JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

Quantification of Entropy-Loss in Replica-Averaged Modeling

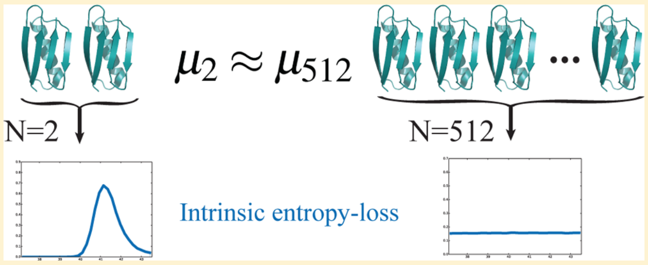
Simon Olsson*,†,‡ and Andrea Cavalli*,†,¶

†Institute for Research in Biomedicine, Via Vincenzo Vela 6, CH-6500 Bellinzona, Ticino, Switzerland

‡Laboratory of Physical Chemistry, Swiss Federal Institute of Technology, ETH-Hönggerberg, Vladimir-Prelog-Weg 2, CH-8093 Zürich, Zürich, Switzerland

¶Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW United Kingdom

ABSTRACT: Averaging across multiple replicas provides a straightforward and rigorous approach to employ averaged experimental data as restraints in molecular simulations. One significant practical obstacle is to optimally choose the number of replicas, *N*. Here, we describe a statistical method to estimate the intrinsic entropy-loss associated with modeling some data using *N* replicas, from an unbiased simulation. We discuss how having such a measure at hand may be used to assess *N* optimally.



INTRODUCTION

Molecular simulations are becoming an increasingly important tool for structural biologists to better understand the molecular underpinnings of biophysical data, in particular when sparse or averaged. While molecular mechanics force fields have undergone impressive improvements in the last years, it is still unclear how well these improvements will transfer not only to larger proteins but also to nucleic acids, intrinsically disordered proteins, and protein−protein interactions.[1−8] A classic approach to remedy the problem is to use experimental data to directly bias molecular simulations.[9−19] Much research has been done in this field, but, recently, we and others have started advocating invoking the Maximum Entropy principle (MEP) to achieve this.[20−25] By employing the MEP we are guaranteed to introduce the least bias possible to ensure agreement with averaged data while also maintaining self-consistency.

Chemical physics has seen numerous successful direct implementations of the MEP to bias simulations of small organic molecules, typically only with a small number of degrees of freedom.[26,27] However, the direct implementation involves estimation of a parameter, or *Lagrange multiplier*, λ_i, for each experimental observable used to bias the simulation. Recent papers describe efficient algorithms for their estimation,[24,28−31] but it remains to be seen how widely applicable these approaches are.

The major novelty in recent works on the MEP in the context of molecular simulations is a direct proof that computationally tractable *replica*-averaged restrained molecular simulations (RAMD) constitute increasingly good approximations of the MEP for an increasing number of replicas, *N*.[20−22] By implementing the MEP in this manner one avoids the estimation of the Lagrange multipliers altogether, however, choosing the ideal number of replicas, *N*, remains an issue. Ideally, *N* should be chosen as large as possible, however, increasing *N* comes at an increasing computational cost, as for each replica, another copy of the entire molecular simulation box is needed. Thus, it appears that any practical implementation of the MEP through RAMD will be a trade-off between accuracy and computational demand. Still, the question remains, how may we gauge the accuracy of a *N*-replica approximation relative to the exact solution. − Further, how does the data set at hand and its uncertainty influence this accuracy? Having such methodology at hand would enable us to more rigorously assess the ideal number of replicas needed for a particular application.

In this Letter we present a simple statistical method to emulate RAMD simulations of *N*-replicas, using unrestrained, unbiased simulations. Consequently, this method may be used to compute the entropy-loss associated with approximating the MEP modeling of some data by using *N*-replicas to form averaged restraints. We discuss how this approach may be used to more efficiently assess *N* for restrained molecular simulations quantitatively.

