Supporting information: "Machine Learning Prediction of Interaction Energy in Rigid Water Clusters"

Samik Bose[1], Diksha Dhawan[1], Sutanu Nandi[2], Ram Rup Sarkar[2] and Debashree Ghosh[1,*]

[1] *Centre for Mathematical, Computational and Data Science, Indian Association for the Cultivation of Science, Kolkata - 700032, West Bengal, India*
[2] *Chemical Engineering & Process Development (CEPD), CSIR-National Chemical Laboratory, Pune - 411008, Maharasthra, India*

Email: debashree.ghosh@gmail.com
Phone: +91-033-24731103

1 Four body term in MBE: Minimal accuracy gain at steep computational cost.

As discussed in the article, most of the previous studies which employ MBE of interaction energy, generally neglect the fourth order and higher terms. In this work, we also take the same route. This is due to two reasons, i) large number of four body and higher body interaction terms in an n-mer. ii) computational cost increases significantly going from three body to four body terms and higher.

In case of water decamer, the number of four body terms is 210, whereas the number of two body and three body terms are 45 and 120 respectively. Moreover, estimation of each of these 210 four body terms requires ∼80 seconds of walltime in compasrison to ∼17 seconds and ∼5 seconds required for a three body and two body term respectively, calculated in same computers. So, the approximate time required for complete four body contribution estimation in a single decamer is $210 \times 80 = 16800$ seconds in comparison to 2160 and 225 seconds for three body and two body respectively.

Now, it should be noted that, Xantheas has already shown that the contribution of four body term is only ∼2% in tetramer interaction energy. Moreover, it is well documented that, the four body and higher order terms have a contribution of ¡ 2% in cluster interaction energies. In order to include that contribution, one needs to incorporate the four and higher body terms. However, as illustrated in the preceding paragraph, the estimation of four body contributions requires significantly large computational power (almost 8 times more than that of three body terms). Hence, the four body and higher terms are neglected.

2 Configuration space of water decamer: 100K and 300K clusters.

The configuration space of water decamer is generated by classical NVT simulation at two different temperatures. The configuration space at 100K is different from that of the 300K. This is because the kinetic energy at the two temperatures are significantly different. At 100K, due to less thermal energy the intermolecular motion of water is significantly less, giving rise to a more compact configuration space. However, at 300K, the thermal energy provided to the system is comparatively larger. Hence the 300K configuration space will be less compact in comparison to its 100K counterpart.

We use both set of configurations for training and testing of SVR to ensure that the scheme is transferable to different type of configurations at different temperatures. Hence, at any point of the phase space of water, we can use the SVR based approach with the same set of optimized parameters to estimate the interaction energies of water cluster. Also it should be noted that training of SVR with a properly represented training set is mandatory for this purpose, i.e., the datapoints of both 100K and 300K should be present in the training set.

3 Optimization of σ with different datasets.

We change the RBF kernel parameter, σ, and train five datasets using SVR. The value of ν is kept at 0.1 throughout. The datasets consist of randomly chosen 2000, 3200, 4000, 6000 and 8000 training points of dimer respectively. Since we always maintain a five fold cross-validation for error analysis, the testing sets consist of randomly chosen 500, 800, 1000, 1500 and 2000 datapoints respectively. Fig. S1 shows that, with σ value 1 and 0.5, the RMSE in dimer interaction energy for training and testing are significantly high. For σ=100, although the training RMSE is lowest, the testing RMSE is considerably higher than the training RMSE. The large difference between training and testing RMSE clearly indicates the overfitting problem with σ=100. Now, with σ=10, we notice that, the training and testing RMSEs are closer to each other (maximum difference is 0.01 Kcal/mol) than in comparison to σ=100. Fig. 4 in the article also indicates σ=10 has lower testing RMSE. Therefore σ=10 is chosen as the optimized value.

