Supporting information for: Solvation Free Energy Calculations with Quantum Mechanics / Molecular Mechanics and Machine Learning Models

Pan Zhang,[†] Lin Shen,[†] and Weitao Yang[*,‡,¶]

†*Department of Chemistry, Duke University, Durham, North Carolina 27708, United States*

‡*Department of Chemistry and Department of Physics, Duke University, Durham, NC 27708, United States*

¶*Key laboratory of Theoretical Chemistry of Environment, Ministry of Education, School of Chemistry and Environment, South China Normal University, Guangzhou 510006, P.R.China*

E-mail: weitao.yang@duke.edu

Gradient Boosting

Here is the basic procedure of component-wise gradient boosting algorithm:[S1]

1. Define the loss function $L(\mathbf{y}, f(\mathbf{x}))$, the base learner $h(x) = ax + b$, and the database $(\mathbf{x}_i, y_i)$ including n samples, where $\mathbf{x}_i$ is the p-dimensional input variable and y_i is the reference value for the i-th sample.

2. Initialize the prediction model $f_0(\mathbf{x})$.

3. For $m = 1$ to m_{stop}:

   (a) Compute the negative gradient of the loss function $\mathbf{r} = \dfrac{\partial L(\mathbf{y}, f_{m-1}(\mathbf{x}))}{\partial f_{m-1}(\mathbf{x})}$ as the residue at the m-th step.

   (b) Fit the residue $\mathbf{r}$ using different one-dimensional base learners $h(x_j)$, respectively, where x_j is the component of the p-dimensional input variables, and j varies from 1 to p.

   (c) Compare the RMSEs using different base learners and choose the best one as $h_{m-1}(\mathbf{x})$.

   (d) Build the boosting model as $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu h_{m-1}(\mathbf{x})$.

4. Determine the final prediction model $f(\mathbf{x})$ through early stopping technique.

At step 3, $\nu \in [0, 1]$ is the shrinkage parameter that reduces the significance of every boosting step in order to slow down weak learning and avoid early overfitting. It is set as 0.01 in this work. The early stopping technique[S2,S3] is applied to select the prediction model with the best performance on the testing set.

Gaussian Model

Here is the detailed process to build Gaussian model. The initial values of the parameters of the k-th Gaussian distribution, including the mean vector $\boldsymbol{\mu}_k$, the covariance matrix $\boldsymbol{\Sigma}_k$, and the percentage P_k, are first calculated using k-means clustering. The following process is then implemented:

1. The probability of a configuration $\mathbf{x}_i$ in the k-th Gaussian distribution, denoted as $P(\mathbf{x}_i \in k)$, is

$$P(\mathbf{x}_i \in k) = \frac{P_k \mathrm{Norm}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k P_k \mathrm{Norm}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \tag{1}$$

   where

$$\mathrm{Norm}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right]}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}}. \tag{2}$$

2. The P_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are updated as

$$P_k = \frac{\sum_i P(\mathbf{x}_i \in k)}{\sum_k \sum_i P(\mathbf{x}_i \in k)}, \tag{3}$$

$$\boldsymbol{\mu}_k = \frac{\sum_i P(\mathbf{x}_i \in k)\mathbf{x}_i}{\sum_i P(\mathbf{x}_i \in k)}, \tag{4}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_i P(\mathbf{x}_i \in k)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_i P(\mathbf{x}_i \in k)}. \tag{5}$$

3. Repeat step 1 and 2 until P_k is unchanged for all Gaussian distributions or the maximum step is reached.

4. Check the configuration of interest $\mathbf{x}$ in MD samplings using

$$P(\mathbf{x}) = \sum_k P_k \mathrm{Norm}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{6}$$

   The database will be extended with the configuration if $P(\mathbf{x})$ is smaller than the predetermined threshold.

