Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software — Reviews

Reviewer(s)' Comments to Author:
Reviewer: 1

Recommendation: Publish after major revisions noted.

Comments:

Manuscript ct-2018-00544z, "Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software" by Loeffler et al. takes on the unthanked, but very important task of checking whether it is possible to obtain identical relative free energy differences within statistical error bars for a set of alchemical transformations and an identical force field with several widely employed biomolecular simulation programs. Since the model task chosen involves solvation free energies, in passing the same is done for absolute solvation free energies.

This should definitely be published, subject to the revisions/suggestions pointed out below:

0) The PDF I downloaded is not displaying the figures. I see them mostly in the online preview though Fig.~1 is shown only partially even there. Whether this is the authors' fault or a bug in ACS' review platform I cannot tell. Curiously, the TOC graphic is shown just fine. Anyways, I saw enough to carry out the review.

We will check this on resubmission and try to sort out any issue that may arise then. (#32)

1) To me, the AFE results are a point of concern. I reproduce a shortened version of Table 1 here, showing just the most negative and most positive result for a compound.

My reading of Table 1 (G..Gromacs, A..Amber, C..Charmm, S..SOMD)

	min AFE	max AFE
methane	2.44(G)	2.52(S)
methanol	-3.73(A)	-3.51(G)
ethane	2.48(G)	2.56(S)
toluene	-0.72(A,G)	-0.55(S)
neopentane	2.58(C,G)	2.71(S)
2-methylfuran	-0.51(G)	-0.39(S)
2-methylindole	-6.35(G)	-6.06(C,S)
2-CPI	-6.54(G)	-6.05(A)
7-CPI	-6.52(G)	-5.66(A) !! almost 1 kcal/mol

The statistical error for the results reported is typically extremely low, a few 1/100 kcal/mol, with +-0.1 kcal/mol being the 'largest' reported statistical uncertainty / error estimate. Thus, I consider all of the above differences statistically significant. In most cases the discrepancy between the 'minimal' and 'maximal' result is 0.1--0.3kcal/mol, which I consider irrelevant for practical purposes. However, I am puzzled by the two CPI results, where the spread is 0.5 and almost 1 kcal/mol, respectively. I would argue that this requires some investigation. At one point (unpublished), I recomputed the Shirts/Pande amino acid side chain AFEs (for the CHARMM force field) with CHARMM as opposed to TINKER, which had been used in the original work. Once I set input options as close as possible to what I could deduce from the Shirts / Pande paper, my results were well within +-0.2kcal/mol of the published value, including the indole (Trp side chain analog).

Some programs seem to use reaction fields, even in the gas phase(?). It would be interesting to see if e.g. programs that can use PME in solution and infinite cutoff in gas phase could be coaxed into using equivalent PME settings (I know, easier said than done) to see whether this could explain the discrepancy. The density differences reported don't seem to be the root cause to me.

The issue with 2-CPI and 7-CPI seems to be due to the AMBER results diverging from the other three packages. Excluding AMBER's free energies gives a min/max variability of 0.3-0.4 kcal/mol. We have been unable to explain this particular discrepancy. It warrants additional

follow up work by the community, including the AMBER developers, which is part of why we are making all the input files for this work freely available online.

Finally, there are some statements surrounding Table 1 I find a bit misleading. P.20, l17-21 mentions CHARMM being the largest 'outlier'. I don't see this -- so either the statement is wrong, or some values in the Table are mixed up.

This statement applied to an earlier version of the CHARMM results that lacked LRC correction terms. This statement has been removed from the manuscript.

