Ionic Force Field Parameterization with Bayesian Optimization

Yuchen Zhu

Delft University of Technology

TU Delft Delft University of Technology

Challenge the future

Ionic Force Field Parameterization with Bayesian Optimization

Yuchen Zhu

in partial fulfillment of the requirements for the degree of

Master of Science

in Mechanical Engineering at Delft University of Technology

Student ID:	4879058	
Report number:	3020	
Thesis committee:	Dr. Ir. R. M. Hartkamp,	TU Delft, supervisor
	Ir. M. F. Döpke,	TU Delft, supervisor
	Dr. Ir. R. Pecnik,	TU Delft
	Dr. O. Moultos,	TU Delft
	Dr. D. J. P. Lahaye,	TU Delft

Complex Fluid Processing, Faculty of 3mE, Delft University of Technology

TU Delft Delft University of Technology

Acknowledgements

Memory flashes back to the very first day when I arrived at TU Delft campus. Now that my master study here almost comes to an end. I would like to thank some fantastic people who accompanied me along this journey.

First of all, I would like to give my sincere gratitude to my supervisors Dr. Ir. R. M. Hartkamp and Ir. M. F. Döpke. They introduced me to the molecular simulation world and taught me valuable knowledge. They provided me kind support and help when I was frustrated about the simulation results. They inspired me continuously on thinking and writing. They taught me how to be an eligible researcher. So, I could have a memory that I will cherish for the rest of my life.

I am thankful to Dr. O. Moultos for imparting me knowledge in the early stage of my project. Talking with him not only broadens my horizon of related topics but also passes his optimism to me. I would also like to thank Dr. Ir. R. Pecnik and Dr. D. J. P. Lahaye for joining the committee and taking time to evaluate my thesis.

I am grateful for all friends I have encountered during my stay at Delft. I really enjoyed the discussions and mutual supports we had in the past two years. I feel lucky to have their accompany. I would also like to thank my family for always supporting my decisions and showing their unconditional love to me.

Last but not least, I want to say thank you to TU Delft, I stumbled, I fell, I stood up, I grew up. I will never forget this amazing journey.

Yuchen Zhu
Delft, Novmber 2020

Abstract

Force field is widely used to model the potential energy in atomistic simulation systems. Despite force fields have a concise mathematical form, a good set of force field parameters usually requires extra care of calibration. Besides, numerous ionic force field parameters are reported from various sources as researchers have specific target properties for their interests. Previous studies mainly used brute force optimization to find the most desired set of parameters in ionic solution. However, these methods are not efficient since the evaluation of the performance of a parameter set is time-consuming.

This work used a stochastic optimization routine in machine learning to tackle the problem of black-box function optimization. This method shows excellent performance of locating the optimum regions of the black-box cost function in only a few iterations. To evaluate the performance of a set of ionic force field parameters, MD simulations are carried out in LAMMPS to compute ionic properties. The solvation free energy and ion oxygen distance are selected as the primary targets while the self-diffusion coefficient and contact ion pairs are regarded as the secondary targets. The optimum region of primary targets are found by direct optimization, then secondary targets are studied with optimized parameters of the primary targets. There have been found discrepancies between the optimum regions of different targeted properties. The dependence studies of individual ionic force field parameters (ϵ, σ, q) are analyzed and parameterization trends are found out. Base on these trends, the final calibration model is proposed.

Contents

List of Figures

List of Tables

Nomenclature

ϵ	Lennard-Jones energy parameter
$\mathcal{N}()$	Normal distribution
$\mu(x)$	Mean at test point x
σ	Lennard-Jones size parameter
$\sigma(x)$	Standard deviation at test point x
ε_0	Permittivity of vacuum
$a()$	Acquisition function
D_i	Self diffusivity
m	Mass of ion
q	Charge of ion
BO	Bayesian Optimization
CDF	Cumulative distribution function
CIP	Contact ion pairs
CN	Coordination number
ECC	Electronic charge correction
GP	Gaussian Processes
IOD	Ion oxygen distance
JC	In Suk Joung and Thomas E Cheatham III
L-J	Lennard-Jones
LAMMPS	Large-scale atomic/molecular massively parallel simulator
MD	Molecular Dynamics
MSD	Mean square displacement
PDF	Probability density function
SFE	Solvation free energy
TI	Thermodynamics Integration

1 Introduction

1.1. Motivation

The water from our daily life always contains a certain amount of dissolved salts. Therefore, research into aqueous electrolyte solutions is of great significance for both industrial applications and research understandings. In industrial applications, ions affect various physical chemistry phenomena, including altering the reaction rate and changing solubility limit of the solution. In seawater, for example, salts can play a big part in determining the solution quality and the subsequent water processing. In life sciences, the study of aqueous electrolyte solutions can be beneficial to understand the thermodynamic properties of living organisms [7, 36, 81]. However, drastic variations of aqueous electrolyte solutions from system to system make the direct experimental measurement unfeasible. Alternatively, computer simulations can be of great assistance to gain a better understanding for various system conditions.

In the early stage of computer simulation, the solvent water was described implicitly as a continuum model with no structure[16]. The presence of water was taken into account only by its dielectric constant. This way of treating solvent has an advantage of low resources demand, since water is the richest component in the aqueous electrolyte solution. However, this way might not work well where the salt concentration is high. Some important ionic properties, e.g., the solubility limit, requires a good representation of ion-ion and ion-water interactions. Therefore, more accurate simulation routine should be carried out.

To acquire more robust information of concerned phenomena, researchers prefer to use systems with extensive sizes, long simulation duration, and complex combinations of species. Therefore it is prohibitively expensive to use quantum mechanics approximations for every configuration one has encountered. The use of empirical formulas is a preferable alternative[2, 21]. Classical molecular simulations employ these empirical formulas to reproduce the results from quantum calculations or experimental measurements[66], and collectively model the potential of an atomistic simulation system[24, 36, 66, 81]. The functional form and parameter sets of these empirical formulas are termed as the force fields. Although the form of empirical force fields is concise, the determination of a good force field is not a trivial task. The accuracy of simulation outcomes depend heavily on the selected force field. If not parameterized with care, a highly biased force field might yield results that deviate the target value by several orders of magnitude. Optimizing a good force

field subsequently has significant research values. The end goal of the force field optimization is to find the simplest essentially representation of an atomistic system, which allows researchers to carry out simulations in an efficient and accurate way.

Once a good force field is developed, researchers can gain understandings for configurations where the experimental measurements are inaccessible. Moreover, simulation reveals information of microscopic properties, e.g. single ion properties, that can be difficult to obtain from experiments.

1.2. Goals

Force field optimization is a tedious process as it requires a lot of trial and errors to find the best parameters for an empirical formula. Many researchers have put much effort to only gain a slight improvement in the performance of a force field. Moreover, due to the empirical nature of the force field, there is no perfect parameterization and trade-offs must be made. The need exists for a more efficient calibration model regarding the time costly parameterization process. To fulfil this purpose, the goals of the present study are:

- (1) *Formulate an efficient alternative approach to sample the force field parameters with the help of a probabilistic Machine Learning model.*
- (2) *Explore the ionic force field parameter space and estimate the parameterization trends of target ionic properties.*
- (3) *Identify the possible cheap simulation setup for speeding up the parameterization process.*
- (4) *Propose the major procedures for ionic force field parameters design. Balance the trade-offs for selecting different properties.*

1.3. Thesis Outline

This thesis is organized as follows. In Chapter 2, we provide a literature survey about the related study in this research community. Then, we discuss the background in two separate chapters. Chapter 3 covers some techniques in Molecular Dynamics (MD) simulation and computational methods for 4 target properties: solvation free energy (SFE), ion-oxygen distance (IOD), ionic self diffusivity (D_i), and contact ion pairs (CIP). Chapter 4 introduces Gaussian Processes (GP) and Bayesian Optimization (BO). In Chapter 5, we will discuss the methodology and the setup for MD simulations. Due to the time-costing parameterization process, we search for the best simulation setup to balance the efficiency vs. accuracy dilemma. In chapter 6, we will first introduce the optimization model. Then we will provide the optimum region of force field parameters we have found for different ionic species and evaluate the performance of our model based on MD results from literature[18, 36]. Next, parameterization trends are found to evaluate different parameter sets for various properties. Finally, we summarize the findings in this study and provide future recommendations.

2

Literature Survey

2.1. Force Fields

Force fields usually refer to the functional forms and parameter sets of the selected potential energy function. Since force fields are usually optimized through empirical fitting, they only approximate the bonded and non-bonded interaction in an atomistic system. Through representing the potential energy of a system in this simple way, we can carry out simulation with different spatial and temporal scales. The most common terms in the functional form are Lennard-Jones, Coulombic, bond, angle, and torsion terms, as it is shown in Equation 2.1.

$$\begin{aligned}\mathcal{U} = &\sum_{\text{bonds}} \frac{1}{2} k_r \left(r_{ij} - r_0\right)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta \left(\theta_{ijk} - \theta_0\right)^2 \\ &+ \sum_{\text{torsions}} \sum_{n} k_{\phi,n} \left[\cos\left(n\phi_{ijk\ell} + \delta_n\right) + 1\right] + \sum_{\text{non-bonded}} \left[\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right]\end{aligned} \tag{2.1}$$

This expression of potential can be viewed as the spring sphere system in classical mechanics plus the non-bonded interactions. The stretch of bonds between molecules, the bend of two bonds, or even the torsion of a single bond can contribute to the potential energy. The very last term in Equation 2.1 represents the non-bonded potential in atomistic systems, i.e. vdW and electrostatic potentials.

The functional forms employed for describing particular terms in the potential function might vary in mathematical format, but the underlying physics is essentially the same. For example, the repulsion in Lennard-Jones potential is represented by r^{-12} like term, whereas in the Buckingham potential, it is represented by an e^{-r} like term. Hence, it should be pointed out the fitting these functional forms might contain biases from the very beginning.

2.2. Water Models

Water molecular models are designed to compute various properties in biophysics simulations. For the ionic force field parameterization, the first concern is to select a good

water model. There are pros and cons for polarizable and non-polarizable water models, as well as the flexible and non-flexible water models. Good summaries can be found in the two references[64, 76]. The polarizable water models theoretically have a better performance in describing the Coulombic interactions. However, a good parameterization of polarizable water model comes at the expense of extra efforts and additional computing resources[7, 81]. Despite this, the performance of a good non-polarizable water model is quite similar to that of a state-of-art polarizable water model[7, 64, 81]. Therefore, rather than putting hope to optimize a perfect polarizable force field[7, 81], many researchers still prefer to devote their energy into parameterizing a good rigid non-polarizable water model[7, 42–44, 79, 81].

The commonly used rigid non-polarizable water models are SPC/E (Berendsen et al., 1987[8]), TIP3P (Jorgensen et al, 1983[35]), TIP4P (Jorgensen et al, 1983[35]), TIP4P-Ew (Horn et al., 2004 [29]), and TIP4P/2005 (Vega et al., 2005[1]). SPC/E and TIP3P have 3 charge sites located at the 3 atom positions of the water molecule, as well as one Lennard-Jones site located at the oxygen atom. The 3-site water models are widely used for their computational efficiency[8, 35, 36]. With the increasing of computing power, more efforts have also been devoted to the parameterization of higher sites water model, i.e, 4 sites. The TIP4P (transferable intermolecular potential with 4 points) water model represents a family of the 4 sites water models[1, 29]. These water models have 2 hydrogen atoms, 1 oxygen atom, and 1 dummy atom along the bisector of HOH angle. One L-J interaction site locates at the oxygen position, while three Coulombic interaction sites locate at two hydrogen positions and one dummy atom position. This treatment allows a better description of electrostatics distribution around water molecules. A variety of 4 site water models have been developed. For examples, the TIP4P-Ew is a reparameterized version of the TIP4P model for the use of Ewald summation methods[29]. TIP4P/2005 water model, proposed by Abascal and Vega, is used to simulate the entire phase diagram of condense water. Table 2.1 gives an overview about force field parameters of these two water models.

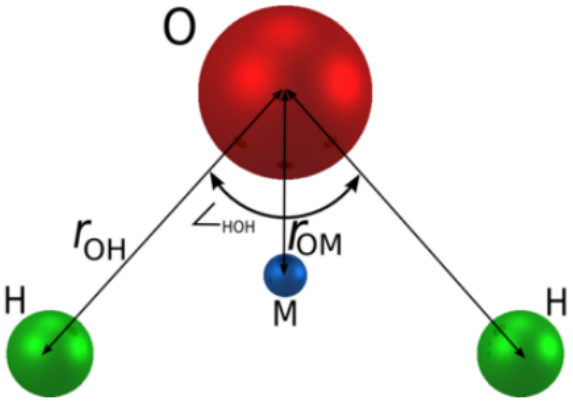


Figure 2.1: The structure of TIP4P water model, picture is taken from [68].

Table 2.1: Force-Field parameters for water models.

	TIP4P-Ew	TIP4P/2005
H-O-H [°]	104.52	104.52
r(O-H) [Å]	0.9572	0.9572
$\sigma_{\text{O-O}}$[Å]	3.16435	3.1589
$\sigma_{\text{H-H}}$[Å]	1.00000	1.00000
r(O-M) [Å]	0.125	0.1546
$\epsilon_{\text{O-O}}$ [Kcal/mol]	0.16275	0.185205
$\epsilon_{\text{H-H}}$ [Kcal/mol]	0.00000	0.00000
q_o[e]	-1.04844	-1.1128
q_H[e]	+0.52422	+0.5564

2.3. Parameterizations

After selecting a good water model, researchers focus on optimizing the ionic force field parameters to match this choice[7, 23, 36, 81]. Then with some so-called combination rules, a good description of water-water, ion-ion, and water-ion interactions can be found. The ion-water interactions play an important part in determining ionic properties like solvation free energy and ion oxygen distances, so a good water model is of vital importance for representative simulations. Many previous studies have developed ionic force field parameters with water models like TIP3P and SPC/E, however, a few studies have optimized ionic force parameters with newly developed water model TIP4P/2005. The parameterization of ionic force field with rigid water models only concerns the non-bonded term, i.e., Coulombic potential and Lennard Jones potential in Equation 2.1. For many years, ionic parameters that fit the target properties were developed through trial and error process.[7, 23, 36, 48, 79, 81]. This is because tuning the empirical parameters of ions to fit quantum calculations or experimental results is a systematic trial-and-error process. Moreover, the conventional ionic force field parameters are optimized for a limited range of species. This way of parameterization is not systematic, and it clearly has some drawbacks like misbalances of the target properties and other properties. The obtained parameter sets cannot reproduce simulation results that match all properties simultaneously since different researchers might have different interests. Therefore an ion force field are developed to serve specific purposes only, e.g., reproduce the solubility according to the experimental measurements[7, 36, 79, 81].

The pioneer studies of ion force field development can be traced back to the work of Smith-Dang-Garrett and Åqvist in the 1990s [3, 15, 69]. The parameterizations of these pioneers are included in famous MD and force field software, e.g. the AMBER [13] and CHARMM[12]. Interestingly, as the ionic parameters are developed, the targeted experimental results have also been improved as more accurate measurements have been achieved. These parameterizations that existed in these packages are somewhat arbitrary and inconsistent. This inconsistency leads to some non-physical phenomena like direct cation-anion clustering even at a low concentration, which can underestimate the solubility limit. [5, 36].

After observing the drawbacks of the existing parameters, Joung and Cheatham[36] performed an early reparameterization study in 2008. They performed L-J parameterizations of different monovalent ions with the solvation free energy and lattice energy of salt as the initial targets. Then the optimization procedures were extended to different combinations

of water models and ions. The significance of their work is that they scanned the L-J parameter space , which shows different combinations of L-J parameters can yield the same solvation result. After Joung's work, many studies related to ionic force field parameterization followed this procedure of finding the correlation of L-J parameters, simply for it provides a good criterion for parameterization [23, 28, 45, 48, 79]. They usually first determine this correlation, then shift the attention to other ionic properties so the final optimal set can be obtained.

