates the other (Fig. 3, top panel). However, it is not at all clear that minimizing the sample variance ought to be preferred in practice due to (1) the inherent error in the sample variance itself and (2) the accompanying bias that might be introduced. Nonetheless, investigating these trade-offs further could

FIG. 3. The squared error components vary as a function of the shift constant ϵ and, on average, attain minima near $\epsilon = \Delta f$ (top). The sample quantities are not, in general, minimized for a given data set (top, illustrative data shown as dots). The mean of multiple simulations shows how this pattern emerges (bottom).

lead to weakly biased estimators that may be preferable to BAR in practice.

C. Bias correction vs bias detection

It is worth distinguishing between two distinct purposes in estimating bias. The obvious inclination is to compute the bias for an estimator so that it can be corrected. This ultimately implies a new estimator, such as the classic Bessel correction for sample variance calculations. However, a corrected estimator can also have different (i.e., worse) variance properties, and so the value of employing the correction should be weighed against the size of the bias that it is meant to correct. An alternative approach is simply to detect the effective presence or absence of bias but otherwise leave the original estimator unchanged. The most prudent use of this work is likely along this latter route. To be sure, predicting the exact bias would be quite spectacular, as it would essentially suggest a new method of free energy estimation that outperforms the best methods to date. The accuracy of Eq. (6) does not appear to rise to this level.

The scatter of estimated vs observed biases indicates rather poor predictive power (Fig. 4). Indeed, the correct sign is only predicted 55%–65% of the time for the three subsample sizes shown. For the ethane-to-ethane transformation data, the distributions of both biases are normally distributed (albeit with different variances), but the distribution of their ratio suggests zero correlation (data not shown). Nonetheless, there is some improvement when considering the root mean square bias (i.e., the magnitude), although this still shows some degree of underestimation (by more than $0.2k_B T$ roughly 20%–40% of the time). Doubling the estimated bias, which is akin to taking a larger confidence interval with the variance, reduces this failure rate to less than 10% and often

FIG. 4. Plots of the estimated vs observed bias show a relatively low correlation when considering absolute bias (top). This is rather improved for the root mean square bias (bottom, taken over repeated subsamples of the given size). All energies are in $k_B T$ units, and dotted lines indicate an interval of ± 0.2 .

eliminates it entirely. Another conservative alternative not explored extensively here might be to take the larger magnitude from the two bounding estimates, although this would discount the effects of bias cancellation.

V. CONCLUSION

Any serious free energy calculation should also include estimates of its statistical and systematic errors. However, nearly all published calculations only supply variance estimates. This approach is frequently justified on the grounds that the bias vanishes in the large-sample limit, and thus, systematic errors are generally not examined any further. This work provides a clear and practical method for making bias estimates due to undersampling and permits a low-cost examination of this assertion. The method works with the widely popular BAR estimator as well as any of a broad class of free energy bridge estimators. Realistic examples show that variance does indeed tend to dominate for large sample sizes and the framework used here also gives insights into why this is so. Overall, the accuracy observed in test data suggests that the method is best when used to verify that the true bias is small. Future work might explore the usage of this framework to potentially introduce bias into new estimators in an effort to reduce variance. While such weakly biased estimators

have been long supposed, the bias defined here provides new ways to think about them.

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for the raw data used in this work.

ACKNOWLEDGMENTS

The author is grateful to Benoît Roux and Christophe Chipot for valuable discussions as well as the anonymous reviewers who offered several insights that improved the manuscript considerably. This work was supported by the National Institutes of Health (Grant No. P41-GM104601).

