Study of Acid Suppressed Thickener Technology Using Density Functional Theory and Machine Learning Techniques

by

Wenkun Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Materials Science and Engineering)
in the university of Michigan
2018

Doctoral Committee:

Professor John Kieffer, Chair
Professor Ronald G. Larson
Assistant Professor Liang Qi
Associate Professor Anish Tuteja

Wenkun Wu

wuwenkun@umich.edu

ORCID iD: 0000-0002-3764-1130

Table of Contents

List of Figures

List of Tables

Abstract

Hydrophobically modified ethylene oxide urethane (HEUR) rheology modifiers, which are water-based polyurethane formulations manufactured by Dow Coating Materials, a division of the Dow Chemical Company, are often added to interior and exterior water-based Latex paint formulations to control their viscosity. The thickening efficiency of the HEUR rheology modifier is controlled by the pH of the solvent, as this affects the protonation-deprotonation equilibrium of the amine hydrophobe group at the end of the rheology modifier polymer chain. The principal quantity characterizing this equilibrium is the acid dissociation constant (pK_a) of the hydrophobe group, which identifies the transition between high and low viscosity of the suspension. To gain a better understanding of the functioning of the hydrophobe molecular groups, and to develop novel hydrophobes that meet specific performance characteristics, it is important to accurately predict the pK_a based on first principles calculations, and use it as a first evaluation criterion for a rapid screening of candidate hydrophobe molecules.

A main source of error in the pK_a calculation is the value of solvation free energy of the molecule in its charged state. We therefore develop new methods to increase the accuracy of the solvation free energy calculation for charged species without excessive increase the computational expense. This includes a hybrid cluster-continuum model approach, where explicit solvent molecules are added to the traditionally employed continuum solvation model, and a molecular dynamics (MD sampling procedure that eliminates the costly energy minimization step.

Using test molecules for pK_a calculations, we systematically examine the convergence behavior in terms of number of explicit water molecules that need to be included in the cluster-continuum model, the influence of the dielectric constant attributed to the continuum, and the placement of a counter ion for charge neutrality for the accurate calculation of the solvation free energy. We establish that the MD sampling method yields results comparable the energy minimization procedure during density functional theory (DFT) calculations, but at 100 times the speed. When calculating the solvation free energy and the pK_a calculation of a known hydrophobe, ethoxylated bis(2-ethylhexy)amine, we find that including explicit water molecules and a fragment of the latex polymer in its local environment both significantly improve the results.

Finally we develop an informatics-based approach that employs a transferable machine learning (ML) model, trained and validated on a limited amount of experimental data, to predict the solvation free energies of new ionic species at a reasonable computational cost. We compare three different ML methods – linear ridge regression, support vector regression and random forest regression, and find that the model trained by the random forest regression method yields the predictions with the lowest mean absolute error. A feature selection analysis shows that the atomic fraction feature, which reflects the chemical constitution of the hydrophobe, plays the most important role in the solvation free energy prediction. Adding the Wiener index, a measure of the molecular topology, and the solvent accessible surface area of the molecules further improve the performance of the model. Accordingly, our ML model predicts the solvation energies of ionic species, including our test hydrophobe molecule, with similar accuracy as

atomistic modeling using first-principles calculations.

Chapter 1

Introduction

1.1 Background and Motivation

Aqueous polymer systems, such as containing emulsion polymer binders used for coating surfaces, typically draw on thickeners to obtain target rheological properties, like the specific degree of viscosity needed for the proper formulation and application. Traditional thickeners used in such coating emulsions, e.g., cellulose, are non-associating thickeners. [1] However, in the last two decades, a new, improved class of thickeners known as associative thickeners have been found to be superior to cellulose, offering properties including improved flow and leveling in aqueous systems. [2]

Thickeners are called associative because their thickening function involves hydrophobic associations among hydrophobic groups in the thickener molecules, as well as between these hydrophobic groups in and other hydrophobic surfaces. Commonly used associative thickeners have a polymeric backbone with hydrophobic functional groups either attached to or incorporated into the backbone. The backbone can be neutral, such as poly(ethylene oxide) (PEO) and poly(acrylamide) (PAM), or charged, such as poly(acrylic acid) (PAA) or partially hydrolyzed poly(acrylamide) (PHPAM). The hydrophobic groups are typically classified as aliphatic, fluorinated, or aromatic. [3]

Among the commercially available associative thickeners are hydrophobically modified ethylene oxide urethane (HEUR) rheology modifiers, which are water-based polyurethane formulations manufactured by Dow Coating Materials, a division of the Dow Chemical Company. They are added to interior and exterior latex (water-based) paint formulations to control the viscosity of the paint. [4] The structure of the HEUR molecule is shown in Figure 1. It is composed of a polyurethane polyether backbone and two amine hydrophobes on each end of the backbone.

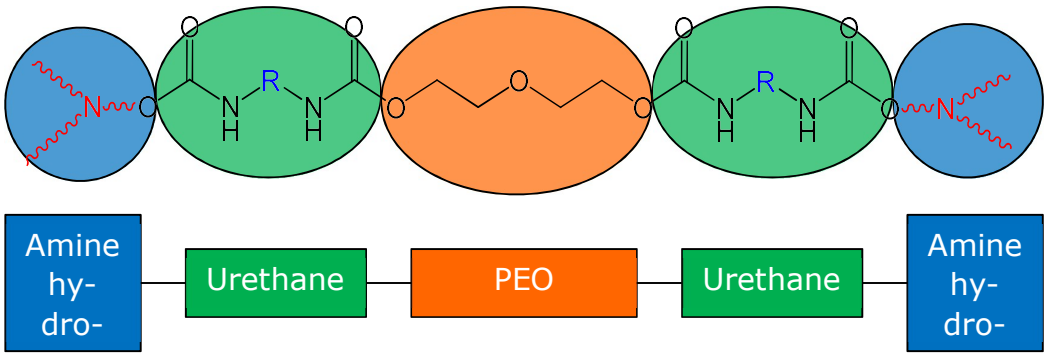


Figure 1-1. Structure of HEUR rheology modifier.

Ethoxylated bis(2-ethylhexy)amine is one of the potential candidates for the hydrophobe group on HEUR molecules. It is a tertiary amine with a NCO reactive group as shown in Figure 2. Indeed, ethoxylated bis(2-ethylhexy)amine provides a unique mechanism for controlling the thickening function of the HEUR molecules, because, depending on the acidity of the aqueous solution, the nitrogen atom on the hydrophobe molecule can either protonate of deprotonate. In a basic environment, most of the ethoxylated bis(2-ethylhexy)amine remains in its deprotonated form, which is more hydrophobic and thereby promotes attraction of the amine to the latex particle surface, effectively anchoring the HEUR molecule to it. This results in a thickening of the suspension. Conversley, when lowering the pH of the solvent, the ethoxylated

bis(2-ethylhexy)amine protonates. The ionized form of the group interacts favorably with the dipole of the water molecule and is more easily solvated, which causes the HEUR molecules to detach from the latex surface, and hence reversing the thickening effect. The governing factor that reveals the point of transition between the two behaviors is the acid dissociation constant K_a, or, as it is reported most commonly, the negative logarithm of this constant, pK_a. In view of further elucidating this process, understanding the underlying mechanisms, and ultimately, achieve a predictive design capability based on a computational approach, it is imperative to be able to accurately calculate the pK_a values for arbitrary hydrophobe molecules.

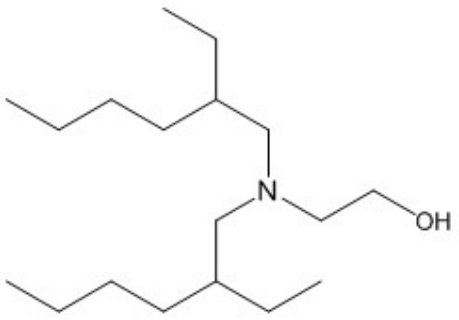


Figure 1-2. Molecular Structure of ethoxylated bis(2-ethylhexy)amine.

1.2 Project Overview

Over the past decade, accurate pK_a predictions using quantum chemical methods have been attempted numerous times. The most common approach is to calculate the pK_a using a thermodynamic cycle involving the gas-phase reaction free energy and solvation free energies of all reaction partners. The calculation details for this approach are outlined in the next section. While the calculation of the gas-phase reaction energies is straightforward, as it only involves the reaction partners themselves, the accurate calculation of free energies in solution remains difficult, because the interactions with the solvent must also be accounted for. In early pK_a

calculations of the solvation free energy, [5–7] a continuum model was used to represent the solvent, which simply amounted to a uniform effective medium with a fixed dielectric constant. This model has been shown to deliver solvation free energies with an accuracy of ±1 kcal/mol for neutral solutes, [8] but the mean unsigned errors for ionic species are around 4 kcal/mol. [9] Since a variation in the solvation free energy by 1.36 kcal/mol results in an unit change of the pK_a value, [7] an error of 4 kcal/mol in the solvation free energy calculation is too large to allow for reliable prediction of pK_a values. Consequently, in more recent investigations, the continuum model has been refined by adding explicit solvent molcules to directly account for local solute solvent interactions that prominently contribute to the solvation energy. [10–12] This cluster-continuum model generally yields more accurate results. Both approaches are well documented in a recent review article by Ho and Coote. [13] However, there are no reported studies of pK_a calculations that account for the influence of the local environment beyond explicit solvent molecules. The overall goal of this thesis is to develop a method so that the solvation free energy and pK_a of the hydrophobe molecule on the HEUR rheology modifier could be accurately calculated without significant increase of computational effort. Specific objectives and milestones towards accomplishing this goal are:

- Gain a fundamental understanding of how to calculate pK_a and solvation free energy.
- Develop an improved method based on the traditional continuum model to calculate the solvation free energy more accurately without increasing much computational effort.

- Apply the method to study the solvation free energy and pK_a of the hydrophobe molecule on HEUR rheology modifier and study how the surroundings affect these properties.
- Develop an ML approach to further decrease the computational cost and give accurate prediction of the solvation free energy, and get an idea of the effectiveness of the hydrophobe molecular design.

1.3 Thesis Overview

The contents of the thesis are divided into 5 chapters. The results of three main projects are presented in Chapter 2, 3, and 4.

Chapter 2 presents the study of solvation free energy calculations with our hybrid cluster-continuum model. We calculate the solvation free energies of novel molecules using two different approaches; analyze the advantages for each of them and discuss the possible factors that could influence the calculation accuracy.

In Chapter 3 we apply the model we have developed in Chapter 3 to study the solvation free energy and pK_a value of the hydrophobe molecule. Besides the influence of explicit solvent molecule, we also study the influences to the solvation free energy and pK_a value, which could be brought by the Latex particle.

In Chapter 4 we further decrease the computation effort/time by developing an informatics based ML model to predict the solvation free energies from some pre-existed solvation data. We also discuss the keys that affect the prediction accuracy and possible ways to further improve the model performance.

Chapter 5 gives a final summary, including major findings and achievements in this thesis study. An outlook for future research in this field is also suggested.

1.4 References

1 Bobsein, B.R., Johnson, M.M., Rabasco, J.J., and Zeszotarski, C., 'Thickener composition and method for thickening aqueous systems,' U.S. patent no. US7741402B2 (2010).

2 Sau, A.C., 'Hydrophobically modified poly(acetal-polyethers),' U.S. patent no. US5574127A (1996).

3 Winnik, M.A. and Yekta, A., 'Associative polymers in aqueous solution,' *Curr. Opin. Colloid Interface Sci* **2**, 424 (1997).

4 The Dow Chemical Company, 'AcrysolTM hydrophobically modified ethylene oxide urethane (HEUR) rheology modifiers,' 2011.

5 Jang, Y.H., Goddard, W.A., Noyes, K.T., Sowers, L.C., Hwang, S., and Chung, D.S., 'First principles calculations of the tautomers and pK(a) values of 8-oxoguanine: Implications for mutagenicity and repair,' *Chem. Res. Toxicol.* **15**, 1023 (2002).

6 Klicić, J.J., Friesner, R.A., Liu, S.-Y., and Guida, W.C., 'Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods,' *The Journal of Physical Chemistry A* **106**, 1327 (2002).

7 Liptak, M.D. and Shields, G.C., 'Accurate pK(a) calculations for carboxylic acids using Complete Basis Set and Gaussian-n models combined with CPCM continuum solvation methods,' *J. Am. Chem. Soc.* **123**, 7314 (2001).

8 Cramer, C.J., '*Essentials of computational chemistry*,' 2nd edition (WILEY, 2004), P579.

9 Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening,' *J. Chem. Theory Comput.* **5**, 2447 (2009).

10 Bryantsev, V.S., Diallo, M.S., and Goddard, W.A., 'Calculation of solvation free energies of charged solutes using mixed cluster/continuum models,' *J. Phys. Chem. B* **112**, 9709 (2008).

11 Pliego, J.R. and Riveros, J.M., 'The cluster-continuum model for the calculation of the solvation free energy of ionic species,' *J. Phys. Chem. A* **105**, 7241 (2001).

12 Zhan, C.G. and Dixon, D.A., 'Absolute hydration free energy of the proton from first-

principles electronic structure calculations,' *J. Phys. Chem. A* **105**, 11534 (2001).

13 Ho, J.M. and Coote, M.L., 'A universal approach for continuum solvent pK(a) calculations: are we there yet,' *Theor. Chem. Acc.* **125**, 3 (2010)

Chapter 2

Hybrid Method for the Calculation of the Solvation Free Energy of Organic Molecules in Aqueous Solutions

2.1 Introduction

Accurate prediction of the solvation free energies is essential in many fields of study, ranging from chemical reactions in solutions to the design of functional molecules in chemistry and biochemistry. However, the reliable determination of the solvation free energies for ionic species can be computationally challenging. Over the past decade, much theoretical effort has been put into the development of methods to calculate the solvation free energy. Explicit solvation models involve actual solvent molecules in the calculation. This provides descriptive and realistic details of the solvation environment, which in principle should give an accurate result. However, the explicit representation of a dense solvent molecule configuration embedding the solute creates a large number of degrees of freedom and makes it difficult for first-principles calculation codes to find the fully relaxed complex structure. [1] Unlike the computationally expensive explicit models, continuum-based, or implicit solvent models, are easy to utilize. Examples are the polarizable continuum solvation model (PCM), [2] its variations like IPCM and SCIPCM, [3] and the SMx solvation models. [4–9] In these models, actual solvent molecules are represented by a uniform polarizable medium of fixed dielectric constant, and the solute molecule is embedded in a suitably shaped cavity. The main shortcoming

of implicit models is that they are inadequate for ionic species, mostly because of the failure to properly account for the strong localized solute-solvent interactions. Even for the recent SMD solvation model, where an additional correction term has been introduced to correct for these localized interactions, the average deviation between calculated and measured solvation energies is still of the order of 4 kcal/mol. [10] To overcome this shortcoming, cluster-continuum models, where explicit solvent molecules are added to the continuum model, have been devised in recent solvation free energy calculations. [11–15]

Cluster-continuum model approaches towards the calculation of solvation free energies in aqueous solution have been undertaken previously for species such as H^+, OH^- and F^- by Zhan, [11, 13, 14] and for H^+ and Cu^{2+} by Bryantsev. [15] These studies produced results close to the experimental solvation free energies are reported upon adding a seemingly arbitrary and relatively small number of solvent molecules in the cluster-continuum model. However, the range of increments in added solvent molecules in these studies is likely too narrow to prove convergence for the reported solvation energy magnitudes. In Zhan's work for H^+ solvation free energy calculation, only three different cluster sizes, i.e., $H^+(H_2O)$, $H^+(H_2O)_4$ and $H^+(H_2O)_{10}$ were considered. Similarly, Bryantsev examined four different sizes of cluster $H^+(H_2O)$, $H^+(H_2O)_6$, $H^+(H_2O)_{10}$ and $H^+(H_2O)_{14}$. Both authors claimed that their optimized structures are the most stable ones, but they only obtained comparable optimized geometries for $H^+(H_2O)_4$. The optimized geometries for $H^+(H_2O)_{10}$ were very different between these independent studies, even though the level of theory and basis set that these researchers were using, B3LYP/6-31++G(d,p) for Zhan and B3LYP/6-311++G(d,p) for Bryantsev, are almost identical, notwithstanding the fact that B3LYP is not the optimal functional that we should

use because it does not describe the Van der Waals interactions well. [16] This suggests that it must be really difficult to find the optimized geometry and that multiple stable structures may exist when the size of the cluster becomes large. It also suggests that the experimentally measured solvation free energy may indeed reflect an average of multiple possible stable clusters.

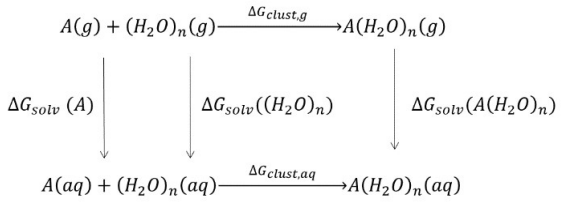
It is of great importance for the cluster-continuum model to establish convergence in the energy calculations as a function of the number of explicit solvent molecules, i.e., to determine the minimum number of explicit water molecules beyond which addition of more water molecules no longer improves the accuracy of the calculation. In the present work we systematically explore this aspect from two different computational approaches – DFT geometry optimization approach and MD sampling approach, using small solvated molecules. The DFT geometry optimization approach is similar to the methods used by Zhan and Bryantsev. We have improved this method by adding more clusters with different sizes and we also find multiple stable geometries for these clusters to calculate the solvation free energy as the size of the cluster increases. We further increase the number of water molecules contained in the cluster of the hybrid model, up to three coordination layers around the solute molecule, when using the MD sampling approach. Finally, we examine the importance of maintaining charge neutrality in these calculations by placing a counter ion to the solute molecule in the cluster, and how do decide on the location of this counter ion. To our knowledge, this issue has been ignored in all of the previous studies described in the literature.

2.2 Theory and Methodology

2.2.1 Calculation Formalisms

Traditional thermodynamic cycles for solvation free energy calculations with the cluster-continuum model, as shown in Scheme 1, have been discussed thoroughly in recent works. [15, 17]

Scheme 1-1: Thermodynamic cycle for solvation free energy calculation using the cluster-continuum model.



The top equation in this thermodynamic cycle describes gas-phase reactions between the solute A and a cluster of *n* water molecules $(H_2O)_n$,

$$A(g) + (H_2O)_n(g) \rightarrow A(H_2O)_n(g) \quad \Delta G_{clust,g}, \quad (1)$$

where $\Delta G_{clust,g}$ is the free energy of forming the gas-phase solute-water cluster $A(H_2O)_n$. It is defined with an ideal gas at 1 atm as the standard state. The vertical equations in the thermodynamic cycles describe the solvation processes, i.e., the immersion into immersion into the continuum with the dielectric constant of the solvent,

$$X(g) \rightarrow X(aq) \quad \Delta G_{solv}(X) \, . \quad (2)$$

Here X represents either the solute molecule, the water cluster, or the solute-water cluster. The $\Delta G_{solv}(X)$ are the solvation free energies associated with these processes.

The bottom equation describes the aqueous-phase reactions between the solute A and the clusters of water molecules,

$$A(aq) + (H_2O)_n(aq) \rightarrow A(H_2O)_n(aq) \quad \Delta G_{clust,aq} \quad (3)$$

The final product $A(H_2O)_n(aq)$ on the bottom right of the cycle is the hybrid cluster-continuum representation of solute in the aqueous phase. The formation of a cluster of solvated molecule in liquid water does not involve any change in free energy, i.e. $\Delta G_{clust,aq} = 0$. [18] Completing the thermodynamic cycle in Scheme 1 requires that all free energy changes add up to zero. Oftentimes, calculated gas-phase free energies are defined with an ideal gas at 1 atm as the standard state,, whereas the solvation free energies are associated with the 1 M (gas) → 1 M (solution) process. Therefore, a correction term ΔG_{corr} is needed to account for the changes in standard reference state. Accordingly, the solvation free energy of a given solute is calculated from the cluster formation free energy in the gas phase, the solvation free energies of the clusters and the standard state correction:

$$\Delta G_{solv}(A) = \Delta G_{clust,g} + \Delta G_{solv}(A(H_2O)_n) - \Delta G_{solv}((H_2O)_n) + \Delta G_{corr} \quad (4)$$

In this work, we assume that there are no significant changes in the geometry and vibrational energy upon transitioning from the gas phase to the aqueous phase. Only electronic contributions are taken into consideration in the DFT calculations. It has been shown that the use of vibrationally corrected free energies of solvation does not necessarily improve the quality of the calculated free energy of solvation, as long as solvation induced changes in the geometry and frequencies are small. [19, 20] Therefore the solvation free energy included in the solvation process $A(g) \rightarrow A(aq)$ can be written as:

$$\Delta G_{solv}(A) = E_{aq}(A) - E_g(A) \quad (5)$$

Here $E_{aq}(A)$ and $E_g(A)$ are the electronic energies of A in the aqueous phase and gas phase respectively. Notice that without the above assumption, we need to carry out the frequency

calculations to obtain the vibrational energies and they are done at the 1 atm standard state in Gaussian09, which is different from the standard state for the 1 M (gas) → 1 M (solution) solvation process. Thus we need a correction term for the changes in standard state. But now since we assume there are no vibration energy changes between the gas phase and the aqueous phase, all calculations are done at the same standard state (1 M) and we no longer need the correction term.

Similar to what we did in Equation 5, furthermore we assume that the vibrational energy change on the two sides of Equation 3 is negligible. [19, 20] Then we can get:

$$E_{aq}(A) = E_{aq}(A(H_2O)_n) - E_{aq}((H_2O)_n) \quad (6)$$

Therefore we can simplify the calculation of the solvation free energy of A in the following form:

$$\Delta G_{solv}(A) = E_{aq}(A(H_2O)_n) - E_{aq}((H_2O)_n) - E_g(A) \quad (7)$$

2.2.2 Cluster Configurations

To generate the initial configurations for solvation, we first carry out molecular dynamics (MD) simulations starting with a random placement of solute and solvent molecules. This procedure is used to produce a dense configuration of the solute and water molecules, in equilibrium at room temperature, and hence, exhibiting the corresponding degree of thermal disorder. Figure 1 shows a system with one solute molecule (NH_4^+) and 215 water molecules in a simulation box of 18.856 Å edge length, subject to periodic boundary conditions at 298.15K, before and after equilibration using MD simulations. During equilibration, the volume of the simulation box is kept constant achieving a density remains of 0.963g/cm^3.

