Bayesian Model Averaging for Ensemble-Based Estimates of Hydration Free Energies

Luke J. Gosink, Christopher C. Overall, Sarah M. Reehl,
Paul D. Whitney, and, David L. Mobley, Nathan A. Baker*

*To whom correspondence should be addressed. Pacific Northwest National Laboratory, Computational and Statistical Analytics Division, PO Box 999, MSID K7-2, Richland, WA 99352. Email: nathan.baker@pnnl.gov, Phone: +1-509-375-3997

Keywords: hydration free energy; SAMPL4; statistical design of an ensemble; prediction

Abstract

This paper investigates the use of an ensemble technique called Bayesian Model Averaging (BMA) to estimate hydration free energies in small molecules. There is a diverse set of methods for predicting hydration free energies, ranging from empirical statistical models to *ab initio* quantum mechanical approaches. However, each of these methods are based on a set of conceptual assumptions that can affect a method's predictive accuracy and generalizability; e.g, uncertainties due to hydrophobicity [7], surface effects [13], and solvent asymmetries [59]. This work presents an iterative statistical process to construct an aggregate estimate. This process optimally selects and combines estimates (e.g., through a weighted average) from an ensemble of methods to form a single, aggregated estimate. The process begins with an initial ensemble of 17 diverse methods; these methods are from the SAMPL4 blind prediction study conducted by Mobley et al. [60]. The design process evaluates the statistical information in each method as well as the performance of the aggregate estimate obtained from the ensemble as a whole. Methods that possess minimal or redundant information are pruned from the ensemble and the evaluation process repeats until aggregate predictive performance can no longer be improved. We show that this process results in a final aggregate estimate that outperforms all individual methods by reducing estimate errors by 28-91%. We also compare the approach to other statistical ensemble approaches and demonstrate that this process reduces estimate errors by 25%-61%. This work provides a new mechanism for designing and improving the accuracy of estimates for hydration free energies and lays the foundation for future work on aggregate models that can balance computational cost with predictive accuracy.

LG: Share a copy with all participants

1 Introduction

Accurate representation of solvent-solute interactions is essential to simulate most biological phenomena. Consequently, solvation methods that can estimate the thermodynamics of hydration are important across many fields of study, ranging from protein structure prediction [50,78,80] to conformational equilibria [5, 6, 15, 91]. These methods also play an important role in tuning energy parameters for binding free energy calculations as they provide a rigorous test of force field accuracy [11, 17, 42]. Research for developing and parameterizing solvation methods has been active for over thirty years [21,27,44–47,87]. Unfortunately, most research has had to rely on small, public databases consisting of monofunctional molecular libraries that don't adequately represent the highly diverse, densely polyfunctional, and polar molecular compounds that are of biological interest [16,51,58,85]. Resultantly, many solvation methods perform substantially worse than expected when estimating hydration free energies of organic molecules (e.g., drug molecules) [57, 63, 68].

To address this challenge, recent research has broadened the range of molecular data available for developing and validating solvation methods. One prominent example of these efforts is found in the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge studies [17, 28,52,57,60,63,68]. These blind studies provide a rigorous test and assessment of multiple, varied solvation methods based on a suite of complex relevant molecules. By presenting a larger range of hydration free energies and molecular weights than seen in common public data sets, these challenges are helping to advance the development of solvation methods to become more robust and accurate in their estimates for a variety of molecular targets [22, 24, 32].

Consistent across many of these challenge studies is the observation that the top performing methods typically come from a wide range of different strategies [17, 57, 60, 63]. Thus, while solvation methods as a whole are generally improving year-to-year, there is no clear consensus for which methodologies are most successful: across different challenges, top performers have included explicit solvation methods [16, 68, 73], implicit solvation methods [56, 91] and, hybrid methods [26,43] that combine mixed quantum mechanics (QM) with molecular mechanical (MM) approaches. In this context, there is still a significant degree of uncertainty associated with how to best select, specify, and evaluate the set of parameters and mathematical systems needed to accurately estimate hydration free energies.

This type of uncertainty, which affects a wide range of scientific and mathematical disciplines, is referred to as *method selection uncertainty* and is arguably the greatest source of error and risk associated with estimation tasks [4, 20, 62, 81]. One of the most powerful ways to address this uncertainty is by combining an ensemble of varied methods (e.g., through a weighted average) to form a single aggregated estimate [8, 34, 67, 82]. The motivation behind ensemble approaches is based on two principles: 1) most methods in the ensemble possess some unique, useful information; and, 2) no single method is sufficient to fully account for all uncertainties. When modeled correctly, the information and strengths of individual methods can be combined, and their corresponding weaknesses and biases can be overcome by the strength of the group [34, 75, 77, 84]. Ensemble-based estimates are therefore expected to be more reliable and accurate than individual methods, an expectation that has been upheld in numerous examples [8, 30, 34, 61, 67, 76, 82, 84, 89, 96].

This paper investigates the utility of an ensemble approach called Bayesian Model Averaging (BMA) [34] to estimate hydration free energy in 45 different molecular compounds. Leveraging data from the SAMPL4 challenge [60], this work demonstrates how BMA can help statistically design an ensemble from an initial set of 17 diverse methods; these methods consist of a range of both implicit, explicit, and hybrid solvation approaches. Though BMA itself has been applied successfully for prediction tasks across many domains [30,61,76,89,95] this is the first application of the BMA approach to this problem domain. Throughout this paper we differentiate between the terms model and method by reserving the term method to indicate what many computational chemists would call a model for predicting hydration free estimates, and reserve the word model to indicate a statistical model used to combine an ensemble of these estimates.

2 Methods

2.1 Model Specification with Bayesian Model Averaging

For estimates of hydration free energies, a basic BMA approach is to consider a set of solvation methods as a linear system [34, 75, 77]. Let y_i for $i = 1, \ldots, N$ be a series of hydration free energy observations for a collection of molecules, and let x_{ij} denote the i^{th} estimate obtained from the j^{th} prediction method for these observations. For example, given that y_i is the experimentally measured hydration free energy of benzaldehyde, each x_{ij} for $j = 1, \ldots, P$ would be a specific method's estimate for this value. Given P solvation methods, the combination of all x_{ij} forms the numerical ensemble estimate matrix that, along with y_i, defines a linear regression model

$$y_i = \sum_{j=1}^{P} x_{ij}\, \beta_j + \varepsilon_i \tag{1}$$

Here, the parameter vector β_j defines the unknown relationship between the ensemble's P constituents and ε_i is the disturbance term that captures all factors (e.g., noise and measurement error) that influence the dependent variable y_i *other* than the regressors x_{ij}.

In evaluating Equation 1, the objective is to estimate the values β_j that will both fit the known hydration free energy data in y_i and facilitate the ability to make inferences on the hydration free energy of unknown molecules. Many different regression techniques can estimate β_j [12, 36, 54, 79]; however, these techniques commonly generate estimates that vary in their ability to model and infer [18, 29, 34, 75, 77].

The risk and uncertainty associated with using one of these estimates over any other estimate (i.e., for statistical inference) is called *statistical model uncertainty*. Like method selection uncertainty, statistical model uncertainly is also a common source of error in predictive modeling [34, 75–77, 90].

BMA addresses the challenge of statistical model uncertainty by first evaluating all $2^P - 1$ possible models that can be formed from the P estimation methods. Next, each model's estimates for β_j is aggregated together through a weighted average. This process generates an aggregated parameter vector, β_j^{BMA} (Equation 2) that can provide more accurate and reliable estimates than any other model, and can also outperform other ensemble strategies (e.g., stepwise regression) [18, 29, 34, 92].

Formally, there are $k = 1, \ldots, 2^P - 1$ distinct combinations of the P estimation methods, each with a corresponding statistical model, $M^{(k)}$, and parameter vector, $\beta_j^{(k)}$. BMA combines each $\beta_j^{(k)}$, through a weighted average that weights each $\beta_j^{(k)}$ by the probability that its statistical model, $M^{(k)}$, is the "true" model.

$$\beta_j^{\text{BMA}} = E[\beta_j \,|\, \mathbf{y}] = \sum_{k=1}^{2^P-1} E[\beta_j^{(k)} \,|\, \mathbf{y}, M^{(k)}] \Pr(M^{(k)}|\mathbf{y}) \tag{2}$$

In Equation 2, $E[\beta_j^{(k)} \,|\, \mathbf{y}, M^{(k)}]$ is the expected value of the posterior distribution of $\beta_j^{(k)}$. This distribution is weighted by the posterior probability, $\Pr(M^{(k)}|\mathbf{y})$, that $M^{(k)}$ is the *true* statistical model given $\mathbf{y}$. The expected posterior distribution of $\beta_j^{(k)}$ is approximated through the linear least squares solution of the given model $M^{(k)}$ and hydration energy response variable, $\mathbf{y}$. The posterior probability term is estimated from information criteria [75]

$$\Pr(M^{(k)}|\mathbf{y}) \propto \frac{e^{-\frac{1}{2}B^{(k)}}}{\sum_{l=1}^{2^P-1} e^{-\frac{1}{2}B^{(l)}}} \tag{3}$$

where $B^{(k)}$ is the Bayesian Information Criteria for model $M^{(k)}$, and the information criteria itself is estimated [75]

$$B^{(k)} \approx N \log(1 - R^{2(k)}) + p^{(k)} \log N \tag{4}$$

Here $R^{2(k)}$ is the adjusted R^2 model $M^{(k)}$ that indicates the model's goodness of fit for the observations, $p^{(k)}$ is the number of methods used by the model (not including the intercept), and N is the number of hydration free energy values to be predicted (i.e., the number of molecules).

