Response to Reviews of "Assessing Conformer Energies"

Geoffrey Hutchison[1] and Dakota Folmsbee[2]

[1]University of Pittsburgh
[2]Affiliation not available

June 11, 2020

Referee #1 (Report will be published at publication of the article)

Summary

> The paper presents some interesting results, but is riddled with missing or misattributed data, typos, grammatical errors (particularly agreements for single and plural nouns) and errors in the references. The paper should be carefully corrected before resubmission.

As the reviewer is, no doubt aware, the article was submitted as part of a special issue for *IJQC*, using the interactive Authorea publishing system, including interactive figures, raw data, etc. We note that almost all the issues noted derived from bugs in generating static PDF files from the interactive website and discuss below. With the help of the Authorea developers, we believe we have fixed most, if not all of these issues.

In the revised manuscript, we have gone through the text thoroughly to improve the quality of the writing, fix a few minor typos, etc.

> The key omission in the paper is any attempt to provide confidence in the deductions made about the differences in accuracy between the methods compared. Confidence intervals on each of the estimators, estimates of success rates and their errors, and pairwise hypothesis tests, at a minimum, must be added before publication. With this data in hand the new version can make quantitative estimates of the differences between the methods.

See comments below.

Detailed Report

> When superiority of one set of results is asserted over another it is simply not acceptable to report raw performance numbers and state that the biggest/smallest is the best. All the results in this paper are sample results and therefore have an associated error, which must be reported. Given this error, when two methods are compared measures of the significance and the impact of the difference must be reported (hypothesis tests, effect sizes, confidence intervals etc.).

Across computational chemistry, there are **countless** benchmark studies similar to ours, often with many fewer data points, which compare multiple methods and do not discuss statistical significance or confidence intervals. Indeed, we are unaware of other published quantum chemistry or machine learning of quantum chemical property papers that report confidence intervals, error bars, etc. as requested by the reviewer.

Nevertheless, we agree with the reviewer and provide such metrics in the revised manuscript and hope the field will broadly adopt such practices. We note that little of our discussion or conclusions required revision

– in part because our sample sizes were intentionally large. We have added a brief discussion to the methods section on calculating confidence intervals using bootstrap methods.

> The R^2s shown throughout the manuscript have an error that can be calculated analytically - this must be provided along with the raw results. Confidence intervals for the estimators like mean and median must also be added.

The reviewer is incorrect – while a confidence interval of an R^2 can be calculated analytically, the appropriate metric would be a confidence interval on the **median** R^2 and Spearman ρ values that we use for comparison and discussion. Since the distributions are unknown and clearly non-Gaussian (e.g. Figure 2) there are no analytical formulas.

In the revised manuscript, we have added a discussion of calculating confidence intervals using bootstrap methods and have added them to appropriate figures and tables. As usual, our code is provided as supporting information in our GitHub repository.

> Figure 2 was missing from the PDF.

As noted above, there were some issues with the Authorea PDF provided to the reviewer. We can assure the reviewer that the Figure is available in the PDF as well as the interactive web version.

> What is the y-axis in Figure 3? Counts of molecules in that timing bin?

In the revised manuscript, we have clarified that the y-axis indeed represents the counts of molecules and provided an updated Figure 3.

> Figure 4: R^2 has a calculable error and should be included in the plot.

> Figure 4: ANI family methods are not labelled/included in the plot, but are mentioned as being there in the text.

The reviewer is using the static PDF image – naturally it is not possible to annotate all 22 points. For example, the density functional methods form a tight cluster and are not all labeled. The interactive web version through Authorea allows anyone to see labels and expand regions as needed. In the previous version, the ANI points were labeled with a unique color, as indicated in the figure caption.

In the revised manuscript, we have added a new PDF graphic for Figure 4 including confidence bars for the median R2 (y-axis) and a text label for the ANI methods.

> Figure 7: Add error for R^2.

In the revised manuscript, both the R^2 and MARE statistics have confidence intervals indicated.

> Figure 8: Missing from the PDF

We can only assume this was an issue with the PDF version the reviewer received. The PDF we generated ourselves (and posted on ChemRxiv) had all figures. Both the interactive Authorea version and the current PDF generated by Authorea have all figures included.

> Test Set selection - "the training set was the first five conformers" - how were these conformers generated and ranked?

As noted in the manuscript, the conformers "were initially created from a set of 250 diverse poses with maximal heavy-atom root mean squared deviation (RMSD) using Open Babel, and at most 10 poses were selected based on the lowest heat of formation calculated by PM7." The five poses chosen for training of BOB, BAT, BATTY models were thus random.

