Small-sample limit of the Bennett acceptance ratio method and the variationally derived intermediates

Martin Reinhardt and Helmut Grubmüller*
Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

(Received 30 August 2021; accepted 28 October 2021; published 24 November 2021)

Free energy calculations based on atomistic Hamiltonians provide microscopic insight into the thermodynamic driving forces of biophysical or condensed matter systems. Many approaches use intermediate Hamiltonians interpolating between the two states for which the free energy difference is calculated. The Bennett acceptance ratio (BAR) and variationally derived intermediates (VI) methods are optimal estimator and intermediate states in that the mean-squared error of free energy calculations based on independent sampling is minimized. However, BAR and VI have been derived based on several approximations that do not hold for very few sample points. Analyzing one-dimensional test systems, we show that in such cases BAR and VI are suboptimal and that established uncertainty estimates are inaccurate. Whereas for VI to become optimal, less than seven samples per state suffice in all cases; for BAR the required number increases unboundedly with decreasing configuration space densities overlap of the end states. We show that for BAR, the required number of samples is related to the overlap through an inverse power law. Because this relation seems to hold universally and almost independent of other system properties, these findings can guide the proper choice of estimators for free energy calculations.

DOI: 10.1103/PhysRevE.104.054133

I. INTRODUCTION

Free energy differences provide detailed insights into the molecular driving forces of biophysical processes, and their accurate calculation is crucial for their successful application, e.g., in pharmaceutical ligand design or material science [1–7]. To calculate the free energy difference between, e.g., two potential drug molecules bound to a receptor, alchemical equilibrium techniques [8] based on simulations with atomistic Hamiltonians are among the most widely used methods. Aside from the two states of interest, these techniques conduct sampling from intermediate states whose Hamiltonians are constructed from those of the end states. The stepwise summation of the individual differences then yields the total free energy difference.

Two choices have to be made that critically affect the accuracy of free energy calculations: First, the choice of the estimator that is used to evaluate the free energy differences between the individual states. Whereas a number of estimators exist that have practical advantages in different situations [8–10], it has been shown that between two states the Bennett acceptance ratio (BAR) method [11] minimizes not only the variance, but also the mean-squared error (MSE) [12]. Remarkably, as will be revisited in the theory section, the Zwanzig formula [9] yields identical MSEs if applied together with an optimally chosen virtual intermediate state in which no sampling is conducted [10,12]. For BAR, the variance and the bias have been extensively analyzed [10,13–16]. As the MSE can be decomposed into variance plus the squared bias and, therefore, accounts for both the variance and the bias, we will focus our analysis in this paper on the MSE. Furthermore, from an application perspective, the MSE is the relevant quantity.

The second choice concerns the functional form of the intermediate states, i.e., how these are constructed from the two end state Hamiltonians. Apart from the conventionally used linear interpolation intermediates, various functional forms have been suggested [17–20] with a particular focus on appearing or vanishing particles in solution [21–25]. In general, when using the Zwanzig formula or BAR as an estimator, and assuming independent samples, the variationally derived intermediates (VI) [12,26,27] have been shown to yield the optimal MSE among all possible functional forms of intermediate states.

However, both BAR and VI have been derived using approximations that strictly hold only for large sample numbers. This question becomes particularly urgent for free energy calculations of large systems or when using quantum mechanics based methods [28–31], which are computationally demanding and, therefore, provide limited sampling. Furthermore, sample points derived from atomistic simulations are time correlated such that the effective number of independent sample points is often orders of magnitude smaller than the number of configurations obtained from a simulation. We, therefore, will analyze how the accuracy of BAR and VI depends on sample

*hgrubmu@gwdg.de

size and show how the obtained relations provide guidance on their proper use.

II. THEORY

Several different derivations of BAR have been published [11,32,33], resting on different assumptions. Here we recapitulate the one with the least restrictive assumptions that also highlights the unexpected relation between estimators and intermediate states [12]. The generalization of this relation to N intermediate states has been used to derive VI. Both approaches rest on the Zwanzig formula [9]. Accordingly, the free energy difference between states A and B with Hamiltonians $H_A(\mathbf{x})$ and $H_B(\mathbf{x})$, respectively, is given by

$$\Delta G_{A,B} = -\ln\langle e^{-[H_B(\mathbf{x})-H_A(\mathbf{x})]}\rangle_A \ , \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{3M}$ denotes the position of all M particles of the simulation system. Only sample points from state A are used, where $\langle\,\rangle_A$ denotes the ensemble average. For ease of notation, all energies are expressed in units of k_BT.

In the following, the free energy estimate governed by Hamiltonian $H_A(\mathbf{x})$ that is obtained when the ensemble average in Eq. (1) is calculated from a finite sample of size n will be denoted by $\Delta G^{(n)}_{A\to B}$, whereas $\Delta G_{A,B}$ denotes the exact free energy difference. For statistically independent samples, the MSE of the free energy calculated via Eq. (1) reads [12]

$$\mathrm{MSE}(\Delta G^{(n)}_{A\to B}) = \mathbb{E}[(\Delta G_{A,B} - \Delta G^{(n)}_{A\to B})^2] \tag{2}$$

$$= \frac{1}{n}\left(\int \frac{[p_B(\mathbf{x})]^2}{p_A(\mathbf{x})}dx - 1\right), \tag{3}$$

where $p_A(\mathbf{x}) = e^{-H_A(\mathbf{x})}/Z_A$ and $p_B(\mathbf{x}) = e^{-H_B(\mathbf{x})}/Z_B$ denote the configuration space densities and Z_A and Z_B denote the partition functions of the respective end states.

