Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations: v1.0

Alan Grossfield[1*†], Paul N. Patrone[2*†], Daniel R. Roe[3*†], Andrew J. Schultz[4*†], Daniel W. Siderius[5*†], Daniel M. Zuckerman[6*†]

[1]University of Rochester Medical Center, Department of Biochemistry and Biophysics; [2]Applied Computational and Mathematics Division, National Institute of Standards and Technology; [3]Laboratory of Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health; [4]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York; [5]Chemical Sciences Division, National Institute of Standards and Technology; [6]Department of Biomedical Engineering, Oregon Health & Science University

This LiveCoMS document is maintained online on GitHub at https://github.com/dmzuckerman/Sampling-Uncertainty; to provide feedback, suggestions, or help improve it, please visit the GitHub repository and participate via the issue tracker.

This version dated December 4, 2017

Abstract The quantitative assessment of uncertainty and sampling quality is essential in molecular simulation. Many systems of interest are highly complex, at the edge of current computational capacity, and simulators must understand and communicate statistical uncertainties so that 'consumers' of the data understand the meaning and limitations of the simulation data. This article covers key analyses appropriate for trajectory data generated by straightforward simulation methods such as molecular dynamics and (single Markov chain) Monte Carlo, as well as providing guidance for analyzing some 'enhanced' sampling approaches. We do not discuss *systematic* errors arising from inaccuracy in the chosen model or force field.

***For correspondence:**
alan_grossfield@urmc.rochester.edu (AG); paul.patrone@nist.gov (PNP); daniel.roe@nih.gov (DRR); ajs42@buffalo.edu (AJS); daniel.siderius@nist.gov (DWS); zuckermd@ohsu.edu (DMZ)

[†]These authors contributed equally to this work.

1 Introduction: Scope and definitions

1.1 Scope

This article is geared toward simulation studies attempting to quantify observables and derive reliable estimates of uncertainty (e.g., error bars) based on 'standard' canonical sampling methods (e.g., molecular dynamics and Monte Carlo) and associated 'enhanced' sampling methods. Some problems and systems may be better studied with cruder techniques and analyses, which will not be covered here. This article also will not cover issues of systematic error arising from inaccuracy in force field (underlying model) or even from the simulation setup. Rather, this article will take raw trajectory data at face value, assuming it is a valid outcome given the underlying model. We emphatically will *not* assume that a trajectory has been sufficiently 'equilibrated'.

1.2 Key Definitions

In order to make the discussion that follows more precise, we define key terms used in subsequent sections. We caution that while many of these concepts are familiar, our terminology follows the *International Vocabulary of Metrology* (VIM)[20], a standard that sometimes differs from the conventional or common language of engineering statistics. For additional information about or clarification of the statistical meaning of terms in the VIM, we suggest that readers consult the *Guide to the expression of uncertainty in measurement* (GUM)[19]. In cases of lexical ambiguity, the reader shall assume that we hold to the definition of terms as given in the VIM. For clarity, we highlight differences between the VIM and conventional terms in the following glossary and then discuss the statistical terms more generally in the subsequent section.

1.2.1 Glossary of Statistical Terms

- **Random quantity**: A quantity whose numerical value is inherently unknowable or unpredictable. Observations or measurements taken from a molecular simulation are treated as random quantities[1].
- **True value:** The value of a quantity that is consistent with its definition and is the objective of an idealized measurement or simulation. The adjective "true" is often dropped when reference to the definition is clear by context[19, 20].
- **Standard Uncertainty**: Uncertainty in a result (e.g. estimation of a true value) as expressed in terms of a standard deviation[2].
- **Arithmetic mean**: An estimate of the (true) expectation value of a random quantity, given by the formula

$$\bar{q} = \frac{1}{n}\sum_{k=1}^{n} q_k \tag{1}$$

  where q_j is an experimental or simulated realization of the random variable and n is the number of samples.

  Remark: This quantity is often called the "sample mean."
- **Experimental standard deviation**[3] : An estimate of the (true) standard deviation of a random variable, given by the formula

$$s\left(q_k\right) = \sqrt{\frac{\sum_{j=1}^{n}\left(q_j - \bar{q}\right)^2}{n-1}} \tag{2}$$

  Remark: This quantity is often called the "sample standard deviation."
- **Experimental standard deviation of the mean**: An estimate of the standard deviation of the distribution of the arithmetic mean, given by the formula

$$s\left(\bar{q}\right) = \frac{s\left(q_k\right)}{\sqrt{n}}. \tag{3}$$

  Remark: This quantity is often called the "standard error."
- **Precision**: The amount of variability in an estimate (e.g. based on repeating a given simulation protocol multiple times).
- **Accuracy**: The degree to which a result agrees with a reference value. The latter may be an experimental measurement or the result of a well-sampled simulation.

- **Raw data**: The numbers that the computer program directly generates as it proceeds through a sequence of states. For example, a MC simulation generates a sequence of configurations, for which there are associated properties such as the instantaneous pressure, temperature, volume, etc.
- **Derived observables**: Quantities derived from 'non-trivial' analyses of raw data, e.g. properties that may not be computed for a single configuration such as free energies.
- **Independent observables**: Random quantities x and y for which the joint probability density can be written as a product of individual probability densities $P(x,y) = P(x)P(y)$.
- **(Linearly) Uncorrelated observables**: If quantities q_j and q_k have mean values $\langle q_j \rangle$ and $\langle q_k \rangle$, then q_j and q_k are linearly uncorrelated if

$$\left\langle \left(q_j - \langle q_j \rangle\right)\left(q_k - \langle q_k \rangle\right)\right\rangle = 0 \tag{4}$$

  where <⋆> denotes the (true) expectation value.

  Remark: Independence of random variables implies that they are linearly uncorrelated. The converse, however, is not true[4].

[1] Most molecular simulations (even though using pseudo-random number generators) are deterministic in that the sequence of visited states is generated by a fixed algorithm and, as such, the simulation output is never truly random. In practice, however, the chaotic nature of the simulation allows for application of the principles of statistics to the analysis of simulation observations. Thus, observations/measurements taken at points along the simulation may be treated as random quantities. See Ref. [**?**] for more discussion of this rather deep point.

[2] The definition of standard uncertainty does not specify how to calculate the standard deviation. This choice ultimately rests with the modeler and should be dictated by the details of the uncertainty relevant to the problem at hand.

[3] We note that the use of the adjective "experimental" here does not necessarily imply that a measurement was made in a physical experiment. In fact, the VIM[20] treats the word "experiment" and its grammatical derivatives as non-defined, primitive, concepts with assumed meaning. For the purposes that follow, a molecular simulation *is* an experiment. Thus, the variation in measurements taken in a simulation may be appropriately called the "experimental standard deviation." The same applies to other terms that include the adjective "experimental."

[4] In practice, it is easier to (empirically) show that random variables are uncorrelated than independent. However, linear correlations are often all that are needed to arrive at uncertainty estimates for various quantities. Cf. the discussion of correlation times and autocorrelation analyses.

- **Correlated observables**: Random quantities that are not independent.
- **Correlation time**: In time-series data of a random quantity $q(t)$ (e.g. a physical property from a MC or MD trajectory), this is the time τ over which $q(t)$ and $q(t+\tau)$ remain (linearly) correlated[5].
- **Two-sided confidence interval**: A statistically derived pair of lower and upper bounds between which the (true) expectation value of a random quantity is likely to fall, as quantified by a *confidence level* (typically given as a percentage, e.g. 95 %). The confidence level, in conjunction with the sample size, determines the *coverage factor*, which is multiplied by $s(\bar{q})$ to assign the bounds of the confidence interval.

1.2.2 Discussion of terminology

As we recognize that the glossary of statistical terms is overly pedantic, it is valuable to subsequently connect those formal terms to the ultimate goal of this guide, to enable the estimation of uncertainty in thermophysical properties obtained via molecular simulation. Additionally, this section will connect the formal VIM definitions to more common usages for accessibility. We again reiterate here that the statistical definitions in the VIM[20] are chosen to both minimize ambiguity in the particular word choice and honestly restrict the statistical capability of their associated mathematical quantities.

Firstly, we point out that a property obtained by analysis of results from a molecular simulation (either directly from raw data or as a derived observable) should be quantified via the arithmetic mean (i.e., the sample mean, using older terminology). This mean property is emphatically *not* the true value; an infinite number of simulations would be necessary to properly obtain the true value. Secondly, the uncertainty in the arithmetic mean should be quantified via the experimental standard deviation of the mean (i.e., the standard error, using older terminology). This standard deviation characterizes the dispersion of the arithmetic mean relative to its true value; it is *not* the statistical variation in the true value itself (e.g. it is not used to compute fluctuation quantities). Lastly, results of a molecular simulation should be reported as 1) the arithmetic mean with 2) an appropriate confidence interval, e.g., as $\bar{x} \pm c$, where $\bar{x}$ is the arithmetic mean (of the property of interest) and c is the size of the confidence interval. Calculation of the confidence interval, which requires preselection of a confidence level based on field-specific heuristics, is described in Section 7.