2) The Introduction describes single vs dual topology in much needed detail. However, the further subdivision of single topology into parameter vs energy scaling is brought up only later (p9). I think this is so fundamental that it should be discussed together with the single/dual topology distinction. These subtle(?) differences see some 'revival', see e.g. 10.1021/acs.jctc.7b01175. I personally tend to classify methods into

dual topology,
single topology (parameter scaling) and
dual-topology-single-coordinate ("single topology with energy scaling")

Moved the relevant discussion on p9 (Methods) to p6 (Introduction)

3) In the discussion of single topology (p6), one finds the following statement about treatment of dummy atoms: "However, it is important to stress that a dummy atom should retain ... to yield correct results." (p6, l27-31). While not really relevant to the primary focus of the manuscript, I disagree strongly with this statement. The conclusion as stated can be found primarily in the work of Shobana et al. (ref. 36 of the manuscript). From a purist's point of view, it has merit, _but_ it may lead to severe problems in practice: Consider the mutation of a methyl group into a hydrogen, i.e., -CH3 -> -HD3. If one follows the advice above, each dummy is attached to the physical H by a bond, an angle and a dihedral. These terms are not sufficient to maintain the tetrahedral geometry of the group; at the -HD3 endpoint, the three dummies will adopt rather arbitrary geometries. The convergence of <dU/dl> at this end point will be extremely poor; if overlooked, this can become a source of systematic error. Bennett's method is less affected, as the few reasonable geometries are likely to suffice. Schroedinger seems to be aware of this; see the supp. information of 10.1021/ja512751q (ref 28 of the manuscript). A more thorough discussion, including the theoretical proof that the

recommendation of Shobana et al. is overly rigid, can be found in https://doi.org/10.1080/08927020211969.

On a related note, while I am aware of the study by Axelsen and Li (ref.33), I wonder whether any dual topology simulations (aside from QM/MM free energy simulations!) are actually carried out in this manner.

We have updated the manuscript to clearly state that the issue of treating bonded terms remain controversial.

“However, it is important to stress that a dummy atom should retain at most only one angle term (Atom1–Atom2–Dummy) and one dihedral term (Atom1–Atom2–Atom3–Dummy) with respect to non-dummy atoms to yield correct results.25,36”

→

“Some practitioners stress that a dummy atom should retain at most only one angle term (Atom1–Atom2–Dummy) and one dihedral term (Atom1–Atom2–Atom3–Dummy) with respect to non-dummy atoms to yield correct results,25,38 but this is somewhat controversial in the literature.39”

4) It is clear that any comparative study has to draw the line somewhere. Nevertheless, it might be useful to briefly describe NAMD's capabilities (though I concede it is a moving target). Further, it is not clear to me whether all recent developments in AMBER have been taken into account (e.g., the already mentioned work by Giese and York (10.1021/acs.jctc.7b01175))

The alchemical FEP code in NAMD implements a pure dual topology only, and for this reason, it has not been considered in this study. This is because FESetup generates only single topologies as it uses the maximum common substructure to determine the common atoms. We have therefore decided not to describe a software we were not using. As for AMBER the version we have used is 16 and as such does not include the latest changes. Again, we would highly encourage a comparison involving NAMD within its dual toplogy framework, and hope that the detailed supporting materials we provide will help encourage further such studies.

5) p14, l11: I don't quite see why Ref. 67 (certainly an interesting publication) is relevant in this context.

This reference has been removed.

6) I do admit that I find the statement in the second paragraph of p

34 (ll11-23) a bit confusing / disconcerting. On the one hand, the gas phase issues mentioned may also have some relevance with respect to the AFEs (my item 1). On the other hand, I am not too surprised that different programs require different protocols, as, e.g., various forms of soft-core potentials have different strengths and weaknesses.

We raised this issue because it may not be obvious to newcomers to the field that small differences in softcore energy functions may have a significant effect on other aspects of a free energy calculation, for instance the selection of a lambda schedule. We have adjusted this discussion to clarify somewhat.

7) The authors point out correctly that CHARMM/PERT/PSSP does not support the Lennard-Jones force switch, which is the preferred method in modern CHARMM force fields. However, I don't think such a force field was used in the present work, so the regular switch should be adequate. Clearly, this limitation is not pertinent to the results presented in the Manuscript.