> "any molecules with fewer than five conformers was omitted" - how many were omitted?

In the revised manuscript, we clarify that these molecules had their conformers added to the training set but are absent from the test set.

It is understood/guessable why some DLPNO calculations did not converge?

As noted in the methods section of the manuscript, we chose the widely-used cc-pVTZ basis set for the DLPNO-CCSD(T) calculations for comparison with a variety of other benchmarks using coupled-cluster thermochemistry. The molecules include some with iodine, which is not available in that basis set.

In the revised manuscript, we have clarified this further in the methods and as well in the results and discussion.

The statement "deriving accurate rankings.." is not supported by the previous sentence. Boltzmann weighting relies entirely on energy, so how is accurate ranking going to help improve "computational predictions"?

As the example in Figure 9 indicates, small deviations in relative energies and thus rankings can significantly affect the Boltzmann-weighted averages for properties such as dipole moment. This was the whole reason for including this section of the manuscript - several discussions of our preliminary work suggested such small differences were irrelevant. The work by Vo and Johnson, our colleagues at Pitt, give an indication of why accurate conformer rankings are needed.

In the revised manuscript, we have omitted the phrase "deriving accurate rankings" as it is redundant with the phrase "accurate relative energies" in the same sentence.

The text in "Comparison of single points vs. DLPNO-CCSD(T)" should be tabulated and is redundant with text elsewhere in the manuscript.

The reviewer is referring to a bullet list in the manuscript. We do not see how providing a table version would be any different than the current formatting. While the methods and citations are of course mentioned in the methods section, we believe it is helpful to the reader to list the computational methods considered in the work. After all, not everyone reads the methods section in its entirety.

(Editor's Note: It is clear that any issue with the references is due to Authorea, so the reviewer is not completely correct here)

The writer appears to be getting to know Endnote's referencing scheme:

The negative tone of this comment is completely unnecessary. As noted above, and by the editor, as part of the special issue, we used Authorea to import references. Before submission, we even checked through the BibTeX data for accuracy, since the platform is new to us. We have changed the format for our Authorea export, but leave the formatting of references to the capable hands of the *IJQC* editorial and production team.

Two citations in the Results page are in text form.

There are errors in the following citations: 7, 37, 59, 61 (content missing entirely here)

The SciPy citation, 59, is not the standard one - justification for using this citation should be provided.

The citation for pybel is a '?'

See above. We note that SciPy now has an official citation in Nature Methods and have updated it accordingly.

Comparison of timing - the details on the hardware should go into the Methods section.

As customary in many papers discussing timing of quantum chemical calculations, the details on the hardware are provided in the context of the timings. In the revised manuscript, we added the details on hardware to the methods section – but have repeated the details in the discussion as an aid to the reader. After all, not every reader will remember every detail from the methods section.

> xray - X-ray

We have revised the text to indicate these are from "experimental crystal structures," since data using cryo-EM methods are becoming available.

> Mllr-PLesset = Moller-Plesset

As noted above, this appears to be an issue with export from Authorea to the "ACS style" template. We have resolved the problem with Authorea and the new export style should provide all correct accents and special characters.

> BATTY/n = BATTY

The reviewer is mistaken - this is not a typo. BATTY/n is an indication of the normalization used with the BATTY descriptor. Based on the results, we strongly encourage the use of such normalization.

In the revised manuscript, we have clarified this description.

> CPU Time in Table 1 cannot be 0.0. Perhaps each method should have time scaled by forcefield time set to 1?

Of course none of our data indicates 0.0 second timings. This appears to be an issue with export from Authorea - the raw data is available in the Authorea version as well as our GitHub version. We trust that production of the final published article will produce timings with appropriate formatting.

Referee #2 (Report will be published at publication of the article)

> Paper by Folmsbee and Hutchison is a great example of reproducible benchmark paper and weel suited for IJQC special issue. ... This paper provides computational chemists with a substantial body of high accuracy data. Overall this paper could serve a solid practical guideline for applying approximate computational methods to a problem of conformer search.

We appreciate the positive comments from the reviewer.

> 1. All ANI models were are fitted to wB97x DFT functional data *minus* D dispersion term. This is done because dispersion is an analytical ad hoc correction. The intention that at the run time dispersion should be added back. D3 could be easily computed with ASED3 code referenced in our GitHub.

We did not find that the ANI models correlated perfectly with ωB97X, and have concerns about applying an empirical correction to ANI.