For many years, solvation and structural properties were regarded as the golden standards for ionic force field parameterization[3, 23, 28, 36, 45, 48, 79]. Until recently some studies have aimed for other exotic properties, e.g., osmotic pressure, activity coefficient, dynamical properties, and solution properties [7, 23, 48, 81]. These studies greatly explored the capability of existing ionic parameters, and they have shown that there are still needs for developing more sound force fields. Leontyev and co-workers recently challenged the conventional idea that ionic charge should be an integer; they believe that non-polarizable models do not account for the electronic contribution to the dielectric constant. By simply scaling ionic charge, one can include the electronic screening effect of the medium and improve the simulation performance. This scaling charge treatment is termed as the Electronic Charge Correction (ECC)[42, 43]

The pioneer studies which applied electronic charge correction to parameterize solution properties are carried out in Vega's group [7, 81]. The force field developed by them is known as the Madrid force field. Benavides et al. proposed a parameter set for sodium chloride with the TIP4P/2005 water model[7]. Their primary target properties were the concentration dependency of solution density and a variation of solubility determined by the difference between the chemical potential in solid and solution states. It is worth noting that they tried a multidimensional Taylor expansion technique to determine optimums for some properties[7]. But this gradient based approach did not worked well due to that the derivative formation for the target properties is a prohibitively expensive task. Their final strategy was to start with a Taylor expansion for a few properties, followed by many trial and error calibrations. Benavides' study have revealed that implementation of ECC indeed have a positive effect on improving the performance of some targets[42, 43]. From the same group, Zeron et al. expanded the investigation of the Madrid force field with scaled charge to a next level[81]. They started by targeting the density of aqueous solution at different concentrations, then targeting the structural properties and the contact ion pairs that could indirectly indicate the solubility limit[81]. It is also important to point out when they were developing the force field parameters; they abandoned the conventional idea that the cross interaction between different species must follow some so-called combination rules. They specifically parameterized the interactions between different species, and therefore the force field performance is improved greatly. Their work has revealed that the scaling charge and directly parameterized ion-ion interaction can solve the unphysical phenomenon that salts will spontaneously aggregate below the solubility limit[5, 81].

Correlation of L-J parameters

In the work of Joung and Cheatham, 2008[36], it was first found that the optimums for solvation free energy locate on a curve like region in the L-J parameter space. This is because the effect of increasing the ions σ parameter can be countered by decreasing the ions ϵ pa-

rameter, so different combinations of (ϵ, σ) can yield the same solvation free energy result. In this type of approach, one can follow a well-designed procedure since the correlation between two L-J parameters can reduce the degrees of freedom of force field parameters by one, because two L-J parameters are correlated with each other, e.g. $\epsilon = f(\sigma)$. First, they find parameters that give the solvation results match the experimental values. Then other targets were taken into account base on the correlation of L-J parameters for solvation energy. The correlation map of two independent variables can give useful guidance for the parameterization process. It provides critical information for selecting the new parameter set. Researchers can also use these maps as good references for their specific purposes: either to optimize a new parameter set that matches the desired property or to avoid exploring the irrelevant region and narrow down the search extent.

Many previous studies that considered the solvation free energy as a target property found this optimum region with brute force methods like the grid search. This is because the ionic target property is a black-box of the force field parameters. For this consideration, the grid search method is reliable at the expense of additional computing resources. Take the study of Joung and Cheatham for an example[36]. To find parameters fitting the experimental solvation free energy, they first chose a large range of ϵ and σ to constrain the calculation domain. Then intervals of sigma and epsilon are chosen to generate the grid points. The total number of grid points in the work of Joung and Cheatham is around 400-500. Next, for each grid point in the mesh, the hydration free energy is calculated. The best matches with experiments are then chosen as the optimal parameter sets. These sets can further be fitted with a mathematical formula as a function of sigma and epsilon. However, one issue is that the quality of the correlation between two variables depends hugely on the density of the mesh. Another drawback of this method is that only a small portion of computer power ($1/L$, L being the mesh length for grid points) is spent near the optimum region for finding the optimums. The grid search samples the data at the grid points, so computing power will increase drastically if the search domain becomes large.

After the work of Joung and Cheatham[36], other studies of designing ionic parameter sets based on the correlation of L-J parameters had also been reported[23, 28, 36, 45, 46, 48]. In the year 2009, Netz and co-workers published a series of papers regarding force field parameter design using L-J parameters correlation [23, 28, 48]. They mapped out the solvation free energy (ΔG_{sol}) and solvation entropy (ΔS_{sol}) hypersurfaces[28] in one of their studies[28]. Then, in their subsequent works, they specifically designed force field parameters for divalent cations[48] by using the solvation free energy hypersurface as a first starting point. Next, the radial distribution function (RDF) is used as a check for the structural properties. Finally, the activity derivatives are calculated via the Kirkwood-Buff solution theory to determine the final optimal L-J parameter[48]. They directly mapped the solvation free energy as a function of L-J parameters, G = $f(\epsilon, \sigma)$. Unlike finding the correlation of (ϵ, σ), the hyper-surface fitting is considered non-robust and prone to errors[22], since the mathematical expression is pre-defined which might fail in capturing the true shape of the underlying data. Li et al[45] also have adopted this way of parameterizing ionic force fields. They extended the design of L-J parameters for +2, +3 and even +4 metal cations[45, 46]. They determined the final parameters set with the help of L-J parameters of noble gas. Their optimization framework has proved that this way of considering solvation as the primary target then fitting the secondary targets according to some criteria can be generalized to optimize exotic force field combinations, e.g. high charge metal ions,

and achieve promising results. Yagasaki et al. revisited this correlation and performed a parameterization study to reproduce the solubility of NaCl and KCl[79]. In their first treatment for starting optimization, the correlation between ϵ and σ is found by first fixing ϵ as a constant and then moving σ to match the experimental results. They reiterate this procedure for different ϵ, so a set of optimal (ϵ, σ) can be found. Finally, the solubility is matched according to crystalline potential energy. However, their optimization process is still essentially a variation of the "grid search". Since it went through combinations of (ϵ, σ) in a one by one manner.

Transferability

Testing performance for combinations of force field parameters from different sources are usually referred as the transferability study. This is typically important for evaluating complex ionic systems. The transferability study of ion parameters has been favored by researchers [18, 23, 36]. This is because: (1) An optimized parameter set only satisfies specific target properties, whereas it requires a lot of efforts. (2) There are many existing parameterization works that have been done previously in this community. (3) Many of these works have optimized the ion parameters independently. Therefore researchers always prefer to cross-evaluate the existing ion parameters than actually carry out systematic parameterization. However, it should be noted that this inconsistent way of parameterizing and mixing parameters from different sources naively is the culprit for degraded performance of complex ionic systems.

Many new parameterizations that have been done are developed based on previous works [7, 36, 48, 79, 81]. One particularly interesting case of all these parameterizations is the Li^+ with TIP4P like water: parameters taken from TIP4P/2005&Madrid (water&ion force field combination) [81] is exactly the same as the parameters taken from TIP4P-Ew&JC [36], both as (ϵ, σ) = (0.10398 Kcal/mol, 1.43970Å).

Machine Learning of Force Fields

Due to its empirical nature, tuning a force field to match experimental results requires tons of trials, as well as insightful understandings into the physical chemistry phenomena from the researchers. Given the blooming of artificial intelligence technologies in recent years, many researchers have put their efforts to find alternative automatic ways of solving complex atomic simulation tasks. These methods can be divided into two categories: (1) Learning complex potential with the help of deep neural networks[27, 70, 77]. (2) Kernel-based surrogate models[6, 51, 52].

The first type of method can directly predict system's potential, and it has gained much popularity in recent years. The most representative work of potential learning with deep neural networks can be found in a series of papers published by Zhang and co-workers[77, 82–84]. The key idea is to feed the pair-wise distances to the deep neural network for learning the system potential. This can avoid the pair-wise potential calculation and gives almost linear scalability of operation complexity to the system size ($\sim \mathcal{O}(N)$). This model requires *ab initio* data, and it is first principle based. Representing the system potential via machine learning methods can not only reduce the time consumed on MD simulations but also make good use of the existing optimized learning algorithms to speed up the calcula-

tions.

The early study of using the Bayesian statistic approach for automated force field development of coarse-grained models can be found in the work of Liu et al.[47], in which they searched a better coarse-grained model from the parameters region where the sampling numbers are limited. Later on, Dequidt[17] also utilized Bayesian inference to select the most desired coarse-grained force field model in their work. Recently, studies of force field parameterization with Bayesian optimization have also been carried out by McDonagh and co-workers[51, 52], they have shown that from millions of discrete candidates, a good force field parameterization can be reached via dozens of Bayesian iterations. It should be pointed out that the force field optimization itself is a time-consuming problem. For example, the time cost of designing a targeted force field is brought down from 16 weeks human calibrations to 1.5 weeks automatic Bayesian searching in McDonagh's work[52]. The key idea of all these Bayesian based methods is to search for a superior force field result with less human labor involved.

3

Molecular Simulation

This chapter includes used MD simulation techniques and computational methods for different ionic properties. Section 3.1 introduces several important concepts of atomistic simulation. Section 3.2, 3.3 and 3.4 introduces the computational methods for solvation free energy, self diffusivity and structure related properties.

3.1. Atomistic Simulation

Molecular Dynamics

This study used Molecular Dynamics simulation to compute ionic properties. Through MD simulations, one can verify the observations from experiments and carry out simulations under extreme conditions, e.g. extreme temperature and toxic environment. The role of simulation is to apply certain amount of assumptions to the mathematical model and obtain results within some accuracy limits according the computation budget. Essentially, the MD simulation solves Newtonian physics through numerical time integration, particles will transfer energy and momentum to each other via electrostatics and van der Waals interaction during the simulation. This way of solving Newtonian dynamics can provide pretty good results for numerous system properties. There are 3 main steps in a typical MD simulation: (1) Initialization. (2) Equilibration. (3) Production.

In the initialization stage, a good starting system should resemble the state of interest in equilibrium.. If one starts the simulation at an irrelevant configuration then he might spend most of resources to equilibrate the system. The configurations of local energy minimum can be found through many minimization algorithms[67]. Then the initial velocities that follow Maxwell-Boltzaman distribution at the given temperature can be assigned to particles. The role of equilibration is to bridge the initial system and the system of interests. After the system reached the thermodynamic equilibrium, one can decide what data to store and develop specific post-processing procedures, this phase is usually referred as the production.

The potential

The system of our interests contains rigid water molecules with 1 L-J sites and 3 point charge sites, ions with 1 L-J site and 1 point charge site. Thus, the total non-bonded potential of the system is the summation of L-J potential and the Coulomb potential, which has an explicit expression:

$$u_{ij} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{l}\sum_{m}\frac{q_l q_m}{4\pi\varepsilon_0 r_{lm}} \tag{3.1}$$

where the u_{ij} represents the potential energy between particle i and particle j, l and m represent the number of point charge sites in each molecule. The double summation in Equation 3.1 represents summing over all charge sites in ith and jth particle. σ_{ij} is the L-J size parameter, it can be viewed as the diameter of the particle. ϵ_{ij} is the energy parameter, which represents the depth of the potential well, or the minimum energy of the L-J potential. ε_0 is the vacuum permittivity.

A detail representation of two L-J parameters can be found in Figure 3.1. At $r = \sigma$, the inter-molecule potential is 0, while at $r_m = 2^{1/6}\sigma$, the potential reaches the energy minimum. The inter-molecule forces can be obtained by taking the derivative of the potential $F(r) = -dU(r)/dr$. The expression of L-J potential contains two terms: the 12th power term is the repulsion term while the 6th power term is the attraction term. When $r < r_{\mathrm{m}}$, the interaction is dominated by repulsion, When $r > r_{\mathrm{m}}$, the interaction is dominated by attraction.

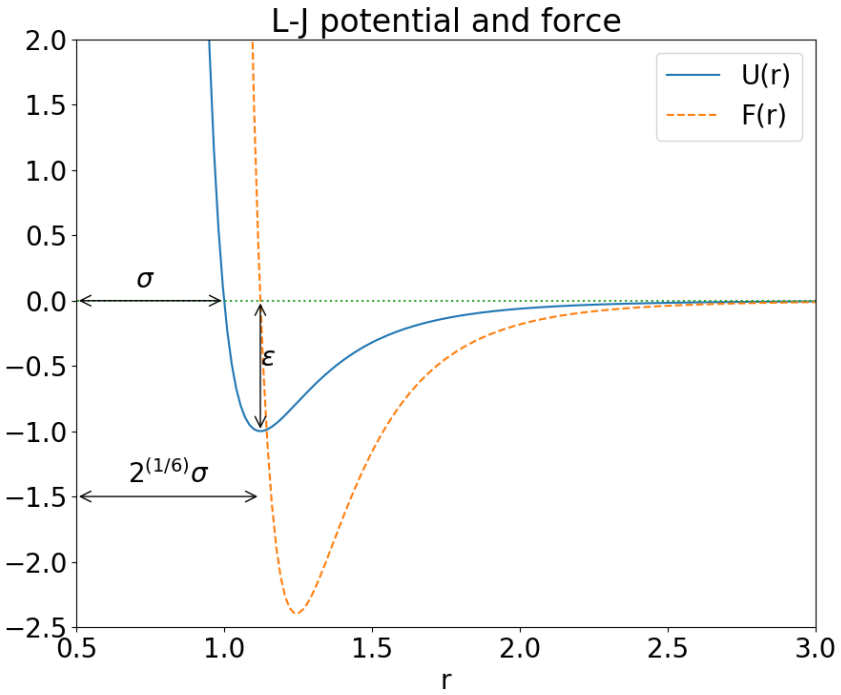


Figure 3.1: L-J potential and corresponding inter-molecule force.

To account the pair-wise potential between different species, the Lorentz-Berthelot (LB)

combining rules can be used:

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2}$$
$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{3.2}$$

Long range interactions

For a system containing N particles, the computational complexity of calculating pair-wise potential is of $\sim \mathcal{O}(N^2)$. The L-J interaction at long distance is negligible since the attraction decrease with $(1/r)^6$. Only a small magnitude of the L-J potential exist at long range, see Figure 3.1. Account all these values whose magnitude close to zero is not necessary. Therefore, a cut off from which the long range potential can be turned off is introduced. Smit and Frenkel suggested a cut off $r_c = 2.5\sigma$ to truncate the potential [21]. This empirical choice of cut off is useful, at $r = r_c$, the potential is only $\epsilon/60$, see Figure 3.1. The expression of a truncated L-J potential is:

$$u_{ij,\text{truncated}}(r) = \begin{cases} u_{ij}(r), & r \leq r_c \\ 0, & r > r_c \end{cases} \tag{3.3}$$

The computational complexity will be reduced to roughly $\mathcal{O}(N)$ after applying the cut off, but the dominant resource requirement is still calculating the pair-wise distances inside the cut off r_c. The remaining tail correction can be calculated and added to the final expression. A shifted form, which moves the entire curve upper, is applied to the potential to avoid the discontinuity at $r = r_c$. It is worth noting that the choice of cut-off should also consider the types of target property [21, 74]. For properties only depend on free energy, it is desirable to choose a long r_c since they are very sensitive to the cut-off distance. While for other properties related to forces, the 2.5σ choice is sufficient. This is due to that the derivative of potential after the truncated modification remains the same within the cut-off range, there is only a small discontinuity of derivative at the cut-off location.

The electrostatic interactions are treated as point charge model which has the expression that scales with $(1/r)$ as it can be seen in Equation 3.1. This decay slowly converges to zero at long distances. So a simple truncation of electrostatic interactions will cause artifacts. The particle-particle-particle-mesh (PPPM) can be employed for dealing with long-range forces. The PPPM method deals with short range interactions in real space, while the long range interactions are transformed into the Fourier space. This allows one to make use of highly efficient numerical Fast-Fourier Transform (FFT) for computing the long-range interactions, which gives a speedup compared to direct summation. The computation complexity drops from $\mathcal{O}(N^2)$ to $\mathcal{O}(N\log(N))$ with PPPM method employed.

Periodic boundary conditions

With a fixed system size, the ratio between surface and volume $(1/L)$ is much larger than it is in the real scenario. To fix the unreal surface volume ratio, periodic boundary conditions (PBC) are implemented. Take a cubic box for example, the box is replicated in all spatial directions and a pair-wise interaction is considered only for the shortest distance of all these images, see figure 3.2. It should be pointed out that the PBC is applied with a spherical cut-off in MD, this is not shown in this figure. If the minimum image convention is used, then

the pair-wise interaction in the nearest image is considered, which results in non-constant potential on the surface [21].

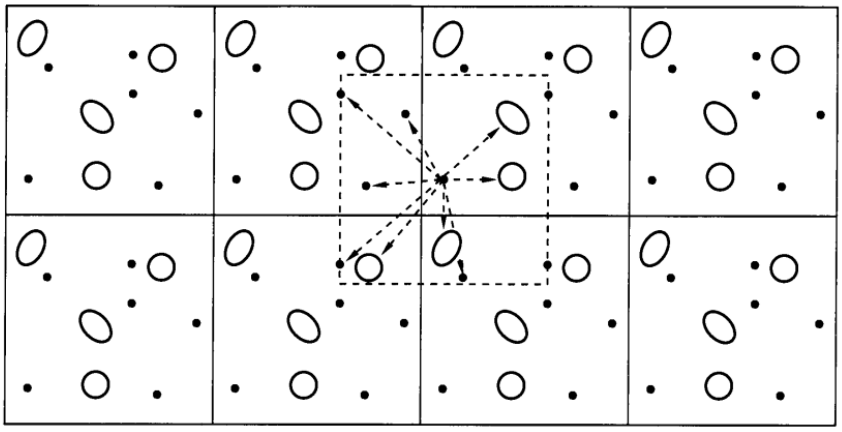


Figure 3.2: Periodic boundary condition, picture is taken from [21].