THEORY

The aim here is to establish a procedure which allows for the estimation of entropy-loss associated with representing an expectation from a conformational distribution by an average of a discrete set of conformations. In general, one may compute this quantity by systematically running simulations with an increasing number of averaging replicas *N*. However, this procedure will naturally be associated with considerable computational effort. In the following, we will introduce the concept of sum probability density functions (spdfs). In combination with a set of samples from an unrestrained simulations, these spdfs may be used to construct posterior pdfs corresponding to the simulations restrained with a term averaging over an increasing number of replicas, *N*. Finally,

Received: June 19, 2015
Published: August 7, 2015

DOI: 10.1021/acs.jctc.5b00579

we will show how sum probabilities may be used to directly compute the entropy-loss associated with a specific N-replica approximation of the MEP.

Sum Probabilities. Let $\{x_i\}$ be a set of samples (*microstates*) from the probability density function (pdf) $p_1(x)$ e.g. the Boltzmann distribution of a potential energy function $\mathcal{E}$ at a given temperature. If α_N is the sum of N random samples from $p_1(x)$, then its samples $\{x_i\}$ may be used to generate samples from the sum distribution $p_N(\alpha_N)$ as

$$\alpha_N \sim \sum_{j=1}^{N} x_j, \quad x_j \overset{\text{i.i.d.}}{\sim} p_1(x) \tag{1}$$

The mean of $p_N(\alpha_N)$ is $N\mu$ where μ is the mean of $p_1(x)$. Alternatively, in cases where the characteristic function of $p_1(x)$ is known and analytically tractable, we may directly obtain the sum distribution $p_N(\alpha_N)$ as the inverse Fourier transform of the N-th power of the characteristic function of $p_1(x)$.

Computing the Probability of a Microstate in a Sum of N+1 Replicas Using Bayes Theorem. We want to specify a probability density function of observing a microstate x as a component of a sum of N+1 microstates given the sum is exactly $\hat{\mu}(N$+1), where $\hat{\mu}$ is the target mean. We may express such a distribution using Bayes theorem, as the ratio between the probability of observing the sum (N+1)$\hat{\mu}$ given one of the components, x (the *likelihood*), and the probability of the sum of N+1 components being (N+1)$\hat{\mu}$ (the *evidence*) times the probability of observing a particular microstate $p_1(x)$ (the *prior*)

$$p_{N+1}(x|x+\sum_{i=1}^{N} x_i = \hat{\mu}(N+1)) = \frac{p_N(\sum_{i=1}^{N} x_i = \hat{\mu}(N+1) - x|x)}{p_{N+1}(\sum_{i=1}^{N+1} x_i = \hat{\mu}(N+1))} p_1(x) \tag{2}$$

This (*posterior*) distribution may be constructed by realizing that the likelihood and evidence may be expressed in terms of sum probability densities introduced above. Specifically, the likelihood may be expressed as $p_N(\alpha_N)$ by choosing $\alpha_N = (N+1)\hat{\mu} - x$, while the evidence is expressed as $p_{N+1}(\alpha_{N+1})$ with $\alpha_{N+1} = (N+1)\hat{\mu}$.

We are free to choose $\hat{\mu}$ for any value in the domain of $p_1(x)$. In the special case where $\hat{\mu} = \mu$, the ratio of the likelihood and the evidence will express the bias associated with the particular N+1 replica approximation. In all other cases, this ratio will have an additional contribution, the MEP bias, which we will quantify next.

Equivalence to the Maximum Entropy Principle. Following the central limit theorem a sum probability distribution $p_N(\alpha_N)$ will be normal for $N\to\infty$. Thus, we may express a sum probability density with a Gaussian pdf as

$$p_N(\alpha_N) \approx \exp(a\alpha_N^2 + b\alpha_N + k) \tag{3}$$

$$= Z^{-1}\exp\left(-\frac{\alpha_N^2}{2\sigma_{\alpha_N}^2} + \frac{N\mu\alpha_N}{\sigma_{\alpha_N}^2}\right) \tag{4}$$

where $\sigma_{\alpha N}$ is the standard deviation of α_N. As the evidence from above is invariant in x we can now express the entire bias by inserting the likelihood from above

$$p_N((N+1)\hat{\mu} - x) = Z^{-1}\exp\left(-\frac{[(N+1)\hat{\mu} - x]^2}{2N\sigma^2} + \frac{N\mu[(N+1)\hat{\mu} - x]}{N\sigma^2}\right) \tag{5}$$