Importantly, the derivation of the MSE of the Zwanzig formula Eq. (3) and, therefore, also the optimization thereof leading to BAR and VI, is based on approximations. As a prior step, we consider the Hamiltonian $H_B(\mathbf{x}) - C$, i.e., the Hamiltonian of end state B shifted by a constant C. Using this Hamiltonian with the Zwanzig formula, Eq. (1), the free energy difference between A and B is calculated as

$$\Delta G_{A,B} = -\ln\langle e^{-[H_B(\mathbf{x})-C-H_A(\mathbf{x})]}\rangle_A + C \ . \tag{4}$$

We now denote the sample-based average from Eq. (4) as

$$y^{(n)}(C) = \frac{1}{n}\sum_{i=1}^{n} e^{-[H_B(\mathbf{x}_i)-C-H_A(\mathbf{x}_i)]}, \tag{5}$$

and the exact ensemble average as

$$y(C) = \int p_A(\mathbf{x})dx\, e^{-[H_B(\mathbf{x})-C-H_A(\mathbf{x})]} \ . \tag{6}$$

For large n, using $C \approx \Delta G_{A,B}$ implies $y^{(n)}(C) \approx y(C) \approx 1$. After expanding the MSE, Eq. (2) (for the full derivation, see Ref. [12]), the expectation value of the estimate based on finite sampling,

$$\mathbb{E}\big[\Delta G^{(n)}_{A\to B}\big] = -\int p_A(\mathbf{x}_1)dx_1 \cdots \int p_A(\mathbf{x}_n)dx_n \ln[y^{(n)}(C)] + C, \tag{7}$$

and its square,

$$\mathbb{E}\big[\big(\Delta G^{(n)}_{A\to B}\big)^2\big] = -\int p_A(\mathbf{x}_1)dx_1 \cdots \int p_A(\mathbf{x}_n)dx_n \times \{\ln[y^{(n)}(C)] + C\}^2 \tag{8}$$

are approximated by using the first-order series expansion of the logarithm $\ln[y^{(n)}(C)] \approx y^{(n)}(C) - 1$ around $y^{(n)}(C) = 1$. Along similar lines, the exact difference and its square are approximated as $\Delta G_{A,B} = -\ln[y(C)] + C \approx -y(C) + 1 + C$ and $(\Delta G_{A,B})^2 = \{-\ln[y(C)] + C\}^2 \approx [-y(C) + 1 + C]^2$ around $y(C) = 1$.

Critically, for small n the averages $y^{(n)}(C)$ and $y(C)$ generally differ, and, therefore, C cannot be chosen such that both are approximately one. If, as in practice, C is evaluated based on the acquired samples such that $y^{(n)}(C) = 1$, then $y(C)$ differs from one and, consequently, the first-order series expansion of $y(C)$ becomes inaccurate. If $y^{(n)}(C)$ and $y(C)$ differ by, e.g, less than 10%, then the relative error of this approximation of the logarithm remains below 5%. However, for larger differences, the neglected higher-order terms will contribute markedly. A similar effect is caused by small configuration space density overlaps of the end states: Due to wider distributions of the exponentially weighted differences $H_B(\mathbf{x}) - H_A(\mathbf{x})$, the variance of the sample-based averages $y^{(n)}(C)$ will increase and, therefore, also the average absolute deviations from $y(C)$.

In the next step, Fig. 1(a) shows how an intermediate state I is used to derive the BAR formula via $\Delta G^{(n)}_{A\rightleftharpoons B} = \Delta G^{(n)}_{A\to I} - \Delta G^{(n)}_{B\to I}$. We refer to I as a *virtual* intermediate because it only serves as an end state for the Zwanzig formula without actually being used for sampling. The derivation based on the above approximations [12] yielded an additive MSE in this case, i.e., the MSE of the total estimate is

$$\mathrm{MSE}(\Delta G^{(n)}_{A\rightleftharpoons B}) = \mathrm{MSE}(\Delta G^{(n)}_{A\to I}) + \mathrm{MSE}(\Delta G^{(n)}_{B\to I}). \tag{9}$$

For easier notation, we assume that the same number of samples n is available for the two end states. Minimizing Eq. (9) through a variational approach leads to the Hamiltonian of the optimal virtual intermediate [12],

$$H_I(\mathbf{x}) = \ln(e^{H_A(\mathbf{x})} + e^{H_B(\mathbf{x})-C}), \tag{10}$$

where the MSE is minimal if $C = \Delta G_{A,B}$ and approaches that minimum as C approaches $\Delta G_{A,B}$. Figure 1(b) shows this virtual intermediate state as a black dashed line for a one-dimensional example where one of the two end Hamiltonians is harmonic (red), and the other is quartic (blue).

Let us compare the result using $\Delta G^{(n)}_{A\rightleftharpoons B} = \Delta G^{(n)}_{A\to I} - \Delta G^{(n)}_{B\to I}$ with intermediate Eq. (10) to the original approach by Bennett [11],

$$\Delta G^{(n)}_{A\rightleftharpoons B} = \ln\frac{\langle w[H_A(\mathbf{x}), H_B(\mathbf{x})]e^{-H_A(\mathbf{x})}\rangle_B}{\langle w[H_A(\mathbf{x}), H_B(\mathbf{x})]e^{-H_B(\mathbf{x})}\rangle_A}, \tag{11}$$

which uses a suitably chosen weight function $w[H_A(\mathbf{x}), H_B(\mathbf{x})]$. Bennett optimized the weighting function with respect to the variance, which yields the widely used