Integrators

We cannot find the analytical solutions to dynamics of complex atomistic system. So we should approximate equations of motion in a discrete manner. Numerical integration can be used to solve discretized equations of motion. During the approximation, one crucial property of integrator is that it must preserve phase-space volume, which makes sure the ensemble of interest is conserved.

There are many well designed algorithms available to be used for time integration [2, 21], among which one popular type is the Verlet like algorithms. These integrators expand the atomic positions up to the second-order term, i.e., including the forces, to update the positions and velocities of particles in the system.[21]

An appropriate timestep should also be selected for the integrator. If the timestep is too large, the integrator might fail to capture fast dynamics. Even worse, the energy might explode due to the overlap of atoms when solving the equations of motion. If the timestep is too small, it will cost too much time to achieve the same simulation length. Even worse, the truncation error due to loss of precision during the scientific computation will be accumulated. To maintain the accuracy while solving equations of motion, one should determine the length of timestep according to the fastest dynamics in the system, which is usually the bond (e.g., O-H bond) vibration. In our simulations, we have used rigid water models with the SHAKE algorithm to constrain the geometry. These rigid models allow a larger timestep to be involved, and therefore the computational efficiency is also improved.

Thermostat

In MD simulation, the common ways of regulating system variables can be sorted to 3 types: (1) Constrain the system variables (e.g. temperature) to some preset distributions. (2) Rescale system variables (3) Extend system variables. Among all these regulating methods, the Nosé-Hoover (NH) thermostat or barostat based on extending system variables to include temperature or pressure are considered to be most popular and reliable. This

method introduces a fictitious degree of freedom which has the dimension of mass and can interact with the system. This way of weak coupling can make sure the dynamics of system is well-preserved. But due to this extra mass is fictitious, the fluctuation of controlled variables is obvious. The performance of a Nosé-Hoover thermostat and other control methods can be found in Figure 3.3.

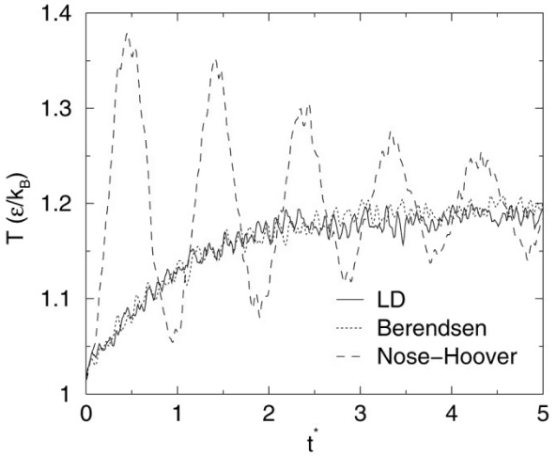


Figure 3.3: Temperature response of different thermostat, picture is taken from [75]

3.2. Solvation Free Energy

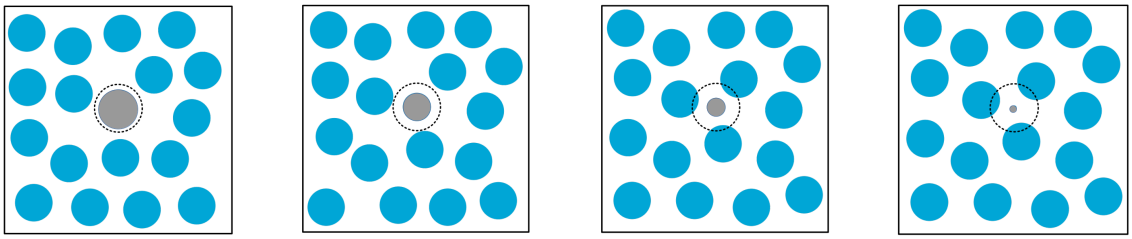
There are many MD based routines to simulate and analyze the solvation result[39], among which the numerical thermodynamic integration with perturbation-based esimation is quite popular[2, 21, 39, 67].

Finite difference thermodynamic integration

To estimate the solvation free energy of ions, one needs to calculate the energy differences between two aqueous states with and without ions. However, the free energy of a state is not merely a function of the coordinates. Instead, it is related to the partition function for that thermodynamic state. Therefore we cannot simply perform algebraic operations between two potential energy values to get free energy difference of two states. A thermodynamic integration can be employed to pave the path between the two thermodynamic states. Then the free energy difference is calculated by constructing a free energy path and integrating over ensemble-averaged enthalpy change along this path[2, 21]. While constructing this path, the conventional treatment is to find as many alchemical intermediate states as possible so that the two end states can be smoothly connected to each other. In term of our case, we have one end state where there is only solvent in the system. While at the other end state, the solute (the ion) interacts with the solvent (water molecules). Water molecules will surround this single ion due to polarity, so a hydration shell of water will be formed.

There are only non-bonded interactions between ions and water molecules in the ionic solutions. They are, specifically, the van der Waals interactions and electrostatic interactions. So a two-stage thermodynamic integration is formalized as follows. In the first stage,

the ion was slowly neutralized by reducing the charge step by step with the soft Coulomb potential, see Figure 3.6a . Then in the second stage, the vdW potential is slowly removed with the soft L-J potential, see Figure 3.6b. In both stages, a coupling parameter λ is introduced that varies from 1 to 0 to bridge the end states. A good representation of effect of this λ can be found in Figure 3.4.



(a) $\lambda = 1$ (b) $\lambda = 0.6$ (c) $\lambda = 0.3$ (d) $\lambda = 0.1$

Figure 3.4: The effect of interaction parameter λ to the system. While $\lambda = 1$ the interaction is fully on. While $\lambda = 0$ the interaction is off. This λ parameter controls how much the ion is allowed to interact with the system.

By considering the Helmholtz free energy as a function of coupling parameter λ, the fundamental theorem of calculus states:

$$A_1 - A_0 = \int_0^1 \frac{\partial A(\lambda)}{\partial \lambda} d\lambda \tag{3.4}$$

The expression for Helmholtz free energy is found to be:

$$A(\lambda) = -k_B T \ln \int \exp(-U(q,\lambda)/k_B T) dq \tag{3.5}$$

where $Z(T,V,N) = \int \exp(-U(q,\lambda)/k_B T) dq$ is the partition function for canonical ensemble. Substitute Equation 3.5 to Equation 3.4 one can get

$$\Delta A = \int_0^1 \left\langle \frac{\partial U(\lambda,q)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{3.6}$$

Note that the accurate ensemble average should be used here is the one considering the volume change, which has the explicit form[2, 67]:

$$\Delta_0^1 A = -kT \sum_{i=0}^{n-1} \ln \frac{\left\langle V \exp\left(-\frac{U(\lambda_{i+1})-U(\lambda_i)}{kT}\right)\right\rangle_{\lambda_i}}{\langle V \rangle_{\lambda_i}} \tag{3.7}$$

In practice, the volume fluctuation in the equilibrated state is negligible[2]. The results of using Equation 3.6 and 3.7 are nearly identical. Therefore Equation 3.6 is chosen for calculating the solvation free energy for the isothermal–isobaric (NPT) ensemble.

To evaluate the derivative of potential energy with respect to λ term inside the ensemble average brackets, one can use the perturbation method[54, 67]. In this method a very small perturbation parameter δ is chosen to approximate the derivative term numerically. Then the free energy is found to be:

$$\int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \approx \sum_{i=0}^{n-1} w_i \left\langle \frac{U(\lambda_i+\delta) - U(\lambda_i)}{\delta} \right\rangle_{\lambda_i} \tag{3.8}$$

where w_i are the weights of numerical quadrature. In a evenly sample λ system, these weights all equal to one. These intermediate states are usually referred as the *alchemical state* due to the fact that they served as the the path between two end states, thus they are considered to be unphysical. Figure 3.5 shows the variations of λ parameters in our two stage TI.

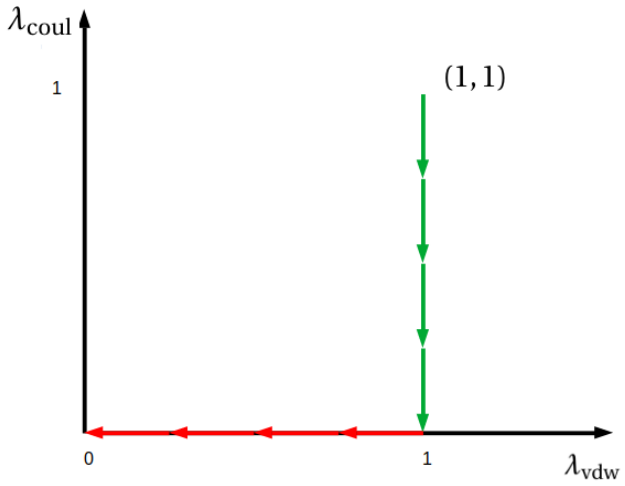


Figure 3.5: The two stages thermodynamic integration. First stage the Coulombic interaction is turned off, in the second stage the vdW interaction is turned off.

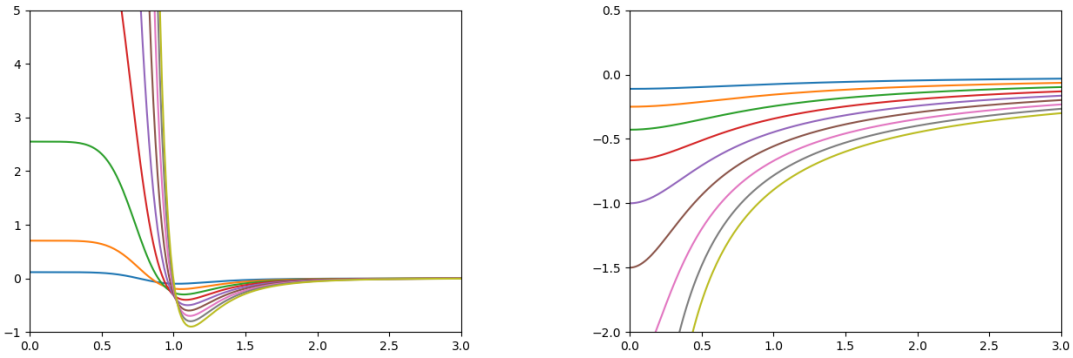
The soft potential

Contrast to the normal 6-12 potential, a soft potential is used to model the system with varying coupling parameter λ as shown in the Figure 3.4. The aim of this special treatment of the soft core is to avoid the singularities scenario where atoms are overlapped during the free energy calculation when the interaction sites are created or eliminated [10]. The modified L-J and Coulomb potential are respectively:

$$E = \lambda^n 4\epsilon \left\{ \frac{1}{\left[\alpha_{\mathrm{LJ}}(1-\lambda)^2 + \left(\frac{r}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha_{\mathrm{LJ}}(1-\lambda)^2 + \left(\frac{r}{\sigma}\right)^6} \right\} \quad r < r_c \tag{3.9}$$

$$E = \lambda^n \frac{C q_i q_j}{\epsilon \left[\alpha_{\mathrm{C}}(1-\lambda)^2 + r^2\right]^{1/2}} \quad r < r_c \tag{3.10}$$

One can see that when $\lambda = 1$, the potentials are degraded to the normal form. While $\lambda = 0$, the interactions are turned off. The different potential shapes with respect to the varying coupling parameter λ can be found in Figure 3.6. The λ in these figures varying from 0.9 to 0.1. The transition between two states are smoothed by the function form in Equation 3.9, and 3.10.



(a) L-J potential

(b) Coulomb potential

Figure 3.6: Plots of soft potential by ranging the interaction parameter λ.

Accuracy vs. efficiency dilemma

Once the alchemical states are determined, equilibrium Molecular Dynamics are carried out for the desired ensemble. Typically, the term $\partial U/\partial\lambda$, the potential change with respect to λ is calculated. The number of alchemical states not only affects the accuracy of the final free energy results, but also influences the resource requirements. With number of these alchemical states dozens times larger than number of end states, calculating the solvation free energy is always considered to be an expensive endeavour [36, 39]. As one needs to spend a whole bunch of additional resources for computing these intermediate states. This is a typical accuracy vs. efficiency dilemma that many researchers have faced in computer simulation. We will discuss the effort to find the most suitable configuration for our case in Chapter 5.

Finite size correction

It has been found that finite correction term for the single ion configuration is negligible[18]. Therefore, the finite size correction for solvation free energy is not considered in our implementation.

3.3. Self Diffusivity

Derivations

The diffusion of a tagged molecule among the same species is usually referred to as the self-diffusivity. From a macroscopic point of view, Fick's second law of diffusion relates the Brownian particle's density ρ at location r and time t:

$$\frac{\partial\rho(r,t)}{\partial t} - D\nabla^2\rho(r,t) = 0 \tag{3.11}$$

ρ is the particle's density, D represents the diffusion coefficient. This equation states that the material balance of a given system. Assume at time equal to zero the tagged molecules

are concentrated at the origin of the coordinate of interest, which can also be presented as the boundary condition: $\rho(r,0) = \delta(r)$ where $\delta(r)$ is the Dirac delta function with the property: $\int \delta(r)\mathrm{d}r = 1$. With this boundary condition, Equation 3.11 has a direct solution of form[2, 21]:

$$\rho(r,t) = \frac{1}{(4\pi D t)^{d/2}} \exp(-\frac{r^2}{4Dt}) \tag{3.12}$$

Where d represents the dimensionality. If we view this solution as a probability density function (PDF) and set the dimenionality as $d = 1$, then the mean and standard deviation of this PDF is $\mu = 0$, and $\sigma = \sqrt{2Dt}$. Hence essentially, the mean squared displacement of a set of particles is the variance of their distribution over time. Figure 3.7 gives these Gaussian profiles of the dimensionality $d = 1$ (single line) case. One can interpret particles' mean squared displacements as the variance (σ^2) of their Gaussian distribution, $\langle r^2(t)\rangle = 2 * 1 * Dt$.

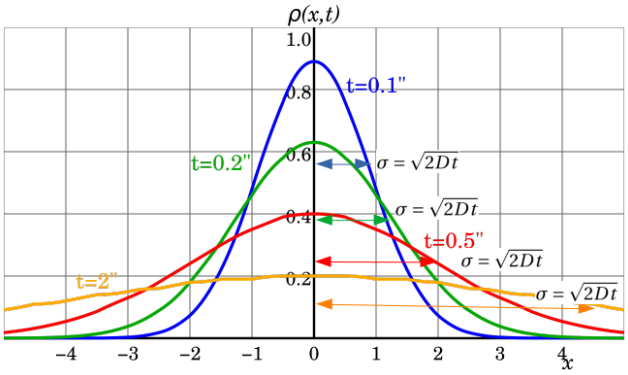


Figure 3.7: Diffusion of density profiles according to time. The concentration profile will diffuse though time, thus the initial sharp Gaussian will become flatten as time progress. The variance or second moment of the particles' distribution, which is 2dDt, representing the level of diffusion through time. It is also the mean square displacement for that time instance.

From the above analysis, we can find the mean square displacement equals the second moment of the Brownian particle's distribution [21]:

$$\langle r^2(t)\rangle \equiv \int \rho(r,t) r^2 \mathrm{d}r \tag{3.13}$$

Multiplying equation 3.11 by r^2 and integrating over all space [21] can give us the expression for time evolution of $\langle r^2(t)\rangle$ after some mathematical manipulation. The detailed proof can be found in the book of Frenkel and Smit[21]:

$$\frac{\partial \langle r^2(t)\rangle}{\partial t} = 2dD \tag{3.14}$$

Equation 3.14 is well known as the Einstein's relation. Although the diffusivity D is a macroscopic transport coefficient, $\langle r^2(t)\rangle$ has a microscopic interpretation: it is the means squared displacement of the tagged molecules at time t. This sheds light on calculating the diffusion coefficient in a computer simulation[21]. For every specie we measured, we can store their mean square displacement at time t, MSD(t), then we plot the figure of mean squared

displacement against time. We can interpret the slope of MSD-t plot to get the self diffusion coefficient of each species. This expressed as:

$$D_{\text{self}} = \lim_{t\to\infty} \frac{1}{6Nt} \left\langle \sum_{i=1}^{N} [\mathbf{r}_i(t) - \mathbf{r}_i(0)]^2 \right\rangle \tag{3.15}$$

where t is the simulation time, N is the number of molecules and angle brackets is the ensemble average. The accuracy of diffusivity would increase with simulation length increases as shown in Figure 3.8.