$$= \hat{Z}^{-1}\exp\left(\frac{x^2 + 2N\hat{\mu}x - 2\hat{\mu}x}{2N\sigma^2} - \frac{\mu x}{\sigma^2}\right) \tag{6}$$

where σ is the standard deviation of $p_1(x)$. In the last step we expand the brackets and absorb all constant terms into the normalization constant $\hat{Z}$. If we now consider the limit of eq 6

$$\lim_{N\to\infty} p_N((N+1)\hat{\mu} - x) \approx \hat{Z}^{-1} \exp\left(\frac{(\hat{\mu} - \mu)}{\sigma^2}x\right) \tag{7}$$

which is simple linear exponential bias equivalent to that derived by the MEP. We observe that this bias will be zero when the target mean $\hat{\mu}$ is equal to the mean of the prior μ, and thus when this term constitutes the MEP bias referred to above when $\hat{\mu} \neq \mu$.

METHODS

Simulation of Posterior Distributions. We used a previously reported 20 μs MD simulation[1] of the 56 residue protein GB3 (the third immunoglobin-binding domain of Fab) in the AMBER99SB-ILDN* force field and Camshift[32] to construct a $p_1(x)$ distribution using chemical shifts as reaction coordinate. $p_1(x)$ has 320 dimensions each of which correspond to a specific nonterminal, backbone atom chemical shift including $\{H^N, N, C^\alpha, C^\beta, C', H^\alpha\}$. In the following we treat each of these dimensions as statistically independent. Using $p_1(x)$ we generated histograms with 12 uniformly spaced bins of the *likelihood* $p_N((N+1)\hat{\mu}-x)$ for N+1 = {2, 4, 8, 16, 32, 64, 128, 256} using eq 1. This was repeated for a number of different sets of *target means* $\hat{\mu}$, specifically, $\{\mu - 0.02\sigma_X, \mu, \mu + 0.02\sigma_X\}$ corresponding to the sample mean of $p_1(x)$ − μ − and this value was perturbed by 2% of the Camshift random prediction error for nuclei X as previously reported.[32] Similarly, we constructed a histogram of the *prior* $p_1(x)$ which in turn allowed us to generate posterior histograms using eq 2. For illustrative purposes we repeated the computation for one chemical shift, C^α of Tyr3, with $\hat{\mu} = \mu$ for N+1 = {2, 4, 8, 16, 32, 64, 128, 256, 512} using histograms with 32 uniformly spaced bins.

Simulating the Influence of Noise. To emulate the influence of Gaussian noise to the entropy-loss, we repeated the procedure above for N+1 = {2, 4, 8, 16, 32, 64}, keeping $\hat{\mu}$ fixed at μ but adding increasing amounts of Gaussian noise to $p_1(x)$. The standard deviations of the Gaussian noise − or noise fractions − added were {0.00, 0.05, 0.10, 1.00} in units of σ_X, which again is the random prediction error of Camshift for the nuclei X.

Measuring Entropy-Loss. We quantify the loss of entropy by using the Kullback−Leibler divergence between the prior and the posterior

$$D_{\mathrm{KL}}(p_1|p_{N+1}) = \int_{-\infty}^{\infty} \mathrm{d}x\, p_{N+1}(x|x+\sum_{i=1}^{N} x_i = \hat{\mu}(N+1)) \ln\frac{p_{N+1}(x|x + \sum_{i=1}^{N} x_i = \hat{\mu}(N+1))}{p_1(x)} \tag{8}$$

