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

Luke J. Gosink,[†] Christopher C. Overall,[†] Sarah M. Reehl,[†] Paul D. Whitney,[‡] David L. Mobley,[¶] and Nathan A. Baker[*,‡,§]

†*Computational and Statistical Analytics Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

‡*Advanced Computing, Mathematics, and Data Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*

¶*Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine, Irvine, CA 92697*

§*Division of Applied Mathematics, Brown University, Providence, RI 02912, USA*

E-mail: nathan.baker@pnnl.gov

Phone: +1-509-375-3997

Abstract

This paper applies the Bayesian Model Averaging (BMA) statistical ensemble technique to estimate small molecule solvation free energies. There is a wide range of methods available for predicting solvation free energies, ranging from empirical statistical models to *ab initio* quantum mechanical approaches. Each of these methods is based on a set of conceptual assumptions that can affect predictive accuracy and transferability. Using an iterative statistical process, we have selected and combined solvation energy estimates using an ensemble of 17 diverse methods from the fourth Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) blind prediction study to form a single, aggregated solvation energy estimate. The ensemble design process evaluates the statistical information in each individual 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 as much as 91% to 1.2 kcal mol^{-1} accuracy. We also compare our iterative refinement approach to other statistical ensemble approaches and demonstrate that this iterative process reduces estimate errors by as much as 61%. This work provides a new approach for accurate solvation free energy prediction and lays the foundation for future work on aggregate models that can balance computational cost with prediction accuracy.

1 Introduction

Accurate calculation of solvent-solute interactions is an important component of robust molecular simulation including protein structure prediction,[1–3] conformational ensemble calculations,[4–7] and binding free energy calculations.[8–10] Solvation free energy methods for atomically detailed molecular models have a long history of development[11–19] and have recently benefited from the curation of experimental small molecule solvation data for blind

prediction challenges such as the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge studies.[10,20–25] By presenting a larger range of solvation free energies and molecular weights than seen in common public datasets, these challenges are helping to advance the development of solvation methods, making their estimates more accurate and robust for a variety of molecular targets.[26–28]

Top-performing methods typically come from a wide range of modeling strategies[10,20,21,25]. Across different challenges, top performers have included explicit solvation methods,[22,29,30] implicit solvation methods,[4,31] and hybrid methods that combine mixed quantum mechanics (QM) with molecular mechanical (MM) approaches.[32,33] In this context, there is 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 solvation free energies: e.g., uncertainties due to hydrophobicity,[34] surface effects,[35] and solvent asymmetries.[36] This type of *method selection uncertainty* affects a wide range of scientific and mathematical disciplines and is arguably the greatest source of error and risk associated with estimation tasks.[37–40] 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.[41–44] 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 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.[44–47] Ensemble-based estimates are therefore expected to be more reliable and accurate than individual methods, an expectation that has been upheld in numerous examples.[41–44,44,45,45–52]

This paper demonstrates the ability of an ensemble approach called Bayesian Model Averaging (BMA)[44] to estimate solvation free energies for 45 small molecules by combining predictions from 17 diverse methods provided by the SAMPL4 challenge dataset.[25] Though BMA has been applied successfully for prediction tasks in many other domains[48,50–53] this

is the first application of the BMA approach to solvation energy prediction.

Throughout this paper, we use the term "method" to refer to a specific method (and associated set of parameters, settings, etc.) used to estimate solvation free energies, and the term 'model' to refer to some approach for estimating solvation free energies that could be a weighted combination of estimates from one or more methods. We also define the term "optimality", or "optimal performance", as the forecasting method that provides the best *overall* performance for targets randomly selected through cross-validation experiments.

2 Methods

This paper presents a framework for accurately estimating small molecule hydration free energies through ensembles of solvation free energy methods. The paper considers two types of ensembles. In the simplest case, we examine unique combinations of methods for estimating solvation free energy. These ensembles of methods are aggregated via a single statistical model to construct an aggregated estimate. In the next level of ensembles, we examine multiple such statistical models; i.e. an ensemble of models. The framework's objective is to identify the best statistical models, that are in turn based on the best combination of methods. Through Bayesian Model Averaging, we construct and aggregate an ensemble of ensembles (i.e., multiple statistical models that uniquely combine multiple methods) to create a final aggregated model. This approach and ensemble have been shown to provide accurate and reliable estimates for numerous problem types.[41–44,44,45,45–52]

The framework begins by evaluating all solvation methods collectively as an ensemble and uses the ensemble to make an aggregated estimate of free energy. The framework then performs an iterative process of ensemble pruning, analyzing, and estimating. The pruning process uses statistical models that collectively assess the information content of each method. During each iteration, this pruning process removes the most redundant or least accurate methods from the ensemble. After pruning, performance for the new ensemble is assessed through a set of cross-

validation studies. The ensemble that provides the best overall performance is defined as optimal and is the best choice for future estimation tasks (e.g., for future SAMPL challenges).

When an aggregate estimate proves to be more accurate than individual methods, understanding how individual methods were combined can help improve method development. Towards this objective, this framework provides quantitative information to characterize the optimal ensemble by answering the following types of questions. Did the ensemble relied solely on implicit or explicit methods, or if it used variety of different methods? Which methods contributed the most useful information in the aggregate estimate? Alternatively, if the ensemble is largely composed of one type of method, how do variations in parameters influence the results?

2.1 Model specification with Bayesian Model Averaging (BMA)

The simplest BMA approach assembles a set methods into a linear system.[44,46,47] Let y_i for $i = 1, \ldots, N$ be a series of hydration free energy observations for a collection of N molecules, and let x_{ij} denote the free energy estimate for the ith molecule obtained from $1 \leq j \leq P$ prediction methods. The x_{ij} values form a numerical ensemble estimate matrix that, along with y_i, defines a linear regression model

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

where the parameter vector β_j defines the unknown relationship between the ensemble's P constituents and ϵ_i is a disturbance term that captures all factors that influence the dependent variable y_i *other* than the regressors x_{ij}. Such factors include latent variables impacting the response y_i, nonlinearity effects, and measurement errors associated with quantifying y_i, as well as stochastic, experimental effects that impact the ability to reproduce y_i. Explicitly solving for the disturbance term is not feasible: ϵ_i is a theoretical, non-observable random term that must be estimated through an analysis of model residuals. Fortunately, if the distribution of $\epsilon_i : i = 1, \ldots, N$ is normal (i.e., $\epsilon \sim \mathcal{N}(0, \sigma)$) the total disturbance term's contribution to the

model's estimate goes to zero as the expectation of random variables with such distributions is $E[\epsilon|X] = 0$.

Under this assumption of normality, Equation 1 estimates the values β_j that will fit both 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;[54–57] however, these techniques commonly generate estimates that vary in their ability to model and infer.[44,46,47,58,59]

BMA addresses the challenge of statistical model uncertainty by first evaluating all $2^P - 1$ possible statistical models that can be formed from the P estimation methods and combining each model's β_j through a weighted average into an aggregated parameter vector, β_j^{BMA} (Equation 2). 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 an average that weights each $\beta_j^{(k)}$ by the probability that its statistical model, $M^{(k)}$, is the "true" model:

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

where $E[\beta_j^{(k)} \,|\, \mathbf{y}, M^{(k)}\,]$ 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 the data $\mathbf{y}$. The expected posterior distribution of $\beta_j^{(k)}$ is approximated through the linear least squares solution of the given model $M^{(k)}$ and solvation energy response variable $\mathbf{y}$. The posterior probability term is estimated from information criteria[47]

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

where $B^{(k)}$ is the Bayesian Information Criteria for model $M^{(k)}$ estimated as

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

The Bayesian Information Criteria (BIC) is a method for ranking models based on goodness of fit, R^2. The R^2 is a statistical measure of how close the data are to the fitted regression line, essentially the percentage of the response variable variation that is explained by a linear mode. The BIC strongly penalizes models for having too many parameters, striking a parsimonious balance between sufficient complexity for a model to be useful and over-fitting the data. In Equation 4, $R^{2(k)}$ is this adjusted R^2 for 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 solvation free energy values to be predicted (i.e., the number of molecules).

The 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 solvation free energy values y_i and values x_{ij} predicted by each solvation method j. Over-fitting by more complicated models is addressed through the Bayes information criterion, which penalizes models with more variables (i.e., statistical models that use more methods in their regression). Perhaps more importantly, β_j^{BMA} can be used to estimate new solvation free energy values for unmeasured molecules by combining new x_{ij} estimates.

