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 February 2, 2018

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

Simulating molecular systems that are interesting by today's standards, whether for biomolecular study, materials science, or a related field, is a challenging task. More ambitious simulations are inevitably performed every year, but even systems which would be considered simple and far from the cutting edge, such as short alkanes in liquid phase, provide challenges [45]. In addition to the various system-specific issues that modelers must address, questions often arise concerning the best way to adequately sample the desired phase-space or estimate uncertainties. While these latter questions are not unique to molecular modeling, their importance cannot be overstated: the usefulness of a simulated result ultimately hinges on being able to confidently and accurately report

uncertainties along with any given prediction. In the context of techniques such as molecular dynamics (MD) and Monte Carlo (MC), these considerations are especially important, given that even large-scale modern computing resources are no guarantee of adequate sampling.

This article therefore aims to provide best-practices for reporting simulated observables, assessing confidence in simulations, and deriving uncertainty estimates (more colloquially, "error bars") based on a variety of statistical techniques applicable to physics-based sampling methods and their associated "enhanced" counterparts. As a general rule, we advocate a tiered approach to computational modeling: workflows should begin with back-of-the-envelope calculations to determine the feasibility of a given computation, followed by the actual simulation(s). Semi-quantitative checks can then be used to check for adequate sampling and assess the quality of data. Only once these steps have been performed should one actually construct estimates of observables and uncertainties. In this way, modelers avoid unncessary waste by continuously gauging the likelihood that subsequent steps will be successful. Moreover, this approach can help to identify seemingly reasonable data that may have little value for prediction and/or be the result of a poorly run simulation.

It is worth emphasizing that in the last few years, many works have developed and advocated for uncertainty quantification (UQ) methods not traditionally used in the MD and MC communities. In some cases, these methods buck trends that have become longstanding conventions in certain communities, e.g. the practice of only using uncorrelated data to construct statistical estimates. One goal of this manuscript is therefore to advocate newer UQ methods when these are demonstrably better. Along these lines, we wish to remind the reader that better results are not only obtained from faster computers, but also by using data more thoughtfully.

In this vein, the reader should be aware that there is not a "one-size-fits-all" approach to UQ. Ultimately, we take the perspective that uncertainty quantification in its broadest sense aims to provide actionable information for making decisions, e.g. in an industrial research and development setting or in planning future academic studies. A simulation protocol and subsequent analysis of its results should therefore take into account the intended audience and/or decisions to be made on the basis of the computation. In some cases, quick-and-dirty workflows can indeed be useful if the goal is to only provide order-of-magnitude estimates of some quantity. We also note that uncertainties can often be estimated through a variety of techniques, and there may not be consensus as to which, if any, are best. *Thus, a critical component of any UQ analysis is communication of the very UQ method, e.g. of the assumptions being made, the tools used, and the way that results are interpreted.* Educated decisions can only be made through an understanding of both the process of estimating uncertainty and its numerical results.

While UQ is a central topic of this manuscript, our scope is limited to issues associated with sampling and related uncertainty estimates. We do not address systematic errors arising from inaccuracy of force-fields, the underlying model, or parametric choices such as the choice of a thermostat time-constant. See, for example, Refs. [25, 37–39] for methods that address such problems. Moreover, we do not consider model-form error and related issues that arise when comparing simulated predictions with experiment. Rather, we take the raw trajectory data at face value, assuming that it is a valid description of the system of interest.[1]

1.2 Key Definitions

In order to make the discussion that follows more precise, we first 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)[21], 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)[20].

For clarity, we highlight differences between conventional terms and the VIM usage employed throughout this article. Readers should study the term "standard uncertainty" which is sometimes estimated by (in common parlance) the "standard error of the mean"; however, the VIM term for the latter is the "experimental standard deviation of the mean." In cases of lexical ambiguity, the reader should assume that we hold to the definition of terms as given in the VIM.

The glossary is presented in a logical, rather than alphabetical order. We strongly encourage reading through the full glossary because of its structure and potentially unfamiliar terminology. Importantly, we also recommend reading the discussion that immediately follows, since this (i) explains the rationale for adopting the chosen language, (ii) discusses the limited relationship between statistics and uncertainty quantification, and (iii) thereby clarifies our perspective on best-practices.

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[2].

[1] In more technical UQ language, we restrict our scope to *verification* of simulation results, as opposed to *validation*.

[2] Most molecular simulations (even those using pseudo-random number

- **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[20, 21].
- **Expectation value**: If $P(x)$ is the probability density of a continuous random quantity x, then the expectation value is given by the formula

$$\langle x \rangle = \int \mathrm{d}x\, P(x) x. \tag{1}$$

  In the case that x adopts discrete values $x_1, x_2, \ldots$, we instead write

$$\langle x \rangle = \sum_i x_i P(x_i). \tag{2}$$

- **Variance**:[3] Taking $P(x)$ as defined previously, the variance of a random quantity is a measure of how much it can fluctuate, given by the formula

$$\sigma_x^2 = \int \mathrm{d}x\, P(x) \left(x - \langle x \rangle\right)^2. \tag{3}$$

  If x assume discrete values, the coresponding definition becomes

$$\sigma_x^2 = \sum_i P(x_i) \left(x_i - \langle x \rangle\right)^2. \tag{4}$$

- **Standard Deviation**: The positive square root of the variance, denoted σ_x.
- **Arithmetic mean**: An *estimate* of the (true) expectation value of a random quantity, given by the formula

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{5}$$

  where x_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." Note that a proper realization of a random variable (with no systematic bias) will yield values distributed according to $P(x)$, so $\bar{x} \to \langle x \rangle$ as $n \to \infty$.
- **Standard Uncertainty**: Uncertainty in a result (e.g. estimation of a true value) as expressed in terms of a standard deviation[4].
- **Experimental standard deviation**[5] : An *estimate* of the (true) standard deviation of a random variable, given by the formula[6]

$$s(x) = \sqrt{\frac{\sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2}{n-1}} \tag{6}$$

  Remark: This quantity is often called the "sample standard deviation." Additionally, $s(x)$ is a statistical property of the specific set of observations $\{x_1, x_2, \ldots, x_n\}$, not of the random quantity x in general. Thus, $s(x)$ is sometimes written as $s\left(x_j\right)$ for emphasis of this property.
- **Linearly Uncorrelated observables**: If quantities x and y have mean values $\langle x \rangle$ and $\langle x \rangle$, then x and y are linearly uncorrelated if

$$\langle \left(x - \langle x \rangle\right) \left(y - \langle y \rangle\right) \rangle = 0 \tag{7}$$

  Remark: The concepts of linear uncorrelation and independence of random variables are often conflated. The latter amounts to the statement that the joint density of two random variables x and y can be decomposed as $P(x,y) = P(x)P(y)$, which is a stronger condition than linear uncorrelation. Empirically testing for independence, however, is not practical, nor is it necessary for any of the estimates discussed in this work.
- **Experimental standard deviation of the mean**: An estimate of the standard deviation of the distribution of the arithmetic mean, given by the formula

$$s(\bar{x}) = \frac{s(x)}{\sqrt{n}}, \tag{8}$$

  where the realizations of x_j are assumed to be linearly uncorrelated.