The resulting parameter vector, β_j^{BMA}, obtained from Equation 2 helps to address model uncertainty by accounting for all systems of linear equations that can model the relationship between the measured hydration free energy values y_i and values x_{ij} predicted by each solvation method j. Perhaps more importantly, β_j^{BMA} can be used to estimate new hydration free energy values for unmeasured molecules by combining new x_{ij} estimates.

2.2 Ensemble Pruning and Statistical Design

2.2.1 Model and Method Pruning

The inclusion of all $2^P - 1$ models in Equation 2 is not necessarily beneficial for predictive performance. While some models will provide accurate information that will boost the ensemble's predictive accuracy and robustness, many models will be misspecified. The cumulative effect of these misspecified models, despite the fact that they are down-weighted via low posterior probabilities, can erode the ensemble's overall performance [34, 53, 55, 61, 66, 71, 77]. Madigan and Raftery [53] present an ensemble pruning approach referred to as Occam's Window that eliminates models based on Bayesian information criteria. The premise of Occam's Window states that if a model's information content is so low that it predicts $\mathbf{y}$ far less well than the best models, it should be removed from aggregation. Formally, Occam's Window defines:

$$\mathbf{A} := \{M^{(k)} \in \mathbf{M} | BIC^{(k)} - BIC^{(min)} < 6\} \tag{5}$$

where $BIC^{(min)}$ denotes the model, $M^{(k)}$, with the lowest BIC; low BIC indicates higher information content. The value "6" is a constant based on Jeffreys' [41] and Raftery's [75] assessment of Bayes factors for comparing models; this constant ensures all $m_k \in \mathbf{A}$ meet a minimum statistical information criteria for the aggregation process. Constraining Equation 5 to $\mathbf{A}$ accelerates the evaluation of Equation 5 and improves BMA's predictive capability [53, 77]. Restricting the ensemble to $\mathbf{A}$, Equation 2 is written:

$$\beta_j^{BMA} = E[\beta_j \,|\, \mathbf{y}] = \sum_{M^{(k)} \in \mathbf{A}} E[\beta_j{}^{(k)} \,|\, \mathbf{y}, M^{(k)}]\, Pr(M^{(k)} \,|\, \mathbf{y}) \tag{6}$$

This work builds on Raftery's model pruning strategy with an analogous process to prune *methods* that are redundant in information content, or that poorly explain observations in $\mathbf{y}$. The utility of a given method, j, is estimated by evaluating how the method's information content is used across the different $2^P - 1$ models. Methods that are predominantly used by statistically significant models (i.e., models with high posterior probabilities) should be preserved. Contrariwise, methods that are found primarily in models with little statistical significance should be iteratively removed from the ensemble design process.

Formally, BMA estimates method j's utility for explaining a set of observations, $\mathbf{y}$, by assessing the probability that the method's coefficient term, β_j^{BMA} will receive a non-zero value. The estimate of this probability is based on the conditioned, cumulative sum of all model posteriors

$$\Pr(\beta_j^{\mathrm{BMA}} \neq 0) = \sum_{M^{(k)} \in \mathbf{A}} \Pr(M^{(k)} | \mathbf{y})\, \mathrm{I(j)} \tag{7}$$

where,

$$\mathrm{I(j)} := \begin{cases} 1, & \text{if model } M^{(k)} \text{ specifies j as a regressor} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The application of Equation 7 for a given set of P methods can be used to sort and prioritize the methods based on their utility; i.e., the probability that the coefficient term weighting a method's estimate will not be 0. The combination of Equations 6 and Equation 7 therefor provide a statistical framework to support an iterative, statistical process for designing an ensemble.

2.2.2 Statistical Design of Ensembles

In statistically designing an ensemble, the objective is to identify the best combination of models and methods that can be used to construct an aggregate estimate for hydration free energy. This design process is shown in Algorithm 1.

The process assumes an initial set of n estimates made by p methods, $\mathbf{x}^{n \times p}$ for observations, $\mathbf{y}^{n \times 1}$. The second and third lines initialize the root mean squared error (RMSE) and the solution for the best ensemble of models and methods. While the algorithm still has more than two methods to evaluate (line 5 - line 19), it will iterate over the following tasks.

First, the performance of the current ensemble is evaluated through *AssessMethods* (line 6). The *AssessMethods* function, shown in Algorithm 2, takes as input the same observation data, $\mathbf{y}$, as Algorithm 1, as well as a set of N estimates made by a subset of $j \leq p$ methods. Algorithm 2 performs 100 iterations of a 2-fold cross-validation and begins by partitioning the estimate matrix, $\mathbf{m}$, equally into training and validation data (line 6). Using the training data, the algorithm uses

Algorithm 1
Ensemble Design

Require: vector $\mathbf{y}^{n\times 1}$, matrix $\mathbf{x}^{n\times p}$

1: $r \leftarrow 1$
2: $rmse \leftarrow \infty$
3: **x_solution** $\leftarrow$ **x**
4: $\beta_{BMA}^{1\times j} \leftarrow \mathbf{0}^{1\times j}$
5: **while** (r < p) **do**
6: $\beta_{BMA}, s, \mathbf{v} \leftarrow AssessMethods$(**y, x**)
7: **if** (s < rmse) **then**
8: **x_solution** $\leftarrow$ **x**
9: $rmse \leftarrow s$
10: **end if**
11: $m \leftarrow 1$
12: **for** ($methods \in \mathbf{x}$) **do**
13: **if** ($\mathbf{v}[methods] < \mathbf{v}[m]$) **then**
14: $m \leftarrow methods$
15: **end if**
16: **end for**
17: **x** $\leftarrow$**x**.delete(m) {remove the least informative method}
18: $r \leftarrow (r+1)$
19: **end while**
20: **return** **x_solution**, β_{BMA}, rmse

Algorithm 2
AssessMethods

Require: vector $\mathbf{y}^{n\times 1}$, matrix $\mathbf{m}^{n\times j} : j \leq p$

1: $cv \leftarrow 100$ {perform 100 rounds of 2-fold cross-validation}
2: $\beta_{BMA}^{1\times j} \leftarrow \mathbf{0}^{1\times j}$
3: $error \leftarrow 0$
4: $\mathbf{v}^{1\times j} \leftarrow \mathbf{0}^{1\times j}$
5: **while** ($index < cv$) **do**
6: $\mathbf{m}_t^{q\times j}, \mathbf{m}_v^{r\times j} \leftarrow \mathbf{m}^{n\times j}$ {split **m** into **t**rain and **v**alidate}
7: $\beta_{BMA} \leftarrow \beta_{BMA}$+BMA($\mathbf{m}_t$) {find β: Equation 6}
8: $error \leftarrow error + RMSE(\mathbf{Y}, \mathbf{m}_v \times \beta_{BMA})$
9: $\mathbf{v} \leftarrow \mathbf{v}$+P$\neq$0($\mathbf{m}_t$) {assess method utility: Equation 7}
10: $index \leftarrow index + 1$
11: **end while**
12: **return** $\frac{\beta_{BMA}}{cv}, \frac{error}{cv}, \frac{\mathbf{v}}{cv}$

Equation 6 to estimate β_{BMA} and then calculates the RMSE of this coefficient vector based on the validation data (lines 7 and 8). Next, the $\Pr(\beta_j^{\text{BMA}} \neq 0)$ of each method are calculated and saved in the vector, **v** (line 9). After 100 iterations, the function returns an estimate for β_j^{BMA}, the mean RMSE of the current ensemble, as well as a list of all methods and their corresponding probability values for $\Pr(\beta_j^{\text{BMA}} \neq 0)$

Next, Algorithm 1 compares and conditionally updates the current "best" performing model, **x_solution**, with the model returned by *AssessMethods* (lines 8 and 9). The Algorithm then identifies the method with the lowest $\Pr(\beta_j^{\text{BMA}} \neq 0)$ and removes this method from the current model (line 17). With this method removed, the process repeats until there are just two methods left.

The output of this process is an ensemble of methods, $\mathbf{x}_{ij}$ that are statistically determined to be the best methods for estimating the observations in **y**, and a statistical model, β_{BMA} that specifies how to best combine these methods.

3 Hydration free energy data and solvation methods

The SAMPL4 challenge consists of 49 submissions representing a total of 19 different research groups [60]. Each of these methods provides hydration free energy estimates for 45 varied small molecule compounds. This challenge is blind in the sense that the hydration free energy values for these molecules were hidden from participants: e.g., energy values are not found in standard hydration free energy test sets, and their values are not readily available in primary literature [37].

In this work, we restrict our analysis to a subset of 17 of these submissions based on the fact that many groups made multiple submissions that were strongly correlated. In these cases, we chose only a single variant to ensure that multicollinearity did not inflate the significance of certain methods during model selection and averaging; such bias can create unstable estimates for β_j^{BMA}

that can reduce BMA's predictive accuracy [14]. These methods are summarized in Table I based on their solvation methodology:

- **Group 1:** Single-conformation implicit solvent methods [3,22,25] based on Poisson-Boltzmann and related methods [19, 23, 35];
- **Group 2:** Multi-conformational implicit solvent methods [1, 31, 38, 49, 86];
- **Group 3:** Methods based on alchemical molecular dynamics simulations in explicit solvent [16, 32, 58, 68] with small molecule force fields [39]; and,
- **Group 4:** Hybrid solvent methods based on data from explicit solvent simulations [48].