2.2 Pruning models with Occam's window

The inclusion of all $2^P - 1$ ensemble models in Equation 2 is not necessarily beneficial for predictive performance. A "misspecified" ensemble model may perform poorly if it omits a method (and resultantly information) that is crucial to estimating y_i. Alternatively, a model may perform poorly if it includes methods that provide inaccurate information about y_i. Despite the fact that they are down-weighted via low posterior probabilities, the cumulative effect of these misspecified models, can erode the ensemble's overall performance.[44,46,50,60–63] Madigan and Raftery[63] developed the Occam's Window ensemble-pruning approach that eliminates under-

performing models based on the BIC discussed above. Occam's Window defines a set of models

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

where $BIC^{(min)}$ denotes the BIC of the model $M^{(k)}$ with the lowest BIC (highest information content). The value 6 is based on Jeffreys'[64] and Raftery's[47] assessment of Bayes factors for model comparison and ensures all $m_k \in \mathbf{A}$ meet a minimum statistical information criteria for the aggregation process. Constraining Equation 5 to the set $\mathbf{A}$ accelerates the evaluation of Equation 5 and improves BMA's predictive capability:[46,63]

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

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

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

Equation 7 sorts and prioritizes 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 7 therefore provide a statistical framework to support an iterative, statistical process for designing an ensemble that will provide the best combination of models and methods for estimating solvation free energy.

2.3 Statistical design of method ensembles

The statistical design process, shown in Algorithms 1 and 2, proceeds in two iterative stages. In both algorithms, we denote matrix and vectors with bold script and indicate row and column dimensions in superscript. For example, $\mathbf{x}^{n\times p}$ is a matrix with n rows and p columns, where as x is a scalar, or real value. Additionally, $\mathbf{0}^{n\times p}$ is a zero matrix and $\mathbf{x}^{1\times p}$ is a zero vector; essentially, a matrix and vector of the specified dimensions where all values are 0.

The first stage, presented in Algorithm 2, uses Equation 7 to evaluate the statistical significance of each method in the current version of the ensemble. This information is based on a set of 2-fold cross-validation tests that randomly create training and validation datasets from the small molecule compounds in the SAMPL4 challenge. In these tests, each method's statistical information is thus evaluated against a variety of random training and validation sets. The results of this first stage are then sent to stage two (Algorithm 1): the ensemble pruning stage.

The second stage evaluates the performance of the current ensemble and then prunes this ensemble by removing that method with the least statistical significance. The process then repeats using the new pruned ensemble to identify the ideal ensemble of methods.

At a more detailed level, the design 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. The algorithm will iterate over the following tasks while it still has more than two methods to evaluate (lines 5 – 19).

First, the performance of the current ensemble is evaluated through *Analyze Methods* (line 6). The *Analyze Methods* function in Algorithm 2 takes as input the same observation data $\mathbf{y}$ as Algorithm 1 as well as a matrix $\mathbf{m}^{n\times j}$ that contains estimates made by a subset of $j \leq p$ methods. The matrix $\mathbf{m}^{n\times j}$ represents the current set of methods from $\mathbf{x}^{n\times p}$ still being analyzed. 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 Equation 6 to estimate β_{BMA} and then calculates the RMSE of

Algorithm 1
Prune and Design Ensemble

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

1: $loopCount = 1$
2: $rmse = null$
3: $\mathbf{x_solution} = \mathbf{x}$
4: $\boldsymbol{\beta}_{BMA}^{1\times j} = \mathbf{0}^{1\times j}$ {initialize coefficients to zero}
5: **while** (loopCount < p) **do**
6: $\quad \boldsymbol{\beta}_{BMA}, s, \mathbf{v} =$ **Analyze Methods**($\mathbf{y}$, $\mathbf{x}$)
7: $\quad$ **if** (s < rmse) or (rmse == null) **then**
8: $\quad\quad \mathbf{x_solution} = \mathbf{x}$
9: $\quad\quad rmse = s$
10: $\quad$ **end if**
11: $\quad m = 1$
12: $\quad$ **for** ($methods \in \mathbf{x}$) **do**
13: $\quad\quad$ **if** ($\mathbf{v}[methods] < \mathbf{v}[m]$) **then**
14: $\quad\quad\quad m = methods$
15: $\quad\quad$ **end if**
16: $\quad$ **end for**
17: $\quad \mathbf{x} = \mathbf{x}$.delete($m$) {prune least significant method}
18: $\quad loopCount+ = 1$
19: **end while**
20: **return** $\mathbf{x_solution}, \boldsymbol{\beta}_{BMA}$, rmse

Algorithm 2
Analyze Methods

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

1: $cv = 100$ {100 cross-validation experiments}
2: $\boldsymbol{\beta}_{BMA}^{1\times j} = \mathbf{0}^{1\times j}$ {initialize coefficients to zero}
3: $error = 0$ {mean RMSE of the cross-validation}
4: $\mathbf{v}^{1\times j} = \mathbf{0}^{1\times j}$ {vector storing each method's p≠0}
5: **while** ($index < cv$) **do**
6: $\quad \mathbf{m}_t^{q\times j}, \mathbf{m}_v^{r\times j} = \mathbf{m}^{n\times j}$ {split $\mathbf{m}$: train and test}
7: $\quad \boldsymbol{\beta}_{BMA} = \boldsymbol{\beta}_{BMA}$+BMA($\mathbf{m}_t$) {Equation 6}
8: $\quad error = error + RMSE(\mathbf{Y}, \mathbf{m}_v \times \boldsymbol{\beta}_{BMA})$
9: $\quad \mathbf{v} = \mathbf{v}$+P≠0($\mathbf{m}_t$) {Equation 7}
10: $\quad index = index + 1$
11: **end while**
12: **return** $\frac{\boldsymbol{\beta}_{BMA}}{cv}, \frac{error}{cv}, \frac{\mathbf{v}}{cv}$

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, $\mathbf{v}$ (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 *Analyze Methods* (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 $\mathbf{y}$ and a statistical model, β_{BMA} that specifies how to best combine these methods.

2.4 Solvation free energy data and solvation methods

The SAMPL4 challenge consists of 49 submissions representing a total of 19 different research groups.[25] Each of these methods provides solvation free energy estimates for 45 small

molecule compounds.[1] The challenge is blind in the sense that the solvation free energy values for these molecules were hidden from participants; e.g., free energy values are not found in standard solvation free energy test sets, and their values are not readily available in the literature.[65] In this work, we restrict our analysis to a subset of 17 method 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 specific methods during model selection and averaging; such bias can create unstable estimates for β_j^{BMA} and that in turn can reduce BMA's estimate accuracy.[66] These methods are summarized in Table I based on the methods used to calculate the solvation energy:

- **Group 1:** Single-conformation implicit solvent methods[26,67,68] based on Poisson-Boltzmann and related methods;[69–71]
- **Group 2:** Multi-conformational implicit solvent methods;[72–76]
- **Group 3:** Methods based on molecular dynamics-based free energy calculations in explicit solvent[22,27,30,77] with small molecule force fields;[78] and,
- **Group 4:** Hybrid solvent methods.[79]

2.5 Training, estimating, and pruning an ensemble with BMA

From this initial set of 17 methods, we iteratively assessed and pruned methods as described in Algorithm 1. The approach began by randomly sampling (without replacement) 26 of the original 52 experimental solvation free energy measurements to form training and validation datasets (Algorithm 2, line 6). Collectively, these sampled values formed the observation vector y_i and the estimates from each of the prediction methods in Table I for

[1] The original SAMPL4 study used 52 samples; thus some samples have labels like SAMPL4_052. However, several compounds were removed during the course of the study based on "problems with experimental values, SMILES strings, or structures for a number of compounds, resulting in removal of some compounds from the challenge." [25]

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. CPU time is an order-of-magnitude estimate of the method's reported time to make a prediction. 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 as described in the text. 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% (i.e., 1.2–9.4 kcal mol^{-1}) in comparison to the individual methods. The final BMA ensemble is indicated by the two italicized 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 shown in Table II.

Reference	Method ID	CPU Time	Methodology	Ensemble Mean RMSE (kcal mol^{-1}) and Std Err Mean	Improvement with BMA (Stage 16)
Coleman et al.[80]	imp-6	seconds	multi-conformation implicit	9.40 ± 0.2	91%
Parsod et al.	alc-5	hours	alchemical MD	4.50 ± 0.071	82%
Sharp et al.[81]	imp-4	minutes	single conformation implicit	3.80 ± 0.045	78%
Jiafu et al.	exp-3	hours	QM/MD	2.80 ± 0.045	71%
Purisima et al.[74]	imp-5	minutes	single conformation implicit	2.55 ± 0.049	68%
Mark et al. 2011	alc-2	hours	alchemical MD	2.40 ± 0.036	66%
Genheden et al.[82]	exp-1	hours	MM-PB/SA	2.00 ± 0.022	59%
Weyang et al.	exp-4	hours	MM-PB/SA	1.84 ± 0.031	55%
Biorga et al.[83,84]	alc-4	hours	alchemical MD	1.65 ± 0.013	50%
Elingson et al.[26]	imp-7	seconds	single conformation implicit	1.52 ± 0.039	46%
Fennell et al.[79]	hyb-2	hours	hybrid	1.52 ± 0.013	46%
Jambeck et al.[85]	alc-1	hours	alchemical MD	1.52 ± 0.020	46%
Park[86]	imp-3	seconds	single conformation implicit	1.44 ± 0.020	43%
Klamt et al.[87]	imp-1	hours	single conformation implicit	1.36 ± 0.030	40%
Gilson et al.[27]	*alc-3*	hours	*alchemical MD*	1.24 ± 0.017	34%
Geballe et al.[26]	imp-8	seconds	single conformation implicit	1.17 ± 0.017	30%
Sandberg et al.[72]	*imp-2*	minutes	*multi-conformation implicit*	*1.15 ± 0.023*	*29%*
BMA (Stage 16)	**NA**	seconds	**ensemble**	**0.82 ± 0.015**	**0**

these measurements defined the ensemble estimate matrix, x_{ij}. The observation vector and the ensemble estimate matrix were used to form the linear system in Equation 1.