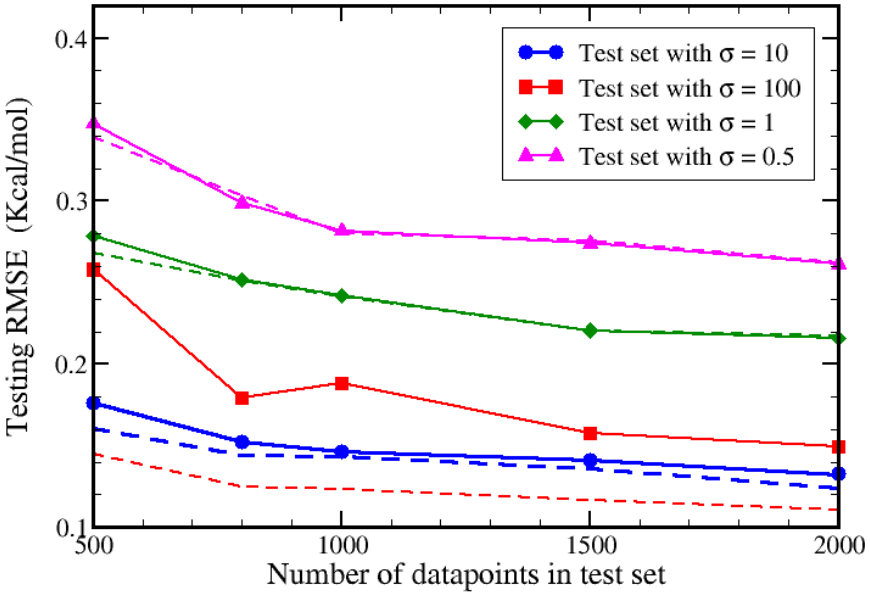


Figure S1: Variance of RMSE with different datasets. The bold and dashed lines represent training and testing RMSE respectively.