3.1 Training, estimating with, and pruning an ensemble with BMA

From this initial set of 17 methods, we iteratively assess and prune methods as described in Algorithm 1. The approach begins by randomly sampling (without replacement) 26 of the original 52 experimental hydration free energy measurements to form training and validation datasets (Algorithm 2, line 6). Collectively, these sampled values form the observation vector y_i and the estimates from each of the prediction methods in Table I for these measurements define the ensemble estimate matrix, x_{ij}. The observation vector and the ensemble estimate matrix form the linear system in Equation 1.

Next, we estimate the β_j^{BMA} parameter from Equation 2 by assessing all possible $(2^{17} - 1)$ statistical models $M^{(k)}$. Each model's information criteria, $B^{(k)}$, is used to identify a reduced set of most informative models per Equation 5. Based on this reduced modeling space, the coefficient terms β_j^{BMA} are estimated through a weighted average of each statistical model's ordinary least squares solution (Equation 6). We use β_j^{BMA} to estimate the remaining 26 hydration free energy measurements that were *not* used to train the BMA model. This task is accomplished by combining the estimates of all methods in Table I for the validation data with β_j^{BMA} to produce an aggregated estimate. A root mean squared error is obtained for the 26 estimates made on validation data.

This total process is repeated 100 times (Algorithm 2) so that all performance for any estimation method is reported as a mean RMSE: i.e., the mean of 100 RMSEs that each represent performance for a 2-fold cross-validation that uses a 26 member validation set. In addition to the mean RMSE, the information content provided by each method in the ensemble is also returned.

Algorithm 1 uses this information to perform two tasks. First, if the performance of this aggregated estimate is better than all previously examined aggregates, the statistical model combining these methods is saved as the new optimal model (Algorithm 1, lines 7-9). Next, Algorithm 1 examines the methods in the ensemble and prunes the method that provided the least amount of information to the aggregated estimate. The pruning process is repeated until just two methods are left (Algorithm 1, lines 12-18). The ensemble of methods whose aggregate forecast has the lowest mean RMSE is saved and returned as the final model for the ensemble. This final model, and the set of methods that correspond to this model, are the final products of the statistically driven ensemble design process. In the Section 4, this model and ensemble of methods is referred to as the BMA-based optimal ensemble.

4 Results and discussion

We exercise our ensemble design process to identify an optimal subset of methods from 17 initial methods that competed in the SAMPL4 challenge. We combine estimates from this optimal subset using a BMA-based model to form aggregated estimates for hydration free energies. We report the results of our study in three stages. Stage 1 contrasts the performance of the initial 17 SAMPL4 methods to the performance of the BMA-based optimal ensemble (Table I). We then present data on the iterative design process used to create this optimal ensemble (Figures 1, 2, and 3 as well as Table II). Stage 2 examines the *conditional* performance of this optimal ensemble according to the individual molecules used in the SAMPL4 challenge (Figures 4 and 5). In this stage we show how the BMA-based optimal ensemble provides more reliable estimates in comparison to individual methods, especially against the more challenging set of small molecules from the SAMPL4 dataset. Stage 3 completes the analysis of the optimal ensemble by contrasting its performance to the performance of alternate statistical techniques that can be used to form aggregated estimates (Table III and Figure 6).

As indicated in Section 3.1, the performance of all methods is based on a mean RMSE: i.e., the mean value of 100 RMSEs generated from the 100 iterations of a 2-fold cross-validation. We report the statistical significance of all performance through a Wilcoxon rank sum paired comparison test [93]. This non-parametric approach tests the hypothesis that the mean RMSE distributions of two approaches are equal: $H_0 : \mu_Y = \mu_X$. To control the familywise error rate of our tests, we applied a Bonferroni correction to determine corrected p-values with a threshold of $\alpha = 0.05$. Thus when comparing BMA to a given method, a Wilcoxon generated p-value greater than 5.0E $-$ 2 indicates we fail to reject H_0: the distributions are thus equal and we conclude that BMA and the method are equivalent in their performance. On the other hand, Wilcoxon generated p-values that are less than 5.0E $-$ 2 indicate we should reject H_0. In this latter case, we then compare the mean RMSE for BMA and the given method to assess performance.

4.1 Stage 1: Comparing estimates from BMA's optimal ensemble to SAMPL4 challenge methods

The performance of all methods used in this work is shown in Table I and Figure 1. The third column in Table I lists the specific methodology behind each estimation approach. The performance results in column four of this table are consistent with results reported by Mobley et al. [60]. The performance of the optimal ensemble is shown in the last row. The corresponding methods that are constituents in this optimal ensemble, alc-3 and imp-2, are highlighted in blue. The final column in Table I lists a direct comparison of each method's performance to the performance of the optimal ensemble: e.g., the ensemble's aggregated estimate reduces estimation errors by as much as 91% in comparison to imp-6 and by 29% in comparison to imp-2.

The box plots in Figure 1 visually illustrate the performance variability across the different methods. These plots also underscore the amount of uncertainty inherent in estimating hydration free energies: e.g., top performing methods vary from single conformation and multi-conformation implicit methods to alchemical explicit methods. In this context, method selection uncertainty plays a confounding factor in estimating hydration free energies. Finally, the red line indicates the mean RMSE of the best performing method: imp-2. This line is also used in Figure 2 to contrast

the iterative improvements obtained during the optimal ensemble's design process.

The ensemble's iterative design process based on Algorithms 1 and 2 is shown in Table II and Figure 2. The process starts at Stage 1 with the 17 initial methods shown in Table I. Each subsequent row in Table II represents an iteration through the pruning process. The second column lists the mean root mean squared error (RMSE) of each stage's ensemble based on its aggregated estimate. At the end of each stage, a method is selected to be pruned from the existing ensemble before proceeding to the next iteration; the specific method that was selected is shown in column 3. For example, at Stage 1 there are 17 methods in the ensemble and imp-6 has been selected to be removed. During the next step, Stage 2, imp-6 has been removed from the ensemble so that there are only 16 methods used to create an aggregated estimate. At the end of this stage, alc-4 has been selected to be pruned for Stage 3. In the final stage, the only remaining methods in the ensemble are imp-2 and alc-3.

This iterative design process is also graphically illustrated in Figure 2. The general trend for the design process in this figure indicates that the selective pruning increases the performance of each successive ensemble. The redline in this figure indicates the performance of the best performing method, imp-2. The benefits of the aggregated estimates becomes apparent after Stage 6 where the ensembles outperform imp-2.

The statistical significance of the iterative design process is listed in columns four and five in Table II; bold p-values in these columns indicate the performance between two distributions are equivalent. Column four lists the Wilcoxon generated p-values based on comparisons of mean RMSE distributions obtained from sequential ensembles. For example, in the second row the p-value for the comparison of Stage 2 vs. Stage 1 indicate that these distributions are equivalent. Contrariwise, in row four, the p-value for the comparison between Stage 4 and Stage 3 indicate that the distributions are not equivalent, and so the mean RMSE in column two indicates that the ensemble of Stage 4 outperforms the ensemble built in Stage 3.

As a second analysis of significance, column five lists the Wilcoxon generated p-values that represent comparisons between each ensemble and the best performing method, imp-2. From these values, Stages 4-6 are seen to be equivalent to imp-2. The mean RMSE distribution of all subsequent stages, however, are not equivalent; based on mean RMSE listed in column two we conclude that these successive ensembles (increasingly) outperform imp-2. Based on the p-values in these columns, and the mean RMSE results, the optimal ensemble is the one created in Stage 16. The final column in Table II lists the performance improvement that the Stage 16 ensemble provides in comparison to each other ensemble that was generated in the design process.

Figure 3 depicts a heat-map that shows the statistical information that drives the pruning process for each stage. In this image, the y-axis represents the different methods. The x-axis (starting from left) indicates the successive stages in the design process. The color scale represents the mean probability, 0 - 100%, that a given method's coefficient term, β_j will not be zero. Thus at Stage 1, all methods are used in the ensemble and their $\Pr(\beta_j^{\text{BMA}} \neq 0)$ range from 40% (imp-6) to 80% (imp-2). As imp-6 has the lowest mean probability of not being zero, imp-6 is pruned after Stage 1 such that in Stage 2 the color map colors this method white to indicate it has been eliminated. The methods listed on the legend of the y-axis are thus ordered by the sequence that they were eliminated in the design process: e.g., imp-6 first, alc-4 second, imp-8 third, etc.

In general the trend across the stages of pruning illustrates that the $\Pr(\beta_j^{\text{BMA}} \neq 0)$ for methods becomes increasingly polarized and move towards lower or higher probability of having a coefficient term of 0. For example, the mean probabilities for the coefficients of imp-7 and exp-4 not

being zero become increasingly lower until they are pruned. Contrariwise, alc-3, and imp-2 remain above 70% and increase throughout the design process indicating a high degree of confidence in the statistical importance of these methods. The final ensemble in Stage 16 consists of just two method constituents: alc-3 and imp-2. As shown in Tables I and II, the BMA-based estimate that aggregates estimates from this ensemble provides the best performance for estimating hydration free energies.

4.2 Stage 2: Performance Analysis Based on Compounds

Figures 4 and 5 depict performance of different methods according to specific SAMPL4 challenge small molecule compounds. In addition to the optimal ensemble, we also show the performance of the first, second and third best-performing methods from the SAMPL4 challenge: imp-2, imp-8, and alc-3. Methods imp-2 and alc-3 are the methods used in the optimal BMA ensemble; exp-3 is the final method eliminated from the ensemble (Stage 15 in Table II).