As this limitation is not pertinent to the results presented in the Manuscript we have not modified the manuscript to discuss this point.

8) As presently presented, it is very difficult to get an overview of the simulation protocols (e.g., how many lambda states, how many simulations steps per lambda state etc.) Could a table giving this info be added, at least to supp. info?

We have added this information in table 1 in the revised manuscript.

Additional Questions:
Please rate the quality of the science reported in this paper (10 - High Quality / 1 - Low Quality).: 8

Please rate the overall importance of this paper to the field of chemical theory or computation (10 - High Importance / 1 - Low Importance).: 9

Reviewer: 2

Recommendation: Publish after minor revisions noted.

Comments:
In this manuscript, the authors compare the results of calculations of absolute and relative solvation free energies for nine rather small organic molecules obtained with thermodynamic integration using four different software, AMBER, CHARMM, GROMACS and SOMD. Alarmingly, the authors report several serious problems in some of the software used, in particular some unresolved for AMBER. Besides this, they show that the four software give results that agree within 0.2 kcal/mol, which is somewhat larger than the reported. This is a most important study, providing much interesting information. The paper should certainly be published. However, the paper is quite poorly written (probably reflecting that it was written independently by several research groups and not thoroughly checked by all at the end). There are also numerous details that need to be fixed.

1. The general design of the experiments is very strange. The authors spend very much time to discuss small technical differences between the various software. However, the most important parameters affecting the results and are trivial to control are not set equal, e.g. number of lambda values, length of simulations and cutoff distance. This is a total mystery to me. Can you really expect that calculations using a 8 or 12 Å cutoff give results that are identical to 0.1 kcal/mol? This must be discussed and explained. It is a most annoying shortcoming of this otherwise so interesting study.

We initially aimed to make these parameters as similar as possible across codes but concluded that unfortunately, it is essentially impossible to use exactly the same protocols with the different software packages employed. Superficially the parameters may seem the same but implementation details can vary a lot. Thus, instead, we ended up selecting protocols that our groups (which are experts in the various packages) knew to work fairly well. This is why we have spend so much time in discussion these details. We would also like to point out that the main point of our study was to test if the simulation codes will agree on the “same” free energy. We believe that indeed we have shown that this is possible. It may perhaps be interesting for developers to conduct follow up work which attempts to make the protocols more similar to assess underlying algorithmic details, and our supporting files may facilitate this, but this is not within the scope of our study.

As for the lambda schedule it is important to understand that we have chosen TI. As the free energy implementations are so different, in particular the exact form of the softcore potentials employed, the resulting dV/dlambda curves will be very different too. Consequently, the lambda schedule will need to be adapted to those curves.

As for the cutoff distances we need to point out that cutoff implementations can be very different too. E.g. AMBER uses a hard cutoff for van der Waals interactions while other codes may use

shifted or scaled potentials and/or forces. This means that none of those implementations can be expected to yield exactly the same energy and hence exactly the same free energy as a function of lambda, meaning that different protocols are in some sense required (#41)

Thus given that it is not possible to make energy functions identical across codes, we felt it important to validate the extent to which free energies calculated with one particular force field are reproducible across codes. That is, while the exact energy and free energy as a function of lambda will differ strongly by code, the overall free energy of solvation ought not to be strongly dependent on the code.

2. In relation to refs. 28 and 29, J. Chem. Inf. Model. 54 (2014) 108 and 54 (2014) 2794 should also be mentioned, as they provide earlier large-scale tests of ligand-binding affinities with RAFE methods (and are not advertisements of a commercial software).

We have added those references.((#42)

3. For Amber, it should be first explained that the software contains two separate implementations of RAFE methods (i.e. what is called sander and pmemd). It is not clear whether the discussion on p. 10 relates to one or both of the implementations. This must be clearly stated.