[5] Generally speaking, MC and MD trajectories generate new configurations from preceding ones. Thus, the correlation time can be interpreted as the time over which the system retains memory of its previous states. Such correlations are often **stationary**, meaning that τ is independent of t. Roughly speaking, the total simulation time divided by the longest correlation time yields an estimate of the number of *uncorrelated* samples generated by a simulation.

2 Best Practices Checklist

3 Pre-simulation sanity checks and planning tips

Sampling a molecular system that is complex enough to be "interesting" in modern science is likely to be extremely challenging. Therefore, a small amount of effort spent planning a study can pay off many, many times over: in the worst case, weeks or months of simulation time followed by additional weeks or months of analysis can lead to insignificant conclusions in a poorly planned study.

If you read this guide through *before* performing a simulation, you will have a much better sense of the criteria applicable to your data – and which indeed *should* be applied by knowledgeable reviewers of your work. Thus we strongly advise understanding the concepts presented here as well as in related reviews [15, 19].

In a generic sense, the overall goal is to be able to draw statistically significant conclusions regarding a particular phenomenon or question of interest. And "good statistics" follow from repeated observations of a phenomenon, which can only happen if the simulation length *exceeds* the pertinent timescales.

One general strategy that will allow you to at least understand your data in a self-consistent way is to perform several repeats of the same simulation protocol: as described below, repeats can be used to assess variance in *any* observable, within the time you have run your simulation. When performing simulation repeats, it is sometimes suggested to use different starting states which are as diverse as possible; then, differences among the runs can be an indicator of inadequate sampling of the equilibrium distribution. Alternatively, performing multiple runs from the same starting state will yield behavior particular to that starting state and potentially equilibrium if the runs are long enough.

What are the pertinent timescales? That's a question which has to be answered individually for each system. You will want to study the experimental and computational literature for your particular system, although we warn that a published prior simulation of a given length does not in itself validate a new simulation of a similar or slightly increased length. In the end, your data must be validated as well as possible by statistical analyses, such as those described below. Be warned that a system may possess states (regions of configuration space) that, although important, are *never* visited in a given simulation set because of insufficient computational time [15] – and this type of error will not be discovered through the analyses presented below. Finally, note that "system" here does not necessarily refer to a complete biological simulation (e.g. protein, solvent, ions, etc) - it can also refer to

QUANTIFYING UNCERTAINTY AND SAMPLING QUALITY IN MOLECULAR SIMULATION

☐ **Plan your study carefully: Start with pre-simulation sanity checks.** *Underlying concept:* There is no guarantee that any method, enhanced or otherwise, can sample the system of interest. See Sec. 3

- ☐ See best-practices papers on simulation background and planning/setup [https://github.com/MobleyLab/basic_simulation_training]
- ☐ Are system timescales known experimentally and feasible computationally based on published literature? If timescales are too long for straight-ahead MD, is there an enhanced method for which there are precedents for systems of similar complexity?
- ☐ Read up on sampling assessment – from this article or another source (e.g., [15]). Understanding uncertainty will help in the *planning* of a simulation.
- ☐ Consider multiple runs instead of a single simulation. Multiple runs may be especially useful in assessing uncertainty for enhanced sampling methods.
- ☐ Check your code/method via a simple benchmark system. See: https://github.com/shirtsgroup/software-physical-validation

☐ **Perform simple, semi-quantitative checks which can rule out (but not ensure) sufficient sampling.** *Underlying concept:* It is easier to diagnose insufficient sampling than to demonstrate good sampling. See Sec. 4.

- ☐ Critically examine the time series of a number of observables – those of interest and then some more. Is each time series fluctuating about an average value or drifting overall? What states are expected and what are seen? Are there a significant number of transitions between states?
- ☐ If multiple runs have been performed, compare time series, distributions, and cluster populations from the different simulations.
- ☐ An individual trajectory can be divided into two parts and analyzed as if two simulations had been run.
- ☐ For individual runs, perform pairwise RMSD visual analysis as described below.

☐ **Remove an 'equilibration' (a.k.a. 'burn in', or transient) portion of a single MD or MC trajectory** and perform analyses only on remaining 'production' portion of trajectory. *Underlying concept:* An initial configuration is unlikely to be representative of the desired ensemble and the resulting relaxation process must be accounted for. See Sec. 5.

☐ **Consider computing a quantitative measure of global sampling, particularly for a biomolecular system** - i.e., attempt to estimate the number of statistically independent samples in a trajectory. *Underlying concept:* Trajectory 'frames' (configurations) are highly correlated because one frame is generated from the preceding one, and estimating the degree of correlation is essential to understanding overall simulation quality. See Sec. 6.

☐ **Quantify uncertainty in specific observables of interest using a confidence interval.** *Underlying concept:* The statistical uncertainty in the estimate of an observable's *average* decreases as more independent samples are obtained, and can be much smaller than the standard deviation – which measures the range of variation in the observable. See Sec. 7.

☐ **Use special care and uncertainty analyses for enhanced sampling methods.** *Underlying concept:* The use of multiple, potentially correlated trajectories within a single enhanced-sampling simulation can invalidate the assumptions underpinning traditional analyses of uncertainty. See Sec. 8.

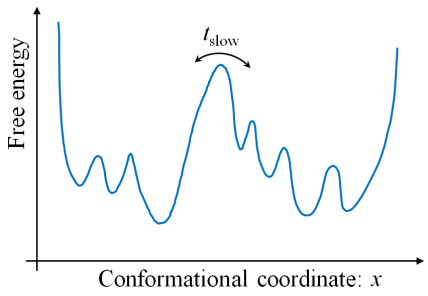

Figure 1. Schematic illustration of a free energy landscape dominated by a slow process. The timescales associated with a system will often reflect "activated" (energy-climbing) processes, although they could also indicate diffusion times for traversing a rough landscape with many small barriers. In the figure, the largest barrier is associated with the slowest timescale t_{slow}, and the danger for conventional MD simulations is that the total length of the simulation may be inadequate to generate the barrier crossing. The presence of stochasticity implies that even a simulation as long as t_{slow} may not yield the key event.

some subset of the simulation for which data is desired. For example, if one is only interested in the dynamics of a binding site in a protein, it probably is not necessary to observe the unfolding and refolding of that protein as well.

A toy model can generically illustrate timescales and their effects on sampling. Consider the double-well free energy landscape shown in Fig. 1, noting the slowest timescale is for crossing the largest barrier. Generally, you should expect that the value of *any* observable (e.g., x itself or another coordinate not shown or a function of those coordinates) will depend on which of the two dominant basins the system occupies; and, in turn, the equilibrium average of an observable will require sampling the two basins according to their equilibrium populations. In order to directly sample the equilibrium populations of the two basins, the length of a trajectory will have to be many times the slowest timescale, i.e. the largest barrier should be crossed multiple times. The relative populations of states are inferred from time spent in each state, which only will be representative if there are many transitions; put another way, the equilibrium populations follow from the transition rates [43] which can be estimated from multiple events. For completeness, we note that there is no guarantee that sampling of a given system will be limited by a dominant barrier. Instead, a system could exhibit a generally rough landscape with many pathways between states of interest. Nevertheless, the same cautions apply.

What should be done if a determination is made that a system's timescales are too long for direct simulation? The two main options would be to consider a more simplified ("coarse-grained") model or an enhanced sampling technique, bearing in mind that enhanced sampling methods are not foolproof but have their own limitations which should be considered carefully.

Lastly, whatever simulation protocol you pursue, be sure to use a well-validated piece of software [https://github.com/shirtsgroup/software-physical-validation]. If you are using your own code, check it against independent simulations on other software for a system that can be readily sampled.

4 Quick-and-Dirty checks that can rule out good sampling

It is difficult to establish with certainty that good sampling has been achieved, but it is not difficult to *rule out* high-quality sampling. Here we elaborate on some quick-and-dirty tests that quickly show inadequacies in sampling.