The analysis of Mobley et al. identify that certain compounds were especially difficult to estimate in the SAMPL4 challenge: SAMPL4_022 (mefenamic acid), SAMPL4_023 (diphenhydramine), SAMPL4_027 (1,3-bis-(nitroxy)propane), SAMPL4_009 (2,6-dichlorosyringaldehyde), and SAMPL4_001 (mannitol) [60]. In looking at the performance of the different methods in Figures 4 and 5 that estimate these compounds, the optimal ensemble outperforms all methods in estimating SAMPL4_022 and SAMPL4_001, and provides the second best estimates for SAMPL4_009, SAMPL4_023 and SAMPL4_027. Note that aside from the optimal ensemble, there is no clear best method for estimating these compounds. For example, while imp-8 is best at estimating SAMPL4_009, it is does not do well at estimating either SAMPL4_023 or SAMPL4_027. Similarly while imp-2 performs well at estimating SAMPL4_023, it does not do as well at estimating SAMPL4_009 or SAMPL4_027. While there are certain compounds that challenge the ensemble (e.g., SAMPL4_017), the general trend in performance suggests that the optimal ensemble provides more consistent and accurate estimates than any specific method.

4.3 Stage 3: Performance Analysis of Alternate Ensemble Techniques

There are other ensemble-based approaches besides BMA that can combine an ensemble of methods to make an aggregate prediction. In our cross-validation study, we evaluated four common approaches for aggregating an ensemble and evaluated their predictive benefits in comparison to BMA. These methods are listed in Table III and include: Random Forest [9], Ridge Regression [33], Lasso [88], and stepwise regression via forward selection. These techniques were chosen as they have all been used successfully for a variety of inference tasks and are readily available for use [10, 74]. As all of these approaches constructe an estimate for β, training and predicting with these approaches was performed identically to how we trained and predicted with BMA based on the cross-validation detailed in Section 3.1. We also follow the same procedure for comparing BMA?s predictive capability to these alternate ensemble-based techniques.

Figure 6 and Table III provide an overview of the cross-validation errors for the various ensemble-based approaches and the BMA-based optimal ensemble. The statistical significance of BMA?s performance in Figure 6 is based on p-values shown in Table III. Based on an $\alpha = 0.05$, Table VI indicates that we reject the null hypothesis for all paired comparison tests. BMA?s mean RMSE

distribution is therefore not equivalent to the mean RMSE distribution of any other ensemble-based technique

As the distributions are not equal, we compared mean RMSE distributions of BMA to the other ensemble-based approaches in Figure 6 and Table III. From these mean RMSE, it is clear that the BMA-based approach outperforms all other ensemble-based prediction approaches: BMA-based estimates reduced error by approximately 60% in comparison to Random Forest and Ridge regression methods. In comparison to Lasso, BMA reduces estimation error by approximately 27%. Finally, in comparison to stepwise regression via forward selection, BMA reduces error by approximately 25%

5 Conclusions

This study demonstrates a proof-of-principle application of how to statistically design and aggregate an ensemble of methods for estimating hydration free energies in small molecules. While the performance of BMA is expected to generalize to a much broader set of small molecule estimation problems, the specific BMA model trained in this study is likely to be dependent on the small molecules used in the SAMPL4 challenge. While our BMA approach is purely statistical in nature, the BMA method described here could be trained to modify the aggregation process based on structural and environmental features (e.g., only look at ensembles of empirical methods for certain structural features and consider all methods for other structures). In future work we will look at penalizing computationally expensive methods that provide minimal accuracy benefits.

Acknowledgments

This research was funded by the National Biomedical Computational Resource (NIH award P41 RR0860516) and NIH grant R01 GM069702 to NAB.

References

[1] K. A., E. F., and D. M. Prediction of the free energy of hydration of a challenging set of pesticide-like compounds. *J Phys Chem B.*, 2009.

[2] K. A. and D. M. Blind prediction test of free energies of hydration with cosmo-rs. *J. Compute. Aided Mol. Des.*, 24(4):357–360, 2010.

[3] N. A., W. S., and G. J. Sampl2 and continuum modeling. *J Comput Aided Mol Des.*, 24(4):293–306, 2010; 24(4):293–306.

[4] G. Apostolakis. The concept of probability in safety assessments of technological systems. *Science*, 250(4986):1359–1364, 1990.

[5] H. Ashbaugh, S. Garde, G. Hummer, E. Kaler, and M. Paulaitis. Conformational equilibria of alkanes in aqueous solution: Relationship to water structure near hydrophobic solutes. *Biophysical Journal*, 77(2):645–654, 1999.

[6] H. Ashbaugh, E. Kaler, and M. Paulaitis. Conformational equilibria of polar and charged flexible polymer chains in water. *Polymer*, 43(5):559–565, 2002.

[7] H. S. Ashbaugh, E. W. Kaler, and M. E. Paulaitis. A universal surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.*, 39(121):9243–9244, 1999.

[8] J. M. Bates and C. W. J. Granger. The combination of forecasts. *Operational Research Quarterly*, 20(4):451–468, 1969.

[9] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[10] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[11] Y. C., S. H., C. J., N.-C. Z., and W. S. Importance of ligand reorganization free energy in protein-ligand binding-affinity prediction. *J Am Chem Soc.*, 131(38):13709–21, 2009.

[12] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–51, 2007.

[13] I. Chorny, K. A. Dill, and M. P. Jacobson. A universal surface area correlation for molecular hydrophobic phenomena. *J. Phys. Chem. B.*, 50(109):24056—24060, 2005.

[14] M. Clyde. Bayesian Model Averaging and Model Search Strategies. In *Bayesian Statistics*, volume 6, pages 157–185, 1999.

[15] Q. Cui and V. Smith. Solvation structure, thermodynamics, and molecular conformational equilibria for n-butane in water analyzed by reference interaction site model theory using an all-atom solute model. *J. Phys. Chem. B,*, 106(25):6554–6565, 2002.

[16] M. D, B. C, C. M, S. M, and D. KA. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J Chem Theory Comput.*, 5(2):350–358, 2009.

[17] M. D. and D. K. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure*, 17(4):489–498, 2009.

[18] I. Davidson and W. Fan. *When Efficient Model Averaging Out-Performs Boosting and Bagging*, volume 4213 of *Lecture Notes in Computer Science*, pages 478–486. Springer Berlin Heidelberg, 2006.

[19] M. Davis and A. McCammon. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, 90(3):509–521, 1990.

[20] J. Devooght. Model uncertainty and model inaccuracy. *Reliability Engineering and System Safety*, 59(2):171–185, 1998.

[21] D. Eisenberg and A. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(1):199–203, 1986.

[22] B. Ellingson, M. Gaballe, S. Wlodek, C. Baylyl, A. Skillman, and A. Nicholls. Efficient calculation of sampl4 hydration free energies using omega, szybki, quacpac, and zap tk. *J. Compute. Aided Mol. Des.*, 28(3):289–298, 2014.

[23] M. Fixman. The Poisson–Boltzmann equation and its application to polyelectrolytes. *Journal of Chemical Physics*, 70(11):4995–146, 1979.

[24] J. Fu, Y. Liu, and J. Wu. Fast prediction of hydration free energies for sampl4 blind test from a classical density functional theory. *J Comput Aided Mol Des*, 28(3):299–304, 2014.

[25] H. G., G. D., L. G., C. C., R. I., S. J., L. J., Z. T., T. J., W. P., and L. B. Amsol.

[26] K. G, P. C, M. Y, and B. R. Predicting hydration free energies with a hybrid qm/mm approach: an evaluation of implicit and explicit solvation models in sampl4. *J Comput Aided Mol Des*, 28(3):245–57, 2014.

[27] E. Gallicchio, L. Zhang, and R. Levy. The sgb/np hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Phys.*, 23(5):517–529, 2002.

[28] M. Geballe, A. Skillman, A. Nicholls, J. Guthrie, and P. Taylor. The sampl2 blind prediction challenge: introduction and overview. *J Comput Aided Mol Des*, 24(4):259–279, 2010.

[29] A. Genell, S. Nemes, G. Steineck, and P. W. Dickman. Model selection in medical research: A simulation study comparing Bayesian model averaging and stepwise regression. *BMC Medical Research Methodology*, 10:108, 2010.

[30] L. Gosink, E. Hogan, T. Pulsipher, and N. Baker. Bayesian model aggregation for ensemble-based estimates of protein pka values. *Proteins*, 82(3):354–363, 2014.

[31] H. H., S. T., and P. E. Exuastive docking and solvated interaction energy scoring: Lessons learned from the sampl4 challenge. *J Comput Aided Mol Des.*, 2014.

[32] M. H., N. Sapra, A. Fenley, and M. Gilson. The sampl4 hydration challenge: Evaluation of partial charge sets with explicit-water molecular dynamics simulations. *J Comput Aided Mol Des*, 28(3):277–287, 2014.

[33] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[34] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

[35] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–1149, 1995.

[36] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989.

[37] G. J. Sampl4, a blind challenge for computational solvation free energies: the compounds considered. *J Comput Aided Mol Des.*, 28(3):151–168, 2014.

[38] R. J. and K. A. Prediction of free energies of hydration with cosmo-rs on the sampl4 data set. *J Comput Aided Mol Des*, 2014.

[39] W. J., W. R., C. J., K. P., and C. D. Development and testing of a general amber force field. *J Comput Chem.*, 25(9):1157–1174, 2004.

[40] J. Jambeck, F. Mocci, A. Lyubartsev, and A. Laaksonen. Partial atomic charges and their impact on the free energy of solvation. *J. Compute. Chem.*, 34(3):187–197, 2013.