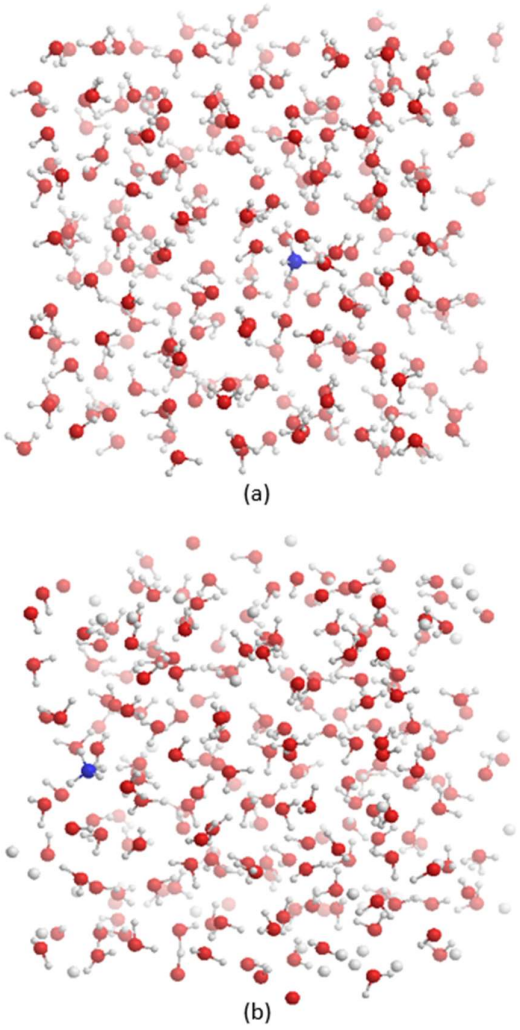


Figure 2-1. Structures of NH4+ and 215 water molecules before (a) and after equilibration using MD simulations (b).

From this point onwards, we have two different approaches to generate a structural model of the solvation cluster. Our goal is to construct a solvent configuration surrounding the solute molecule that is sufficiently detailed to comprise all essential local molecular interactions. One approach is that we extract a cluster that contains about two concentric coordination shells of water molecules surrounding the solute molecule from the MD simulation. Then we further eliminate molecules to reach the target cluster size and optimize the geometry using DFT energy minimization, ultimately resulting in the free energy of the fully relaxed cluster

at zero Kelvin. The other approach is that we sample a large number of distinct clusters from a series of different instantaneous configurations produced by the MD simulations and only carry out single point calculations to get a statistical distribution of cluster energies, reflecting the systems natural sampling of energies at finite temperatures. To this end, the calculation is repeated many times for configurations collected along the MD trajectory, appropriately spaced in time. Each configuration represents a snapshot of a structure subject to thermal motion, and from the distribution of calculated solvation energies we can evaluate room-temperature thermal averages. The first approach tends to result in specific molecular configurations constrained by coordination requirements that strongly depend on the cluster size. The second approach is computationally more efficient, thus allowing us to consider larger clusters, and at the same time it accounts for the effects of thermal disorder. Importantly, we have ascertained that the two approaches yield equivalent results.

2.2.3 Computational Methods

The AMBER force field [21] in LAMMPS [22] is used to describe the interactions between all species in the MD simulations. All DFT calculations are carried out using Gaussian09 [23], employing the M06-2X functional with the 6-31++G(d,p) basis set. The M06-2X functional is a highly non-local functional with double amount of non-local exchange (2X), and it is parameterized only for non-metals. We choose M06-2X because it was parameterized to allow for an approximate modeling of vdW interactions at short-range, [24] which is of great importance for our calculations about clusters.

The solution phase free energies are determined using self-consistent reaction field (SCRF) calculations based on the SMD model. SMD is a continuum solvation model based on the

quantum mechanical charge density of a solute molecule interacting with a continuum description of the solvent. In the SMD model a reaction field calculation is performed using the integral equation formalism model (IEFPCM) with radii and non-electrostatic terms from Truhlar and co-workers. [10]

2.3 Results and Discussion

In the following sections, we analyze the calculations of solvation free energies of cations (NH_4^+, $CH_3NH_3^+$) and anions (OH^-, SH^-) using the two different approaches introduced above, and examine how the number of explicit water molecules included in the cluster-continuum model, the inclusion of a counter ion to achieve charge balance in the system, and the placement of this counter ion affects the reliability of the calculations. When carrying out geometry optimization, we vary the number of explicit water molecules from 1 to 12. For the direct MD sampling approach as many as 48 explicit water molecules are included in the calculations.

2.3.1 Geometry Optimization Approach.

In this approach, the maximum number of water molecules that we include in a cluster is 12 because for small solute molecules, 12 water molecules are about two concentric coordination layers around it. The significant local solute-solvent interactions are well considered. Moreover, the computational cost of geometry optimization involving van der Waals interactions increases significantly with the number of molecules in the system. The solution-phase free energy of the solute A can be calculated from the difference between the solution-phase free energy of the cluster $A(H_2O)_n$ and that of the water cluster $(H_2O)_n$.

2.3.1.1 Water Clusters.

Most stable geometries of the water clusters $(H_2O)_n$, with n = 2 to 12, in the solvent configuration, as obtained using M06-2X/6-31++G(d,p) calculations are shown in Figure 2. The structures of these optimized water clusters are similar to those in the gas phase predicted by others, such as a linear dimer and a cyclic trimer, tetramer, and pentamer structure. When there are five or fewer water molecules included in a cluster, the minimum-energy geometries of the clusters tend to be two-dimensional regular polygons, as this maximizes the number of hydrogen bonds. The prism hexamer marks the transition from two-dimensional cyclic structures to three-dimensional structures. [25] Bryantsev claimed the cyclic hexamer to be the most stable geometry for a cluster of six water molecules, which should not be the case because the cyclic hexamer only have six hydrogen bonds compared to the nine hydrogen bonds in the prism hexamer. Notice that we cannot observe any three-dimensional structures as the most stable configurations when n<6 because the bond angle of the water molecule is about 109°, the out-of-plane water molecule cannot form more than one hydrogen bond with the in-plane water molecules. The structures of the octamer and decamer are the same as those described in Bryantsev's work. [15]

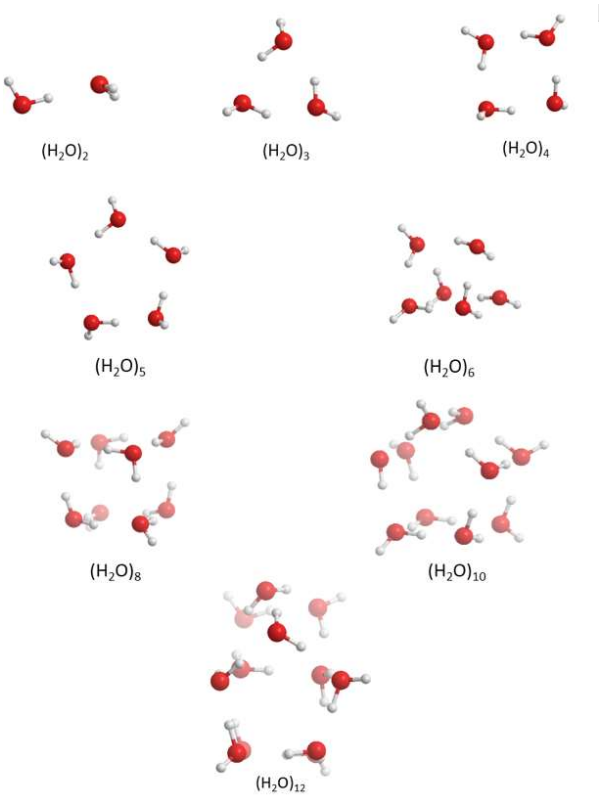


Figure 2-2. Solution-phase structures of water molecule clusters containing 2 to 12 water molecules.

The calculated solution-phase energies and the solvation free energies of water clusters are shown in Table 1. When the number of water molecules included in the cluster increases, the solvation free energy of water clusters per molecule decreases because of the commensurate number of hydrogen bonds that form within the clusters. Bryantsev and coworkers observed similar trends. [15]

Table 2-1. Solution-phase free energies (hartree), solvation free energies (kcal/mol) of water clusters, and solvation free energies (kcal/mol) of water clusters per molecule.

n	$G_s((H_2O)_n)$	$\Delta G_{solv}((H_2O)_n)$	$\Delta G_{solv}((H_2O)_n)/n$
1	-76.41	-9.03	-9.03
2	-152.82	-14.77	-7.39
3	-229.24	-13.68	-4.56

4	-305.66	-16.67	-4.17
6	-458.50	-19.30	-3.22
8	-611.34	-19.11	-2.39
10	-764.17	-23.41	-2.34
12	-917.01	-28.75	-2.39

2.3.1.2 Test Solute Molecules

In the following we discuss the results of solvation free energy calculations for four test molecular groups, NH_4^+, $CH_3NH_3^+$, HS^- and OH^-, i.e., two cations and two anions. These molecules are sufficiently small to be completely enveloped in water molecules at an affordable computational cost, while still providing a meaningful variation in size. The cations share a common protonation/deprotonation group, namely amine. The anions are structurally simple but vary in size and electronegativity of the chalcogen. Finally, for all molecules solvation free energies are available in the literature for comparison.

2.3.1.3 Ion/Water Cluster.

For each ion/water cluster $A(H_2O)_n$ (n=1 to 12), four to six initial conformations are extracted from the MD simulations. We keep the n nearest water molecules around A and eliminate the rest. The final solution-phase free energy is the average of these conformations after geometry optimization. Compared to previous studies where only one conformation was used, [26] we find that taking the average of more than one conformation improves the reliability of the solvation free energy calculation. Such an approach was also used to successfully calculate the acid dissociation constants of dicarboxylic acids by Marenich. [27]

Figure 3 shows the distribution of water molecules around the central solute NH_4^+. Each peak represents one concentric layer of water molecules. The first two concentric layers contain 12 water molecules, which is the reason why we choose n = 1 to 12.

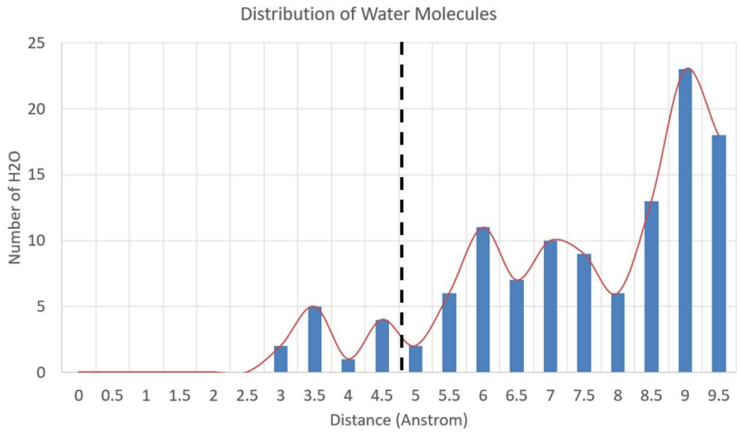


Figure 2-3. Distribution of water molecules around the solute ion NH4+. The black dashed line is the cutoff for the closest two concentric layers.

Examples of optimized solution-phase geometries of cluster $NH_4^+/(H_2O)_n$, n = 1 to 12, as obtained using M06-2X/6-31++G(d,p) calculations are shown in Figure 4. And their corresponding electronic energies are summarized in Table 2.

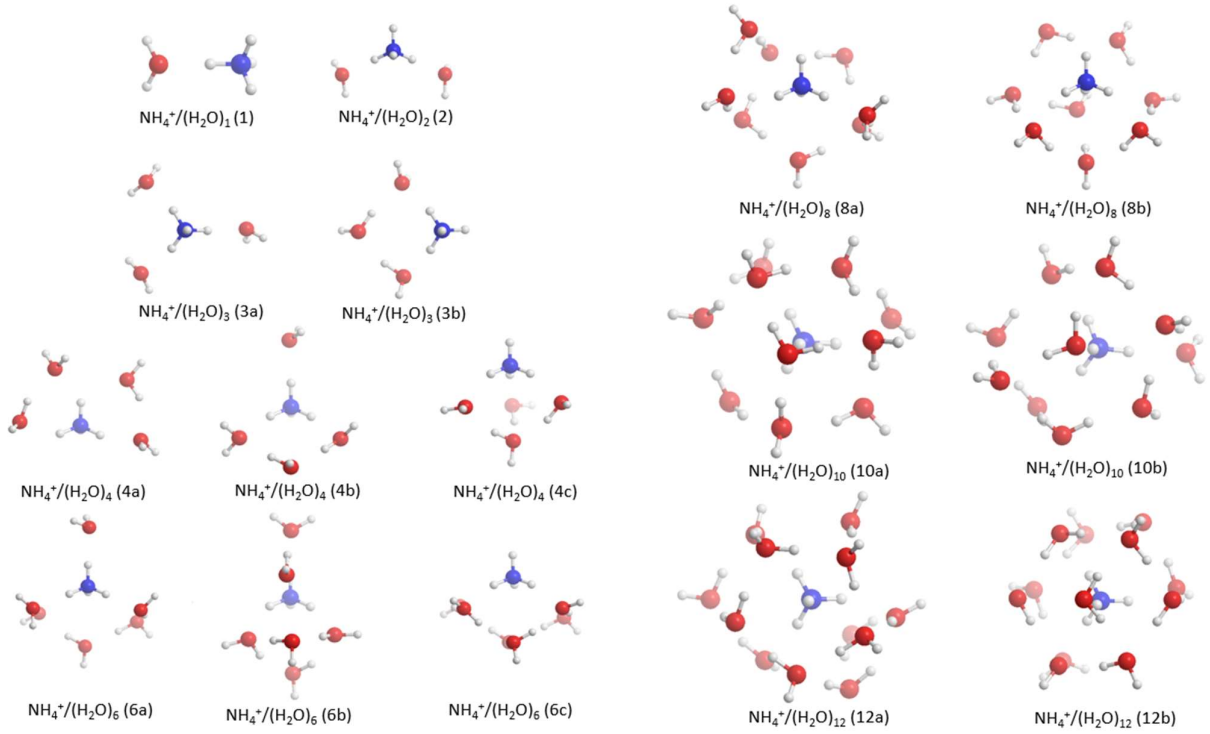


Figure 2-4. Solution-phase structures of NH4+/(H2O)n clusters containing 1 to 12 water molecules. For each number of water molecules, the electronic energy of the cluster decreases from left to right.

Table 2-2. Solution-phase free energies (hartree) of NH4+/(H2O)n clusters; a, b, c correspond to different conformations of the solvation cluster as shown in Fig. 4.

n	$G_s((NH_4^+H_2O)_n)\ (a)$	$G_s\ (b)$	$G_s\ (c)$
1	-133.41421		
2	-209.83217		
3	-286.24427	-286.24967	
4	-362.66316	-362.66491	-362.66783
6	-515.49830	-515.50068	-515.50788
8	-668.34189	-668.34365	
10	-821.18180	-821.18255	
12	-974.01937	-974.02124	

The low-energy structure has been previously reported by Wang et al. [28–30] Our results for small ammonium clusters (n=1-3) in the solution agree with the previous work by Pickard, [29] but above n = 4 we observe different low-energy structures. Instead of the previously reported cluster where the four water molecules bond to one hydrogen each on the ammonium cation, [29] we find $NH_4^+/(H_2O)_4$ has the minimum-energy structures shown in Fig. 4 (4c), where the water molecules tend to form more hydrogen bonds among each other. If we compare the three low-energy structures for $NH_4^+/(H_2O)_4$. Fig. 4 (4a) has four hydrogen bonds, results in the highest electronic energy. Fig. 4 (4b) has five hydrogen bonds with lower electronic energy and Fig. 4 (4c) has six hydrogen bonds and the lowest electronic energy. The minimum-energy structures for larger ammonia clusters $NH_4^+/(H_2O)_n$ (n>4) follow the same rule. Clusters have lower energies if they form more hydrogen bonds.

One of the advantages of our methodology is that the solvation free energy of the charged solute can be calculated directly from the solution-phase energies of the clusters and the gas-phase energy of the solute, we do not need to calculate the gas-phase cluster formation energy. This helps to reduce the amount of calculation and further reduce the error. We calculate the gas-phase energy of the NH_4^+ ion, $E_g(NH_4^+)$, as -56.867 hartree using the same functional and basis set M06-2X/6-31++G(d,p). Combining the solution-phase energies of $(H_2O)_n$ and $NH_4^+/(H_2O)_n$, we get the solvation free energy of NH_4^+ ion using Equation 7 shown in Figure 5.

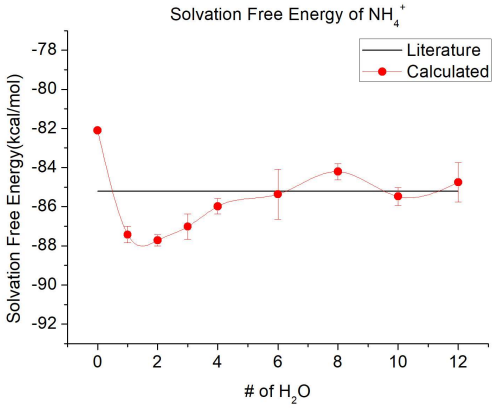


Figure 2-5. Solvation free energy of NH4+ ion as a function of the number of water molecules in the cluster. The black line is the experimental value from the Minnesota solvation database. [31, 32] The red line is our calculated result, where the red dots are the average solvation free energy for each number of water molecule, and the error bar represents the standard error.

The label $n = 0$ indicates that the solvation free energy is calculated using the traditional continuum model. It is about 3 kcal/mol higher than the experimental value. After we add one water molecule to the ammonia ion, the solvation free energy drops below the experimental value by a significant amount, i.e., it increases in magnitude as a result of the formation of hydrogen bonding in configurations that permit unobstructed access of water molecules to the solute molecule. As we increase the number of water molecules included in the cluster, the solvation free energy increases and, gradually approaches the experimental value. When the size of the cluster exceeds four water molecules, the calculated results start to converge and oscillate around the experimental value. Therefore, the minimum number of explicit water molecules one should include in the ammonia water cluster is four. At and beyond this number the calculated solvation energy of the ammonia is sufficiently close to the experimental value -85.2 kcal/mol to be deemed an accurate result.

We apply the same method and obtain the corresponding results for $CH_3NH_3^+$, HS^- and OH^- ions as shown in Figure 6. For each case, the solvation free energy is higher than the experimental value when evaluated using the continuum model, and drops as soon as explicit water molecules are included in the structure. Values start to converge when the size of the ion/water cluster increases and contains between four and eight water molecules. The converged values are in good agreement with the experimental solvation free energy in all cases, except for the OH^- ion. For this ion we obtain a value that is about 7.3 kcal/mol below the experimental solvation free energy. The reason for this discrepancy is further analyzed below. The number of explicit water molecules needed for the $CH_3NH_3^+$ ion is around eight, which is larger than for the other cases ($n \approx 4$). This is to be expected simply based on the larger size of the $CH_3NH_3^+$ ion compared to those of the other three ions. More water molecules are needed to adequately account for all possible local solute-solvent interactions.

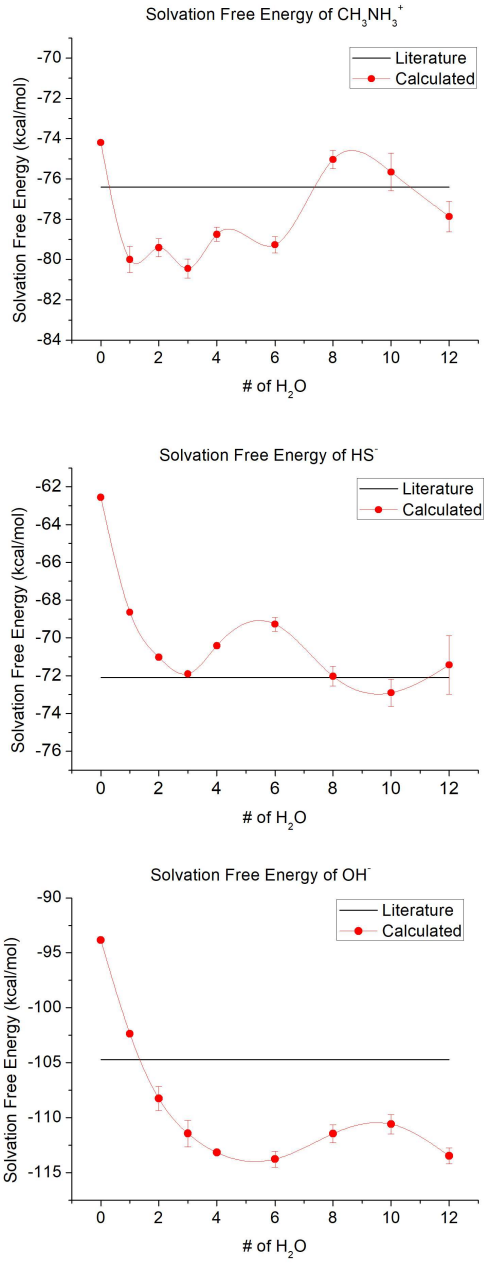


Figure 2-6. Solvation free energy of CH3NH3+, HS- and OH- ions as a function of the number of water molecules in the cluster.

The geometry optimization approach is very accurate for calculating the solvation free energy of small ions, because the configuration necessary to satisfy all local bonding requirements can be achieved with a relatively small water cluster. The final converged values yield a very good estimate for the solvation free energy. However, when the solute ion gets larger, we find

that more and more explicit water molecules need to be added to achieve convergence with this calculation approach. This significantly increases the computation time. Take for example the ammonia ion, the CPU time (eight processors) to optimize the geometry of $NH_4^+/(H_2O)_1$ cluster is about 6 minutes. But it takes at least 3 days of CPU time (optimization time varies for different initial structures) to finish the geometry optimization of large $NH_4^+/(H_2O)_{12}$ cluster. In order to treat larger solute ions we endeavored to find a more efficient approach, in which the computational effort is significantly reduced.

2.3.2 Direct MD Sampling Approach.

The most time consuming aspect of the geometry optimization approach is to determine the minimum-energy configuration, and this process becomes increasingly more costly the larger the solvation cluster has to be. Instead, for each cluster size, we only carry out single point calculations at the M06-2X/6-31++G(d,p) level to get two groups of aqueous-phase energies for a large number of water clusters and ion-water clusters, sampled from the MD trajectories of a pure water system and a system containing the solute surrounded by water, respectively (Fig.7 (a) and (b)). It is to be expected that the free energies calculated for the same clusters in their optimized geometries lie just beyond the low-energy tails of these distributions, i.e., quantities that cannot be pinpointed with great certainty. On the other hand, the target solvation energy is obtained as the difference between the energies of the water and water-ion cluster. Hence, for as long as the standard deviations in the two distributions are comparable, the same difference can be obtained by subtracting the average values of the two distributions from each other.