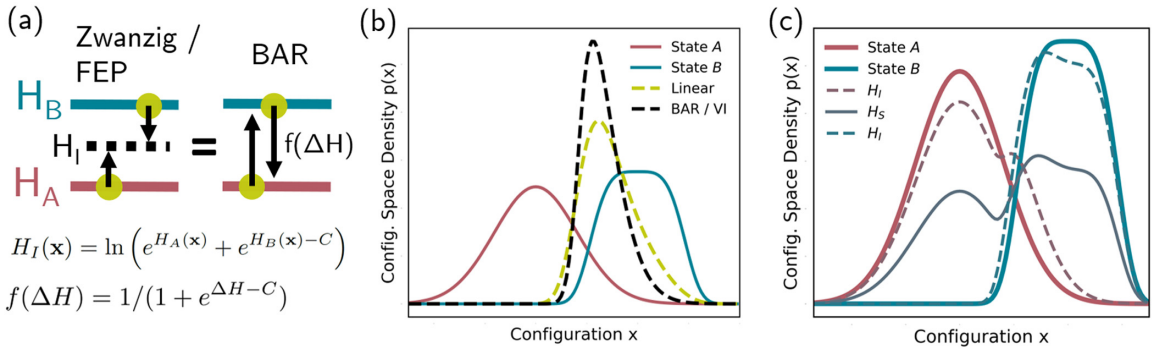

FIG. 1. (a) Two schemes of free energy estimators. Left: Using the Zwanzig formula to calculate the free energy difference from the two end states to a virtual intermediate state in which no sampling is conducted. Right: Using BAR where a weighting factor is applied to the difference in Hamiltonians. The two schemes are identical if the expressions shown beneath the schemes are used for the Hamiltonian of the virtual intermediate and the weighting function of BAR. (b) Configuration space densities of the virtual intermediate states corresponding to the linear estimator (green dashed line) and BAR (black dashed line). The densities of the harmonic end state $H_A(\mathbf{x}) = ax^2$ and the quartic end state $H_B(\mathbf{x}) = b(x - x_0)^4$ are shown in red and blue, respectively. (c) VI. States in which sampling is conducted are indicated through solid lines, whereas virtual intermediates are indicated through dashed lines.

BAR result,

$$\Delta G^{(n)}_{A,B} - C = \ln \frac{\langle f[H_A(\mathbf{x}) - H_B(\mathbf{x}) - C]\rangle_B}{\langle f[H_B(\mathbf{x}) - H_A(\mathbf{x}) + C]\rangle_A}, \tag{12}$$

where $f(x) = 1/(1+e^x)$ is the Fermi function and $C \approx \Delta G_{A,B}$ has to be determined iteratively.

From Eq. (11) and $\Delta G^{(n)}_{A\rightleftharpoons B} = \Delta G^{(n)}_{A\to I} - \Delta G^{(n)}_{B\to I}$ with Eq. (1) follows that the two approaches are equivalent if the weighting function relates to the Hamiltonian of the virtual intermediate state through

$$w[H_A(\mathbf{x}), H_B(\mathbf{x})] = e^{-H_I(\mathbf{x})+H_A(\mathbf{x})+H_B(\mathbf{x})}. \tag{13}$$

Therefore, any Hamiltonian of a virtual intermediate state corresponds to a weighting function.

The variance of BAR [11] is given by

$$\mathrm{Var}(\Delta G^{(n)}_{A,B}) = \frac{2}{n}[\Omega^{-1} - 1], \tag{14}$$

$$\Omega = \int dx \frac{2p_A(\mathbf{x})p_B(\mathbf{x})}{p_A(\mathbf{x}) + p_B(\mathbf{x})}, \tag{15}$$

where Ω can be interpreted as an overlap measure. Within the limits of the approximations discussed above, Bennett's variance Eq. (14) equals the MSE Eq. (3) of using Zwanzig in two steps as shown in Appendix A.

This link between BAR and VI Eq. (13) allows creating different estimators and transforming them between the formalism of using an intermediate state or a weighting function. Here, we will apply this result and compare BAR to the estimator that uses $H_I(\mathbf{x}) = \frac{1}{2}[H_A(\mathbf{x}) + H_B(\mathbf{x})]$ as the virtual intermediate state. Because $H_I(\mathbf{x})$ is a linear interpolation, we will refer to the resulting estimator as the "linear estimator," also known as the simple overlap sampling method [34,35]. The resulting configuration space density is shown by the green dashed line in Fig. 1(b). As shown in Appendix B, our MSE for the Zwanzig formula Eq. (3) yields the MSE for the linear estimator,

$$\mathrm{MSE}(\Delta G^{(n)}_{A,B}) = \frac{2}{n}\left[\left(\int p_A(\mathbf{x})^{1/2} p_B(\mathbf{x})^{1/2} dx\right)^{-2} - 1\right]. \tag{16}$$

The term in the round parentheses of Eq. (16) can be interpreted as an overlap measure, different from above, which equals one for two identical configuration space densities, and zero for disjunct supports.

Next, any number of optimal intermediate states can be derived by extending Eq. (9) with the MSEs of additional steps. Here, we focus our analysis on only one intermediate state S for sampling, i.e., calculations of the form $A \to I \leftarrow S \to I \leftarrow B$. The optimization with variational calculus with respect to all intermediate Hamiltonians yields the VI. These consist of, first, Eq. (10) (the BAR equivalent) as the optimal Hamiltonian of the virtual intermediates and second, the optimal sampling Hamiltonian $H_S(\mathbf{x})$, which is determined through the solution of

$$H_S(\mathbf{x}) = -\frac{1}{2}\ln\left[\left(e^{H_A(\mathbf{x})}\frac{Z_A}{Z_S} + e^{H_S(\mathbf{x})}\right)^{-2} + \left(e^{H_B(\mathbf{x})}\frac{Z_B}{Z_S} + e^{H_S(\mathbf{x})}\right)^{-2}\right]. \tag{17}$$

The initially unknown ratios of the partition sums are determined iteratively, similar to the constant C for BAR. The converged VI states for the harmonic and quartic end states are shown in Fig. 1(c). For molecular systems, such as the electrostatic decoupling of butanol or nitrocyclohexane [12,27], a sufficiently accurate estimate of the partition sum ratios such that VI yields better MSEs than conventional intermediates has been obtained within a few percent of the overall simulation time. Whether this holds true for complex molecular systems with large time correlations has, however, not been validated as of now. In order to disentangle such effects from the ones resulting from inaccurate approximations in the derivation, we here focus on cases with perfectly independent sample points.