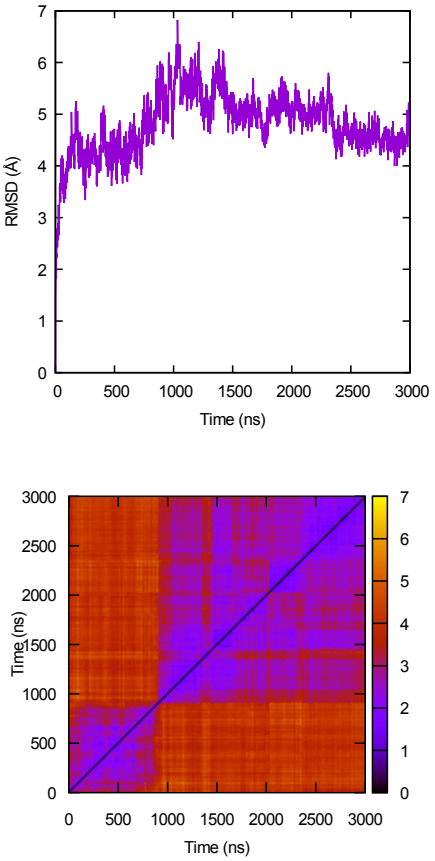

Figure 2. RMSD as a measure of convergence. The upper panel shows the α-carbon RMSD of the protein rhodopsin from its starting structure as a function of time. The lower panel shows the all-to-all RMSD map computed from the same trajectory. Data from Leioatts, et al [23].

4.1 Zeroth-order system-wide tests

The simplest test for poor sampling is lack of equilibration: if the system is still noticeably relaxing from its starting conformation, statistical sampling has not even begun, and thus by definition is poor. As a result, the very first test should be to verify that the basic equilibration has occurred. To check for this, one should inspect the time series for a number of simple scalar values, such as potential energy, system size (and area, if you are simulating a membrane or other system where one

dimension is distinct from the others), temperature (if you are simulating in the NVE ensemble), and/or density (if simulating in the isothermal-isobaric ensemble). Often, simple visual inspection is sufficient to determine that the simulation is systematically changing, although more sophisticated methods have been proposed by Chodera [7]. If *any* value appears to be systematically changing, then the system is not equilibrated.

4.2 Tests based on configurational distance measures - e.g., RMSD for biomolecules

We will use the standard biomolecular RMSD (root mean-squared difference) as a generic distance measure for illustrative purposes. Alternatives to RMSD could be a dihedral-angle distance or another measure specific to your system of interest. Note that RMSD, like any distance in a high-dimensional space, becomes "degenerate" for larger values: given a reference configuration, there are a large number of configurations which differ from the reference by a given large RMSD. This is analogous to the increasing number of points in three-dimensional space with increasing radial distance from a reference point, except much worse because of the dimensionality. For a detailed exploration of expected RMSD distributions for biomolecular systems see the work of Pitera.[28]

Some qualitative tools for assessing global sampling based on RMSD were reviewed in prior work [15]. The classic time series plot of RMSD with respect to a crystal or other single reference structure can immediately indicate whether the structure is still systematically changing. Although this kind of plot was historically used as a sampling test, it should really be considered as another equilibration test like those discussed above. Moreover, it's not even a particularly good test of equilibration, because the degeneracy of RMSD means you can't tell if the simulation is exploring new states that are equidistant from the chosen reference. The upper panel of Figure 2 shows a typical curve of this sort, taken from a simulation of the G protein-coupled receptor rhodopsin [23]; the curve increases rapidly over the few nanoseconds and then roughly plateaus. It is difficult to assign meaning to the other features on the curve.

A better RMSD-based convergence measure is the all-to-all RMSD plot; taking the RMSD of each snapshot in the trajectory with respect to all others allows you to use RMSD for what it does best, identifying very similar structures. The lower panel of Figure 2 shows an example of this kind of plot, applied to the same trajectory before. By definition, all such plots have values of zero along the diagonal, and occupation of a given state shows up as a block of similar RMSD along the diagonal; in this case, there are 2 main states, with one transition occuring roughly 800 ns into the trajectory. Off diagonal "peaks" (regions of low RMSD between structures sampled far apart in time) indicate that the system is revisiting previously sampled states, a necessary condition for good statistics. In this case, the initial state is never sampled after the first transition, but there are a number of small transitions within the second state.

4.3 Assessing Convergence

Convergence in the context of biomolecular simulations typically refers to the overlap of two independent measurements of the same property. If a measure is not converged it is a strong indication that sampling is poor. There are two common ways to try to obtain independent measurements. Arguably the best way is to have multiple independent simulations, each with varying initial starting conditions. Ideally these starting conditions should in some way span the space to be sampled; this way one can have confidence that simulations are not being trapped in a local minimum. For example, say the goal is to sample the phi and psi torsions of alanine dipeptide: the phi and psi angles for one simulation could be started from an alpha-helical conformation, while another simulation could be started from a polyproline II conformation. It is important to note that the starting conditions only need to be varied enough so that the desired space is sampled. For example, if the goal is to sample protein folding and unfolding, there should be some simulations started from the folded conformation and some from the unfolded, but if it is not important to consider protein folding no initial conformation needs to be unfolded.

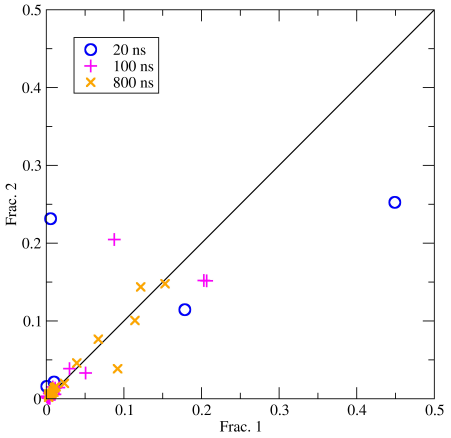

Figure 3. Combined clustering between two independent trajectories as a measure of convergence. The X axis is the population of a cluster from trajectory 1, while the Y axis is the population of that cluster from trajectory 2. Cluster populations are show after 20, 100, and 800 ns of sampling. The simulations used to generate the data used in this plot are described in Roe et al.[30]

Another way to obtain independent measurements is to divide a single simulation into two or more subsets. However this can at times be problematic because it can be more difficult to tell if the system is trapped in a local minimum

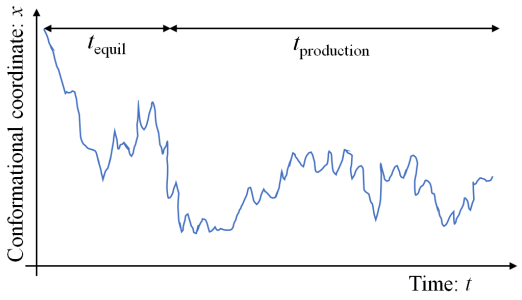

Figure 4. The equilibration and production segments of a trajectory. "Equilibration" over the time t_{equil} represents transient behavior while the initial configuration relaxes toward configurations more representative of the equilibrium ensemble. Readers are encouraged to select t_{equil} in a systematic way based on published literature. Note that strong sensitivity of "production" data to the choice of t_{equil} suggests additional sampling is required.

since there is a single starting point. Those employing this approach should take extra care to assess their results (see in particular sections on 'block averaging' below).

There are two simple ways to compare independent measurements of a property. The simplest is to compare the arithmetic means and standard deviations. If these values do not overlap then convergence has not been achieved. However, since this assumes that the values are normally distributed, a better way is compare the overlap of the probability distributions (i.e. the histograms). This can be done via Kullback-Leibler or Jensen-Shannon divergence.

4.3.1 Combined Clustering

One useful technique for evaluating convergence of structure populations is so-called "combined clustering". Briefly, in this method two or more independent trajectories are combined into a single trajectory, on which cluster analysis is performed. The resulting clusters are then split according to the trajectory they originally came from. If simulations are converged then each simulation will have the same population for any given cluster. Figure 3 shows results from combined clustering of two independent trajectories as a plot of cluster population fraction from the first trajectory compared to the second. If the two independent trajectories are perfectly converged then all points should fall on the X=Y line. As simulation time increases the cluster populations from the independent trajectories are in better agreement, which indicates the simulation is converging. For another example of performing combined cluster analysis see Bergonzo et al.[2]

5 Determining and removing an equilibration or 'burn-in' portion of a trajectory

The 'equilibration' or 'burn-in' time t_{equil} represents the initial part of a single continuous trajectory (whether from MD or MC) that is *discarded* for purposes of data analysis; the remaining trajectory data is often called 'production' data. See Fig. 4. Discarding data may seem counter-productive, but there is no reason to expect that the initial configurations of a trajectory will be important in the ensemble ultimately obtained. Including early-time data, therefore, can systematically *bias* results.

