Physics-based Machine Learning to Predict Hydration Free Energies for Small Molecules with a minimal number of descriptors: Interpretable and Accurate

Ajeet Kumar Yadav [†], Marvin V. Prakash [†], and Pradipta Bandyopadhyay[*]

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi 110067, INDIA

E-mail: praban07@gmail.com

[†]These authors contributed equally

Abstract

Hydration free energy (HFE) of molecules is a fundamental property having importance throughout chemistry and biology. Calculation of the HFE can be challenging and expensive with classical molecular dynamics simulation-based approaches. Machine learning (ML) models are increasingly being used to predict HFE. Although the accuracy of ML models for datasets for small molecules is impressive, these models suffer from lack of interpretability. In this work, we have developed a physics-based ML model with only six descriptors, which is both accurate and fully interpretable, and applied it to a database for small molecule HFE, *FreeSolv*. We have evaluated the electrostatic energy by an approximate closed form of the Generalized Born (GB) model and polar surface area. In addition, we have $logP$ and hydrogen bond acceptor

and donors as descriptors along with the number of rotatable bonds. We have used different ML models such as random forest and extreme gradient boosting. The best result from these models has a mean absolute error of only 0.74 kcal/mol. The main power of this model is that the descriptors have clear physical meaning and it was found that the descriptor describing the electrostatics and the polar surface area, followed by the hydrogen bond donors and acceptors, are the most important factors for the calculation of hydration free energy.

Introduction

Solvation free energy is one of the key quantities in chemistry and biology as most of the molecular phenomena occur in different solvents.[1,2] Water being the most versatile solvent, determination or calculation of hydration free energy (HFE) is the most important step in understanding any complex process. The calculation of HFE is typically done using either a quantum mechanical description of the solute in a dielectric continuum[3–8] or a classical description of the solute where water is treated either with explicit models[9–14] or implicit models (which are mostly dielectric continuum).[10,15–17] The classical forcefields are usually used for macromolecules, and for small molecules interacting with macromolecules, classical forcefields are used for compatibility. Henceforth, all discussions on the calculation of hydration free energy will be with the classical models. Although the physics-based methods are well developed, there are outstanding issues in getting accurate solvation free energy and it has been an active area of new developments and improvements.[11,18–20] The most rigorous calculations are alchemical methods like thermodynamic integration[21,22] and free energy perturbation.[23] However, these calculations are time-consuming and generally speaking, are not suitable for a large set of molecules.

Continuum solvent models are often preferred methods when dealing with a larger number of small molecules, typical in a drug design project. The Poisson-Boltzmann (PB) and Generalized Born (GB) are the two most common models used in continuum solvent models.

In the PB approach, PB eq. (or Poisson eq. in the absence of any salt concentration) is solved by defining an interior dielectric constant for the solute and an external dielectric constant for the solvent. The Generalized Born approach also defines internal and external dielectric constants; however, here no electrostatic equation is solved, rather an expression, obtained from the generalization of the Born equation for a single ion, is evaluated.[15,24,25] In both PB and GB approaches, the molecular surface area is calculated and in GB, the so-called Born radii is calculated, which can be time-consuming. Also, there are some inherent limitations for continuum models as they neglect the molecular nature of water. There have been several attempts to build cluster-continuum models.[26,27]

From an entirely different perspective, several machine learning (ML) models are developed, in the last couple of years,[28–39] to predict solvation free energy using experimental data in the *FreeSolv* database.[9] The faster speed of ML models compared to physics-based models is advantageous and can be used for large databases of small molecules used in drug discoveries. However, ML models often suffer from a lack of interpretability and the reasons why they work (or do not work) are often not clear. There have been attempts to define descriptors having clear physical meaning. For instance, Zhang et al. used electron density (obtained from quantum mechanical calculations) based descriptors.[37] In some of the other representative works, Alibakhshi et al. have combined ML models with PCM model to predict solvation free energy in different solvents using the components of the PCM calculations as the features of the ML model,[32] Pattnaik et al. have developed an ML model to predict relative solvation free energy in forty-one solvents.[38] Vyboishchikov has developed a few NN-based models based on the GB model of solvation.[39–41] The effective Born radii, charges are used as the features in these models. Machine learning has also been used to predict the HFE obtained from MD simulation-based methods.[30]

In the current work, our motivation is to use a minimal number of physically interpretable and simple descriptors for predicting HFE. Starting from the GB expression and with an approximate analytical calculation of Born radii,[25] we evaluate the HFE (after adding five

more descriptors) and got accuracy almost as good as the paper by Zhang et al.[37] for the *FreeSolv* dataset. The power of our method is that it is completely physics-based, hence fully interpretable. It has only six descriptors, an electrostatics term (GB term summed with Coulomb electrostatic), polar surface area, number of donor and acceptor atoms, $logP$, and the number of rotatable bonds. We have used four different models, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GradBoost), and Light Gradient Boosting Machine (LightGBM). Our best result is a mean absolute error (MAE) of 0.74 kcal/mol comparable to the work of Zhang et al.[37] This method can be used for large datasets used in drug designing and the reason for specific HFE values can be understood clearly as opposed to most of the ML models.