[41] H. Jeffreys. *Theory of Probability*. Oxford: Oxford University Press, 1961.

[42] W. K. and S. M. Flooding enzymes: Quantifying the contributions of interstitial water and cavity shape to ligand binding using extended linear response free energy calculations. *J Chem Inf Model.*, 53(9):2349–2359, 2013.

[43] S. Kamerlin, M. Haranczyk, and A. Warshel. Are mixed explicit/implicit solvation models reliable for studying phosphate hydrolysis? a comparative study of continuum, explicit and mixed solvation models. *Chemphysche*, 10(7):1125–1134, 2009.

[44] Y. Kang, D. Gibson, G. Nemethy, and H. Scheraga. Free energies of hydration of solute molecules. 4. revised treatment of the hydration shell mode. *J. Phys. Chem*, 92(16):4739–4742, 1988.

[45] Y. Kang, G. Nemethy, and H. Scheraga. Free energies of hydration of solute molecules. 1. improvement of the hydration shell model by exact computations of overlapping volumes. *J. Phys. Chem*, 91(15):4105–4109, 1987.

[46] Y. Kang, G. Nemethy, and H. Scheraga. Free energies of hydration of solute molecules. 2. application of the hydration shell model to nonionic organic molecules. *J. Phys. Chem*, 91(15):4109–4117, 1987.

[47] Y. Kang, G. Nemethy, and H. Scheraga. Free energies of hydration of solute molecules. 3. application of the hydration shell model to charged organic molecules. *J. Phys. Chem*, 91(15):4118–4120, 1987.

[48] L. L., D. K., and F. C. Testing the semi-explicit assembly model of aqueous solvation in the sampl4 challenge. *J Comput Aided Mol Des.*, 28(3):259–264, 2014.

[49] S. L. Predicting hydration free energies with chemical accuracy: The sampl4 challenge. *J Comput Aided Mol Des.*, 2013.

[50] R. Levy, L. Zhang, E. Gallicchio, and A. Felts. On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute - solvent interaction energy. *J. AM. CHEM. SOC*, 125:9523–9530, 2003.

[51] J. Li, T. Zhu, G. Hawkins, P. Wingnet, D. Liotard, C. Cramer, and D. Truhlar. Extension of the platform of applicability of sm5.42r universal solvation model. *Theoretical Chemistry Accounts*, 103(1):9–63, 1999.

[52] G. M and G. J. The sampl3 blind prediction challenge: transfer energy overview. *J Comput Aided Mol Des.*, 26(5):489–496, 2012.

[53] D. Madigan and A. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *J. Am. Stat. Assoc.*, 89(1):1335–1346, 1994.

[54] C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

[55] G. Martinez-Munoz, D. Hernandez-Lobato, and A. Suarez. An analysis of ensemble pruning techniques based on ordered aggregation. *Trans. Pattern Analysis and Machine Intelligence*, 31(2):245–259, 2009.

[56] B. Mennucci and R. Camm, editors. *Continuum solvation models in chemical physics: from theory to applications*. John Wiley & Sons, 2007.

[57] D. Mobley, C. Bayly, M. Cooper, and K. Dill. Predictions of hydration free energies from all-atom molecular dynamics simulations. *J Phys Chem B.*, 113(10):4533–4537, 2009.

[58] D. Mobley, E. Dunmont, J. Chodera, and K. Dill. Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B.*, 11(9):2242–2254, 2007.

[59] D. L. Mobley, A. E. Barber, C. J. Fennell, and K. A. Dill. Charge asymmetries in hydration of polar solutes. *J. Phys. Chem. B.*, 8(112):2405—2414, 2008.

[60] D. L. Mobley, K. L. Wymer, N. M. Lim, and J. P. Guthrie. Blind prediction of solvation free energies from the sampl4 challenge. *J. Compute. Aided Mol. Des.*, 3(28):135—150, 2014.

[61] E. Morales-Casique, S. P. Neuman, and V. V. Vesselinov. Maximum likelihood Bayesian averaging of airflow models in unsaturated fractured tuff using Occam and variance windows. *Stochastic Environmental Research and Risk Assessment*, 24(6):863–880, 2010.

[62] S. Neuman and P. Wierenga. A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, 2003.

[63] A. Nicholls, D. Mobley, J. Guthrie, J. Chodera, C. Bayly, M. Cooper, and V. Pande. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *J. Med. Chem.*, 51(4):769–779, 2008.

[64] B. O., F. A., and I. B. Prediction of hydration free energies for the sampl4 diverse set of compounds using molecular dynamics simulations with the opls-aa force field. *J. Compute. Aided Mol. Des.*, 28(3):265–276, 2014.

[65] B. O. and I. B. Prediction of hydration free energies for aliphatic and aromatic chloro derivatives using molecular dynamics simulations with the opls-aa force field. *J Comput Aided Mol Des*, 26(5):635–645, 2012.

[66] L. Onorante and A. Raftery. Dynamic model averaging in large model spaces using dynamic occam's window. Technical Report 628, University of Washington, 2014.

[67] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[68] K. P. and M. D. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J Comput Aided Mol Des.*, 24(4):307–316, 2010.

[69] H. Park. Extended solvent-contact model approach to sampl4 blind prediction challenge for hydration free energies. *J. Compute. Aided Mol. Des.*, 28(3):175–186, 2014.

[70] Y. Q. and S. K. Atomic charge parameters for the finite difference poisson-boltzmann method using electronegativity neutralization. *J. Chem. Theory Comput.*, 4(2):1152–1167, 2006.

[71] C. Quan, Y. Yu, and Z. Zhou. Pareto ensemble pruning. In *Conference on Artificial Intellegince*, volume 29, pages 2935– 2941, 2015.

[72] C. R., S. T., and W. D. Sampl4 & dock3.7: lessons for automated docking procedures. *J. Compute. Aided Mol. Des.*, 28(3):201–209, 2014.

[73] L. R. and G. E. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu Rev Phys Chem*, 49(11):531–567, 1998.

[74] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[75] A. Raftery. Bayesian Model Selection in Social Research. *Sociological Methodology*, 25:111–163, 1995.

[76] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

[77] A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1998.

[78] S. Rakhmanov and V. Makeev. Atomic hydration potentials using a monte carlo reference state (mcrs) for protein solvation modeling. *BMC Struct Biol.*, pages 7–19, 2007.

[79] P. T. Reiss, L. Huang, J. E. Cavanaugh, and A. K. Roy. Resampling-based information criteria for best-subset regression. *Annals of the Institute of Statistical Mathematics*, 64(6):1161–1186, 2012.

[80] G. Robinson and C. Cho. Role of hydration water in protein unfolding. *Biophys J*, 77(6):3311–3318, 1999.

[81] R. Rojas, O. Batelaan, L. Feyen, and A. Dassargues. Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal – North Chile. *Hydrology and Earth System Sciences*, 14(2):171–192, 2010.

[82] L. Rokach. Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1-2):1–39, 2010.

[83] G. S., C. M., C. M., and E. J. Extensive all-atom monte carlo sampling and qm/mm corrections in the sampl4 hydration free energy challenge. *J. Compute. Aided Mol. Des.*, 28(3):187–200, 2014.

[84] G. Seni and J. F. Elder. Ensemble methods in data mining: Improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–126, 2010.

[85] D. Shivakumar, J. Williams, W. Damm, J. Shelley, and W. Sherman. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the opls force field. *J. Chem. Theory Comput.*, 6(5):1509–1519, 2010.

[86] S. T. and P. E. Predicting hydration free energies of polychlorinated aromatic compounds from the sampl-3 data set with fish and lie models. *J Comput Aided Mol Des.*, 26(5):661–667, 2011.

[87] C. Tan, L. Yang, and R. Lou. How well does poisson-boltzmann implicit solvent agree with explicit solvent? a quantitative analysis. *J. Phys. Chem.*, 110(37):18680–18687, 2006.

[88] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stats. Soc.*, 58(12):267–288, 1994.

[89] M. Vlachopoulou, L. Gosink, T. Pulsipher, T. Ferryman, N. Zhou, and J. Tong. An ensemble approach for forecasting net interchange schedule. In *IEEE PES General Meeting*, page n. pag., 2013.

[90] C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal. Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):433–448, 1997.

[91] J. W., U. J., and T.-R. J. Free energies of hydration from a generalized born model and an all-atom force field. *J. Phys. Chem. B*, 108(41):16264–16270, 2004.

[92] D. Wang, W. Zhang, and A. Bakhai. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, 23(22):3451–3467, 2004.

[93] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[94] S. Witham, K. Talley, L. Wang, Z. Zhang, S. Sarkar, D. Gao, W. Yang, and E. Alexov. Developing hybrid approaches to predict pK_a values of ionizable groups. *Proteins*, 79(12):3389–3399, 2011.

[95] M. Ye, S. P. Neuman, and P. D. Meyer. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, 40(5):n. pag., 2004.