Examples can help to illustrate the idea. In the biomolecular realm, consider the initial configuration of a protein, which has been energy-minimized and otherwise relaxed from a crystal structure. Such an initial structure might seem to be intrinsically valuable, but remember that configurations important in a crystal environment may not be equally pertinent to an aqueous environment; also, any minimization/relaxation that has been performed likely will be very local and keep the configuration within or nearby the initial basin. See the RMSD trace in Fig. 2, and note the very rapid RMSD increase in the first ∼200 ns. In materials science, if a simulation is designed to study thermally induced fluctuations or defects in a (mostly) ordered system, an initial fully ordered configuration is not likely to be representative of the targeted ensemble.

Accepting that some data should be discarded, it is not hard to see that we want to avoid discarding too much data, given that many systems of interest are extremely expensive to simulate. Putting those two competing goals into statistical language, we want to remove bias but also minimize uncertainty (variance).

Before briefly describing a procedure for estimating the optimal equilibration time, it should be emphasized that the very notion of separating a trajectory into equilibration and production segments only makes sense if the system has indeed reached configurations important in the equilibrium ensemble. While it is generally impossible to guarantee this has occurred, some easy checks for determining that this has not occurred are described in Sec. 4. *It is essential to perform those basic checks before analyzing data with a more sophisticated approach* that may assume a trajectory has a substantial amount of true equilibrium sampling.

The essential ideas of equilibration analysis are discussed carefully in prior work [7, 40]. The key procedure is to analyze data as a function of the amount of data removed - i.e., as t_{equil} increases from zero. As described by Chodera [7], both the bias and the variance can be monitored, as well as the effective sample size, which is roughly quantified by the total simulation time considered divided by the auto-correlation time. Chodera indicates that the effective sample size peaks at the optimal t_{equil}, making the latter easy to discern. We caution that estimating the correlation time may require care, and readers may want to consider the global 'decorrelation time' [24] described in Sec. 6. Further, if values of observables estimated from the production phase depend sensi-

tively on the choice of t_{equil}, it is likely that further sampling is required.

6 Quantification of Global Sampling

With ideal trajectory data, one would hope to be able compute arbitrary observables with reasonably small error bars. During a simulation, it is not uncommon to monitor specific observables of interest, but after the data is obtained, it may prove necessary to compute observables not previously considered. These points motivate the goal of estimating global sampling quality, which can be framed most simply in the context of single-trajectory data: "Among the very large number of simulation frames (snapshots), how many are statistically independent?" From a dynamical perspective, which also applies to Monte Carlo data, how long must one wait before the system completely loses memory of its prior configuration? The methods noted in this section build on ideas already presented in the section on "quick-and-dirty" qualitative sampling analysis, but attempt to go a step further to quantify sampling quality.

We emphasize that *no single method described here has emerged as a clear best practice.* However, because the global assessment methods provide a powerful window into overall sampling quality – which could easily be masked in the analysis of single observables (Sec. 7) – we strongly encourage their use. The reader is encouraged to try one or more of the approaches in order to understand the limitations of their data.

6.1 Scope and a key caveat

The discussion here will focus largely on biomolecular systems, or more precisely, on systems for which it is straightforward to define a meaningful scalar distance between configurations.

A key caveat is needed before proceeding. Analysis of trajectory data generally cannot make inferences about parts of configuration space not visited [15]. It is generally impossible to know whether configurational states absent from a trajectory are appropriately absent because they are highly improbable (extremely high energy) or because the simulation simply failed to visit them because of a high barrier or bad luck.

6.2 Global sampling assessment for a single trajectory

Two methods applicable for a single trajectory have been previously introduced by some of the authors, exploiting the fact that trajectory correlations are sequential. That is, each configuration evolves from and is most similar to the immediately preceding configuration. This picture holds for standard MD and Markov-chain MC.

Lyman and Zuckerman proposed a global "decorrelation" analysis by mapping a trajectory to a discretization of configuration space and analyzing the resulting statistics [24]. Configuration space is discretized into bins based on Voronoi cells of structurally similar configurations - e.g., based on an RMSD criterion. The analysis method is based on the observation that the variance for any bin of a multinomial distribution is known, given a specified number of independent samples drawn from the discretized distribution. The knowledge of the expected variance allows testing of increasing waiting times between configurations drawn from the trajectory to determine when and if the variance approaches that expected for independent samples. The minimum waiting time yielding agreement with ideal statistics yields an estimate for the decorrelation/memory time, which implies an overall effective sample size.

A second method, employing block covariance analysis (BCOM), was presented by Romo and Grossfield [33] building on ideas by Hess [17]. In essence, the method combines two standard error analysis techniques, block averaging [12] and bootstrapping [11], with a quantitative assessment of the similarity of modes determined from principal component analysis, covariance overlap [17]. The principal components are computed from subsets of the trajectory, and the similarity of the modes evaluated as a function of subset size; as the subsets get larger, the resulting modes get more similar. This is done both for contiguous blocks of trajectory data (block averaging), and again for randomly chosen subsets of trajectory frames (bootstrapping); taking the ratio of the two values as a function of block size yields the degree of correlation in the data. Fitting that ratio to a sum of exponentials allows one to extract the relaxation times in the sampling. The key value of this method over others is that it implicitly takes into account the number of substates; the longest correlation time is the time required not to make a transition, but to sample a scattering of the relevant states. This method is implemented as part of LOOS [32, 34].

6.3 Global sampling assessment for multiple independent trajectories

When sampling is performed using multiple independent trajectories (whether MD or MC), additional care is required. Analyses based solely on the assumption of sequential correlations may break down because of the unknown relationship between separate trajectories.

Zhang et al. extended the decorrelation/variance analysis noted above, while still retaining the basic strategy of inferring sample size based on variance [42]. To enable assessment of multiple trajectories, the new approach focused

on conformational state populations, arguing that the states fundamentally underlie equilibrium observables. Employing a fairly simple kinetic-clustering technique to automatically define states, the approach then uses the variances in state populations among trajectories to estimate the effective sample size.

Nemec and Hoffmann proposed related sampling measures geared specifically for analyzing and comparing multiple trajectories [25]. These measures again do not require user input of specific observables but only a measure of the difference between conformations, which was taken to the be the RMSD. Nemec and Hoffmann provide formulas for quantifying the conformational overlap among trajectories (were the same configurational states sampled?) and the density agreement (were conformational regions sampled with equal probabilties?).

6.4 Global sampling assessment for enhanced sampling methods

The family of enhanced equilibrium sampling methods, including replica exchange and variants [36–38], metadynamics [6, 22], adaptive biasing force [8–10] among other methods, are complex and the resulting data may have a highly nontrivial correlation structure. In replica exchange, for example, the ensemble at a temperature of interest will be based on multiple return visits of different sequentially correlated trajectories.

Given the subtleties of these sampling approaches, we suggest taking a 'bottom line' approach, and assessing sampling based on multiple independent runs. The variance among these runs, if the approach is not biased, should be a measure of the overall sampling. Hence any method applicable to multiple trajectories should be valid for analyzing multiple runs of an arbitrary method. A caveat for the approach of Zhang et al. [42] is that some dynamics trajectory segments would be required to perform state construction by kinetic clustering.

7 Computing error in specific observables

7.1 Basics

"What error bar should I report?" Here we address this simple but critical question.

In general, there is no one-best practice for choosing error bars. However, in the context of simulations, we can nonetheless identify common goals when reporting such estimates: (i) to help authors and readers better understand uncertainty in data; and (ii) to provide readers with realistic information about the reproducibility of a given result.

With this in mind, we recommend the following: (a) in fields where there is a definitive standard for reporting uncertainty, the authors should follow existing conventions; (b) otherwise, such as for biomolecular simulations, *authors should report (and graph) their best estimates of 90% confidence intervals.* As explained in the glossary above, a 90% confidence interval is a range of values that is expected to bracket 90% of the computed predictions *if statistically equivalent simulations are repeated a large number of times;* (c) when feasible, consider plotting all of the points or a histogram instead of an average with error bars.

We emphasize that as opposed to standard uncertainties (reported as a standard deviation σ), confidence intervals have several practical benefits that justify their usage. In particular, they directly quantify the statistical frequency with which we expect a given outcome, which is more relatable to everyday experience than moments of a probability distribution. As such, confidence intervals can help authors and readers better understand the implications of an uncertainty analysis. Moreover, downstream users/readers of a given paper may include less statistically-oriented readers for whom confidence intervals are a more meaningful measure of variation.

In a related vein, error bars expressed in terms of $n\ \sigma$ can be misinterpreted as unrealistically under or overestimating uncertainty if taken at face value. For example, reporting 3 σ uncertainties for a normal random variable amounts to a 99.7 % confidence interval, which is likely to be a significant overestimate for many applications. On the other hand, 1 σ uncertainties only correspond to a 68 % confidence interval, which may be too low. Given that many readers may not take the time to make such conversions in their heads, we feel that it is safest for modelers to do this work up front.

