We would like to thank the reviewers for their careful reading of the manuscript and their helpful comments. Revisions to the text are marked in red color. Specific changes in response to reviewer comments are provided below.

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.

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 2.1 and 3 we show (via a Shapiro-Wilk test) that based on assumptions of normality, errors associated with measurement error do not affect aggregated forecasting results.

R1:2. The method ID of exp is somehow confusing. I suggest to replace exp with other words such as hyb(rid).

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 have, however, changed the actual hybrid models that were previously flagged (exp) to (hyb). In the new manuscript, we have made these changes and clarified the syntax.

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.

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.

The suggested 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.

We have added the relevant information in the last paragraph of Section 3.2.

R1:6. Page 13, line 42, typo: Was -> was

This and other typos have been fixed.

R1:7. Please unify the format of references.

References have been unified.

Referee #2

R2:1. 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?

We address this question more clearly in the revised manuscript (see the new Sections 2 and 3.3 of the revised manuscript). The goal and benefit of this framework is twofold. The first goal and benefit is focused on performance: we demonstrate that the approach is able combine different methods to provide more accurate and reliable performance for estimating hydration free energy (Sections 2). Additionally, this benefit provides the ability to potentially replace more computationally expensive methods with an aggregated ensemble of computationally cheaper methods (Sections 3.2). The second goal and benefit of this framework, which is perhaps more significant, is to provide a rigorous tool for assessing and tuning methods. More specifically, the benefit of a statistical model built through this framework is that the model defines a consistent, interpretable linear relationship between its constituents that can be used to test and refine these individual methods independently and in-concert of each other. In this way the optimal ensemble is a tool that can help facilitate an iterative design process to help methods gradually converge and eventually replace the statistical black box.

R2:2. In my opinion, I'm very suspicious about the generalization ability of this method for unseen data.

The concern for the generalizability of the method is a fair point. In the revised manuscript, Section 3.3, we show that the method is successful in generating a very robust, viable model for the near entirety of the SAMPL4 challenge suite of chemical compounds. This effort is based on 100 iterations of cross-validation studies. We do see some indications in two chemical compounds where the model loses performance and in these examples we elaborate on how to interpret these findings and how the model itself can be used to further tune and refine the methods that it combines (last paragraphs of Section 3.3).

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?

The computational effort for deriving the optimal ensemble is very low and is comparable to most standard regression methods. Additionally, estimating with the ensemble is very straightforward. The bulk of the computational efforts are thus based on the individual methods (i.e, explicit and implicit methods).

In principle, the answer to the question is yes. It is feasible to combine several computationally inexpensive methods to outperform a single, more expensive method. In the revised manuscript, Section 3.3, we comment on this very point. As an example supporting this concept, we note in this section that the optimal ensemble combines a computationally cheaper method (imp-2) to a more expensive method (alc-3) to provide a 34% improvement to predictive performance. The benefit to alc-3 is significant and comes at little additional computational cost.

R2:4. 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.

We thank the referee for pointing this omission out. In the revised manuscript we directly address the Bayesian Information Criteria and provide more complete definitions for the R^2 term (see revised manuscript Sections 2.1 and 2.2)

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.

We have added significantly more detail in Section 3.4. In current version and revised paper, we describe how the performance for these alternate ensemble-based approaches is calculated identically to BMA (paragraph 7, Section 3.4). The referee brings up an excellent point with respect to specifying the actual ensembles that were generated in this comparison. We now provide this information in Table III of the revised paper.

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?

This is a good point that was overlooked in the original version of the paper. 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).*

We thank the referee for providing us an opportunity to address this observation. The referee's assumptions are correct: the performance for the listed compounds - including the referees highlighted molecules 002, 017, 024, 030, and 047 – are based on estimates for these molecules when these compounds are NOT in the training set. The only addition to add to the assumption list is that these steps are repeated over 100 iterations (*see Section 3 of the current manuscript*). Specifically, the optimal model is evaluated 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 in Section 3 are based on a mean RMSE of these 100 iterations.

To help clarify some of the potential confusion, in the revised manuscript we explicitly define the notion of optimality and "best performance": see *second paragraph in Section 2.* Further, in *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.*

We have reviewed the manuscript and fixed this (and other) typos.