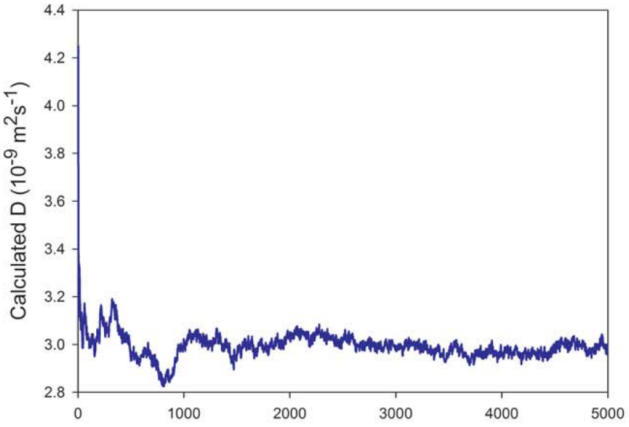


Figure 3.8: Diffusivity as time steps progress, picture is taken from [78]. The deviation of diffusivity reduces as the simulation time increase.

Ensemble average

As the Einstein relation is widely used in computer simulations, most diffusion coefficient calculations follow a MD-based evaluation analysis. The angle brackets for the mean squared displacement calculation have the interpretation as the ensemble average. From statistical scope, to have good results, one needs to use the mathematical expected value for squared displacement, i.e. $E(\langle r^2 \rangle)$, in the analysis. This expected value means one needs to average over all different molecules, time origins, and diverse independent cases as well to get the final expression of MSD. However, having multiple solute molecules in the system means we need to increase the number of solvent molecules or the concentration would be changed. Another way to go for improving the statistics is to have multiple independent cases to average the MSD data or to have long simulation length, both of which are resources demanding. Equation 3.14 requires a statistical limit of sampling as time approaches infinity. The conventional treatment is to select a specific segment to sample data and apply the Einstein Equation 3.14 [21].

Order n like algorithm

The conventional way to measure the transport coefficient with Einstein relation or velocity autocorrelation function (VACF) is inefficient as the measurement frequency is predefined[21, 31, 38]. Using a short time interval to sample data may capture all dynamics in trajectory, but the CPU and memory needs are high. While using a large time interval may

miss the fast dynamics correlations[21, 31]. To solve this dilemma, we can use an efficient sampling algorithm like order-n algorithm[21] or recently equivalent implementations like Window technique (WT) and multiple window technique (MWT) suggested by Dubbeldam et al[19].

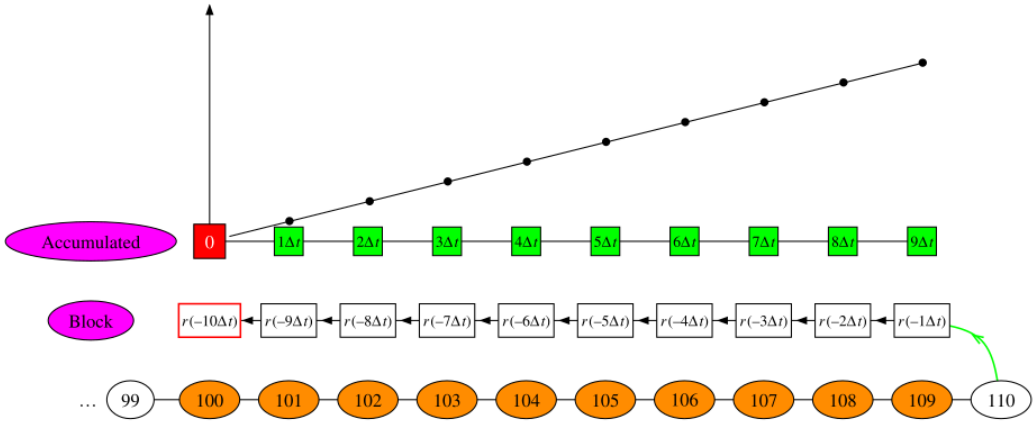


Figure 3.9: Window technique with MSD approach[19]. Red square element is the position value at time step 100 and the first element in the window $r(-10\Delta t)$, therefore it serves as the time origin. The green square data are correlated to the time origin to compute the MSD. The accumulated MSD divided by the number of window sampled is the current average MSD.

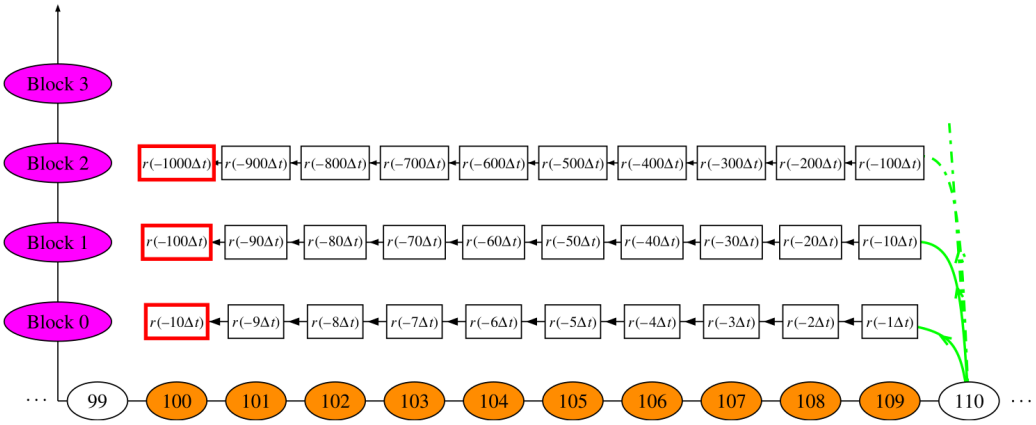


Figure 3.10: Multiple window technique with MSD approach[19]. Different blocks store data at different time scales. Block 0 samples data at every Δt, block 1 samples data at every $10\Delta t$, and block 2 sample data at every $100\Delta t$. Each block is used to compute the MSD for that time interval and diffusivity. Each time interval contribute to a part of the final MSD plot.

In the window technique approach, a block with a pre-defined size is chosen to store the position data. The first element in the block serves as the origin in time, and other positions are relative to the origin. At the update stage, this block will be shifted and sample different time origins. Figure 3.9 shows schematically the window technique. One problem with window technique is that the block size is fixed, thus unflexible. The fixed sampling frequency cannot capture the dynamics at different time scales [21, 31]. Dubbeldam et al.[19] proposed multiple window algorithm following the same idea behind conventional order-n algorithm.

Figure 3.10 shows schematically how MWT works. The key idea of MWT is still to have as many time origins as possible while keeping good use of the data available[19, 21]. The traditional order n algorithm coarse-grained velocity at previous time in different blocks to update the positions, $\Delta r = \Delta t \times \sum v$, while the implementation of multi-window technique from Dubbeldam et al. coarse-grained positions rather than the sum of velocities to yield the mean-square displacement.[19, 21]. Both of these two algorithms give a speed up in terms of memory requirement compared to the conventional sampling strategy of diffusion coefficient. The multiple window technique uses different sizes of blocks to sample the phase space, typically at a different order of time scales[19].

Finite size correction

The accuracy of calculated property under periodic boundary conditions (PBC) can have a vast dependence on the system size[2]. This dependence is due to the slowly decaying of Coulombic interactions. One direct way to tackle this problem is to increase the system size so that the screening effect of neighbor atoms is reduced[2]. However, the problem accompanying with larger system size is much higher computer power requirement since simulation time is proportional to the square of the number of particles: t $\sim N^2 \sim L^6$.

Yeh and Hummer proposed an analytic correction formula for diffusion coefficient based on the hydrodynamic self-interaction effect[80].

$$D_{i,\text{self}} = D_{i,\text{PBC}} + \frac{\xi k_B T}{6\pi\eta L} \tag{3.16}$$

Here, ξ is the Madelung constant in a Wigner lattice, k_B is the Boltzmann constant, T is the temperature in Kelvin. η is the viscosity of the solvent. Equation 3.16 has lightened the way to counter the finite size effect for calculating diffusivity. We can simply add a correction term $\frac{\xi k_B T}{6\pi\eta L}$ to diffusivity obtained from periodic boundary condition simulations. Figure 3.11 shows the dependence of the magnitude of this correction term to the system size for calculating the diffusivity of TIP3P water.

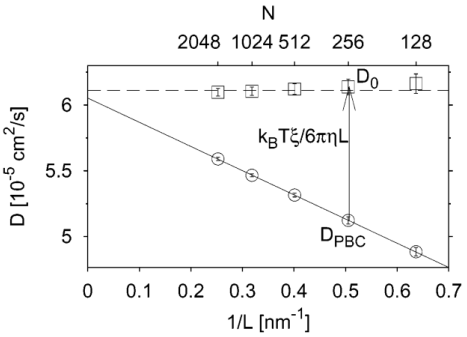


Figure 3.11: Finite size dependence of self-diffusivity[80]. The circles are the calculated diffusion coefficient from periodic boundary condition simulations, the squares are the results after applying finite size correction of Equation 3.16

3.4. Structural Properties

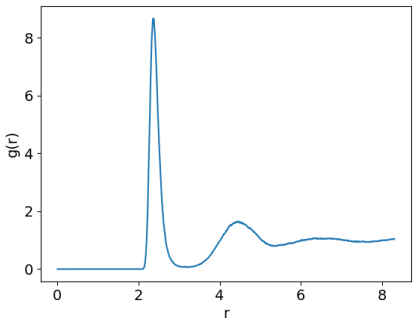


Figure 3.12: A typical radial distribution function.

A radial distribution function (Figure 3.12) describes the number density variation from the targeted center particle. From the RDF, one can derive structure related properties like ion oxygen distance (IOD), coordination numbers (CNs), and contact ion pairs (CIPs).

The ion oxygen distance (IOD) is defined as the distance between the center ion and the oxygen site of first hydration shell. This property measures the size of hydration shell and gives information about ion's local structure. It is also considered as an important parameterization criterion in many former studies[3, 36, 81]. The common way of obtaining the IOD is through finding the first peak location of the ion oxygen RDF.

The coordination numbers (CNs) is defined as the number of water molecules in the first hydration shell as:

$$\mathrm{CNs} = 4\pi\rho_{\mathrm{O}} \int_0^{r_{\min}} g_{\text{ion-O}_w}(r) r^2 \mathrm{d}r' \tag{3.17}$$

where ρ_{O} is the number density of water molecules, $g_{\text{ion-O}_w}$ is the ion-oxygen RDF, and $r_{\min}$ is the location of first minimum in $g_{\text{ion-O}_w}$. It should be pointed out the coordination number is closely correlated to the IOD for a same ionic solution system. This is simply due to the fact that they are extracted from the same RDF. Moreover, The experimental values of CNs also vary a lot from different sources because accurately determine the integration upper bond of Equation 3.17 is difficult[18, 36, 49, 79].

Contact ion pairs are defined as:

$$n^{\mathrm{CIP}} = 4\pi\rho_{\pm} \int_0^{r_{\min}} g_{\pm}(r) r^2 \mathrm{d}r' \tag{3.18}$$

where $\rho_{\pm}$ is the number density of cations or anions, $g_{\pm}$ is the ion-ion RDF, and $r_{\min}$ is the location of first minimum in $g_{\pm}$. Notice that when counting the CIP number, one can either use the cation or anion as the center particle. Many former studies prefer to use the cation as the center[7, 23, 79, 81], so we follow the same choice for consistency.

It has been reported in literature [5, 7, 36, 48, 79, 81] that there is invariably unreal ion precipitation phenomenon in simulations even at low concentrations. These aggregations of molecules are not observed in experiments; they are considered to be a nonphysical phenomena that happens in molecular simulations. The contact ion pairs (CIP) can indicate the level of precipitation in the ionic solution.

4

Bayesian Optimization

This chapter will discuss the Machine Learning technique used in this project. It starts with Section 4.1 introducing Bayesian Optimization. Then it covers the two major parts, namely, the Gaussian Process in Section 4.2 and acquisition function in Section 4.3. Section 4.3 also discusses the exploration vs. exploitation trade-offs.

4.1. Introduction

Bayesian Optimization is used in many applications, including robotics, deep learning architecture configuration search, reinforcement learning, drug trial, and particle physics [30, 50, 73]. This optimization routine was initially suggested by Kushner [41] and Močkus [55]. Then it was refined and made popular by Jones and coworkers [34].

This stochastic optimization routine is typically used to optimize functions with high costs (long time to evaluate). The strength of Bayesian Optimization is that it can perform optimization without any presumed knowledge about the target function. This type of target function is of subtle importance and they can be seen in our everyday pictures. For example, an advertising system always ask customers to rate. The preference of a customer can be viewed as the target function. In this process the information is exchanged only through querying customers some questions. And here the target function is essentially a "black-box", since the advertising system can only probe the viewers' preference. It is difficult to build a precise mathematical model to describe such a black-box function. More specifically, in practice these *black-box functions* usually require a high cost to evaluate, they do not have any closed-form expression, the derivative of the black-box function is not accessible, and their convexity properties are unknown[11, 20]. Bayesian optimization routines are well known for their efficiency of sampling the unknown black-box function, which stems from the fact that they use a probability model to approximate the search and target space[11].

It is called Bayesian because the fundamental idea is the same as the Bayes' theorem:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} \tag{4.1}$$

$$P(H \mid E) \propto P(E \mid H) \cdot P(H) \tag{4.2}$$

where H represents a hypothesis of a model and E represents the evidence found. The Bayes' theorem states the *posterior* of a hypothesis given evidence $P(H \mid E)$, is proportional to the *likelihood* $P(E \mid H)$, and the *prior* of the hypothesis $P(H)$. $P(E)$ is the *marginal likelihood*, since this term is same for different hypothesis considered, it does not contribute to the probability of different hypothesis.

Formation

A typical optimization problem in the engineering world requires a methodology to evaluate the *objective* function in a constrained domain, which states

$$\max_{\mathbf{x} \in \mathcal{A} \subset \mathbb{R}^d} f(\mathbf{x}) \tag{4.3}$$

where $\mathcal{A}$ is the optimization region , d is the dimensionality and $f(x)$ is the objective function, the following considerations apply to our force field optimization scheme:

- The inputs $\mathbf{x}$ (parameters for empirical force field formula) are bonded inside a constrained domain.
- $f(x)$ is expensive to evaluate via MD simulation.
- No closed or analytical form are known regard to $f(x)$ and no convexity information can be presumed. This function is known as a "black-box function".
- Every time we query $f(x)$ with a new point x_t, we get a corresponding function value.
- We search the global optimum of function $f(x)$ inside the bonded domain.

In Bayesian Optimization, the prior distribution of functions and likelihood will be used to construct a posterior distribution of functions that best describe the function algorithm knows. As the number of observations grows, the quality of the posterior distribution improves. The algorithm becomes more certain which region in the parameter space is worth exploring and which is not. Assume we sample black-box $f(x)$ in our optimization process and get results: $f(x_i), \quad i = 1, 2, ..., t$. The observations are accumulated in a set $\mathcal{D}_{1:t} = \{\mathbf{x}_{1:t}, f(\mathbf{x}_{1:t})\}$. Then the Bayes's theorem (Equation 4.2) can be re-written as:

$$P\left(f \mid \mathcal{D}_{1:t}\right) \propto P\left(\mathcal{D}_{1:t} \mid f\right) P(f) \tag{4.4}$$

where the posterior distribution is proportional to the likelihood and prior distribution. As the optimization goes on, more information of the objective function will be acquired and optimums can be approached by an external statistical model.

Core parts

Two parts construct the Bayesian Optimization: one statistical model called *gaussian process* for estimating the objective function, and the *acquisition function* (or *utility function*) to define which point to query in each sampling stage.

4.2. Gaussian Process

4.2.1. Fundamental Concepts

The *multivariate Gaussian distribution* or *multivariate normal distribution* is a variation of one dimensional Gaussian distribution by ascending the order to higher dimensions[11, 20, 59, 62]. It is described by a vector mean μ and *covariance matrix* Σ. A random variable X subjects to multivariate Gaussian is usually written as:

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{4.5}$$

where $\mathcal{N}(\mu, \Sigma)$ stands for a normal distribution with μ as the mean, and Σ as the covariance matrix. The covariance defines the shape of the distribution in terms of the expected value:

$$\Sigma = \mathrm{Cov}(x_i, x_j) = \mathbb{E}\left[(x_i - \mu_i)(x_j - \mu_j)^T\right] \tag{4.6}$$

To represent function values, the best treatment is to use a high dimensional vector. For example, if a n dimensional variable follows a n-dimensional multivariate Gaussian distribution, then the i-th subdimension variable can be used to represent a discrete function value. In other words, we can model a function $f(x)$ such that the discrete points of $f(x)$ are samples which follow the multivariate Gaussian distribution. The number of independent variables in our function equals to the dimension of the multivariate Gaussian distribution. Figure 4.1 gives a schematic view of a sample from a 100-dimensional Gaussian distribution. Each point in this figure is a sub-dimension of that 100-dimensional Gaussian variable. Though this way of representation, different function values can be easily approximated by a Gaussian sample and most importantly, they are cheap. The prediction of Gaussian process narrows down to draw samples from a multivariate Gaussian distribution.