Methodology

Database Description

The experimental hydration free energy database, *FreeSolv*, prepared by Mobley et al.,[9] has been widely used and benchmarked by various physical solvation models as well as machine learning (ML) and deep learning models. The *FreeSolv* database has 643 small organic molecules with their experimental HFE and their SMILES (simplified molecular-input line-entry system). The database also includes the calculated HFE, enthalpy, and entropy data from explicit molecular dynamics simulations. These calculations utilized the GAFF force field,[42] AM1-BCC partial charges,[43,44] and the TIP3P water model.[45] The experimental HFE values have the mean and standard deviation as -3.82 and 4.84 kcal/mol, respectively. We have divided the total dataset into nine different groups based on the functional group or presence of a specific atom in the molecule. The eight groups are *Alkanol, Alkanone, Alkene, Alkyl Alkanoate, Halo Alkane, Aromatic, Aliphatic cyclic, N-based Aliphatic,* and the ninth, *misc*, is the group for molecules that do not come under any of the previous eight groups. We have assessed the performances of our models both for the whole dataset and these different

groups.

Descriptor Generation

One of the primary objectives of this work is to utilize a minimal number of descriptors while ensuring they possess physical interpretability. To achieve this, we have used only six descriptors: polar surface area, hydrogen bond donors, hydrogen bond acceptors, the number of rotatable bonds, $logP$, and an electrostatic term which we call as the *pol term* (GB term summed with Coulomb electrostatic). The first five descriptors were calculated using the RDKit[46] package in Python, while the last descriptor is calculated as described below.[25]

The simplified polar energy is the sum of two terms, the Coulombic electrostatic energy, and a generalized Born energy term. The generalized Born energy term is calculated by the Generalized Born equation as follows

$$\Delta G_{pol} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon_w}\right)\sum_{i<j}\frac{q_i q_j}{f_{GB}} \tag{1}$$

where ϵ_w is the dielectric constant of water (the process being moving a solute from vacuum to water), q_i, and q_j are the charges of atoms i and j. And f_{GB} is a function, dependent on the distance between the atoms i and j, that interpolates between the distance r_{ij} and the Born radii. Most widely used functional form of f_{GB} is given below

$$f_{GB} = \left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}}{4R_i R_j}\right)\right]^{\frac{1}{2}} \tag{2}$$

where R_i and R_j are the effective Born radii of atoms i and j. The main challenge in the implementation of the GB model is the calculation of the effective Born radius.

Approximate analytical evaluation of Born Radius

When the accessible surface of the atoms of a molecule are non-overlapping, then it can be shown that the following (eq. 3) is an analytical expression for Born radius, a_i being the sum of the radii of the atom i and water.[25]

$$R_i^{-1} = a_i^{-1} - \sum_j \frac{a_j}{2\left(r_{ij}^2 - a_j^2\right)} - \frac{1}{4r_{ij}} \log\left(\frac{r_{ij} - a_j}{r_{ij} + a_j}\right) \tag{3}$$

The charge and radius of atoms are taken from the Generalized Amber ForceField (GAFF) forcefield[42] (water radius was taken as 1.4 Angstrom). Although, eq. (3) is valid only for non-overlapping atoms, this can act as an excellent descriptor in an ML model. In the evaluation of eq. (3), the numerator of the second term can be negative for overlapping accessible surfaces of the atoms. To circumvent this problem, we have performed the sum over the pairs of atoms with non-overlapping accessible surfaces only. This approximation provides a fast estimation of the polar part of the solvation free energy. Our results show that using this approximation as a descriptor, ML-based methods perform well for the *FreeSolv* database.

Machine Learning Models

Figure 1 illustrates the workflow for predicting the hydration-free energy. It highlights that the six descriptors are calculated for the *FreeSolv* database first. Then different ML methods are trained and HFE is predicted. Regression algorithms will enable ML models to make predictions based on the information represented by each chemical feature. After calculating RDKit descriptors, the database was divided into two subsets. We used an 80:20 split to divide the dataset into training and testing sets, ensuring that this ratio was consistently maintained across the nine predefined groups. These subsets were utilized to develop, train, and statistically evaluate the model using different ML algorithms. We applied `StandardScaler` to standardize the features, which helps improve model performance

by scaling data to have a mean of zero and a standard deviation of one. This ensures that all features contribute equally to the model training process and in improving convergence during optimization. After these preprocessing steps, different machine learning models are trained on the training set to learn the crucial relationships for making predictions and then tested on the unseen testing set to assess prediction accuracy.

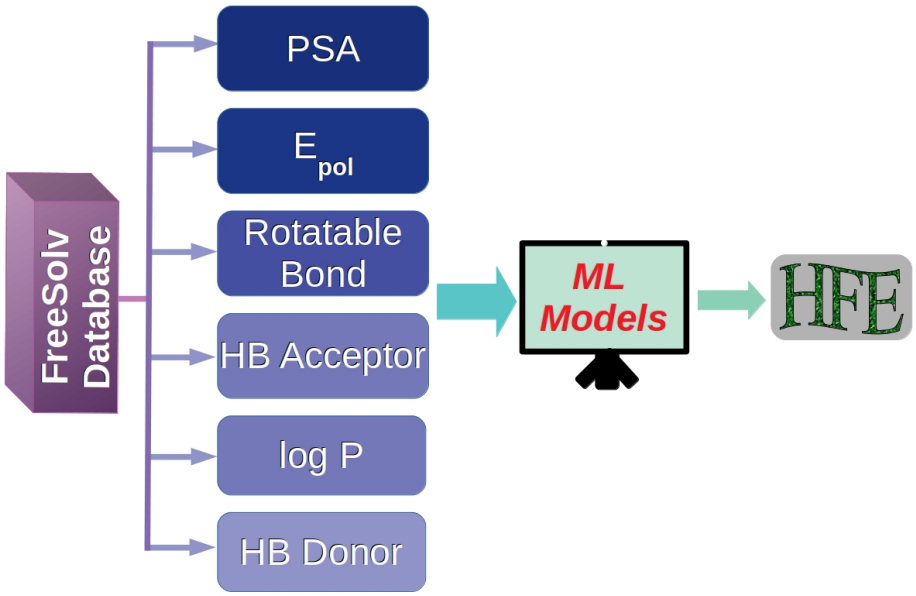

Figure 1: Workflow to predict ΔG_{Hyd} using ML-based models. It involves the calculation of descriptors, ML model training and then predicting HFE.