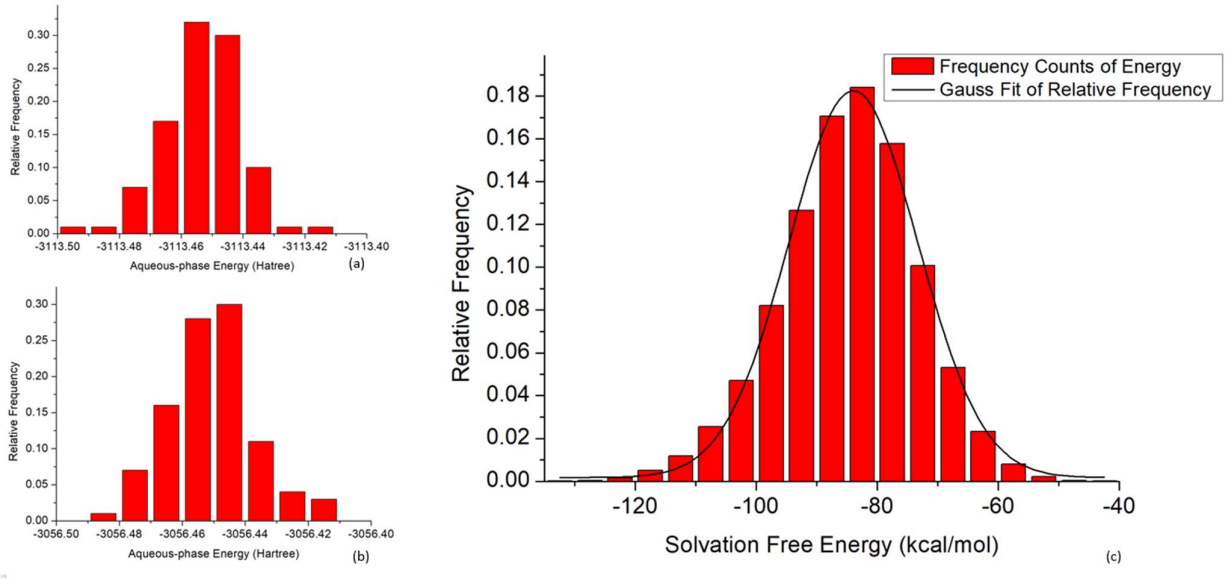


Figure 2-7. Aqueous-phase energy of NH4+/(H2O)40 cluster (a), aqueous-phase energy of (H2O)40 cluster (b), and the solvation free energy distribution of NH4+ calculated from these clusters (c).

A Shapiro-Wilk's test ($p>0.05$) of the histograms in Fig. 7 (a) and (b) shows that these energies approximate normal distributions very well, confirming the visual impression. Furthermore, with 0.0123 Hartree and 0.0131 Hartree, respectively, the variances of the two distributions are very similar. They differ by only 0.5 kcal/mol, which is well within the uncertainty attributed to achieving convergence when using the above geometry optimization approach. Consequently, instead of calculating the difference between estimated tail end values of the distributions in Fig. 7 (a) and (b), we calculate the average difference between the two distributions, for which we employ the statistical equivalence that the average of all differences among values in a distribution is equal to the difference in the averages. Take for example the n=40 case: 100 cluster configurations are sampled from different moments in time of the MD simulations for both $(H_2O)_{40}$ clusters and $NH_4^+/(H_2O)_{40}$ clusters, allowing us to calculate 100 aqueous-phase energies for each of the two clusters (Fig.7 (a) and (b)). Subtracting each energy of the $(H_2O)_{40}$ clusters from each energy of the $NH_4^+/(H_2O)_{40}$ clusters yields 10,000

(100×100) solvation free energy values. The distribution of these values for the ammonia ion embedded in 40 explicit water molecules is shown in the form of a histogram in Fig. 7 (c). The average value of this distribution is equal to the difference between the averages of distributions 7(a) and 7(b). [33] In addition, the variance in the distribution provides a measure of the thermal effects at finite temperatures.

The average value of the distribution in Fig. 7 (c) is a better approximation of the solvation free energy for NH_4^+ calculated from the n=40 case compared to taking the difference of the minimum-energy aqueous-phase configurations of the two clusters calculated using the above geometry optimization approach. This is because with increasing cluster size it becomes more difficult to ascertain that the minimum-energy configuration has indeed been achieved during geometry optimization. For complex structures, the configuration could likely become stuck in a local energy minimum. Furthermore, the MD sampling approach allows us to significantly increase the number of explicit water molecules included in the cluster. As a representative example, Figure 8 shows the average solvation free energy of ammonia ion as a function of the number of water molecules in the cluster generated by the direct MD sampling method up to n = 48.

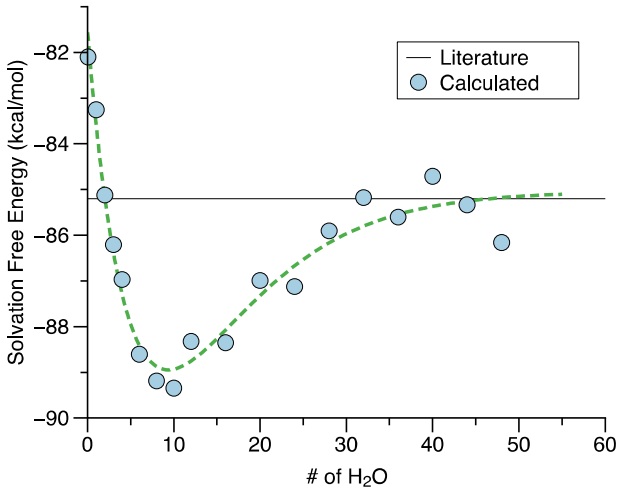


Figure 2-8. Solvation free energy of NH4+ ion as a function of the number of water molecules in the cluster calculated with the direct MD sampling method. Fit lines serve as a guide to the eye. The initial drop is fitted using a polynomial, whereas the approach towards the convergence value is fitted using an exponential function.

It can be seen that the solvation energy first drops with the increase of the number of water molecules. It reaches its minimum at 10 water molecules and then increases until 28 water molecules are included in the cluster. At this point the solvation free energy starts to fluctuate about the convergence value of about -85.5 kcal/mol, which again, is very close to the experimental value -85.2 kcal/mol.

Compared to the geometry optimization approach, the computation time for the solvation free energy of the $NH_4^+/(H_2O)_{12}$ cluster is about 200 times faster. This direct MD sampling method allows us to calculate the solvation free energy of large clusters, such as $NH_4^+/(H_2O)_{48}$. The computation time of single point calculation for the $NH_4^+/(H_2O)_{48}$ cluster is still about 10 times faster than that of geometry optimization for the $NH_4^+/(H_2O)_{12}$ cluster.

We apply the same approach to the other ions. We observe convergence when the number of water molecules gets large in all cases. The converged solvation free energy for the $CH_3NH_3^+$ ion is about -73.2 kcal/mol, 3.2 kcal/mol higher than the experimental value. The converged

solvation free energy for the OH^- ion is about -95.6 kcal/mol, 9.1 kcal/mol higher than the experimental value. However, the converged solvation free energy for the HS^- ion is about 35.2 kcal/mol higher, which is quite different from the experimental value and suggests a procedural insufficiency in our calculations. Whether we perform a geometry optimization or MD trajectory sampling, seemingly results for the small negatively charged ions exhibit the strongest deviations from experimental values.

We consider two possible causes for this discrepancy - the fact the dielectric constant in the hybrid model may differ from that used for the effective medium in the continuum model, and the fact that the requirement for charge neutrality in the system had so far been ignored in our approach, as well as in the literature. The calculations in solution phase above are all using the same dielectric constant of -78.5533ε_0, which is the default dielectric constant for continuum water solution model in Gaussian09. Since we are using the hybrid model, which includes explicit water molecules in addition to the effective medium of the default continuum model. The dielectric constant should shift towards the dielectric constant appropriate for an explicit molecular model, which is equal to the dielectric constant of vacuum. Figure 9 shows how the solvation free energy changes with different dielectric constant of the NH_4^+ and HS^- system.

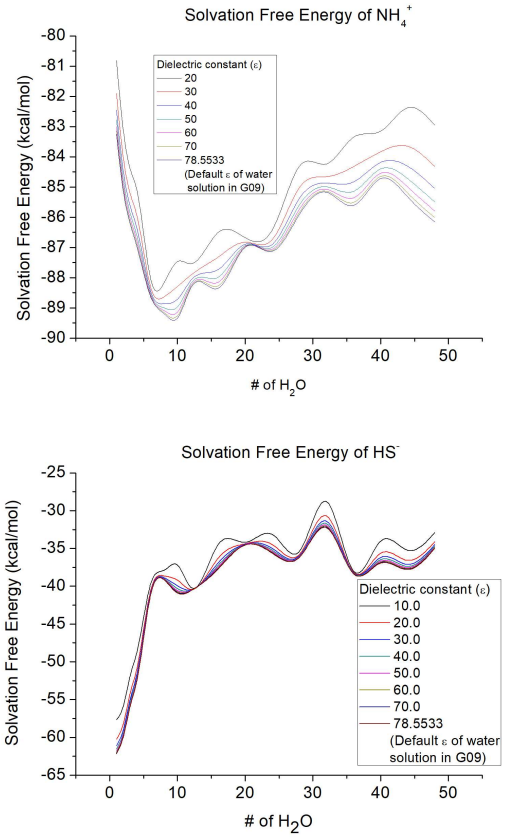


Figure 2-9. Solvation free energy of OH- ion as a function of the number of water molecules in the cluster calculated with the direct MD sampling method.

For both systems, the dielectric constants have a minor influence on the solvation energy, especially when it is not very different from the default value. Hence, we conclude that this does not explain the large difference between the calculated and the experimental solvation energy of the HS^- particle.

Another possible cause is the need for including the counter ion explicitly into the molecular configuration, so as to achieve charge neutrality in the system. In a real solution, there cannot be only one type of ionic species. The counter ion, which has the opposite charge of the solute ion, must also exist. The position of the counter ion, or the distance between the counter

ion and the central solute ion may have an influence on the resulting solvation energy. In addition to the central solute ion and the water molecules, we add one OH^- and one H_3O^+ as the counter ion to the positive charged and negative charged system, respectively. To begin with, we place the counter ion at a close distance from its counterpart. To calculate the solvation free energy we include the necessary terms for the counter ion in equation 2 to yield

$$\Delta G_{solv}(A) = G_s(A(\mathrm{H}_2O)_nC) - G_s((H_2O)_n) - G_g(A) - \Delta G_{solv}(C) \quad (3)$$

where C is the counter ion OH^- or H_3O^+, for $\Delta G_{solv}(C)$ we use the experimental value, $\triangle G_{solv}(OH^-)$= -104.7kcal/mol and $\triangle G_{solv}(H_3O^+)$= -110.3kcal/mol. [31, 32]

Then we pull the counter ion away from the central ion and repeat the calculation for different distances. In the HS^- ion case, HS^- and H_3O^+ start with a distance around 3.0 Å. Upon increasing this distance, the convergence value of the solvation energy approaches the experimental value, and begins to level out at around 10.0Å. The final result is around -77.1kcal/mol, 5.0kcal/mol lower than the experimental value when we have more than 32 water molecules in the cluster. This is a huge improvement compared to the no counter ion situation. Figure 10 shows a comparison between the solvation energies of HS^- ion without the counter ion, with the counter ion close to the central ion, and with the counter ion far away from the central ion.

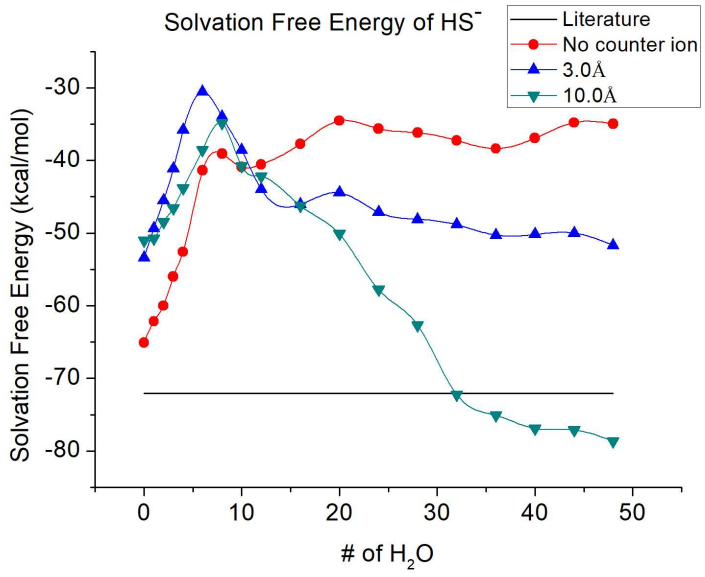


Figure 2-10. Solvation energy of HS- without the counter ion (gray), with the counter ion close to the central ion (yellow) and with the counter ion far away from the central ion (blue), respectively.

For NH_4^+ ion and $CH_3NH_3^+$ ion, adding a counter ion does not result in a significant improvement of the accuracy for the solvation free energy calculation. For the OH^- ion, when the counter ion is located farther away from the central ion, the solvation free energy converges closer to the experimental value when the number of water molecules is larger than 28. The error decreases from 9.1kcal/mol to 4.9kcal/mol, which indicates a good improvement.

As for the calculation speed, if we compare the CPU time (8 processors) that it takes to get the solvation free energy for the $NH_4^+/(H_2O)_{12}$ cluster, the MD sampling approach (20 minutes) is about 200 times faster than the DFT geometry optimization approach (3 days). Based on our findings, the MD sampling approach requires a larger number of explicit water molecules for convergence compared DFT geometry optimization. We attribute this to the fact that in configurations removed from the energy minimum, charge and dipole interactions are as not ideally balanced as they would be with molecular positioning and orientation, and

additional coordination layers are required to compensate for electrical field leakage. To illustrate this, consider how much the low-energy configurations of the $NH_4^+/(H_2O)_6$ cluster in Figure 4 differ from to those of the $NH_4^+/(H_2O)_8$ cluster. This shows how an increasing number of H_2O molecules the environment surrounding the solute provide more effective shielding to the field emanating from the central charged species, and results in the convergence of local electrostatic interactions. However, despite the larger number of water molecules to account for, the CPU time it takes to get the solvation free energy for the $NH_4^+/(H_2O)_{48}$ cluster is about 11 hours, still a much shorter time than is required for smaller size of clusters to relax using the DFT geometry optimization approach.

2.4 Conclusions

A hybrid cluster/continuum model is devised to calculate solvation energy of four ionic molecular groups in aqueous solution. This approach combines a higher accuracy resulting from the consideration of detailed local interactions that are specific to the structure of each solute molecule, while maintaining the computational speed resulting from an effective medium formulation. As the starting configurations we extract water solute clusters of the desired sizes from a large bulk configuration generated using MD simulation, subject to periodic boundary conditions. A systematic variation of the number of water molecules included in these calculations reveals that, depending on the solute size, between four and ten explicit water molecules must be included in the hybrid model in order to account for the most essential local interactions. The larger the solute, and the more complex its structure, the larger is this threshold number of explicit solvent molecules. In a first approach, the cluster geometry is opti-

mized using DFT energy minimization, which yields very accurate solvation energy evaluations about NH_4^+, $CH_3NH_3^+$, HS^- ions and good approximation about OH^-. However, this approach is very time-consuming and can only be reasonably applied to small ions. To encompass a wider range of molecular sizes and structures, we explored a second approach based on DFT single point calculations of a large number of configurations of a given system, sampled along the trajectory from an MD simulation. This procedure yields distributions of solvation energies with comparable variances. Since the desired measure is constructed from the difference between the solvation energies of a water and a water/solute cluster, we can use the most probable value of each distribution instead of the lowest energy value. Eliminating the need for energy minimization in DFT calculations improves the calculation speed and finite temperature can also be accounted for. The solvation energy tends to converge beyond certain size of clusters. Finally, the inclusion of a counter ion to achieve charge balance has proven necessary for the accurate calculation of the solvation energy calculation in the case of some systems like HS^- and OH^-. In that case, it is also important to identify the correct distance between the counter ion and the central ion.

2.5 References

1 Cramer, C.J., '*Essentials of Computational Chemistry*,' 2nd edition (WILEY, 2004), P449.

2 Cossi, M., Barone, V., Cammi, R., and Tomasi, J., 'Ab initio study of solvated molecules: A new implementation of the polarizable continuum model,' *Chem. Phys. Lett.* **255**, 327 (1996).

3 Foresman, J.B., Keith, T.A., Wiberg, K.B., Snoonian, J., and Frisch, M.J., 'Solvent effects.5. Influence of cavity shape, truncation of electrostatics, and electron correlation ab initio reaction field calculations,' *J. Phys. Chem.* **100**, 16098 (1996).

4 Chambers, C.C., Hawkins, G.D., Cramer, C.J., and Truhlar, D.G., 'Model for aqueous solvation based on class IV atomic charges and first solvation shell effects,' *J. Phys. Chem.* **100**, 16385 (1996).

5 Cramer, C.J. and Truhlar, D.G., 'AM1-SM2 and PM3-SM3 parameterized SCF solvation models for free energies in aqueous solution,' *J. Comput. Aided Mol. Des.* **6**, 629 (1992).

6 Cramer, C.J. and Truhlar, D.G., 'An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation,' *Sci.* **256**, 213 (1992).

7 Giesen, D.J., Hawkins, G.D., Liotard, D.A., Cramer, C.J., and Truhlar, D.G., 'A universal model for the quantum mechanical calculation of free energies of solvation in non-aqueous solvents,' *Theor. Chem. Acc.* **98**, 85 (1997).

8 Kelly, C.P., Cramer, C.J., and Truhlar, D.G., 'SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute-water clusters,' *J. Chem. Theory Comput.* **1**, 1133 (2005).

9 Li, J.B., Hawkins, G.D., Cramer, C.J., and Truhlar, D.G., 'Universal reaction field model based on ab initio Hartree-Fock theory,' *Chem. Phys. Lett.* **288**, 293 (1998).

10 Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening,' *J. Chem. Theory Comput.* **5**, 2447 (2009).

11 Zhan, C.G. and Dixon, D.A., 'Absolute hydration free energy of the proton from first-principles electronic structure calculations,' *J. Phys. Chem. A* **105**, 11534 (2001).

12 Pliego, J.R. and Riveros, J.M., 'The cluster-continuum model for the calculation of the solvation free energy of ionic species,' *J. Phys. Chem. A* **105**, 7241 (2001).

13 Zhan, C.G. and Dixon, D.A., 'First-principles determination of the absolute hydration free energy of the hydroxide ion,' *J. Phys. Chem. A* **106**, 9737 (2002).

14 Zhan, C.G. and Dixon, D.A., 'Hydration of the fluoride anion: Structures and absolute hydration free energy from first-principles electronic structure calculations,' *J. Phys. Chem. A* **108**, 2020 (2004).

15 Bryantsev, V.S., Diallo, M.S., and Goddard, W.A., 'Calculation of solvation free energies of charged solutes using mixed cluster/continuum models,' *J. Phys. Chem. B* **112**, 9709 (2008).

16 He, X., Fusti-Molnar, L., Cui, G.L., and Merz, K.M., 'Importance of Dispersion and Electron Correlation in ab Initio Protein Folding,' *J. Phys. Chem. B* **113**, 5290 (2009).

17 Riccardi, D., Guo, H.B., Parks, J.M., Gu, B.H., Liang, L.Y., and Smith, J.C., 'Cluster-Continuum Calculations of Hydration Free Energies of Anions and Group 12 Divalent Cations,' *J. Chem. Theory Comput.* **9**, 555 (2013).

18 Lewis, G.N., Randall, M., Pitzer, K.S., and Brewer, L., '*Thermodynamics*,' 2nd edition (McGraw-Hill: New York, 1961), P272.

19 Ribeiro, R.F., Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation,' *J. Phys. Chem. B* **115**, 14556 (2011).

20 Ho, J.M., 'Are thermodynamic cycles necessary for continuum solvent calculation of pK(a)s and reduction potentials,' *Phys. Chem. Chem. Phys* **17**, 2859 (2015).

21 Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A., 'A second generation force field for the simulation of proteins, nucleic acids, and organic molecules,' *J. Am. Chem. Soc.* **117**, 5179 (1995).

22 Plimpton, S., 'Fast parallel algorithms for short-range molecular-dynamics,' *J. Comput. Phys.* **117**, 1 (1995).

23 Gaussian 09, Revision D.01, M. J. Frisch, G.W.T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

24 Zhao, Y. and Truhlar, D.G., 'The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals,' *Theor. Chem. Acc.* **120**, 215 (2008).

25 Dunn, M.E., Pokon, E.K., and Shields, G.C., 'Thermodynamics of forming water clusters at various temperatures and pressures by gaussian-2, gaussian-3, complete basis set-QB3, and complete basis Set-APNO model chemistries; Implications for atmospheric chemistry,' *J. Am. Chem. Soc.* **126**, 2647 (2004).

26 da Silva, E.F., Svendsen, H.F., and Merz, K.M., 'Explicitly Representing the Solvation Shell in Continuum Solvent Calculations,' *J. Phys. Chem. A* **113**, 6404 (2009).

27 Marenich, A.V., Ding, W.D., Cramer, C.J., and Truhlar, D.G., 'Resolution of a Challenge for Solvation Modeling: Calculation of Dicarboxylic Acid Dissociation Constants Using Mixed Discrete-Continuum Solvation Models,' *J. Phys. Chem. Lett.* **3**, 1437 (2012).

28 Wang, Y.S., Chang, H.C., Jiang, J.C., Lin, S.H., Lee, Y.T., and Chang, H.C., 'Structures and isomeric transitions of NH4+(H2O)(3-6): From single to double rings,' *J. Am. Chem. Soc.* **120**, 8777 (1998).

29 Pickard, F.C., Dunn, M.E., and Shields, G.C., 'Comparison of model chemistry and density functional theory thermochemical predictions with experiment for formation of ionic clusters of the ammonium cation complexed with water and ammonia; Atmospheric implications,' *J. Phys. Chem. A* **109**, 4905 (2005).

30 Morrell, T.E. and Shields, G.C., 'Atmospheric Implications for Formation of Clusters of Ammonium and 1-10 Water Molecules,' *J. Phys. Chem. A* **114**, 4266 (2010).

31 Kelly, C.P., Cramer, C.J., and Truhlar, D.G., 'Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton,' *J. Phys. Chem. B* **110**, 16066 (2006).

32 Marenich, A.V.K., C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database - version 2012*, University of Minnesota, Minneapolis, 2012.

33 Weisstein, E.W. "Normal Difference Distribution." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/NormalDifferenceDistribution.html

Chapter 3

Accurate Acid Dissociation Constant Calculations for Hydrophobes in the Rheology Modifiers

3.1 Introduction

Aqueous polymer systems, such as containing emulsion polymer binders used for coating surfaces, typically draw on thickeners to obtain target rheological properties, like the specific degree of viscosity needed for the proper formulation and application. Traditional thickeners used in such coating emulsions, e.g., cellulose, are non-associating thickeners. [1] However, in the last two decades, a new, improved class of thickeners known as associative thickeners have been found to be superior to cellulose, offering properties including improved flow and leveling in aqueous systems. [2]

Thickeners are called associative because their thickening function involves hydrophobic associations among hydrophobic groups in the thickener molecules, as well as between these hydrophobic groups in and other hydrophobic surfaces. Commonly used associative thickeners have a polymeric backbone with hydrophobic functional groups either attached to or incorporated into the backbone. The backbone can be neutral, such as poly(ethylene oxide) (PEO) and poly(acrylamide) (PAM), or charged, such as poly(acrylic acid) (PAA) or partially hydrolyzed poly(acrylamide) (PHPAM). The hydrophobic groups are typically classified as aliphatic, fluorinated, or aromatic. [3]

Among the commercially available associative thickeners are hydrophobically modified ethylene oxide urethane (HEUR) rheology modifiers, which are water-based polyurethane formulations manufactured by Dow Coating Materials, a division of the Dow Chemical Company. They are added to interior and exterior latex (water-based) paint formulations to control the viscosity of the paint. [4] The structure of the HEUR molecule is shown in Figure 1. It is composed of a polyurethane polyether backbone and two amine hydrophobes on each end of the backbone.