To summarize, for small n, BAR and VI result from the accurate optimization of an inaccurate MSE. Naturally, this does not ensure that better estimators and intermediate sampling states exist, which is, therefore, the subject of our test simulations.

III. METHODS

In the first step, we assess the MSEs of different estimators. To this aim, we consider the one-dimensional system with end states consisting of a harmonic and a quartic Hamiltonian as shown in Fig. 1(b). Based on n sample points drawn from the configuration space density of A and B, the free energy estimate $\Delta G^{(n)}_{A\rightleftharpoons B}$ is obtained and compared to the exact difference $\Delta G_{A,B}$. Rejection sampling is used to obtain uncorrelated sample points. The MSE Eq. (2) is then calculated by averaging over 10^6 of such realizations. We use $n = 1,\ 20$, and 1000 sample points per end state. For each n, we consider 82 different setups for which the potential of end state B is moved horizontally away from A by varying x_0, thereby considering a range of overlap Ω, which is obtained through numerical integration of Eq. (15).

With this procedure, we compare three variants: To separate the effects of an inaccurate estimate of C, first, BAR is used where C has been set to the (in practice unknown) exact free energy difference. Second, using BAR, where C is iteratively determined based on the sample set as performed in practice. Third, the linear estimator.

In the second step, aside from sampling in the end states, sampling is also conducted in one intermediate state S and a similar procedure as above is used to evaluate the MSEs of different Hamiltonians $H_S(\mathbf{x})$. Separate sample sets in S are used to evaluate the free energy differences to either end state as using the same sample set would introduce correlations between the two stepwise free energy estimates that would require a different analytic approach as the one described above [26]. Again, three variants are compared: First, VI, i.e., Eqs. (10) and (17). For simplicity, only exact estimates for C and the ratios of the partition sums are considered. Second, as a comparison, two variants with a linearly interpolated sampling Hamiltonian: One using the linear estimator, and another one using BAR to evaluate the stepwise free energy difference. Again, the procedure was conducted for $n = 1,\ 20$, and 1000 sample points per sample set.

IV. RESULTS AND DISCUSSION

The MSEs of the three estimator variants are shown in Figs. 2(a)–2(c) for different configuration space density overlaps Ω between the harmonic and the quartic end state. The panels show this relation for different sample sizes n. As can be seen, for $n = 1$ both variants of BAR (blue and green) are suboptimal for all Ω as they yield a worse (larger) MSE than the linear estimator (yellow). For $n = 20$, it depends on Ω whether BAR is suboptimal. Here, a turning point exists, i.e., the linear estimator is only better for approximately $\Omega < 10^{-1}$, whereas both BAR variants yield better MSEs for the larger Ω. For $n = 1000$, this turning point shifts towards smaller Ω. Here, the BAR variants perform better for around $\Omega > 10^{-3}$. Note that as the end states are different in form, the largest achievable overlap is $\Omega = 0.935$, and, therefore, no MSE of zero can be seen in Figs. 2(a)–2(c), which would be expected for $\Omega = 1$.

Unexpectedly, whereas for most n's and Ω's both BAR variants have very similar MSEs, the one in blue where $C = \Delta G_{A,B}$ (i.e., the exact free energy difference) was used yields slightly worse MSEs than the variant that uses a sample-based estimate of C (green). This finding is in contrast to the widespread belief that an estimation for C that deviates from $\Delta G_{A,B}$ is a major contribution to the inaccuracy of BAR. The reason for this behavior lies in the first-order series expansions of $\ln y(C)$ and $\ln y^{(n)}(C)$ as shown in the context of Eqs. (7) and (8) in the theory section. For small n, $y^{(n)}(C)$, and $y(C)$ differ, and C can, therefore, not be chosen such that the requirement is met that both are close to one. As a consequence, even if $C = \Delta G_{A,B}$ such that $y(C) = 1$, then the first-order series expansion of $\ln y^{(n)}(C)$ becomes inaccurate, and the same holds true for the subsequent derivation of BAR.

The dashed lines in Figs. 2(a)–2(c) show the predicted MSEs for BAR, i.e., Eq. (14), whereas the dotted lines show the ones of the linear estimator Eq. (16). As can be seen from Fig. 2(a), for $n = 1$ the prediction completely underestimates the actual MSEs. Furthermore, BAR is predicted to have a better (smaller) MSE than the linear estimator which is, however, not the case for the results of the test simulations. For $n = 20$, the MSEs start to agree for large Ω but still deviate substantially for small Ω. For BAR with $n = 1000$, the MSEs agree well for most Ω's. For the linear estimator, the prediction is still mostly only accurate for large Ω. Interestingly, unlike at $n = 1$, Eq. (16) predicts a MSE that is worse than the one from the test simulations for $n = 1000$. These results show that BAR is only optimal in cases where the predicted MSE is close to the actual one. In cases where the predicted MSE is inaccurate, BAR as the optimization thereof becomes suboptimal.