In recommending 90 % confidence intervals, we are admittedly attempting to address a social issue that nevertheless has important implications for science as a whole. In particular, the authors of a study and the reputation of their field do not benefit in the long run by under-representing uncertainty, since this may lead to incorrect conclusions. But perhaps just as importantly, many of the same problems can arise if uncertainties are reported in a technically correct but obscure and difficult-to-interpret manner. For example, 1 σ error bars may not overlap and thereby mask the inability to statistically distinguish two quantities, since the corresponding confidence intervals are only 68 %. With this in mind, we therefore wish to emphasize that visual impressions conveyed by figures in a paper are of primary importance. Regardless of what a research paper may explain carefully in text, error bars on graphs create a lasting impression and must be as informative and accurate as possible. If 90% confidence intervals are reported, the expert reader can easily estimate the smaller standard uncertainty (especially if it is noted in the text), but showing a graph with overly small error bars is bound to mis-

Pressure (MPa)	Density (mol/L)
0.001	$3.007(3) \times 10^{-4}$
0.010	0.003 011(2)
0.100	0.030 39(16)
1.000	52.1(5)

Table 1. Density as a function of pressure for water at a temperature of 400K

lead most readers – even experts who do not search out the fine print.

[DMZ: SHOULD WE INCLUDE THIS (AND TABLE) GIVEN THAT IT SEEMS FIELD-SPECIFIC? IF SO, WE SHOULD BE SPECIFIC ABOUT WHICH FIELDS USE THIS CONVENTION - CERTAINLY NOT BIOMOLECULAR SIMULATION.] When reporting quantities with uncertainties, only the digits that are significant should be included and the standard error should be placed after the value in parenthesis as the uncertainty in the last digit. In general only one digit of uncertainty need be reported unless the digit is 1, in which case, it is helpful to report two digits (in order to avoid up to 50% roundoff error in the uncertainty). See the ***Table 1*** for examples from hypothetical isobaric simulations of water.

7.2 Overview of procedures for computing a confidence interval

We remind readers that before attempting to quantify uncertainty via an error bar, the "quick-and-dirty" checks on sampling should be performed. If the observable of interest is not fluctuating about a mean value but largely increasing or decreasing during the course of a simulation, it is unlikely that a reliable quantitative estimate for the observable (or an error bar) can be obtained.

For observables passing the qualitative tests noted above in Sec. 4, we advocate obtaining confidence intervals in one of two ways:

- For observables that are approximately Gaussian-distributed, an appropriately chosen *coverage factor k* (typically about 2, but see below) multiplying the standard uncertainty s^* yields an estimate for a 90% confidence interval.
- For non-Gaussian observables, a *bootstrapping* approach (described below) should be used. An example of a potentially non-Gaussian observable is a rate-constant, which must be positive but could exhibit significant variance, so a confidence interval estimated with a coverage factor may lead to an unphysical negative lower limit. Multimodal distributions certainly are not Gaussian. Bootstrapping does not assume an underlying distribution but instead constructs a confidence interval based on the recorded data values, and the limits cannot fall outside the extreme data values. Bootstrapping (or block averaging) is also necessary for cases where the observable isn't computed from a single frame, but rather from a collection of frames; this is true for quantities such as time correlation functions, and anything derived from them.

Below we describe approaches for estimating the standard uncertainty s^* and coverage factor k from a single trajectory, as well as discussing the bootstrapping approach for direct confidence-interval estimation.

Whether using a coverage factor and standard uncertainty or bootstrapping, one requires an estimate for the independent number of observations in a given simulation. This requires care, but may be accomplished based on the effective sample size described in Sec. 6 or via block averaging, described below. However, both methods have their limitations, and must be used with caution. Primarily, both require the user to look at the correct quantity in order to produce the correct answer. Block averaging will produce different effective sample sizes for different quantities; to produce robust answers, one must identify and track the slowest relevant degree of freedom in the system, which can be a non-trivial task. Even apparently fast-varying properties may have significant statistical error if they are coupled to slower varying ones, and this error in uncertainty estimation may not be readily identifiable by solely examining the fast-varying time series.

In the absence of a reliable estimate for the number of independent observations, one can perform n independent simulations and calculate the standard deviation σ among these, yielding a standard uncertainty of $s^* = \sigma/\sqrt{n}$. When computing the uncertainty with this approach, it is important to ensure that each starting configuration is also independent - or else to recognize and indicate in publication that the uncertainty refers to simulations started from a particular configuration. The means to obtain independent starting configurations is system-dependent, but might involve repeating the protocol used to construct a configuration (solvating a protein, inserting liquid molecules in a box, etc.), using a new random seed. However, readers are cautioned that *for complex systems, it may be effectively impossible to generate truly independent starting configurations pertinent to the ensemble of interest.* For example, a simulation of a protein in water will nearly always start from the experimental structure, which introduces some correlation in the resulting simulations even when the remaining simulation components (water, salt, etc) are regenerated de novo.

7.3 Assessing qualitative behavior of data

[DMZ: SINCE THIS IS A QUALITATIVE ISSUE, SHOULDN'T IT GO IN THE QUICK-AND-DIRTY SECTION?]

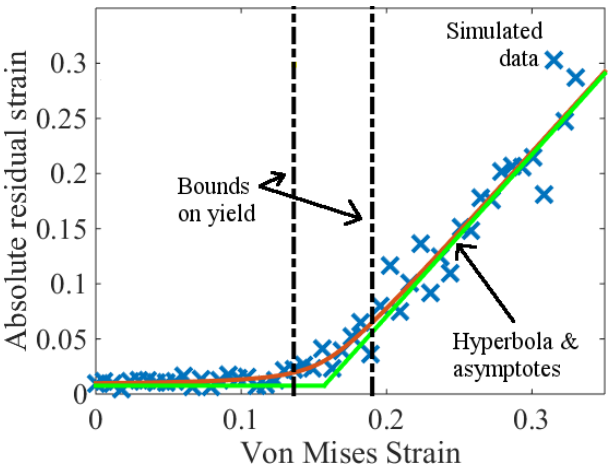

Figure 5. Yield strain tn ϵ_r as a function of applied strain ϵ. Blue $\times$ denote simulated data, whereas the smooth curve is a hyperbola fit to the data. The green lines are asymptotes; their intersection can be taken as an estimate of ϵ_y. Bounds on yield are computed by the synthetic data method discussed in the next section. *From, "Estimation and uncertainty quantification of yield via strain recovery simulations," P. Patrone, CAMX 2016 Conference Proceedings. Reprinted courtesy of the National Institute of Standards and Technology, U.S. Department of Commerce. Not copyrightable in the United States.*

Generally speaking, analysis routines that extract derived quantities from raw data are often formulated on the basis of physical intuition about how that data should behave. Before proceeding to data analysis, it is therefore useful to assess the extent to which raw data conforms to these expectations and the requirements imposed by either the modeler or the analysis routines. Such tasks help reduce subjectivity of predictions and offer insight into when a simulation protocol should be revisited to better understand their meaningfulness [26]. Unfortunately, general recipes for assessing data quality are impossible to formulate, owing to the range of physical quantities of interest to modelers. Nonetheless, a short example will help clarify the matter.

In the context of materials science, understanding when a structural material fails is critical for many engineering applications. In the past decade, scientists have increasingly focused on modeling the yield-strain ϵ_y, which is (loosely speaking) the amount of stretching (or strain) at which it deforms irreversibly. Intuition and experiments tells us that upon deforming a material by a fraction $1 + \epsilon$, it should recover its original dimensions if $\epsilon \leq \epsilon_y$ and have a residual strain $\epsilon_r = \epsilon - \epsilon_y$ if $\epsilon \geq \epsilon_y$ [27]. Owing to the time-scale limitations of MD, it is also reasonable to expect that the transition in ϵ_r around yield will be smooth and not piecewise linear. Thus, if we fit residual strain to a hyperbola, the proximity of data to the asymptotes illustrates the extent to which simulated ϵ_r conforms to the expectation that $\epsilon_r = 0$ when $\epsilon < \epsilon_y$. See Fig. 5.

7.4 Block averaging for estimating the standard uncertainty

The standard uncertainty of a simulation observable can be estimated from the fluctuations in the quantity along with the fact that (in the absence of correlation) the squared uncertainty will be inversely proportional to the number of samples. Block averages can be used instead of individual samples to avoid correlation in the samples while using all the data so long as the trajectory "blocks" (i.e., segments) are long enough to be essentially uncorrelated [12–15]. The uncertainty is expressed as

$$s^* = \frac{\sigma_{\text{blocks}}}{\sqrt{M}} \tag{5}$$

where σ_{blocks} is the standard deviation of the block averages and M is the number of *independent* blocks.