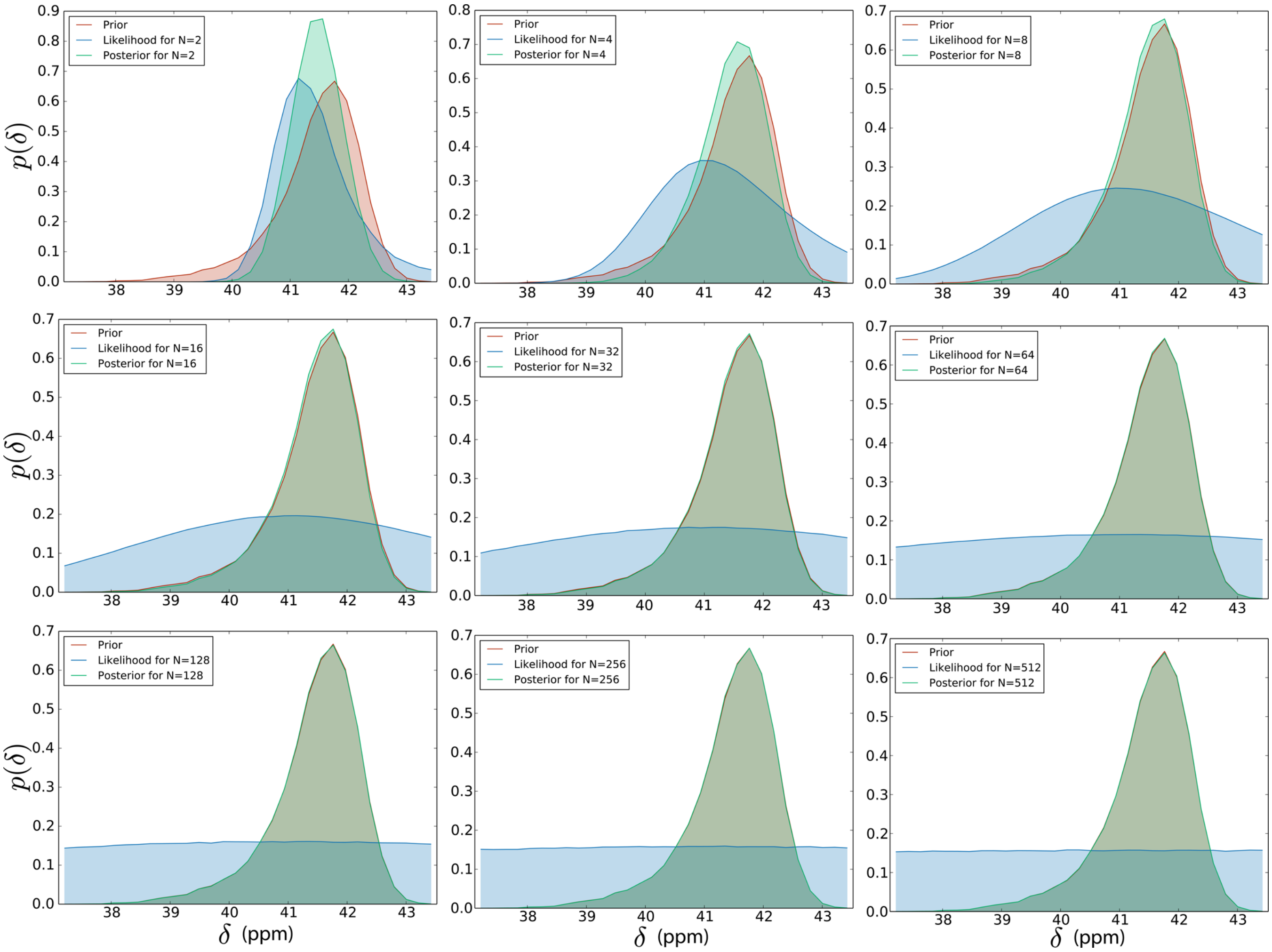


Figure 1. Histograms for prior (red), likelihood (blue), and posterior (green) distributions of Tyr3 C^{α} chemical shift for increasing number of replicas N, with $\hat{\mu} = \mu$.

This quantity will be zero if and only if the ratio of the likelihood and evidence in eq 2 is uniform in x. The values reported here are averages across the whole data set computed as

$$\overline{KL} = \frac{1}{M} \sum_{i=1}^{M} D_{\mathrm{KL}}(p_1^i \,|\, p_{N+1}^i) \tag{9}$$

where M is the number of dimensions of x_1.