We have added text to explain the different implementations in section 2.1 'AMBER'. Sander has been used for vacuum calculations, pmemd for solution simulations. Pmemd does not support vacuum TI in version 16. (#43)

4. Is there any difference in which atoms in the perturbed groups that are allowed to have different coordinates between the different software? It is not obvious that topologically equivalent atoms (the common substructure) should have identical coordinates, at least not for binding free energies.

The implementations we tested here use a single-coordinate framework so the perturbed groups have identical coordinates.

5. What is PSSP softcore?

PSSP is the name of the CHARMM keyword. We use it as an identifier to distinguish it against the softcore implementations in other codes. This is now explained in the text. (#45)

6. It is better to say that 9 transformations were considered in the two directions ("18 transformations" is confusing).

We have clarified this.(#46)

7. What parameters were missing for these simple organic molecules? Atom types and charges should be provided for all molecules in the SI.

We provide this information through the SI. All input files are referenced in the SI. (#47)

8. The switched protocol needs to be somewhat better explained in the main article. If both end states involve dummy atoms the more common approach is to first zero all charges, then run the LJ perturbation and finally insert the new charges, a three-step approach, rather than the two-step approach described. This is shortly discussed on p. 35, but it should be mentioned already in the Methods section.

We discuss this in much detail in the SI. This explanation seemed to be too long for the main text. The reviewer is right that the protocol indeed requires 3 steps but with software like GROMACS this can be simulated with only two input files i.e. 2 steps in the meaning of running only 2 independent simulations. (#48)

9. How were "steep variations in gradients" detected and what criterion was used? What gradient?

This was done by visual inspection of TI gradients without a firm quantitative criterion.

We have updated the text in the manuscript accordingly.

“In some instances, steep variations in gradients were observed with this protocol and additional windows were added to obtain smoother integration profiles.”
→
“In some instances, steep variations in TI gradients were observed by visual inspection with this protocol and additional windows were added to obtain smoother integration profiles.”

10. What exactly is meant by "perturbed group"? The whole molecules shown in Figure 2 or only atoms that are dummy atoms in one of the end states?

We have explained this on page 4. The “perturbed group” are all those atoms which vary by at least one force field parameter between the states. (#50)

11. How can a time step of 2 fs be used if some atoms are not SHAKEn?

We have clarified in section 2.3 SOMD that this is possible due to the use of high atomic masses for perturbed hydrogens.

12. Were all simulations NTP?

Yes. We clarified this in the Methods. (#52)

13. Why was a different PME cutoff used for the various simulations?
14. Why were different lambda values employed for GROMACS?
15. Why were the length of the simulation different for the different software?

This has been answered in our response to Reviewer 1.

16. The "20-step alchemical protocol" should be better described.

We have updated the text with a more detailed description.

17. Why was not SHAKE used for waters with Gromacs.

Waters were constraint as with the other simulation codes. GROMACS uses LINCS though.

18. Why were structures minimized first with SOMD, but not with the other software?

All starting structures were prepared and minimized with FESetup. Individual workers chose to use their “standard” protocol to start off with simulation. We believe this has not further bearing on the result. (#58)

19. What is "a suitable numerical integration method"?

We have updated the text to clarify we used cubic spline interpolation to integrate the TI profiles.

20. The statement "All data was sub-sampled to eliminate correlated data." should be better described.

We explain this in the Methods with reference to the actual algorithm used to eliminate correlated data. (#60)

21. What is n in Eqns. 2 and 4? Are they the same? A SEM error could (also?) be calculated from the triplicate samples? Is it similar (larger or smaller)?

We explain this now and use consistent terminology to avoid confusion. (#61)

22. The first sentence in the results section is totally unexpected, considering that the previous text in the article has concentrated on relative binding free energies (for example the abstract talks only about relative free energies).

We have changed the wording to make it clearer that the purpose of the absolute simulations is to provide reference values to validate the relative simulations. (#62)

23. Ref. 79 should be described in the introduction, rather than in the results section.

We have moved the sentence describing ref 79 to the introduction.