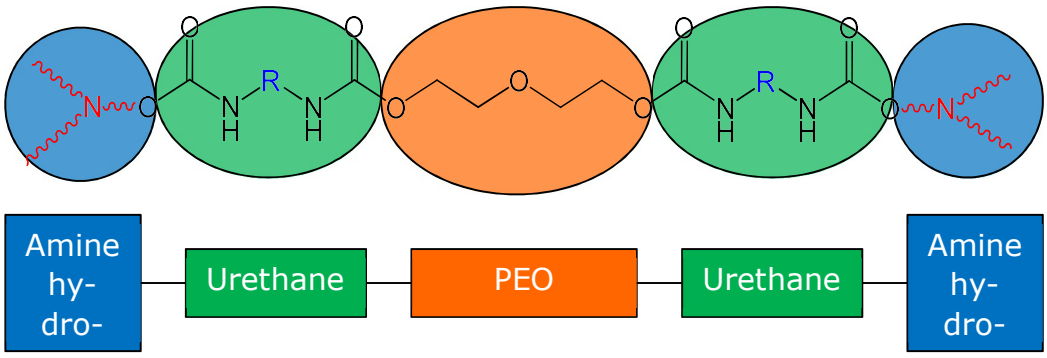


Figure 3-1. Structure of HEUR rheology modifier.

Ethoxylated bis(2-ethylhexy)amine is one of the potential candidates for the hydrophobe group on HEUR molecules. It is a tertiary amine with a NCO reactive group as shown in Figure 2. Indeed, ethoxylated bis(2-ethylhexy)amine provides a unique mechanism for controlling the thickening function of the HEUR molecules, because, depending on the acidity of the aqueous solution, the nitrogen atom on the hydrophobe molecule can either protonate of deprotonate. In a basic environment, most of the ethoxylated bis(2-ethylhexy)amine remains in its deprotonated form, which is more hydrophobic and thereby promotes attraction of the amine to the latex particle surface, effectively anchoring the HEUR molecule to it. This results in a thickening of the suspension. Conversley, when lowering the pH of the solvent, the ethoxylated

bis(2-ethylhexy)amine protonates. The ionized form of the group interacts favorably with the dipole of the water molecule and is more easily solvated, which causes the HEUR molecules to detach from the latex surface, and hence reversing the thickening effect. The governing factor that reveals the point of transition between the two behaviors is the acid dissociation constant K_a, or, as it is reported most commonly, the negative logarithm of this constant, pK_a. In view of further elucidating this process, understanding the underlying mechanisms, and ultimately, achieve a predictive design capability based on a computational approach, it is imperative to be able to accurately calculate the pK_a values for arbitrary hydrophobe molecules.

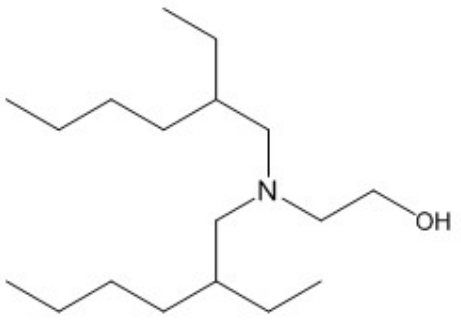


Figure 3-2. Molecular Structure of ethoxylated bis(2-ethylhexy)amine.

Over the past decade, accurate pK_a predictions using quantum chemical methods have been attempted numerous times. The most common approach is to calculate the pK_a using a thermodynamic cycle involving the gas-phase reaction free energy and solvation free energies of all reaction partners. The calculation details for this approach are outlined in the next section. While the calculation of the gas-phase reaction energies is straightforward, as it only involves the reaction partners themselves, the accurate calculation of free energies in solution remains difficult, because the interactions with the solvent must also be accounted for. In early pK_a

calculations of the solvation free energy, [5–7] a continuum model was used to represent the solvent, which simply amounted to a uniform effective medium with a fixed dielectric constant. This model has been shown to deliver solvation free energies with an accuracy of ±1 kcal/mol for neutral solutes, [8] but the mean unsigned errors for ionic species are around 4 kcal/mol. [9] Since a variation in the solvation free energy by 1.36 kcal/mol results in an unit change of the pK_a value, [7] an error of 4 kcal/mol in the solvation free energy calculation is too large to allow for reliable prediction of pK_a values. Consequently, in more recent investigations, the continuum model has been refined by adding explicit solvent molcules to directly account for local solute solvent interactions that prominently contribute to the solvation energy. [10–12] This cluster-continuum model generally yields more accurate results. Both approaches are well documented in a recent review article by Ho and Coote. [13] However, there are no reported studies of pK_a calculations that account for the influence of the local environment beyond explicit solvent molecules. In this study, we focus on the accurate pK_a prediction of ethoxylated bis(2-ethylhexy)amine, considering its local environment, not only explicit water molecules but also a fragment of the latex polymer that the hydrophobe could be in contact with. Using the computational procedure described in our previous chapter, which involves a combination of density functional theory (DFT) calculations and molecular dynamics (MD) simulations, we first ascertain convergence in the energy calculations, i.e., establish the minimum number of explicit water molecules and the shortest latex segment needed to obtain pK_a values that no longer change in a considerable way upon further increasing the complexity of the structure. We then examine in detail the most significant contributions to the calculated pK_a value and

document the extent to which different computational strategies improve the results. Finally, we analyze our findings in terms of the effectiveness of the hydrophobe molecular design.

3.2 Theory and Methodology

3.2.1 pK_a Calculations

For acid dissociation reaction

$$HA^+_{(aq)} \xrightarrow{\Delta G^\circ_{deprot,aq}} A_{(aq)} + H^+_{(aq)} \quad (1)$$

the pK_a is defined as

$$pK_a = -logK_a, \quad (2)$$

and since at equilibrium, the standard free energy of deprotonation of HA^+ in the aqueous phase can be written as

$$\Delta G^\circ_{deprot,aq} = -RTlnK_a = -2.303RTlogK_a \quad (3)$$

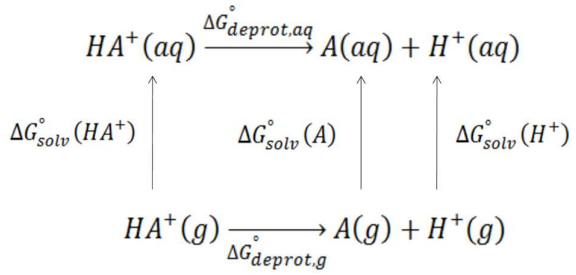
The pK_a of an acid HA^+ is then given by

$$pK_a = \frac{1}{2.303RT}\Delta G^\circ_{deprot,aq} \quad (4)$$

where R is the ideal gas constant and T is the temperature, RT is equal to 0.593 kcal/mol at room temperature (298 K). To compute the free energy associated protonation-deprotonation reaction in aqueous solution it is impractical and unnecessary to replicate the naturally occurring atomic-scale rearrangements during this process. Instead, it is sufficient to consider the beginning state (hydriphobe and proton) and ending state (protonated hydrophobe) of this reaction. Calculating the free energy change associated with the protonation of the isolated hydorphobe molecule, which corresponds to the reaction in the gas phase, is straightforward, as it only involves the reaction partners themselves. To obtain the corresponding quantity for

the reaction occurring in aqueous solution requires the additional knowledge of the solvation free energies of reactants and product. This can be illustrated using the thermodynamic cycle shown in Scheme 1.

Scheme 2-1. Thermodynamic cycle for pK_a calculation.



In this thermodynamic cycle, the upper equation is the acid dissociation reaction in the solution as shown in Eqn. 1 and the lower equation is the corresponding reaction in the gas phase. The bridges that connect these two reactions are the solvation processes of every participant in the reaction. Since the sum of the Gibbs free energy following a complete thermodynamic cycle must vanish, the standard deprotonation free energy in the aqueous phase can be written as:

$$\Delta G^\circ_{deprot,aq} = \Delta G^\circ_{deprot,g} + \Delta G^\circ_{solv}(A) + \Delta G^\circ_{solv}(H^+) - \Delta G^\circ_{solv}(HA^+) \quad (5)$$

where $\Delta G^\circ_{deprot,g}$ is the standard gas-phase deprotonation free energy, also known as the gas-phase basicity, *GB*, and ΔG°_{solv} represents the solvation free energy.

3.2.2 Gas-phase Basicity

From the bottom equation of the thermodynamic cycle in Scheme 1 we know that

$$GB = G_g(A) + G_g(H^+) - G_g(HA^+) = \left(H_g(A) - H_g(HA^+)\right) - \mathrm{T}\left(S_g(A) - S_g(HA^+)\right) + G_g(H^+) = \left(U_g(A) - U_g(HA^+)\right) - \mathrm{T}\left(S_g(A) - S_g(HA^+)\right) + G_g(H^+) \quad (6)$$

Assuming the thermal contribution to the energies of A and HA^+ cancel out, we only care about the energy difference at 0 K. The electronic energy and zero-point energy (ZPE) are a realistic measure of the energy difference. [14] The gas-phase basicity is then represented as

$$GB = \left[E_g(A) + \mathrm{ZPE}(A)\right] - \left[E_g(HA^+) + \mathrm{ZPE}(HA^+)\right] + G_g(H^+) \quad (7)$$

where $E_g(A)$ and $E_g(HA^+)$ are the gas-phase electronic energies of the deprotonated and protonated forms of the molecules, which are calculated at the optimized geometry from DFT calculations, and $\mathrm{ZPE}(A)$ and $ZPE(HA^+)$ are their respective zero-point energies which can be obtained from the vibrational frequency calculations using the DFT method. The free energy term for H^+, $G_g(H^+)$, can be calculated using the standard equations of thermodynamics and the Sackur-Tetrode equation. [15] A proton contains no electronic, vibrational or rotational energy. Only translational energy contributes to the internal energy of the proton $U_g(H^+)$, which is equal to 3/2RT, i.e., 1/2RT for each translational degree of freedom. The enthalpy of the proton, $H_g(H^+) = U_g(H^+) + PV = U_g(H^+) + RT = 5/2RT$, or 1.48 kcal/mol at 298K. Use of the Sackur-Tetrode equation yields the entropy, $TS_g(H^+) = 7.76\ kcal/mol$. Finally since $G_g(H^+) = H_g(H^+) - TS_g(H^+)$, the Gibbs free energy of the proton is equal to -6.28 kcal/mol.

All the DFT calculations are carried out using Gaussian09 [16] at the M06-2X/6-31++G(d,p) level. M06-2X functional is a highly non-local functional with double amount of non-local exchange (2X), and it is parameterized only for nonmetals. We chose M06-2X because it is parameterized to allow for an approximate modeling of van der Waals (vdW) interactions at

short range, [17] which is of great importance for our calculations pertaining to molecular clusters when we calculate the solvation free energies.

3.2.3 Solvation Free Energies

From Equation 5 we know that three solvation free energies are needed to calculate the standard deprotonation free energy in the aqueous phase, i.e., the solvation free energies of the neutral molecule, the proton, and the charged molecule.

The widely accepted solvation free energy of the proton, $\Delta G^{\circ}_{solv}(H^{+})$, is equal to -265.9 kcal/mol, which was determined by Tissandier al. in 1998 using correlations between the solvation free energy of neutral ion pairs and experimental ion-water clustering data. [18] Kelly et al. confirmed this value in 2006 using a similar method but a larger data set. [19] The solvation free energy of a proton can also be calculated using the approach described below. However, in this project we did not conduct this part of calculation and used -265.9 kcal/mol for all calculations reported in this paper.

While the continuum model for the calculation of the solvation free energy of the solute molecule yields relative accurate results for the solute in its charge-neutral, the method incurs a significant error when the solute is charged. Therefore limit use of the traditional continuum model to calculating $\Delta G^{\circ}_{solv}(A)$ in water. Assuming that there are no significant changes in the geometry and vibrational energy upon transitioning from the gas phase to the aqueous phase, only electronic contributions are considered in the DFT calculations. In fact, it has been shown that the use of vibrationally corrected free energies of solvation does not necessarily improve the quality of the calculated free energy of solvation, as long as solvation induced changes in the geometry and frequencies are small. [20, 21] We also prove below that the

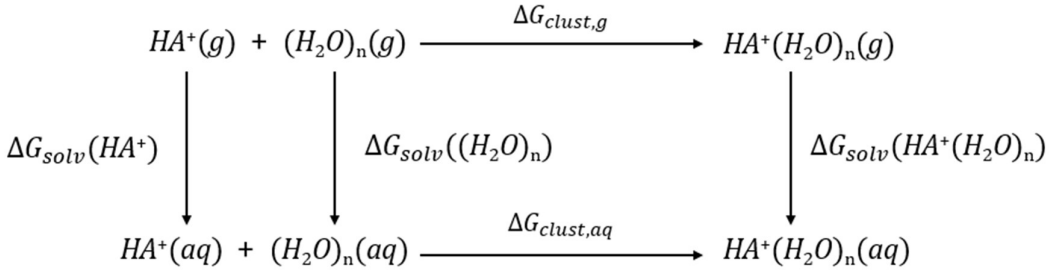
small geometry change and vibrational energy difference indeed have negligible influence on the solvation free energy calculations of amines. Therefore, the solvation free energy of the neutral solute $\Delta G^{\circ}_{solv}(A)$ can be expressed as:

$$\Delta G_{solv}(A) = G_{aq}(A) - G_g(A) \approx E_{aq}(A) - E_g(A) \quad (8)$$

where $E_{aq}(A)$ and $E_g(A)$ are the electronic energies of A in the aqueous phase and gas phase respectively.

To improve the calculation accuracy when calculating $\Delta G^{\circ}_{solv}(HA^+)$ we use the cluster-continuum model, which includes enough explicit solvent molecules to adequately account for local solute-solvent interactions. The thermodynamic cycle for solvation free energy calculations with the cluster-continuum model is shown in Scheme 2.

Scheme 2-1. Thermodynamic cycle for solvation free energy calculation using the cluster-continuum model.



Assuming again that there are no significant changes in geometry and vibrational energy in going from the gas phase to the aqueous phase, only the electronic contributions of the energy are taken into consideration, we obtain the equation to calculate the solvation free energy of HA^+ as detailed in our previous chapter,

$$\Delta G_{solv}(HA^+) = E_{aq}(HA^+(H_2O)_n) - E_{aq}((H_2O)_n) - E_g(HA^+) \quad (9)$$

where $E_{aq}(HA^+(H_2O)_n)$ is the electronic energy of the solute-water cluster in the aqueous phase, $E_{aq}((H_2O)_n)$ is the electronic energy of the water cluster in the aqueous phase and $E_g(HA^+)$ is the electronic energy of the charged solute in the gas phase.

Similarly, to account for the chemical effects of Latex molecules when the hydrophobe is adjacent to a Latex particle surface, we include a Latex polymer segment explicitly in the molecular configuration and use the cluster-continuum model to calculate both $\Delta G^{\circ}_{solv}(A)$ and $\Delta G^{\circ}_{solv}(HA^+)$. The underlying formalism is essentially identical to that in Equation 9, except that we substitute the water cluster with the Latex polymer segment.

All of the aqueous phase free energies are determined using self-consistent reaction field (SCRF) calculations based on the SMD model. SMD is a continuum solvation model (SM) that treats the solvent as a uniform polarizable medium of fixed dielectric constant having a solute molecule placed in a suitably shaped cavity. The letter "D" in the acronym stands for "density" to denote that the full solute electron density is used without defining partial atomic charges. In the SMD model a reaction field calculation is performed using the integral equation formalism of the polarizable continuum model (IEF-PCM) with radii and non-electrostatic terms from Truhlar and coworkers. [9]

3.3 Results and Discussion

In the following, we calculate the pK_a of the hydrophobe molecule ethoxylated bis(2-ethylhexy)amine using the method introduced above. We describe the selection process that led to the choice of the M06-2X/6-31++G(d,p) level calculations for the determination of pK_a values, which is based on a series of gas-phase basicity calculations for test molecules using

different combinations of functional and basis set. Then we use this functional and basis set to calculate the gas-phase basicity and the solvation free energy of the hydrophobe molecule. When calculating the solvation free energy, we systematically examine how adding explicit solvent molecules and explicit surrounding particle in the environment like the Latex segment affect the reliability of the calculations. Finally we compare our calculated pK_a with the experimental rheology data from the Dow Chemical Company.

3.3.1 Gas-phase Basicity Calculation

3.3.1.1 Functional/Basis Set Selection

First we validate the DFT calculation procedure for determining gas-phase basicity using various test molecules for which experimental data are available. Most of the test molecules share a common protonation/deprotonation group with the hydrophobe molecule, namely amine. The results are summarized in Table 1, and a more intuitive representation is provided in Figure 3.

Table 3-1. Gas-phase basicity for test molecules with different methods in Gaussian09. The experimental values are from "Evaluated gas phase basicities and proton affinities of molecules: An update".[22]

Molecule	Gas-phase Basicity (kcal/mol)					
	Experiment	B3LYP/ 6-31G(d,p)	B3LYP/ 6-31++G(d,p)	CBS-QB3	M06/6-31 ++G(d,p)	M06-2X/ 6-31++G(d,p)
H_2O	159.0	164.9	157.3	162.7	157.6	158.4
NH_3	195.6	202.6	196.9	200.1	196.1	196.2
CH_3NH_2	205.7	212.4	207.6	210.2	206	206.5
$C_2H_5NH_2$	208.5	216.1	211.1	213.	209.4	209.5

				4		
n- $C_3H_7NH_2$	210.1	217.3	212.3	214.6	210.5	210.8
$(CH_3)_3N$	217.3	222.2	218.6	219.8	215.9	217.3
$(CH_3)_2NH$	212.8				212.0	212.9
(n-$C_3H_7)_2NH$	219.7				219.6	220.2

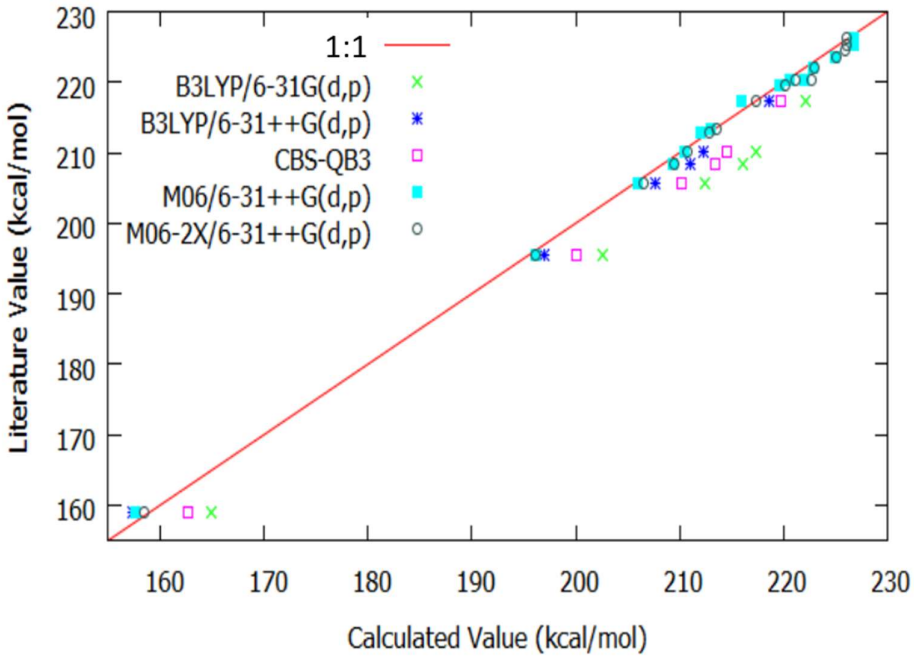


Figure 3-3. Calculated gas-phase basicity values vs. experimental gas-phase basicity values.

We first confirm that diffuse functions are necessary for accurate gas-phase basicity calculations. If we compare the data from B3LYP/6-31G(d,p) (green crosses) with those from B3LYp/6-31++G(d,p) (blue asterisks), we find that adding the diffuse functions significantly decreases the mean unsigned error between the calculated gas-phase basicity and the experimental value from 6.55 kcal/mol to 1.83 kcal/mol. This is because the gas-phase basicity for

the molecules we are testing is a property that is closely related to the protonation and deprotonation of atoms with lone pairs of electrons, like nitrogen or oxygen, in which case adding diffuse functions dramatically improves the calculation result. [23] In addition, it is also necessary to include the diffuse functions in the solvation free energy calculations that involve long-range interactions between the solute and its surrounding molecules. As mentioned in the previous section, the M06-2X functional takes into account of the Van der Waals interaction, which is of great importance in accounting for dispersive interactions for our solvation free energy calculations. By comparing the black circles and the blue asterisks data in Figure 3, we also find that M06-2X functional results in a slightly more accurate gas-phase basicity calculation compared to the B3LYP functional. The mean unsigned error of M06-2X functional with diffuse basis set 6-31++G(d,p) is only 0.49 kcal/mol. Bryantsev et al. conducted similar evaluations of B3LYP and M06-class density functionals for predicting the binding energies of neutral, protonated and deprotonated water clusters, and concluded that the M06-class density functionals yielded more accurate binding energies. [10] Overall, we have concluded that the M062X/6-31++G(d,p) method leads to the most accurate results among these approaches. The gas-phase basicity values we obtain for the test molecules using M062X/6-31++G(d,p) are very close to the experimental values, with an average error of 0.75 kcal/mol. We therefore apply this functional and basis set for both the gas-phase basicity and solvation free energy calculations of the hydrophobe molecules.

3.3.1.2 Gas-phase Basicity of the Hydrophobe Molecule

To calculate the gas-phase basicity, we first determine the optimized geometry and vibrational characteristics of the ethoxylated bis(2-ethylhexy)amine in both deprotonated and protonated

states using standard numerical procedures for minimizing the energy of the configurations. The M062X/6-31++G(d,p) optimized geometries of the neutral ethoxylated bis(2-ethylhexy)amine (deprotonated form) and that of the charged ethoxylated bis(2-ethylhexy)amine (protonated form) are shown in Figure 4.

