# Referee # 1

* **R1:1. In this manuscript, the experiment/observed values (y) are used for computing RMSE of each model. However, no specific information regarding y is given. Without knowing the experimental method and associated uncertainties, it is difficult to judge how significant the RMSE of each model is. That being said, I understand that is not the main scope of this manuscript and the information may be revealed elsewhere. My suggestion is that the authors address this issue explicitly either with a few sentences or in a separate supplementary file.**
  + Author response:
    - This is an excellent point and we thank the reviewer for bringing this topic to our attention. The uncertainties and errors that are associated with experimental measurements for (y) are very relevant. We do mention this topic very briefly in the first manuscript in Section 2.1 (second paragraph) where we discuss the notion of the epsilon term that captures all factors that influence (y) aside from the regressors, (x). This treatment of the topic, however, is too brief. In the revised manuscript, *second paragraph in Section 2.1*, we explicitly address both the importance of measurement uncertainty and show how under conditions of normality these uncertainties may be nullified for statistical models*. In Section 3 we further report (via a Shapiro-Wilk test and a normal probability plot) that our assumptions of normality are valid for this analysis and that such uncertainties and errors are a net zero impact to the forecasting results.*
* **R1:2. The method ID of exp is somehow confusing. I suggest to replace exp with other words such as hyb(rid).**
  + Author response:
    - Mapping all explicit modeling methods to (hyb) is not completely accurate and MM-PB/SA methods are technically not the same as the hybrid entries into the SAMPL4 challenge. We are, however, happy to change the actual hybrid models that were previously flagged (exp) to (hyb). *In the new manuscript, we have made these changes.*
* **R1:3. In Figure 1 and 2, the 'Best performing method' simply refers to imp-2. It would be better to include the standard deviations with the red lines, such as a striped area, to complement the authors' argument of pruning process.**
  + Author response:
    - Excellent suggestion. *These changes have been added in the new manuscript to Figures 1 and 2.*
* **R1:4. For Figure 3, it would be better to move the method ID to the left.**
  + Author response: *Changes have been made to Figure 3*
* **R1:5. Some of the compounds were difficult to estimate in SAMPL4 challenge. It would be appreciated if the authors can address the probable reasons with simple notes. This may be analyzed by Mobley et al, so a few words such as geometry/configurations might be sufficiently informative.**
  + Author response: Good additional information to add. *We have added the relevant information in the last paragraph of Section 3.2.*
* **R1:6. Page 13, line 42, typo: Was -> was**
  + Author response:
    - *Typos have been addressed*
* **R1:7. Please unify the format of references.**
  + Author response:
    - *References have been unified*

# Referee #2

* **R2:1. First of all, I do not think that the Journal of Physical Chemistry is an appropriate journal for presenting this work since this paper does not provide any new physical insight on solvation free energy. In addition, at least in my opinion, this method is not so attractive. In the end, what is achieved is the slight improvement in RMSE of solvation free energies from 1.24 kcal/mol (alc-3) and 1.15 kcal/mol (imp-2), which are already good in light of experimental uncertainties and force field inaccuracies, to 0.82 kcal/mol. And in addition I’m very suspicious about the generalization ability of this method for unseen data.**
  + Author Response
* **R2:2. What is the ultimate goal of this approach assuming that a large number of solvation free energy data are available? Is it to propose a single combination of two (or so) specific computational methods with the corresponding beta value for the general use? When can one say that this is achieved?**
  + Author Response
* **R2:3. The computational efficiency is another concern. Would it be possible, just combining quite cheep computational methods, to achieve a better RMSE value than the one from a single expensive but accurate method?**
  + Author Response
* **R2:4. Jargons are frequently used to average readers who are not familiar with the concepts in machine learning approaches. For example, the author should briefly explain what is the Bayesian Information Criterion, and how and why it is related to the posterior probability and information content of the model. In particular, this quantity is not well defined in the manuscript since R^2 in equation 4 is unspecified.**
  + Author Response
* **R2:5. The comparison with different ensemble approaches presented in Section 3.3 is not illuminating since there is no explanation on those approaches. For example, is each of them using just a single combination of the two best methods as in BMA? If so, are those two methods the same as those in BMA (alc-3 and imp-2)? More explanation should be added for a fair comparison of different approaches.**
  + Author Response:
* **R2:6. (Minor) In the beginning of the manuscript, it is stated that the method is applied to the solvation free energies for 45 small molecules. However, there are 52 solvation free energy data in Figures 4 and 5. Why are these numbers different?**
  + Author Response:
    - This is a good point that was overlooked. The confusion lies in the fact that the original SAMPL4 challenge used 52 samples; this is why one of the samples is labeled SAMPL4\_052. However, several molecules were dropped from the final study due to a variety of factors: “problems with experimental values, SMILES strings, or structures for a number of compounds were uncovered, resulting in removal of some compounds from the challenge.” [Mobley, 2014]. In the revised manuscript we now point out this detail in *Section 2.4*.