[96] G. P. Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

Table I: This table lists the solvation methods used in our ensemble design process. Method ID indicates the identification number of the method that is referenced throughout this paper. Sampling strategies include quantum mechanical (QM), molecular dynamics (MD), and molecular mechanics with Poisson-Boltzmann surface area solvation (MM-PBSA). The listed performance is based on 100 iterations of a 2-fold cross-validation study. This performance is also shown graphically in Figure 1. The last column is a comparison of each method to the optimal BMA ensemble, BMA (Stage 16). This column indicates that the ensemble design approach presented in this work is able to reduce estimation errors by 29% to 91% in comparison to the individual methods. The final BMA ensemble is indicated by the two blue highlighted methods: imp-2 and alc-3. Wilcoxon based p-values for BMA's mean RMSE distribution vs. the best method's mean RMSE distribution (imp-2) are show in Table II

Reference	Method ID	Methodology	Ensemble Mean RMSE and Standard Deviation	Performance Improvement Provided by BMA (Stage 16)
Coleman et al. [72]	imp-6	multi-conformation implicit	9.40 ± 2.03	91%
Parsod et al.	alc-5	alchemical MD	4.50 ± 0.71	82%
Sharp et al. [70]	imp-4	single conformation implicit	3.80 ± 0.45	78%
Jiafu et al.	exp-3	QM/MD	2.80 ± 0.45	71%
Purisima et al. [31]	imp-5	single conformation implicit	2.55 ± 0.49	68%
Mark et al. 2011	alc-2	alchemical MD	2.40 ± 0.36	66%
Genheden et al. [83]	exp-1	MM-PB/SA	2.00 ± 0.22	59%
Weyang et al.	exp-4	MM-PB/SA	1.84 ± 0.31	55%
Biorga et al. [64, 65]	alc-4	alchemical MD	1.65 ± 0.13	50%
Elingson et al. [22]	imp-7	single conformation implicit	1.52 ± 0.39	46%
Fennell et al. [48]	exp-2	hybrid	1.52 ± 0.13	46%
Jambeck et al. [40]	alc-1	alchemical MD	1.52 ± 0.20	46%
Park [69]	imp-3	single conformation implicit	1.44 ± 0.20	43%
Klamt et al. [2]	imp-1	single conformation implicit	1.36 ± 0.30	40%
Gilson et al. [32]	alc-3	alchemical MD	1.24 ± 0.17	34%
Geballe et al. [22]	imp-8	single conformation implicit	1.17 ± 0.17	30%
Sandberg et al. [49]	imp-2	multi-conformation implicit	1.15 ± 0.23	29%
BMA (Stage 16)	**NA**	**ensemble**	$\mathbf{0.82 \pm 0.15}$	**0**

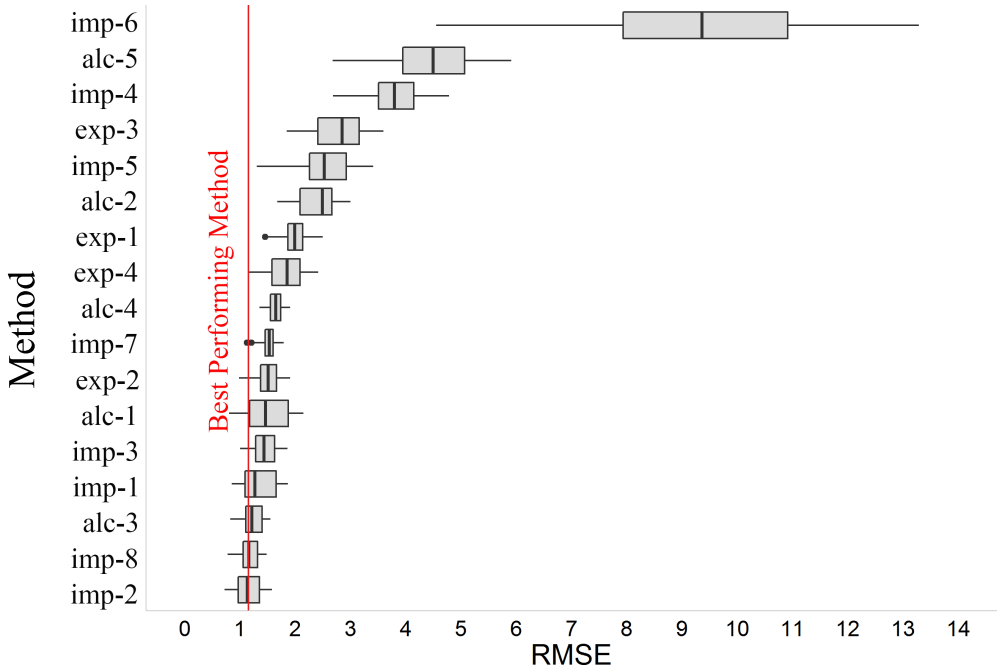

Figure 1: This figure depicts the mean root mean squared error, min, max, first and third quartiles for the 17 initial methods used in our ensemble design process. Method performance is based on 100 iterations of the 2-fold cross-validation experiment detailed in Algorithm 2. The red line is used to indicate the mean performance of the best method, imp-2; this line is referenced again in Figure 2 to show how the ensemble tuning process compares to the performance of this best solvation method.

Table II: This table lists the performance of the aggregated estimates obtained from the different ensembles created during the design process in Section 2.2. The second column lists the mean root mean squared error (RMSE) for each ensemble's aggregated estimate based on 100 iterations of a 2-fold cross-validation; this performance is also shown in Figure 2. The third column lists the method that was selected to be pruned from the ensemble at the *next* stage of the design process. The Wilcoxon generated p-values in column four are based on comparisons of mean RMSE distributions obtained from sequential ensembles and their aggregated estimates. Based on an $\alpha = 0.05$, p-values that are greater than 0.05 are bolded and indicate distributions that are equivalent (e.g., Stage 3 and 2). Similarly, column five lists the Wilcoxon generated p-values reflecting comparisons between each stage and the best performing method, imp-2. As with column 4, bolded p-values indicate that the performance between imp-2 and a given stage are equivalent (e.g., Stages 4-6). Based on mean RMSE and p-values from this table, the optimal ensemble is the one created in Stage 16; the final column in this table lists the performance improvement this ensemble provides in comparison to the ensembles generated in previous stages.

	Ensemble mean RMSE ± Standard Deviation	Method Selected To Be Pruned	p-value (sequential stages)	p-value (stage vs. imp-2)	Performance Benefit of BMA (Stage 16)
Stage 1	1.40 ± 0.60	imp-6	NA	0.000	41%
Stage 2	1.32 ± 0.53	alc-4	**0.411** (vs. Stage 1)	0.002	38%
Stage 3	1.29 ± 0.46	imp-8	**1.000** (vs. Stage 2)	0.004	36%
Stage 4	1.19 ± 0.31	imp-5	0.012 (vs. Stage 3)	**0.356**	31%
Stage 5	1.14 ± 0.29	imp-3	0.000 (vs. Stage 4)	**0.438**	28%
Stage 6	1.13 ± 0.26	imp-7	**0.883** (vs. Stage 5)	**0.306**	27%
Stage 7	1.09 ± 0.25	alc-5	0.015 (vs. Stage 6)	0.027	25%
Stage 8	1.06 ± 0.19	alc-2	**0.055** (vs. Stage 7)	0.003	23%
Stage 9	1.03 ± 0.20	imp-1	**1.000** (vs. Stage 8)	0.001	20%
Stage 10	0.99 ± 0.17	imp-4	0.000 (vs. Stage 9)	0.001	17%
Stage 11	0.96 ± 0.18	alc-1	0.000 (vs. Stage 10)	0.000	15%
Stage 12	0.93 ± 0.18	exp-4	**1.000** (vs. Stage 11)	0.000	12%
Stage 13	0.90 ± 0.17	exp-2	**1.000** (vs. Stage 12)	0.000	09%
Stage 14	0.90 ± 0.18	exp-1	**1.000** (vs. Stage 13)	0.000	09%
Stage 15	0.91 ± 0.18	exp-3	0.001 (vs. Stage 14)	0.000	10%
Stage 16	0.82 ± 0.15	NA	0.000 (vs. Stage 15)	0.000	0%

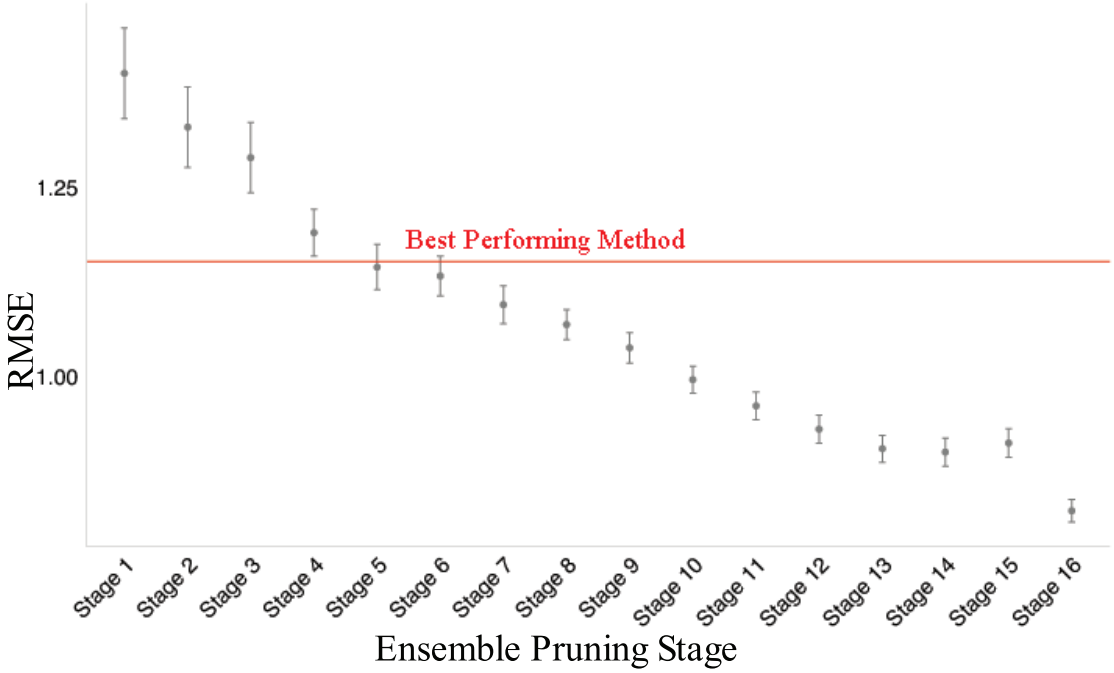

Figure 2: This plot illustrates the iterative pruning process discussed in Section 2.2. The y-axis depicts the mean root mean squared error (RMSE) of the different aggregated estimates based on ensembles formed during the different iterations of pruning. The x-axis indicates the stages of pruning. In general, the variance and overall mean RMSE reduces with each iteration. The performance of the different ensembles are compared to the best performing method through the red line at $y = 1.15$; all iterations past Stage 7 outperform the best method in the ensemble. Based on Wilcoxon generated p-values, the significance in the distributions of mean RMSE between the different ensembles are presented in Table II. Based on mean RMSE, the optimal ensemble is created at Stage 16.