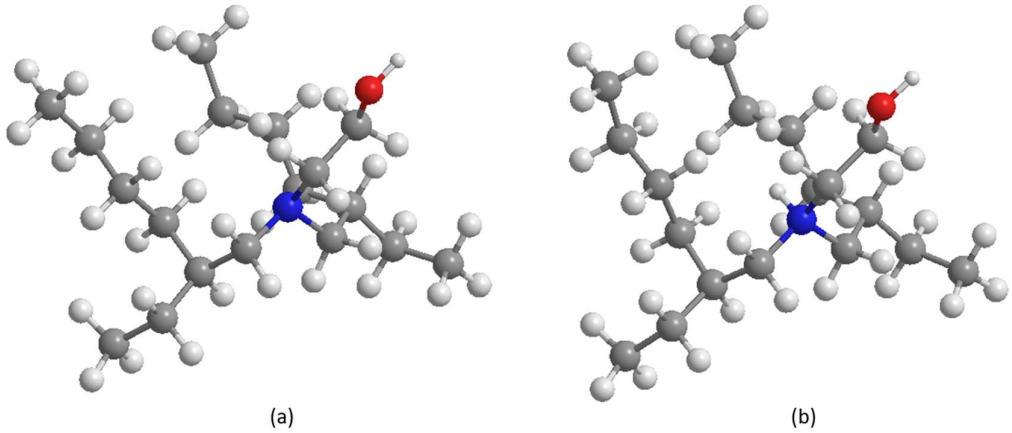


Figure 3-4. Optimized geometry of the ethoxylated bis(2-ethylhexy) amine in deprotonated state (a) and protonated state (b).

It can be seen that the two optimized geometries are quite similar, except that there is an additional proton connected to the nitrogen atom in the protonated form. Accordingly, we expect comparable energies for the two different forms. The calculated electronic and zero-point energies for the deprotonated form is -839.0169 Hartree and 0.5561 Hartree, respectively. Those for the protonated form are -839.4077 Hartree and 0.5721 Hartree, respectively. Based on Eqn. 7, we determine the gas-phase basicity of ethoxylated bis(2-ethylhexy) amine to be 228.91 kcal/mol.

3.3.2 Solvation Free Energy Calculation

3.3.2.1 Continuum model

Generally, the continuum model can deliver accurate solvation free energies for neutral solutes. [8] We have verified this with some of our test molecules. The solvation free energies of the test amines calculated using the continuum model are shown in Table 2.

Table 3-2. Calculated solvation free energies (in kcal/mol) for neutral test molecules using the continuum model. The experimental values are from the Minnesota solvation database. [19, 24]

A. Molecule	B. Experiment	C. Calculated I	D. Calculated II	E. Calculated III
NH_3	-4.29	-4.31	-4.38	-4.54
CH_3NH_2	-4.56	-4.33	-4.53	-4.48
$(CH_3)_2NH$	-4.29	-4.08	-4.46	-4.20
$(CH_3)_3N$	-3.23	-3.37	-3.81	-3.36

Note that besides the column containing the experimental data, Table 2 has three columns for the calculated data. All of them use the same SMD continuum model. Column C assumes that there are no geometry changes upon transitioning from the gas phase to the aqueous phase, i.e. we use the optimized geometry in the aqueous phase for the energy calculations of both phases. Conversely, for the data listed in column D, geometry optimization is carried out separately for both gas and aqueous phase to get the solvation free energy. Finally, column E also does not take the geometry change into account, but it includes the vibrational energy part for the solvation free energy calculation. We can see that for NH_3, the calculated result from column C has a minimum error of 0.02 kcal/mol, for CH_3NH_2 including the geometry changes gives us a slightly more accurate result with an error of 0.03 kcal/mol, for

$(CH_3)_2NH$ adding vibrational energy term yields the most accurate result with an error of 0.09 kcal/mol, and finally for $(CH_3)_2N$ columns C and E have very similar results with an error of 0.14 kcal/mol and 0.13 kcal/mol respectively. The overall mean unsigned error for the three calculation methods are 0.15 kcal/mol, 0.22 kcal/mol and 0.14 kcal/mol, respectively, with negligible differences. Therefore, including the geometry change or the vibrational energy term does not produce a significant improvement for these solvation free energy calculations. When calculating solvation free energies for amines, it is reasonably accurate to apply the continuum model and only consider the electronic energy difference between the gas and aqueous phase using the same geometry that is optimized in the aqueous phase.

We use the same method for the solvation free energy calculation of the ethoxylated bis(2-ethylhexy) amine in the deprotonated state. The aqueous-phase electronic energy is -839.0208 Hartree. Recall that the gas-phase electronic energy that we have calculated in the previous section is equal to -839.0169 Hartree. Using the continuum model, the calculated solvation free energy of the deprotonated ethoxylated bis(2-ethylhexy) amine is thus -2.45 kcal/mol based on Eqn. 8.

We have also done tests with the solvation free energies of the corresponding charged solutes. The results are shown in Table 3.

Table 3-3. Calculated solvation free energies (in kcal/mol) for charged test molecules using the continuum model. The experimental values are from the Minnesota solvation database.[19, 24]

A. Molecule	B. Experiment	C. Calculated I	D. Calculated II	E. Calculated III
NH_4^+	-85.2	-82.1	-82.1	-82.6
$CH_3NH_3^+$	-76.4	-74.1	-74.2	-74.3

$(CH_3)_2NH_2^+$	-68.6	-67.2	-67.4	-67.5
$(CH_3)_3NH^+$	-61.1	-61.6	-61.7	-62.0

Comparing the calculated solvation free energies with one another, the differences are all within 0.5%. This further proves that including the geometry change or vibrational energy difference does not necessarily improve the accuracy of solvation free energy calculations, even for the charged solutes. However, the mean unsigned error increases to 1.8 kcal/mol for the charged test amines. Therefore we endeavor to improve the implicit continuum model by adding explicit surrounding solvent molecules. The calculated solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine using the continuum model is -57.06 kcal/mol. Below we examine how surrounding the hydrophobe with explicit water molecules in the environment affects this value.

3.3.2.2 Explicit Solvent Effect

We first add explicit water molecules around the protonated ethoxylated bis(2-ethylhexy) amine to take the significant local solute-solvent interactions into account. To calculate the solvation free energy with the influence of explicit solvent effect we use Equation 9. In our case solute HA^+ is the protonated ethoxylated bis(2-ethylhexy) amine. $E_{aq}(HA^+(H_2O)_n)$, $E_{aq}((H_2O)_n)$ are the aqueous-phase electronic energies of the solute-water cluster and water cluster, respectively, and $E_g(HA^+)$ is the gas-phase electronic energy of the solute molecule. The underlying theory and computational details can be found in our previous chapter.

We carry out molecular dynamics (MD) simulations starting with a random placement of the protonated ethoxylated bis(2-ethylhexy) amine and water molecules. Figure 5 shows a system with one protonated ethoxylated bis(2-ethylhexy) amine, one OH^- counter ion (to keep the system charge neutral) and 1860 water molecules in a simulation box of 34.8 Å edge length, subject to periodic boundary conditions at 298.15 K, as the initial conditions for the MD simulations. The number of water molecules and the size of the simulation box are chosen so that the configuration is dense and the solute molecules in the periodic boundary condition are far away enough and not affected by each other. During equilibration, the volume of the simulation box is kept constant with a density of 1.20 g/cm^3. The AMBER force field[CORNELL et al., 1995, #61377] in LAMMPS[PLIMPTON, 1995, #58695] is used to describe the interactions between all species in the MD simulations.

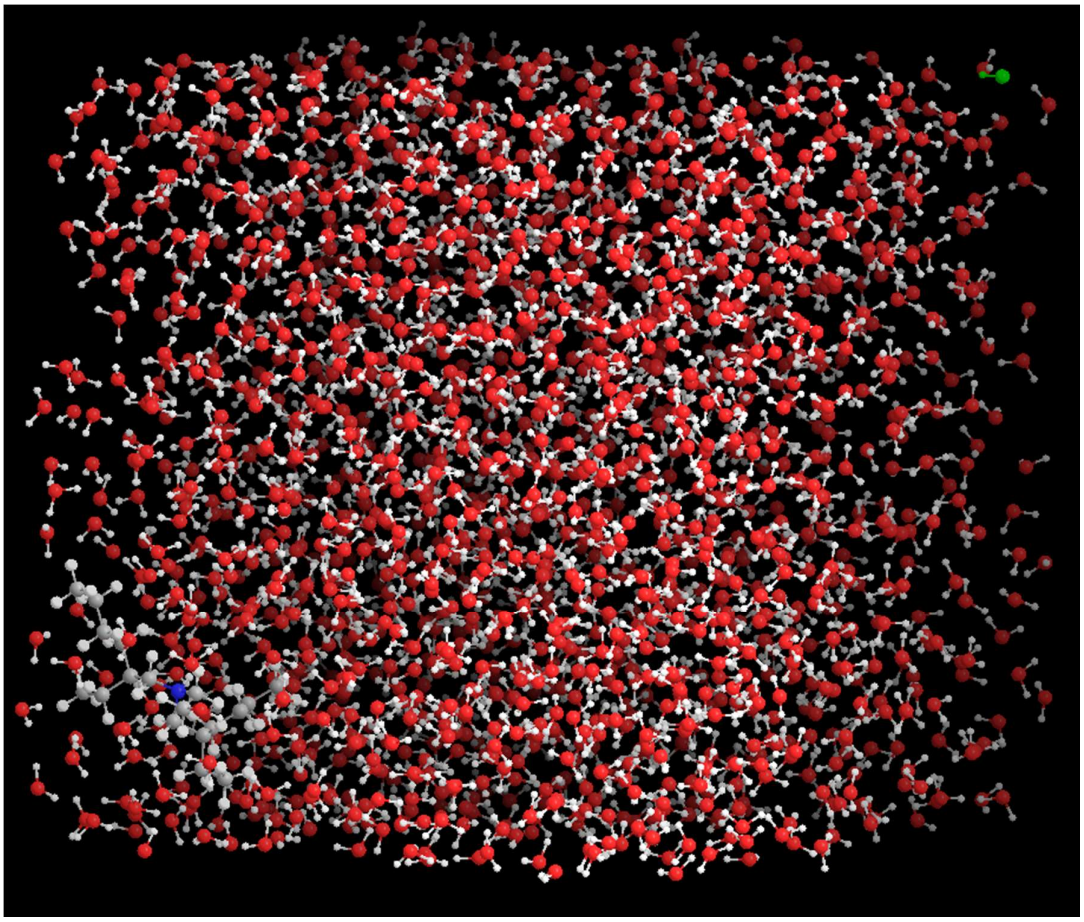


Figure 3-5. MD simulation system of the protonated ethoxylated bis(2-ethylhexy) amine (left bottom), the counter ion (top right in green) and 1860 water molecules.

From this point onwards, we use two different approaches to generate a structural model of the solvation cluster. Our goal is to construct a solvent configuration surrounding the solute molecule that is sufficiently detailed to comprise all essential local molecular interactions. In the first approach, which involves geometry optimization, we extract a cluster that contains as many as six water molecules surrounding the protonated ethoxylated bis(2-ethylhexy) amine from the MD simulation. We keep the water molecules nearest to the protonation/deprotonation site – the nitrogen atom of the ethoxylated bis(2-ethylhexy) amine – and eliminate the rest. We are able to include up to six water molecules because the computational cost of geometry optimization involving van der Waals interactions increases significantly with the

number of molecules in the system. We then optimize the geometry using DFT energy minimization, resulting in the free energy of a fully relaxed cluster at zero Kelvin. For each solute-solvent cluster $HA^+(H_2O)_n$ (n=1 to 6), four to six initial configurations are extracted from the MD simulations. The final aqueous-phase free energy is the average of these configurations after geometry optimization, which improves the reliability of the solvation free energy calculation. Representative examples of optimized aqueous-phase geometries, as obtained using M06-2X/6-31++G(d,p) calculations are shown in Figure 6. The calculated solvation free energies of the protonated ethoxylated bis(2-ethylhexy) amine based on these optimized structures are plotted in Figure 7 as a function of the number of explicit water molecules in the system.

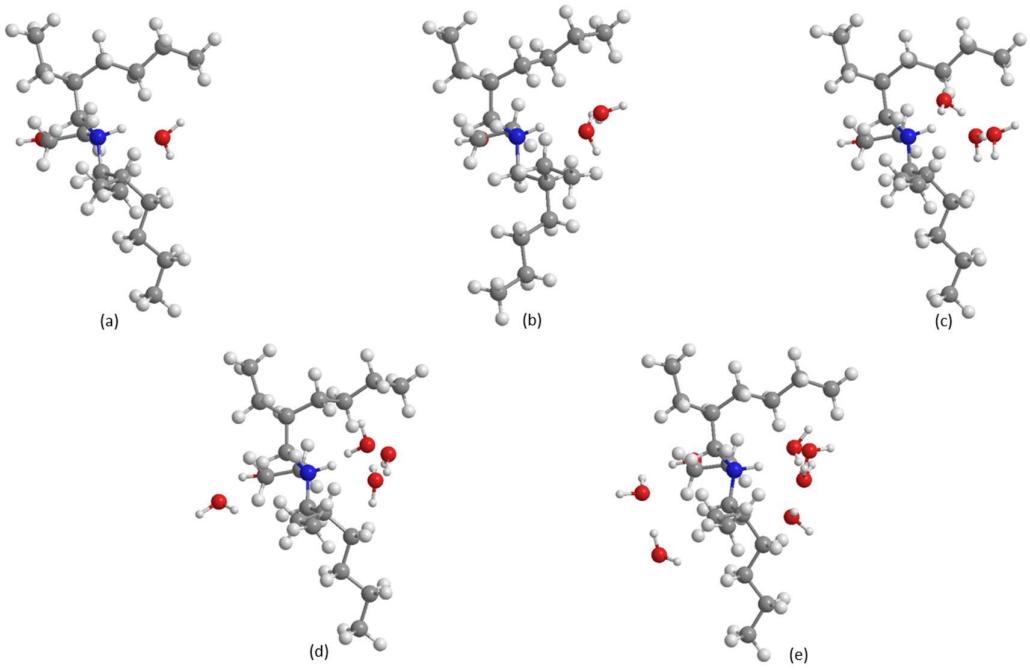


Figure 3-6. Aqueous-phase structures of protonated ethoxylated bis(2-ethylhexy) amine/water clusters containing (a) 1 water molecule, (b) 2 water molecules, (c) 3 water molecules (d) 4 water molecules and (e) 6 water molecules.

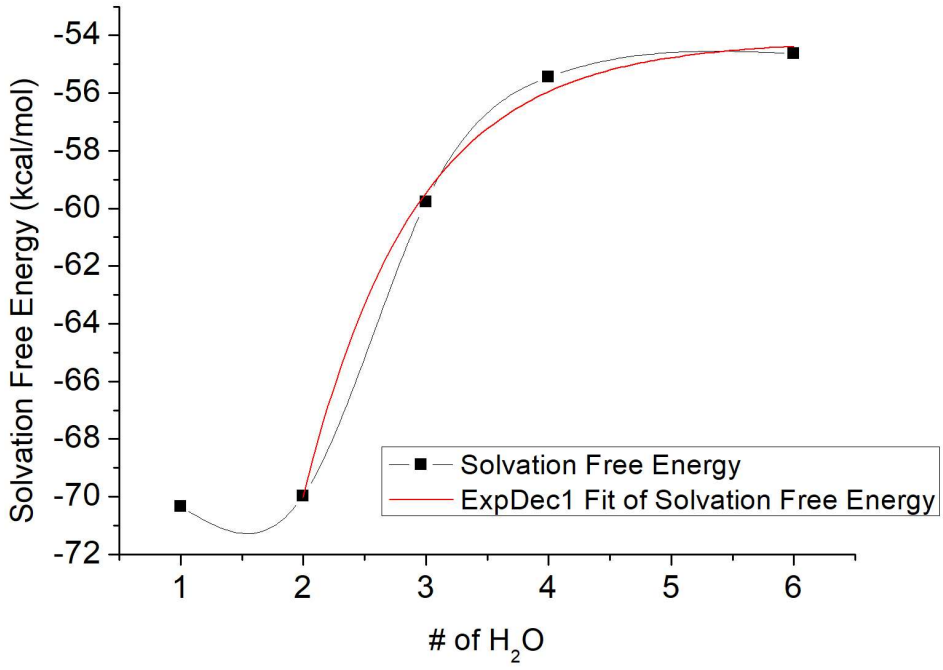


Figure 3-7. Solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine as a function of the number of water molecules in the cluster, calculated by using the geometry optimization approach. The black squares are the average solvation free energy for each number of water molecules, and the red line is an exponential fit for the data when n>1.

From Fig.6 (a) we can see that the stable low-energy configuration for the solute/one-water cluster is the one in which the water molecule forms a hydrogen bond with the nitrogen atom, intrinsically the most effective way to satisfy the local solute-solvent coordination needs. As we increase the number of water molecules included in the cluster, the water molecules wrap around the nitrogen atom to form a more impermeable shield for the field emanating from the protonated site. Figure 7 shows that the solvation energy increases and levels off as the number of water molecules included in the cluster increases. Similar to the approach taken by Bryantsev to determine the solvation free energy of H^+ and Cu^+, [10] we estimate the solvation

free energy of the protonated ethoxylated bis(2-ethylhexy) amine by extrapolating the calculated energies to infinite water molecules using an exponential fit function, which is equal to -54.18 kcal/mol.

The most time consuming aspect of the geometry optimization approach is to determine the minimum-energy configurations, and this process becomes increasingly more costly as the solvation cluster gets larger. For example, the CPU time (eight processors) to optimize the geometry of the cluster containing the protonated ethoxylated bis(2-ethylhexy) amine and six water molecules is about 29 days and 7 hours. In order to treat larger clusters, we must resort to a more efficient approach – the MD sampling approach, in which the computational effort is significantly reduced. Instead of the lengthy geometry optimization, we only carry out single point calculations at the M06-2X/6-31++G(d,p) level to get two groups of aqueous-phase energies for a large number of water clusters and protonated hydrophobe-water clusters, sampled from the MD trajectories of a pure water system and a system containing the solute surrounded by water. By taking the average of the difference distribution from these two groups of energy, we get the solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine as a function of the number of water molecules in the cluster as shown in Figure 8.

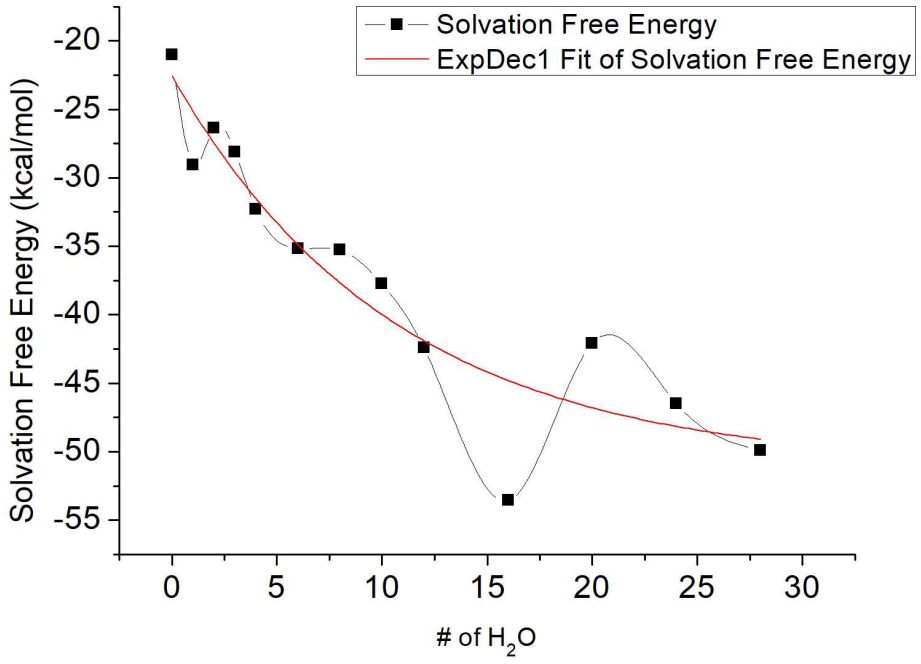


Figure 3-8. Solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine as a function of the number of water molecules in the cluster calculated by using the MD sampling approach. The black squares are the average solvation free energy for each number of water molecules, and the red line is an exponential fit.

It can be seen that the calculated solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine is very high when the number of water molecules included in the cluster is small. It gradually decreases with increasing number of water molecules. Again, extrapolating the observed trend towards the convergence value using an exponential fit, we obtain for the solvation energy of the protonated ethoxylated bis(2-ethylhexy) amine a value of -51.13 kcal/mol, comparable to that calculated using the geometry optimization approach. Based on our findings, the MD sampling approach requires a larger number of explicit water molecules for convergence compared DFT geometry optimization. We attribute this to the fact that in configurations removed from the energy minimum, charge and dipole interactions are as not ideally balanced as they would be with molecular positioning and orientation, and additional

coordination layers are required to compensate for electrical field leakage. However, despite the larger number of water molecules to account for, the CPU time it takes to get the solvation free energy for the cluster of protonated ethoxylated bis(2-ethylhexy) amine and 28-water molecules is about 1 day and 1 hour, which is one and a half orders of magnitude shorter than is required for smaller size of clusters to relax using the DFT geometry optimization approach.

3.3.2.3 Explicit Latex Particle Sample Effect

After studying the effects of explicit solvent molecules on the solvation free energy of the hydrophobe molecule ethoxylated bis(2-ethylhexy) amine, we now examine how the presence of a Latex polymer segment influences its solvation free energy.

First we create a Latex polymer fragment large enough to be significant, but still small enough to accommodate computationally. The main components of the Latex particle are methyl methacrylate (MMA) and butyl arcylate (BA). We use MMA trimer and BA trimer as simple representations for the Latex particle. The geometry-optimized structure of these two trimers are shown in Figure 9. Since both MMA trimer and BA trimer are neutral molecules, we can use the continuum model to get their minimum aqueous-phase energies, i.e., -1077.5608 Hartree and -1313.3568 Hartree, respectively. These energies are used in our solvation free energy calculations later.

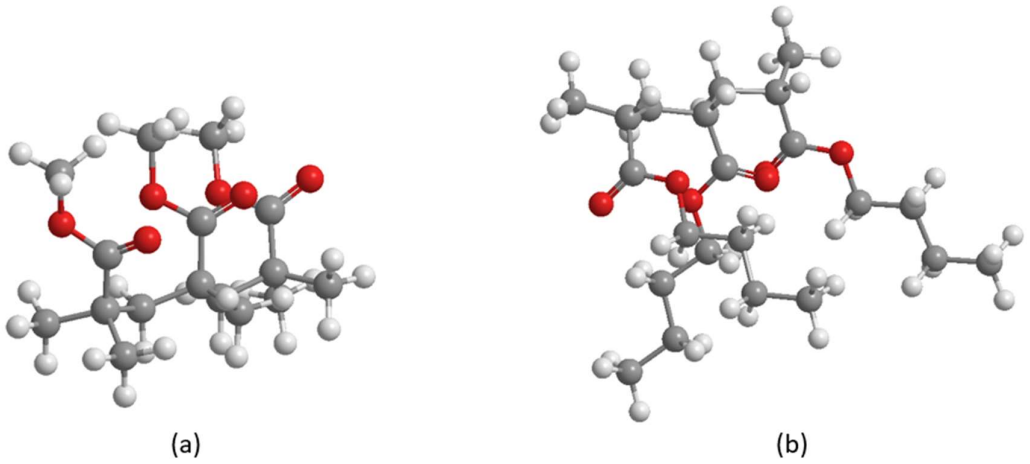


Figure 3-9. Geometry optimized structures of MMA trimer (a) and BA trimer (b).