APPENDIX A: DETAILED DERIVATION OF BRIDGE SAMPLING BIAS

The full bounds of integration implied in Eq. (1) can be divided into those from Eq. (4) plus an integral over each of the tail regions

$$e^{\varepsilon-\Delta f} = \frac{\int_{-\infty}^{\Delta u_f} d\Delta u \alpha(\Delta u - \varepsilon) p_f(\Delta u) + \int_{\Delta u_f}^{\infty} d\Delta u \alpha(\Delta u - \varepsilon) p_f(\Delta u)}{\int_{-\infty}^{\Delta u_b} d\Delta u \alpha(\varepsilon - \Delta u) p_b(\Delta u) + \int_{\Delta u_b}^{\infty} d\Delta u \alpha(\varepsilon - \Delta u) p_b(\Delta u)} \quad (\text{A1})$$

[note the use of the identity $\exp(x)\alpha(x) = \alpha(-x)$]. Making further use of Eq. (4), the quantity in the denominator can be factored out of the numerator and vice versa. This immediately reduces to factors involving $\varepsilon - \Delta \hat{f}$,

$$e^{\varepsilon-\Delta f} = e^{\Delta \hat{f}-\varepsilon} \frac{\int_{-\infty}^{\Delta u_f} d\Delta u \alpha(\Delta u - \varepsilon) p_f(\Delta u)}{\int_{-\infty}^{\Delta u_b} d\Delta u \alpha(\varepsilon - \Delta u) p_b(\Delta u)} + e^{\varepsilon-\Delta \hat{f}} \frac{\int_{\Delta u_f}^{\infty} d\Delta u \alpha(\varepsilon - \Delta u) p_f(\Delta u)}{\int_{\Delta u_b}^{\infty} d\Delta u \alpha(\varepsilon - \Delta u) p_b(\Delta u)}. \quad (\text{A2})$$

The numerators of the remaining integrals can be transformed by first using the equality $p_f(\Delta u) = \exp(\Delta u - \Delta f) p_b(\Delta u)$ and then the same identity imposed on α (this is only used for brevity and can be undone later). The integrands then become

$$\begin{aligned} \alpha(\Delta u - \varepsilon) p_f(\Delta u) &= \alpha(\Delta u - \varepsilon) e^{\Delta u - \varepsilon} p_b(\Delta u) e^{\varepsilon - \Delta f} \\ &= \alpha(\varepsilon - \Delta u) p_b(\Delta u) e^{\Delta f - \varepsilon} \end{aligned}$$

and

$$\begin{aligned} \alpha(\varepsilon - \Delta u) p_b(\Delta u) &= \alpha(\varepsilon - \Delta u) e^{\varepsilon - \Delta u} p_f(\Delta u) e^{\Delta f - \varepsilon} \\ &= \alpha(\Delta u - \varepsilon) p_f(\Delta u) e^{\Delta f - \varepsilon}, \end{aligned}$$

which (up to a constant factor) are the same as the integrands in their denominator. Using the assumptions about normality and noting that $\Delta u_f < \Delta u_b$ permits the integral ratios to be written in terms of CDFs with the PDF definitions from the main text. Substituting these definitions and multiplying through by common factors leads to

$$e^{\Delta \hat{f} - \Delta f} = \frac{P_b^\alpha(\Delta u_f; \varepsilon) e^{\Delta \hat{f} - \Delta f} + 1}{1 + \bar{P}_f^\alpha(\Delta u_b; \varepsilon) e^{\Delta f - \Delta \hat{f}}}, \quad (\text{A3})$$

which can be solved in terms of $\Delta \hat{f} - \Delta f$ to give the result in the main text.

APPENDIX B: SAMPLE SIZE DEPENDENCE

The approximate sample size dependence of the bias can be determined by estimating Δu_f and Δu_b as sample minima and maxima. For independently drawn samples, these have well-defined distributions in terms of the PDF and CDF as follows:

$$p_b^{\min}(\Delta u_f; N_b) = N_b p_b(\Delta u_f) \bar{P}_b(\Delta u_f)^{N_b-1}, \quad (\text{B1a})$$

$$p_f^{\max}(\Delta u_b; N_f) = N_f p_f(\Delta u_b) P_f(\Delta u_b)^{N_f-1}. \quad (\text{B1b})$$

Rearranging and taking the logarithm of each side yields

$$\ln[1 - P_b(\Delta u_f)] = \frac{\ln\left[\frac{p_b^{\min}(\Delta u_f; N_b)}{N_b p_b(\Delta u_f)}\right]}{N_b - 1}, \quad (\text{B2a})$$

$$\ln[1 - \bar{P}_f(\Delta u_b)] = \frac{\ln\left[\frac{p_f^{\max}(\Delta u_b; N_f)}{N_f p_f(\Delta u_b)}\right]}{N_f - 1}, \quad (\text{B2b})$$

which, on the left-hand side, are essentially identical to the FEP bias expressions.