4 Comparison between SVR and TIP3P force fields predicted dimer interaction energy.

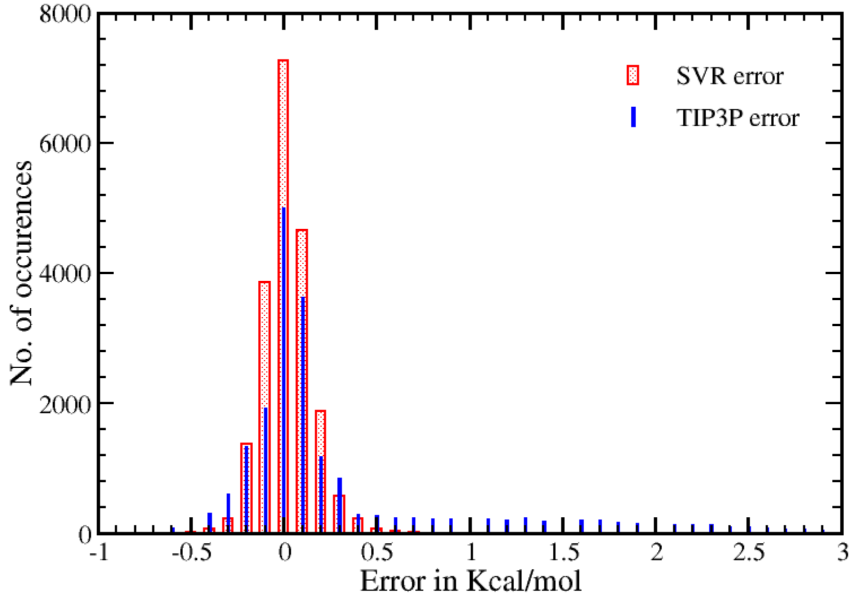
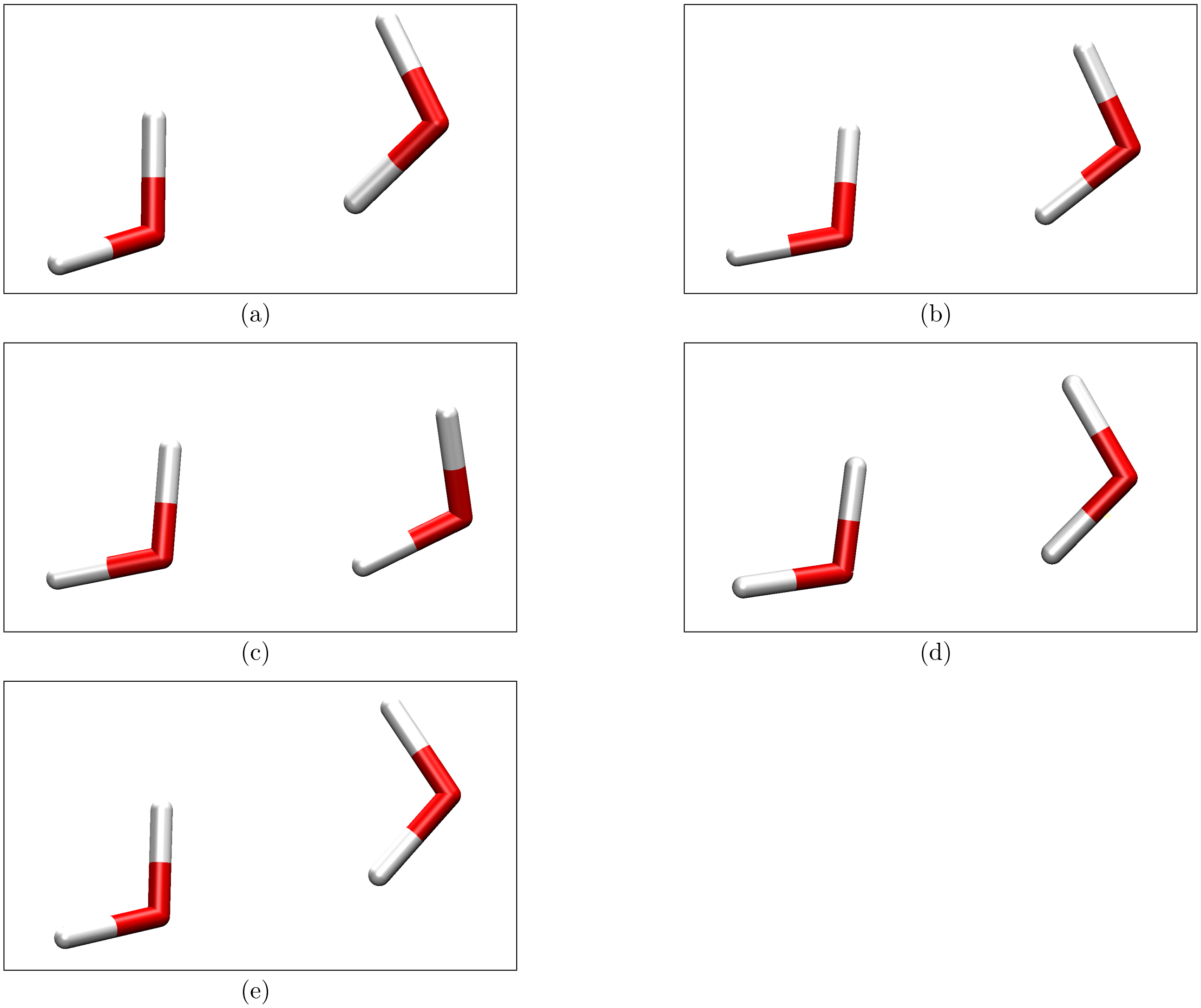


Figure S2: Distribution of error in interaction energy of water dimers: Comparison between SVR and TIP3P force fields.

5 Structure and interaction energy of dimers with high prediction error.

The five dimer structures with highest error in SVR prediction of interaction energy are shown in fig. S3. The O-O bond distances, O-H-O bond angle and interaction energy of each structure are presented in table S1. We notice the O-O bond distances are around the H-bond distance cut-offs. The O-H-O angles are in the range of weak H-bond. However, the interaction energies of these stuctures are significantly lesser than H-bonded water dimers (interaction energy: 3.5 - 6.5 Kcal/mol).


(a)

(b)

(c)

(d)

(e)

Figure S3: The five dimer structures with highest error in the prediction of interaction energy.

Table S1: Interaction energies, O-O distances, O-H-O angle of five dimer structures with highest prediction error.