MEP Bias. The MEP bias is the minimum amount of bias necessary to ensure agreement with experimental data if the unrestrained simulations do not agree with the data *a priori*. It is computed by explicit estimation of Lagrange multipliers by steepest decent

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + \Delta t(\hat{\mu}_i - \delta_i^{\lambda^{(n)}}) \tag{10}$$

with

$$\delta_i^{\lambda^{(n)}} = Z^{-1}(\lambda^{(n)}) \int \mathrm{d}x\, \delta_i(x) \exp(\sum_i \lambda_i^{(n)} \delta_i(x)) \tag{11}$$

where $Z^{-1}(\boldsymbol{\lambda}^{(n)})$ is a normalization constant, and $\Delta t = 0.1$. $\lambda_i^{(n)}$ is the nth Lagrange multiplier estimate for the i chemical shift, which for an instantaneous state x is represented by $\delta_i(x)$. Convergence is defined as the normalized mean error $Q = \langle(|\hat{\mu}_i - \delta_i(x)|)/(\delta_i(x))\rangle_i$ being $<10^{-12}$. After C iterations convergence is reached, and histograms are computed using samples, x, weighed according to

$$\omega(x) \propto \exp\left(\sum_i \lambda_i^{(C)} \delta_i(x)\right) \tag{12}$$

RESULTS AND DISCUSSION

Visualization of Entropy-Loss Due to Replica-Averaged Restraints. To illustrate how replica-averaged restraints introduce an intrinsic entropy-loss into simulations we determined prior, likelihood, and posterior distribution estimates for the α carbon chemical shift in Tyr3 of GB3 as a function of the number of replicas N, see Figure 1. This was achieved by using an unrestrained 20 μs molecular dynamics simulation of GB3 in explicit solvent to generate $p_1(x)$, which in turn was used to generate the likelihood and posterior distributions. As we here wish to consider only bias due to replica-averaging, we assume that neither systematic or random errors are present, that is $\hat{\mu} = \mu$. In this case, we expect the likelihood to converge to a uniform distribution for $N \to \infty$ − that is, in this limit there will be no bias, and therefore the prior and the posterior distributions will coincide.

DOI: 10.1021/acs.jctc.5b00579
J. Chem. Theory Comput. 2015, 11, 3973−3977

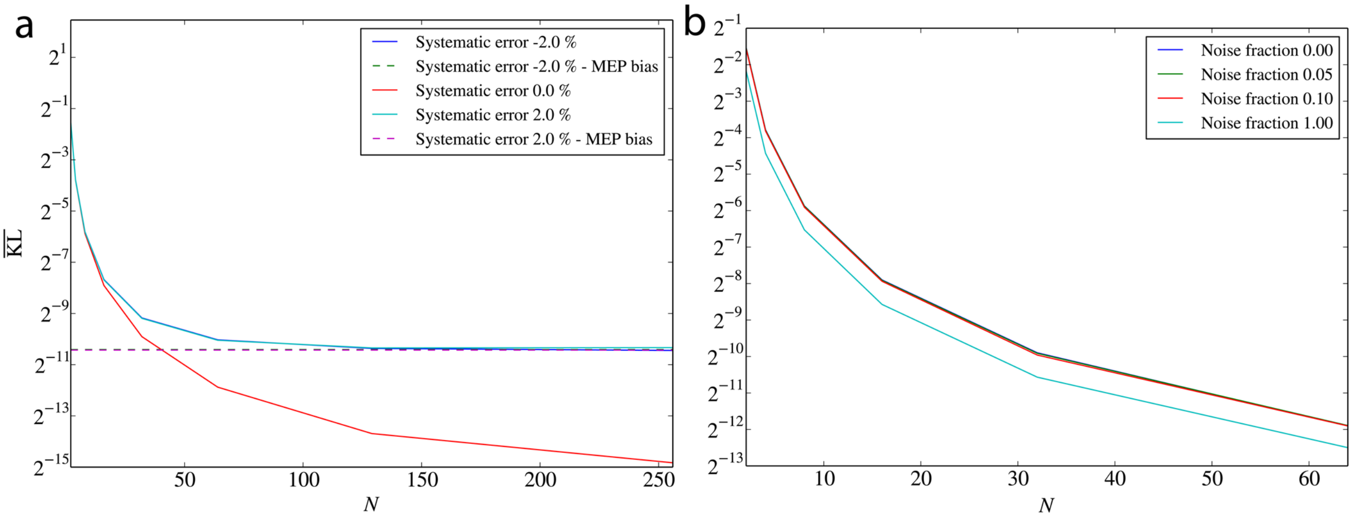


Figure 2. Average entropy-loss (average Kullback-Liebler divergences) associated with using replica-averaged restraints as a function of the number of replicas, N, as well as the influence of systematic errors (a) and random noise (b). Dashed lines in (a) indicate MEP biases i.e. in the absence of entropy-loss due to replica averaging.