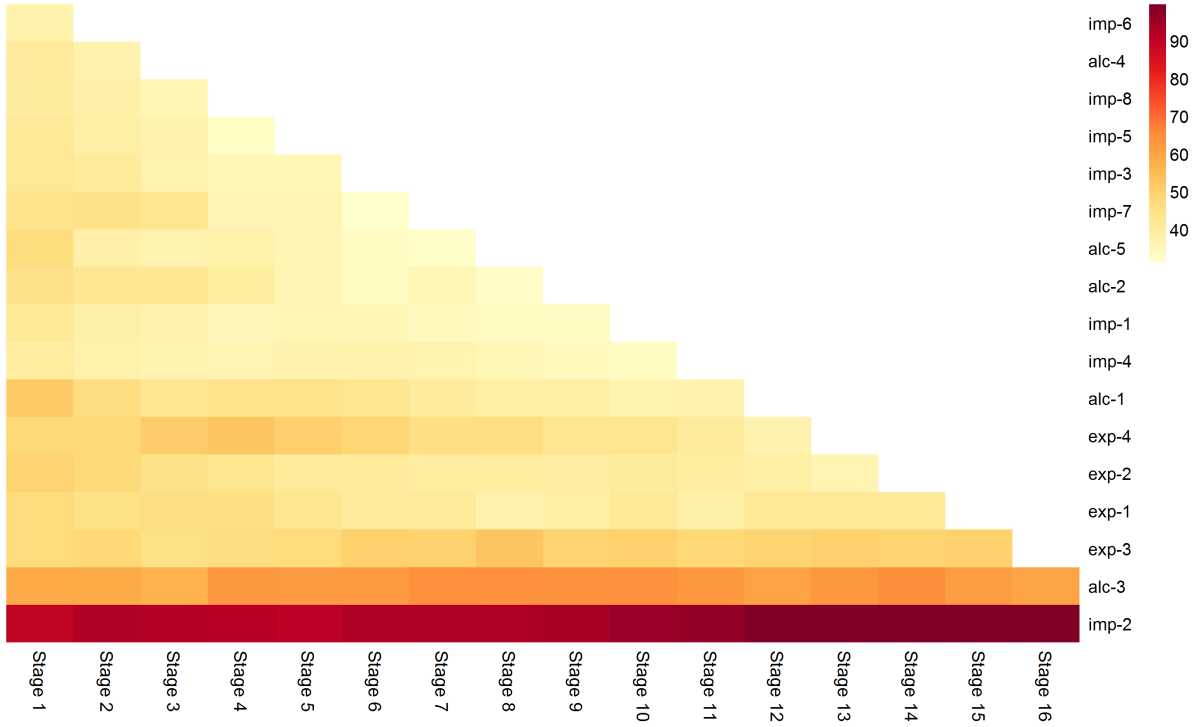

Figure 3: This image provides a graphical depiction of the ensemble design process performed on the initial ensemble of 17 methods. The color scale represents the probability, 0 - 100%, that a given method's coefficient term, β_j will not be zero. The stages in the design process, starting at the left and moving to the right, are shown on the x-axis. The methods in each ensemble are shown on the y-axis; note that the methods are listed (from top to bottom) in the order that they are pruned in the design process. Thus at Stage 1, all methods are used in the ensemble and their $\Pr(\beta_j^{\text{BMA}} \neq 0)$ range from 40% (e.g., imp-6) to 90% (imp-2). By Stage 3, imp-6 and alc-4 have been pruned from the ensemble and the probability values have adjusted accordingly as shown in the colored column above Stage 3. In general the trend across the different stages of pruning illustrates that methods become increasingly lighter, i.e., the ensemble design process becomes increasingly confident that these methods (e.g., imp-7 and exp-4) are not needed in the ensemble. Contrariwise, the $\Pr(\beta_j^{\text{BMA}} \neq 0)$ for a few methods (e.g., alc-3, and imp-2) remain above 70% and even increase throughout the design process indicating a high degree of confidence in the statistical significance of these methods.

Table III: This table lists the performance of different ensemble approaches in comparison to the optimal ensemble designed in this work through BMA. The performance of these ensemble approaches, given as the mean root mean squared error with standard deviation, are based on the 100 iterations of the 2-fold cross-validation experiment discussed in Section 2.2. Based on an $\alpha = 0.05$, the Wilcoxon based p-values indicate that BMA's improved performance is statistically significant to the other ensemble approaches for combining methods to make an aggregated estimate. The last column indicates the improvement in estimation that the optimal ensemble provides to these alternate techniques: estimation accuracy is improved from 25% to 61%.

Ensemble Model	Ensemble Mean RMSE and Standard Deviation	p-value	Performance Improvement Provided by BMA (Stage 16)
Random Forest	2.08 $\pm$1.00	0.00	61%
Ridge	2.06 $\pm$0.75	0.00	60%
Lasso	1.12 $\pm$0.34	0.00	27%
Forward Selection	1.09 $\pm$0.26	0.00	25%
BMA (Stage 16)	0.82 $\pm$0.17	NA	0%