We have employed four different machine learning models: Random Forest (RF),[47] Extreme Gradient Boosting (XGBoost),[48] Gradient Boosting (GradBoost),[49] and Light Gradient Boosting Machine (LightGBM).[50] These models are trained on the training set to learn the crucial relationships of the descriptors with the HFE which is the target property. Although all the above four machine learning models— RF, XGBoost, GradBoost, and LightGBM — use ensemble techniques, their approaches to prediction differ. Random Forest is a bagging-based technique that builds multiple decision trees independently, averaging their predictions to reduce variance and improve stability. Gradient Boosting applies a boosting strategy to further train the weak learners one after another in a sequence to minimize the loss function while correcting the mistakes of the preceding one. The other two methods –

XGBoost, and LightGBM – are more advanced versions of the Gradient Boosting algorithm. XGBoost uses regularization, parallel processing, and tree-trimming methods to overcome the over-fitting. LightGBM, another variant of Gradient Boosting, was developed to handle large amounts of data, it uses leaf-wise growth and histogram-based learning to speed up and reduce memory usage. Collectively, these models use various techniques to merge decision trees to balance the prediction accuracy and computing time. We have trained and tested our ML models on two classes of data: 1) full data set, and 2) without outliers. We have used the interquartile range (IQR) method to define outliers in the experimental hydration-free energy, with bounds set at $Q_1 - 1.5 * IQR$ and $Q_3 + 1.5 * IQR$. $Q1$ and $Q3$ are the first and third quartiles, respectively. This leaves 628 number of molecules in the second dataset set.

We have also utilized `GridSearchCV` with 5-fold cross-validation to optimize the hyperparameters of our model. This method involves systematically searching through a grid of hyperparameter values to identify the best settings for our model. In conjunction with 5-fold cross-validation, the dataset is divided into 5 folds. The model is trained on 4 folds for each hyperparameter combination and evaluated on the remaining fold. This process is repeated 5 times, each time using a different fold as the test set. We ensure that the selected hyperparameters provide robust and generalizable model performance by averaging the performance across these iterations. In Table S1 in SI, we have listed the optimized parameters of the four models we have used.

To evaluate the performance of our model, we employed several metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Pearson correlation coefficient (Pr), and R^2 score. RMSE and MAE provide insights into the magnitude of prediction errors, while Pr assesses the strength of the linear relationship between observed and predicted values, and R^2 indicates the proportion of variance explained by the model. These metrics were used to evaluate our model's accuracy and robustness rigorously, and the results were compared with those from other studies to benchmark our model's performance against existing methods. The feature importance for each descriptor was calculated using the mean

decrease of impurity.

Results and discussion

Performance of the simplified GB model

First, we have assessed the performance of the approximated GB model alone in figure 2. The figure shows that the model has relatively high values of RMSE and MAE indicating that this model alone is not accurate enough and needs further refinement.

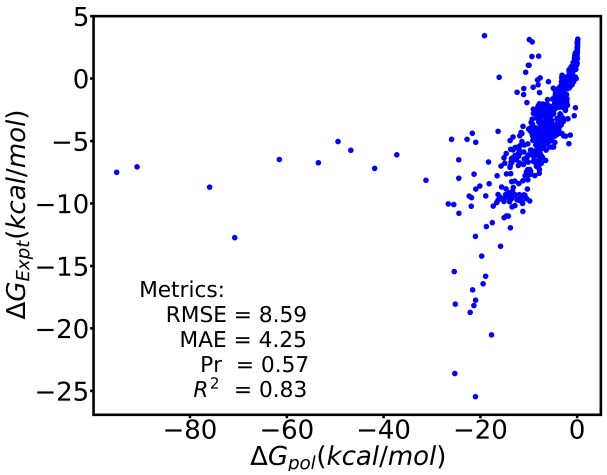

Figure 2: The scatter plot compares the experimental hydration free energy with predicted values.

Machine Learning-based Model Performance

To select the appropriate model for our study, we have compared the hydration-free energy prediction performance of the Random Forest (RF) and XGBoost models in Figure 3. Performances of the other two models are shown in Figure S1 in Supporting Information (SI). For the RF model, the test set Root Mean Squared Error (RMSE) is 1.30 kcal/mol, and the coefficient of determination (R^2) is 0.89, indicating that the model explains approximately 89% of the variance in the test set. The Pearson correlation coefficient (Pr) is 0.94, demonstrating a strong linear correlation between predicted and experimental HFE values.

The Mean Absolute Error (MAE) of 0.83 kcal/mol reflects that the model provides accurate predictions overall.