For $N = 2$ we observe considerable overlap between the prior (*target*) and posterior distributions; however, the skewness present in prior is not readily captured. While $N = 4$ and $N = 8$ do capture the prior skewness, the posterior distributions still are overtly narrowly distributed. From $N = 16$ and on it becomes difficult to visually distinguish the posterior from the prior distribution; however, until $N = 256$ the likelihood is visually distinct from a uniform distribution.

These results show that employing replica-averaged restraints with a small number of replicas (in this case, up to around $N = 16$) may introduce a considerable bias even if the prior, that is the force field, would agree with the data in a free simulation. This result is not immediately intuitive and underlines the importance of testing whether the force field at hand already is sufficiently accurate to reproduce the available experimental data. While such tests may involve significant simulation time, if the force field indeed agrees with the data restraining is not necessary in the first place. On the other hand, if it turns out that the simulation systematically disagrees with experimental data it may be used to emulate RAMD simulations for an arbitrary number N using our formalism, as we discuss next.

Quantification of Entropy-Loss and the Influence of Systematic Errors and Random Noise. Having established some intuition about how the intrinsic bias of replica-averaged restraints diminishes as a function of the number of replicas N in the one-dimensional case we now turn to investigate this bias more quantitatively for a number of full, synthetic backbone chemical shift data sets.

We first consider the case also studied above, in the absence of random and systematic error, that is, $\hat{\mu} = \mu$ (Figure 2a, red line). As anticipated, the average entropy-loss ($\overline{KL}$) is monotonically decreasing as a function of the number of replicas, N. As noted above, this function will approach zero for $N \to \infty$.

To emulate a practical case where a simulation does not agree with experimental data *a priori*, we introduce a systematic error, $\hat{\mu} \neq \mu$. In this case, the entropy-loss will not approach zero for $N \to \infty$ but rather converge to a finite bias approximately given by eq 7. Indeed, we observe that the cases where systematic errors in $\hat{\mu}$ were introduced (green and cyan lines in Figure 2a) converge to a nonzero bias. For comparison we obtain numerical estimates of the MEP bias, from eqs 9 and 12, and plot these values (Figure 2a, dashed green and purple lines). Indeed, the MEP bias estimates coincide with the finite biases to which the entropy-loss converges for the emulated replica-averaged simulations. Interestingly, the entropy-loss converges to the MEP bias for a lower N in the presence of systematic errors than in their absence (red line, Figure 2a).

Quantification of the influence of experimental noise on replica-averaged restrained simulations is an important topic but has until yet only received modest treatment.[21,25,33] If we consider random errors (Figure 2b) using our framework we find that the bias is systematically reduced as a function of the amplitude of the added random noise. It appears that the influence of the bias introduced by replica-averaged restraints are partially overcome when the amplitude of the experimental noise is substantial. However, as this effect is proportional to the noise, and many modern biophysical measurements allow for highly accurate measurements, it seem likely that systematic deviations from force fields will be the primary source of error in the coming future.[34]

■ CONCLUSION

RAMD is a straightforward way to integrate experimental restraints into molecular simulations in a manner consistent with the MEP. However, while the direct MEP solution is unambiguous, an ambiguity is introduced by the replica-averaged implementation, namely the number of replicas N. So far, no rigorous guidelines have existed to assess the number of replicas to use. Consequently, mostly *ad hoc* validations, or computationally expensive cross-validation tests and practical considerations, have gone into deciding this number in the literature. Here we present a statistical method to assess the additional bias, or *entropy-loss*, induced by the use of replica-averaged restraints for a number of different theoretical scenarios. This approach moves the community closer to a quantitative and rigorous determination of N and may also be used to assess the validity of particular simulations either before or after they are performed.

An implementation of the method described herein is available upon request from the authors or at https://github.com/cavallilab/EntroLoss.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: simon.olsson@phys.chem.ethz.ch.

*E-mail: andrea.cavalli@irb.usi.ch.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.O. is funded by an Independent Postdoc grant from The Danish Council for Independent Research | Natural Sciences (ID: DFF-4002-00151).

REFERENCES

(1) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS One* **2012**, *7*, e32131.