To calculate the solvation energy of the protonated ethoxylated bis(2-ethylhexy) amine adjacent to an explicit MMA trimer, we use the formalism of Equation 9, but we substitute the explicit water cluster with the MMA trimer. Equation 9 then becomes

$$\Delta G_{solv}(HA^+) = E_{aq}(HA^+(MMA)_3) - E_{aq}((MMA)_3) - E_g(HA^+) \quad (10)$$

Similar to what is described in the previous section, we first carry out MD simulations. This time we start with a protonated ethoxylated bis(2-ethylhexy) amine, a MMA trimer, a counter ion OH^- and 1843 water molecules in the simulation box subject to the periodic boundary conditions. The volume of the box is the same, with an edge length of 34.8 Å, and it remains constant during the equilibration. After the MD simulation, we extract the cluster that contains the protonated ethoxylated bis(2-ethylhexy) amine and the MMA trimer and perform the geometry optimization using DFT method. The configurations before and after the optimization are shown in Figure 10.

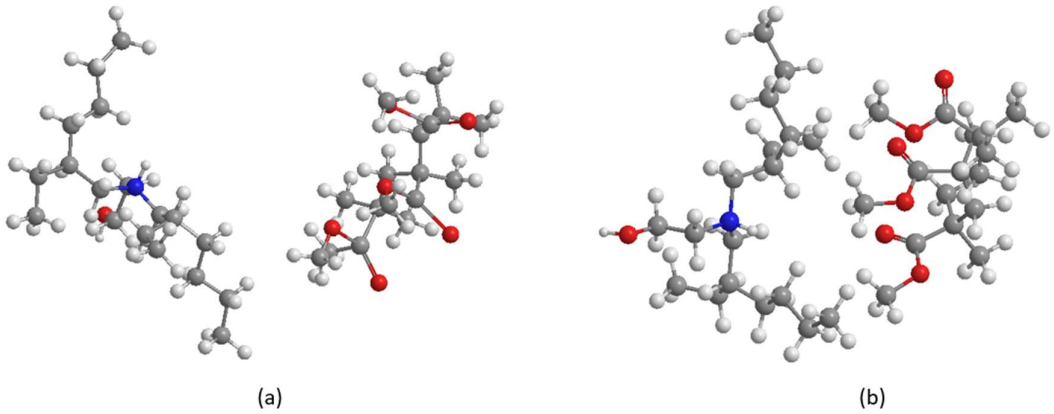


Figure 3-10. Geometry optimized structures of MMA trimer (a) and BA trimer (b).

For the configuration in Fig.10 (b) we determine the aqueous-phase energy of $HA^+(MMA)_3$, $E_{aq}(HA^+(MMA)_3)$ as -1917.0646 Hartree. We then calculate the solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine $\Delta G_{solv}(HA^+)$ using Equation 10 and obtain -60.30 kal/mol.

Using the same method, we also calculated the solvation free energy of the deprotonated ethoxylated bis(2-ethylhexy) amine with the presence of explicit MMA trimer, as well as the solvation free energy of protonated/deprotonated ethoxylated bis(2-ethylhexy) amine with the presence of explicit BA trimer. The results are summarized in Table 4.

Table 3-4. Calculated solvation free energies of the ethoxylated bis(2-ethylhexy) amine with and without the explicit Latex particle sample. A represents the deprotonated form and HA^+ represents the protonated form.

	Without Explicit Latex Sample	With Explicit MMA Trimer	With Explicit BA Trimer
$\Delta G_{solv}(A)$ (kcal/mol)	-2.45	-10.83	-3.26
$\Delta G_{solv}(HA^+)$ (kcal/mol)	-57. 06	-60.30	-52.76
$\Delta G_{solv}(A) - \Delta G_{solv}(HA^+)$ (kcal/mol)	54.61	49.47	49.50

We can see that the presence of the MMA trimer lowers the solvation free energies of both the deprotonated and protonated forms of the hydrophobe, but it influences the deprotonated form more. The presence of the BA trimer also decreases the solvation free energy of the deprotonated form a bit but it increases the solvation free energy of the protonated form, making it less soluble. However, the the solvation energies of the deprotonated and protonated forms of the hydrophobe interacting with these Latex fragments are very close: 49.47 kcal/mol and 49.50 kcal/mol, respectively, which indicates that the two types of Latex fragments have for all practical purposes the same influence on the pK_a of the ethoxylated bis(2-ethylhexy) amine, and for further analyses we can limit ourselves to one of the trimer types, as the two are essentially interchangeable .

Next we study whether the chain length of the Latex particle has a significant influence on the solvation free energy of the ethoxylated bis(2-ethylhexy) amine. We examine this by increasing the repeating units of our earlier samples up to five. Take MMA for an example, Table 5 shows the calculated solvation free energies of the ethoxylated bis(2-ethylhexy) amine interacting with MMA polymer segments of different lengths.

Table 3-5. Calculated solvation free energies of the ethoxylated bis(2-ethylhexy) amine with explicit Latex polymer segments $(MMA)_n$ of different lengths (n=3, 4 ,5). A represents the deprotonated form and HA^+ represents the protonated form.

	$(MMA)_3$	$(MMA)_4$	$(MMA)_5$
$\Delta G_{solv}(A)$ (kcal/mol)	-10.83	-9.09	-10.87
$\Delta G_{solv}(HA^+)$ (kcal/mol)	-60.30	-58.60	-60.04
$\Delta G_{solv}(A) - \Delta G_{solv}(HA^+)$ (kcal/mol)	49.47	49.51	49.17

We find that increasing the number of repeat units of MMA polymer has little influence on the solvation free energies of the ethoxylated bis(2-ethylhexy) amine. The differences between the solvation free energies of deprotonated and protonated hydrophobe groups are also very small, which means that increasing the chain length of Latex particle hardly influences the pK_a calculation outcomes. Note that all calculations involving explicit fragments were done without explicit water molecules for the ease of the computation.

3.3.3 Comparison of the pK_a Calculation Results with Experimental Data

Combining the calculations we have done in Section 3.1 and Section 3.2, we get the pK_a of the ethoxylated bis(2-ethylhexy) amine using Equations 4 and 5. Recall that the solvation free energy of H^+ is a constant equal to -265.9 kcal/mol and the gas-phase basicity is always 228.91 kcal/mol for the ethoxylated bis(2-ethylhexy) amine. Using the solvation free energy of the protonated ethoxylated bis(2-ethylhexy) amine we have calculated in Section 3.2.1, we get 12.9 for the pK_a for the ethoxylated bis(2-ethylhexy) amine on the basis of the continuum model. Applying the hybrid cluster-continuum model results in a pK_a value of 11.1 using the geometry optimization approach, and 8.86 using the MD sampling approach. Finally, taking into account the effect that the presence of a Latex polymer segment has on the pK_a calculation results we observe the following progression: the pK_a of the ethoxylated bis(2-ethylhexy) amine without any Latex particle is 12.9. With $(MMA)_3$ adjacent to the hydrophobe it is 9.43 and with $(BA)_3$ it is 9.45. The difference between these two polymer segment types is negligible. Adding either of them changes the pK_a by 3.5 compared to the original implicit continuum model. Furthermore, the pK_a is 9.43, 9.46, and 9.28 for the calculations that involve $(MMA)_3$, $(MMA)_4$, and $(MMA)_5$, respectively, which again represents very small differences.

Figure 11 shows the experimental pK_a of the ethoxylated bis(2-ethylhexy) amine inferred from the Brookfield viscosity test as the inflection point of the Henderson-Hasselbalch equation fitting the data. The *pH* value of the solution at which the rise in viscosity is steepest is generally considered to coincide with the changeover between protonated and deprotonated states of the hydrophobe. Calculated pK_a results are shown in the same figure as symbols labeled (b) through (e), showing the improvements that each procedural refinement described above has resulted in.

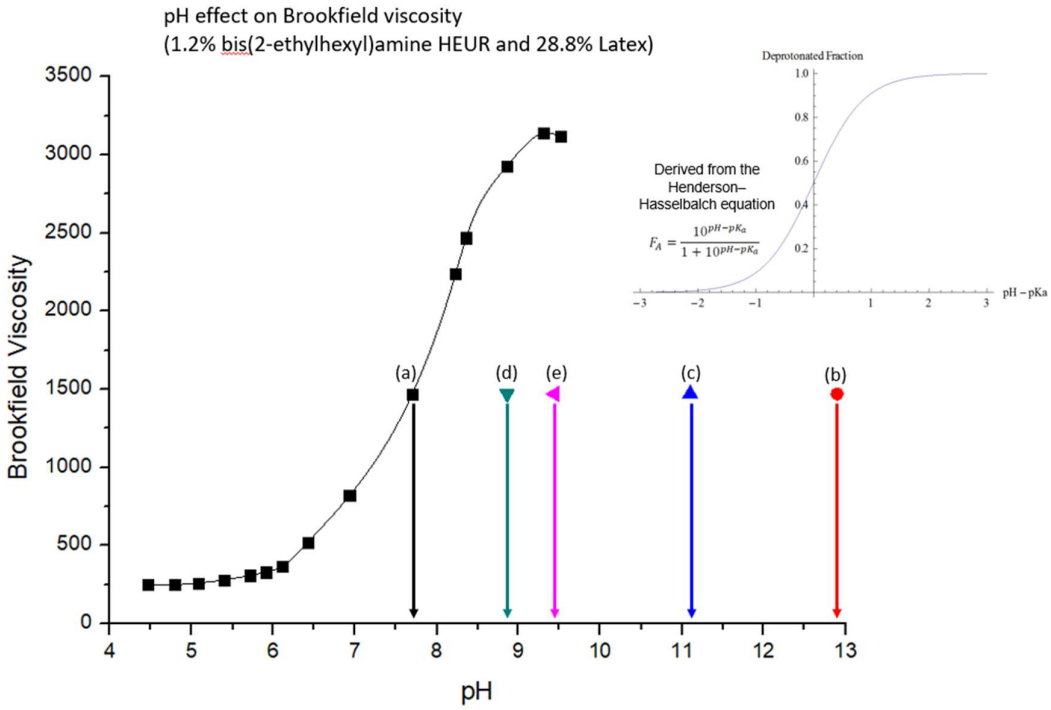


Figure 3-11. Brookfield viscosity of a solution that contains 1.2% ethoxylated bis(2-ethylhexy) amine HEUR and 28.8% Latex as a function of the solvent pH measured by the Dow Chemical Company are shown as black squares.The value labeled with (a) is the experimental pKa derived from the viscosity curve, the value labeled (b) is the calculated pKa using the continuum model, (c) is the calculated pKa using the cluster-continuum model containing explicit water using the DFT geometry optimization approach, (d) is the calculated pKa using the cluster-continuum model containing explicit water using the MD sampling approach, (e) is the calculated pKa using the cluster-continuum model containing explicit Latex fragments using the DFT geometry optimization approach. The plot on the top right is the deprotonation fraction vs. (pH – pKa), derived from the Henderson-Hasselbalch equation.

Theoretically, the pK_a is equal to the current pH value of the solution when the depronated fraction of the ethoxylated bis(2-ethylhexy) amine is equal to 0.5 according to the Henderson-Hasselbalch model. We can see that the experimental data follows a curve of similar shape as the theoretical curve on the top right. Thus the experimental pK_a can be read from the point with the steepest slope on the curve, which is around pH = 7.7. At the beginning, when we only apply the continuum model for our calculations, the resulting pK_a value differs from the experimental one by 5.2. After applying the hybrid cluster-continuum model and take into account the explicit solvent effect, we achieve a significant improvement. The MD sampling approach yields a pK_a value that more closely matches the experimental one than the DFT geometry optimization approach. Considering the effect of explicit Latex particle segments will also give us a closer result to the experimental value compared to only considering about the explicit water molecules using the same calculation method.

3.4 Conclusions

HEUR thickeners are widely used in the Latex paint to control its rheological properties. The depronation/protonation of the hydrophobe on the HEUR molecules controls the thickening functionality. We have used different models, and considering different environments surrounding the hydrophobe ethoxylated bis(2-ethylhexy) amine when calculating the corresponding pK_a . Our analysis shows that the traditional continuum model cannot provide an accurate prediction of the pK_a. Instead, we need to include the explicit surrounding molecules to properly account for local interactions and thereby improve the calculation accuracy. Based on our finds, adding either explicit water molecules or explicit Latex particle fragments to the system improves the pK_a calculation for ethoxylated bis(2-ethylhexy) amine.

When using the same DFT geometry optimization approach, surrounding the hydrophobe with explicit Latex fragments results in a pK_a value that matches the experimental one more closely than surrounding it with explicit water molecules. However, the calculated pK_a that matches the experimental value best is obtained when surrounding the hydrophobe with explicit water molecules and using the MD sampling approach. The future work could include both explicit water and Latex sample in the solvation free energy calculation. This could be a challenging task because the degree of freedom increases significant thus it would be really difficult to find the low-energy structure for the DFT geometry optimization approach. Even the MD sampling approach may require a large number of water molecules to converge which could also be very computationally expensive.

3.5 Reference

1 Bobsein, B.R., Johnson, M.M., Rabasco, J.J., and Zeszotarski, C., 'Thickener composition and method for thickening aqueous systems,' U.S. patent no. US7741402B2 (2010).

2 Sau, A.C., 'Hydrophobically modified poly(acetal-polyethers),' U.S. patent no. US5574127A (1996).

3 Winnik, M.A. and Yekta, A., 'Associative polymers in aqueous solution,' *Curr. Opin. Colloid Interface Sci* **2**, 424 (1997).

4 The Dow Chemical Company, 'AcrysolTM hydrophobically modified ethylene oxide urethane (HEUR) rheology modifiers,' 2011.

5 Jang, Y.H., Goddard, W.A., Noyes, K.T., Sowers, L.C., Hwang, S., and Chung, D.S., 'First principles calculations of the tautomers and pK(a) values of 8-oxoguanine: Implications for mutagenicity and repair,' *Chem. Res. Toxicol.* **15**, 1023 (2002).

6 Klicić, J.J., Friesner, R.A., Liu, S.-Y., and Guida, W.C., 'Accurate prediction of acidity constants in aqueous solution via density functional theory and self-consistent reaction field methods,' *The Journal of Physical Chemistry A* **106**, 1327 (2002).

7 Liptak, M.D. and Shields, G.C., 'Accurate pK(a) calculations for carboxylic acids using

Complete Basis Set and Gaussian-n models combined with CPCM continuum solvation methods,' *J. Am. Chem. Soc.* **123**, 7314 (2001).

8 Cramer, C.J., '*Essentials of computational chemistry*,' 2nd edition (WILEY, 2004), P579.

9 Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening,' *J. Chem. Theory Comput.* **5**, 2447 (2009).

10 Bryantsev, V.S., Diallo, M.S., and Goddard, W.A., 'Calculation of solvation free energies of charged solutes using mixed cluster/continuum models,' *J. Phys. Chem. B* **112**, 9709 (2008).

11 Pliego, J.R. and Riveros, J.M., 'The cluster-continuum model for the calculation of the solvation free energy of ionic species,' *J. Phys. Chem. A* **105**, 7241 (2001).

12 Zhan, C.G. and Dixon, D.A., 'Absolute hydration free energy of the proton from first-principles electronic structure calculations,' *J. Phys. Chem. A* **105**, 11534 (2001).

13 Ho, J.M. and Coote, M.L., 'A universal approach for continuum solvent pK(a) calculations: are we there yet,' *Theor. Chem. Acc.* **125**, 3 (2010).

14 Lewars, E.G., '*Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics, Second Edition*,' (Springer, 2011), P317.

15 McQuarrie, D.M., '*Statistical Mechanics*,' (Harper and Row, New York, 1970), P86.

16 Gaussian 09, Revision D.01, M. J. Frisch, G.W.T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

17 Zhao, Y. and Truhlar, D.G., 'The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals,' *Theor. Chem. Acc.* **120**, 215 (2008).

18 Tissandier, M.D., Cowen, K.A., Feng, W.Y., Gundlach, E., Cohen, M.H., Earhart, A.D.,

Coe, J.V., and Tuttle, T.R., 'The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data,' *J. Phys. Chem. A* **102**, 7787 (1998).

19 Kelly, C.P., Cramer, C.J., and Truhlar, D.G., 'Aqueous solvation free energies of ions and ion-water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton,' *J. Phys. Chem. B* **110**, 16066 (2006).

20 Ho, J.M., 'Are thermodynamic cycles necessary for continuum solvent calculation of pK(a)s and reduction potentials,' *Phys. Chem. Chem. Phys* **17**, 2859 (2015).

21 Ribeiro, R.F., Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Use of Solution-Phase Vibrational Frequencies in Continuum Models for the Free Energy of Solvation,' *J. Phys. Chem. B* **115**, 14556 (2011).

22 Hunter, E.P.L. and Lias, S.G., 'Evaluated gas phase basicities and proton affinities of molecules: An update,' *J. Phys. Chem. Ref. Data* **27**, 413 (1998).

23 Del Bene, J.E., 'Basis-set effects on computed acid-base interaction energeis using the Dunning correlation-consistent polarized split-valence basis sets,' *J. Mol. Struct.* **307**, 27 (1993).

24 Marenich, A.V.K., C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database - version 2012*, University of Minnesota, Minneapolis, 2012.

Chapter 4

Accurate Acid Dissociation Constant Calculations for Hydrophobes in the Rheology Modifiers

4.1 Introduction

Accurate prediction of the solvation free energies is essential in many fields of study, ranging from chemical reactions in solutions to the design of functional molecules in chemistry and biochemistry. However, the reliable determination of the solvation free energies for ionic species can be computationally challenging. Typical neutral molecules have a solvation free energy of less than 10 kcal/mol, while many small charged ions exhibit values in the 50-100 kcal/mol range. This means that achieving a small absolute error, which is important for example in pK_a calculations [1], is of great difficulty. Over the past decade, much theoretical effort has been put into the development of methods to calculate the solvation free energy. Traditional explicit solvation models are very computationally expensive because of the large number of configurational degrees of freedom. Thus it is really difficult to find the fully relaxed complex structure. [2] Continuum-based implicit solvent models [1, 3–5] represent actual solvent molecules as a uniform polarizable medium of fixed dielectric constant, and the solute molecules is embedded in a suitably shaped cavity. While producing reasonable results for neutral solute molecules, these have been found to be inadequate for ionic species because

of the failure to account for the strong localized solute-solvent interactions. The recently developed hybrid cluster-continuum model, which combines the explicit and implicit model by adding explicit solvent molecules to the continuum model, has improved calculation accuracy and speed. [6–10] However, it is still not computationally efficient enough, especially when dealing with large solute species requiring a great number of surrounding water molecules. In such cases DFT calculations, especially when attempting geometry optimization, become impractical. Therefore, it is advisable to explore an informatics-based machine learning (ML) model, which is trained using a limited amount of experimental data, to predict the solvation free energy accurate and fast, especially for the charged species.

ML has been widely applied in materials informatics recently due to its low computational cost. Investigators use ML techniques to train models based on known properties learned from the existing data (also called training data in the ML field) to make predictions of the properties for the new data. Typically training an ML model involves two steps – first extracting key features from materials in an existing dataset, which are quantitative attributes that describe their relevant characteristics; and then mapping these features to the property of interest. The first step requires significant expertise and knowledge of the materials and the second step is purely numerical in nature. [11] There are numerous features that can be used to train the model, such as basic atomic information that can be gained from the chemical formula, molecular geometry information that can be described by the chemical graph theory, or energies that need to be calculated using first-principles calculations. It is very important to choose the right features that are effectually related to the property of interest in order for the

model to give an accurate prediction of the property. Usually features extracted from molecules should fulfill several basic requirements: (1) complete, (2) descriptive, (3) simple and (4) unique. [12] The goal in creating a "complete" representation is to provide enough information that is relevant to the property of interest to sufficiently differentiate materials. For example, the atom type by itself cannot be a complete choice of features for solvation free energy predictions since many solutes share the same type of atoms. "Descriptive" means similar materials should have similar features. For example, the electronegativity difference between elements in a compound could be a good feature for its formation energy prediction. One would expect similar electronegativity differences for bonds with similar formation energies. "Simple" means that the computational representation of the features should be fast to accomplish. This is more important when we are trying to use features that are calculated using the first-principles method. If the computational cost of describing features themselves is very high, then we lose the point of using the informatics-based approach. Lastly "unique" means any material should have exactly one representation. The features should be invariant to certain transformations. If a given material has multiple representations, it is possible to predict different properties for the same material.

Studies of solvation free energy prediction using ML techniques have been carried out recently by Moorthy [13] and Bao Wang. [14] Moorthy did a classification study of solvation free energies of organic molecules using ML methods like support vector machine, random forest and decision tree. His analysis was performed with easily obtainable features such as atom count, topological measures, surface area, and molecular access system (MACCS) finger-

prints that account for the presence of absence of particular functional groups, atoms or fragments in different molecules. Different models were built by selecting different subsets of features from the 188 total features. These models correctly classified >95% of the molecules as having highly favorable or less favorable solvation free energies. Despite the good classification accuracy, Moorthy did not carry out feature importance analysis so it remains unknown whether all the 188 features are equally important or some are actually not that necessary. In addition, further regression analysis can be done to predict the specific solvation free energy of a molecule instead of doing classification only, which is more important in many chemistry and biochemistry studies. Wang has made a great contribution to the solvation free energy prediction using the ML approach. He extracted features from the solvation free energy calculation procedure employed in the implicit solvent model, e.g., electrostatic features like atomic charge and reaction field energy, as well as nonpolar features like atomic surface area. For each target molecule, he adopted an ML algorithm to search for its nearest neighbor, based on the selected features. Then from the features of nearest neighbors so determined, he constructed a functional of solvation free energy, which is employed to predict the solvation free energy of the target molecule. [14] He also analyzed the importance of the nonpolar features to show that they are necessary to improve the prediction results. However, the polar features require first-principles calculations and the analysis failed to show whether all of them are of equal importance or not.

In this study, we apply ML techniques to predict the solvation free energies, mainly for charged species. In contrast to Moorthy's work, we start with fewer features that we think are the keys to the solvation free energy to keep the model simple. We then gradually add

features and examine whether they provide noticeable improvement to the accuracy of the prediction results. For the feature selection, unlike what Wang did in his work, we try to avoid using results from first-principles calculations to lower the computational cost and make it easier for possible future software development. More importantly, the electrostatic term dominates the total solvation free energy of the molecule, although it does not always indicate a high affinity to the solvent. [15] We compare the performance of three different ML methods – linear ridge regression, support vector regression, and random forest regression in terms of solvation free energy prediction accuracy. We train each model with the help of the scikit-learn package in Python. We find that using the atomic fraction extracted form the chemical formula, the Wiener index that is gained from the molecular topology, and the solvent accessible surface area (SASA) are sufficient to give a relatively accurate prediction with the random forest regression method. Compared to the 4 kcal/mol mean unsigned error for ions from the recent continuum solvation model based on electron density (SMD), [16] our ML model performs quite well with a mean unsigned error of 4.43 kcal/mol for charged molecules. Finally we have tried to use our model to predict the solvation free energy of the hydrophobe molelecule on the HEUR rheology modifier, and it shows a good agreement with our previous result calculated using first-priciples methods.