Simulation Details

To calculate the solvation free energy (SFE), the solute molecule was solvated in a cubic water box of 64 × 64 × 64 Å^3 and treated as QM subsystem. The surrounding water molecules were treated as MM subsystem. The TIP3P water model[S4] was applied under periodic boundary condition. The cutoff distance for nonbonded interactions was set as 12 Å. The self-consistent charge density functional tight binding with second-order formulation and MIO basis (DFTB2/MIO)[S5,S6] was employed as the low-level SQM model, and the DFT method with the B3LYP hybrid functional[S7,S8] and the 6-31G(d) basis set was employed as the high-level ab initio QM model. The soft-core potential was used to describe the van der Waals (vdW) interaction between QM and MM subsystems and expressed as[S9]

$$E^{\mathrm{vdw}}_{\mathrm{QM/MM}}(\lambda_{\mathrm{vdw}}) = 4\lambda_{\mathrm{vdw}} \sum_{i\in \mathrm{QM}} \sum_{j\in \mathrm{MM}} \epsilon_{ij} \left[\left(\frac{\sigma_{ij}^6}{\alpha(1-\lambda_{\mathrm{vdw}})^2\sigma_{ij}^6 + r_{ij}^6} \right)^2 - \frac{\sigma_{ij}^6}{\alpha(1-\lambda_{\mathrm{vdw}})^2\sigma_{ij}^6 + r_{ij}^6} \right], \tag{7}$$

where r_{ij} is the distance between QM atom i and MM atom j, ϵ and σ are the potential well depth and vdW radius, which were chosen from the CHARMM22 force field[S10] in this work, λ_{vdw} is the switching parameter varying from 0 to 1 for thermodynamic integration, and α is the soft-core parameter as 0.5 in this work. For each system, 21 MD simulations were performed to calculate the SFE with thermodynamic integration, which includes 11 trajectories with $\lambda_{\mathrm{ele}} = 0.0$ and $\lambda_{\mathrm{vdw}} = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ and 10 trajectories with $\lambda_{\mathrm{vdw}} = 1.0$ and $\lambda_{\mathrm{ele}} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. For the low-level DFTB2/MIO/MM and high-level B3LYP/6-31G(d)/MM simulations, the total time of each MD trajectory was 110 ps, which consists of 10 ps of equilibration and 100 ps of sampling. In the initialization stage of QM/MM ML, 50 configurations were randomly selected from each trajectory of DFTB2/MIO/MM MD simulations. Then the B3LYP/6-31G(d)/MM potential energies of all 1050 configurations were calculated in order to build the initial database. In the subsequent QM/MM ML simulations, the total time of each MD trajectory was 110 ps, which consists of 50 ps in the update stage and 60 ps in the finalization

stage. For all MD simulations, the integration time step was set as 1 fs, and the system temperature was maintained at 300 K using a Langevin thermostat.[S11] All simulations were implemented using the in-house QM4D program[S12] combined with GAUSSIAN 03 program for DFT calculations.[S13]

Table S1. Hyperparameters η (bohr^{-2}) and ζ of Symmetry Functions for Different Elements of Six Molecules.

Molecules	η				ζ			
	C	H	O	N	C	H	O	N
Acetic acid	0.20	0.60	0.10		1.00	1.00	0.05	
Acetamide	0.30	0.60	0.05	0.05	0.40	0.05	1.00	0.05
Acetone	0.70	0.20	0.20		0.05	0.05	0.05	
Benzene	0.70	0.05			0.90	0.05		
Ethanol	0.20	0.60	0.05		1.00	0.05	0.05	
Methylamine	0.05	0.40		0.30	0.05	0.05		0.05

Table S2. Hyperparameters $n_{\max}$, $l_{\max}$, R_n (Å) and α (Å^{-2}) of Power Spectrum for Six Molecules.