(2) Beauchamp, K. A.; Lin, Y.-S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.

(3) Krepl, M.; Havrila, M.; Stadlbauer, P.; Banas, P.; Otyepka, M.; Pasulka, J.; Stefl, R.; Sponer, J. *J. Chem. Theory Comput.* **2015**, *11*, 1220−1243.

(4) Best, R. B.; Zheng, W.; Mittal, J. *J. Chem. Theory Comput.* **2014**, *10*, 5113−5124.

(5) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.

(6) Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A. *J. Chem. Theory Comput.* **2015**, *11*, 2−7.

(7) Henriques, J.; Cragnell, C.; Skepö, M. *J. Chem. Theory Comput.* **2015**, *11*, 3420−3431.

(8) Stanley, N.; Esteban-Martína, S.; Fabritiis, G. D. *Prog. Biophys. Mol. Biol.* **2015**, DOI: 10.1016/j.pbiomolbio.2015.03.003.

(9) Clore, G. M.; Nilges, M.; Sukumaran, D. K.; Brünger, A. T.; Karplus, M.; Gronenborn, A. M. *EMBO J.* **1986**, *5*, 2729−2735.

(10) Torda, A. E.; Scheek, R. M.; van Gunsteren, W. F. *Chem. Phys. Lett.* **1989**, *157*, 289−294.

(11) Kim, Y.; Prestegard, J. H. *Biochemistry* **1989**, *28*, 8792−8797.

(12) Kemmink, J.; Scheek, R. M. *J. Biomol. NMR* **1995**, *6*, 33−40.

(13) Hess, B.; Scheek, R. M. *J. Magn. Reson.* **2003**, *164*, 19−27.

(14) Best, R. B.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 8090−8091.

(15) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291−3299.

(16) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128−132.

(17) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476−477.

(18) Bouvignies, G.; Markwick, P.; Brüschweiler, R.; Blackledge, M. *J. Am. Chem. Soc.* **2006**, *128*, 15100−15101.

(19) Vendruscolo, M. *Curr. Opin. Struct. Biol.* **2007**, *17*, 15−20.

(20) Pitera, J. W.; Chodera, J. D. *J. Chem. Theory Comput.* **2012**, *8*, 3445−3451.

(21) Cavalli, A.; Camilloni, C.; Vendruscolo, M. *J. Chem. Phys.* **2013**, *138*, 094112.

(22) Roux, B.; Weare, J. *J. Chem. Phys.* **2013**, *138*, 084107.

(23) Olsson, S.; Frellsen, J.; Boomsma, W.; Mardia, K. V.; Hamelryck, T. *PLoS One* **2013**, *8*, e79439.

(24) Olsson, S.; Vögeli, B. R.; Cavalli, A.; Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K.; Hamelryck, T. *J. Chem. Theory Comput.* **2014**, *10*, 3484−3491.

(25) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. *PLoS Comput. Biol.* **2014**, *10*, e1003406.

(26) Catalano, D.; Emsley, J. W.; La Penna, G.; Veracini, C. A. *J. Chem. Phys.* **1996**, *105*, 10595−10605.

(27) Berardi, R.; Spinozzi, F.; Zannoni, C. *J. Chem. Phys.* **1998**, *109*, 3742−3759.

(28) Olsson, S.; Ekonomiuk, D.; Sgrignani, J.; Cavalli, A. *J. Am. Chem. Soc.* **2015**, *137*, 6270−6278.

(29) Habeck, M. *Phys. Rev. E* **2014**, *89*, 052113.

(30) White, A.; Voth, G. *J. Chem. Theory Comput.* **2014**, *10*, 3023−3030.

(31) White, A. D.; Dama, J. F.; Voth, G. A. *J. Chem. Theory Comput.* **2015**, *11*, 2451−2460.

(32) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 13894−13895.

(33) Fenwick, R. B.; Esteban-Martín, S.; Salvatella, X. *J. Phys. Chem. Lett.* **2010**, *1*, 3438−3441.

(34) Vögeli, B.; Segawa, T. F.; Leitz, D.; Sobol, A.; Choutko, A.; Trzesniak, D.; van Gunsteren, W.; Riek, R. *J. Am. Chem. Soc.* **2009**, *131*, 17215−17225.