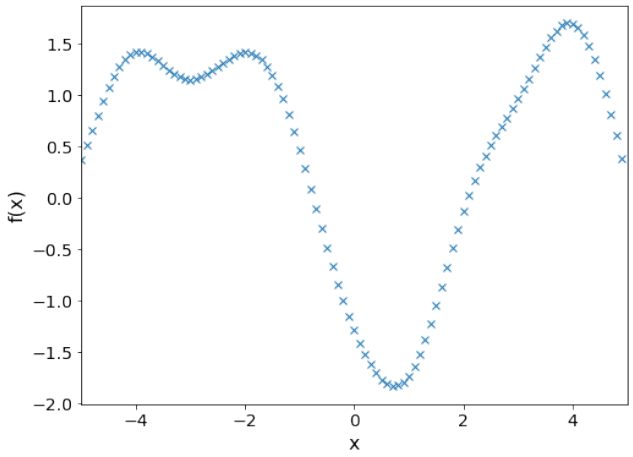


Figure 4.1: A 100 dimensional Gaussian sample for function representation.

Note that for two neighbor points $\mathbf{x}$, and $\mathbf{x}'$, they are considered to be very similar in the kernel function so the corresponding function value $f(\mathbf{x})$, $f(\mathbf{x}')$ generated by GP are nearly identical. Therefore it looks like as if it is a continuous function.

Covariance

A *covariance function* or *kernel function* measures how similar different variables are[11, 59, 61]. It pair-wise measures this similarity among all data points by returning a scalar. The covariance matrix is constructed by going through all pairwise combinations of test points as its matrix entries. A very common choice of covariance function is the squared exponential function[11, 59]

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) \tag{4.7}$$

The returns are close to 1 if $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar, while the returns are close to 0 if $\mathbf{x}_i$ and $\mathbf{x}_j$ are different. In the covariance matrix, the diagonal entries σ_{ii} represent the variance of the i-th variable, while the off diagonal entries σ_{ij} are the correlation between i-th and j-th variable. The covaiance matrix has the property of being symmetric and positive semi-definite[59].

Condition and marginalization

A crucial operation in Gaussian statistics is the *conditioning*. The Gaussian distribution has the algebraic property of being *closed* after conditioning[4, 59], meaning that a conditioned Gaussian is also Gaussian. Condition operation is usually denoted as "|", e.g., P(X | Y). The conditioning operation can be regarded as cutting through a higher dimension Gaussian distribution. The result of this cutting is another Gaussian distribution at a lower dimension. This property is of crucial importance since it allows the Bayesian inference in Gaussian process.

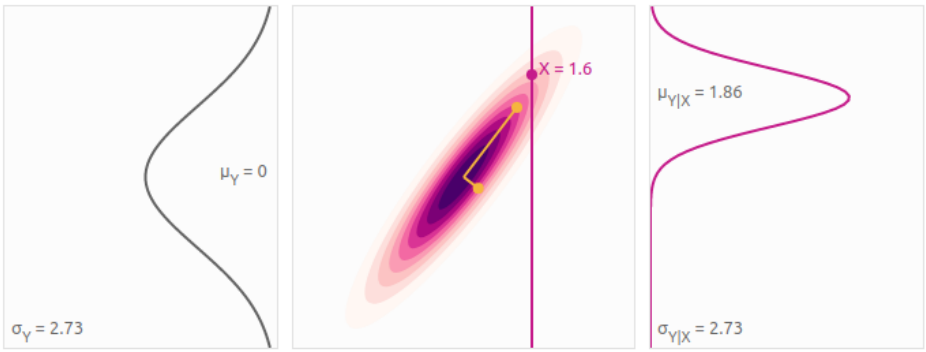


Figure 4.2: Marginalization (left) and Conditioning (right) of a 2D Gaussian distribution, picture is taken from [4].

Another important operation is *marginalization*, in which one can retrieve the partial information of a distribution. We can marginalize a joint Gaussian distribution by integrating all the possible outcomes of one variable. For example in Equation 4.8, we can marginalized the joint distribution by integrating through all possible situation of x to get marginalized distribution of y. Figure 4.2 illustrates the marginalized (left) and conditioned (right) distribution of a joint 2D Gaussian distribution.

$$p_Y(y) = \int_x p_{X,Y}(x,y)dx = \int_x p_{Y|X}(y \mid x)p_X(x)dx \tag{4.8}$$

4.2.2. Gaussian Processes

Gaussian Process is a stochastic process that extends the formation of multivariable Gaussian distribution to the infinite dimension[11, 59]. Any finite linear combination would still follow a Gaussian distribution. Like the Gaussian distribution is a distribution of variables, determined by its mean and covariance, the Gaussian process is a distribution over functions, specified by it's mean function and covariance function.

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right) \\ m(x) &= \mathbb{E}[f(\mathrm{x})] \\ k\left(\mathbf{x}, \mathbf{x}'\right) &= \mathbb{E}\left[\left(f(\mathbf{x}) - m(\mathbf{x})\right)\left(f\left(\mathbf{x}'\right) - m\left(\mathbf{x}'\right)\right)^T\right] \end{aligned} \tag{4.9}$$

Here the $m(\mathbf{x})$ is mean function for the given $\mathbf{x}$, $k\left(\mathbf{x}, \mathbf{x}'\right)$ is the kernel function which measures the similarity between two variables.

Bayesian inference

For simplicity and focus on the effect of kernel function, we assume the mean function of a GP is zero, $m(\mathbf{x}) = 0$, and covariance matrix is $\mathbf{K}$. If we sampled $\mathbf{x}_{1:t}$ and obtained their corresponding function values $f(\mathbf{x}_{1:t})$. These samplings will give us a set $\{\mathbf{x}_{1:t}, f(\mathbf{x}_{1:t})\}$ that satisfy a multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{K})$,

$$\mathbf{f}_{1:t} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{4.10}$$

where the $\mathbf{K}$ is the kernel matrix:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_t) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_t, \mathbf{x}_1) & \dots & k(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix} \tag{4.11}$$

To update our sampling with bayesian optimization, we first look at the previous sampled points and decide what point should be sampled in the next query. The mechanism that updating the next query stems from an outside model (acquisition function) that will be discussed in section 4.3. Here we only focus on getting the mean and covariance of the predicted distribution. Assume the next sample point is $\{x_{t+1}, f(x_{t+1})\}$. From the properties of GP we know that $\mathbf{f}_{1:t}$ and f_{t+1} are jointly Gaussian, with 0 mean and another covariance matrix.

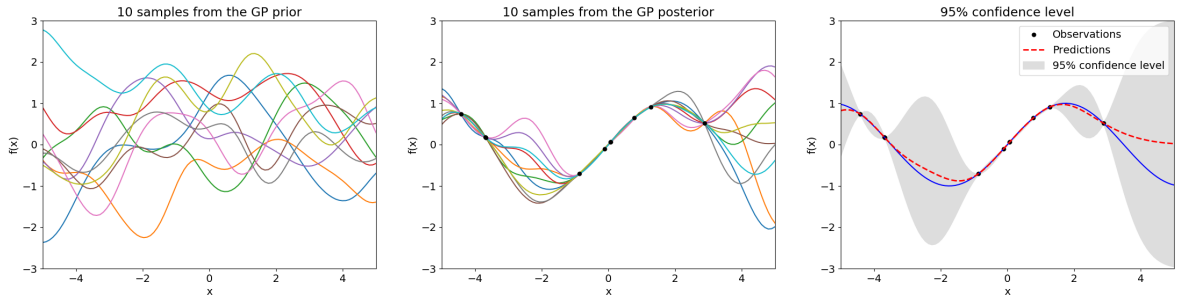
$$\begin{bmatrix} \mathbf{f}_{1:t} \\ f_{t+1} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{k} \\ \mathbf{k}^T & k(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}) \end{bmatrix}\right) \tag{4.12}$$

where the cross term is: $\mathbf{k}^T = [k(\mathbf{x}_{t+1}, \mathbf{x}_1), \quad k(\mathbf{x}_{t+1}, \mathbf{x}_2), \quad \dots, \quad , k(\mathbf{x}_{t+1}, \mathbf{x}_t)]$. Then the predicted distribution of function value at x_{t+1} can be obtained through conditioning the prior distribution

$$\begin{aligned} P\left(f_{t+1} \mid \mathcal{D}_{1:t}, \mathbf{x}_{t+1}\right) &\sim \mathcal{N}\left(\mu(\mathbf{x}_{t+1}), \sigma^2(\mathbf{x}_{t+1})\right) \\ \text{where} \qquad \mu(\mathbf{x}_{t+1}) &= \mathbf{k}^T \mathbf{K}^{-1} \mathbf{f}_{1:t} \\ \sigma^2(\mathbf{x}_{t+1}) &= k(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}) - \mathbf{k}^T \mathbf{K}^{-1} \mathbf{k} \end{aligned} \tag{4.13}$$

For the sake of brevity the derivation of Equation 4.13 is not shown, the details of derivation of this formula can be found in Williams or Murhpy's book[59, 63]. The key takeaway

from Equation 4.13 is that we can follow this relation to get the mean and variance of a conditioned distribution. This is the key of Bayesian inference.



(a) (b) (c)

Figure 4.3: Functions draw from prior distribution and the conditioned distribution of functions, sub-figure (c) gives the predicted mean and 95% confidence level.

This process can be better understood in Figure 4.3. Sub-figure 4.3a gives the 10 sample of functions generated by a Gaussian process with zero mean and exponential 2 norm kernel (Equation 4.7). These functions will revolve around the mean (which is zero here), and the kernel function will specifically determine the shape of the generated functions. To find the posterior distribution, one can think that the functions that do not pass the sampled data points are eliminated. There will be another distribution after this elimination (or conditioning), which can give some sampled functions in Sub-figure 4.3b. Hence, the new sampled data will constrain the prior distribution and the posterior distribution is updated at each step of Bayesian optimization. Equation 4.13 allows us to get the mean and covariance of one test point. If we extent other test points instead of just one, we get the figure 4.3c. The predicted distributions of unknowns are modeled by the shaded area and the solid line gives a surrogate prediction of the mean for the objective function.

All the assumptions so far have not considered the case with noise. In a real world scenario, the measurement might not always be accurate so the training data can contain noise. One can take the noise into account by adding a noise term to the function values:

$$y = f(x) + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma_y^2) \tag{4.14}$$

This can be substituted into Equation 4.13, which gives a modified mean and variance of predicted distribution

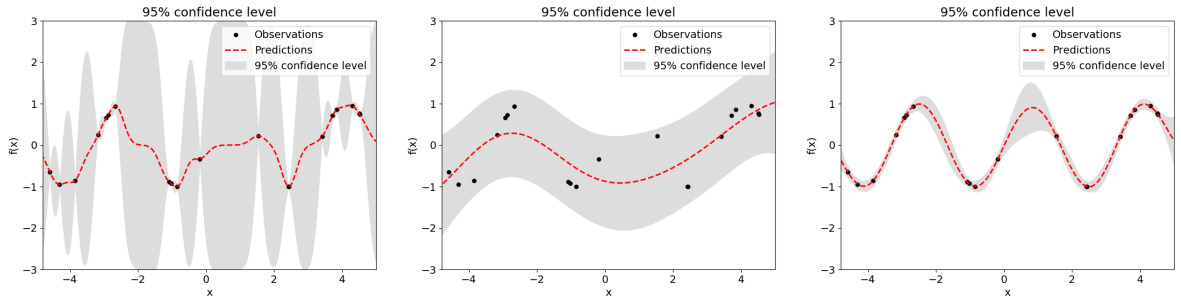
$$\begin{aligned} \mu(\mathbf{x}_{t+1}) &= \mathbf{k}^T (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{y}_{1:t} \\ \sigma^2(\mathbf{x}_{t+1}) &= k(\mathbf{x}_{t+1}, \mathbf{x}_{t+1}) - \mathbf{k}^T (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} \mathbf{k} \end{aligned} \tag{4.15}$$

4.2.3. Learning the kernel parameters

With the kernel being selected, there are still hyper-parameters to be determined for a good fit of data. Hyperparameters in kernel function are introduced to give control over the it's behaviour for measuring similarity. For example, the Squared exponential kernel has the following form of hyperparameters:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2l^2} \right) + \sigma_n^2 \mathbf{I} \tag{4.16}$$

where $l, \sigma_f^2, \sigma_n^2$ are three hyperparameters. Figure 4.4 gives the different behaviour of predicted distribution with this kernel under different hyperparameters. Clearly, parameters in figure 4.4c has the largest likelihood here.



(a) $(l, \sigma_f^2, \sigma_n^2) = (0.223, 1, 0.01)$ (b) $(l, \sigma_f^2, \sigma_n^2) = (2.236, 1, 0.316)$ (c) $(l, \sigma_f^2, \sigma_n^2) = (1, 1, 0.031)$

Figure 4.4: The effect of varying kernel's hyper-parameters.

These hyperparameters are learned by maximizing the log marginal likelihood of function observation given the kernel parameters input. The term marginal likelihood here refers to marginalization (integration) over all noise. Since the function values follows a normal distribution: $\mathbf{y} \sim \mathcal{N}(0, \mathbf{K} + \sigma_n^2 \mathbf{I})$ and it has an explicit expression of probability density function in the form like Equation C.1. Hence the expression of log marginal likelihood can be obtained as:

$$\log p(\mathbf{y} \mid \mathbf{X}) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log\left(\left|\mathbf{K} + \sigma_n^2 \mathbf{I}\right|\right) - \frac{N}{2}\log(2\pi) \tag{4.17}$$

note here $p(\mathbf{y} \mid \mathbf{X})$ represents the marginal likelihood with $\mathbf{y}$ as observations, $\mathbf{X}$ as the model inputs containing the kernel parameters. $\mathbf{K}$ is the noise free kernel matrix. In Equation 4.17, the first term is the data fit term, the second term is the complexity penalty, the last term is a constant depending on the number of samples [59, 63]. The penalty term works with the fitting term to avoid *over-fitting*. It makes sure the model not being too complex, so it can fit the data "just right".

4.3. Acquisition Function

The GP serves as the role of modelling the objective function. With GP we can update the prior distribution as the new observations are discovered. Another part of Bayesian Optimization is called *acquisition function*, which serves the role of guiding the algorithm to search the optimums. The reason it is called acquisition function is because it evaluates the acquisitions for the potential improvement at given points in the objective function. A high mean (expected value) or high variance (uncertainty) can both give rise to high acquisitions. To some extents, the original goal of optimizing objective function can be achieved by a surrogate model which optimizes the much cheaper acquisition function at each optimization step. This section will introduce 2 acquisition functions involved in this project: *expected improvement* and the *upper bond confidence*.

4.3.1. Expected improvement

The expected improvement (EI) acquisition function is an improved version of the probability improvement (PI, Appendix C.2) . The green shaded area in Figure 4.5 represents the probability of getting a better result than the current optimum for one test point. Its magnitude will be close to 1 near the optimums. In Figure 4.5, at the near right region of x^+ the CDF will yield higher return, compared with the far right region where the uncertainty is high. This is likely to cause the algorithm exhaustively search around the local optimum[11, 33].

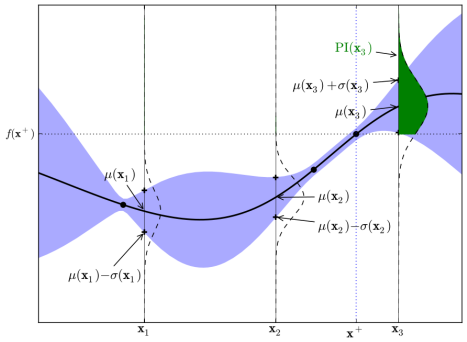


Figure 4.5: Schematic figure of the probability of improvement, figure is taken from [11]. The black dot points are the observations with x^+ being the maximum. Integrate the predicted distribution from the maximum of observations gives the probability of getting higher function value for the test point.