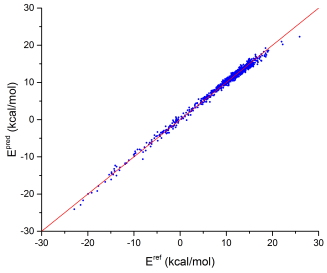
Molecules	QM			MM			α
	$n_{\max}$	$l_{\max}$	R_n/n	$n_{\max}$	$l_{\max}$	R_n/n	
Acetic acid	4	4	0.5	4	2	2.0	0.3
Acetamide	4	4	0.5	4	2	2.0	0.5
Acetone	4	4	0.5	4	2	2.0	0.2
Benzene	4	4	0.5	4	4	2.0	0.5
Ethanol	4	4	0.5	4	2	2.0	0.8
Methylamine	4	4	0.5	4	2	2.0	0.3

Table S3. Solvation Free Energies with Standard Deviations (kcal/mol) from MD Simulations Using DFTB2/MIO/MM, B3LYP/6-31G(d)/MM and QM/MM ML Models (Symmetry Functions and Power Spectrum) Updated Using Output-Check (Model 1), k-Means Clustering (Model 2) or Gaussian Model (Model 3) to Detect New Configurations.

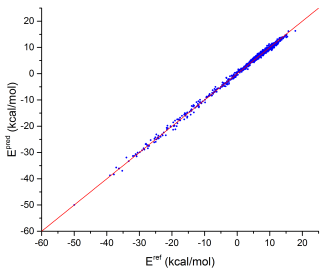
Molecules	DFTB/MM	B3LYP/MM	Symmetry Functions			Power Spectrum		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Acetic acid	-5.0 ± 0.2	-7.5 ± 0.3	-6.8 ± 0.2	-6.7 ± 0.5	-7.0 ± 0.4	-6.9 ± 0.3	-6.8 ± 0.2	-7.3 ± 0.5
Acetamide	-9.0 ± 0.2	-12.1 ± 0.3	-11.4 ± 0.2	-11.3 ± 0.2	-11.6 ± 0.2	-10.9 ± 0.3	-10.9 ± 0.3	-11.7 ± 0.4
Acetone	-2.3 ± 0.2	-4.3 ± 0.2	-3.6 ± 0.3	-3.9 ± 0.4	-3.8 ± 0.3	-3.9 ± 0.2	-3.8 ± 0.3	-3.9 ± 0.3
Benzene	1.0 ± 0.3	-0.6 ± 0.2	-0.3 ± 0.4	-0.2 ± 0.3	-0.3 ± 0.5	-0.2 ± 0.4	-0.5 ± 0.4	-0.6 ± 0.4
Ethanol	-1.0 ± 0.3	-4.8 ± 0.3	-4.6 ± 0.5	-4.6 ± 0.2	-4.3 ± 0.3	-4.0 ± 0.3	-4.0 ± 0.3	-4.6 ± 0.7
Methylamine	0.9 ± 0.1	-5.2 ± 0.2	-3.8 ± 0.1	-4.0 ± 0.5	-4.5 ± 0.5	-2.4 ± 0.2	-2.2 ± 0.2	-2.5 ± 0.4

Table S4. Solvation Free Energies (kcal/mol) and Percentages of New Configurations Sampled during MD Simulations (in Parentheses) from MD Simulations with QM/MM ML Models (Symmetry Functions and Power Spectrum) Updated Using Boundary of Input Variables to Detect New Configurations.

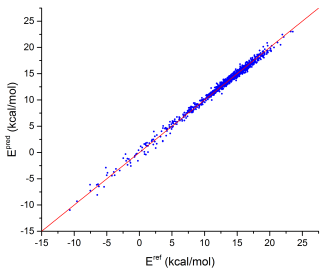
Molecules	Symmetry Functions	Power Spectrum
Acetic acid	-6.8 (3.0%)	-6.5 (59.4%)
Acetamide	-11.4 (2.7%)	-11.3 (61.3%)
Acetone	-3.9 (1.6%)	-3.8 (59.0%)
Benzene	-0.2 (1.2%)	-0.2 (56.8%)
Ethanol	-4.1 (3.1%)	-4.3 (61.1%)
Methylamine	-4.2 (1.6%)	-2.9 (57.6%)