For BAR, the discrepancy between the predicted MSEs and the actual ones also explains the common experience of users of free energy calculations that the error is often largely underestimated. Naturally, for atomistic simulations factors that violate the assumption of independent sample points, such as time correlations or starting configurations of several states that all remain close to the initial structure contribute to an underestimated uncertainty. However, our paper shows that even in the absence of all of these factors and for perfectly independent samples, the error is largely underestimated for small n due to the approximations in the derivation discussed in this paper. For example, for $n = 20$ independent sample points, an overlap of $\Omega = 0.1$, which is not uncommon, already leads to an actual uncertainty that is almost ten times worse than predicted by the uncertainty estimate of Bennett [11], i.e., Eq. (14).

As the turning point Ω above which BAR becomes optimal varies with n, the question arises for the relation between the required n for different Ω's and how this relation compares for different systems. Therefore, in the next step we test how many sample points are required for BAR to achieve a better MSE than the linear estimator, depending on the configuration space density. To this aim, the first variant is used (C exact). Starting with $n = 1$, the MSEs of both BAR and the linear estimator are calculated, and n is gradually increased until the turning point is found. In addition to the setup consisting of end states with a harmonic and a quartic Hamiltonian, three other diverse systems are considered. The configuration space densities $p_A(\mathbf{x})$ and $p_B(\mathbf{x})$ of their end states are shown in red and blue, respectively, in Fig. 2(d). Again, for each system

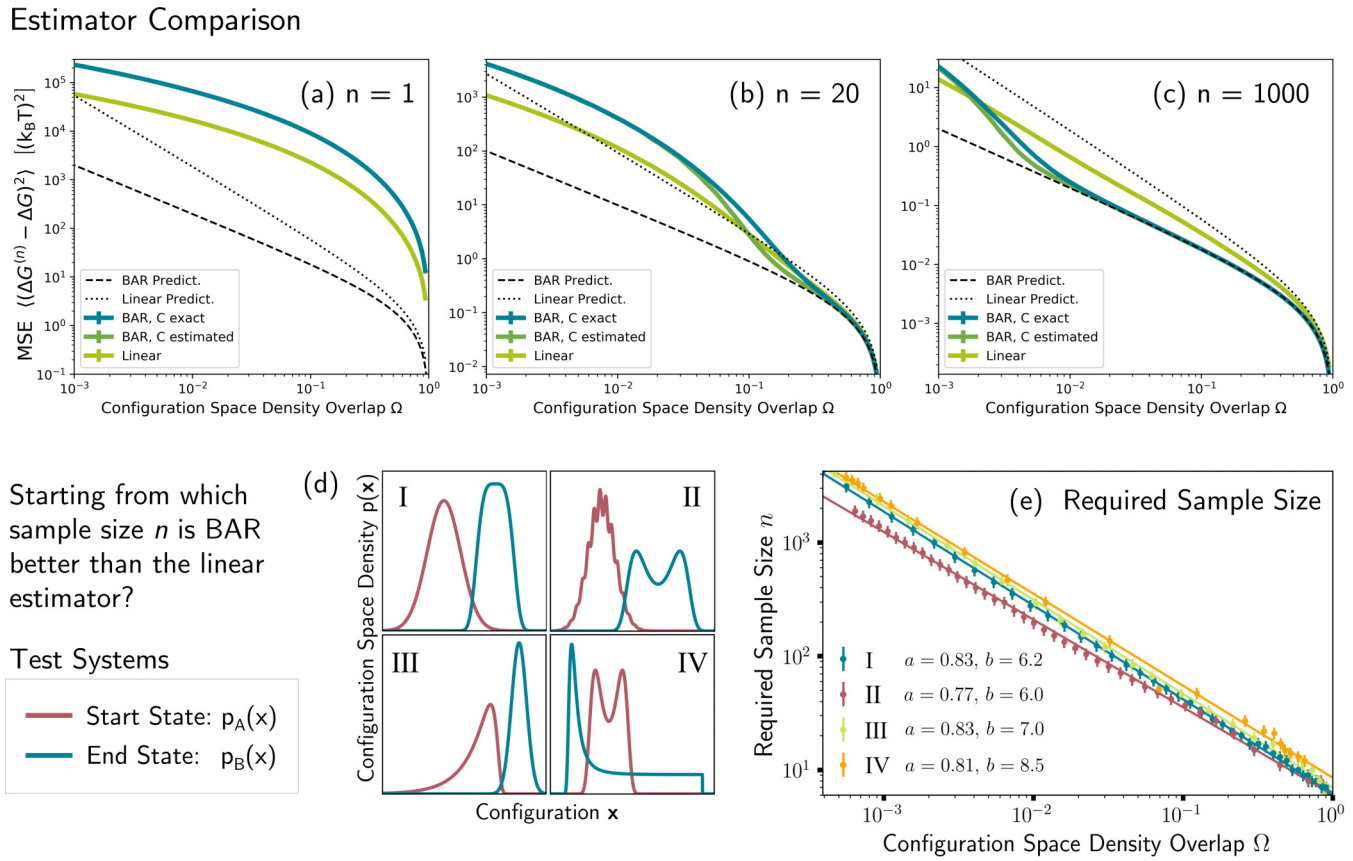

FIG. 2. Comparison of BAR and the linear estimator. (a)–(c) MSEs obtained from test simulations based on the setup shown in Fig. 1(b) for sample sizes of $n = 1$, 20, and 1000. The MSEs are shown as a function of the configuration space density overlap Ω where different Ω's were obtained by varying x_0 of the quartic end state. The results of two variants of BAR are shown: First, using a constant C that equals the exact free energy difference (blue), and second, for C that was iteratively determined for each set of samples (green). The MSE of the linear estimator is shown in yellow. The dashed and the dotted lines show the analytical MSEs calculated based on approximations for BAR and the linear estimator, respectively, i.e., Eqs. (14) and (16). (d) Setups used for the test simulations. The configuration space densities of the start and end states are shown in red and blue, respectively. Setup I is identical to the one in Fig. 1(b). (e) The minimum sample size n required such that the BAR with an exact C yields a better (smaller) MSE than the linear estimator is shown as a function of Ω. The Roman numbers indicate the underlying test system shown in (d). The solid lines show the function $n = b\,\Omega^{-a}$ fitted to the data points in the respective colors. The fit coefficients a and b are provided in the legend.

different horizontal shifts are used to vary Ω. The definitions and parameters of these systems are described in Appendix C.