Apart from only accounting for the probability of improvement, the EI method also accounts for the magnitude of improvement [11, 57]. The utility function proposed by Mockus is [56]:

$$u(x) = \max\{0, f(x) - f(x^+)\} \tag{4.18}$$

where the $u(x)$ is the difference between the probed prediction and the best function value that has been found. If the probed prediction is higher than the best function value from current observation, $u(x)$ returns positive values, otherwise 0. Then the acquisition function is the expected utility:

$$\begin{aligned} a(x) &= \mathbb{E}[u(x) \mid \mathcal{D}_{1:t}] = \int_{f(x^+)}^{\infty} (f(x) - f(x^+))\mathcal{N}(\mu(x), \sigma(x))\, df(x) \\ &= \sigma(x)\left[\frac{\mu(x) - f\left(x^+\right) - \xi}{\sigma(x)}\Phi\left(\frac{\mu(x) - f\left(x^+\right) - \xi}{\sigma(x)}\right) + \phi\left(\frac{\mu(x) - f\left(x^+\right) - \xi}{\sigma(x)}\right)\right] \end{aligned} \tag{4.19}$$

where Φ and ϕ stand for the cumulative distribution function and probability density function. Notice here the ξ parameter is artificially added by the users to balance the exploration vs. exploitation trade off, which will be discussed in a later section. Then one can select the next query point by finding $x_{t+1} = \underset{x}{\text{argmax}} a(x)$.

4.3.2. Upper bound confidence

We can also direct balance the exploration vs. exploitation trade off using an so called upper confidence bound[11, 71, 73] criteria. The acquisition function is formulated as:

$$a(x) = \mu(x) + \kappa\sigma(x) \tag{4.20}$$

where $\kappa > 0$ is the balance parameter. Equation 4.20 explicitly balances the exploration and exploitation trade-offs through κ. The formation κ is to minimize the cumulative regret in the optimization process. In the study of Srinivas[71], one can find the details of this trade off parameter. The idea of upper bond confidence bound is that the $\sigma(x)$ term is a measure of uncertainty of the test points. The acquisition function being maximized is thus sort of the upper bound on the possible true value of the test point[73].

4.3.3. Exploration vs. exploitation trade-off

In a GP based Optimization, any test point has two quantities derived from the statistical model: mean $\mu(x)$ and uncertainty $\sigma(x)$. In principle, if the acquisition function regards the mean as more important, then it will more likely to probe the region where the predicted mean is high. This way of optimization is usually referred to as exploitation since the algorithm tries hard to exploit the known information at current circumstances to find better optimums. On the other hand, if the acquisition function regards the uncertainty as more important, then it will more likely to probe the region where the predicted uncertainty is high. Therefore the whole landscape of the domain is more likely to be inspected thoroughly, and possibly a better solution can emerge from the black-box function.

Figure 4.6 gives a good example of well-balanced optimization strategy. One can see after 10 iterations, the algorithm successfully found the locations of two local maximum, and it spent most of resources near the optimum region. Maximizing acquisition function is used to select which point to sample in each optimization step. The sub-figure in the second grid of Figure 4.6 gives the acquisition at the current optimization step, from which it suggests the next most promising query point locates at $x = 8$.

It is of great interest to balance the exploration and exploitation trade-off when optimizing the objective function. Find a good balance between this trade-off is the key to many leaning based optimization routines[11, 73]. This balance is essentially achieved by adding a controllable parameter in the final acquisition function, for example, ξ in Equation 4.19 and κ in Equation 4.20. These controllable variables change the formation of acquisition function so the search strategies are also changed.

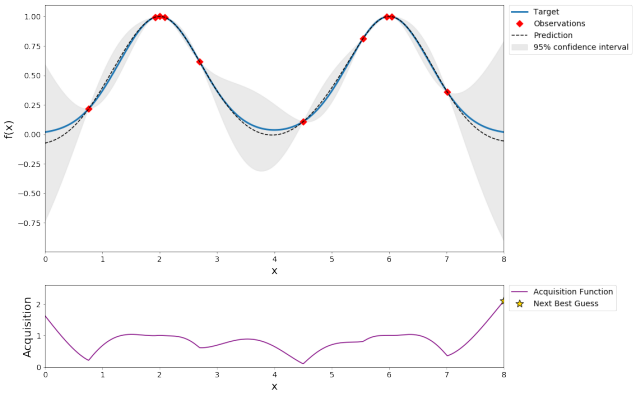


Figure 4.6: Exploration vs. exploitation dilemma in optimization problem[62].

5

Methodology

This chapter discusses the proposed approaches. Section 5.1 and 5.2 cover the methodology for carrying out the parameterization. Sections 5.3-5.6 introduce the detailed setup for calculating different ionic properties, the search of the most appropriate configuration, and the validation of a single simulation. Formulating an accurate and efficient MD simulation scheme first so we can use them in the final optimization stage.

5.1. Optimization Formation

In order to generalize the cost function so it can be used for different properties and ion species, the absolute relative error is employed to evaluate the cost between simulation and experimental results, see Equation 5.1. In the optimization stage, optimization algorithm will minimize this cost function so that the error between the experimental and simulation results will be brought down to the minimum. For each ionic property and combination of ion species, the cost function to be optimized is given by:

$$\mathcal{F}_i(\epsilon,\sigma) = \left| 1 - \frac{\Omega_{\text{isim}}(\epsilon,\sigma)}{\Omega_{\text{iexp}}} \right| \tag{5.1}$$

Here, Ω_{sim} is the simulation result, while Ω_{exp} is the experimental result, subscript i represents different properties. For simulation results that match the experimental results perfectly, Equation 5.1 will be zero.

Gradient free optimization

The most straightforward optimization procedure would be going for a derivative-based routine, like gradient descent method[25, 61]. The problem is to find force field parameters satisfy: $\nabla\mathcal{F} = \left(\frac{\partial\mathcal{F}}{\partial\epsilon}, \frac{\partial\mathcal{F}}{\partial\sigma}\right) = 0$. However, the derivatives $\partial F/\partial\epsilon$, and $\partial F/\partial\sigma$ of our formalized cost function are inaccessible. The function $\mathcal{F}$ is only known each time we query it with a set of force field parameters (ϵ,σ). Therefore, we use gradient-free optimization to deal with this black-box function. In our problem, ionic properties are costly to evaluate. We can query the cost function with input parameter sets and get its corresponding costs every time we need, but we cannot find its derivatives, and in particular, we cannot presume any

analytical form of $\mathcal{F}$. The ultimate objective for this force field parameters design problem is to find the optimum sets $(\epsilon,\sigma)_{\text{opt}} = \text{argmin}(\mathcal{F})$ with less computational resources spent than the conventional trial-error scheme [7, 36, 79, 81].

Compare search methods

In black-box optimization, grid search method is the least favorable. This is because grid search method repeatedly sample the same value for a given dimension, so it can miss potential better results easily. In contrast, the random search can probe the black-box function more thoroughly, so better results could be found. Suppose we have a black-box function with some importance distributions along the vertical and horizontal axes as it is shown in Figure 5.1. Nine trials using grid search only give 3 distinct samples over the horizontal dimension. In contrast, nine trials using random search give 9 distinct samples over the horizontal dimension, and one almost gets the highest reward. The inefficient sampling of grid search is even severe for force field optimization problems with higher dimensions. Guided random search like Bayesian Optimization can have greater potentials to track optimums.

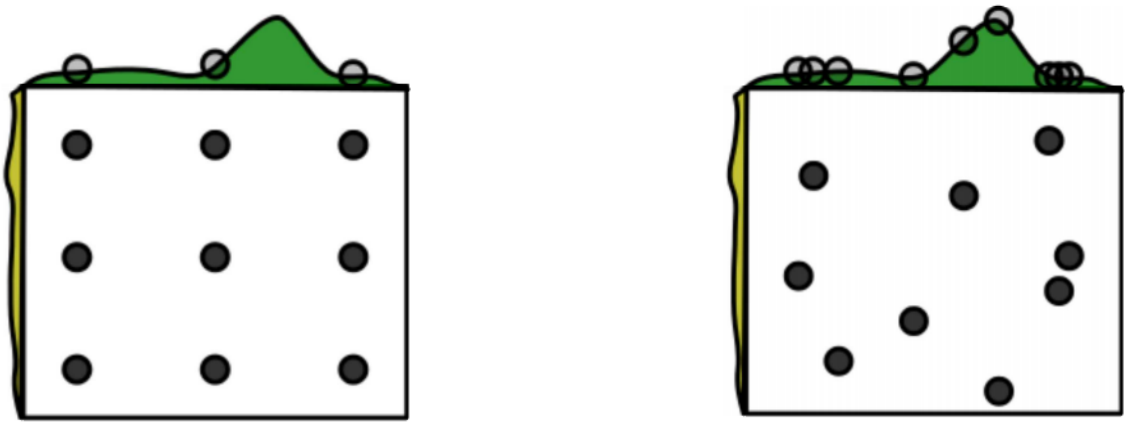


Figure 5.1: Comparison of grid search and random search method. The green and yellow shaded area represent an approximated importance at horizontal and vertical sub dimension. Picture is taken from Ref.[9]

5.2. Method

In this section we will introduce the procedures for carrying out the optimization.

Mapping the isolines with Bayesian

An gradient free optimization algorithm can be employed to minimize Equation 5.1[60, 62]. The calculated property values are compared with results found in literature[18, 36, 49, 79, 81]. It has been examined that our configuration can yield results with a low level of noise (less than 1%) for solvation free energy and ion oxygen distance.

In contrast to the conventional grid search, we have used well-designed optimization algorithms with heavy engineered Gaussian Processes to perform the search [11, 62, 73].

The Bayesian Optimization is implemented in the framework of Nogueira et al.[62], which utilizes scikit-learn as the backend for GP. We first train a model with normalized data to balance the exploration and exploitation trade-off. The determined balance parameter is applicable to all optimization cases because we normalized the data into same order of magnitude ($\sim \mathcal{O}(1)$). Then this model is implemented with the LAMMPS simulation package, which will perform the Molecular Dynamics simulations. The Bayesian optimization will guide the package to sample different parameter sets during the computation. The well-trained model can significantly reduce the optimization steps needed to reach an optimum. After observing enough data, we can use the prediction from our model to formalize the correlation of L-J parameters. Properties need human interaction are considered as secondary targets. They are optimized with brute force using optimum parameters of primary targets. The flow chart of this process can be found in Figure 5.2.

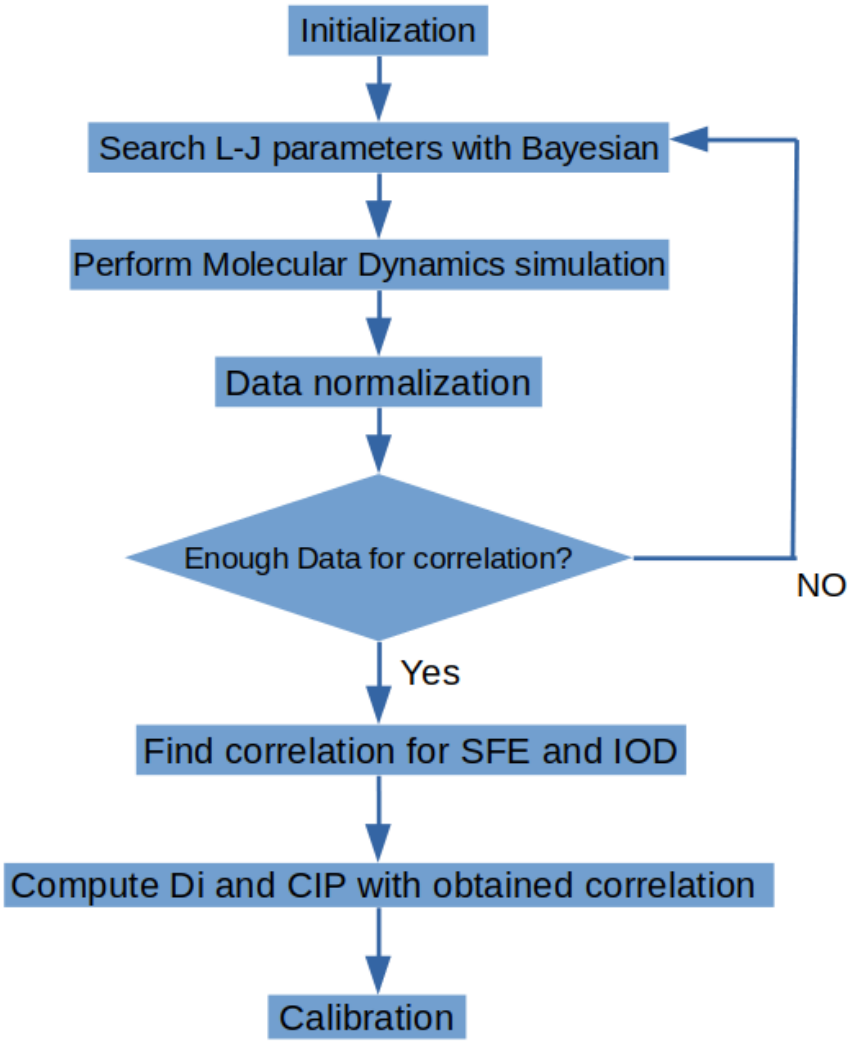


Figure 5.2: Flowchart of the parameterization process

Find parameterization trends of diffusivity and contact ion pairs

Ionic self diffusivity and contact ion pairs are not suitable to be included in the Bayesian optimization framework. So we consider them as the secondary properties. In other words, we optimize these properties with brute force based on optimum parameters of primary properties we have found.

Diffusivity calculation contains high uncertainties, which will greatly deteriorate performance of the optimization algorithm. Different starting conditions could lead to different isoline locations. Using a long simulation length can significantly reduce the level of uncertainty, but the computer power requirement could be too high to be affordable. This high certainty has been reported in many previous studies[7, 58, 81], regardless they have employed considerable long simulation length (up to 100 ns) and large system size (around 4000 molecules).

We cannot formulate an automatic way to read CIP result from machine. The CIP calculation requires human interaction to determine the minimum location, since height of the first peak in ion-ion RDF is varying for different parameter sets. The strategy to deal with self diffusivity and CIP is to study their dependence of force field parameters.

5.3. Simulation Details

Molecular Dynamics simulations are carried out in the open-source Large-scale Atomic/-Molecular Massively Parallel Simulator[67]. A cubic box of length 25.5 Å with periodic boundary conditions was used in simulations for solvation free energy and RDF related properties. The cut-off is set to be 8 Å for both Lennard-Jones and Coulombic interactions. Except for simulating anions with large particle diameters, where the cut off is set to be 12Å. The neighbor list distance is set as 2Å. The particle-particle particle-mesh (PPPM) method with a relative precision of 1e-3 is used to treat long-range electrostatic interactions. This box length will result in a system containing 555 water molecules in total. This choice of system size is useful because 10 ions in the system corresponds to 1M concentration of that species. Moreover, computer power requirements are acceptable with this choice. The molecules are first put into a lattice then ions are inserted into the system. The Polak-Ribiere version of the conjugate gradient algorithm is used to find the energy minimum for provided coordinates. There are two stages in the bulk simulation, where the equilibrium and structural properties are determined. The isothermal–isobaric (NPT) ensemble is used for equilibrating the system, the pressure for equilibrating the system is set to 1 atm and the temperature is fixed at 298 K. After the density is stable, the canonical (NVT) ensemble is used to perform the production run. As for the thermodynamics integration cycle, the isothermal-isobaric (NPT) ensemble is used alone to mimic the experimental solvation condition. The timestep was fixed as 2 fs. The Nose-Hoover thermostat and barostat have been used, with relaxation parameters set to be 100 dt for NVT simulation and 1000 dt for NPT simulations. The Verlet algorithm is used as the time integrator for solving Newton's equation of motion[67]. For all systems, the SHAKE algorithm has been used to constrain the geometry of water molecules[67]. In all simulations, the Lorentz-Berthelot mixing rules are used for pair potential.

There are also some specific configuration setups for each concerned property to speed up parameterization process. They will be discussed in detail in the subsequent sections,

here we first include them for consistency. For solvation free energy, The thermodynamic integration contains 8 stages, i.e, $\lambda = 1.000, 0.666, 0.333, 0.000$ both for Coulombic potential disappearing stage and vdW potential disappearing stage. Each sub-state is first equilibrated 100 ps, followed by another 100 ps production. The perturbation parameter for the finite different TI cycle is settled as 0.002. For diffusivity calculation with OCTP[31], the number of block elements is chosen as 15. Two independent cases are used to get better statistics of diffusivity. For structural properties, the number of bins is selected as 1000 for RDF representation. The final RDF data is sampled from the last 500ps with a frequency of every 2fs. The system is equilibrated for 200 ps for diffusivity and RDF related calculations; this choice has guarantees the density will reach a stable level. The production length is 10 ns for diffusivity and 1 ns for structural properties.