It is reasonable to approximate $p_b^{\min}(\Delta u_f)$ and $p_f^{\max}(\Delta u_b)$ as (small) constants with respect to sample size since the threshold ansatz assumes that Δu_f and Δu_b do not lie too far into the tails of the opposing distribution. In any event, it is improbable that these terms grow with sample size. With this approximation, the log sample size terms in the numerators mean that both FEP biases fall off no faster than $\sim 1/N$. However, substituting instead the new bridge estimator PDFs and combining to get Eq. (6) leads to the cancellation of the log terms. Assuming for simplicity that $N_f = N_b = N/2$ gives

$$\Delta \hat{f} - \Delta f = \frac{2 \ln\left[\frac{p_f^{\alpha, \max}(\Delta u_b) p_b^\alpha(\Delta u_f)}{p_f^\alpha(\Delta u_b) p_b^{\alpha, \min}(\Delta u_f)}\right]}{N - 2}, \quad (\text{B3})$$

where the N and ε dependence of the PDFs has been suppressed. This is $\sim 1/N$ as is usually expected of the bias. The PDF ratios in the logarithm lead to oscillations in the sign since, in most instances, $p_f^{\alpha, \max}(\Delta u_b) > p_f^\alpha(\Delta u_b)$ and $p_b^{\alpha, \min}(\Delta u_f) > p_b^\alpha(\Delta u_f)$. The sign of the bias is dependent on which of these inequalities is more dramatic.

APPENDIX C: BRIDGE ESTIMATE BIAS RELATIVE TO FEP

The bridge sampling bias estimate is composed of two contributions of opposite signs which cancel and improve the net bias. Nonetheless, the two individual components are themselves almost always less than the corresponding FEP bias. This can be demonstrated by comparing the magnitudes of P_b^α and $\bar{P}_f^\alpha$ relative to P_b and $\bar{P}_f$. We shall focus here on the backward CDFs and simply outline the similar arguments for the forward CDFs.

In order to show that $P_b^\alpha(\Delta u_f; \varepsilon) \leq P_b(\Delta u_f)$, it suffices to show that $p_b^\alpha(\Delta u; \varepsilon) < p_b(\Delta u)$ for $\Delta u \leq \Delta u_f$. Rearranging the definition of $p_b^\alpha(\Delta u; \varepsilon)$ from Eq. (5), one has

$$\frac{p_b^\alpha(\Delta u; \varepsilon)}{p_b(\Delta u)} = \frac{\alpha(\varepsilon - \Delta u)}{\int_{-\infty}^{\Delta u_b} d\Delta u' \alpha(\varepsilon - \Delta u') p_b(\Delta u')}. \quad (\text{C1})$$

Applying Jensen's inequality to the denominator (this again requires the assertion that the PDF is correctly normalized on the integration interval), consequently,

$$\frac{P_b^\alpha(\Delta u; \varepsilon)}{p_b(\Delta u)} \leq \frac{\alpha(\varepsilon - \Delta u)}{\alpha(\varepsilon - \langle \Delta u \rangle_b)}. \quad (\text{C2})$$

Since $\alpha(-x)$ is monotonically increasing, the right-hand side is always less than or equal to unity *provided that* $\Delta u \leq \langle \Delta u \rangle_b$. This is not a guaranteed outcome, but it can be easily verified for a given sample. A similar argument follows for the forward distributions when $\Delta u_b \geq \langle \Delta u \rangle_f$ since $\alpha(x)$ is monotonically decreasing. The results can be summarized as

$$P_b^\alpha(\Delta u_f; \varepsilon) \leq P_b(\Delta u_f) \quad \text{given that} \quad \Delta u_f \leq \langle \Delta u \rangle_b, \quad (\text{C3a})$$

$$\overline{P}_f^\alpha(\Delta u_b; \varepsilon) \leq \overline{P}_f(\Delta u_b) \quad \text{given that} \quad \Delta u_b \geq \langle \Delta u \rangle_f, \quad (\text{C3b})$$

which give clear conditions on when the bias components are smaller than those from FEP.