The required number of sample points n is shown in dependence of Ω in Fig. 2(e). The four colors indicate the different test systems with corresponding roman numbers from Fig. 2(d). The required n closely follows a linear relation in the log-log plot, indicating a relation of the form $n = b\,\Omega^{-a}$. Fits of this form are shown as solid lines, and the fit coefficients are provided in the legend of Fig. 2(e). Interestingly, the relation between n and Ω is very similar for all four test systems, suggesting that Ω and n are almost the sole factors that determine which estimator is superior.

Figures 3(a)–3(c) compares MSEs for different intermediate sampling states S as a function of the overlap Ω between A and B for $n = 1$, 20, and 1000 per sample set. For $n = 1$, the linear intermediate combined with the linear estimator (yellow) yields the best MSE, followed by the linear intermediate with BAR (red) and VI (blue) that includes BAR as an estimator. For $n = 20$ and $n = 1000$, VI yields the best MSE for all Ω's. For the linear intermediate sampling state, for $n = 20$ a turning point exists ($\Omega \approx 5 \times 10^{-2}$), above which BAR is superior, and below which the linear estimator is superior. For $n = 1000$, BAR yields better MSEs at all Ω's.

Again, for $n = 1$ the predicted MSEs largely underestimate the actual error. However, already for $n = 20$, the actual MSE for VI is only slightly worse than the prediction and matches perfectly for $n = 1000$. For the linear intermediate, for $n = 20$ both the predictions for BAR and the linear estimator hold only for larger overlaps. For $n = 1000$, the one for BAR matches the actual MSEs very well, whereas for the linear estimator the prediction reproduces the trend but slightly overestimates the MSEs for small overlaps. We also tested how many sample points n are required per state for VI to be optimal. Whereas for systems with large Ω's, two or three sample points per state suffice, in no case does the required number of sample points exceed seven per state (data, therefore, not shown).

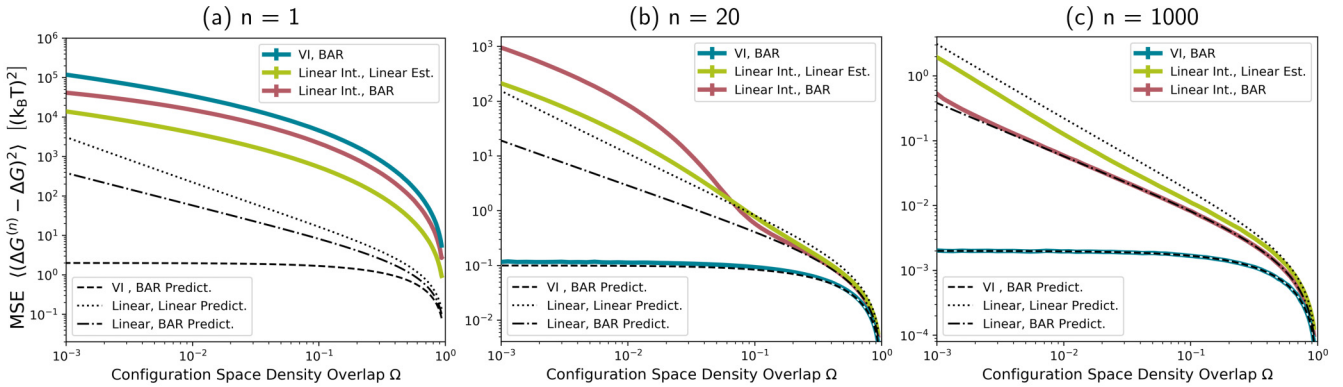

FIG. 3. Comparison of the MSEs between using a linear intermediate state and VI. As for Figs. 2(a)-2(c), test simulations with a harmonic and a quartic end state were used, and (a)–(c) show the results for samples size of $n = 1$, 20, and 1000, respectively, in each state as a function of the configuration space density overlap Ω between the end states. The results of two variants using a linear intermediate state are shown: First, using the linear estimator (yellow) and second, using BAR (red) to evaluate the stepwise free energy differences. The MSE of VI, which includes using virtual intermediate states that correspond to BAR as shown in Fig. 1(c) is shown in blue. The respective analytical MSEs are shown as black dashed, dotted, and dashed-dot lines.

These results show that, again, the predicted MSEs are inaccurate for small n's. As a consequence, VI, which has been derived as an optimization thereof, is suboptimal. However, using an intermediate sampling state, the MSEs become accurate, and VI becomes optimal for much fewer n's than for BAR. We attribute this unexpected result mainly to the fact that for VI the sampling intermediate still maintains a large overlap with both end states, even if their configuration space densities are entirely disjunct.

V. SUMMARY AND CONCLUSION

We have shown that for small sample sizes n the analytically calculated MSEs of free energy estimates based on the Zwanzig formula become increasingly inaccurate due to approximations in its derivation. As a consequence, BAR and VI, which have been derived as an optimization thereof, become suboptimal for small n, which was demonstrated through the existence of better alternatives. For BAR, as explained in the theory section following Eq. (8), even if the constant C is set to the exact free energy difference this suboptimality remains, and was even slightly worse in our test simulations than when C was estimated based on the samples.