24. What is "the recommended protocol for each MD code". Recommended by who? References should be given. This should be clarified already in the Methods section. I am not aware that there is any "recommended protocol" for the software I am using.

We rephrased the text to read ‘’The table shows the data from simulations conducted with the protocol our groups considered most trustworthy for the respective MD code used, as discussed in detail in the following subsections.’’ As each group tested several different variations of a protocol for a given MD code, with some giving clearly incorrect results, we merely stress here that the protocols reported are the better ones in our hands for this particular study.

25. What are the error estimates? Are they not standard errors (SE) of the mean? Then a 95% confidence interval would be about twice the SE. Thus, the sentence "are the only protocol consistent within uncertainty estimates" need to be modified and supplemented by a confidence level.

Good point, we have converted the errors into a 95% confidence interval and updated Table 6 to clarify this, and updated the discussion around Table 6.

26. Sections 3.1-3.5 need an introduction.

We added a subsection 3.1 - overall comparison and briefly introduce 3.2-3.5 for code specific considerations.

27. Are the problem with the methane ? neopentane energies not connected to the observed end-point structure problem? This must be discussed.

We have added to the text in 3.1 the sentence

“It is possible that the discrepancies observed with AMBER are partly due to inconsistencies in the end point geometries (see section 3.2)”

28. Is it not very surprising that the absolute free energies seem to be more reproducible than the relative free energies? I would expect the opposite. Why is it so?

We have added a paragraph in the Discussion and Conclusions section of the manuscript to discuss why this may be the case.

29. "It appears to be the most efficient approach for GROMACS" is an inaccurate statement. In efficiency, you normally consider also the time consumption and it is twice as large for the split protocol. Therefore, the unified protocol should have been run for twice the time or twice the number of lambda values before you can say anything about efficiency.

We have changed the wording to make this point clean. (#69)

30. It seems totally meaningless to discuss efficiency as the various simulations were not set up the same way (same number of lambda values and same length of simulations) and consequently they did not give the same precision. For a fair comparison (which indeed would be very interesting), these parameters must be fully controlled.

We agree that one cannot draw solid conclusions from the data presented here given the variability in protocols. We have included a paragraph to briefly discuss this issue because inevitably readers may seek to make this analysis. We clearly write ‘’The primary focus of this work was to achieve low statistical errors to establish if codes are able to reproduce free energies. We have not investigated the efficiency of the respective protocols as this would require further, complex investigations.’’

31. "all in red" on p. S-3 in the SI does not apply.

Fixed. (#71)

32. "van der Walls" on the on p. S-3.

Fixed. (#72)

33. I suppose that what is called 1-step and separated protocol in the SI is the same as the unified and split protocols in the main article. The same nomenclature should be used throughout the article.

Fixed. (#73)

34. Should not results in Tables S2 be given with two decimals (cf. Table S5), reflecting the precision of the data. The precision of each calculation should be explicitly stated.

Fixed to have 2 significant decimals. (#74)

35. Tables numbering in the supplement is inconsistent and many Tables and Figures are missing important information (in particular which software was used). This made it very hard to follow these sections.

Fixed. (#75)

36. It is said that input files are provided, but I do not find any.

The hyperlink to the GitHub repository is in the information for the SI. We have also added this link a few times in the main text. (#76)

37. Likewise, it is said that some general recommendations will be given, but it is not clear which they are.

We rephrased slightly a sentence in the introduction

“We will discuss the reversible work results obtained with these packages and make recommendations on (...)”
→
“We will discuss the reversible work results obtained with these packages and make observations on (...)”

38. It would be interesting to see whether some overlap measures would indicate any of the problems reported in this study.

We added to the conclusion the following sentence

→ In particular it may be interesting to apply overlap measures to explore the relative efficiency of the different protocols.81

Additional Questions:
Please rate the quality of the science reported in this paper (10 - High Quality / 1 - Low Quality).: 6

Please rate the overall importance of this paper to the field of chemical theory or computation (10 - High Importance / 1 - Low Importance).: 10

Reviewer: 3

Recommendation: Publish after minor revisions noted.