APPENDIX D: DETERMINATION OF THRESHOLD VALUES

As described in the main text, the threshold values can be chosen quite generally as the q th and $(100 - q)$ th percentile values

FIG. 5. Plots of the estimated vs observed bias do not seem to show a discernible pattern as a function of the percentile cutoff, q . This seems to hold for a range of sample sizes. All energies are in $k_B T$ units, and dotted lines indicate an interval of ± 0.2 .

of each data set. However, for small data sets, extracting a specific value from the observed data can lead to repeat results (e.g., the 1st–10th percentile of a 10 element set would all yield the first sorted element). For more flexibility, we consider a set of (sorted) data points $\{\Delta u_n\}$, $n = 1, 2, \dots, N$ and compute the q th percentile Δu_q as a weighted combination of two elements. That is, elements $i = \lfloor q(N - 1)/100 + 1 \rfloor$ and $j = \min(i + 1, N)$ from the set yield the estimate $\Delta u_q = (1 - w)\Delta u_i + w\Delta u_j$, where $w \equiv [q(N - 1)/100 + 1 - i]$. This is the default algorithm implemented, for example, in NumPy.

Clearly, q must be chosen to be relatively small, as the goal is to compromise between a reasonable estimate of the sample minima/maxima and an accurate representation of the density. However, within this constraint, the data shown here do not seem to suggest any clear correlation between the value of q and the quality of the bias estimate (Fig. 5). That is, increasing q leads to neither a clear increase nor decrease in the accuracy of the bias estimate.

One minor concern for small data sets is that a small value of q may lead to the entirety of the opposing data set being in the tail region. In this case, $\overline{P}_f^\alpha$ and/or P_b^α would evaluate to one and the bias estimate diverges. This can of course also happen if the underlying distributions only weakly overlap and thus may indicate a deeper pathology in the simulation setup. Larger values of q may recover a finite estimate of the bias in these cases, but it seems unlikely that this would be considerably more informative.

REFERENCES

- M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1987).
- D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic Press, San Diego, CA, 2002).
- Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in Chemical Physics Vol. 86, edited by C. Chipot and A. Pohorille (Springer, New York, 2007).
- J. G. Kirkwood, "Statistical mechanics of fluid mixtures," *J. Chem. Phys.* **3**, 300–313 (1935).
- R. W. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," *J. Chem. Phys.* **22**, 1420–1426 (1954).
- B. Widom, "Some topics in the theory of fluids," *J. Chem. Phys.* **39**, 2808–2812 (1963).
- J. P. Valleau and D. N. Card, "Monte Carlo estimation of the free energy by multistage sampling," *J. Chem. Phys.* **57**, 5457–5462 (1972).
- C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," *J. Comput. Phys.* **22**, 245–268 (1976).
- G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," *J. Comput. Phys.* **23**, 187–199 (1977).
- K. S. Shing and K. E. Gubbins, "The chemical potential in dense fluids and fluid mixtures via computer simulations," *Mol. Phys.* **46**, 1109–1128 (1982).
- A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis," *Phys. Rev. Lett.* **63**, 1195–1198 (1989).
- S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comput. Chem.* **13**, 1011–1021 (1992).
- N. Lu, J. K. Singh, and D. A. Kofke, "Appropriate methods to combine forward and reverse free-energy perturbation averages," *J. Chem. Phys.* **118**, 2977–2984 (2003).
- M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J. Chem. Phys.* **129**, 124105 (2008).