Whether BAR and VI are optimal depends, aside from n, on the configuration space density overlap Ω, because for small Ω the fluctuations in the exponential averages increase. However, whereas BAR is suboptimal even for $n > 1000$ if $\Omega < 10^{-3}$, VI is already better than all other tested variants for $n = 7$ independent of Ω, owing to the fact that the overlap between adjacent states is largely increased when using an intermediate state. For BAR, Ω was almost the sole factor that determined how many sample points were required to be better than the linear estimator. The relation follows an inverse power law of the form $n = a\Omega^{-b}$ with very similar coefficients a and b for all four test systems considered.

For BAR, the discrepancy between the predicted MSEs and the actual ones also explains the well-known and frequent finding that the error of free energy calculations is often drastically underestimated.

For applications, instead of monitoring the variance or MSE directly (as implemented in many simulation software packages), we recommend to first consider Ω. Second, packages, such as ALCHEMLYB.PY [36–38] analyze the time correlations between sample points and give an estimate for the number of independent ones. Then, third, the relation between the required n and Ω from this paper will indicate whether BAR is optimal or whether another estimator, such as the linear one should be used instead.

We should reemphasize that in atomistic simulations subsequent sample points are correlated, whereas the theory developed and tested in this paper assumes independent sample points. Therefore, the critical number of sample points n identified here for which BAR becomes optimal will typically refer to the *effective* number of statistically independent sample points, which, due to long correlation times, is typically much smaller than the actual sample size. The small number effects on the MSE assessed here, therefore, are likely to be relevant also for the (seemingly) quite large sample sizes used in typical macromolecular free energy calculations.

To summarize, whereas BAR will remain the optimal estimator in many cases, our findings offer guidance in choosing the optimal estimator particularly for challenging applications.

APPENDIX A: PROOF OF MSE EQUIVALENCE TO BAR VARIANCE

The Zwanzig formula [9] Eq. (1) is used in two steps as shown in Fig. 1(a). The MSE of a single step is given through Eq. (3). Therefore, the total MSE is calculated through

$$\mathrm{MSE}\left(\Delta G^{(n)}_{A\rightleftharpoons B}\right) \tag{A1}$$

$$= \mathrm{MSE}\left(\Delta G^{(n)}_{A\to I}\right) + \mathrm{MSE}\left(\Delta G^{(n)}_{B\to I}\right) \tag{A2}$$

$$= \frac{1}{n}\left[\int [p_I(\mathbf{x})]^2\left(\frac{1}{p_A(\mathbf{x})} + \frac{1}{p_B(\mathbf{x})}\right)dx - 2\right]. \tag{A3}$$

Using the configuration space density of the optimal virtual intermediate Eq. (10),

$$p_I(\mathbf{x}) = \frac{[p_A(\mathbf{x})^{-1} + p_B(\mathbf{x})^{-1}]^{-1}}{\int dx [p_A(\mathbf{x})^{-1} + p_B(\mathbf{x})^{-1}]^{-1}} \tag{A4}$$

leads to

$$\begin{aligned} \mathrm{MSE}\big(\Delta G^{(n)}_{A\rightleftharpoons B}\big) &= \frac{1}{n}\frac{\int dx [p_A(\mathbf{x})^{-1} + p_B(\mathbf{x})^{-1}]^{-1}}{\left(\int dx [p_A(\mathbf{x})^{-1} + p_B(\mathbf{x})^{-1}]^{-1}\right)^2} - \frac{2}{n} \\ &= \frac{1}{n}\left(\int dx \frac{1}{p_A(\mathbf{x})^{-1} + p_B(\mathbf{x})^{-1}}\right)^{-1} - \frac{2}{n} \\ &= \frac{1}{n}\left(\int dx \frac{p_A(\mathbf{x})p_B(\mathbf{x})}{p_A(\mathbf{x}) + p_B(\mathbf{x})}\right)^{-1} - \frac{2}{n}, \end{aligned} \tag{A5}$$

which equals the variance from Bennett [11], Eq. (14).

APPENDIX B: MSE DERIVATION OF THE LINEAR ESTIMATOR

The linear estimator uses the linear interpolation $H_I(\mathbf{x}) = \frac{1}{2}[H_A(\mathbf{x}) + H_B(\mathbf{x})]$ as the virtual Hamiltonian. The corresponding MSE is calculated by inserting the configuration space density,

$$p_I(\mathbf{x}) = \frac{e^{-(1/2)[H_A(\mathbf{x})+H_B(\mathbf{x})]}}{Z_I}, \tag{B1}$$

into the expression of the MSE for using Zwanzig in two steps Eq. (A3) which yields

$$\mathrm{MSE}_{\mathrm{lin}}\big(\Delta G^{(n)}_{A\rightleftharpoons B}\big) = \frac{1}{n}\left\{\int \left[\frac{e^{-[H_A(\mathbf{x})+H_B(\mathbf{x})]}}{\left(\int e^{-(1/2)[H_A(\mathbf{x})+H_B(\mathbf{x})]}dx\right)^2} \times \left(\frac{Z_A}{e^{-H_A(\mathbf{x})}} + \frac{Z_B}{e^{-H_B(\mathbf{x})}}\right)\right]dx - 2\right\} \tag{B2}$$

$$= \frac{1}{n}\left(\frac{\int (Z_A e^{-H_B(\mathbf{x})} + Z_B e^{-H_A(\mathbf{x})})dx}{\left(\int e^{-(1/2)[H_A(\mathbf{x})+H_B(\mathbf{x})]}dx\right)^2} - 2\right) \tag{B3}$$

$$= \frac{1}{n}\left(\frac{2Z_A Z_B}{\left(\int e^{-(1/2)[H_A(\mathbf{x})+H_B(\mathbf{x})]}dx\right)^2} - 2\right) \tag{B4}$$

$$= \frac{2}{n}\left[\left(\int p_A(\mathbf{x})^{1/2} p_B(\mathbf{x})^{1/2} dx\right)^{-2} - 1\right]. \tag{B5}$$

APPENDIX C: PARAMETERS OF TEST SYSTEMS

The test systems shown in Fig. 2(d) are based on the Hamiltonians provided below. These were used to determine the results shown in Fig. 2(e), i.e., the minimum required number of sample points n as a function of Ω such that BAR yields a smaller MSE than the linear estimator.