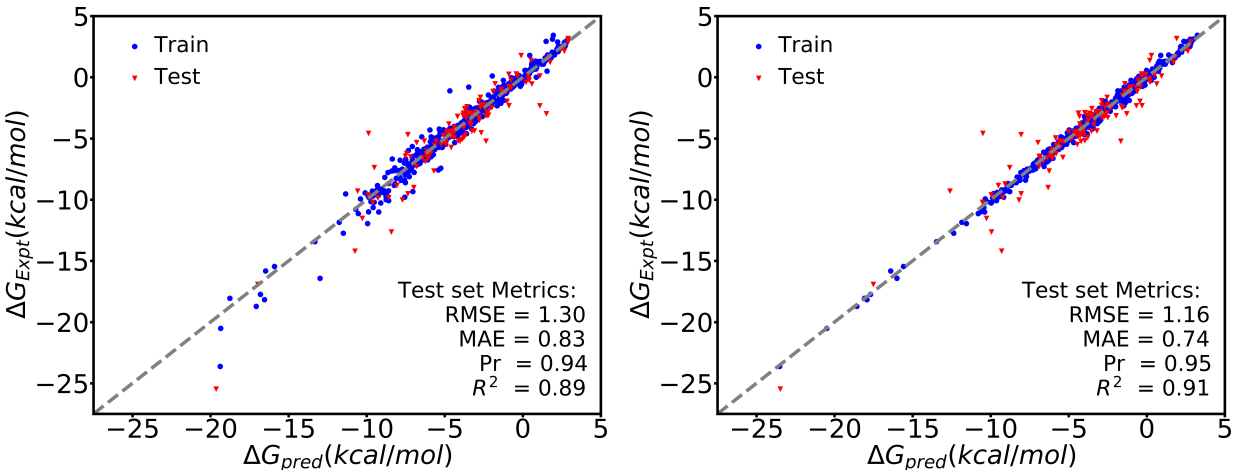

Figure 3: Comparison of hydration free energy predictions using Random Forest (left) and Extreme Gradient Boosting (right) models for the full data set.

In comparison, the XGBoost model outperforms Random Forest, with a lower RMSE of 1.16 kcal/mol and a higher R^2 value of 0.91, explaining about 91% of the variance in the test set. The Pearson correlation coefficient for XGBoost is $Pr = 0.95$, indicating a very strong linear relationship between predicted and experimental values. The MAE of 0.74 kcal/mol confirms XGBoost's improved predictive accuracy and precision compared to Random Forest. Both models exhibit strong agreement between predicted and experimental values, with data points clustered along the diagonal line in the parity plots. However, XGBoost shows a more concentrated distribution, particularly at lower ΔG values, suggesting it provides a better fit overall. The Gradient Boosting model and LightGBM, shown in Figure S1 in SI, perform comparably to the Random Forest and XGBoost models. For Gradient Boosting, with an RMSE of 1.28 kcal/mol and $R^2 = 0.89$, it explains about 89% of the variance in the test set, similar to Random Forest. The Pearson correlation coefficient of $Pr = 0.95$ indicates a strong linear relationship between predicted and experimental values. The MAE of 0.81 kcal/mol highlights its reliable predictive performance. On the other hand, LightGBM has an MAE of 0.84 kcal/mol. The analysis demonstrates that XGBoost and Gradient Boosting outperform

Random Forest and LightGBM, with XGBoost offering the most accurate predictions overall.

All the models display slight deviations indicating potential areas for further improvement. To overcome the deviations, we retrained the models to assess their performance without outliers. We have compared the hydration free energy prediction performance of the RF and XGBoost models without outliers in Figure 4. Performances of the other two models are shown in Figure S2 in SI. The Random Forest model had an RMSE of 1.27 kcal/mol, an MAE of 0.75 kcal/mol, an R^2 of 0.80, and a Pearson correlation coefficient of 0.90. Despite the removal of outliers, the model's performance slightly decreased compared to the original dataset, especially in terms of the correlations i.e. R^2 value and Pr. While, the XGBoost model maintained strong predictive performance without outliers, with an improved RMSE of 1.16 kcal/mol, an R^2 value of 0.83, and a Pearson correlation coefficient of 0.92. The MAE for XGBoost decreased to 0.72 kcal/mol, reflecting only a slight increase in accuracy compared to its performance on the full dataset. The LightGBM model's performance was comparable to Random Forest, with an RMSE of 1.27 kcal/mol, an MAE of 0.78 kcal/mol, an R^2 value of 0.80, and a Pearson correlation coefficient of $Pr = 0.90$. In contrast, the Gradient Boosting model performed significantly better without outliers, achieving an RMSE of 1.11 kcal/mol and an R^2 value of 0.85, along with a Pearson correlation coefficient of $Pr = 0.92$. The MAE for this model was 0.70 kcal/mol, indicating improved accuracy compared to the full dataset. While LightGBM showed similar results to Random Forest, it performed slightly worse compared to Gradient Boosting and XGBoost in terms of both RMSE and MAE. The removal of outliers generally led to improved accuracy for most models, making Gradient Boosting the strongest performance on this cleaner dataset. The following table 1 summarizes the metrics for the RF, GradBoost, XGBoost, and LightGBM models with and without outliers.