It is crucial when performing this analysis to do so systematically as a function of block size; as the blocks get longer, they should become independent and the s^* should plateau [12, 15]. Another approach is to measure the block correlation and to use it to improve the uncertainty estimate [21].

7.5 Autocorrelation method for estimating the standard uncertainty

In this section, describe the procedure for using an autocorrelation function to estimate the standard uncertainty in an observable.

$$s^* = \frac{\sigma_x}{\sqrt{N_{ind}}} \tag{6}$$

$$\hat{C}_x(\tau) = \frac{\langle [f(t) - \langle f \rangle]\,[f(t+\tau) - \langle f \rangle] \rangle}{\sigma_x^2} \tag{7}$$

$$N_{ind} = \frac{t_{run}}{2\int_0^\infty \hat{C}(\tau)\,d\tau} \tag{8}$$

$$N_{ind} = \frac{N_{MC}}{1 + 2\sum_{i=1}^{N_{lags}} \hat{C}(i)} \tag{9}$$

7.6 Propagation of uncertainty

The quantities we are most interested in may not be simulation observables. For instance, the free energy difference between two states might be measured by free energy perturbation, expressed as a function of the average of another quantity [39].

$$\beta\Delta A = -\ln \langle \exp(-\beta\Delta U) \rangle \tag{10}$$

Although $\exp(-\beta\Delta U)$ can be measured during the simulation and its uncertainty can be estimated directly using block averages as described above, $\beta\Delta A$ cannot be handled the same way. If we compute $\beta\Delta A$ for each block, the values will tend to take extremely positive whenever the perturbation does

poorly (where ΔU is consistently large). In the pathological case, ΔU might be ∞ for every sample in a block and the $\beta\Delta A = -\ln 0$ cannot be computed.

Instead of using block averages for $\beta\Delta A$, its uncertainty can be expressed as a first-order Taylor series expansion

$$\sigma_{\beta\Delta A} = \sigma_{\exp(\beta\Delta U)} / \langle \exp(-\beta\Delta U) \rangle \quad (11)$$

Propagation of uncertainty is needed whenever the derived quantity can be expressed as a function of other random observables. It might also be needed when the derived quantity is of a function of quantities measured in separate simulations, such as $< U(T_2) > - < U(T_1) >$. If a derived quantity is a function of multiple observables measured within a single simulation, then terms must be included to account for the correlation between those observables.

The Taylor series approach works well in most cases and easy to use, but does have limitations. Because this approach is based on a first-order Taylor series, propagation of uncertainty can fail in cases where a non-linear formula is used and the uncertainty is very large or the distribution of input averages is not Gaussian. For instance, the uncertainty in $\beta\Delta A$ as prescribed by Eq. 11 cannot exceed unity no matter how short the simulation is or how bad the sampling is. If there is doubt as to the quality of the computed uncertainty, the uncertainty can be computed with alternative approaches such as bootstrapping to verify the Taylor series results or to identify an alternative approach that works better.

7.7 From standard uncertainty to confidence interval for Gaussian variables

Once a standard uncertainty value is obtained for a Gaussian-distributed random variable with mean $\langle x \rangle$, and the number of independent samples n has been estimated, the 90%-confidence interval ($\langle x \rangle - k\,s^*$, $\langle x \rangle + k\,s^*$) can be constructed on the basis of an established look-up table for the coverage factor k based on n. The theoretical basis for the table is the "Student" or "t" distribution, which is *not* a Gaussian, but governs the behavior of an *average* derived from n independent Gaussian variables [19]. See Table 2.

When $n \leq 5$ [SHOULD THIS BE 10?], we recommend showing all data points, as a confidence interval is likely not statistically meaningful.

As a reminder, multi-modally distributed variables with multiple peaks in their distributions cannot be considered Gaussian random variables. Variables with a strict upper or lower limit (such as a positive-definite quantity) and long-tailed distributions likely are not Gaussian. These cases should be treated with bootstrapping.

n (independent samples)	k (coverage factor)
6	2.02
11	1.81
16	1.75
21	1.72
26	1.71
51	1.68
101	1.66

Table 2. Coverage factors k required for a 90% confidence interval for a Gaussian variable [19].

7.8 Bootstrapping

Bootstrapping is an approach to uncertainty estimation that does not assume a particular distribution for the observable of interest or a particular kind of relationship between the observable and variables directly obtained from simulation [11]. In nonparametric bootsrapping, new, "synthetic" data sets (corresponding to hypothetical simulation runs) are created by drawing n samples (configurations) from the original collection that was generated during the actual run. The same sample may be selected twice, while others may not be selected at all in a process called "sampling with replacement." In doing so, these synthetic sets will be different even though they all have the same number of samples and draw from the same pool of data. Having created a new set, the data is analyzed to determine the derived quantity of interest, and this process is repeated to produce multiple estimates of the quantity. The distribution of 'synthetic' observables can be directly used to construct a 90% confidence interval from the 5%ile to the 95%ile value.

As with bootstrapping, the key issue is how many samples to draw from the original collection - what value of n should be used? It should be clear that a larger n will yield less variation and hence a tighter confidence interval: implicitly all n samples will be regarded as independent. Hence, n should be chosen to represent the number of *independent* samples present in the original simulation, which can be gauged using a correlation-time method [7, 24] or implicitly using a block-averaging procedure – see above.

The process described above assumes that the original simulation data is uncorrelated. If this is not the case, then the resampling method can be reformulated in one of two ways. The first option is to estimate the number of independent samples in the original set and to pull only that many samples to create the new data sets. The second option is to group the samples into blocks that are uncorrelated and to then use the block averages as the samples for bootstrapping.

Alternatively, one could use the difference between errors estimated via block averaging and bootstrapping as a

measure of the correlation; if one tracks the bootstrapped and block averaged estimates of a quantity's uncertainty as a function of block size, the only difference between the two modes of calculation is whether the data is correlated. The decay in the ratio of the two quantities as a function of time is a measure of the correlation time in the sample [33].

An alternate approach that can directly account for correlations is called parametric bootstrapping. The main idea behind this method is to model the original data as a deterministic function (which can be zero, constant, or have free parameters) plus additive noise. The parameters of this model, including the structure of the noise (i.e. its covariance), can be determined through a statistical inference procedure. Having calibrated the model, random number generators can be used to sample the noise, which is then added back to the trial function to generate a synthetic data set. As with the nonparametric bootstrap, the generated data can be used to compute the derived quantity of interest, and the uncertainty can be obtained from the statistics of the values compute with different generated sets.

To further clarify the procedure of parametric boostrapping, consider the simplest case in which the data is a collection of uncorrelated random variables fluctuating about a constant mean. In this situation, one could estimate (I) the deterministic part of a parametric model using the sample mean μ of the data, and (II) the stochastic part as a Gaussian random variable whose variance equals the sample variance. If instead the data are correlated (e.g. as in a time-series of simulated observables), one can postulate a covariance function to describe the structure of this randomness. Often these covariance functions are formulated with free parameters (often called "hyperparameters") that characterize properties such as the noise-scale and characteristic length of correlations [29]. In such cases, determining the hyperparameters may require more sophisticated techinques such as maximum likelihood analyses or Bayesian approaches; see, for example, Ref. [29]. See also Refs. [4, 5, 26, 27] for examples and practical implementations applied to cases in which the determinsitic component of the data is not constant.

It is important to note that various bootstrapping approaches can and often are used as uncertainty propagation tools. Nonetheless, care should be exercised when using such methods with nonlinear functions. In the free energy example, setting $\langle \exp(-\beta\delta U)\rangle = 1 \pm 0.5$ and generating new estimates from a Gaussian centered at 1 with a width of 0.5 will eventually output negative numbers, which is mathematically nonsensical and problematic for any function that takes strictly non-negative inputs. Thus, one should be aware of any distributional assumptions imposed either by the physics of the problem or the analyses of synthetic data.

7.9 Dark uncertainty analyses

In some cases, multiple simulations of the same physical observable τ may yield predictions whose error bars do not overlap. This situation can arise, for example, in simulations of the glass transition temperature when undersampling the crosslinked network structure of certain polymers. In such cases, it is reasonable to postulate an unaccounted for source of uncertainty, which we colorfully refer to as "dark uncertainty." In the context of a statistical model, we postulate that the probability of a simulation output depends on the unobserved or "true" mean value $\bar{\tau}$, an uncertainty σ_i^2 whose value is specific to the simulation (estimated, e.g. according to uncertainty propagation), and the unaccounted-for dark uncertinty y^2. (For simplicity, the σ_i^2 and y^2 should be treated as variances.)