  Remark: This quantity is often called the "standard error."
- **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 "nontrivial" analyses of raw data, e.g. properties that may not be computed for a single configuration such as free energies.

generators) are deterministic in that the sequence of visited states is generated by a fixed algorithm. 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. [25] for more discussion of this rather deep point.

[3] The true probability density $P(x)$ is inherently unknowable, given that we can only collect a finite amount of data about x. As such, we can only estimate the expectation value of x and its variance.

[4] 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. Intuitively, this quantity should reflect the degree to which an estimate would vary if recomputed using new and independent data.

[5] The term "experimental" can refer to simulated data, since these are the results of numerical experiments.

[6] The factor of $n - 1$ (as opposed to n) appearing in the demonimator of Eq. 6 is needed to ensure that the estimate is *unbiased*, meaning that on average $s(x)^2$ is equal to the true variance. Physically, we can interpret the –1 as accounting for the fact that one degree-of-freedom (e.g. piece of data) is lost via the appearance of $\bar{x}$ in the definition of $s(x)$. Equivalently, it accounts for the fact that the arithmetic mean is linearly correlated with each x_j (cf. Linearly Uncorrelated Observables).

- **Correlation time**: In time-series data of a random quantity $x(t)$ (e.g. a physical property from a MC or MD trajectory), this is the time τ over which $x(t)$ and $x(t+\tau)$ remain (linearly) correlated[7].
- **Two-sided confidence interval**: An interval, typically stated as $\langle x \rangle = \bar{x} \pm U$, which is expected to contain the possible values attributed to $\langle x \rangle$ given the experimental measurements of x_j and a certain *level of confidence*, denoted p. The size of the confidence interval, known as the *expanded uncertainty*, is defined by $U = ks(\bar{x})$ where k is *coverage factor*[21].[8] The level of confidence p is typically given as a percentage, e.g. 95 %. Hence, the confidence interval is typically described as "the p % confidence interval" for a given value of p.
- **Coverage Factor**: The factor k which is multiplied by the experimental standard deviation of the mean $s(\bar{x})$ to obtain the expanded uncertainty, typically in the range of 2 to 3. In general, k is selected based on the chosen level of confidence p and probability distribution that characterizes the measurement result x_j. For Gaussian-distributed data, k is determined from the t-distribution, based on the level of confidence p and the number of measurements in the experimental sample.[9] See Sec. 7.5 for further discussion on the selection of k and the resultant computation of confidence intervals.

1.2.2 Terminology and its relation to our broader perspective on uncertainty

As surveyed by Refs. [20, 21], the discussion that originally motivated many of theses definitions appears rather philosophical. However, there are practical issues at stake related to both the content of the definitions as well as the need to adopt their usage. We review such issues now.

At the heart of the matter is the observation that any uncertainty analysis, no matter how thorough, is inherently subjective. This can can be understood, for example, by noting that the arithmetic mean is itself actually a random quantity that only approximates the true expectation value.[10] Because its variation relative to the true value depends on the number of samples (notwithstanding a little bad luck), one could therefore argue that a better mean is always obtained by collecting more data. We cannot collect data indefinitely, however, so the *quality* of an estimate necessarily depends on a *choice* of when to stop. Ultimately, this discussion forces us to acknowledge that *the role of any uncertainty estimate is to facilitate decision making,* and, as such, the thoroughness of any analysis should be tailored to the decision at hand.

Practically speaking, the definitions as put forth by the VIM attempt to reflect this perspective while also capturing ideas that the statistics community have long found useful. For example, the concept of an "experimental standard deviation of the mean" is nothing more than the "standard error of the mean." However, the adjective "experimental" explicitly acknowledges that the estimate is in fact obtained from observation (and not analytical results), while the use of "deviation" in place of "error" emphasizes that the latter is unknowable. Similar considerations apply to the term "experimental standard deviation," which is more commonly referred to as the "sample standard deviation."

It is important to note that subjectivity as identified in this discussion does not arise just from questions of sampling. In particular, methods such as parametric bootstrap and correlation analyses (discussed below) invoke modeling assumptions that can never be objectively tested. Moreover, experts may not even agree on how to compute a derived quantity, which leads to ambiguity in what we mean by a "true value."[31] That we should consider these issues carefully and assess their impacts on any prediction is reflected in the definition of the "standard uncertainty," which does not actually tell us how to compute uncertainties. *Rather it is the task of the modeler to consider the impacts of their assumptions and choices when formulating a final uncertainy estimate. To this end, the language we use plays a large role in how well these considerations are communicated.*

As a final thought, we reiterate that the goal of an uncertainty analysis is not necessarily to perform the most thorough computations possible, but rather to communicate clearly and openly what has been assumed and done. We cannot predict every use-case for data that we generate, nor can we anticipate the decisions that will be made on the basis of our predictions. The importance of clearly communicating therefore rests on the fact that in doing so, we allow others to decide for themselves whether our analysis is sufficient or requires revisiting. To this end, consistent and precise use of language plays an important, if understated role.

2 Best Practices Checklist

The self-contained checklist is presented on the following page.

[7] 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 order-of-magnitude estimate of the number of (linearly) *uncorrelated* samples generated by a simulation. See Sec. 7.3.1.

[8] This conceptual description of a confidence interval is only applicable when certain conditions are met, including the important stipulation that all uncertainty contained in $s(\bar{x})$ is determined only by statistical evaluation of the random experimental measurements of x_j[20].

[9] For discussion regarding the selection of k for non-Gaussian-distributed data, consult Annex G of Ref. [20].

[10] Notably, the same observation applies to the experimental standard deviation and the corresponding experimental standard deviation of the mean.