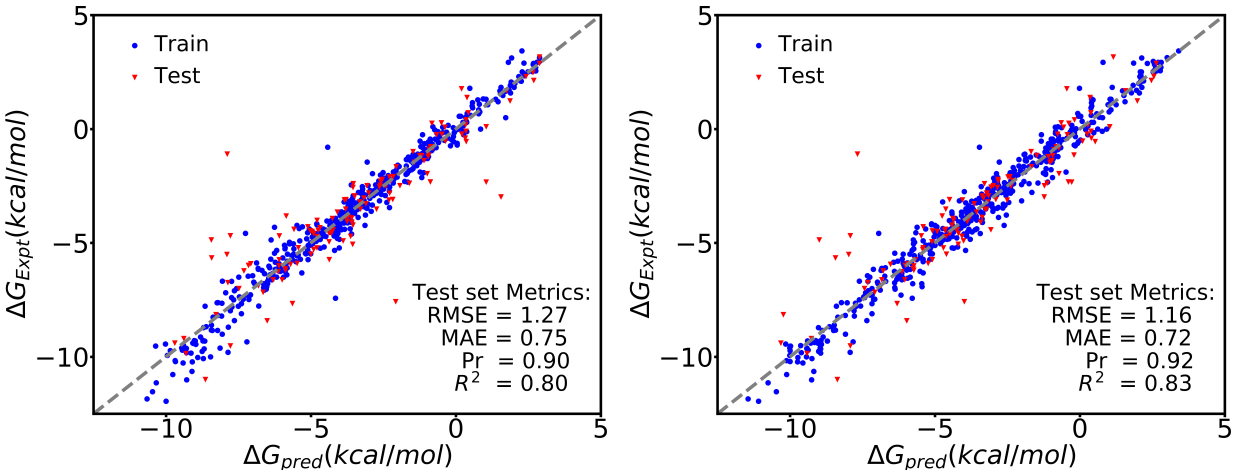

Figure 4: Comparison of hydration free energy predictions Random Forest (left) and Extreme Gradient Boosting (right) models for the data set without outliers.

Table 1: Performance metrics for models with and without outliers (RF: Random Forest, GradBoost: Gradient Boosting, XGBoost: Extreme Gradient Boosting, LGBM: Light Gradient Boosting Machine).

With Outliers				
	RF	**GradBoost**	**XGBoost**	**LightGBM**
RMSE (Train/Test)	0.60 / 1.30	0.51 / 1.28	0.47 / 1.15	0.93 / 1.32
MAE (Train/Test)	0.37 / 0.83	0.38 / 0.81	0.34 / 0.74	0.64 / 0.84
R^2 (Train/Test)	0.98 / 0.89	0.98 / 0.89	0.98 / 0.91	0.94 / 0.88
Rp (Train/Test)	0.99 / 0.94	0.99 / 0.94	0.99 / 0.96	0.97 / 0.94
Without Outliers				
	RF	**GB**	**XGB**	**LGBM**
RMSE (Train/Test)	0.56 / 1.27	0.68 / 1.11	0.56 / 1.16	0.85 / 1.27
MAE (Train/Test)	0.37 / 0.75	0.49 / 0.70	0.41 / 0.72	0.60 / 0.78
R^2 (Train/Test)	0.97 / 0.80	0.96 / 0.85	0.97 / 0.83	0.93 / 0.80
Rp (Train/Test)	0.99 / 0.90	0.98 / 0.92	0.99 / 0.92	0.97 / 0.90

Descriptor performance

The feature importances for the different models highlight the varying roles each descriptor plays in predicting the target variable. We have shown the feature importance for RF and XGBoost in Figure 5 and for Gradient Boosting and LightGBM in Figure S3 in SI. In both the Random Forest and Gradient Boosting models, the polar surface area (psa) emerges

as the most important feature, contributing around 50%, followed by the *pol term*, which accounts for approximately 30%. It is to be noted that polar surface area and non-polar surface area are complementary features. In this work, we have taken PSA; however, taking PSA as a feature implicitly includes non-polar surface area also. Hence, the importance of PSA as a feature indicates the importance of polarity of surface areas in general. These polarity of surface area and the pol term dominate the prediction capabilities of these models, suggesting that molecular surface properties play a crucial role in the prediction task. Other descriptors like the number of hydrogen bond donors (n_donors), rotatable bonds (nrotb), acceptors ($n\_acceptors$), and $logP$ contribute significantly less. This highlights a strong dependence on molecular polarity and surface area in these ensemble tree-based models.

Interestingly, the XGBoost model demonstrates a different feature importance distribution, where the number of acceptors ($n\_acceptors$) becomes the most dominant feature, contributing 30% to the predictions. *Pol term* and psa play smaller but still significant roles, contributing around 18-22%. This indicates that the XGBoost model is more sensitive to hydrogen bond acceptor characteristics compared to the other models. LightGBM also highlights psa, $logP$, and *pol term* as the most critical features. These results suggest that while molecular surface and polarity remain crucial across models, each model places a different emphasis on these features based on their algorithmic structure.