While details are beyond this scope of this document, such a model motivates an estimate of $\bar{\tau}$ of the form

$$\bar{\tau} \approx \mathcal{T} \propto \sum_i \frac{T_i}{\sigma_i^2 + y^2}, \tag{12}$$

where T_i is the prediction from the ith simulation, σ_i^2 is its associated "within-simulation" uncertainty, and y^2 is the dark or between-simulation uncertainty; note that the latter does not depend on i. The variable y^2 can be estimated from a maximum-likelihood analysis of the data and amounts to numerically solving a relatively simple nonlinear equation (see Ref. [26]). Equation (12) is useful insofar as it weights simulated results according to their certainty while reducing the impact of overconfident predictions (e.g. having small σ_i^2). Additional details on this method are provided in Ref. [26] and the references contained therein.

- Basics: how to report, what goal to shoot for, significant figures - When should you not trust uncertainties - Unknown unknowns - If calculating a derived quantity, consider error in conversion from raw data - Propagation of error (this doesn't mean publishing your results!) - Taylor series expansion can handle cases where derived quantity is a direct function of measured data - Wikipedia: Propagation of uncertainty - Bootstrapping [Andrew/Dan S] used for cases where the derived quantity is not simple function of the measured data. An Introduction to the Bootstrap - Need to know correlation to correctly estimate sample size - Otherwise just gives relative uncertainty - Correlation time analysis [Dan Z] - Block averaging [Dan Z/Alan?/Dan S] Flyvjberg and Petersen - MOST OF THESE ALGORITHMS FAIL IF THE TRAJECTORY IS WAY TOO SHORT - If you miss the timescale by enough, you can't tell - YOU HAVE TO THINK ABOUT THIS IN ADVANCE - Link out to transport doc

8 Assessing Uncertainty in Enhanced Sampling Simulations

While recent advances in computational hardware have allowed MD simulations of systems with biological relevance to routinely reach timescales ranging from hundreds of ns to µs, in many cases this is still not long enough to obtain equilibrated (i.e. Boltzmann-weighted) structural populations. Intrinsic timescales of the systems may be much longer. Enhanced sampling methods can be used to obtain well-converged ensembles faster than conventional MD. In general, enhanced sampling methods work through a combination of modifying the underlying energy landscape and/or thermodynamics parameters to increase the rate at which energy barriers are crossed along with some form of reweighting to recover the unbiased ensemble [43]. However, such methods do not guarantee a converged ensemble, and care must be taken when using and evaluating enhanced sampling methods.

Generally speaking, uncertainty analysis is more challenging for data generated by an enhanced sampling method. Before performing such a simulation, consider carefully whether the technique is needed, and consult the literature for best practices in setting up a simulation. Even a straightforward MD simulation requires considerable planning, and the complexity is much greater for enhanced techniques.

8.1 Replica Exchange Molecular Dynamics

One of the most popular enhanced sampling methods is replica exchange MD (REMD) [37]; see also [38]. Broadly speaking, REMD consists of running MD simulations a number of non-interacting replicas of a system, each with a different Hamiltonian and/or thermodynamics parameters (e.g., temperature), and periodically exchanging system coordinates between replicas according to a Metropolis criterion.

In order to assess the results of a REMD simulation, it is important to consider not just the overall convergence of the simulation to the correct Boltzmann-weighted ensemble of structures (via combined clustering, combined PC projection overlap analysis, etc.), but how efficiently the REMD simulation is doing so. These concepts are termed "thermodynamic efficiency" and "mixing efficiency" by Abraham and Gready,[1] and it is quite possible to achieve one without the other; both must be assessed. In order for sampling to be efficient, coordinates must be able to move freely in replica space.

Below we will refer to both "coordinate trajectories" and "replica trajectories". A "coordinate trajectory" follows an individual system's continuous trajectory as it traverses replica space (e.g., a system experiencing multiple temperatures as it is exchanged during a temperature REMD simulation). A "replica trajectory" is the sequence of configurations corresponding to a single replica under fixed Hamiltonian and thermodynamic conditions, (e.g., all structures at a temperature of 300 K in a temperature REMD simulation). Thus, a replica trajectory consists of concatenated coordinate-trajectory segments and *vice versa*.

- Exchange acceptance. The exchange acceptance rate (i.e. the number of exchanges divided by the number of exchange attempts) between neighboring replicas should be roughly equivalent to each other and to the target acceptance rate. If the exchange acceptance is too low then replica spacing may need to be decreased or additional replicas used, and vice versa if the exchange acceptance is too high.
- Replica round-trips. The time taken for a coordinate trajectory to travel from the lowest replica to the highest and back is called the replica "round trip" time. Over the course of a REMD simulation, any given coordinate trajectory should make multiple round trips. In addition one can look at the average, minimum, and maximum round trip times: these should be roughly equivalent for any given set of coordinates. See e.g. Figure 6 in [30].
- Replica residence time. The time coordinate trajectory spends at a replica is called the "replica residence time". For replica sampling to be efficient, the average replica residence time for each set of starting coordinates at each replica should be roughly equivalent. See e.g. Figure 7 in Roe et al..[30]
- Distributions of quantities calculated from coordinate trajectories. If all coordinates are moving freely in replica space, they should eventually converge to the same ensemble of structures. Comparing distributions of various quantities from coordinate trajectories can provide a measure of how converged the simulation is. For example, one can compare the distribution of RMSD values of coordinate trajectories to a common reference structure; see e.g. Figure 8 in Henriksen et al..[16] Poor overlap can be an indication that replica efficiency is poor or the simulation is not yet converged.

All of the above quantities (replica residence time, round trip time, lifetimes etc) can be calculated with CPPTRAJ,[31] which is freely available from https://github.com/Amber-MD/cpptraj or as part of AmberTools (http://ambermd.org).

It may also be useful to perform multiple REMD runs. Using the standard uncertainty among runs can quantify uncertainty and provide the basis for a confidence interval with an appropriate coverage factor - see definitions in Sec. 1. If the ensembles produced depend significantly on the set of starting configurations, that is a sign of incomplete sampling.

8.2 Weighted Ensemble simulations

The weighted ensemble (WE) method orchestrates an ensemble of trajectories that are intermittently pruned or replicated in order to enhance sampling of difficult-to-access regions of configuration space [18]. The final set of trajectories can be visualized as a tree structure based on the occasional replication and pruning events. WE is an unbiased method that can be used to sample rare transient behavior [41] as well as steady states [3] including equilibrium [35].

Like other enhanced sampling methods, WE's tree of trajectories has a complex correlation structure requiring care for uncertainty analysis. It is important to understand the basic theory and limitations of the WE method, as is discussed in this overview document.

From a practical standpoint, the safest way to assess uncertainty in WE simulations is to run multiple instances (ideally set up in the same way) from which a variance and standard uncertainty in any observable can be calculated; see definitions. Note particularly that WE tracks the time evolution of observables as the system relaxes (perhaps quite slowly) to equilibrium or another steady state [41]; hence, the variance computed in an observable from multiple runs should be based on values at the same time point.

When it is necessary to estimate uncertainty based on a single WE run, the user should treat the (ensemble-weighted) value of an observable measured over time much like an observable in a standard single MD simulation; this is because the correlations in ensemble averages are sequential in time. First, as discussed in Sec. 4, the time trace of the observable should be inspected for relaxation to a nearly constant value about which fluctuations occur. A transient/equilibration period should be removed in analogy to MD - see Sec. 5 - and then best practices for single obervable uncertainties should be followed as described in Sec. 7. Despite this rather neat analogy to conventional MD, experience has shown that run-to-run variance in WE simulations can be surprisingly large, so multiple runs are strongly advised.

9 Acknowledgments

The authors appreciate helpful discussions with Pascal T. Merz. DMZ acknowledges support from NIH Grant GM115805.

References

[1] **Abraham MJ**, Gready JE. Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations. Journal of Chemical Theory and Computation. 2008; 4(7):1119–1128. http://dx.doi.org/10.1021/ct800016r, doi: 10.1021/ct800016r, pMID: 26636365.

[2] **Bergonzo C**, Henriksen NM, Roe DR, Swails JM, Roitberg AE, Cheatham TE. Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide. Journal of Chemical Theory and Computation. 2014; 10(1):492–499. http://dx.doi.org/10.1021/ct400862k, doi: 10.1021/ct400862k, pMID: 24453949.

[3] **Bhatt D**, Zhang BW, Zuckerman DM. Steady state via weighted ensemble path sampling. Journal of Chemical Physics. 2010; 133:14110.

[4] **Boettinger WJ**, Williams ME, Moon KW, McFadden GB, Patrone PN, Perepezko JH. Interdiffusion in the Ni-Re System: Evaluation of Uncertainties. Journal of Phase Equilibria and Diffusion. 2017 Oct; 38(5):750–763. https://doi.org/10.1007/s11669-017-0562-7, doi: 10.1007/s11669-017-0562-7.