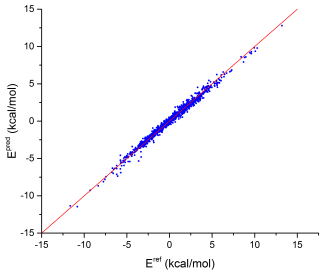
Acetic acid



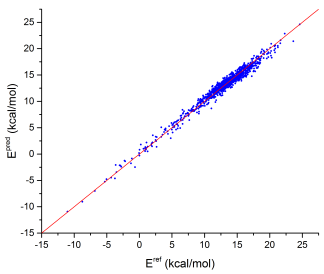
Acetamide



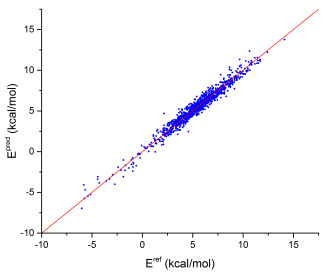
Acetone



Benzene

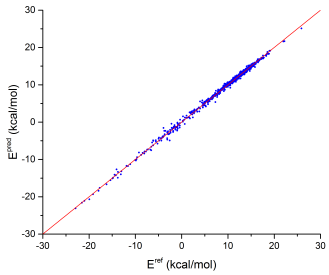


Ethanol

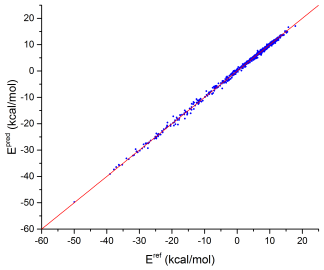


Methylamine

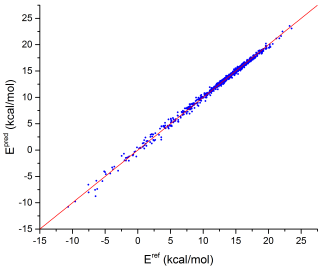
(a) Symmetry Functions



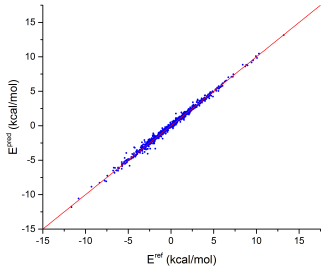
Acetic acid



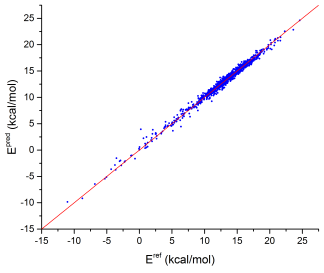
Acetamide



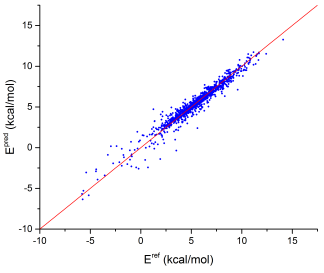
Acetone



Benzene



Ethanol



Methylamine

(b) Power Spectrum

Figure S1. Comparison of the ML-predicted potential energies (E^{pred}) using different ML models (symmetry functions and power spectrum) with the reference values at the B3LYP/6-31G(d)/MM level (E^{ref}) for six molecules.

References

(S1) Schmid, M.; Hothorn, T. *Comput. Stat. Data Anal.* **2008**, *53*, 298–311.

(S2) Zhang, T.; Yu, B. *Ann. Statist.* **2005**, *33*, 1538–1579.

(S3) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*; New York, NY: Springer, 2009.

(S4) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(S5) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(S6) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(S7) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(S8) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(S9) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. *J. Chem. Theory Comput.* **2010**, *6*, 1509–1519.

(S10) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(S11) Gunsteren, W. F. V.; Berendsen, H. J. C. *Mol. Simul.* **1988**, *1*, 173–185.

(S12) Hu, X.; Hu, H.; Yang, W. QM4D: An integrated and versatile quantum mechanical/molecular mechanical simulation package. http://www.qm4d.info/.

(S13) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03, Revision D.02. Gaussian, Inc., Wallingford, CT, 2004.