UNITED STATES DEPARTMENT OF COMMERCE
National Institute of Standards and Technology
Gaithersburg, Maryland 20899

29 August 2018

Daniel W. Siderius
Chemical Sciences Division
100 Bureau Dr M.S. 8320
Gaithersburg, Maryland 20899-8320
USA

Prof. Dr. Chris Oostenbrink
Institute of Molecular Modeling and Simulation
Muthgasse 18
1190 Vienna
AUT

Dear Prof. Oostenbrink,

I would like to submit for your consideration in *The Living Journal of Computational Molecular Science* our revised manuscript:

Title: Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations: v1.0
Corresponding Author: D. W. Siderius Authors: A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. M. Zuckerman

We first thank the reviewers for their comments that we have used to revise and improve our manuscript. We address the reviewers comments below and provide responses (color-coded blue) to address those comments.

Reviewer 1

1. p2: Maybe one should also include software/programming errors (see e.g. Ref 1, and a number of other recent references reporting "bugs" in heavily-used MD codes like GROMACS)

We have revised the footnote #1 on page 2 to point the reader to other in-progress papers in LiveCoMS that address basic simulation issues and validation of simulation results.

2. p3: In a didactical sense, maybe it would be worth stating (footnote?) that a probability distribution P(x) has units which are inverse of those of x, and that it should be normalized (i.e. integrate to one); "probability" graphs which are not normalized or/and reported in "arbitrary units" or in "percent" are a very common "bad practice", I think

The Expectation Value definition on page 3 has been revised to note the need for units and normalization of P(x).

3. p3/footnote 3: "As such, we can only" I agree that it is not possible to fully characterize a P(x) based on finite sampling. But if you say "estimate", I don't know why you should limit the statement to expectation value and variance (one can "estimate" higher moments as well, and P(x) itself as a histogram approximation)

Footnote #3 on page 3 has been revised to clarify the mean and variance are example properties that can be estimated from finite sampling.

4. p3: The notation $s(x_j)$ is mathematically unclean (especially for a "best practice" paper), as it suggests that "s" is a function of "x_j". What is meant is more like $s(x_j)$ or $s(xv)$ where xv is the x_j-vector

The reviewer is correct that the $s(x_j)$ notation is clunky; however, we note that we do not actually use $s(x_j)$ in the paper. Our preferred notation for the experimental standard deviation, $s(x)$, actually departs from the VIM/GUM convention that uses $s(x_j)$. Thus, the paper includes a brief remark to clarify that our notation departs from the VIM/GUM convention in contrast to most situations where we use the VIM/GUM notation unchanged.

5. p3: "which is a stronger condition" - there is a beautiful counter example (maybe for a footnote?): data that is perfectly aligned on a circle! (perfect correlation, but zero linear correlation)

Good suggestion. The example of circular data has been employed in the expanded remark following eq. 7 to illustrate limitations of that equation in defining linear correlation.

6. p4: At "correlation time", maybe add a footnote for the MC case (as there is no "time" there); later, a footnote 11 says this, but it is a bit late

Good idea. The definition of Correlation time on page 4 has been revised to mention that the sequence of Monte Carlo trial moves in a Markov Chain is treated identical to time-series data from, e.g., Molecular Dynamics.

7. p10: I think that it is dangerous to oppose a "strategy 1" where you do N "independent" repeats and use stddev/sqrt(N) to a "strategy 2" where you do a single simulation and need to worry about correlation times. As a beginner, I would go for strategy 1 because it seems you have much fewer things to worry about. And if I am a "greedy" beginner and I can do 1000 potential evaluations, I would go for 1000x1 timestep, because then my error is reduced. The only problem is that my probabily distribution is entirely determined by the arbitrary way I select my "independent" samples. In other words: you also have very much to worry about correlation times in strategy 1 !!! And it may turn out that if you add sufficient equilibration time for all your replicas, the multi-repeat approach ends up as complicated and more expensive than the single-simulation way.

Valid point. The reviewers concern is addressed on page 10 in a new paragraph beginning

with However, the main short trajectories strategy

8. p11: "Relaxation should therefore" -¿ I don't understand the sentence.

The purpose of removing a non-equilibrated portion of the trajectory is stated more clearly in a short revision to page 11:

Note that relaxation/equilibration should be viewed as a means to an end: for equilibrium sampling, we only care that the relaxed state is representative of *any* local energy minimum that the system might sample, not how we arrived at that state, which is ultimately why in general data generated during equilibration can be discarded.

9. p15: "block veraging" (typo)

This typo was corrected.