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

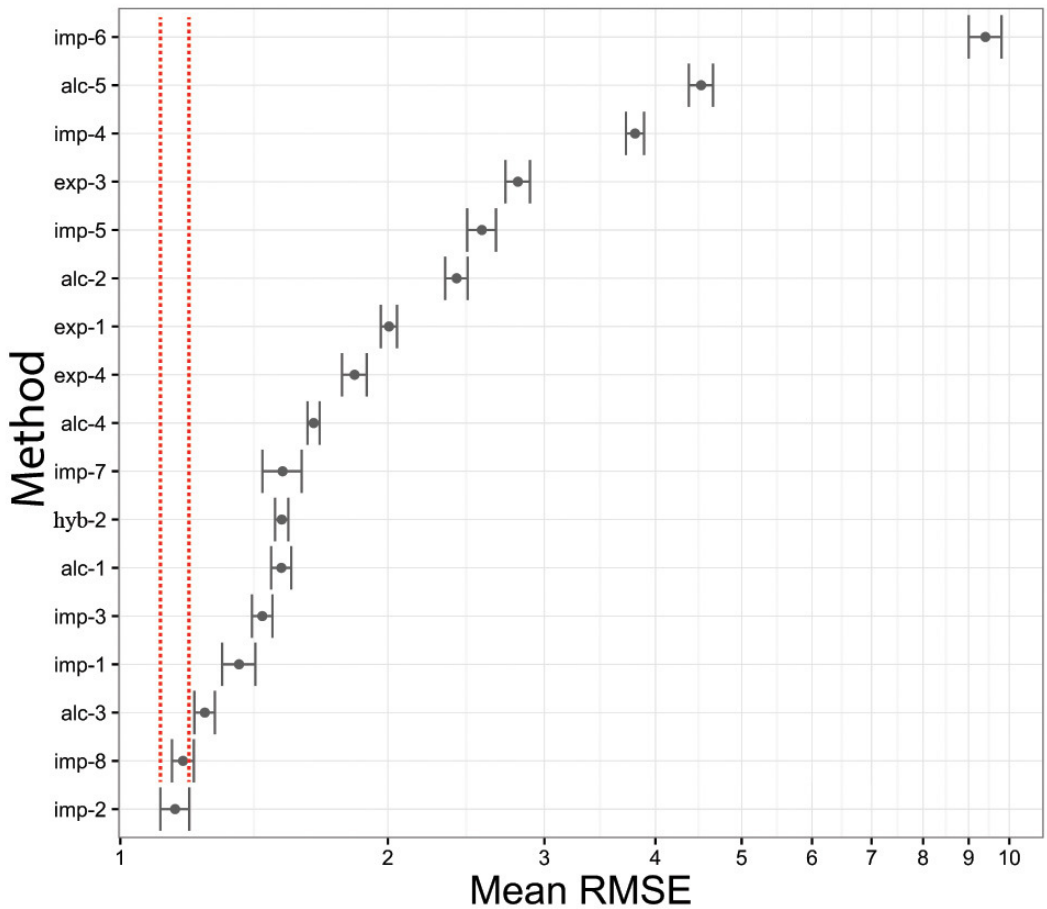

Figure 1: This figure uses a squadron of Imperial Tie Fighters[88] to depict the mean root mean squared error (RMSE) for the 17 initial methods used in our ensemble design process. Additionally, the 95 % confidence interval (based on standard error) for the expected mean RMSE is shown for each method. Thus at 95% statistical confidence, the true expected performance for each method is depicted according to the 100 iterations of the 2-fold cross-validation experiment we detail in Algorithm 2.

with β_j^{BMA} to produce an aggregated estimate. A RMSE was obtained for the 26 estimates made on validation data.

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

Algorithm 1 used this information to perform two tasks. First, if the performance of

this aggregated estimate was better than all previously examined aggregates, the statistical model combining these methods was saved as the new optimal model (Algorithm 1, lines 7-9). Next, Algorithm 1 pruned a method from the ensemble that provided the least amount of information to the aggregated estimate. The pruning process was repeated until just two methods are left (Algorithm 1, lines 12-18). The ensemble of methods whose aggregate forecast has the lowest mean RMSE was saved and returned as the final model for the ensemble. This final model, and the set of methods that correspond to this model, were the final products of the statistically driven ensemble design process, referred to as the BMA-based optimal ensemble.

3 Results and Discussion

We report the results of our BMA-based ensemble in three stages. First, we contrast the performance of the initial 17 SAMPL4 methods to the performance of the BMA-based optimal ensemble. We then present data on the iterative design process used to create this optimal ensemble. Second, we examine the *conditional* performance of this optimal ensemble according to the individual molecules used in the SAMPL4 challenge. 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. Finally, we complete 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.

In all analyses, we report the performance based on the mean RMSE associated with the 100 $\times$ 2-fold cross-validation. Further, we report the standard error of the mean (SEM) to better characterize the expected performance for these estimates. In contrast to standard deviation, which quantifies the spread of the cross-validation distribution, the SEM metric directly maps cross-validation performance into confidence intervals that bound the true expected performance

of each RMSE.

We report the statistical significance of all performance data through a Wilcoxon rank sum paired comparison test.[89] This non-parametric approach tests the hypothesis that the mean RMSE distributions of two approaches are equal: $H_0 : \mu_Y = \mu_X$. Thus when comparing BMA to a given method, a Wilcoxon generated p-value greater than 0.05 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 0.05 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.

To control the family-wise error rate of our tests, we applied a Bonferroni correction to determine p-values with a threshold of $\alpha = 0.05$. This correction addresses a problem that arises in statistical analysis for a large number of concurrent comparisons. Larger comparison spaces are more likely to yield a statistically significant observation by chance, resulting in a false positive. Bonferroni corrections adjust the false positive rate to account for this effect.

Finally, a Shapiro-Wilk test[90] was used to test the hypothesis, H_0, that residuals in the SAMPL4 challenge are not significantly different from a normal population. This test is used to determine if samples from our distribution or residuals are distributed normally. Thus, under an $\alpha = 0.05$, a p-value statistic of 0.9556 was obtained from this test, indicating that there was insufficient information to reject H_0. Without evidence to reject H_0, assumptions of normality indicate that the expectation for the disturbance terms ϵ_i in Equation 1 (e.g., sources associated with measurement error, stochastic effects, etc.) sum to 0 and we ignore these disturbance terms.

3.1 Comparing estimates from BMA's optimal ensemble to SAMPL-4 challenge methods

The performance of all methods used in this work is shown in Table I and Figure 1. Column 3 in Table I lists the specific method for each estimation approach. The performance results in

column 4 of this table are consistent with results reported by Mobley et al.[25] 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 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 illustrate the performance variability across the different methods. These plots also underscore the amount of uncertainty inherent in estimating solvation free energies: method-selection uncertainty plays a confounding factor in estimating solvation 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 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 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 were 17 methods in the ensemble and imp-6 was selected to be removed. During the next step, Stage 2, imp-6 was removed from the ensemble so that only 16 methods were used to create the aggregated estimate. At the end of this stage, alc-4 was selected to be pruned for Stage 3. In the final stage, the only remaining methods in the ensemble were imp-2 and alc-3. This iterative design process is also graphically illustrated in Figure 2 showing how selective pruning increases the performance of each successive ensemble. The red line in this figure indicates the performance of the best performing method, imp-2. The benefits of the aggregated estimates become apparent after Stage 6 where the ensembles outperform imp-2.