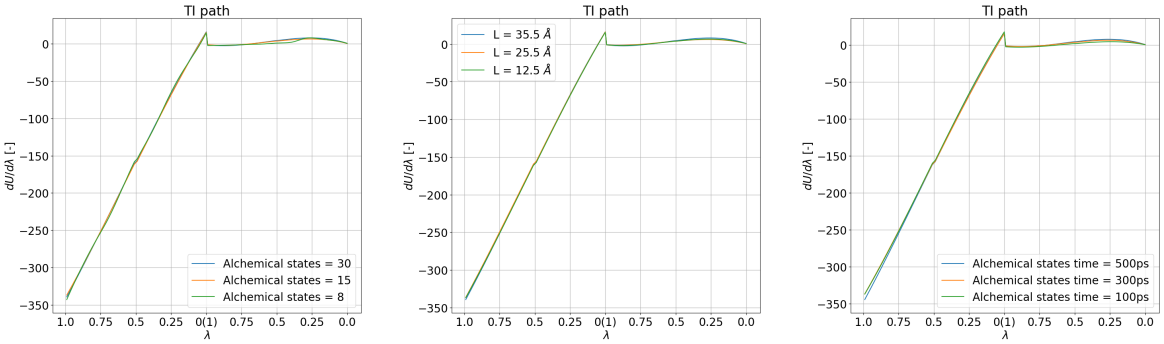
The effect of scaled charge was also put into investigation. As it is was adopted for TIP4P/2005 water model by Vega's group[1, 7, 81], the charge scaling ratio for monvalent ions is 0.85, and for divalent ions is 1.7[43, 44].

5.4. Solvation Free Energy

The main resource requirement of solvation free energy calculation is due to the need for alchemical states for smooth transitions. In thermodynamic integration, the free energy is computed by a weighted sum of ensemble averages of the free energy change with respect to the coupling parameter λ. The analysis of the final solvation energy results requires one to interpolate the discrete sampled intermediate alchemical states along the thermodynamic path. The performance analysis depends on the nature of underlying data structures, i.e., the shape of $\partial U/\partial\lambda$ curve. Therefore, the search of the cheapest configuration while maintaining the accuracy of results should be carried out.

We first perform a simulation with a relatively expensive configuration. A Na^+ taken from TIP4P-Ew&JC force field is used to reproduce the solvation free energy result. The system size is fixed as 25.5Å, and the production time for each alchemical state is 500ps. Totally 30 intermediate alchemical states are selected to pave the path for thermodynamics integration. The corresponding result we get is 88.1 kcal/mol, which is considered to be quite close to the experimental value of 87.2 kcal/mol from Marcus[49]. Then we decrease the number of intermediate alchemical states, box size, and simulation time for each substates window to narrow down search of the cheapest configuration setup.

The thermodynamic integration paths for different configurations are shown in Figure 5.3. One can find the results of solvation free energy by integrating this curve with respect to λ. Clearly, we can exploit the system size, sub-state simulation length, and the number of intermediate states to get a more efficient configuration. The underlying data structure of the integration path is very close to linear relationship for the Coulomb potential disappearing stage in Figure 5.3. This is why we can use a more sparse sampling of alchemical states while keeping the accuracy. It is also worth noting that the free energy contribution from the van der Waals disappearing stage is rather small. In fact, this part is nearly negligible in the whole picture.



(a) TI path as a function of Alchemical states (b) TI path as a function of system size (c) TI path as a function of simulation time

Figure 5.3: Effects of different configuration setups to the two stages TI. From left to right along x axis, in the first stage the coulomb potential was turned off, in the second stage the Van der Waals potential was turned off. The shape of TI path with different numbers of intermediate alchemical states, system size, and the simulation time of sub-states are compared with each other. The cut-off is also changed accordingly to be less than half of the box size.

Table 5.1 gives a numerical comparison of results from this wok, MD results from Döpke et al. Ref.[18], and the experiment values from Marcus Ref.[49]. Force field combination are TIP4P-Ew&JC and TIP4P-2005&Mamatkulov for validation.

Table 5.1: Solvation free energy obtained of different ions for validation. Force field combination are TIP4P-Ew&JC and TIP4P-2005&Mamatkulov for MD simulation of ours and Ref[18]. Units are Kcal/mol.

	Li^+	Na^+	K^+	Rb^+	Cs^+	F^-	Cl^-	Br^-	Mg^{2+}	Ca^{2+}	Ba^{2+}
This work	113.7	89.2	71.3	66.5	62.0	120.2	88.4	80.3	409.3	342.7	285.3
Ref.[18]	114.2	89.0	71.1	66.1	60.5	119.4	89.8	83.1	415.0	344.0	289.4
Ref.[49]	113.5	87.2	70.5	65.7	59.8	111.1	81.3	75.3	439.3	362.1	300.7

5.5. Self Diffusion Coefficient

The simulation length for diffusivity is to be determined. Different initial configurations of molecules are formed to realize the independent simulation cases. If the simulation time is too short, the obtained results will be very noisy since ions have no preferential direction in this system. In contrast, if the simulation time is long, the high requirement of resources will become unaffordable.

It is well-known that the beginning stage of MSD-time figure is the ballistic region and it will overpredict the diffusivity value. Thus it should be excluded from the calculation. On the other hand, data from later simulation segment only contain a few sampled cases, these data are not good for determining the MSD-t slope in order-n algorithm framework. Therefore, a log-log plot can be used to determine which region the MSD follows a linear relation with respect to time. A typical log-log plot can be found in Figure 5.4. From this plot, we determine the region that has slope 1 to sample data. From the investigation of different simulation lengths, it has been found that the third block of our OCTP setup shows good linear trends, corresponding to 5 ps - 68 ps section in Figure 5.4. After verifying this, we can fix the choice of the block in simulations and let the machine automatically read data.

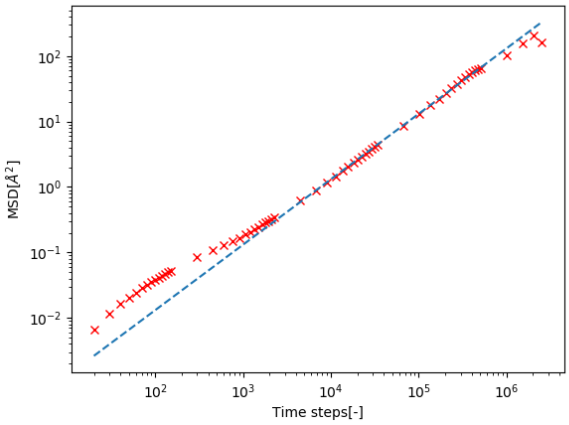


Figure 5.4: OCTP LAMMPS plugin of MWT from Jamali et al.[31], notice here we have several blocks together formulating this whole log-log plot. The linear region is chosen to determine the slope for diffusion coefficient calculation. No finite size correction has been added yet.

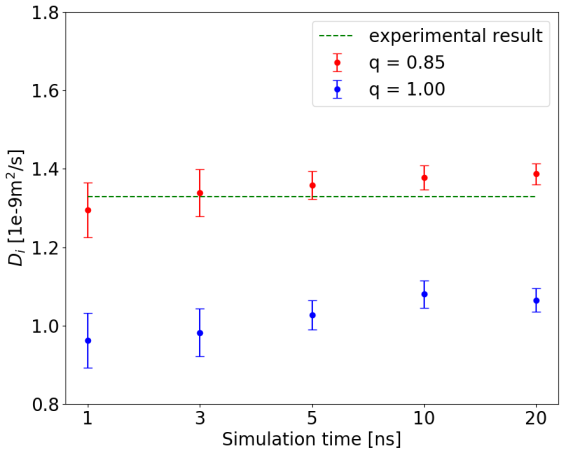


Figure 5.5: Different diffusivity of TIP4P2005/Zeron Na^+ as a function of the simulation time duration. The green dash line represents the experimental results from Marcus[49].

To quantify the level of uncertainty, we compute the deviation of diffusivity for different simulation lengths by performing 10 independent runs. Then the deviation level is calculated empirically from these 10 data sets. From a 1ns simulation with a single Na^+ and TIP4P/2005&Madrid force field combination, a deviation of 0.06956 [1e-9 m^2/s] is obtained. The simulation time is then increased to 3ns, 5ns, 10ns, and 20ns. The corresponding deviations are found as 0.06037, 0.03535, 0.03038, and 0.02687 [1e-9 m^2/s]. Finally, the simulation length is chosen as 10 ns for calculating diffusivity. It should be noted that this way of determining the deviation might not be accurate due to the limitation of sample cases. Figure 5.5 shows the diffusion coefficient under different configuration setups with force field combination of TIP4P/2005&Madrid for Na^+. We can see that getting an accurate result requires a long simulation length. This figure also shows the diffusivity dependence on the ionic charge, which will be discussed further in the results chapter.

The performance of two diffusivity calculation algorithms (MW and MWT) are compared with each other for efficiency improvement. Window technique is implemented in the post-processing stage using the MD analysis package[53], while the Multi-window Technique implementation is from the LAMMPS plugin of Jamali et al. [31]. We use a force field combination of TIP4P/2005&Madrid for validating Na^+ diffusivity. It has been found that MWT and WT both give results that match the experimental values. These two methods give similar diffusivity results. But WT requires a pre-defined window length, and the obtained slope contains noise, which could harm the quality of regression. Therefore, the MWT is employed. To further validate our configuration across all different chemical species. Table 5.2 gives a numerical comparison of results from this wok, MD simulation from Döpke et al. Ref.[18] and the experimental values from Marcus Ref.[49]. It shows clearly the ionic self diffusivity of MD simulations deviates from source to source.

Table 5.2: Self diffusivity obtained of different ions for validation. Force field combination are TIP4P-Ew&JC and TIP4P-2005&Mamatkulov for MD simulation of ours and Ref[18]. Units are 1e-9 m^2/s.

	Li^+	Na^+	K^+	Rb^+	Cs^+	F^-	Cl^-	Br^-	Mg^{2+}	Ca^{2+}	Ba^{2+}
This work	1.28	1.08	1.85	2.00	1.89	1.17	1.73	1.77	0.84	0.95	0.80
Ref.[18]	1.12	1.04	1.41	1.58	1.85	1.04	1.40	1.56	0.85	0.86	0.75
Ref.[49]	1.03	1.33	1.96	2.07	2.06	1.48	2.03	2.08	0.71	0.79	0.85

5.6. Ion Oxygen Distance

The structural properties of ion oxygen distance, coordination numbers and contact ion pairs are obtained in equilibrium MD through the either ion-water RDF or ion-ion RDF. The configuration setup is described as before in section 5.3. Table 5.3 gives IOD results calculated by our configuration setup and reference values from Döpke et al. [18] and Marcus [49].

Table 5.3: Obtained ion-oxygen distance for validation. Force field combination are TIP4P-Ew&JC and TIP4P-2005&Mamatkulov for MD simulation of ours and Ref. [18]. Units are Å.

	Li^+	Na^+	K^+	Rb^+	Cs^+	F^-	Cl^-	Br^-	Mg^{2+}	Ca^{2+}	Ba^{2+}
This work	1.92	2.35	2.72	2.88	3.03	2.69	3.16	3.31	2.00	2.36	2.71
Ref.[18]	1.91	2.34	2.70	2.86	3.02	2.69	3.15	3.30	1.99	2.36	2.70
Ref.[49]	2.08	2.36	2.80	2.89	3.14	2.63	3.19	3.37	2.09	2.41	2.75

6

Results and Discussions

This chapter contains optimization search results of different ions with TIP4P/2005 water model. We implement the Bayesian optimization algorithm to efficiently sample the (ϵ, σ) parameter space and identify the optimum region. Section 6.1 presents the search map, optimization setup and the corresponding optimum regions for solvation free energy (SFE) and ion-oxygen distance (IOD). Next, in section 6.2, we evaluate the dependence of diffusivity upon single force field parameter. This treatment allows us to have a peek over the hyper-surface of diffusivity with limited computing power. After that, the optimization results of structural properties are discussed in section 6.4. RDFs of ion-water and cation-anion are shown. Then we conclude the dependence of ionic properties on force field parameters in section 6.5, which can provide a good guide for the parameterization. Finally, we select the parameter set base on our model and evaluate the performance in section 6.6.

6.1. The Correlation Maps

This section starts with introducing the procedures of carrying out the parameter search in sub-section 6.1.1. Then the results are presented for monovalent cations, monovalent anions and divalent cations in the sub-section 6.1.2, 6.1.3, and 6.1.4 respectively. Section 6.1.5 summaries the findings of the correlation maps. For each ion group, we present the optimum regions for SFE and IOD, each under two different charge conditions, i.e. q = 0.85, and q = 1.00. With different charge conditions, we can understand the effect of scaling charge for ionic force field parameterization. We also apply a SFE correction treatment suggested by Döpke et al[18]. This correction multiplies the obtained SFE results by ϵ_{el} (high frequency dielectric constant). The derivation of electronic charge correction for MD simulation can be found in Appendix D.1. For each obtained map, we use the isolines to identify the location of a selected parameter set from literature. Then we compare the prediction of GP with MD results appeared in the literature. [18, 36, 48].

6.1.1. Optimization Setup

To find the location of optimum regions, we first determine the search domain by checking common force field parameters found in the literature [36, 79, 81]. It is also desirable to go

for a larger search domain as a first step. So the log(ϵ) range is chosen as [-3, 0], [-4, -1], and [-3, 0] for monovalent cations, anions and divalent cations respectively. The large search extent will provide us a good transferability for the whole domain because the Gaussian Process prediction will not work well where the observations are not made. Moreover, a large search extent may find potentially good results. Data normalization scales the results from different orders to relative errors, e.g. $\Delta G_{sol} \sim \mathcal{O}(100)$ while $r_{io} \sim \mathcal{O}(1)$. This normalization not only gives us intuitively insights about the hypersurface, but also improves the performance of optimization. The effective dimensionality of ϵ is low, so we used a base 10 logarithm to scale this feature.

When the search starts, log10(ϵ) and σ parameters are fed to the LAMMPS MD simulation package. First 3 random samples are drawn among the parameter space to initiate the optimization, followed by 27 optimization steps to query the parameter space. This choice is more than enough for locating the optimum regions. The kernel function is chosen as the MA5 kernel, Appendix C.3. In most cases, the acquisition function is selected to be the expected improvement with the balance parameter of ξ = 3e-3. In some rare cases which the algorithm overexploits one optimum, ξ is increased to 1e-2, or the upper bond confidence acquisition function is employed. The same configuration is used to calculate the SFE and IOD, while L-J parameters are determined by the optimization framework and charge of ions are predetermined to be either unscaled or scaled.

Figure 6.1 breaks down the search process into steps. Bayesian optimization framework will use acquisition function as a guidance to search the optimums of the given black-box function. Contour plots in this figure represent the GP predicted mean and standard deviation (test points) of the cost function. The black circles are sampled observations (training points) in this optimization run. As the number of observations grows, one can see the Gaussian Processes become more confident about the shape of the objective function. The standard deviations are brought down to a very low level near the optimum region where most of the observations are made. By emphasizing search on the optimum domain, one can obtain more solid information for the desired region so that the final isoline can be obtained in an efficient and reliable way. The convergence of the L-J correlations can be found in Figure 6.2. In this figure, there is a slight difference between the isoline using 20 and 30 observations. The location of this difference happens to be the region where the uncertainty is high in Figure 6.1. There are no significant improvements for the obtained correlation after about 30 observations.

Because there is no standard L-J correlation function with TIP4P/2005 water model available, we use the best correlation we have found as a reference to measure the efficiency gain. We calculate the R2 score using the correlation in each optimization iteration while setting the best correlation as the reference. The random search method and its result is also shown for comparison. Figure 6.3 compares the performance of these two search methods for Na^+. Clearly, as a guided search strategy, Bayesian Optimization can converge faster than the random search method.