- ¹⁵A. de Ruiter and C. Oostenbrink, "Efficient and accurate free energy calculations on trypsin inhibitors," *J. Chem. Theory Comput.* **8**, 3686–3695 (2012).
- ¹⁶Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy, "Theory of binless multi-state free energy estimation with applications to protein-ligand binding," *J. Chem. Phys.* **136**, 144102 (2012).
- ¹⁷C. Jarzynski, "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach," *Phys. Rev. E* **56**, 5018–5035 (1997).
- ¹⁸M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, "Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods," *Phys. Rev. Lett.* **91**, 140601 (2003).
- ¹⁹G. Hummer, "Fast-growth thermodynamic integration: Error and efficiency analysis," *J. Chem. Phys.* **114**, 7330–7337 (2001).
- ²⁰J. Gore, F. Ritort, and C. Bustamante, "Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements," *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12564–12569 (2003).
- ²¹M. R. Shirts and V. S. Pande, "Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration," *J. Chem. Phys.* **122**, 144107 (2005).
- ²²A. Pohorille, C. Jarzynski, and C. Chipot, "Good practices in free-energy calculations," *J. Phys. Chem. B* **114**, 10235–10253 (2010).
- ²³H. Paliwal and M. R. Shirts, "A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods," *J. Chem. Theory Comput.* **7**, 4115–4134 (2011).
- ²⁴S. Bruckner and S. Boresch, "Efficiency of alchemical free energy simulations. I. A practical comparison of the exponential formula, thermodynamic integration, and Bennett's acceptance ratio method," *J. Comput. Chem.* **32**, 1303–1319 (2011).
- ²⁵S. Bruckner and S. Boresch, "Efficiency of alchemical free energy simulations. II. Improvements for thermodynamic integration," *J. Comput. Chem.* **32**, 1320–1333 (2011).
- ²⁶Z. Tan, "On a likelihood approach for Monte Carlo integration," *J. Am. Stat. Assoc.* **99**, 1027–1036 (2004).
- ²⁷D. M. Zuckerman and T. B. Woolf, "Theory of a systematic computational error in free energy differences," *Phys. Rev. Lett.* **89**, 180602 (2002).
- ²⁸N. Lu and D. A. Kofke, "Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling," *J. Chem. Phys.* **114**, 7303–7311 (2001).
- ²⁹N. Lu and D. A. Kofke, "Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics," *J. Chem. Phys.* **115**, 6866–6875 (2001).
- ³⁰N. Lu and T. B. Woolf, "Understanding and improving free energy calculations in molecular simulations: Error analysis and reduction methods," in *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in Chemical Physics Vol. 86, edited by C. Chipot and A. Pohorille (Springer, New York, 2007), Chap. 6, pp. 199–247.
- ³¹G. E. Crooks, "Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems," *J. Stat. Phys.* **90**, 1481–1487 (1998).
- ³²G. E. Crooks, "Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences," *Phys. Rev. E* **60**, 2721–2726 (1999).
- ³³As noted by Lu and Woolf,³⁰ there is actually an additional approximation being made here. The data cannot be exactly drawn from the true densities and instead come from the sample densities $\hat{p}_f(\Delta u) \equiv p_f(\Delta u) / \int_{\Delta u_f}^{\infty} d\Delta u' p_f(\Delta u')$ and $\hat{p}_b(\Delta u) \equiv p_b(\Delta u) / \int_{-\infty}^{\Delta u_b} d\Delta u' p_b(\Delta u')$. By assuming the denominators are roughly unity, the densities effectively become interchangeable.
- ³⁴To be completely general, the definition of $p_b^a(\Delta u; \epsilon)$ in Eq. (5) should be modified with the substitution $a(\epsilon - \Delta u) = \exp(\Delta u - \epsilon)\alpha(\Delta u - \epsilon)$, but this identity is appealingly compact and holds commonly enough that it is used here regardless.
- ³⁵J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.* **26**, 1781–1802 (2005).
- ³⁶P. Liu, F. Dehez, W. Cai, and C. Chipot, "A toolkit for the analysis of free-energy perturbation calculations," *J. Chem. Theory Comput.* **8**, 2606–2616 (2012).