The statistical significance of the iterative design process is listed in columns 4 and 5 in

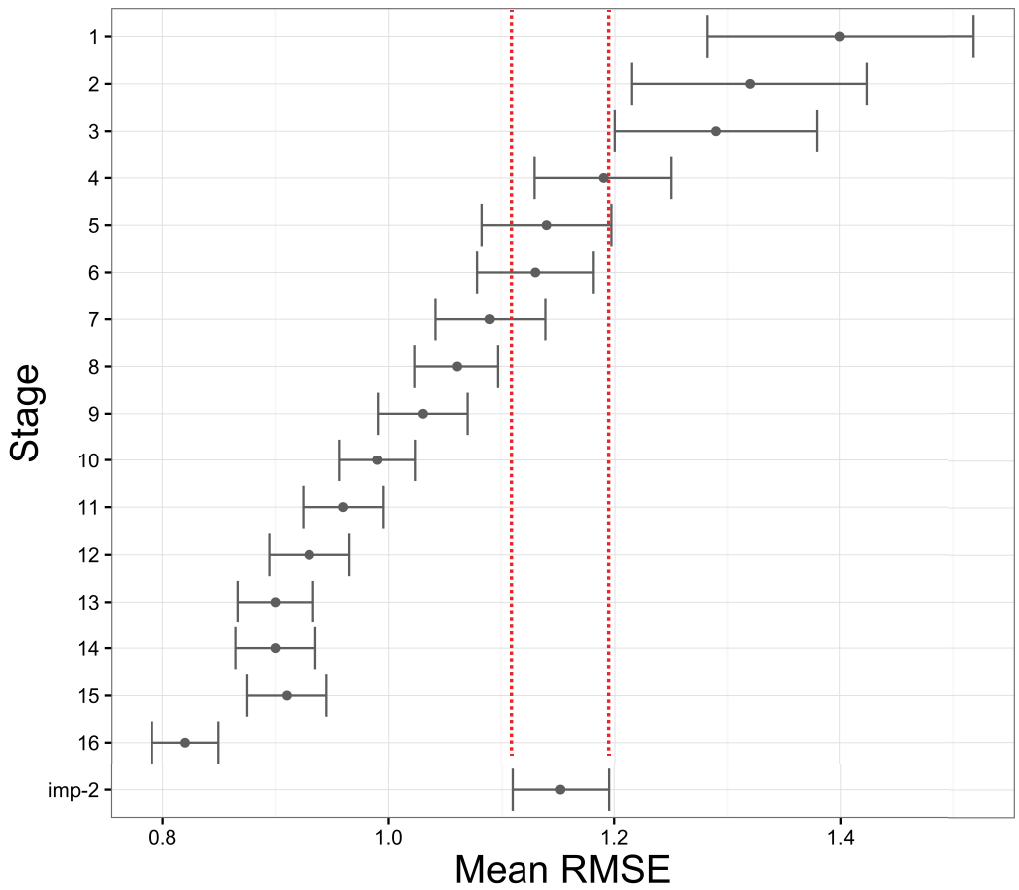

Figure 2: This plot illustrates the iterative pruning process discussed in Section 2.2. The y-axis depicts the mean root mean squared error (RMSE) of the different aggregated estimates based on ensembles formed during the different iterations of pruning. The x-axis indicates the stages of pruning. In general, the mean RMSE and the 95% confidence intervals for each mean (based on standard error for the mean SEM) were reduced 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.

Table II; bold p-values in these columns indicate the performance between two distributions are equivalent. Column 4 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 indicates that these distributions are equivalent. Contrariwise, in row 4, the p-value for the comparison between Stage 4 and Stage 3 indicates that the distributions are not equivalent; the mean RMSE in column 2

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. p-values that are greater than 0.05 are shown and indicate distributions that are equivalent (e.g., Stage 3 and 2) based on an $\alpha = 0.05$. 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, bold 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.

Stage	Ensemble RMSE (kcal mol^{-1}) Mean and Std Err Mean	Method Selected for Pruning	p-value (sequential stages)	p-value (stage vs. imp-2)	Improvement with BMA (Stage 16)
1	1.40 ± 0.060	imp-6	NA	0.000	41%
2	1.32 ± 0.053	alc-4	**0.411** (vs. Stage 1)	0.002	38%
3	1.29 ± 0.046	imp-8	**1.000** (vs. Stage 2)	0.004	36%
4	1.19 ± 0.031	imp-5	0.012 (vs. Stage 3)	**0.356**	31%
5	1.14 ± 0.029	imp-3	0.000 (vs. Stage 4)	**0.438**	28%
6	1.13 ± 0.026	imp-7	**0.883** (vs. Stage 5)	**0.306**	27%
7	1.09 ± 0.025	alc-5	0.015 (vs. Stage 6)	0.027	25%
8	1.06 ± 0.019	alc-2	**0.055** (vs. Stage 7)	0.003	23%
9	1.03 ± 0.020	imp-1	**1.000** (vs. Stage 8)	0.001	20%
10	0.99 ± 0.017	imp-4	0.000 (vs. Stage 9)	0.001	17%
11	0.96 ± 0.018	alc-1	0.000 (vs. Stage 10)	0.000	15%
12	0.93 ± 0.018	exp-4	**1.000** (vs. Stage 11)	0.000	12%
13	0.90 ± 0.017	hyb-2	**1.000** (vs. Stage 12)	0.000	09%
14	0.90 ± 0.018	exp-1	**1.000** (vs. Stage 13)	0.000	09%
15	0.91 ± 0.018	exp-3	0.001 (vs. Stage 14)	0.000	10%
16	0.82 ± 0.015	NA	0.000 (vs. Stage 15)	0.000	0%

indicates that the ensemble of Stage 4 outperforms the ensemble built in Stage 3.

As a second analysis of significance, column 5 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. However, the mean RMSE distributions of all subsequent stages are not equivalent: based on mean RMSE listed in column 2, we conclude that these successive ensembles (increasingly) outperform imp-2. The p-values in these columns and the mean RMSE results demonstrate that the optimal ensemble is created in Stage 16. The final column in Table II lists the performance improvement that the

Stage 16 ensemble provides in comparison to all the other ensembles generated in the design process.

Figure 3 depicts a heat-map that shows the statistical information that drove the pruning process for each stage. In this image, the y-axis represents the different methods; the x-axis

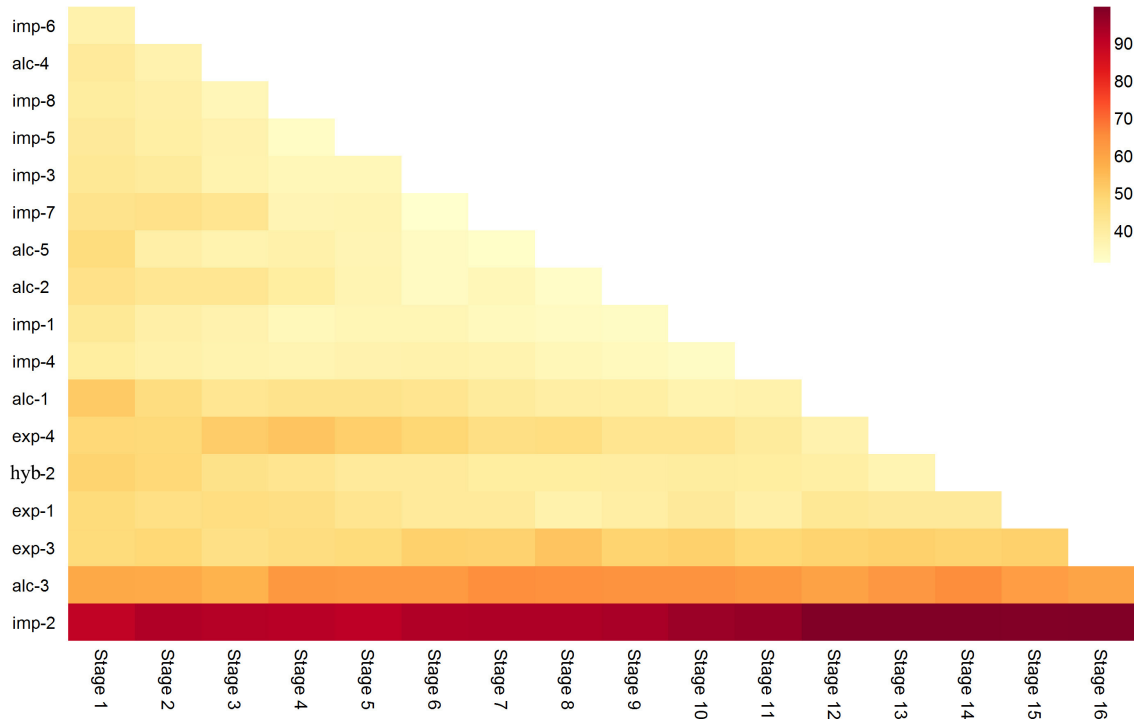

Figure 3: This image provides a graphical depiction of the ensemble design process performed on the initial ensemble of 17 methods. The color scale represents the probability, 0 - 100%, that a given method's coefficient term, β_j will not be zero. The stages in the design process, starting at the left 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.