System I: $H_A(\mathbf{x}) = 0.75x^2$ and $H_B(\mathbf{x}) = (x - x_0)^4$ using 46 values for x_0 with $0 \leqslant x_0 \leqslant 4.5$.

System II: $H_A(\mathbf{x}) = 0.1\sin(20x) + x^2$ and $H_B(\mathbf{x}) = 0.3x^4 - 0.8(x - x_0)^2$ using 47 values for x_0 with $0 \leqslant x_0 \leqslant 23$.

System III: $H_A(\mathbf{x}) = e^x - x$ and $H_B(\mathbf{x}) = 0.15(x - x_0)^2$ using 24 values for x_0 with $0 \leqslant x_0 \leqslant 9$.

System IV: $H_A(\mathbf{x}) = 0.3x^4 - 0.8(x - x_0)^2$ and $H_B(\mathbf{x}) = 4\epsilon[(\frac{\sigma}{x-x_0})^{12} - (\frac{\sigma}{x-x_0})^6]$ for $0 < x - x_0 \leqslant 15$ and $H_B(\mathbf{x}) = \infty$ otherwise, using $\epsilon = 2.0446$ and $\sigma = 3.405$ and 22 values for x_0 with $0 \leqslant x_0 \leqslant 4.03$.

[1] *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, edited by C. Chilpot and A. Pohorille, Springer Series in Chemical Physics Vol. 86 (Springer, Berlin/Heidelberg, 2007).
[2] H. Ge and H. Qian, Phys. Rev. E **94**, 052150 (2016).
[3] T. Sun, J. P. Brodholt, Y. Li, and L. Vočadlo, Phys. Rev. B **98**, 224301 (2018).
[4] Z. Cournia, B. Allen, and W. Sherman, J. Chem. Inf. Model. **57**, 2911 (2017).
[5] T. D. Swinburne and M.-C. Marinica, Phys. Rev. Lett. **120**, 135503 (2018).
[6] Ä. Baumeler and S. Wolf, Phys. Rev. E **100**, 052115 (2019).
[7] K. A. Armacost, S. Riniker, and Z. Cournia, J. Chem. Inf. Model. **60**, 1 (2020).
[8] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[9] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[10] D. Wu and D. A. Kofke, J. Chem. Phys. **123**, 054103 (2005).
[11] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[12] M. Reinhardt and H. Grubmüller, J. Chem. Theory Comput. **16**, 3504 (2020).
[13] M. R. Shirts and V. S. Pande, J. Chem. Phys. **122**, 144107 (2005).
[14] A. M. Hahn and H. Then, Phys. Rev. E **80**, 031111 (2009).
[15] G. König and S. Boresch, J. Comput. Chem. **32**, 1082 (2011).
[16] A. J. Schultz and D. A. Kofke, Mol. Simul. **47**, 379 (2021).
[17] A. Blondel, J. Comput. Chem. **25**, 985 (2004).
[18] C. D. Christ and W. F. van Gunsteren, J. Chem. Phys. **126**, 184110 (2007).
[19] J. W. Perthold and C. Oostenbrink, J. Phys. Chem. B **122**, 5030 (2018).
[20] G. König, N. Glaser, B. Schroeder, A. Kubincová, P. H. Hünenberger, and S. Riniker, J. Chem. Inf. Model. **60**, 5407 (2020).
[21] T. Steinbrecher, D. L. Mobley, and D. A. Case, J. Chem. Phys. **127**, 214108 (2007).

[22] T. T. Pham and M. R. Shirts, J. Chem. Phys. **135**, 034114 (2011).
[23] T. T. Pham and M. R. Shirts, J. Chem. Phys. **136**, 124120 (2012).
[24] F. P. Buelens and H. Grubmüller, J. Comput. Chem. **33**, 25 (2012).
[25] V. Gapsys, D. Seeliger, and B. L. de Groot, J. Chem. Theory Comput. **8**, 2373 (2012).
[26] M. Reinhardt and H. Grubmüller, Phys. Rev. E **102**, 043312 (2020).
[27] M. Reinhardt and H. Grubmüller, Comput. Phys. Commun. **264**, 107931 (2021).
[28] F. R. Beierlein, J. Michel, and J. W. Essex, J. Phys. Chem. B **115**, 4911 (2011).
[29] T. J. Giese and D. M. York, J. Chem. Theory Comput. **15**, 5543 (2019).
[30] P. Zhang, L. Shen, and W. Yang, J. Phys. Chem. B **123**, 901 (2019).
[31] R. Hall, T. Dixon, and A. Dickson, Front. Mol. Biosci. **7**, 106 (2020).
[32] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[33] M. Habeck, Phys. Rev. Lett. **109**, 100601 (2012).
[34] N. Lu, J. K. Singh, and D. A. Kofke, J. Chem. Phys. **118**, 2977 (2003).
[35] N. Lu, D. A. Kofke, and T. B. Woolf, J. Comput. Chem. **25**, 28 (2004).
[36] P. V. Klimovich, M. R. Shirts, and D. L. Mobley, J. Comput.-Aided Mol. Des. **29**, 397 (2015).
[37] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[38] J. D. Chodera, J. Chem. Theory Comput. **12**, 1799 (2016).