Comparison with Other ML Models used for *FreeSolv* dataset

We have compared our models with previous models trained on the *FreeSolv* database. In comparison to several previous models, such as CIGIN[34] (0.76), MLSolvA[33] (0.76), and MoleculeNet[29] (1.15), our XGB model achieves a lower test MAE (0.74). At the same time, there are models (e.g. the A3D-PNAConv-FT[35] with the MAE of 0.42) having lower MAE than ours. However, essentially all previous models for predicting HFE use complex and a large number of descriptors (and more complex predictors) making the interpretation difficult. Our model stands out from the others while competitive with other models in terms

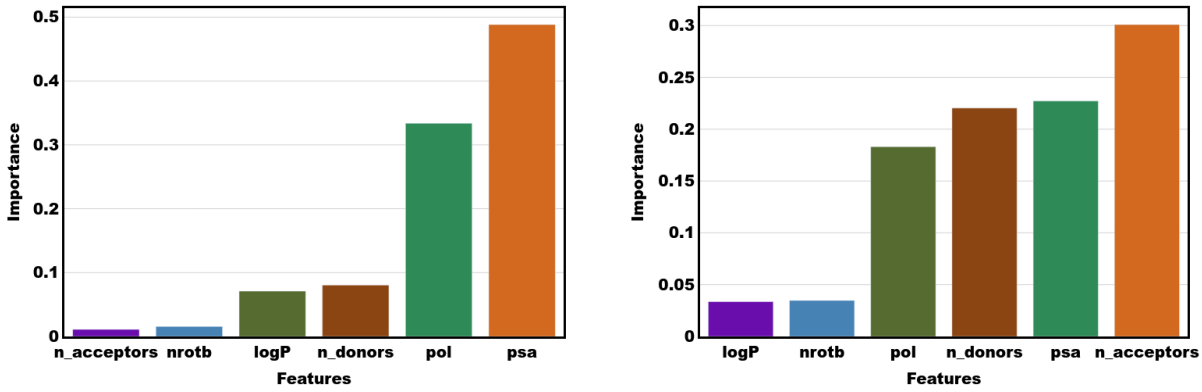

Figure 5: Comparison of feature importances for Random Forest (left) and XGBoost (right) models. The bars represent the relative importance of each feature in predicting the target variable.

of performance, it is completely physics-based and fully interpretable.

Performance against different functional groups

We have assessed the performance of our models on the nine groups defined in the methods section. Figure 6 shows the performance metrics for RF regression model for the nine groups. The RMSE and MAE for the ninth group i.e. misc (molecules not categorized in any of the previous eight groups) show the highest deviation in the prediction with their values of 0.86 and 0.54 kcal/mol respectively. But the correlation metrics i.e. R^2 and Pr show different behaviour than the error metrics. The correlation metrics for this group ($R^2 = 0.9$ and $Pr = 0.95$) indicate that this group's performance closely agrees with experimental hydration free energy. These two contradicting metrics show that there is a systematic error in the model both in the training and testing phases. The same contradicting trend is also observed in the case of *aromatic* group. Except for these two groups, our models perform well across different groups with relatively low RMSE (less than 0.53 kcal/mol) and low MAE (less than 0.36 kcal/mol). For the correlation metrics except for *alkanone* group, all other groups are highly correlated with their corresponding experimental hydration free energy. The R^2 is

always more than 0.85 and Pr is always greater than 0.95 except for *alkanone* group which signifies the performance of our model across the groups.