(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. At Stage 1, all methods were used in the ensemble and their $\Pr(\beta_j^{\text{BMA}} \neq 0)$ ranged

from 40% (imp-6) to 80% (imp-2). As imp-6 had the lowest mean probability of not being zero, imp-6 was pruned after Stage 1; the color map colors this method white in Stage 2 to indicate it has been eliminated. The methods listed on the legend of the y-axis are ordered by the sequence that they were eliminated in the design process: imp-6 first, alc-4 second, imp-8 third, etc.

⋆⋆ **Check section title capitalization** ⋆⋆

3.2 Performance analysis based on compounds

Figures 4 and 5 depict the 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 and exp-3 is the final method eliminated from the ensemble (Stage 15 in Table II).

Mobley et al.[25] note that prediction error generally increases as molecular complexity and size increases. Specifically, polyfunctional molecules with several interacting groups 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). Figures 4 and 5 show that 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. 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. The general trend in performance indicates that the optimal ensemble provides more consistent and accurate estimates than any specific method.

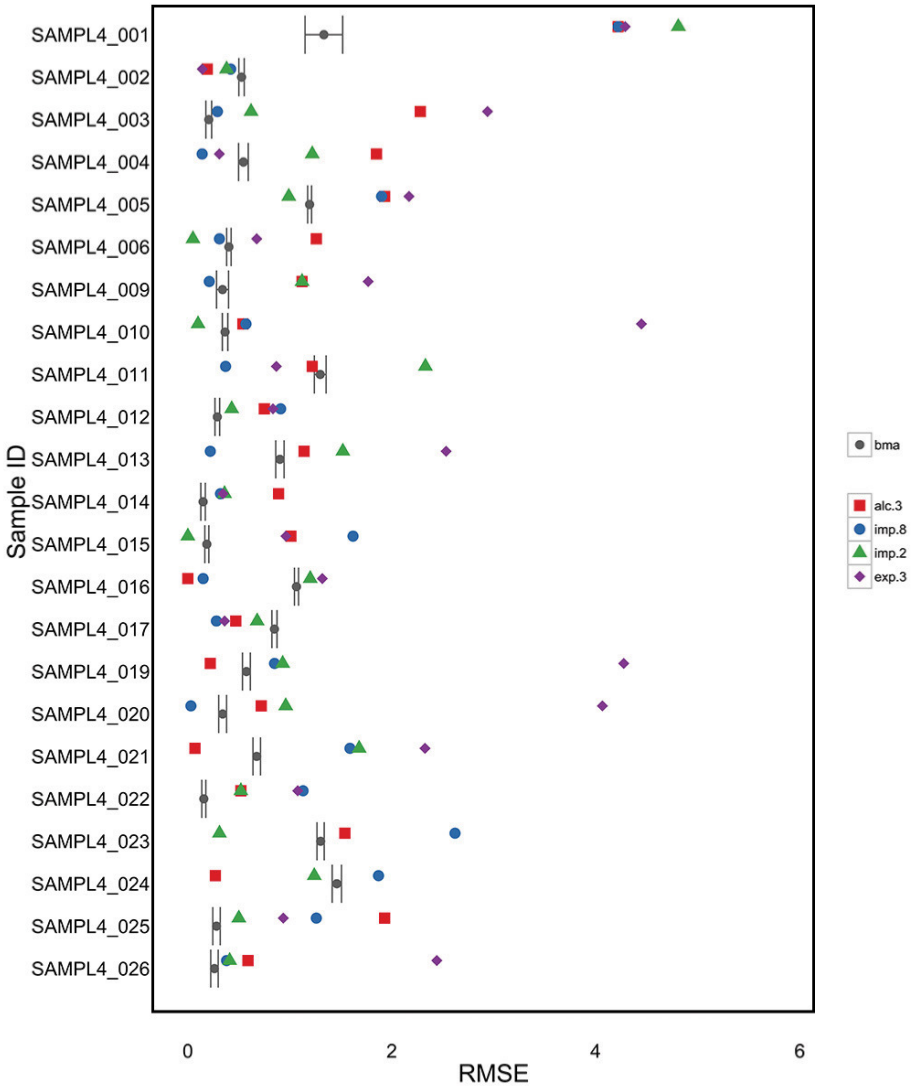

Figure 4: This figure is one of two figures (Figures 4 and 5) that depict the RMSE (kcal mol^{-1}) performance of several methods based on the individual compounds taken from the SAMPL4 challenge: the first, second, and third best-performing methods (i.e., imp-2, imp-8, and alc-3) as well as exp-3. Note 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 RMSE for estimates made in our 2-fold cross-validation analysis. Of 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.[25] 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.

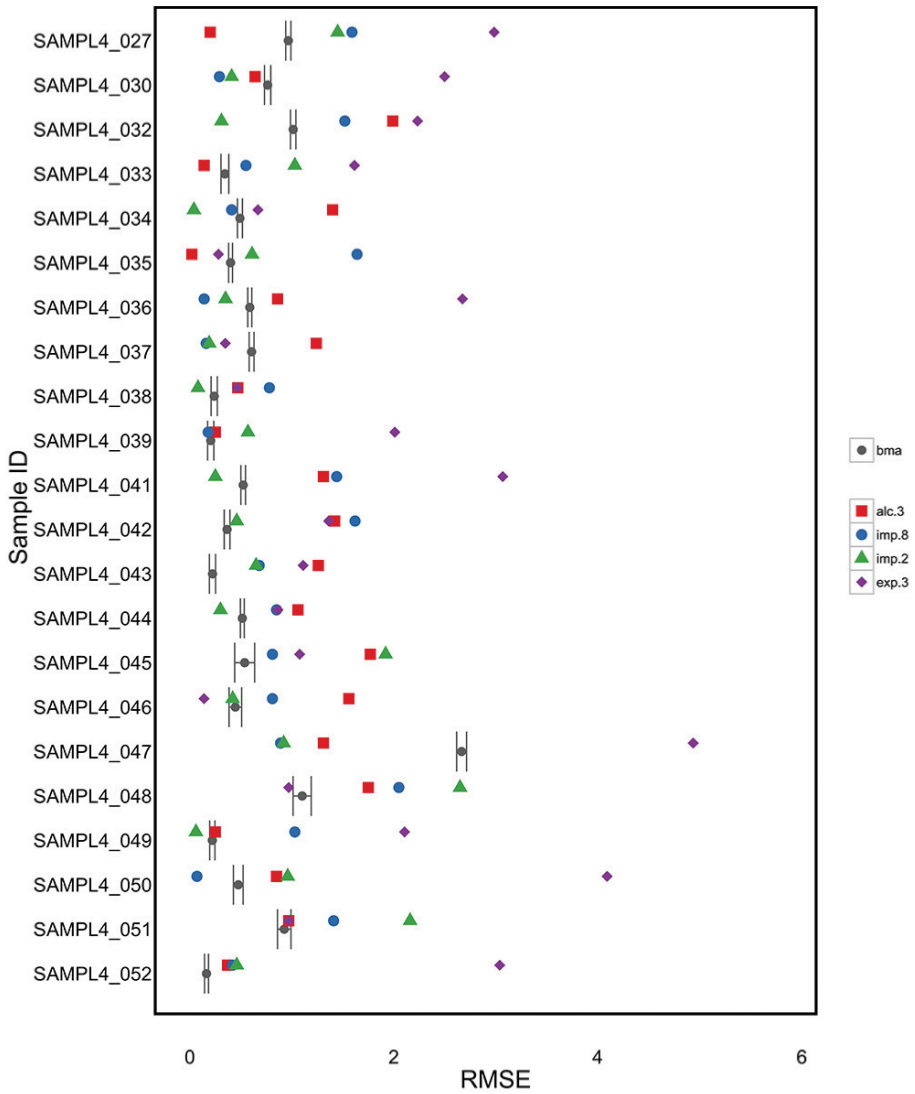

Figure 5: This figure is one of two figures (Figures 4 and 5) that depict the RMSE (kcal mol^{-1}) performance of several methods based on the individual compounds taken from the SAMPL4 challenge: the first, second, and third best-performing methods (i.e., imp-2, imp-8, and alc-3) as well as exp-3. See Figure 4 for more information about both plots.

3.3 Interpreting aggregate estimate performance

Aggregate estimate performance is degraded by both variable (i.e., inconsistent) predictive performance of its method constituents and the uncertainties associated with statistical model construction. These sources of error manifest in different ways and we can characterize trends in estimation accuracy and consistency to support our analysis of these sources.

If methods are sensitive to functional discrepancies between training and testing molecules, then aggregate estimates that rely on these methods may perform poorly on chemical compounds that are outside the training set. While some of these sensitivities may be obvious to identify (e.g., methods that are only accurate for certain chemical classes), most are much more subtle and difficult to detect. Cross-validation tests, like those performed in Section 2.3, help assess the more obvious class or functional group-based sensitivities. More specifically, under cross-validation tests the overall performance of the aggregate estimate will be significantly poorer in comparison to other methods. This trait will be most notable when comparing the aggregate against its method constituents on molecules where constituents are very accurate. In general, as method sensitivity increases, estimate accuracy and eventually consistency are negatively impacted.