QUANTIFYING UNCERTAINTY AND SAMPLING QUALITY IN MOLECULAR SIMULATION

- ☐ **Plan your study carefully by starting 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
  - ☐ Consult best-practices papers on simulation background and planning/setup. See: https://github.com/MobleyLab/basic_simulation_training
  - ☐ Estimate whether system timescales are known experimentally and feasible computationally based on published literature. If timescales are too long for straight-ahead MD, investigate enhanced-sampling methods for systems of similar complexity. The same concept applies to MC, based the number of MC trial moves instead of actual time.
  - ☐ Read up on sampling assessment and uncertainty estimation, from this article or another source (e.g., Ref. [16]). Understanding uncertainty will help in the *planning* of a simulation (e.g., ensure collection of sufficient data).
  - ☐ Consider multiple runs instead of a single simulation. Diverse starting structures enable a check on sampling for equilibrium ensembles, which should not depend on the starting structure. Multiple runs may be especially useful in assessing uncertainty for enhanced sampling methods.
  - ☐ Check and validate your code/method via a simple benchmark system. See: https://github.com/shirtsgroup/software-physical-validation
- ☐ **Do not "cherry-pick" data that provides a hoped-for outcomes.** This practice ethically questionable and, at a minimum, can significantly bias your conclusions. Use all of the available data unless there is an objective and compelling reason not to, e.g. the simulation setup was incorrect or a sampling metric indicated that the simulation was not equilibrated.
- ☐ **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, both those of interest *and others*. 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 results (e.g., time series, distributions, etc.) from different simulations.
  - ☐ An individual trajectory can be divided into two parts and analyzed as if two simulations had been run.
- ☐ **Remove an "equilibration" (a.k.a. "burn in", or transient) portion of a single MD or MC trajectory** and perform analyses only on the remaining "production" portion of trajectory. *Underlying concept:* An initial configuration is unlikely to be representative of the desired ensemble and the system must be allowed to relax so that low probability states are not overrepresented in collected data. See Sec. 5.
- ☐ **Consider computing a quantitative measure of global sampling**, i.e., attempt to estimate the number of statistically independent samples in a trajectory. *Underlying concept:* Sequential configurations are highly correlated because one configuration is generated from the preceding one, and estimating the degree of correlation is essential to understanding overall simulation quality. See Secs. 6 and 7.3.1.
- ☐ **Quantify uncertainty in specific observables of interest using confidence intervals.** *Underlying concept:* The statistical uncertainty in *arithmetic mean* of an observable decreases as more independent samples are obtained and can be much smaller than the experimental standard deviation of that observable. See Sec. 7.
- ☐ **Use special care when designing uncertainty analyses for simulations with 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.
- ☐ **Report a complete description of your uncertainty quantification procedure, detailed enough to permit reproduction of reported findings.** Describe the meaning and basis of uncertainties given in figures or tables in the captions for those items, e.g., "Error bars represent 95% confidence intervals based on bootstrapping results from the independent simulations." Provide expanded discussion of or references for the uncertainty analysis if the method is non-trivial. Consider publishing unprocessed simulation data (measurements/observations) and post-processing scripts, perhaps using public data or software repositories, so that readers can exactly reproduce the processed results and uncertainty estimates. *Underlying concept:* The non-uniformity of uncertainty quantification procedures in the modern literature underscores the value of clarity and transparency going forward.

3 Pre-simulation "sanity checks" and planning tips

Sampling a molecular system that is complex enough to be "interesting" in modern science is often extremely challenging – not to mention the difficulties in studying "simple" systems [45]. Therefore, a small amount of effort spent planning a study can pay off many times over. In the worst case, a poorly planned study can lead to weeks or months of simulations and analyses that yield questionable results.

With this in mind, one of the objectives of this document is to provide a set of benchmark practices against which reviewers and other scientists can judge the quality of a given work. If you read this guide in its entirety *before* performing a simulation, you will have a much better sense of what constitutes (in our minds) a thoughtful simulation study. Thus, we strongly advise that readers review and understand the concepts presented here, as well as in related reviews [16, 20, 29]

In a generic sense, the overall goal of a computational study is to be able to draw statistically significant conclusions regarding a particular phenomenon. To this end, "good statistics" usually follow from repeated observations of a quantity-of-interest. While such information can be obtained in a number of ways, time-series data is a natural output of many simulations and is therefore a commonly used to achieve the desired sampling. Several observations follow.

For one, time-series data generally displays a certain amount of autocorrelation in the sense that the numerical values of nearby points in the series tend to cluster close to one another. Intuition dictates that correlated data does not reveal fully "new" information about the quantity-of-interest, so that we require uncorrelated samples to achieve meaningful sampling [30].[11]

Thus, it is critical to ask: *what are the pertinent timescales of the system?* Unfortunately, this question must 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 the autocorrelation analysis 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 [16] – 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 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.

[11] This intuition is actually somewhat misguided in the sense that anti-correlated data actually *increases* our knowledge of a given random quantity. See, for example, the discussion in Ref. [30].

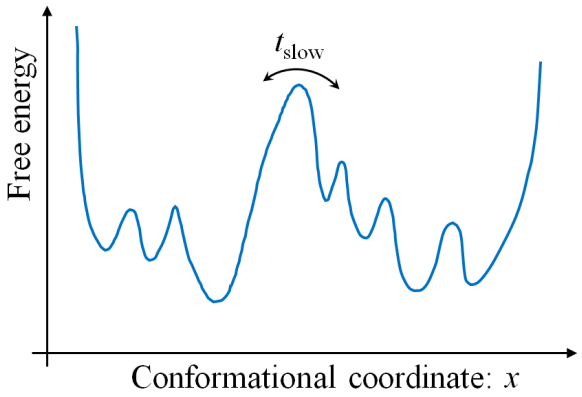

Figure 1. Schematic illustration of a free energy landscape dominated by a slow process. The timescales associated with a system will often reflect "activated" (energy-climbing) processes, although they could also indicate diffusion times for traversing a rough landscape with many small barriers. In 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.

One general strategy that will allow you to understand the relevant timescales in a system 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 generally advised 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 information if the runs are long enough.

A toy model illustrates some of these timescale-issues and their effects on sampling. Consider the "double-well" free energy landscape shown in Fig. 1, and note that the slowest timescale is associated with 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. In turn, the equilibrium average of an observable will require sampling the two basins according to their equilibrium populations. In order to directly sample these basins, however, the length of a trajectory will

have to be many times the slowest timescale, i.e. the largest barrier should be crossed multiple times. Only in this way can the relative populations of states are inferred from time spent in each state. Stated differently, the equilibrium populations follow from the transition rates [8, 23, 55] 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 [[ADD REFS]] or an enhanced sampling technique [[ADD REFS]], 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 Qualitative and semi-quantitative 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 relatively simple tests that can quickly bring out inadequacies in sampling.

Generally speaking, analysis routines that extract information from raw simulated data are often formulated on the basis of physical intuition about how that data should behave. Before proceeding to quantitative data analysis and uncertainty quantification, it is therefore useful to assess the extent to which 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 its meaningfulness [31]. Unfortunately, general recipes for assessing data quality are impossible to formulate, owing to the range of physical quantities of interest to modelers. Nonetheless, several example procedures will help clarify the matter.

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 – see Sec. 5. If *any* value appears to be systematically changing, then the system may not be equilibrated and further investigation is warranted.