# Referee # 3

* ***R3:1 - I am slightly irked by some of the algorithm's results shown in Figures 4 and 5.  Perhaps it is because I am misunderstanding some part of the Methods section, but nevertheless here is my observation: the authors' optimal model (labeled as "bma") actually performs worse than its two constituent methods ("alc-3" and "imp-2") on not just one occasion, but rather several (molecules 002, 017, 024, 030, and 047).  Based on my reading of the manuscript, I am assuming the following:***

***(1) A training set consisting of solvation free energies is created by randomly selecting 26 of the 52 SAMPL molecules (and the set doesn't change from this point forward).***

***(2) Each possible model consists simply of a linear combination of estimators (i.e., methods).***

***(3) The estimators are pruned from the possible models using the training set until an optimal model consisting of just two estimators remains.***

***(4) The optimal model therefore consists of some sort of linear combination of alc-3 and imp-2 (or rather a family of such linear combinations).***

***If all of the above are true, then it is difficult for me to see how the (median) RMSE for bma could be larger than the maximum RMSE of {alc-3, imp-2} for a given molecule \*unless\* that molecule was not in the training set.  So long as the authors clarify these points and explain these observations in a satisfactory way, then I would approve of publishing their manuscript.  Perhaps they could also indicate in Figures 4 and 5 which molecules are/are not included in the training set (provided that assumption #1 above is correct).***

* + Author response:
    - This is an astute point and we thank the referee for providing us an opportunity to address this observation. The referee’s assumptions (1-4 above) are correct but missing an important piece of contextual information. The only addition to add to the list is that these steps are repeated over 100 iterations (*as we indicate in the beginning of Section 3 of the current manuscript*). Specifically, optimal models are constructed on a randomly selected set of molecules (i.e., the training set). The compounds not used to train the model form the test set. After performance is assessed, this process is repeated such that a new random selection of molecules becomes the training set for the model. Reported results I represent a mean RMSE of these 100 iterations.

The referee’s assumption is therefore correct: the performance for the listed compounds - including the referees highlighted molecules 002, 017, 024, 030, and 047 – are based on performance for these molecules when these compounds are NOT in the training set.

* + - In the revised manuscript, we discuss the notion of optimality and “best performance” in both the *first paragraph in Section 2* and when discussing the results in *Section 3.2 (see second and third paragraphs).* Further, in *the new section 3.4.,* we directly address some of the subtleties of interpreting performance for aggregated estimates; this section directly addresses the concerns expressed by the referee.
* ***R3:3 Brief postscript: I found a few small typos here and there (e.g., second sentence in the abstract should read "wide range \*of\* methods), but these can be corrected in the proof stage.***
  + Author response:
    - All typos have been addressed

# David Mobley’s Comments

1. Domain of applicability: In some sense this is a type of machine learning, so the ability to make predictions hinges on domain of applicability, and especially on ensuring that the compounds to be predicted are drawn from the same distribution (i.e. chemical space covered by the prediction compounds) as that for the training compounds. Here you do this by taking the set of molecules from SAMPL4 and dividing them randomly into two groups. However, this is much harder in general, since (a) people deliberately try to design new compounds to be different from old ones (such as in drug discovery), and (b) people/techniques have particular domains of expertise/interest/relevance, so experimental measurements from researcher or method X are likely to be on different compounds than those from researcher/method Y. Did you consider assessing the impact of that here, or at least discussing it? Particularly, you could imagine instead of breaking the SAMPL4 set into two groups at random, you could cluster compounds by chemical similarity and take the 26 "most similar" as the training set and the 26 "least similar" as the test set to get an idea of how set-dependent the results would be. Perhaps there's no time for that here, but likely this should at least be discussed as a possible pitfall?

2. Education: It seems like a major point of this work is to educate people on how to do this type of thing/show that it can be done. If that's the case, I think more work needs to go into clarity of terminology/definitions/etc. Perhaps source code could also be provided. To be specific, I find the code in "Algorithm 1" and "Algorithm 2" very hard to read and there are many things there I don't understand. For example, what does the left arrow mean? What is "r" in algorithm 1? Why is the rmse set to infinity initially? I think it would be good to try and carefully edit the whole thing with an eye to how to explain to/show to a novice what this means and how to do it... (I could edit at that level if given enough time, but presumably there's not enough time for that right now.)

3. Predictions: It's worth discussing somewhere how one would imagine using this to actually make predictions of some target property, let's say solvation free energies. Are you actually going to have to bring all of the methods of interest online in-house? It seems like it's one thing to do this when you already have all of the data from people who have run their different approaches, but what would you do if you wanted to predict solvation free energies, or some other property, for, say, a new SAMPL challenge?

5. Pitfalls: I think (along the lines of #1) the paper could benefit from a bit more discussion of the potential pitfalls of this type of approach. i.e., if you're going to do this carefully, what do you have to do to avoid overfitting and try and maximize the odds that (a) the result actually can be used to make predictions, and (b) you learn something in the process...

Also, just to confirm one other thing -- you're always training on the selected 26 compounds and testing by how well the other 26 are predicted, and the second 26 have no impact at all on what models are selected, correct? (Seems to be what it says at end of p9, but confirming.) And all statistics such as in Table I are for only the "prediction 26"?