[5] **Boettinger WJ**, Williams ME, Moon KW, McFadden GB, Patrone PN, Perepezko JH. Interdiffusion in the Ni-Re System: Evaluation of Uncertainties. Journal of Phase Equilibria and Diffusion. 2017 Oct; 38(5):750–763.

[6] **Bussi G**, Laio A, Parrinello M. Equilibrium free energies from nonequilibrium metadynamics. Phys Rev Lett. 2006 mar; 96(9):90601.

[7] **Chodera JD**. A simple method for automated equilibration detection in molecular simulations. Journal of chemical theory and computation. 2016; 12(4):1799–1805.

[8] **Comer J**, Gumbart JC, Hénin J, Lelievre T, Pohorille A, Chipot C. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. Journal of Physical Chemistry B. 2015; 119(3):1129–1151. doi: 10.1021/jp506633n.

[9] **Darve E**, Pohorille A. Calculating free energies using average force. The Journal of Chemical Physics. 2001; 115(20):9169–9183. http://aip.scitation.org/doi/10.1063/1.1410978, doi: 10.1063/1.1410978.

[10] **Darve E**, Rodríguez-Gómez D, Pohorille A. Adaptive biasing force method for scalar and vector free energy calculations. Journal of Chemical Physics. 2008; 128(14). doi: 10.1063/1.2829861.

[11] **Efron B**, Tibshirani RJ. An introduction to the bootstrap. Boca Raton: Chapman and Hall/CRC; 1998.

[12] **Flyvbjerg H**, Petersen HG. Error estimates on averages of correlated data. J Chem Phys. 1989; 91:461–466.

[13] **Frenkel D**, Smit B. Understand Molecular Simulation: From Algorithms to Applications. New York: Academic Press; 2002.

[14] **Friedberg R**, Cameron JE. Test of the Monte Carlo Method: Fast Simulation of a Small Ising Lattice. The Journal of Chemical Physics. 1970; 52(12):6049–6058. doi: 10.1063/1.1672907.

[15] **Grossfield A**, Zuckerman DM. Quantifying uncertainty and sampling quality in biomolecular simulations. Annu Rep Comput Chem. 2009 jan; 5:23–48. http://dx.doi.org/10.1016/S1574-1400(09)00502-7, doi: 10.1016/S1574-1400(09)00502-7.

[16] **Henriksen NM**, Roe DR, Cheatham TE. Reliable Oligonucleotide Conformational Ensemble Generation in Explicit Solvent for Force Field Assessment Using Reservoir Replica Exchange Molecular Dynamics Simulations. The Journal of Physical Chemistry B. 2013; 117(15):4014–4027. http://dx.doi.org/10.1021/jp400530e, doi: 10.1021/jp400530e, pMID: 23477537.

[17] **Hess B**. Convergence of sampling in protein simulations. Physical Review E. 2002; 65(3):31910.

[18] **Huber GA**, Kim S. Weighted-ensemble {B}rownian dynamics simulations for protein association reactions. Biophys J. 1996; 70:97–110.

[19] **JCGM**. JCGM 100: Evaluation of measurement data - Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology; 2008.

[20] **JCGM**. JCGM 200: International vocabulary of metrology - Basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology; 2012.

[21] **Kolafa J**. Autocorrelations and subseries averages in Monte Carlo Simulations. Molecular Physics. 1986; 59(5):1035–1042. doi: 10.1080/00268978600102561.

[22] **Laio A**, Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Reports on Progress in Physics. 2008; 71(12):126601. http://stacks.iop.org/0034-4885/71/i=12/a=126601?key=crossref.2dcba907622222368898e968e8d42594, doi: 10.1088/0034-4885/71/12/126601.

[23] **Leioatts N**, Romo TD, Danial SA, Grossfield A. Retinal Conformation Changes Rhodopsin's Dynamic Ensemble. Biophys J. 2015 Aug; 109(3):608–617. http://dx.doi.org/10.1016/j.bpj.2015.06.046, doi: 10.1016/j.bpj.2015.06.046.

[24] **Lyman E**, Zuckerman DM. On the Structural Convergence of Biomolecular Simulations by Determination of the Effective Sample Size. J Phys Chem B. 2007; 111(44):12876–12882.

[25] **Nemec M**, Hoffmann D. Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems. Journal of Chemical Theory and Computation. 2017; 13(2):400–414. doi: 10.1021/acs.jctc.6b00823.

[26] **Patrone PN**, Dienstfrey A, Browning AR, Tucker S, Christensen S. Uncertainty quantification in molecular dynamics studies of the glass transition temperature. Polymer. 2016; 87:246 – 259. http://www.sciencedirect.com/science/article/pii/S003238611630074X, doi: https://doi.org/10.1016/j.polymer.2016.01.074.

[27] **Patrone PN**, Tucker S, Dienstfrey A. Estimating yield-strain via deformation-recovery simulations. Polymer. 2017; 116(Supplement C):295 – 303. http://www.sciencedirect.com/science/article/pii/S0032386117303117, doi: https://doi.org/10.1016/j.polymer.2017.03.046.

[28] **Pitera JW**. Expected Distributions of Root-Mean-Square Positional Deviations in Proteins. The Journal of Physical Chemistry B. 2014; 118(24):6526–6530. http://dx.doi.org/10.1021/jp412776d, doi: 10.1021/jp412776d, pMID: 24655018.

[29] **Rasmussen CE**, Williams CKI. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press; 2005.

[30] **Roe DR**, Bergonzo C, Cheatham TE. Evaluation of Enhanced Sampling Provided by Accelerated Molecular Dynamics with Hamiltonian Replica Exchange Methods. The Journal of Physical Chemistry B. 2014; 118(13):3543–3552. http://dx.doi.org/10.1021/jp4125099, doi: 10.1021/jp4125099, pMID: 24625009.

[31] **Roe DR**, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. Journal of Chemical Theory and Computation. 2013; 9(7):3084–3095. http://dx.doi.org/10.1021/ct400341p, doi: 10.1021/ct400341p, pMID: 26583988.

[32] **Romo TD**, Grossfield A. LOOS: A lightweight object-oriented software library. . 2017; Version 2.3.2. http://loos.sourceforge.net,https://github.com/GrossfieldLab/loos.

[33] **Romo TD**, Grossfield A. Block covariance overlap method and convergence in molecular dynamics simulation. Journal of Chemical Theory and Computation. 2011; 7(8):2464–2472. doi: 10.1021/ct2002754.

[34] **Romo TD**, Leioatts N, Grossfield A. Lightweight object oriented structure analysis: tools for building tools to analyze molecular dynamics simulations. J Comput Chem. 2014 Dec; 35(32):2305–2318. http://dx.doi.org/10.1002/jcc.23753, doi: 10.1002/jcc.23753.

[35] **Suárez E**, Lettieri S, Zwier MC, Stringer CA, Subramanian SR, Chong LT, Zuckerman DM. Simultaneous Computation of Dynamical and Equilibrium Information Using a Weighted Ensemble of Trajectories. J Chem Theory Comput. 2014 jul; 10(7):2658–2667. http://dx.doi.org/10.1021/ct401065r, doi: 10.1021/ct401065r.

[36] **Sugita Y**, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. J Chem Phys. 2000; 113:6042–6051.

[37] **Sugita Y**, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters. 1999; 314(1):141 – 151. http://www.sciencedirect.com/science/article/pii/S0009261499011239, doi: https://doi.org/10.1016/S0009-2614(99)01123-9.

[38] **Swendsen RH**, Wang JS. Replica {M}onte {C}arlo Simulation of Spin-Glasses. Phys Rev Lett. 1986; 57:2607–2609.

[39] **Taylor JR**. An Introduction to Error Analysis. Sausalito, California: University Science Books; 1997.

[40] **Yang W**, Bitetti-Putzer R, Karplus M. Free energy simulations: Use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence. Journal of Chemical Physics. 2004; 120(6):2618–2628. doi: 10.1063/1.1638996.

[41] **Zhang BW**, Jasnow D, Zuckerman DM. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. J Chem Phys. 2010 feb; 132(5):54107. http://dx.doi.org/10.1063/1.3306345, doi: 10.1063/1.3306345.

[42] **Zhang X**, Bhatt D, Zuckerman DM. Automated sampling assessment for molecular simulations using the effective sample size. Journal of Chemical Theory and Computation. 2010; 6:3048–3057.

[43] **Zuckerman DM**. Equilibrium sampling in biomolecular simulations. Annu Rev Biophys. 2011 jun; 40:41–62. http://dx.doi.org/10.1146/annurev-biophys-042910-155255, doi: 10.1146/annurev-biophys-042910-155255.