In the case of visually ambiguous datasets, autocorrelation analyses can be used to better understand the extent to which a time-series represents an equilibrated system. In particular, systems in steady-state (which includes equilibrium) by definition have statistical properties that are time-invariant. Thus, correlations between a single observable at different times depend only on the relative lag between the timesteps. That is, the correlation function has the stationarity property

$$C(x_j, x_{j+\tau}) = \langle (x_j - \langle x \rangle)\,(x_{j+\tau} - \langle x \rangle) \rangle = C(\tau) \qquad (9)$$

where $C(\tau)$ is independent of the time-step j. With this in mind, one can partition a given time-series into continuous blocks, compute the autocorrelation for a collection of lags τ, and compare between blocks. Estimates of $C(\tau)$ that are independent of the block suggest an equilibrated (or at least a steady-state) system, whereas significant changes in the autocorrelation may indicate an unequilibrated system. Importantly, this technique can help to distinguish long-timescale trends in otherwise equilibrated data from truly non-equilibrated systems.

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

PNP comment: as more of a material scientist, I don't really know what this section is saying. It might be nice to have a brief intro paragraph (or at least a few sentences) saying what RMSD is and how we can understand it physically. What's the executive summary of why it can be used to assess equilibration? Maybe then launch into a more detailed discussion

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

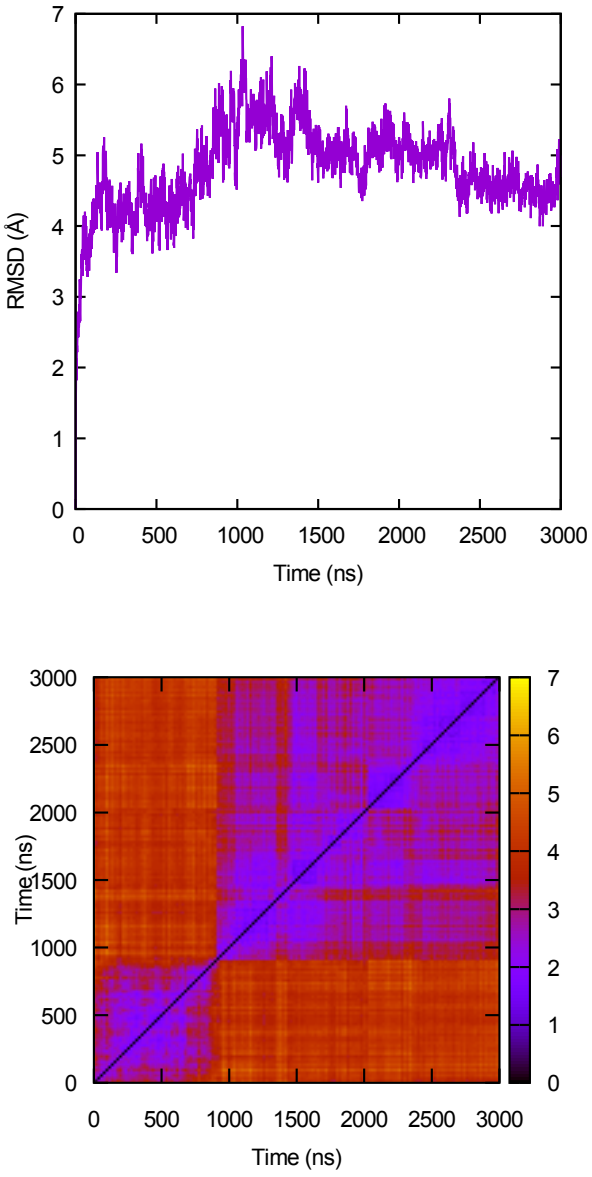

Figure 2. RMSD as a measure of convergence. The upper panel shows the α-carbon RMSD of the protein rhodopsin from its starting structure as a function of time. The lower panel shows the all-to-all RMSD map computed from the same trajectory. Data from Leioatts, et al [26].

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 [33].

Some qualitative tools for assessing global sampling based on RMSD were reviewed in prior work [16]. 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 [26]; 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 rhodopsin trajectory. 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 Analyzing the qualitative behavior of data

In many cases, analysis of simulated outputs relies on determining or extracting information from a regime in which data is expected to behave a certain way. For example, we might anticipate that a given dataset should have linear regimes or more generically look like a convex function. However, typical sources of fluctuations in simulations often introduce noise that can distort the character of data and thereby render such analyses difficult or even impossible to approach objectively. It is therefore often useful to sytematically assess the extent to which raw data conforms to our expectations and requirements.

In the context of materials science, simulations of yield-

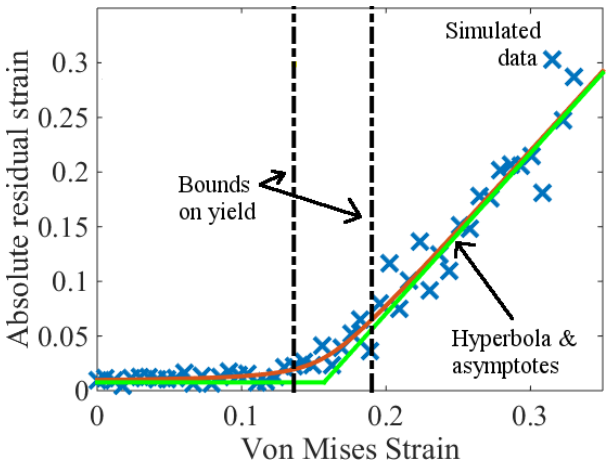

Figure 3. Yield strain tn ϵ_r as a function of applied strain ϵ. Blue $\times$ denote simulated data, whereas the smooth curve is a hyperbola fit to the data. The green lines are asymptotes; their intersection can be taken as an estimate of ϵ_y. Bounds on yield are computed by the synthetic data method 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.*

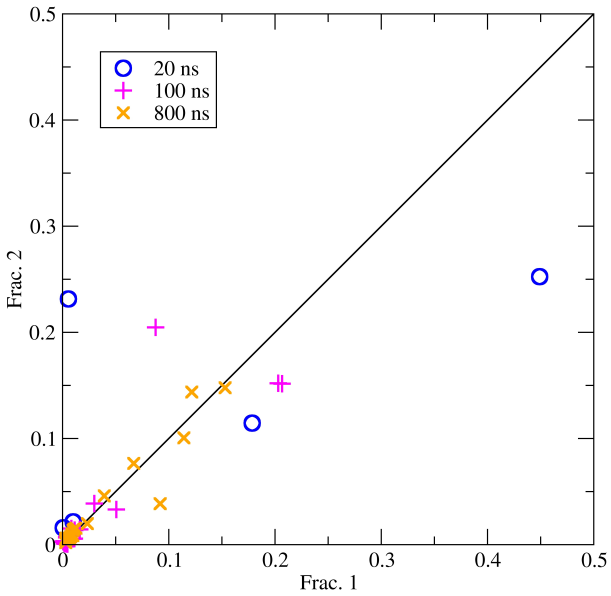

Figure 4. Combined clustering between two independent trajectories as a measure of convergence. The X axis is the population of a cluster from trajectory 1, while the Y axis is the population of that cluster from trajectory 2. Cluster populations are show after 20, 100, and 800 ns of sampling. The simulations used to generate the data used in this plot are described in Roe et al.[40]

strain ϵ_y (loosely speaking, the deformation at which a material fails) provide one such example. In particular, 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 $\epsilon \geq \epsilon_y$ [32]. Thus, residual-strain data should exhibit bilinear behavior, with slopes indicating whether the material is in the pre- or post-yield regime.