4.2 Theory and Methodology

4.2.1 Data Set

The Minnesota solvation database – version 2012 [17] is used to train our ML model. It contains 3037 experimental solvation free energies for 790 unique solutes (541 neutrals and 249 singly-charged ions) in 92 solvents (including water). The database focuses on the solvation

free energy of organic molecules and all of the 790 solutes in this database contain at most the following elements: H, C, N, O, F, Si, P, S, Cl, Br, I. In this work, we are not going to use all these 3037 experimental data because we focus on the solvation free energies of singly charged ions in the water. We expect very different solvation free energies between neutral and charged solutes, and therefore we believe it is better to use different models to predict their solvation free energies. After removing the data for which the solutes are neutral molecules and the solvent is not water, 112 aqueous free energies of solvation for 112 singly charged ions are retrieved. All data are experimentally measured and available in the literature. [16, 18, 19]

4.2.2 Feature Selection

As mentioned above, feature selection is very important to build an ML model that can accurately predict the target property. Ideally, we want to choose the features in which the property of interest is fundamentally rooted. Instead of applying all of the available features from software like PaDEL-Descriptor, we use representative features from three main feature categories for chemoinformatics: physico-chemical features, 2D molecular features, and 3D molecular features.

4.2.2.1 Physico-chemical Features

This category contains a lot of features that may be easily obtained from the chemical formula like atom count, molecular weight, atomic fraction, etc., or features that need to be calculated using first-principles calculations like the reaction field energy and atomic charge used in Bao Wang's study. [14] We believe that different atoms and their atomic fractions have different effects on the solvation free energy. For example, one fully expects to find that the

solvation free energy of NH_4^+ is different from that of H_3O^+ because they contain different atoms. The same is true for methylamine and n-propylamine as they have different atomic fractions. However, adding too many of these features may not be effective, as they could be strongly correlated, and the additional features increase the complexity of the model while it does not help much to improve the prediction accuracy. For example, adding the electronegativity as an additional feature might not be a good idea since this characteristic is straightforwardly mapped to the period table of elements and may therefore already by reflected in the atom type information. The features that need to be calculated from the first-principles simulations are computationally expensive, which is discrepant from the "simple" feature selection principle. Furthermore, they may also be violating the "unique" feature selection principle. For example, when the molecule size increases, first-principles calculations with different initial configurations may give different results for the atomic charges, and this could result in different properties for the same molecule. We first choose the atomic fraction as the representative for this category of features. To be specific, we define a material using a vector wherein n^{th} component of the vector represents the atomic fraction of the n^{th} element from the database ($n \leq 10$, in the order of H, C, N, O, F, Si, P, S, Cl, Br, I). For example, the atomic fraction features for H_3O^+ would be [0.75, 0, 0, 0.25, 0, 0, …]. In the future, more features from this category may be examined/added to train the model if they are not already highly correlated with the atomic fraction.

4.2.2.2 2D Molecular Features

This category usually contains features that can be extracted from the molecular topology. Indeed the feature selection is not "complete" by just including the physico-chemical features

for our solvation free energy prediction model sometimes. For example, isomers with the same atomic fraction may have different solvation free energies. Therefore, adding topological features is necessary. We choose one of the classical topological features, the Wiener index as our 2D molecular feature. The Wiener number is the sum of the distances between all pairs of vertices in a graph. To calculate the Wiener number, we first need to map the molecule to a graph. The nodes in the graph are the non-hydrogen atoms in the molecules and the edges are the bonds in the molecules. Then we find the distance matrix of the graph and we compute the Wiener number by adding the entries in the upper triangular part of the distance matrix. The Wiener index is the average of the distances between all pairs of vertices in a graph. For a graph having n vertices, $Mean\ Wiener\ index = Wiener\ number/\binom{n}{2}$. In the future, more features from this category like the Hosoya index, Randić's molecular connectivity index may also be examined/added to train the model.

4.2.2.3 3D Molecular Features

This category usually contains finer features that describe the geometry of the materials compared to the 2D molecular features. We add the solvent accessible surface area feature like both Moorty and Bao Wang did in their studies for our model training. According to Junmei Wang's study, the solvent accessible surface area is one of the most important features to describe the solvation free energy very well. Good results have been achieved by just using this feature. [15] However, due to the "complete" feature selection principle, we believe that it is necessary to include the features we have discussed in the previous sections, since different molecules may have similar solvent accessible surface areas. We use the GePol method that is described in Pascual-ahuir's work [20] to calculate the solvent accessible surface area. In

short, it is a numerical method in which a set of points is placed reasonably uniformly on the surface of the spheres surrounding each atom. Any point on a sphere that lay within the volume of another sphere is discarded, and the ratio of the exposed surface area to the total area of all the spheres is set equal to the number of nondiscarded points divided by the original number of points. More details of this method can be found in his paper "GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface." [20]

4.2.3 ML Methods

We have tried different ML methods to train the model so as to accurately connect the features we have extracted in the previous section and the solvation free energy property.

4.2.3.1 Linear Ridge Regression

In statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables $\boldsymbol{x}$. In our case, we have a set of molecular features $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in R^M$, N is the number of sample data and M is the number of features, and corresponding target solvation free energies $y_1, \dots\ y_N$. A linear regression model assumes that the relationship between y and $\boldsymbol{x}$ is linear. We want to learn a function $f(\boldsymbol{x}_i, \boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{x}_i$ to predict future solvation free energies, where $\boldsymbol{\omega}$ is the model parameter with dimension M that we need to determine to minimize the error between the predicted value and the N real property value. We find $\boldsymbol{\omega}$ using the standard linear least-squares optimization algorithms, which is to minimize the cost function: $L(\boldsymbol{\omega}) = \sum_i^N (y_i - \boldsymbol{\omega}^T \boldsymbol{x}_i)^2$. In the ridge regression procedure another term is added to in the cost function to penalize for size of

the regression coefficient, $L(\boldsymbol{\omega}) = \sum_i^N (y_i - \boldsymbol{\omega}^T \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{\omega}\|^2$. λ is referred to as a hyperparameter in the cost function, which requires tuning to achieve the best performance of the model.

4.2.3.2 Support Vector Regression

The goal of support vector regression is to find a function $f(\boldsymbol{x}) = \boldsymbol{\omega}^T \boldsymbol{x} + b$ that deviates from y by a value within the ε band for each training data point $\boldsymbol{x}$. It also uses slack variables ξ to overcome noise and outliers in the data. (Figure 1)

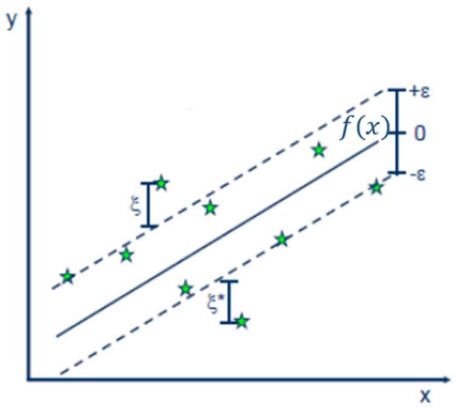


Figure 4-1. Picture of ε band with slack variables for support vector regression.[21]

The support vector regression is formulated as minimization of the following functional:

$$\min \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*),$$

subject to $y_i - f(\boldsymbol{x}_i) \le \varepsilon + \xi_i^*; \quad f(\boldsymbol{x}_i) - y_i \le \varepsilon + \xi_i; \quad \xi_i, \xi_i^* \ge 0.$

This optimization problem can be transformed into a dual problem and its solution is given by

$$\min \quad L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\boldsymbol{x}_i, \boldsymbol{x}_j) + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*)$$

$$s.t.\ \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0;\quad 0 \le \alpha_i \le C;\quad 0 \le \alpha_i^* \le C$$

Where $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is called the kernel function that maps $\boldsymbol{x}$ to a high-dimensional space. We will be tuning hyperparameters C, the kernel function type K and the kernel coefficient gamma for the kernel function later in the discussion section.

4.2.3.3 Random Forest Regression

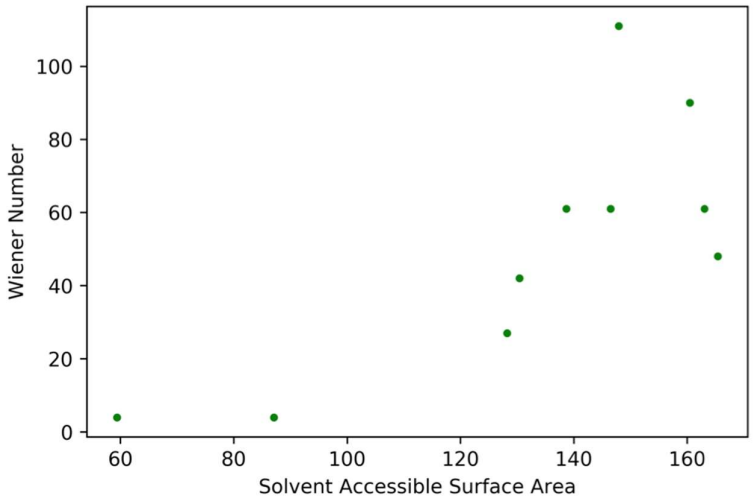
Random forests, proposed by Breiman, is an ensemble learning method that operates by constructing a number of decision trees at training time. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. For classification, the measure based on which the optimal condition is chosen is Gini impurity or information gain. And for regression trees it is residual sum of squares (RSS): $RSS = \sum_{left}(y_i - \overline{y_L})^2 + \sum_{right}(y_i - \overline{y_R})^2$, where $\overline{y_L}$ and $\overline{y_R}$ are the mean value of the property of interest for the left node and right node, respectively. To make it clearer, next we illustrate how to construct a simple regression tree using just two features -- solvent accessible surface area (SASA) and Wiener number (W). We choose 10 random samples from the pre-existed data set as shown in Table 1. A 2D feature plot is also shown in Figure 2(a). Then we tried to find the best horizontal and vertical split of the data by taking the average SASA and W between each adjacent data point and calculate their corresponding RSS. In this scenario either the horizontal split at W=15 or the vertical split at SASA=107.63 will result in a minimum RSS. We choose the vertical split as our first branch of the regression tree. (Figure 2(b)) Now we have the most simple regression tree with only one layer. Then we continue splitting to construct the second layer of the tree. For the left branch, since

only two data points left it is easy to split them. They share the same Wiener number so we split them at SASA=73.21. By far the left branch has reached to the end, which are usually called leaf in the regression tree. The values on the leaf are the solvation free energy for that data. By applying the same optimization algorithm, we can find the best split for the right branch, which is at SASA=164.24. Then the SASA>164.24 branch reach the leaf and we can get the solvation free energy of this leaf as -54.6 kcal/mol. So far we have finished constructing the second layer of the regression tree (Figure 2(c)). Similarly, we could continue constructing the third layer of the regression tree by finding the best split as W=35. (Figure 2(d)) If we stop here, we will end up with a regression tree with three layers. We could also keep going until we separate every single sample in the data. Later we will be tuning this hyperparameter -- number of layers (max_depth). A random forest is simply composed by multiple regression trees like this. Each time a regression tree could use different combinations of features and select different samples from the original data for the data splitting. If we are trying to predict the solvation free energy of a new molecule, we first see which leaf will this molecule end up with for all of the regression trees. The final predicted solvation free energy is the average value of the solvation free energy on the leaf. We will also be tuning this hyperparameter – number of trees (n_tree) in the later section.

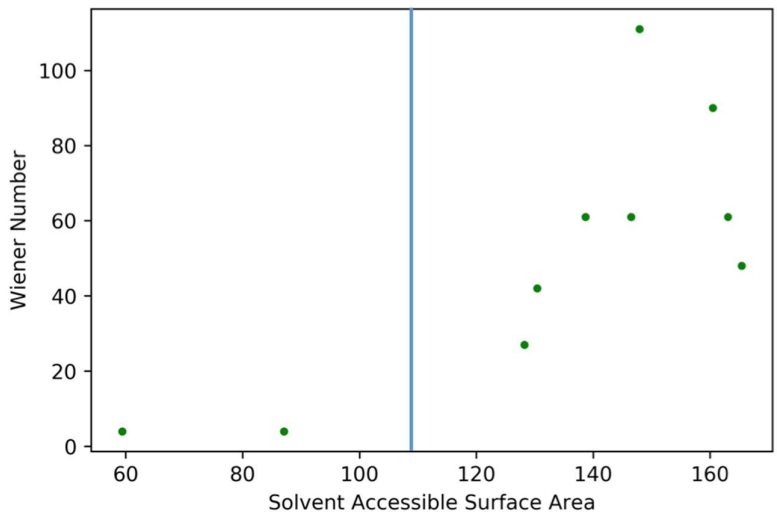
Table 4-1. 10 random samples selected from the original data set.

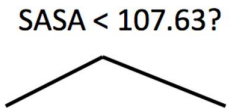
	Molecule	SASA (Å^2)	Wiener Number	G_{solv}
1	3-methylphenol	138.66	61.0	-71.1
2	1,1,1,3,3,3-hexafluoropropan-2-ol	147.92	111.0	-65.5

3	triethlamine	165.41	48.0	-54.6
4	dimethylether	87.05	4.0	-79.7
5	aniline	130.40	42.0	-72.4
6	formicacid	59.38	4.0	-76.2
7	3-chloroaniline	163.07	61.0	-74.7
8	piperidine	128.21	27.0	-64.2
9	3-aminoaniline	146.46	61.0	-65.8
10	4-methoxyaniline	160.46	90.0	-71.2

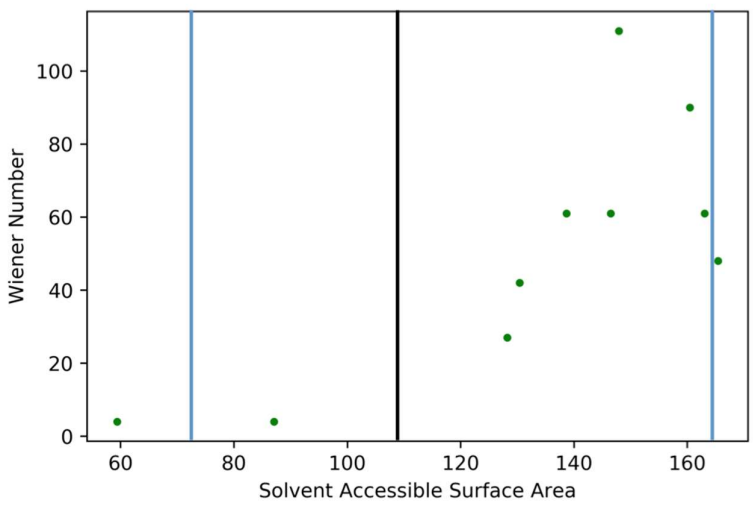


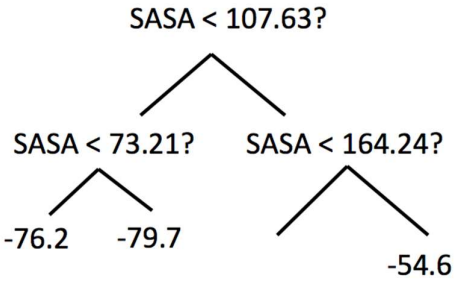
(a)


Wiener Number
Solvent Accessible Surface Area
0
20
40
60
80
100
60
80
100
120
140
160


SASA < 107.63?

(b)


Wiener Number
Solvent Accessible Surface Area
0
20
40
60
80
100
60
80
100
120
140
160


SASA < 107.63?
SASA < 73.21?
SASA < 164.24?
-76.2
-79.7
-54.6

(c)

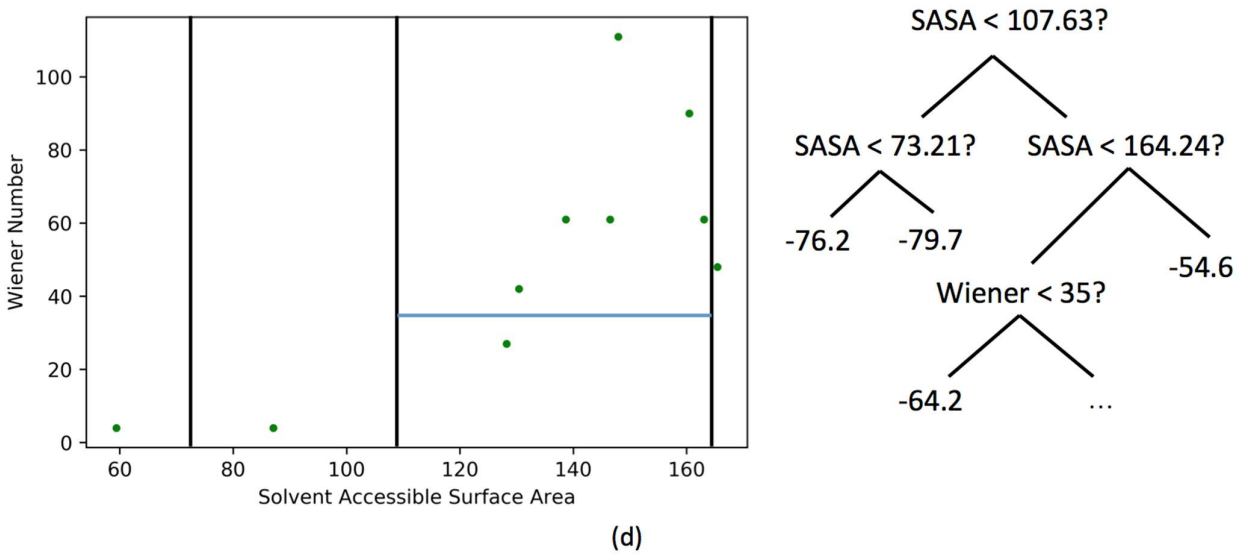


Figure 4-2. (a) 10 random samples selected from the original data set. And the first split (b), second split (c), third split (d) of the data. The numbers on the leaves (end of the branch) are the solvation free energies.

4.2.4 Workflow

To illustrate how the training and predicting work, we have summarized a workflow as shown in Figure 1 below.

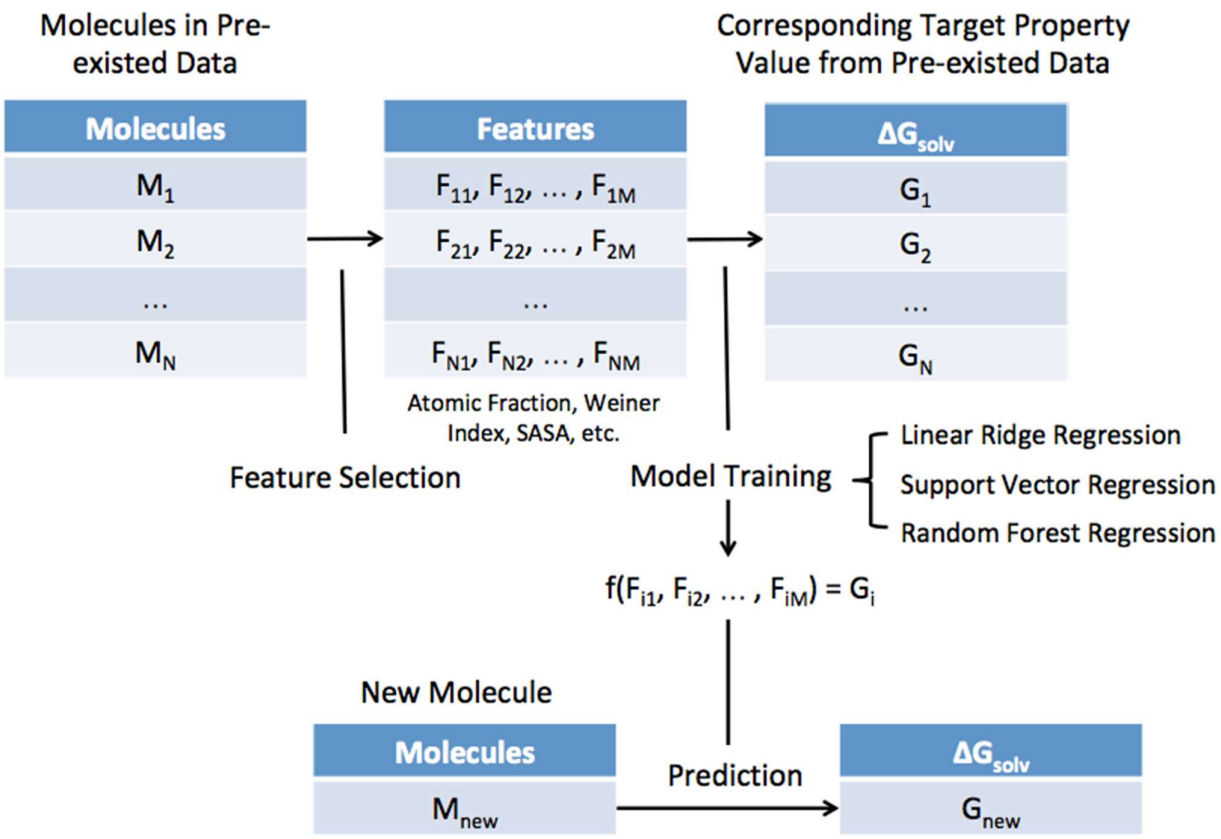


Figure 4-3. Workflow of the solvation free energy prediction using ML techniques.

We start from the molecules of which the experimental solvation free energies are stored in the Minnesota solvation database. The first step to build the solvation free energy prediction model is to select and extract useful features that we have discussed in previous sections from these molecules. Then we train the model using different ML methods to find the mapping between these features and the solvation free energy value. At last, we apply this model on the new molecules that we are interested in and predict its solvation free energy.

4.3 Results and Discussion

In this section we describe how we tune and train the models using each of the three ML methods based on pre-existing data. We compare the accuracy of these methods and pick the one with the lowest test mean absolute error (MAE). We also carry out feature importance

analysis to show which of the features are more important for the accurate model and which are not. We also show why it is not guaranteed that including as many features as one can to train the model necessarily improves its predictive accuracy. And finally we show that our ML prediction of the solvation free energy for a ethoxylated bis(2-ethylhexy)amine hydrophobe molecule is comparable to that calculated using the first-principles methods.