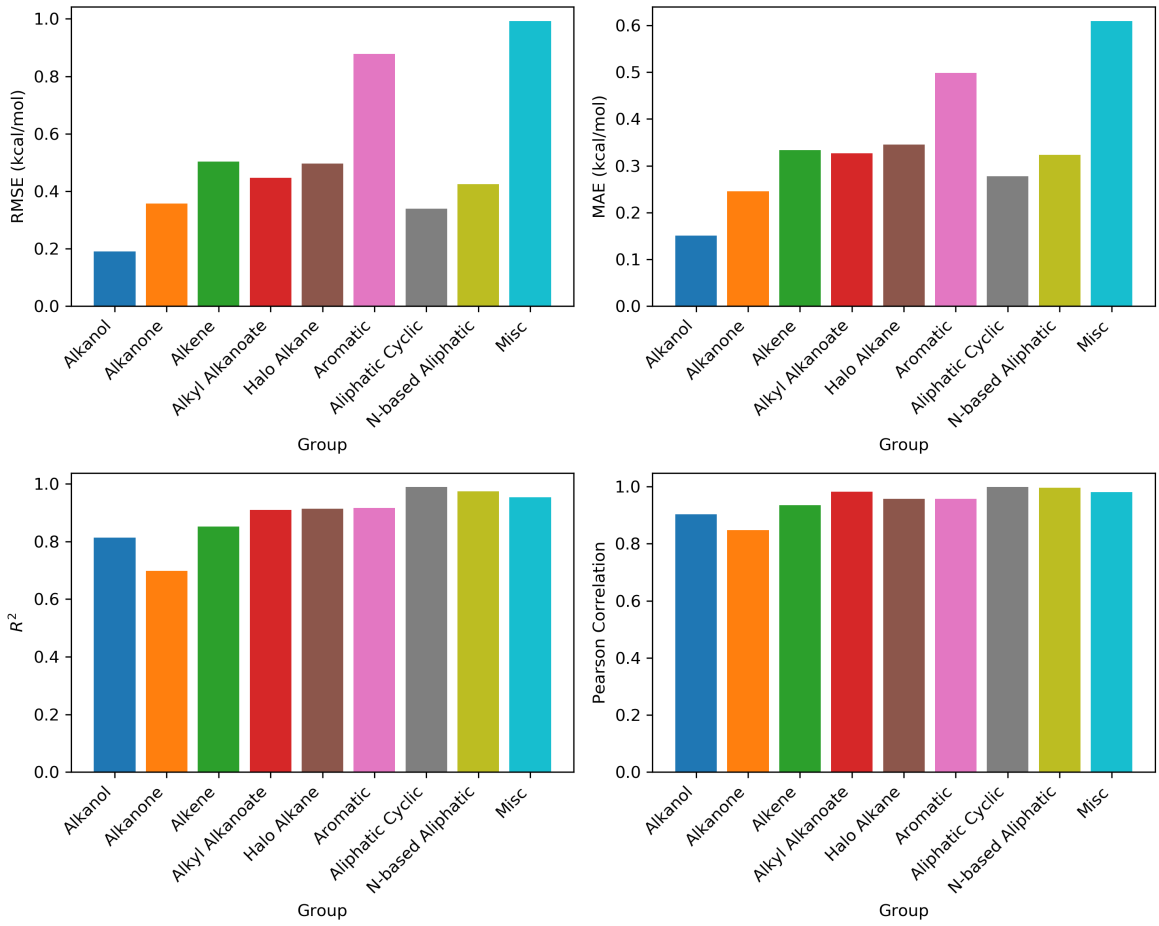

Figure 6: Group-wise performance

Conclusions

In this work, we have developed a physics-based and interpretable machine learning model for predicting hydration free energy of small molecules with only six descriptors. Our results compare well with other works with this dataset. However, the advantage of our method is that the results are fully interpretable, which is often the issue with the ML models. Our models perform well across different chemical groups signifying their applicability to larger databases such as those used in drug discoveries.

Acknowledgement

This work is supported by a MATRICS grant from SERB (MTR/2021/000365) awarded to P.B. This work was also partially supported by grants from the DBT (BT/PR/40251/BITS/137/11/2021) awarded to the Centre for Computational Biology and Bioinformatics, Jawaharlal Nehru University and by the Indo-Slovenia bilateral research grant from DST (DST/ICD/Indo-Slovenia/2022/02(G)). The authors thank Prof. Tomaz Urbic for insightful discussions.

Data Availability

The codes used in this work are available from the corresponding author.

Supporting Information Available

Additional plots comparing hydration free energy using Gradient Boosting and Light Gradient Boosting methods are given in Supporting Information.

References

(1) Brini, E.; Fennell, C. J.; Fernandez-Serra, M.; Hribar-Lee, B.; Luksic, M.; Dill, K. A. How water's properties are encoded in its molecular structure and energies. *Chemical reviews* **2017**, *117*, 12385–12414.

(2) Perlovich, G. L. Thermodynamic approaches to the challenges of solubility in drug discovery and development. *Molecular Pharmaceutics* **2014**, *11*, 1–11.

(3) Mennucci, B. Polarizable continuum model. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *2*, 386–404.

(4) Klamt, A. The COSMO and COSMO-RS solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 699–709.

(5) Mennucci, B.; Tomasi, J. Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. *The Journal of chemical physics* **1997**, *106*, 5151–5158.

(6) Luukkonen, S.; Belloni, L.; Borgis, D.; Levesque, M. Predicting hydration free energies of the FreeSolv database of drug-like molecules with molecular density functional theory. *Journal of Chemical Information and Modeling* **2020**, *60*, 3558–3565.

(7) Voityuk, A. A.; Vyboishchikov, S. F. A simple COSMO-based method for calculation of hydration energies of neutral molecules. *Physical Chemistry Chemical Physics* **2019**, *21*, 18706–18713.

(8) Kriz, K.; Rezac, J. Reparametrization of the COSMO solvent model for semiempirical methods PM6 and PM7. *Journal of Chemical Information and Modeling* **2019**, *59*, 229–235.

(9) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* **2014**, *28*, 711–720.

(10) Shivakumar, D.; Deng, Y.; Roux, B. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *Journal of Chemical Theory and Computation* **2009**, *5*, 919–930.

(11) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the prediction of absolute solvation free energies using the next generation OPLS force field. *Journal of chemical theory and computation* **2012**, *8*, 2553–2558.

(12) Nerenberg, P. S.; Jo, B.; So, C.; Tripathy, A.; Head-Gordon, T. Optimizing solute–water van der Waals interactions to reproduce solvation free energies. *The Journal of Physical Chemistry B* **2012**, *116*, 4524–4534.

(13) Riquelme, M.; Lara, A.; Mobley, D. L.; Verstraelen, T.; Matamala, A. R.; Vohringer-Martinez, E. Hydration free energies in the FreeSolv database calculated with polarized iterative Hirshfeld charges. *Journal of chemical information and modeling* **2018**, *58*, 1779–1797.

(14) Heyden, M. Disassembling solvation free energies into local contributions—Toward a microscopic understanding of solvation processes. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2019**, *9*, e1390.

(15) Onufriev, A. V.; Case, D. A. Generalized Born implicit solvent models for biomolecules. *Annual review of biophysics* **2019**, *48*, 275–296.

(16) Tan, C.; Yang, L.; Luo, R. How well does Poisson- Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *The Journal of Physical Chemistry B* **2006**, *110*, 18680–18687.

(17) Aguilar, B.; Onufriev, A. V. Efficient computation of the total solvation energy of small molecules via the R6 generalized Born model. *Journal of chemical theory and computation* **2012**, *8*, 2404–2411.

(18) Lang, E. J.; Baker, E. G.; Woolfson, D. N.; Mulholland, A. J. Generalized Born implicit solvent models do not reproduce secondary structures of de novo designed Glu/Lys peptides. *Journal of chemical theory and computation* **2022**, *18*, 4070–4076.