A *post hoc* dispersion D3 correction to the ANI-1x and ANI-2x models slightly improve the performance over the non-dispersion corrected models, although the results were not statistically significant. On the other hand, it is **not** possible to calculate ωB97X dispersion corrections using any of the standard calculators (e.g. ASED3) as suggested by the reviewer, since no existing D3 calculator has parameters for ωB97X. In our manuscript, we can provide the corrections, since we have run the same molecules / geometries with ωB97X-D3 and can extract the dispersion correction directly.

In the case of the new D4 dispersion correction, we can use the `dftd4` calculator to apply the correction for ωB97X to the ANI-1x and ANI-2x models - but the results surprisingly get **worse**. Again, the results are not statistically significant.

In the revised manuscript, we added a new table summarizing these findings, as well as a short section discussing the challenge of doing *post hoc* dispersion correction to a ML model.

> 2. ANI timings are simply wrong. Therefore the TOC and Figure 4 are misleading. ANI timing is at least 100 times faster. The author's script is re-loading all python dependencies and compiles

> the neural network model for every conformer. This takes 2.45 out of 2.5 seconds of the run. Even with sequential energy evaluation on a CPU, it should be around 0.05s for the 2x model and probably ˜0.025s for 1x/ccx . . . . Therefore the recommended use is to load all conformers and evaluate them at once.

We have discussed this point with the ANI developers repeatedly. Many of the quantum codes also require time to load code libraries, create atomic orbital basis sets, etc. Programs like ORCA have a wide variety of options that change performance (e.g., DFT grid size, use of RI approximations, etc.) Furthermore, many force fields perform atom typing once per molecule and also have a fixed-time overhead.

Our timings for all methods do **not** involve loading python dependencies or launching a program. Traditional quantum chemical programs report time for the calculation, which is what we have used in our scripts for ML methods (i.e., creating a neural network model is similar to creating an initial guess for the self-consistent field iterations). We believe we provide a fair comparison for single-point energy evaluations across all our methods. If the ANI authors feel a different comparison is worthwhile, they are welcome to publish a rebuttal or alternative benchmark - our data and scripts are all online.

In the revised manuscript, we have added a section on "batch evaluation" for force field and ANI methods. We note this does not change the accuracy of the methods, nor our resulting conclusions. It only somewhat "bends the curve" since the force field and ANI methods improve in speed – but since the x-axis of our figure is logarithmic, and force field methods are still 200-300x faster, qualitative descriptions are similar.

> I think there is a fundamental flow of logic here, that ultimately hurts the value of this paper. In practical research settings where conformed sampling is used, there is no access to 3D geometries obtained with high-level QM methods. Therefore, I think the meaningful comparison would be conformed energies with geometries obtained by respective approximate methods.

As noted in the introduction, we performed this exact analysis in our previous paper in *IJQC (Kanal et al., 2017)*. We found reviewers (across several journals) and many readers found the analysis of differing optimized geometries confusing. Consequently, we focused this work on a limited question - "how well do single points correlate?" This limited question was, in fact, suggested by multiple reviewers of our previous paper.

We agree with the reviewer that analysis of optimized geometries from different methods is a useful concern, but believe it is beyond the scope of this paper – which already considers ˜6500 single points and over 30 methods.

In the revised manuscript, we have discussed the consequences for conformer sampling in the conclusions – that is, we suggest readers use methods such as the ANI models, GFN methods, or even faster density functional methods such as B97-3c to optimize and rank relative conformers.

> 4. A comparison between BOB/BAT/BATTY and ANI is also one-sided. BOB models are just molecular scorers, they just give you a number. In contrast, ANI and force fields are true automatic potentials with forces and analytic hessian. We can do geometry minimization, MD, etc.

This is not quite correct. While BoB, BATTY, etc. were designed to solely produce a molecular energy, it is, in principal, possible to perform optimizations. We are currently working on a new work, providing comparisons of different ML methods on the optimization task. Again, that is beyond the scope of this present work.

> 5. There is also a small pesky bug in the authors' scripts. They use different conversion factors au to kcal/mole in different places, therefore some of the energies are inaccurate to ˜0.2 kcal.

The reviewer highlights a key reason we focus on the R^2 and Spearman metrics for this work - such scaling differences are removed in both metrics. Moreover, the ANI developers noted a difference between their

published version of the ANI-2x model and the pre-release version we used initially. For example, in the initial manuscript, molecules containing chlorine were not included.

In the revised manuscript, we have revised the ANI-2x data, fixing conversion factors as well as using the final ANI-2x model on all available molecules. Very little changed in the tables, figures, or conclusions.

References

A sobering assessment of small-molecule force field methods for low energy conformer predictions. (2017). *International Journal of Quantum Chemistry*, *118*(5), e25512. https://doi.org/10.1002/qua.25512