In experimental data, these regimes are generally distinct and connected by a sharp transition. In simulated data, however, the transition in ϵ_r around yield is generally smooth and not piecewise linear, owing to the timescale limitations of MD. Thus, it is useful to perform analyses that can objectively identify the asymptotic regimes without need for input from a modeler. One way to achieve this is by fitting residual strain to a hyperbola. In doing so, 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. 3 and Refs. [31, 32] for more examples and discussion.

While extending this approach to other types of simulations invariably depends on the problem at hand, we recognize a few generic principles. In particular, it is sometimes possible to test the quality of data by fitting it to *global* (not piecewise or local!) functions that exhibit characteristics we desire of the former. By testing the goodness of this fit, we can assess the extent to which the data captures the entire structure of the fit-function and therefore conforms to expectations. We note that this task can even be done in the absence of a known fit function, given only more generic properties such as convexity. See, for example, the discussion in Ref. [30].

4.4 Tests based on independent simulations and related ideas

When estimating any statistical property, multiple measurements are often required to capture the uncertainties involved. Consider, for example, the probability that an unbiased coin will land heads-up as described in terms of the relative fraction coin-flips that give this result. This fraction approximated in terms of a single flip (measurement) will always yield a grossly incorrect probability, since only one outcome (heads or tails) can ever be represented by this procedure. However, as more flips (measurements) are made, the relative fraction of outcomes will converge to the correct probability, i.e. the former represents and increasingly good estimate of the latter.

In an analogous way, we often use "convergence" in the context of simulations to describe the extent to which an estimator (e.g. an arithmetic mean) approaches some true value (i.e. the corresponding expectation of an observable) with increasing amounts of data. In many cases, however, the true value is not known *a priori*, so that we cannot be sure what value a given estimator should be approaching. In such cases, it is common to use the overlap of independent estimates and confidence intervals as a proxy for convergence because

the associated clustering suggests a shared if unknown mean. Conversely, lack of such “convergence” is a strong indication that sampling is poor.

There are two approaches to obtaining independent measurements. Arguably the best is to have multiple independent simulations, each with different initial conditions. Ideally these conditions should be chosen so as to span the space to be sampled, which provides confidence that simulations are not being trapped in a local minimum. Consider, for example, the task of sampling the phi and psi torsions of alanine dipeptide. To accomplish this, one could initialize these angles in the alpha-helical conformation and then run a second simulation initialized in the polyproline II conformation. It is important to note, however, 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.

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 biased by initial conditions (e.g. trapped in a local energy minimum) since there is only a single starting point. Those employing this approach should take extra care to assess their results (see in particular sections on ’block averaging’ below).

PNP comment: I have some serious concerns about the next paragraph. Overlap of standard deviations has nothing to do with normality. I actually think this is a critical point. I think this language conflates issues of statistics with what I would call "value judgements." Generally speaking, an overlap of error bars (or lack thereof) is ONLY a statement about second moments. Second moments do not fully characterize a distribution (except for Gaussians). So given a lack of overlap of error bars, we can’t say with 100% certainty that two random quantities (e.g. estimators) don’t share a common mean. I think the best we can do is say that the situation is unlikely, and if we’re good at probability and make a few assumptions, we can say how unlikely. However, the issue of convergence here is really a subjective one: do we feel that the sampling is good enough based on overlap of error bars (or lack thereof). I call this a "value judgement" because the answer depends what my priorities / concerns are about the data are and how I plan to use conclusions I draw from it. If there is a tiny amount of overlap, that may be good enough in some cases. I recommend changing the message of the next paragraph to reflect the distinction between “convergence” as a value judgement and overlap of error bars as a purely statistical observation.

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 [[SHOULD HAVE REFS FOR THIS]].

Combined Clustering

PNP comment: I think it would be useful to describe what a cluster analysis is. What is the main idea, and briefly sketch what the relevant math would look like, or at least point to a later section where we discuss these issues

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 (or a single trajectory is divided into two or more parts), on which cluster analysis is performed. The resulting clusters are then split according to the trajectory (or part of the trajectory) they originally came from. If simulations are converged then each part will have similar populations for any given cluster. Indications of poor convergence are large deviations in cluster populations, or clusters that show up in one part but not others. Figure 4 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 simulations are 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 of *equilibrium properties*; the remaining trajectory data is often called ‘production’ data. See Fig. 5. 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.

PNP Comment: I rephrased the next two paragraphs consistent with my understanding of what they were trying to say. Original text is commented out in latex. Please verify what I

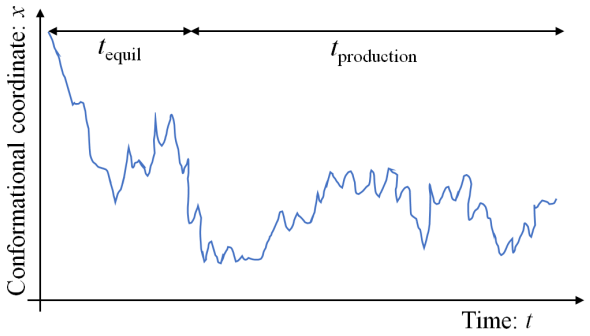

Figure 5. The equilibration and production segments of a trajectory. "Equilibration" over the time t_{equil} represents transient behavior while the initial configuration relaxes toward configurations more representative of the equilibrium ensemble. Readers are encouraged to select t_{equil} in a systematic way based on published literature. If you find strong sensitivity of "production" data to the choice of t_{equil}, this suggests additional sampling is required.

have written is representative of what was intended.

To illustrate these points, consider the process of relaxing an initial, crystalline configuration of a protein to its amorphous counterpart in an aqueous environment. While the initial structure might seem to be intrinsically valuable, remember that configurations representative of the crystal structure may never appear in an aqueous system. Perhaps worse, the force field used for aqueous interactions will in general be different from its crystalline counterpart, so that we cannot realistically hope to even model the latter. As a result, the initial structure may be subject to unphysical forces and/or transitions that provide useless, if not misleading information about the system behavior.[12] Relaxation should therefore be viewed as a means to an end: we only care that the final state is representative of *any* local energy-minimum that the system might sample, not how we arrived at that state.