4.3.1 Regression Model Selection

As mentioned in Section 2.1, regression models with the ability to predict the solvation free energy of molecules are developed based on a dataset comprised of 112 ions and their solvation free energies. A 5-fold cross-validation method is applied to these data to prevent from overfitting. One cannot evaluate the quality of a model by examining the error it achieves on the data on which it was trained. Therefore we divided the original data set into five subsets of the same size. Each time we picked four out of five subsets as the training dataset, which we will use to train our model, and the leftover subset is used as a test dataset to check the quality of our model. We repeat this process five times until all subsets have been used as a test dataset once. All of the features that have been mentioned in the previous section are used to find the best training method.

Before we start to train our model, we first need to tune the hyperparameters within the model so that it gives us the best results. As opposed to regular model parameters, which are used to relate selected features to the predicted property, hyperparameters are those that ascertain proper convergence in the parameter optimization and confine their values to within allowed parameter ranges. For example, in the linear ridge regression method, λ is hyperparameter in the model that penalizes for exagerated size of the regression coefficient and thus

prevents from overfitting. When $\lambda = 0$, the linear ridge regression becomes the normal linear regression and there is no penalty for a large regression coefficient, which will likely cause the overfitting of the model. On the other hand, a really large λ will force the minimization to focus on minimizing $\|\beta\|^2$, thus diverting from the true fit and this is usually called under-fitting. Figure 2 shows the validation curve for the linear ridge regression method, which tells us how the training error and validation error change with the choice of different hyperparameter λ. Notice that in Python, the MAE score is simply the negative value of the mean absolute error. The sign is flipped because of the "greater is better" principle in the Python scoring system.

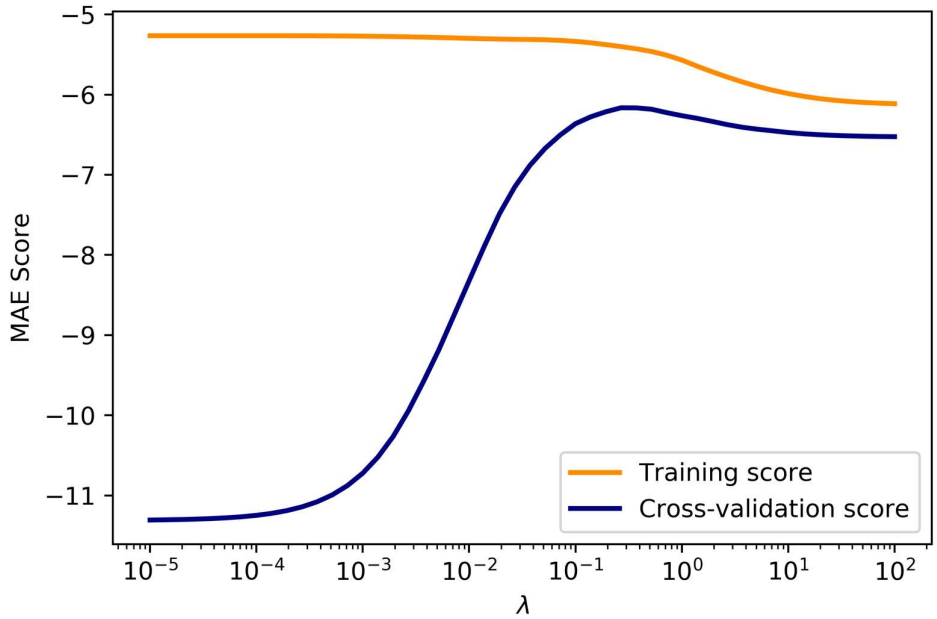


Figure 4-4. Validation curve with linear ridge regression. Orange line and blue line show how the training error and validation error change with the hyperparameter λ, respectively.

We can see that when λ is small, the model has a small training error and a really large validation error. This means that the model is overfitted -- it describes the 80% of data used for

training relatively well but fails to give good predictions to the remaining 20% of validation data. When we increase the value of λ, the training error goes up because we emphasize the term that generalizes the model more and more. Mathematically, we are far from minimizing the term purely based on the training dataset $\sum_i^n (y_i - \boldsymbol{X_i}^T \beta)^2$. As for the validation error, it has a minimum value near $\lambda = 0.1$, which is a balancing point in the tradeoff between minimizing $\sum_i^n (y_i - \boldsymbol{X_i}^T \beta)^2$ and minimizing the regularization term $\|\beta\|^2$. With the help of the GridSearchCV function in Python, we fine-tune the hyperparameter λ equal to 0.193 for the model that gives us the best prediction capability.

We apply a similar technique to the support vector regression and random forest regression method. Everything is the same except that they are much more complicated methods with many more hyperparameters. We only include the tuning of most important ones here – Penalty parameter of the error term (C), kernel type (k_type), kernel coefficient (gamma) for the support vector regression method, and number of regression trees in a random forest (n_tree), maximum depth of the regression tree (max_depth), the minimum number of samples required to split an internal node (min_split) for the random forest regression method. We use the default settings for the rest of the hyperparameters in Python. We adjust the above set of hyperparamters by trying out a wide range of values so that we have a grid of hyperparamters. Instead of getting a 2D plot in Figure 2, we try each combination of the hyperparameter in the grid and find the following combination gives us the best result: C=0.5, k_type = 'linear', gamma = 10^{-5} for support vector regression and n_tree = 50, max_depth = 20, min_split = 2 for random forest regression.

Now we have found the best hyperparameters for all ML methods. We then use the three ML methods with their best hyperparameters to train the models. We record the performance of the two models in Table 1.

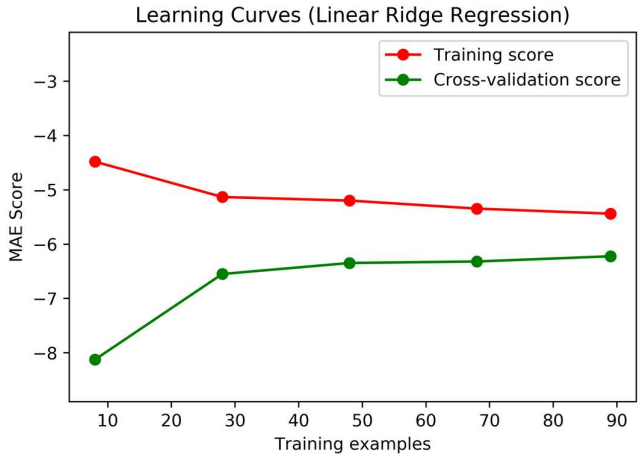
Table 0-2. Summary of model performance with different ML method.

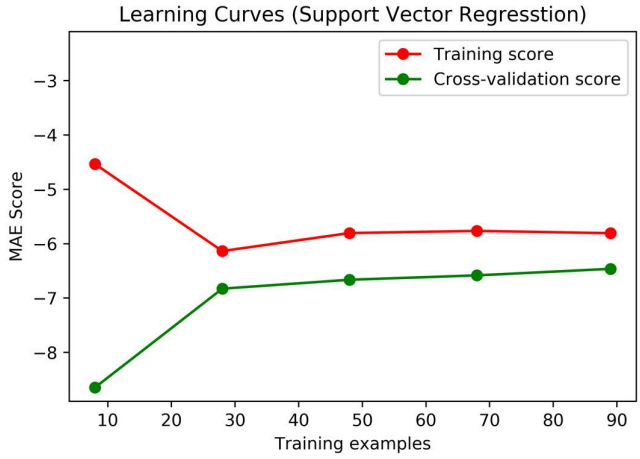
Model #	ML Method	Hyperparameters	Prediction Result (MAE)
1	Linear Ridge Regression	λ=0.193	4.99 kcal/mol
2	Support Vector Regression	C=0.5, k_type= 'linear', gamma = 10^{-5}	5.67 kcal/mol
3	Random Forest Regression	n_tree=50, max_depth=20, min_split=2	4.43 kcal/mol

It is clear that the third model, the random forest regression, has the lowest mean absolute error, which is equal to 4.43 kcal/mol. This is a very promising result since the mean absolute error of solvation free energies calculated by the first-principles calculation is 4 kcal/mol. Our third model therefore has a prediction accuracy on par with the traditional first-principles calculations, but the required computational effort is much less and the calculation speed is much faster. Before we conclude that this is the go-to model for the solvation free energy prediction.

Next we conduct and additional analysis to further prove that random forest regression is indeed superior to the other two regression methods. Figure 3 shows the learning curves of the

three methods. A learning curve shows the validation and training score of a method as a function of the number of training samples.





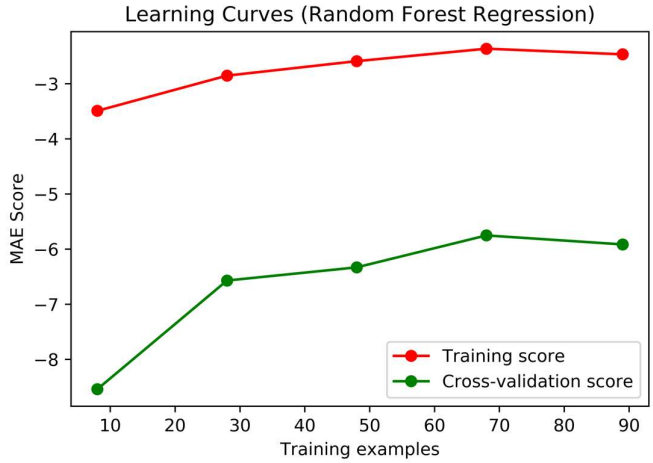


Figure 4-5. Learning curves with linear ridge regression (top), support vector regression (middle) and random forest regression (bottom). Green line and red line show how the training error and validation error change with the number of training samples, respectively.

The validation error decreases in all three cases. The decrease of the validation error is due to the model's ability to generalize. Only random forest regression has a slightly increasing training score, which means the training error decreases as we increase the size of the training data. The other two regression methods, linear ridge and support vector, both have increasing training error with increasing training samples. This means that, as the data get more complex, it is more difficult for these two methods to learn a model that correctly represents all the training data. Thus the random forest regression method is superior. There is some more important information we can get from the learning curve. We can see that although both training error and validating error of the random forest regression method decrease as the sample size increases, the training error is still much smaller than the validation error even for the maximum number of training samples we have included here. This means that adding more training samples most likely increases generalization and further decreases the validation error, make the model better the more it is trained. After all, 112 is a really small number for the size of the ML data. This suggests that measuring more solvation free energies of ions and increasing the original dataset will further improve the accuracy of our model.

4.3.2 In-depth Discussion of Feature Selection

After establishing the best ML method along with well-tuned hyperparameters for the solvation free energy prediction, we now discuss the importance of selecting the right number of features and type of feature to include in the model training in terms of prediction accuracy. Previously, we used all three types of features in the model training – (1) atomic fraction, (2)

Wiener number and (3) the SASA of the molecule. Table 2 shows how different combinations of these features affect the solvation free energy prediction.

Table 4-3. Summary of model performance with different combinations of features.

Combination #	Selected Feature(s)	MAE (kcal/mol)
1	(1), (2), (3)	4.43
2	(1)	5.76
3	(2)	7.17
4	(3)	6.38
5	(1), (2)	4.34
6	(1), (3)	4.91
7	(2), (3)	7.02

We can see that using only one of features from (1) (2) (3) does not give a good prediction result. This because none of them individually can describe the solvation free energy well. Adding either (2) or (3) to (1) improves the result with a mean absolute error of less than 5 kcal/mol, while combining (2) and (3) still results in a model with poor performance. From this we know that feature (1) atomic fraction plays a more important role in the solvation free energy prediction compared to the other two features. If we take a further look on Table 2, using only feature (1) and (2) actually gives us a very similar, even slightly better result than using all of them. This shows that it is not guaranteed that adding more features necessarily increases the performance of the model. Sometimes it may not be best to include as many

features as possible. The reasons for this could be: (1) the additional features might be irrelevant to the property of interest, (2) they are just highly correlated with the previous features, or (3) the training sample size is just not big enough to differentiate between these features. In our case, it is more likely to be the latter two reasons because using the model with each feature individually actually yields a better result than just guessing the average solvation free energy for all molecules, in which case the MAE is 8.45 kcal/mol. This means that each of them does give us positive information about the solvation free energy. If we look more closely into the Wiener number and the SASA of the molecules, they actually convey similar information. For example, a large molecule generally yields a bigger Wiener number as well as a bigger SASA. To this extent they are correlated. Most importantly, we only have 112 pre-existed data, and 80% of them are used for model training, the small sample size may not allow us to differentiate between the two features very well. We believe that by increasing the size of the pre-existed data, combination #1 would have a better performance than combination #5, where only feature (1) and (2) are included. In this study, we still keep all of the three features for our solvation free energy prediction model. In the future, additional features are encouraging to be examined and added to the model if they indeed improve the model performance, however it is not recommended to add arbitrary features randomly to the model without careful consideration.

4.3.3 Solvation Free Energy Prediction of the Charged Hydrophobe Molecule on the HEUR Rheology Modifier

We have developed the ML based solvation free energy prediction model. Now we can apply this model to predict the solvation free energies of molecules of interest of which the solvation free energy has not been measured. In our previous paper we have calculated the solvation free energy of the charged hydrophobe molecule, ethoxylated bis(2-ethylhexy) amine on the HEUR rheology modifier as -54.18 kcal/mol. To make the prediction using our new ML model, we first extract features from the hydrophobe molecule. The chemical formula of protonated ethoxylated bis(2-ethylhexy) amine is $C_{18}H_{40}O_1N_1$. Thus the corresponding atomic fraction vector is: [0.6667, 0.3, 0.0167, 0.0167, 0, 0, …]. Its Wiener number is 930 and the solvent accessible surface area is 395.678 $Å^2$. Then we can predict its solvation free energy as -55.77 kcal/mol, which is only 1.59 kcal/mol different from the first-principles calculation result.

4.4 Conclusions

In conclusion, we have applied three different machine learning techniques to predict the solvation free energies of charged species. We found that after tuning the hyperparameters, the random forest regression leads to a model that predicts the solvation free energies most accurately, with a mean absolute error of 4.43 kcal/mol. Compared to the traditional first-principles calculations, which has a comparable mean absolute error of 4 kcal/mol, our machine-learning based model greatly reduces the computational cost and time without losing much accuracy. Our analysis was performed with easily obtainable features: The atomic frac-

tion feature plays the most important role in the solvation free energy prediction, adding Wiener number and solvent accessible surface area of the molecules further improves the performance of the model. In the future, increasing the sample size and adding more clearly non-correlated features that are relevant to the solvation free energy could further improve the model performance. Finally we use our model to predict the solvation free energy of the hydrophobe molecule on the HEUR rheology modifier, and it shows a good agreement with the previous result calculated using the first-principles method with only 1.59 kcal/mol difference.

4.5 Reference

1 Liptak, M.D. and Shields, G.C., 'Accurate pK(a) calculations for carboxylic acids using Complete Basis Set and Gaussian-n models combined with CPCM continuum solvation methods,' *J. Am. Chem. Soc.* **123**, 7314 (2001).

2 Cramer, C.J., '*Essentials of Computational Chemistry*,' 2nd edition (WILEY, 2004), P449.

3 Cramer, C.J. and Truhlar, D.G., 'AM1-SM2 and PM3-SM3 parameterized SCF solvation models for free energies in aqueous solution,' *J. Comput. Aided Mol. Des.* **6**, 629 (1992).

4 Cramer, C.J. and Truhlar, D.G., 'An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation,' *Sci.* **256**, 213 (1992).

5 Foresman, J.B., Keith, T.A., Wiberg, K.B., Snoonian, J., and Frisch, M.J., 'Solvent effects.5. Influence of cavity shape, truncation of electrostatics, and electron correlation ab initio reaction field calculations,' *J. Phys. Chem.* **100**, 16098 (1996).

6 Zhan, C.G. and Dixon, D.A., 'Absolute hydration free energy of the proton from first-principles electronic structure calculations,' *J. Phys. Chem. A* **105**, 11534 (2001).

7 Bryantsev, V.S., Diallo, M.S., and Goddard, W.A., 'Calculation of solvation free energies of charged solutes using mixed cluster/continuum models,' *J. Phys. Chem. B* **112**, 9709 (2008).

8 Zhan, C.G. and Dixon, D.A., 'First-principles determination of the absolute hydration free energy of the hydroxide ion,' *J. Phys. Chem. A* **106**, 9737 (2002).

9 Zhan, C.G. and Dixon, D.A., 'Hydration of the fluoride anion: Structures and absolute hydration free energy from first-principles electronic structure calculations,' *J. Phys. Chem. A* **108**, 2020 (2004).

10 Pliego, J.R. and Riveros, J.M., 'The cluster-continuum model for the calculation of the solvation free energy of ionic species,' *J. Phys. Chem. A* **105**, 7241 (2001).

11 Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C., 'Machine learning in materials informatics: recent applications and prospects,' *npj Computational Materials* **3**, 54 (2017).

12 Ward, L. and Wolverton, C., 'Atomistic calculations and materials informatics: A review,' *Curr. Opin. Solid State Mater. Sci.* **21**, 167 (2017).

13 Moorthy, N.S.H.N., Martins, S.A., Sousa, S.F., Ramos, M.J., and Fernandes, P.A., 'Classification study of solvation free energies of organic molecules using machine learning techniques,' *RSC Adv.* **4**, 61624 (2014).

14 Wang, B., Wang, C., Wu, K., and Wei, G., 'Breaking the polar-nonpolar division in solvation free energy prediction,' *J. Comput. Chem.* **39**, 217 (2018).

15 Wang, J., Wang, W., Huo, S., Lee, M., and Kollman, P.A., 'Solvation model based on weighted solvent accessible surface area,' *J. Phys. Chem. B* **105**, 5055 (2001).

16 Marenich, A.V., Cramer, C.J., and Truhlar, D.G., 'Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions,' *J. Phys. Chem. B* **113**, 6378 (2009).

17 Marenich, A.V.K., C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database - version 2012*, University of Minnesota, Minneapolis, 2012.

18 Marenich, A.V., Olson, R.M., Kelly, C.P., Cramer, C.J., and Truhlar, D.G., 'Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges,' *J. Chem. Theory Comput.* **3**, 2011 (2007).

19 Kelly, C.P., Cramer, C.J., and Truhlar, D.G., 'SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute− water clusters,' *J. Chem. Theory Comput.* **1**, 1133 (2005).

20 Pascual-ahuir, J.-L., Silla, E., and Tunon, I., 'GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface,' *J. Comput. Chem.* **15**, 1127 (1994).

21 Sayad, S. 'Support Vector Machine - Regression (SVR).' Biomarkers.ai. http://www.saedsayad.com/support_vector_machine_reg.htm

Chapter 5

Conclusion

A hybrid cluster/continuum model is devised to calculate solvation energy of four ionic molecular groups in aqueous solution. This approach combines a higher accuracy resulting from the consideration of detailed local interactions that are specific to the structure of each solute molecule, while maintaining the computational speed resulting from an effective medium formulation. As the starting configurations we extract water solute clusters of the desired sizes from a large bulk configuration generated using MD simulation, subject to periodic boundary conditions. A systematic variation of the number of water molecules included in these calculations reveals that, depending on the solute size, between four and ten explicit water molecules must be included in the hybrid model in order to account for the most essential local interactions. The larger the solute, and the more complex its structure, the larger is this threshold number of explicit solvent molecules. In a first approach, the cluster geometry is optimized using DFT energy minimization, which yields very accurate solvation energy evaluations about NH_4^+, $CH_3NH_3^+$, HS^- ions and good approximation about OH^-. However, this approach is very time-consuming and can only be reasonably applied to small ions. To encompass a wider range of molecular sizes and structures, we explored a second approach based on DFT single point calculations of a large number of configurations of a given system, sampled along the trajectory from an MD simulation. This procedure yields distributions of solvation

energies with comparable variances. Since the desired measure is constructed from the difference between the solvation energies of a water and a water/solute cluster, we can use the most probable value of each distribution instead of the lowest energy value. Eliminating the need for energy minimization in DFT calculations improves the calculation speed and finite temperature can also be accounted for. The solvation energy tends to converge beyond certain size of clusters. Finally, the inclusion of a counter ion to achieve charge balance has proven necessary for the accurate calculation of the solvation energy calculation in the case of some systems like HS^- and OH^-. In that case, it is also important to identify the correct distance between the counter ion and the central ion.

HEUR thickeners are widely used in the Latex paint to control its rheological properties. The depronation/protonation of the hydrophobe on the HEUR molecules controls the thickening functionality. We have used different models, and considering different environments surrounding the hydrophobe ethoxylated bis(2-ethylhexy) amine when calculating the corresponding pK_a . Our analysis shows that the traditional continuum model cannot provide an accurate prediction of the pK_a. Instead, we need to include the explicit surrounding molecules to properly account for local interactions and thereby improve the calculation accuracy. Based on our finds, adding either explicit water molecules or explicit Latex particle fragments to the system improves the pK_a calculation for ethoxylated bis(2-ethylhexy) amine. When using the same DFT geometry optimization approach, surrounding the hydrophobe with explicit Latex fragments results in a pK_a value that matches the experimental one more closely than surrounding it with explicit water molecules. However, the calculated pK_a that

matches the experimental value best is obtained when surrounding the hydrophobe with explicit water molecules and using the MD sampling approach.

We have applied three different machine learning techniques to predict the solvation free energies of charged species. We found that after tuning the hyperparameters, the random forest regression leads to a model that predicts the solvation free energies most accurately, comparable to the traditional first-principles calculations, which has a comparable mean absolute error of 4 kcal/mol, our machine-learning based model greatly reduces the computational cost and time without losing much accuracy. Our analysis was performed with easily obtainable features: The atomic fraction feature plays the most important role in the solvation free energy prediction, adding Wiener number and solvent accessible surface area of the molecules further improves the performance of the model. We use our model to predict the solvation free energy of the hydrophobe molecule on the HEUR rheology modifier, and it shows a good agreement with the previous result calculated using the first-principles method with only 1.59 kcal/mol difference.

In the future, for the theoretical solvation free energy calculations, Bayesian analysis could be applied to the MD sampling method to further reduce the number of samples we need to get from the MD simulation, this could greatly increase the calculation speed. And it could help us to determine the number of explicit water molecules needed to reach the convergence of the solvation free energy. As for the pK_a study of the hydrophobe molecule, we could include both explicit water and Latex sample fragment in the solvation free energy calculation, and see how the solvation free energy and pK_a change with such local environment. This could be a challenging task because the degree of freedom increases significant thus it would be really

difficult to find the low-energy structure for the DFT geometry optimization approach. MD sampling method would be a better approach, but it may require a large number of water molecules to converge. Lastly, the prediction accuracy of the machine learning model could be further improved by increasing the sample size and adding more clearly non-correlated features that are relevant to the solvation free energy.

As soon as we validate a more accurate way to calculate/predict pK_a, we could examine more candidate amine hydrophobe molecules as well as construct a series of new candidate molecules and evaluate their usefulness based on their pK_a. Once we have identified a most promising subset of candidate hydrophobe molecules we could simulate the protonation/deprotonation process and idenfity the charge regulation mechanisms for given hydrophobe molecular architectures. The insights gained in this detailed investigation will advance our fundamental understanding of thickener functionality and efficacy, and will allow for a rational design of novel hydrophobe on the thickener.