The overall performance of the aggregate estimate is better than its constituents (Table 1, Figures 4 and 5,) and significantly better than other methods. Based on these observations, large method-based sensitivities do not appear to be an issue for the SAMPL4 challenge data and so we next proceed to assess the sensitivities associated with statistical model construction.

Statistical sensitivities in a model manifest in estimate consistency in a cross-validation study. In this work, the 95% confidence intervals of the expected mean provide a quantitative metric for assessing these types of sensitivities. Under the assumption of minimal method sensitivity, a perfect statistical model will provide very precise and accurate estimates, regardless of training and testing data. The further away a statistical model is from this hypothetical "perfect" model, the more uncertainty will impact how weights are assigned in its coefficient vector, β_j. As a result, changes in training or testing data will results in an increasingly wider distribution of estimates (i.e., wider confidence intervals) and will impact the model's accuracy. In extreme cases, these

distributions may even become bimodal.

The overall performance of the aggregate estimate is very consistent across chemical compounds (Figures 4 and 5,) with the exception of SAMPL4_001 and SAMPL4_045. For SAMPL4_001 and SAMPL4_045, the 95% confidence intervals for the expected mean are noticeably wider then other compounds. As these estimates are still very accurate in comparison to the ensemble's constituents, it is reasonable to assume that the underlying statistical model of the optimal is largely valid across the space of chemical compounds defined by the SAMPL4 challenge. Thus, while there appears to be indications of statistical sensitivity, it may be more effective to address the performance of methods for these two compounds first before refining the statistical model.

Given the assumption of minimal method and statistical sensitivities, there are several chemical compounds that deserve specific attention: SAMPL4_002, SAMPL4_017, SAMPL4_024, SAMPL4_030, and SAMPL4_047. In all of these cases, aggregate's estimate is outperformed by its constituents. For three of the compounds, SAMPL4_002, SAMPL4_017, and SAMPL4_030, the degree of error is very small in comparison to the ensemble's constituents. It is likely that statistical uncertainties are playing a role in degrading performance for these three such that the underlying model has lost some accuracy, but maintained consistency. Addressing these concerns is usually best done by increasing the size of training and testing pools to provide the statistical design process more exemplar data.

In comparison to the previous compounds, aggregate estimate performance for SAMPL4_024 and SAMPL4_047 is notably poor. Further, the usual indicators of method and statistical sensitivities do not provide evidence to explain this performance loss. For example confidence bounds for these estimates are narrow, indicating that estimates are consistent regardless of the cross-validation training and testing sets. Additionally, total mean estimation errors are largely restricted to these two specific chemicals, indicating that overall the model is valid and accurate for the SAMPL4 challenge data.

One explanation for these trends may be that the current optimal ensemble experiences a more subtle statistical or method-based sensitivity for these specific chemical compounds. For exam-

ple, Mobley et al.[25] indicate that these two specific chemical compounds are the largest MW compounds evaluated by the SAMPL4 challenge: SAMPL4_024 is amitriptyline (MW 277.41) and SAMPL4_047 is 1-(2-hydroxyethylamino)-9,10-anthraquinone (MW 267.28). Additionally, SAMPL4_047 is listed as the second largest hydration free energy in the challenge. It is thus feasible that the ensemble design process has identified a statistical model that is valid for lighter weight, lower hydration free energy compounds but is not sufficiently accurate for larger compounds with higher free energy values. To test this hypothesis, the ensemble design process would need to expand to include more chemical exemplars that are similar to SAMPL4_024 and SAMPL4_047.

This statistical framework has designed a single model for combining two methods in order to estimate hydration free energies for any compound in the SAMPL4 challenge. These method constituents are based on two different approaches for predicting hydration free energy: one is based on a multi-conformational implicit method and requires tens of minutes to make a prediction, where as the other is based on molecular dynamics and requires many hours to make the same prediction. One benefit of this approach therefore is not only in improving estimation accuracy, but in providing this increase in accuracy by combining computationally expensive methods with cheaper methods. We envision that through this type of framework, specific ensembles of methods will be able to be developed for different classes of chemical compounds. In this way, a stratification of the chemical landscape (e.g., based on MW, functionality and polarization, etc.) is possible where specific classes of compounds are mapped to distinct ensembles, and these ensembles are able to provide the best estimates for hydration free energies at the lowest possible computational cost.

3.4 Performance analysis of alternate ensemble techniques

There are alternative approaches to BMA that can combine an ensemble of methods into an aggregate estimate. The notion of method and statistical sensitivities are also points of concern for these alternate approaches. In our cross-validation study, we evaluated four common

approaches for aggregating an ensemble and evaluated their estimated benefits in comparison to BMA. These methods are listed in Table IV and include: Random Forest,[91] Ridge Regression,[92] Lasso,[93] and stepwise regression via forward selection.

The difference between the BMA approach and these other approaches is in their design process. More specifically, each approach has a strategy for performing model specification (e.g., Section 2). As we have addressed the notion of method sensitivities in the last section, the objective for this section is to compare these different approaches for model construction to highlight the effects of model sensitivities. As each method will select its own methods to form its own ensemble, the difference in performance for these approaches will highlight the benefits of addressing model sensitivities. In other words, relying on a modeling approach that does not fully address these sensitivities can mislead analysis of ensemble performance and thwart subsequent ensemble design.

The random forest approach is an ensemble learning method similar to BMA. It operates by constructing multiple decision trees at training time and outputs an estimate based on the mean prediction of the individual trees. Where as the random forest algorithm relies on the non-informative weighting, BMA's approach relies on a Bayesian weighting for each estimate. The random forest approach is well established in computational biology and bioinformatic applications.

The ridge regression approach is a technique for analyzing regression data that may suffer from multicollinearity. When multicollinearity occurs, it can create inaccurate estimates of the regression coefficients, β_j, resulting in an increase of standard error. Ridge regression addresses the concerns for high variance by adding a degree of bias to the regression estimates that will reduce overall estimate errors.

Lasso and forward-selection are both strategies for selecting methods to use in a linear regression model (Equation 1). In forward-selection, the approach starts with the assumption that all coefficients, β_j are equal to zero. Next, it selects the method, x_ij most correlated with y_i, and adds it into the model. It then assesses the performance of the model and continues to add

Table III: This table lists the different methods used by each modeling approach in Section 3.4. The performance for each of these methods is listed in Table IV and Figure 6. Starting from the top and working down, the total number of methods used in each ensemble are: Random Forest 14, Ridge Regression 7, Lasso 4, Forward Selection 4, and BMA 2.

Model	imp-1	alc-5	imp-4	exp-3	imp-5	alc-2	exp-1	exp-4	alc-4
Random Forest	X	X	-	X	-	X	X	X	-
Ridge	-	X	-	-	-	X	X	-	-
Lasso	-	-	-	-	-	-	-	-	-
Forward Selection	-	-	-	-	-	-	-	X	-
BMA (Stage 16)	-	-	-	-	-	-	-	-	-

Model	imp-7	hyb-2	alc-1	imp-3	imp-1	alc-3	imp-8	imp-2
Random Forest	X	X	X	X	X	X	X	X
Ridge	-	-	X	X	-	X	-	X
Lasso	-	-	X	X	-	X	-	X
Forward Selection	-	-	-	X	-	X	-	X
BMA (Stage 16)	-	-	-	-	-	X	-	X

the next highest correlated methods to the model until performance no longer increases. Lasso follows the same general approach, but doesn't add a method fully into the model. Instead, Lasso penalizes the absolute size of the regression coefficients until the coefficient, β_j for the method being added is no longer the one most correlated method with overall performance.

Table III lists the methods that were selected by each modeling approach. From largest to smallest ensemble the modeling ensembles are: Random Forest 14, Ridge Regression 7, Lasso 4, Forward Selection 4, and BMA 2.