The RMSD trace in Fig. 2 illustrates typical behavior of a system undergoing relaxation. Note the very rapid RMSD increase in the first ∼200 ns. Part of this increase is simply entropic: the volume of phase space within 1 Å of a protein structure is extremely small, so that the process of thermalizing rapidly increases the RMSD from the starting structure, *regardless of how favorable or representative that structure is*. Thus, examining that initial rapid increase is not helpful in determining an equilibration time. However, in this case, the RMSD continues to increase past 3 Å, which is larger than the amplitude of simple thermal fluctuations (shown by Fig. 2B), indicating an initial drift to a new structure, followed by sampling.

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. In statistical terms, we want to remove bias but also minimize uncertainty (variance) through adequate sampling. Before addressing this problem, however, we emphasize 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.

PNP comment: The below paragraph has a few ambiguities. See .tex file for my comments

A robust approach to determining the equilibration time is discussed in Refs. [7, 52]. The key idea 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' [27] described in Sec. 6. Further, if values of observables estimated from the production phase depend sensitively 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 task 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? PNP comment: I left a question about this section in the corresponding issue thread. The methods noted in this section build on ideas already presented in Sec. 4 on qualitative sampling analysis, but attempt to go a step further to quantify

[12] In MD modeling of structural polymers (e.g. thermoset polymers), the problem of unphysical forces can be so severe that simulations become numerically unstable and crash. This frequently manifests as systems that explode and/or tear themselves apart. As a result, relaxation is often performed using Monte Carlo moves that minimize energy without reference to velocities and forces.

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

PNP comment; I left a question about this section in the corresponding issue thread. 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 [16]. 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 trajectories are often correlated in time. 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 [27]. Would be helpful to clarify what this configuration space is. Am I incorrect in assuming it is the full 6N-dimensional phase-space? Other comments embedded in the .tex file as commented lines. 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. Is this really about decorrelation times? Perhaps I'm understanding wrong, but it seems like you're trying to address the problem of testing that all of phase-space has been sampled multiple times.

A second method, employing block covariance analysis (BCOM), was presented by Romo and Grossfield [43] building on ideas by Hess [18]. In essence, the method combines two standard error analysis techniques, block averaging [13] and bootstrapping [12], with a quantitative assessment of the similarity of modes determined from principal component analysis, covariance overlap [18]. 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.

These methods are implemented as part of LOOS [42, 44].

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 [54]. 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. What is the effective sample size? Number of independent trajectory frames?

Nemec and Hoffmann proposed related sampling measures geared specifically for analyzing and comparing multiple trajectories [28]. 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 (addressing whether the same configurational states were sampled) and the density agreement (addressing whether conformational regions were sampled with equal probabilities).

6.4 Global sampling assessment for enhanced sampling methods

The family of enhanced equilibrium sampling methods, including replica exchange and variants [47–49], metadynamics [6, 24], adaptive biasing force [9–11] 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. [54] 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

Here we address the simple but critical question, "What error bar should I report?" 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: 1) to help authors and readers better understand uncertainty in data; and 2) 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 95% confidence intervals.* As explained in the glossary above, a 95 % confidence interval is a range that contains 95 % of the possible values that could be attributed to the observed quantity *if statistically equivalent simulations are repeated a large number of times;* (c) when feasible and especially for a small number of independent measurements ($n < 10$), authors should consider plotting all of the points or a histogram instead of an average with error bars.

We emphasize that as opposed to standard uncertainties [$s(\bar{x})$], confidence intervals have several practical benefits that justify their usage. In particular, they directly quantify the range in which the average value of an observed quantity is expected to fall, which is more relatable to everyday experience than, say, the moments of a probability distribution. As such, confidence intervals can help authors and readers better understand the implications of an uncertainty analysis. Moreover, downstream consumers 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 integer multiples of $s(\bar{x})$ can be misinterpreted as unrealistically under or overestimating uncertainty if taken at face value. For example, reporting $3s(\bar{x})$ uncertainties for a normal random variable amounts to a 99.7 % level of confidence, which is likely to be a significant overestimate for many applications. On the other hand, $1s(\bar{x})$ uncertainties only correspond to a 68 % level of confidence, 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 explicitly state the confidence level of their error bar or reported confidence interval.

In recommending 95 % 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. 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, $1s(\bar{x})$ 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 95 % 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 mislead most readers – even experts who do not search out the fine print.

As a final note, we remind readers that only significant figures should be reported. Additional digits beyond the precision implicit in the uncertainty are unhelpful at best, and potentially misleading to readers who may not be aware of the limitations of simulations or statistical analyses generally.

7.2 Overview of procedures for computing a confidence interval

We remind readers that they should perform the semi-quantitative sampling checks (Sec. 4) before attempting to quantify uncertainty. If the observable of interest is not fluctuating about a mean value but largely increasing or decreasing during the course of a simulation, a reliable quantitative estimate for the observable or its associated uncertainty cannot 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 Gaussian-distributed (or assumed to be, as an approximation or due to lack of information), an appropriately chosen *coverage factor k* (typically in the range of 2 to 3; see Sec. 7.5 for further details) is multiplied by the standard uncertainty $s(\bar{x})$ to yield the expanded uncertainty, which estimates the 95 % confidence interval.
- For non-Gaussian observables, a *bootstrapping* approach (Sec. 7.6) should be used. An example of a potentially non-Gaussian observable is a rate-constant, which must be positive but could exhibit significant variance. As such, a confidence interval estimated with a coverage factor may lead to an unphysical negative lower limit. In contrast, 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 is also sometimes useful for estimating uncertainties associated with derived observables.

Below we describe approaches for estimating the standard uncertainty $s(\bar{x})$ from a single trajectory with a coverage factor k as well as 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, via block averaging, or by analysis of a time-correlation function. However, these methods have their limitations and must be used with caution. In particular, both block averaging and autocorrelation analyses will produce effective sample sizes that depend on the quantity of interest. To produce reliable answers, one must therefore 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 $s(x)$ for quantity x (which could be the ensemble average of a raw data output or a derived observable) among the n simulations, yielding a standard uncertainty of $s(\bar{x}) = s(x)/\sqrt{n}$. When computing the uncertainty with this approach, it is important to ensure that each starting configuration is also independent or else to recognize and report 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 Dealing with correlated time-series data

When samples of a simulated observable are independent, Eq. 8 for the experimental standard deviation of the mean can be used as an estimate of the corresponding standard uncertainty. Due to correlations in typical time-series data, however, the number of independent samples in a simulation is neither equal to the number of observations nor known *a priori*; thus Eq. 8 is not directly useful. To overcome this problem, a variety of techniques have emerged that attempt to estimate the effective number of independent samples in a dataset. Generally speaking, these method fall into roughly two categories: (i) autocorrelation analyses, which directly estimate the number of independent samples in a time-series; and (ii) block-averaging, which projects a time-series onto a smaller dataset of (approximately) independent samples. We now discuss these methods in more detail.