(19) Bass, L.; Elder, L. H.; Folescu, D. E.; Forouzesh, N.; Tolokh, I. S.; Karpatne, A.; Onufriev, A. V. Improving the Accuracy of Physics-Based Hydration-Free Energy Predictions by Machine Learning the Remaining Error Relative to the Experiment. *Journal of chemical theory and computation* **2023**, *20*, 396–410.

(20) He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. A fast and high-quality charge model for the next generation general AMBER force field. *The Journal of Chemical Physics* **2020**, *153*.

(21) Martins, S. A.; Sousa, S. F.; Ramos, M. J.; Fernandes, P. A. Prediction of solvation free energies with thermodynamic integration using the general amber force field. *Journal of Chemical Theory and Computation* **2014**, *10*, 3570–3577.

(22) Yu, Z.; Batista, E. R.; Yang, P.; Perez, D. Acceleration of Solvation Free Energy Calculation via Thermodynamic Integration Coupled with Gaussian Process Regression and Improved Gelman–Rubin Convergence Diagnostics. *Journal of Chemical Theory and Computation* **2024**, *20*, 2570–2581.

(23) Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *The Journal of Chemical Physics* **1954**, *22*, 1420–1426.

(24) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **1990**, *112*, 6127–6129.

(25) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annual review of physical chemistry* **2000**, *51*, 129–152.

(26) Bandyopadhyay, P.; Gordon, M. S. A combined discrete/continuum solvation model: application to glycine. *The Journal of Chemical Physics* **2000**, *113*, 1104–1109.

(27) Bandyopadhyay, P.; Gordon, M. S.; Mennucci, B.; Tomasi, J. An integrated effective fragment—polarizable continuum approach to solvation: Theory and application to glycine. *The Journal of chemical physics* **2002**, *116*, 5023–5032.

(28) Zhang, P.; Shen, L.; Yang, W. Solvation free energy calculations with quantum me-

chanics/molecular mechanics and machine learning models. *The Journal of Physical Chemistry B* **2018**, *123*, 901–908.

(29) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

(30) Bennett, W. D.; He, S.; Bilodeau, C. L.; Jones, D.; Sun, D.; Kim, H.; Allen, J. E.; Lightstone, F. C.; Ingólfsson, H. I. Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning. *Journal of Chemical Information and Modeling* **2020**, *60*, 5375–5381.

(31) Chen, D.; Gao, K.; Nguyen, D. D.; Chen, X.; Jiang, Y.; Wei, G.-W.; Pan, F. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature communications* **2021**, *12*, 3521.

(32) Alibakhshi, A.; Hartke, B. Improved prediction of solvation free energies by machine-learning polarizable continuum solvation model. *Nature Communications* **2021**, *12*, 3584.

(33) Lim, H.; Jung, Y. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *Journal of Cheminformatics* **2021**, *13*, 56.

(34) Pathak, Y.; Mehta, S.; Priyakumar, U. D. Learning atomic interactions through solvation free energy prediction using graph neural networks. *Journal of Chemical Information and Modeling* **2021**, *61*, 689–698.

(35) Zhang, D.; Xia, S.; Zhang, Y. Accurate prediction of aqueous free solvation energies using 3D atomic feature-based graph neural network with transfer learning. *Journal of chemical information and modeling* **2022**, *62*, 1840–1848.

(36) Low, K.; Coote, M. L.; Izgorodina, E. I. Explainable solvation free energy prediction combining graph neural networks with chemical intuition. *Journal of Chemical Information and Modeling* **2022**, *62*, 5457–5470.

(37) Zhang, Z.-Y.; Peng, D.; Liu, L.; Shen, L.; Fang, W.-H. Machine Learning Prediction of Hydration Free Energy with Physically Inspired Descriptors. *The Journal of Physical Chemistry Letters* **2023**, *14*, 1877–1884.

(38) Pattanaik, L.; Menon, A.; Settels, V.; Spiekermann, K. A.; Tan, Z.; Vermeire, F. H.; Sandfort, F.; Eiden, P.; Green, W. H. ConfSolv: Prediction of Solute Conformer-Free Energies across a Range of Solvents. *The Journal of Physical Chemistry B* **2023**, *127*, 10151–10170.

(39) Vyboishchikov, S. F. Predicting Solvation Free Energies Using Electronegativity-Equalization Atomic Charges and a Dense Neural Network: A Generalized-Born Approach. *Journal of Chemical Theory and Computation* **2023**, *19*, 8340–8350.

(40) Vyboishchikov, S. F. Dense Neural Network for Calculating Solvation Free Energies from Electronegativity-Equalization Atomic Charges. *Journal of Chemical Information and Modeling* **2023**, *63*, 6283–6292.

(41) Vyboishchikov, S. F. Solvation Enthalpies and Free Energies for Organic Solvents through a Dense Neural Network: A Generalized-Born Approach. *Liquids* **2024**, *4*, 525–538.

(42) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **2004**, *25*, 1157–1174.

(43) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry* **2000**, *21*, 132–146.

(44) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* **2002**, *23*, 1623–1641.

(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

(46) Landrum, G.; others Rdkit: Open-source cheminformatics software. **2016**,

(47) Ho, T. K. Random Decision Forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995; pp 278–282.

(48) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(49) Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* **2013**, *7*, 21.

(50) Brownlee, J. Gradient boosting with scikit-learn, xgboost, lightgbm, and catboost. *Machine Learning Mastery* **2020**,