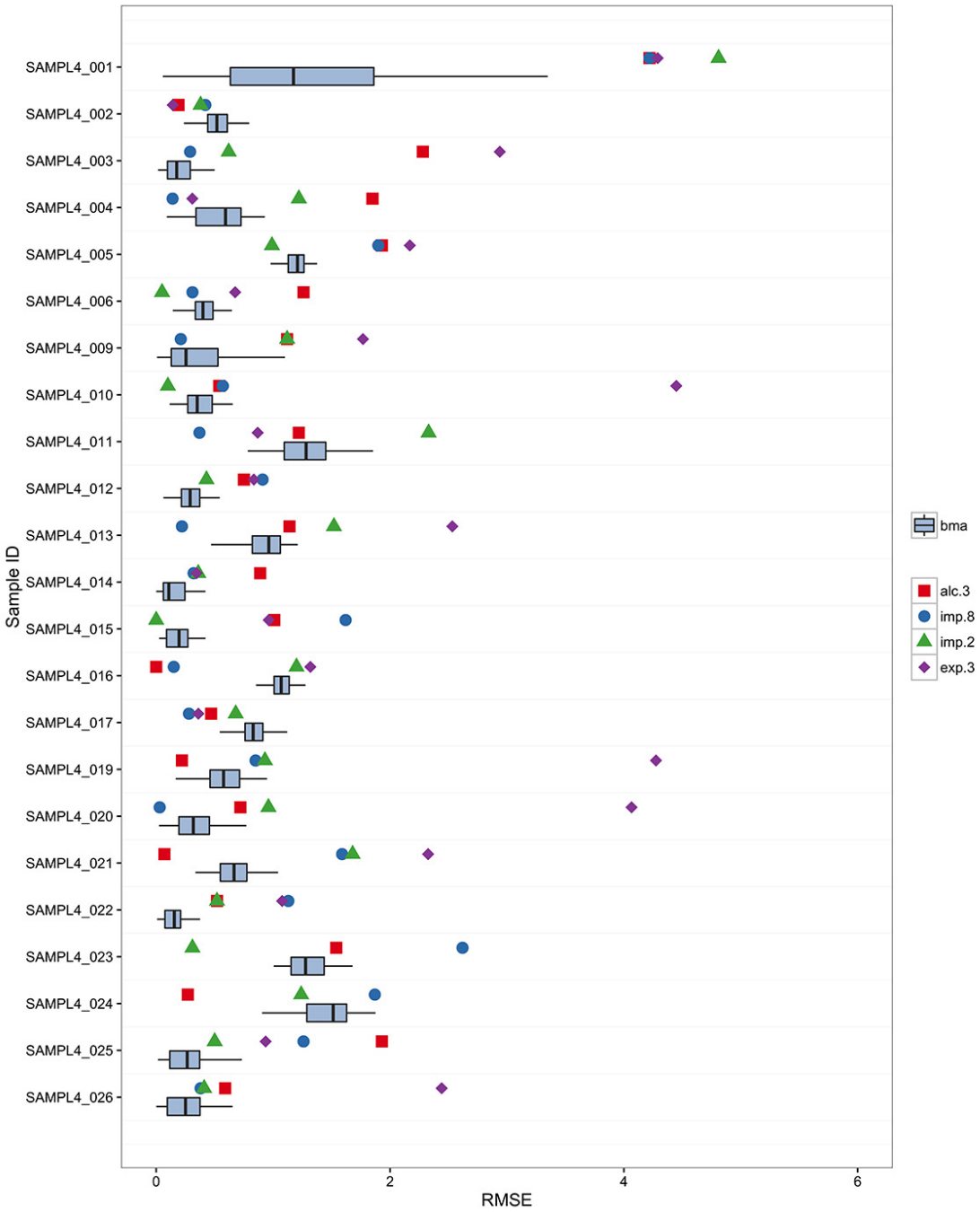

Figure 4: This figure is one of two figures (Figures 4 and 5) that depict the performance of several methods based on the individual compounds taken from the SAMPL4 challenge: the first, second and third performing methods (i.e., imp-2, imp-8, and alc-3) as well as exp-3. Note that imp-2 and alc-3 are the methods used in the optimal BMA ensemble (Stage 15) and exp-3 is the final method eliminated from the ensemble (Table II) in Stage 16. BMA's performance based on the optimal ensemble is shown based on its distribution of the mean root mean squared error for estimates made in our 2-fold cross-validation analysis. Of specific note is the performance of the different methods for SAMPL4_022 (mefenamic acid), SAMPL4_023 (diphenhydramine), SAMPL4_027 (1,3-bis-(nitroxy)propane), SAMPL4_009 (2,6-dichlorosyringaldehyde), and SAMPL4_001 (mannitol). These are the most challenging compounds for methods to estimate based on the analysis of Mobley et al. of the SAMPL4 data [60]. The benefits of the ensemble is clearly demonstrated here as the BMA ensemble outperforms all methods in estimating SAMPL4_022, SAMPL4_009, and SAMPL4_001. For SAMPL4_023 and SAMPL4_027 the ensemble provides the second best performance, and in this context provides more consistent performance than the other methods: e.g., alc-3 is better at estimating SAMPL4_027, but is third at estimating SAMPL4_23.

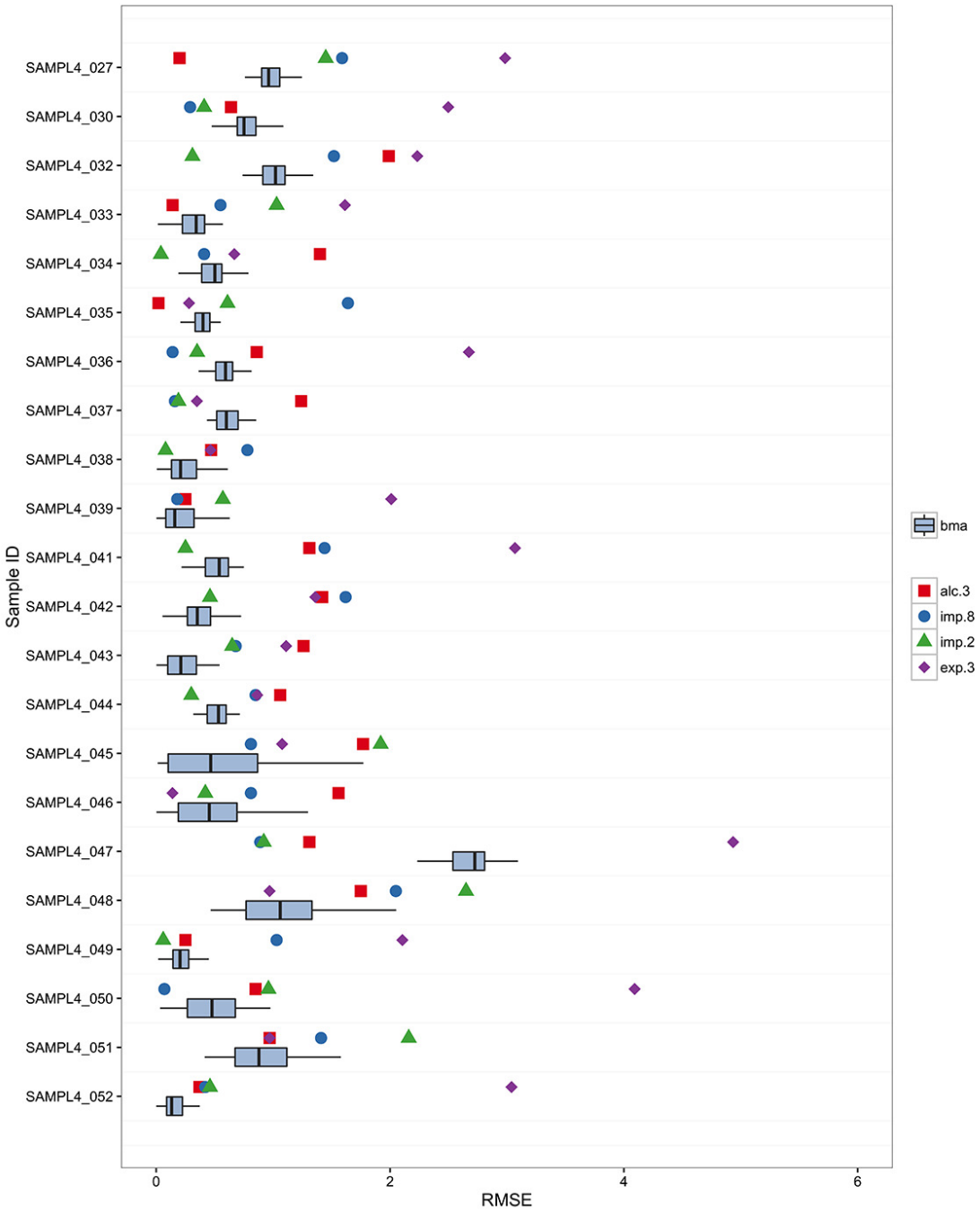

Figure 5: This figure is one of two figures (Figures 4 and 5) that depict the performance of several methods based on the individual compounds taken from the SAMPL4 challenge: the first, second and third performing methods (i.e., imp-2, imp-8, and alc-3) as well as exp-3. Note that imp-2 and alc-3 are the methods used in the optimal BMA ensemble (Stage 15) and exp-3 is the final method eliminated from the ensemble (Table II) in Stage 16. BMA's performance based on the optimal ensemble is shown based on its distribution of the mean root mean squared error for estimates made in our 2-fold cross-validation analysis. Of specific note is the performance of the different methods for SAMPL4_022 (mefenamic acid), SAMPL4_023 (diphenhydramine), SAMPL4_027 (1,3-bis-(nitroxy)propane), SAMPL4_009 (2,6-dichlorosyringaldehyde), and SAMPL4_001 (mannitol). These are the most challenging compounds for methods to estimate based on the analysis of Mobley et al. of the SAMPL4 data [60]. The benefits of the ensemble is clearly demonstrated here as the BMA ensemble outperforms all methods in estimating SAMPL4_022, SAMPL4_009, and SAMPL4_001. For SAMPL4_023 and SAMPL4_027 the ensemble provides the second best performance, and in this context provides more consistent performance than the other methods: e.g., alc-3 is better at estimating SAMPL4_027, but is third at estimating SAMPL4_23.

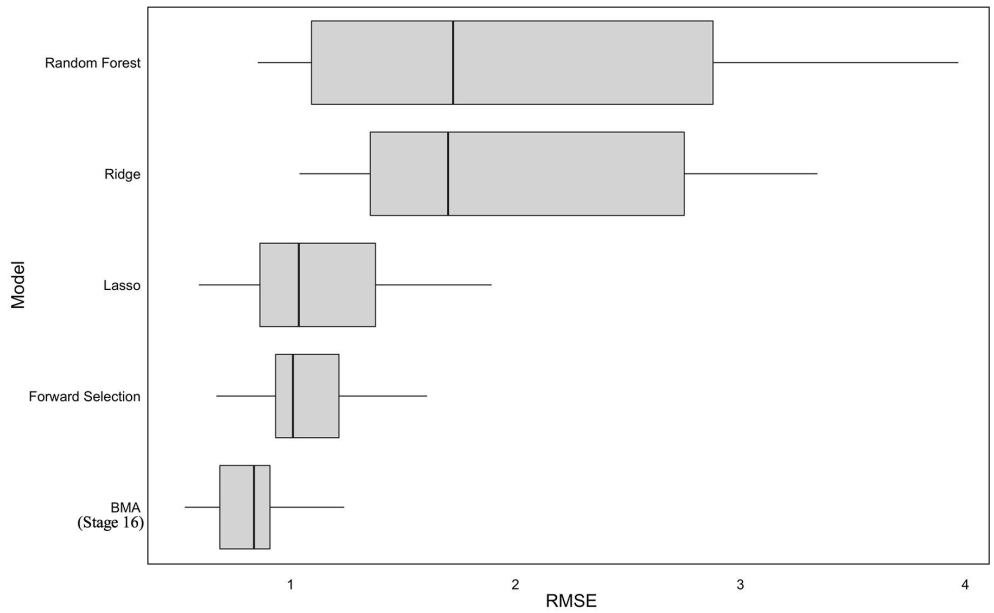

Figure 6: This figure displays the performance of different ensemble approaches in comparison to the optimal ensemble designed in this work, BMA (Stage 16). The mean root mean squared error, min, max, first and third quartiles of these ensemble approaches are shown based on the 100 iterations of the 2-fold cross-validation experiment discussed in Section 2.2. Based on an $\alpha = 0.05$, the Wilcoxon based p-values (Table III) indicate that BMA's improved performance is statistically significant to the other approaches that can combine an ensemble of methods to make an aggregated estimate.