7.3.1 Autocorrelation method for estimating the standard uncertainty

Conceptually, autocorrelation analyses aim to recover the usefulness of Eq. 8 by replacing the number of samples N with an effective number N_{ind} that takes into account "redundant" (or even possibly new) information arising from correlations.[13] The main idea behind this approach is to invoke the fact

[13] It is worth pointing out that correlations do not always provide redundant information. Consider, for example, the time-series 1, –1, 1, –1, 1, –1, ... In the limit that the number of elements goes to infinity, the arithmetic mean also converges to zero. However, a block of $2n$ entries also has a mean of zero, so that (anti)correlations effectively increase the amount of information. See also Ref. [30].

that the statistical properties of steady-state simulations (e.g. those in equilibrium or non-equilibrium steady-state) are, by definition, time-invariant. As such, correlations between an observable computed at two different times depends only on the lag (i.e. difference) between those times, not their absolute values.

This observation motivates one to compute an autocorrelation function. Specifically, given a sequence of observations $\{x_1, ..., x_N\}$, the autocorrelation function C is defined for a set of lags j via:

$$C_j = \frac{\overline{\left(x_k - \bar{x}\right)\left(x_{k+j} - \bar{x}\right)}}{s(x)^2} \tag{10}$$

where the denominator is the square of the experimental standard deviation of x given by Eq. 6. Then, the number of independent samples is estimated by[14]

$$N_{ind} = \frac{N}{1 + 2\sum_{i=1}^{N_{lags}} C_j} \tag{11}$$

where N_{lags} is the number of lags for which the C_j was computed. Note that N_{ind} need not be an integer Finally, the standard uncertainty is estimated via

$$s(\bar{x}) = \frac{s(x)}{\sqrt{N_{ind}}} \tag{12}$$

We note that the experimental standard deviation of the observable x is used in Eq. 12 to estimate the uncertainty. Strictly speaking, the standard uncertainty should be estimated using the true standard deviation of x (e.g. σ_x); given that the true standard deviation is unknown, the experimental standard deviation is used in its place as an *unbiased estimate* of σ_x [30].

7.3.2 Block averaging method for estimating the standard uncertainty

The main idea behind block-averaging is to permit the direct usage of Eq. (6) by projecting the original dataset onto one comprised of only independent samples, so that there is no need to compute N_{ind}. Acknowledging that typical MD time-series have a finite-correlation time τ, we recognize that a continuous block of M data-points will only be correlated with its adjacent blocks through its first and last τ points, provided $\tau \ll M$ is small compared to the block size. That is, correlations will be on the order of τ/M, which goes to zero in the limit of large blocks.

[14]The reader should note that both the autocorrelation function (Eq. 10) and the number of independent samples (Eq. 11) may be written in different forms[7, 16]. Our convention here presents the observations as a list $\{x_j\}$ in which the time interval (Molecular Dynamics) or trial spacing (Monte Carlo) of adjacent x_j is implicitly fixed. For time-series data, one could alternately write both the observations and autocorrelation function as continuous functions of time, e.g. $x(t)$ and $C(\tau)$ where τ is the lag time. In that case, N_{ind} is written as a division of the total simulation time by the time integral of $C(\tau)$[16].

This observation motivates block-averaging as follows [13–16]. Briefly, the set of N observations $\{x_1, ..., x_N\}$ are converted to a set of M "block averages" $\left\{x_1^b, ..., x_M^b\right\}$, where a block average x_j^b is the arithmetic mean of n (the block size) sequential measurements of x:

$$x_j^b = \frac{\sum_{j=1+(k-1)n}^{kn} x_j}{n} \tag{13}$$

From this set of block averages, one may then compute the arithmetic mean of the block averages, $\bar{x}^b$, which is an estimator for $\langle x \rangle$. Following, one computes the experimental standard deviation of the block averages, $s\left(\bar{x}^b\right)$, using Eq. 6. Lastly, the standard uncertainty of $\bar{x}^b$ is just the experimental standard deviation of the mean given the set of M block averages:

$$s\left(\bar{x}^b\right) = \frac{s\left(x^b\right)}{\sqrt{M}} \tag{14}$$

This standard uncertainty may then be used to calculate a confidence interval on $\bar{x}^b$.

It is important to note that for statistical purposes, the blocks must all be of the same size in order to identically distributed, and thereby satisfy the requirements of Eq. 8. It is alost important systematically assess the impact of block-size on the corresponding estimates. In particular, as the blocks get longer, the block averages should decorrelate and $s\left(\bar{x}^b\right)$ should plateau [13, 16]. Another approach is to measure the block correlation and to use it to improve the selection of the block size and, hence, uncertainty estimate [22]. We stress that this final step of adjusting the block size and recomputing the block standard uncertainty is absolutely necessary. Otherwise, the blocks may be correlated, yielding an uncertainty that is not meaningful.

7.4 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 [50].

$$\beta\Delta A = -\ln\left\langle \exp\left(-\beta\Delta U\right)\right\rangle \tag{15}$$

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

$$\sigma_{\beta\Delta A} = \sigma_{\exp(\beta\Delta U)} / \left\langle \exp\left(-\beta\Delta U\right)\right\rangle \quad (16)$$

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 is 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. 16 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 estimated with alternative approaches such as bootstrapping to validate the Taylor series results or to identify an alternative approach that works better.

7.5 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 95 %-confidence interval $\left[\bar{x} - k\,s\left(\bar{x}\right), \bar{x} + k\,s\left(\bar{x}\right)\right]$ can be constructed on the basis of an established look-up table (or a statistics software model) for the coverage factor k based on n. The theoretical basis for the table is the "Student" or "t" distribution, which is *not* Gaussian, but governs the behavior of an *average* derived from n independent Gaussian variables [20]. Table 1 lists k for two-sided 95 % confidence intervals for select values of n.

When $n \leq 10$, we recommend showing all data points, as a confidence interval may not be 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 are also not Gaussian. These cases should be treated with bootstrapping.

n (independent samples)	k (coverage factor)
6	2.57
11	2.23
16	2.13
21	2.09
26	2.06
51	2.01
101	1.98

Table 1. Coverage factors k required for a two-sided 95 % confidence interval for a Gaussian variable [20].