Structure no. in fig. S3	O-O distance in the dimer (in Å)	O-H-O angle in the dimer	Interaction energy of the dimer (Kcal/mol)
(a)	2.20	133°	-0.91
(b)	2.13	134°	-0.87
(c)	1.78	151°	-1.09
(d)	1.87	149°	-0.88
(e)	2.11	141°	-0.49

6 Error distribution of dimer interaction energy by CCSD and ωB97X-D.

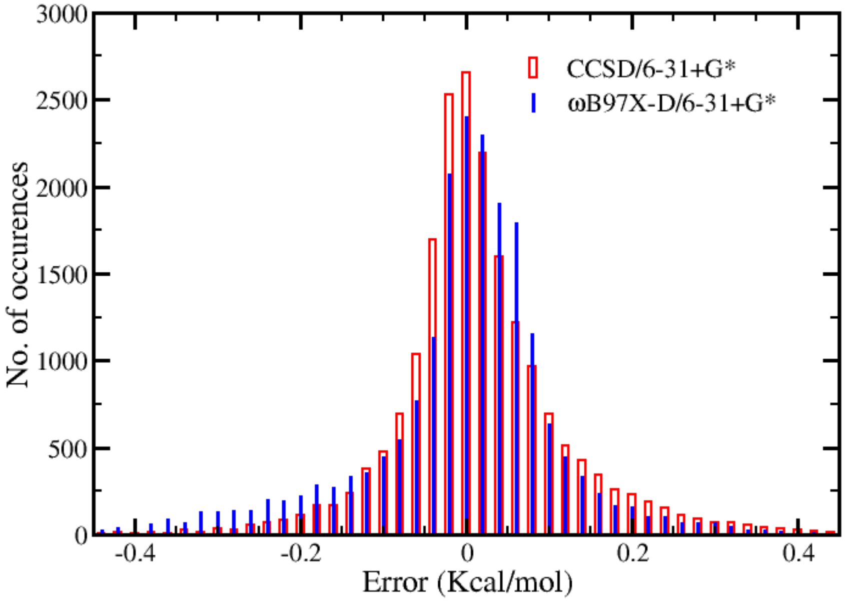


Figure S4: Comparison of error distribution of dimer interaction energy, as estimated by CCSD and ωB97X-D.

Fig. S4 shows error distribution in the prediction of dimer interaction energies of full dataset with the two QM methods. We see that the two distributions are comparable, with both having maxima at 0 Kcal/mol and similar spread. This indicates that the SVR-MBE scheme does function well irrespective of the underlying QM method.

7 Total Error in the prediction of decamer interaction energy.

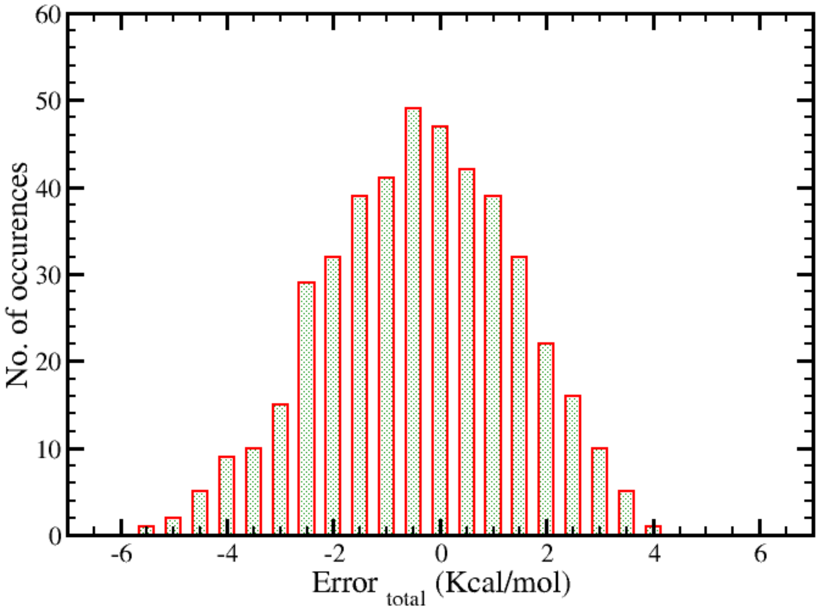


Figure S5: Distribution of total error in ML-MBE scheme.

Fig. S5 shows the distribution of error in ML-MBE scheme in predcition of decamer interaction energy. As reported in the article, we notice an RMSE of 1.98 kcal/mol, which amounts to only 2.78% error in our prediction. From the error distribution, it is evident that, ∼75% of the total set of decamers have error less than the RMSE. Also, we note that only 6% of the decamers have higher error (>5%) in prediction of interaction energy.