10. p11/footnotes 14,15: Very often, one encounters TCFs that have a damped-oscillatory form. It is not clear at all to me how to extract a "tau" from these. And the time integral of C(tau) is only expected to deliver tau if C(tau) is exponential. What if it is not? Finally, if I agree that the effective number of sample is proportional to N_{config}/τ, is the appropriate prefactor really one? (I have never worked out the math, but I am sure this has been done)

Fundamentally the definition of the correlation time encompasses the case in which the TCF is a damped-oscillatory function; that is, it is the time beyond which linear correlations are zero to within statistical uncertainty. Thus, one can in essence determine tau by inspection, irrespective of how it decays to zero, although automated methods could be devised as well.

As regards the time integral of C, the point is not to determine tau from the integral of C, but rather to integrate C from $t = 0$ to $t = \tau$, since the latter sets an upper limit on the domain in which the integral is non-trivial. That being said, it may be difficult to identify tau, so some authors use and upper bound that is significantly larger than the correlation time (sometimes as large as the number of lags computed), recognizing that $C(t) \approx 0$ for any $t > \tau$. However, in the case of finite sampling, $C(t)$ may contain small amounts of noise that, when integrated, give rise to Brownian-type motion that is unrelated to the underlying correlation function one wishes to integrate.

Because of the complexity of these issues, as well as the fact that the community has yet to settle on best practices for using correlation functions, our main goal is therefore to highlight what techniques are used and the various assumptions underlying them. So instead of recommending a best practice per se, we point the reader to relevant references that summarize current perspectives on using autocorrelation functions.

Regarding the prefactor of N_{config}/τ, we are somewhat unsure as to what prefactor the reviewer has in mind. Conceptually the formula for the effective number of samples is a

definition. So the prefactor is one by construction.

11. p16/footnote 16: This footnote is unclear to me. Seems to me the blocks are non-overlapping and tiling segments of the entire time series (the possibility that it may be otherwise was never mentioned in the text; and I don't think anyone would do this in practice

We have edited the footnote to make it less ambiguous. In light of practices that we have personally seen in the community (especially among beginners), we feel that it is useful to point out the implications of not using all of the data when constructing blocks. In particular, we are aware of instances in which students chose block sizes that were not factors of the number of simulations steps. They were surprised when the block averages were different from the trajectory averages. Moreover, knowing when equality should hold provides a simple verification tool for checking that a code is written correctly.

12. p19: The authors might want to add local-elevation, which predates the two cited methods.

Thank you. The paragraph on page 19 beginning with Generally speaking has been revised to include local-elevation as a technique for assessing enhanced sampling.

Reviewer 2

1. The density of information contained in different sections can vary strongly. That makes the manuscript partially harder to read. While sections 1-5 are rather stretched out and new information is presented slowly, section 6 contains much more information in much less text. Moreover, while in sections 1-5 also rather basic concepts are introduced with care, in section 6 more advanced concepts are just introduced very quickly or not even explained at all. Examples for this are Voroni cells mentioned at the beginning of section 6.1 on page 12 or PCA mentioned on page 13. This makes section 6 partially hard to understand. Figure 6 helps to understand this section, but is not referenced in the text.

First, we would like to acknowledge the reviewers concern regarding asymmetries in information density in the paper; this is indeed true. We deliberately chose to balance readability with completeness in the paper; whenever possible, our discussion is extensive and, perhaps, more detailed than necessary (cf. the definitions section, which could rely on an external document like the GUM) when we wish to firmly establish concepts and techniques. However, when it proves impractical to fully describe a technique in this document (e.g., global uncertainty analysis in Section 6 or the bootstrap in Section 7.6), we provide a higher-level discussion and accompanying references for the reader, else the document would be too long to be a practical best practices guide. In summary, we feel it is impractical and inadvisable to make the discussion details fully uniform.

Section 6, as noted by the reviewer, is fairly dense perhaps because it tackles an important

but less familiar and more difficult subject. As noted above, full explanations would make the overall paper unwieldy so we compromised with a briefer discussion. Regarding the reviewers specific points, Figure 6 was originally referenced in the paragraph beginning Lyman and Zuckerman on page 13, and now is referenced twice in that paragraph. The definition of a Voronoi cell is now given in a footnote - our thanks for pointing that out.

2. In section 5, the authors suggest to remove non-equilibrated trajectory regions from the beginning of simulations to improve simulation results. In my experience, this can indeed be very important and improve convergence of results a lot. Two references (27,28) for the determination of the time needed for equilibration are given, which both discuss algorithms for automated determination. I personally also find another, probably simpler approach proposed by Klimovich , Shirts & Mobley (Guidelines for the analysis of free energy calculations. J Comput Aided Mol Des. 2015) very useful. Simply speaking, in this approach forward and reverse cumulative averages are compared and data is discarded from the beginning of the trajectory until both match. With this approach, also visualization of convergence is nicely possible.