7.6 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 [12]. 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 95 % confidence interval from the 2.5 %ile to the 97.5 %ile value.

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 (e.g. using an autocorrelation method [7, 27]) 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 based on analyzing varying block sizes (see above) 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 [43].

7.6.1 Bootstrapping variants

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 de-

terministic 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 $\bar{x}$ 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 [36]. In such cases, determining the hyperparameters may require more sophisticated techinques such as maximum likelihood analyses or Bayesian approaches; see, for example, Ref. [36]. See also Refs. [4, 5, 31, 32] for examples and practical implementations applied to cases in which the determinsitic component of the data is not constant.

An alternative to bootstrapping is the "jackknife" method[34, 35, 51] (also outlined in Ref. [12]). It operates similar to the bootstrap as a resampling technique, but it uses synthetic data sets created by subtraction of samples rather than replacement; as such it is often categorized as a variant of the bootstrap (even though it predates the bootstrap). Since it operates by sample deletion, it may be better suited to smaller data sets for which a few samples may be overrepresented in the synthetic data set created by the bootstrap replacement technique. Ultimately, though, the results are similar in that the jackknife technique creates a distribution of derived observables that can be used to compute both an arithmetic mean and estimate of the standard uncertainty.

7.6.2 Bootstrapping and uncertainty propagation

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.

7.7 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

$$\bar{\tau} \approx \mathcal{T} \propto \sum_i \frac{T_i}{\sigma_i^2 + y^2}, \tag{17}$$

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. [31]). Equation (17) 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. [31] and the references contained therein.

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 [55]. 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) [48]; see also [49]. Broadly speaking, REMD consists of running parallel MD simulations on a number of non-interacting replicas of a system, each with a different Hamiltonian (PNP comment: see latex for question) and/or thermodynamics parameters (e.g., temperature), and periodically exchanging system coordinates between replicas according to a Metropolis criterion. PNP comment: I think we need a sentence or two explaining this 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.

In practical settings, several metrics are often used to assess these two efficiencies, a few of which we list below. In these definitions, note that we 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. A low exchange acceptance suggests that configuration space may not be sampled adequately. In such cases, the replica spacing may need to be decreased or additional replicas used. Conversely, a high exchange ratio may indicate that each trajectory does not adequetely sample a given replica condition and/or that more resources than necessary are being used to simulate replicates.PNP comment: One problem with this definition is that "high" and "low" are not really defined. I suspect these are user-defined ideas, but do they bear any relation to the other UQ we have discussed?
- 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. PNP question: why should any trajectory 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 [40]. PNP comment: not clear to me why these should be equivalent. Isn't there a distribution of times?
- Replica residence time. The time a 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..[40] Can you elaborate briefly? What does equivalent mean in this context (same order of magnitude?) and why should they be equivalent?
- 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..[17] 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,[41] 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 [19]. 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 [53] as well as steady states [3] including equilibrium [46].

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 a WE overview document.

From a practical standpoint, the safest way to assess uncertainty in WE simulations is to run multiple instances (which can be seeded from identical or different starting structures depending on the desired calculation) 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 [53]; 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 of challenging systems can be large, so multiple runs are advised. In the future, variance-reduction techniques may alleviate the need for multiple runs.

9 Acknowledgments

The authors appreciate helpful discussions with Pascal T. Merz and comments from Lillian T. Chong. 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] **Chou T**, Mallick K, Zia RKP. Non-equilibrium statistical mechanics: from a paradigmatic model to biological transport. Reports on Progress in Physics. 2011; 74(11):116601. http://stacks.iop.org/0034-4885/74/i=11/a=116601.

[9] **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.

[10] **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.

[11] **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.

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

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

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

[15] **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.

[16] **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.

[17] **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.

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

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

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

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

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

[23] **Kolmogoroff A**. Zur Theorie der Markoffschen Ketten. Mathematische Annalen. 1936; 112:155–160. http://eudml.org/doc/159823.

[24] **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.

[25] **Leimkuhler B**, Matthews C. Molecular Dynamics with Deterministic and Stochastic Numerical Methods. Switzerland: Springer International Publishing; 2015.

[26] **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.

[27] **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.

[28] **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.

[29] **Patrone PN**, Dienstfrey A. Uncertainty Quantification for Molecular Dynamics. ArXiv e-prints. 2018 Jan; .

[30] **Patrone P**, Kearsley A, Dienstfrey A. The role of data analysis in uncertainty quantification: Case studies for materials modeling. In: 2018 AIAA Non-Deterministic Approaches Conference AIAA SciTech Forum, American Institute of Aeronautics and Astronautics; 2018. https://doi.org/10.2514/6.2018-0927, doi: 10.2514/6.2018-0927, 0.

[31] **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.

[32] **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.

[33] **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.

[34] **Quenouille MH**. Approximate Tests of Correlation in Time-Series. J Roy Stat Soc B Met. 1949; 11:68–84. doi: 10.1017/S0305004100025123.

[35] **Quenouille MH**. Notes on Bias in Estimation. Biometrika. 1956; 43:353–360. doi: 10.1093/biomet/43.3-4.353.

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

[37] **Rizzi F**, Jones RE, Debusschere BJ, Knio OM. Uncertainty quantification in MD simulations of concentration driven ionic flow through a silica nanopore. II. Uncertain potential parameters. Journal of Chemical Physics. 2013; 138(19):194105. doi: http://dx.doi.org/10.1063/1.4804669.

[38] **Rizzi F**, Najm HN, Debusschere BJ, Sargsyan K, Salloum M, Adalsteinsson H, Knio OM. Uncertainty Quantification in MD Simulations. Part I: Forward Propagation. Multiscale Modeling & Simulation. 2012; 10(4):1428–1459. doi: 10.1137/110853169.

[39] **Rizzi F**, Najm HN, Debusschere BJ, Sargsyan K, Salloum M, Adalsteinsson H, Knio OM. Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-Field Parameters. Multiscale Modeling & Simulation. 2012; 10(4):1460–1492. doi: 10.1137/110853170.

[40] **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.

[41] **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.

[42] **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.

[43] **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.

[44] **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.

[45] **Schappals M**, Mecklenfeld A, Kröger L, Botan V, Köster A, Stephan S, García EJ, Rutkai G, Raabe G, Klein P, Leonhard K, Glass CW, Lenhard J, Vrabec J, Hasse H. Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom. J Chem Theory Comput. 2017; 13:4270–4280. doi: 10.1021/acs.jctc.7b00489.

[46] **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.

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

[48] **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.

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

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

[51] **Tukey JW**. Bias and confidence in not quite large samples (abstract). Ann Math Stat. 1958; 29:614. doi: 10.1214/aoms/1177706647.

[52] **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.

[53] **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.

[54] **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.

[55] **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.