These techniques were chosen as they have all been used successfully for a variety of inference tasks and are readily available for use.[94,95] As all of these approaches construct an estimate for β, training and estimating with these approaches was performed identically to BMA, using the cross-validation detailed in Section 2.2. More specifically, the performance reflects the result of 100-iterations of a 2-fold cross-validation. We also followed the same procedure for comparing BMA's estimating capability to these alternate ensemble-based techniques. Figure 6 and Table IV 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 IV. Based on an

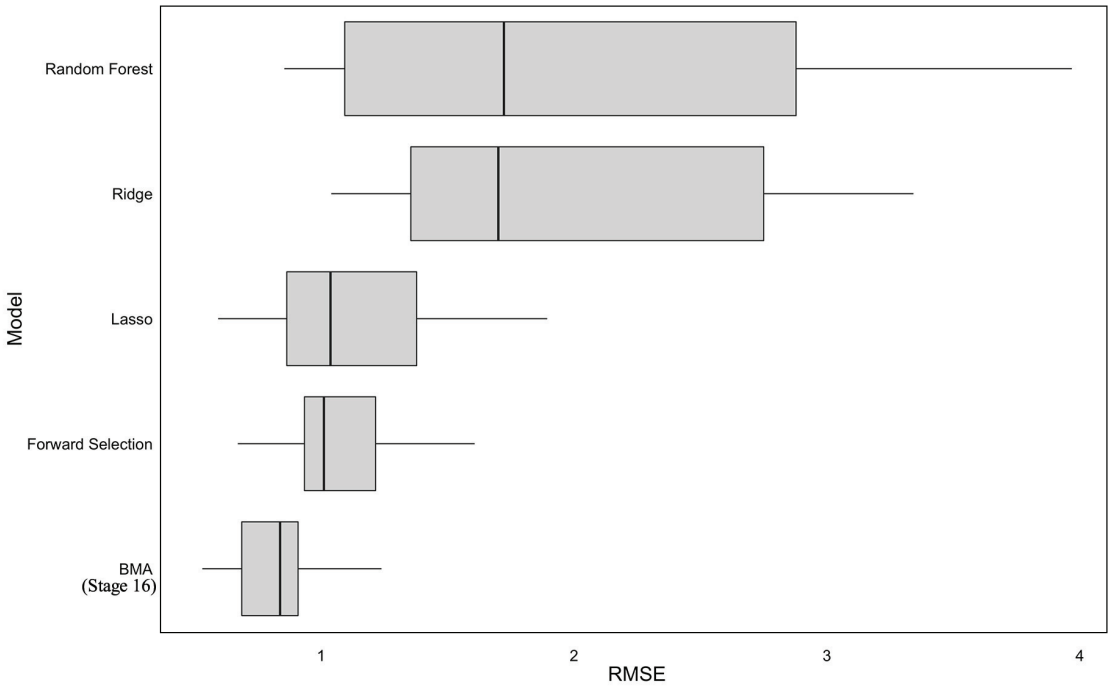

Figure 6: This figure displays the performance of different ensemble approaches in comparison to the optimal ensemble designed in this work, BMA (Stage 16). The mean RMSE, min, max, first, and third quartiles (kcal mol^{-1}) of these ensemble approaches are shown based on the 100 iterations of the 2-fold cross-validation experiment discussed in Section 2.2. Based on an $\alpha = 0.05$, the Wilcoxon based p-values (Table IV) 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.

Table IV: 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 RMSE 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 Std Err Mean	p-value	Improvement by BMA (Stage 16)
Random Forest	2.08 ±1.00	0.00	61%
Ridge	2.06 ±0.75	0.00	60%
Lasso	1.12 ±0.34	0.00	27%
Forward Selection	1.09 ±0.26	0.00	25%
BMA (Stage 16)	0.82 ±0.17	NA	0%

$\alpha = 0.05$, Table IV 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 IV. From these mean RMSEs, it is clear that the BMA-based approach outperforms all other ensemble-based estimation 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%.

4 Conclusions

This study demonstrates a proof-of-principle application of how to statistically design and aggregate an ensemble of methods for estimating solvation 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.

In future work, we will investigate the application of this method to broader regions of chemical space. Specifically, we envision that this type of framework can help stratify the small molecule chemical landscape based on chemical classes defined by molecular weight, functionality and polarization, available surface area, etc. This stratification will map specific classes of compounds to distinct ensembles that provide the best estimates for hydration free energies at the lowest possible computational cost. Towards this task, we are looking at penalizing computationally expensive methods that provide minimal accuracy benefits to optimize both the accuracy of the estimates and the efficiency of the calculation.

Acknowledgments

We thank the anonymous reviewers for their helpful questions and suggestions, and the SAMPL participants and organizers for making the data available. This research was funded by NIH grant R01 GM069702.

References

(1) Levy, R.; Zhang, L.; Gallicchio, E.; Felts, A. On the Nonpolar Hydration Free Energy of Proteins: Surface Area and Continuum Solvent Models for the Solute - Solvent Interaction Energy. *J. AM. CHEM. SOC* **2003**, *125*, 9523–9530.

(2) Robinson, G.; Cho, C. Role of Hydration Water in Protein Unfolding. *Biophys J* **1999**, *77*, 3311–3318.

(3) Rakhmanov, S.; Makeev, V. Atomic hydration potentials using a Monte Carlo Reference State (MCRS) for protein solvation modeling. *BMC Struct Biol.* **2007**, 7–19.

(4) Jorgensen, W.; Ulmschneider, J.; Tirado-Rives, J. Free Energies of Hydration from a

Generalized Born Model and an All-Atom Force Field. *J. Phys. Chem. B* **2004**, *108*, 16264–16270.

(5) Cui, Q.; Smith, V. 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,* **2002**, *106*, 6554–6565.

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

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

(8) Yang, C.; Sun, H.; Chen, J.; Nikolovska-Coleska, Z.; Wang, S. Importance of ligand reorganization free energy in protein-ligand binding-affinity prediction. *J Am Chem Soc.* **2009**, *131*, 13709–13721.

(9) Whalen, K.; Spies, 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.* **2013**, *53*, 2349–2359.

(10) Mobley, D.; Dill, K. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009**, *17*, 489–498.

(11) Eisenberg, D.; McLachlan, A. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199–203.

(12) Kang, Y.; Nemethy, G.; Scheraga, H. Free energies of hydration of solute molecules. 1. Improvement of the hydration shell model by exact computations of overlapping volumes. *J. Phys. Chem* **1987**, *91*, 4105–4109.

(13) Kang, Y.; Nemethy, G.; Scheraga, H. Free energies of hydration of solute molecules. 2. Application of the hydration shell model to nonionic organic molecules. *J. Phys. Chem* **1987**, *91*, 4109–4117.

(14) Kang, Y.; Nemethy, G.; Scheraga, H. Free energies of hydration of solute molecules. 3. Application of the hydration shell model to charged organic molecules. *J. Phys. Chem* **1987**, *91*, 4118–4120.

(15) Kang, Y.; Gibson, D.; Nemethy, G.; Scheraga, H. Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell mode. *J. Phys. Chem* **1988**, *92*, 4739–4742.

(16) Tan, C.; Yang, L.; Lou, R. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *J. Phys. Chem.* **2006**, *110*, 18680–18687.

(17) Gallicchio, E.; Zhang, L.; Levy, R. 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.* **2002**, *23*, 517–529.

(18) Thomas, D.; Chun, J.; Chen, Z.; Wei, G.; Baker, N. Parameterization of a Geometric Flow Implicit Solvation Model. *Journal of Computational Chemistry* **2013**, *34*, 687–695.

(19) Lei, H.; Yang, X.; Zheng, B.; Baker, N. Constructing Surrogate Models of Complex Systems with Enhanced Sparsity: Quantifying the influence of conformational uncertainty in biomolecular solvation. *Multiscale Modeling and Simulation* **2015**, *13*, 1327–1353.

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

(21) Mobley, D.; Bayly, C.; Cooper, M.; Dill, K. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J Phys Chem B.* **2009**, *113*, 4533–4537.

(22) Klimovich, P.; Mobley, D. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J Comput Aided Mol Des.* **2010**, *24*, 307–316.

(23) Geballe, M.; Guthrie, J. The SAMPL3 blind prediction challenge: transfer energy overview. *J Comput Aided Mol Des.* **2012**, *26*, 489–496.

(24) Geballe, M.; Skillman, A.; Nicholls, A.; Guthrie, J.; Taylor, P. The SAMPL2 blind prediction challenge: introduction and overview. *J Comput Aided Mol Des* **2010**, *24*, 259–279.

(25) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput. Aided Mol. Des.* **2014**, *3*, 135–150.

(26) Ellingson, B.; Gaballe, M.; Wlodek, S.; Baylyl, C.; Skillman, A.; Nicholls, A. Efficient calculation of SAMPL4 hydration free energies using OMEGA, SZYBKI, QUACPAC, and Zap TK. *J. Comput. Aided Mol. Des.* **2014**, *28*, 289–298.

(27) Muddana, H.; Sapra, N.; Fenley, A.; Gilson, M. The SAMPL4 hydration challenge: Evaluation of partial charge sets with explicit-water molecular dynamics simulations. *J Comput Aided Mol Des* **2014**, *28*, 277–287.

(28) Fu, J.; Liu, Y.; Wu, J. Fast prediction of hydration free energies for SAMPL4 blind test from a classical density functional theory. *J Comput Aided Mol Des* **2014**, *28*, 299–304.

(29) Levy, R.; Gallicchio, E. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu Rev Phys Chem* **1998**, *49*, 531–567.

(30) Mobley, D.; Bayly, C.; Cooper, M.; Shirts, M.; Dill, K. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J Chem Theory Comput.* **2009**, *5*, 350–358.