See our reply to Reviewer 1 in comment 8. Also, we added a reference to the useful paper by Klimovich et al - our thanks for pointing us to that paper.

3. In section 7.4 (propagation of uncertainty) expressions for the uncertainty of derived quantities are derived. While it is nice to see the derivation, I personally dont think it is very necessary as the aim of the manuscript is to give basic orientation in the field of uncertainty analysis. It is the only derivation in the whole manuscript.

We disagree with the reviewer that the derivation of linear propagation of uncertainty is unnecessary in this best practices document. It is our experience that very few practitioners of uncertainty propagation actually understand the limitation and/or assumptions (linearization and assumed non-correlation) of the method. By presenting a short derivation, we are able to highlight those assumptions and the limitations of the method, as they follow directly from the mathematical description of the uncertainty propagation equation. Accordingly, we left this section unaltered.

4. However, section 7.4 is completely omitting the important topic of error propagation through multiple steps of simulations/experiments. I think it is very important to properly quantify errors if multiple simulation steps are needed calculate a certain quantity (this happens often in the field of free energy calculations). While the propagation of the experimental standard deviation of the mean might be straightforward, the authors should explain how such error propagation works if the uncertainty is quantified with confidence intervals, as the authors suggest.

We addressed the reviewers comment by adding a new subsection on page 19 entitled Propagation across multiple steps.

Reviewer 3

PDF Issue: The pdf I received had substantial formatting issues. It appears that the letters "fi" often was reproduced as "!" and many other such curious formatting issues appeared. For example, there where "squares" where letters should be and "Definitions" has an with no fi. I encourage the editors / authors make sure these formatting issues do not appear in the final version.

We have been unable to replicate the issue reported by the reviewer. If the editors think this is a real issue, would they please contact the reviewer to determine her/his PDF viewer configuration so that we can attempt to replicate those conditions?

1. For eqn 8, the authors may wish to discuss when n and n-1 is needed. I find this is a source of confusion with students.

Good point. We included a new footnote, #6 on page 3, that briefly addresses the difference between the n and n-1 factors that appear at different times in statistical analyses.

2. In the checklist (page 6) "based the number of MC" =¿ "based on the number of MC"

Thank you. This text error was corrected.

3. In the checklist on page 6 the authors say "Consider publishing unprocessed simulation data" Perhaps the authors would like to be a little stronger here? Strongly urge? I think this is important and perhaps they do too.

The final item in the checklist on page 6 has been modified to state that We strongly urge publication of unprocessed simulation data

4. The lower panel of Fig 2 could use a little more explanation. The caption should define the colors which I think are RMSD angstroms.

The caption of Figure 2 has been revised to clarify the color scale.

5. Section 6.1 contains great information but I was left wanting more information and maybe a practical example to help solidify concepts.

See response to Reviewer 2, point 1. It would be difficult to treat every method fully, so we opted to give a high-level discussion and refer the reader to more complete discussions. As we noted explicitly, there is no true broadly-accepted best practice for assessing global sampling, so at present a fuller description does not seem warranted at present.

6. Page 15 first column: "block verging" => "block averaging". This could just be related to the formatting problem I mentioned earlier.

This typo was corrected as noted previously.

7. Section 7.6 is key, and again I was wanting to see more details - this topic is confusing to students and if the authors could provide more details with an example, it would be great.

The reviewer has a real concern here that the bootstrap section does not have a worked-out example. Our group opinion is, however, that a fuller description/recipe of the bootstrap method would be too lengthy to be included in a best practices document. We did, however, add references in the first paragraph of Section 7.6 (paragraph begins on page 17, new text and references are on page 18) to point the reader to valuable reviews of bootstrap.

8. Eqn 23: Define $\mathcal{T}$

$\mathcal{T}$ has been identified / defined in the revised text on page 19

9. Some pitfalls could be mentioned. Students often take the average from a trajectory and then the standard deviation and assume this is the "error bar"! They don't realize that the standard deviation here is just a measure of the fluctuations, which is just related to system size. It might be worth mentioning somewhere.

Important point. We added an additional sentence in the definitions section to make it clear that the standard deviation reflects the width of the distribution, not statistical error.

Overall, we took the Reviewers comments to be quite favorable to publication, with only minor changes responses needed. We thank you for the opportunity to resubmit this manuscript and hope that our responses are satisfactory.

Yours sincerely, on behalf of all the authors,

Daniel W. Siderius, Ph.D.