Comments:
This manuscript reports benchmark studies on hydration free energy calculations of small organic solutes using molecular simulation softwares including AMBER, CHARMM, GROMOS, and SOMD. With carefully validated simulation setups, all simulation engines can achieve decent consistency in free energies of hydration. Caveats and plausible source that led to discrepancies among different simulation protocols are given to facilitate future free energy studies to ensure comparability.

This is a solid piece of work; the modeling results are carefully analyzed and interpreted. I have no major revision suggestions except a few comments.

1. It is noticed that the free energy simulation setting, such as number of lambda windows, simulation timescale, truncation distance of long range electrostatics interactions, etc., are very different for the four MD software packages examined. Since these parameters will significantly affects the consistency of the free energy results obtained, it might be useful to discuss how they were chosen. In the conclusion section, the manuscript discusses that the simulation protocols used for CHARMM and SOMD is much computational efficient than that for GROMACS, though the latter gives the most accurate results. I wonder how much accuracy is brought about from more extensive sampling. Has any comparison been taken to compare the free energy results from simulations of similar timescale in different MD packages.

We have addressed these issues in answering reviewer 2 question 1.

2. For absolute hydration free energy calculation results reported in table 1, free energies obtained by SOMD exhibits the largest number of extreme values compared to other simulations protocols, the uncertainties are also larger than others. Is it not yet fully clear to me whether this is due to convergence issue or bond constraint. Could you further clarify?
For tables 1 and 3, it could be useful to include experimental hydration free energy values for the sake of precision comparison.

Regarding the extreme values for the absolute hydration free energy (tab. 1) we have this situation:

molecule	min fe	max fe
methane	GRO (2.44)	SOMD (2.52)

methanol	GRO (3.51)	SOMD (3.70)
ethane	GRO (2.48)	SOMD (2.56)
toluene	SOMD (0.55)	AMB & GRO (0.72)
neoP	CH & GRO (2.58)	SOMD (2.71)
2 MF	SOMD (0.39)	GRO (0.51)
2 MIND	CH & SOMD (6.06)	GRO (6.35)
2 CPI	AMB (6.05)	GRO (6.54)
7 CPI	AMB (5.66)	GRO (6.52)

where *min fe* is the minium value in abs. hydration free energy, *max fe* the maximum one and I reported the absolute value.

From here we can see that SOMD has 7 extreme numbers out of 9 (4 times is the max 3 times the min), GROMACS has 9 extreme numbers out of 9 (5 times is the max 4 the min).

Amber appears only three times and Charmm twice. So GROMACS and SOMD have the largest extreme values. This does not seem a code specific issue, and we do not unfortunately have a clear explanation. We hope that by making all our input files availables others may be able to progress some unresolved issues of our reproducibility study.

In our view a comparison with the experimental values is irrelevant, as accuracy relative to experiment is determined by the force field. It is entirely possible (given force field limitations) that a better protocol might give worse results relative to experiment, or vice versa.

3. The fact that the MADs of relative free energy simulations are larger than those of absolute free energy simulations is somewhat unexpected to me, considering that perturbation simulations generally would have better cancellation of systematic errors and better overlap of conformation space. Any explanations?

We have added a brief discussion of this in the Discussion section.

4. There a few inconsistent use of notations, e.g. tab. vs. Table, fig. vs. Figure. In addition, the volume, year, and page information is missing in ref. 22. All references should also be double checked to make sure that they comply with the ACS citation format.

We have addressed and fixed those issues. (#82)

Additional Questions:
Please rate the quality of the science reported in this paper (10 - High Quality / 1 - Low Quality).: 9

Please rate the overall importance of this paper to the field of chemical theory or computation (10 - High Importance / 1 - Low Importance).: 9