(31) Mennucci, B., Camm, R., Eds. *Continuum solvation models in chemical physics: from theory to applications*; John Wiley & Sons, 2007.

(32) Konig, G.; Pickard, C.; Mei, Y.; Brooks, 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* **2014**, *28*, 245–57.

(33) Kamerlin, S.; Haranczyk, M.; Warshel, A. Are Mixed Explicit/Implicit Solvation Models Reliable for Studying Phosphate Hydrolysis? A Comparative Study of Continuum, Explicit and Mixed Solvation Models. *Chemphysche* **2009**, *10*, 1125–1134.

(34) Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. A universal surface area correlation for molecular hydrophobic phenomena. *J. Am. Chem. Soc.* **1999**, *39*, 9243–9244.

(35) Chorny, I.; Dill, K. A.; Jacobson, M. P. A universal surface area correlation for molecular hydrophobic phenomena. *J. Phys. Chem. B.* **2005**, *50*, 24056–24060.

(36) Mobley, D. L.; Barber, A. E.; Fennell, C. J.; Dill, K. A. Charge Asymmetries in Hydration of Polar Solutes. *J. Phys. Chem. B.* **2008**, *8*, 2405–2414.

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

(38) Apostolakis, G. The concept of probability in safety assessments of technological systems. *Science* **1990**, *250*, 1359–1364.

(39) Devooght, J. Model uncertainty and model inaccuracy. *Reliability Engineering and System Safety* **1998**, *59*, 171–185.

(40) Neuman, S.; Wierenga, P. A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites. 2003.

(41) Bates, J. M.; Granger, C. W. J. The combination of forecasts. *Operational Research Quarterly* **1969**, *20*, 451–468.

(42) Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research* **1999**, *11*, 169–198.

(43) Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39.

(44) Hoeting, J.; Madigan, D.; Raftery, A.; Volinsky, C. Bayesian model averaging: a tutorial. *Statistical Science* **1999**, *14*, 382–417.

(45) Seni, G.; Elder, J. F. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery* **2010**, *2*, 1–126.

(46) Raftery, A.; Madigan, D.; Hoeting, J. Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **1998**, *92*, 179–191.

(47) Raftery, A. Bayesian Model Selection in Social Research. *Sociological Methodology* **1995**, *25*, 111–163.

(48) Gosink, L.; Hogan, E.; Pulsipher, T.; Baker, N. Bayesian model aggregation for ensemble-based estimates of protein pKa values. *Proteins* **2014**, *82*, 354–363.

(49) Zhang, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **2003**, *50*, 159–175.

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

(51) Raftery, A.; Gneiting, T.; Balabdaoui, F.; Polakowski, M. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review* **2005**, *133*, 1155–1174.

(52) Vlachopoulou, M.; Gosink, L.; Pulsipher, T.; Ferryman, T.; Zhou, N.; Tong, J. An Ensemble Approach for Forecasting Net Interchange Schedule. IEEE PES General Meeting. 2013; p n. pag.

(53) Ye, M.; Neuman, S. P.; Meyer, P. D. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research* **2004**, *40*, n. pag.

(54) Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*; Wiley: New York, 1989.

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

(56) Mallows, C. L. Some comments on C_p. *Technometrics* **1973**, *15*, 661–675.

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

(58) Genell, A.; Nemes, S.; Steineck, G.; Dickman, P. W. Model selection in Medical Research: A simulation study comparing Bayesian Model Averaging and Stepwise Regression. *BMC Medical Research Methodology* **2010**, *10*, 108.

(59) Davidson, I.; Fan, W. *Knowledge Discovery in Databases: PKDD 2006*; Lecture Notes in Comput. Science; Springer Berlin Heidelberg, 2006; Vol. 4213; pp 478–486.

(60) Quan, C.; Yu, Y.; Zhou, Z. Pareto Ensemble Pruning. Conference on Artificial Intellegince. 2015; pp 2935– 2941.

(61) Martinez-Munoz, G.; Hernandez-Lobato, D.; Suarez, A. An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation. *Trans. Pattern Analysis and Machine Intelligence* **2009**, *31*, 245–259.

(62) Onorante, L.; Raftery, A. *Dynamic Model Averaging in Large Model Spaces Using Dynamic Occam's Window*; 2014.

(63) Madigan, D.; Raftery, A. Model selection and accounting for model uncertainty in graphical models using Occam's Window. *J. Am. Stat. Assoc.* **1994**, *89*, 1335–1346.

(64) Jeffreys, H. *Theory of Probability*; Oxford: Oxford University Press, 1961.

(65) Guthrie, J. SAMPL4, a blind challenge for computational solvation free energies: the compounds considered. *J Comput Aided Mol Des.* **2014**, *28*, 151–168.

(66) Clyde, M. Bayesian Model Averaging and Model Search Strategies. Bayesian Statistics. 1999; pp 157–185.

(67) Nicholls, A.; Wlodek, S.; Grant, J. SAMPL2 and continuum modeling. *J Comput Aided Mol Des.* **2010; 24(4):293–306.**, *24*, 293–306.

(68) Hawkins, G.; Giesen, D.; Lynch, G.; Chambers, C.; Rossi, I.; Storer, J.; Li, J.; Zhu, T.; Thompson, J.; Winget, P. et al. AMSOL.

(69) Fixman, M. The Poisson–Boltzmann equation and its application to polyelectrolytes. *J. Chem. Phys.* **1979**, *70*, 4995–146.

(70) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144–1149.

(71) Davis, M.; McCammon, A. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.* **1990**, *90*, 509–521.

(72) Sandberg, L. Predicting hydration free energies with chemical accuracy: The SAMPL4 challenge. *J Comput Aided Mol Des.* **2013**, *28*, 211–9.

(73) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of the Free Energy of Hydration of a Challenging Set of Pesticide-Like Compounds. *J Phys Chem B.* **2009**, *113*, 4508–4510.

(74) Hogues, H.; Sulea, T.; Purisima, E. Exhaustive docking and solvated interaction energy scoring: Lessons learned from the SAMPL4 challenge. *J Comput Aided Mol Des.* **2014**, *28*, 417–27.

(75) Sulea, T.; Purisima, 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.* **2011**, *26*, 661–667.

(76) Reinisch, J.; Klamt, A. Prediction of free energies of hydration with COSMO-RS on the SAMPL4 data set. *J Comput Aided Mol Des* **2014**, *28*, 169–73.

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

(78) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. Development and testing of a general AMBER force field. *J Comput Chem.* **2004**, *25*, 1157–1174.

(79) Li, L.; Dill, K.; Fennell, C. Testing the semi-explicit assembly model of aqueous solvation in the SAMPL4 challenge. *J Comput Aided Mol Des.* **2014**, *28*, 259–264.

(80) Coleman, R.; Sterling, T.; Weiss, D. SAMPL4 & DOCK3.7: lessons for automated docking procedures. *J. Comput. Aided Mol. Des.* **2014**, *28*, 201–209.

(81) Yang, Q.; Sharp, K. Atomic Charge Parameters for the Finite Difference Poisson-Boltzmann Method Using Electronegativity Neutralization. *J. Chem. Theory Comput.* **2006**, *4*, 1152–1167.

(82) Genheden, S.; Cabedo, M.; Criddle, M.; Essex, J. Extensive all-atom Monte Carlo sampling and QM/MM corrections in the SAMPL4 hydration free energy challenge. *J. Comput. Aided Mol. Des.* **2014**, *28*, 187–200.

(83) Beckstein, O.; Iorga, 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* **2012**, *26*, 635–645.

(84) Beckstein, O.; Fourrier, A.; Iorga, B. Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the OPLS-AA force field. *J. Comput. Aided Mol. Des.* **2014**, *28*, 265–276.

(85) Jambeck, J.; Mocci, F.; Lyubartsev, A.; Laaksonen, A. Partial atomic charges and their impact on the free energy of solvation. *J. Comput. Chem.* **2013**, *34*, 187–197.

(86) Park, H. Extended solvent-contact model approach to SAMPL4 blind prediction challenge for hydration free energies. *J. Comput. Aided Mol. Des.* **2014**, *28*, 175–186.

(87) Klamt, A.; Diedenhofen, M. Blind prediction test of free energies of hydration with COSMO-RS. *J. Comput. Aided Mol. Des.* **2010**, *24*, 357–360.

(88) Lucas, G. Star Wars. Episode IV. A New Hope. 1977.

(89) Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1945**, *1*, 80–83.

(90) Shapiro, S.; WIlk, M. An analysis of variance test for normality. *Biometrika* **1965**, *52*, 591–611.

(91) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(92) Hoerl, A.; Kennard, R. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **2000**, *42*, 80–86.

(93) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. Royal Stats. Soc.* **1994**, *58*, 267–288.

(94) R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2008.

(95) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J. et al. API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013; pp 108–122.