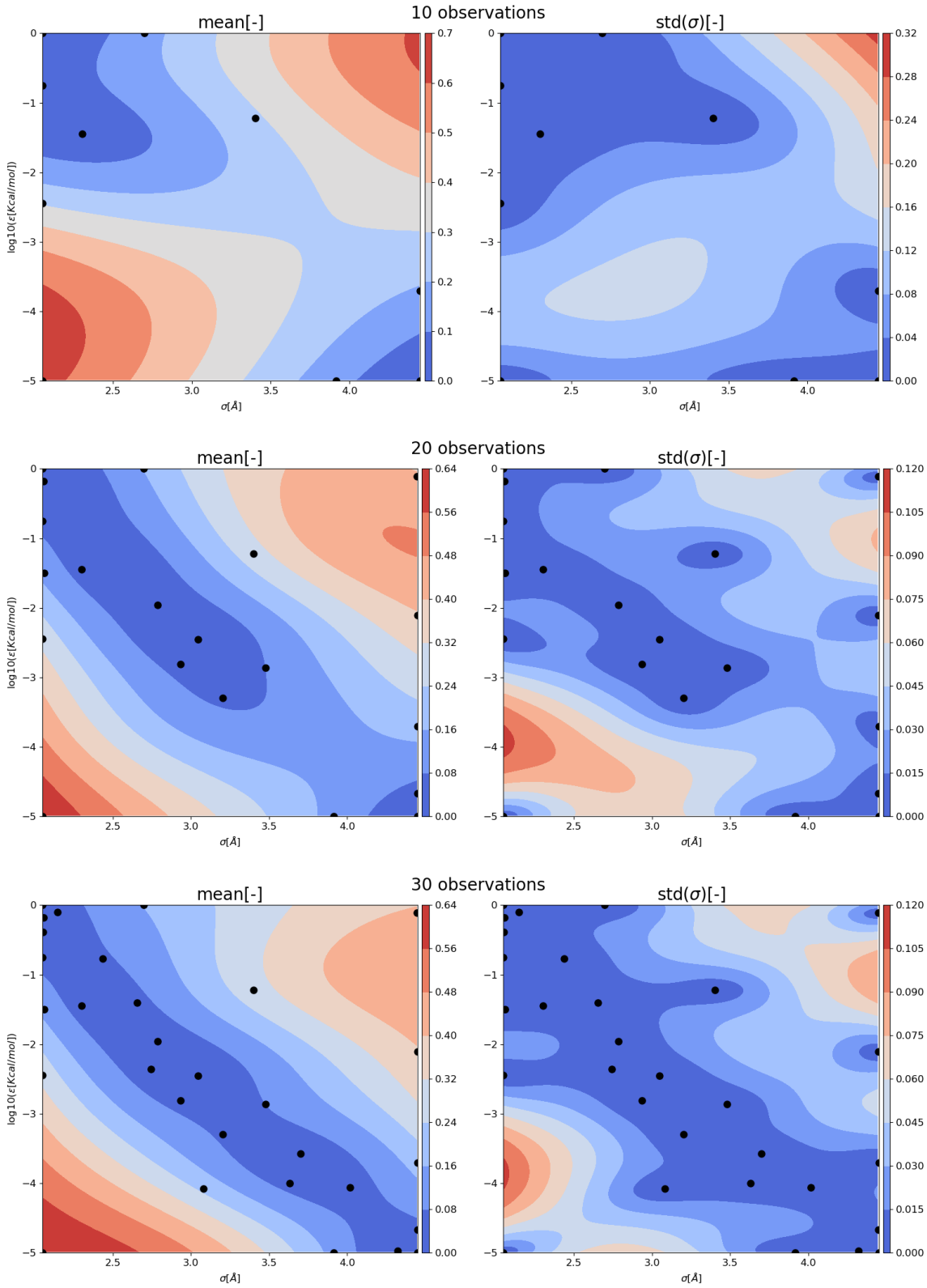


Figure 6.1: The mean and standard deviation of Gaussian Process for different observations during search. This biased sampling reveals more information near the optimum region.

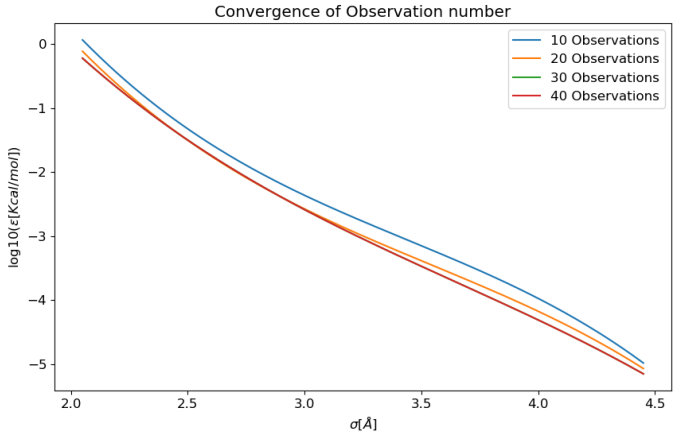


Figure 6.2: The obtained correlation for different observations.

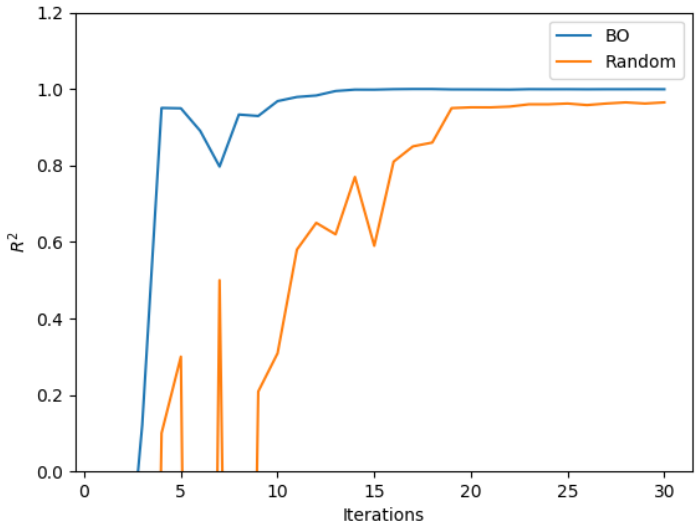


Figure 6.3: The R2 score between our model and random search. Target is Na^+ SFE isoline.

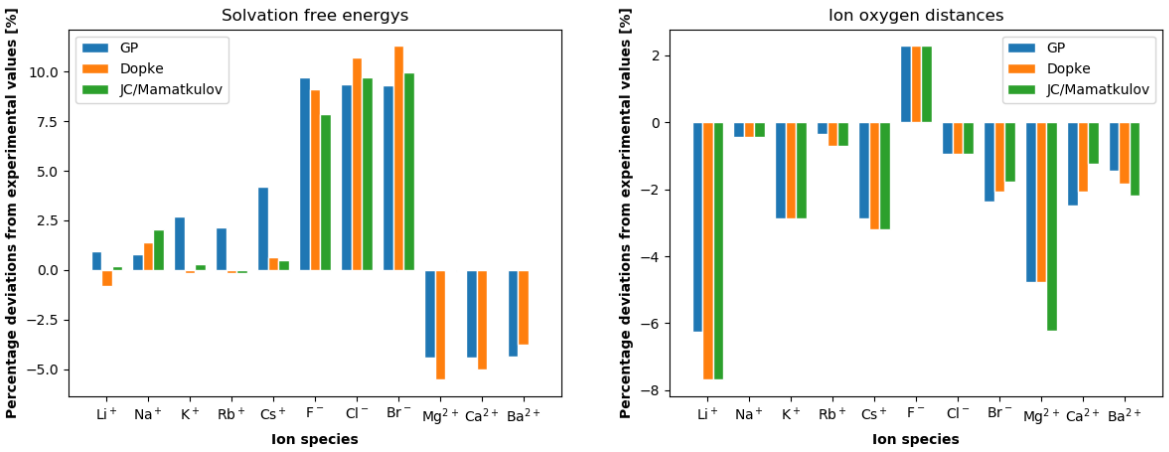
Extend methodology to other ion species and properties

With the optimization framework being settled, we can extend this methodology for different ion species: Li^+, Na^+, K^+, Rb^+, Cs^+, Cl^-, F^-, Br^-, Mg^{2+}, Ca^{2+}, Ba^{2+}. We also find the optimum regions with 2 charge conditions and 1 charge correction [18], their notations are defined as: (1) Unscaled ion charge case, $q = 1.00$. (2) Scaled charge proposed by Leontyev et al.[42, 43] $q = 0.85$. (3) Scaled charge with the SFE compensation treatment proposed by Döpke et al.[18] $q = 0.85^{\text{crc}}$. For all simulation cases, the experimental values are taken from Marcus et al.[49] unless stated otherwise. The search history for a particular ion species can be found in the Appendix A.3.

Compare the results with other studies

Figure 6.4 give an overview about performance of test parameter sets [18, 36, 48] reported from different sources. Due to the empirical nature, even the most accurate simulation result with force fields can deviate from the experimental values. So rather than considering the experimental results as the sole reference, it is more rational to also refer to MD simulation results in other studies. These test parameter sets [18, 36, 48] are shown as circle points in Figure 6.5, 6.6, 6.8. The prediction results of these test parameter sets can be found in Table B.1, B.2, and B.3. Discussions are carried out with the isoline map in the subsequent sections.

Note the water model employed is different for these selected references. Our training points are generated with TIP4P/2005 water model. Ref.[36] used TIP4P-Ew water model. Ref.[48] used SPC/E water combining with divalent cations. Ref.[18] used TIP4P/2005 water model. We compare force fields with different water models not only because there are very few studies that have reported the MD results of the TIP4P/2005 water model, but also because it has been found that there is a good transferability between TIP4P like water models[18].



(a) (b)

Figure 6.4: Using test parameter sets to evaluate the prediction accuracy of SFE (left) and IOD (right), compared with results from previous studies. Numerical values can be found in Table B.1, B.2, and B.3.

6.1.2. Monovalent cations

Figure 6.5 gives an overview of the obtained isolines for monovalent cations: Li^+, Na^+, K^+, Cs^+, Rb^+. Figure A.13 and A.14 give the search history, prediction from GP and corresponding found isolines for the monovalent alkali metal ions. Monovalent cation is believed to be the most representative ion species group. Many previous studies have parameterized these ions with early developed water models [7, 23, 36, 81]. In the correlation maps, the parameter sets from the left side of the SFE isoline give SFE results larger than the experimental values, while the sets from the right side of SFE isoline give results smaller than experimental values. The trend for IOD isoline is inverted. The parameter sets from the left side give IOD results smaller than experimental values, and parameter sets from the right

side give IOD results larger than experimental values. One can also obtain the prediction of target ionic properties for test points from search figures in Appendix A.3.

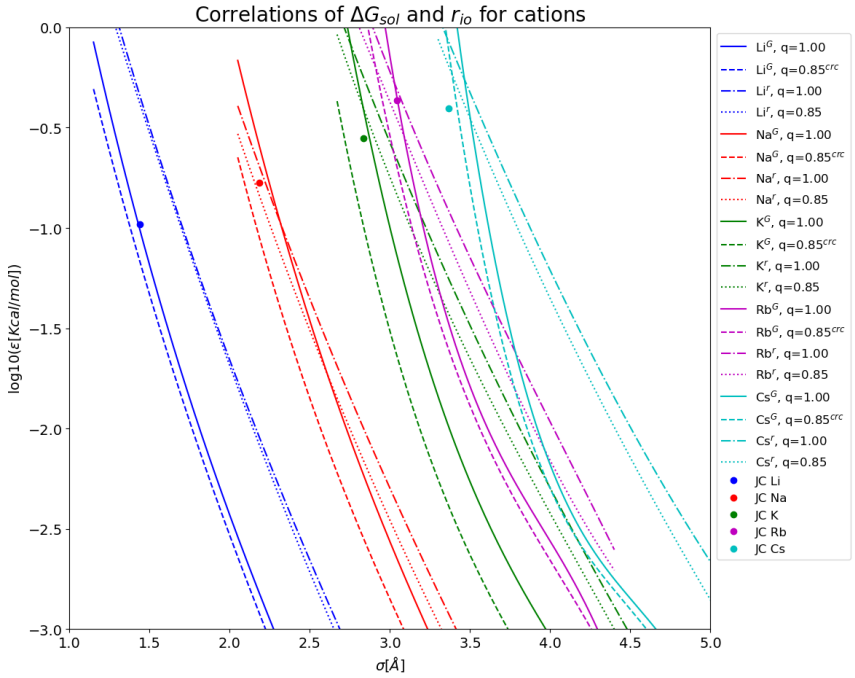


Figure 6.5: Isolines for monovalent cations. The solid and dash-dot lines, e.g. Li^G, $q = 1.00$ and Li^r, $q = 1.00$, represent the isolines of SFE and IOD under charge equal to 1. The dash and dot lines, e.g. Li^G, $q = 0.85^{\mathrm{crc}}$ and Li^r, $q = 0.85$, represent the isolines of SFE and IOD under charge equal to 0.85, with SFE result applied compensation treatment. The experimental value are from Marcus[49].

Lithium: In the Figure 6.5, JC Li^+ (test parameter set) [36] lies near the SFE isoline, while it is relatively far left of the IOD isoline. This can be explained by that JC's parameterization of Li^+ prioritize SFE. It is also very interesting to find the parameterization study from Zeron [81] also used exactly same parameter set for Li^+. For this parameter set, deviations from experimental IOD result in these two references are found as -7.7% and -5.8%, while prediction of our model gives a deviation of -6.2% [18, 36].

Sodium: Location of JC Na^+ is close to the intersection of SFE and IOD optimum regions in Figure 6.5, suggesting this parameter can be safely transferred to use with TIP4P/2005 water model. Using this parameter set, the deviations are found as +0.8% and -0.4% for SFE and IOD.

Potassium: In the Figure 6.5, JC K^+ matches the SFE isoline, while it locates to the left of the IOD isoline. This can be explained by that JC K^+ gives lower IOD results as it is verified in these two references (-3% and -3% respectively)[18, 36].

Rubidium: In Figure 6.5, one can find that the JC Rb^+ is near the intersection of SFE and IOD isoline. It further suggests that JC's parameterization is perfect in terms of SFE and IOD property. As this parameter set can simultaneously yield results that match the experimental values for these two properties. This agreement can also be found in these two references [18, 36].

Caesium: Our model overpredicts the experimental SFE result by 4.1% and underpredicts experimental IOD value by 2.9%, see Figure 6.4a and 6.4b. The IOD result from our

model agrees with Ref.[18] and Ref.[49], but the SFE result overpredicts both of these two references.

6.1.3. Anions

Figure 6.6 shows the correlation map found for F^-, Cl^-, and Br^-. The anions have a larger discrepancies between different target properties, compared with the monovalent cations. Moreover, due to more dispersive electron cloud structure, the anions generally have large size parameters.

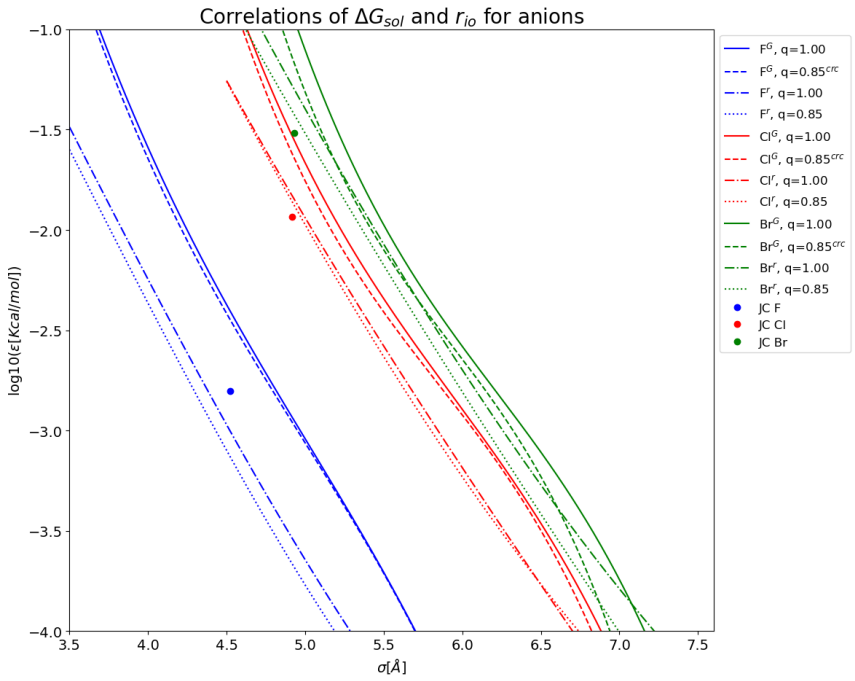


Figure 6.6: Isolines for anions. The solid and dash-dot lines, e.g. F^G, $q = 1.00$ and F^r, $q = 1.00$, represent the isolines of SFE and IOD under charge equal to 1. The dash and dot lines, e.g. F^G, $q = 0.85^{\text{crc}}$ and F^r, $q = 0.85$, represent the isolines of SFE and IOD under charge equal to 0.85, with SFE result applied compensation treatment. The experimental value are from Marcus[49].

Fluoride: In Figure 6.6, the JC F^- locates to the right of IOD isoline and to the left of SFE isoline. Specifically, our GP mean prediction deviates +9.7% for SFE and +2.3% for IOD. Ref.[18] suggests these two deviations are +9% and +2.3% and Ref.[36] gives these two deviations as +8% and +2.3% .

Chloride: The JC Cl^- locates to the left of both SFE and IOD isoline, which indicates our predictions are lower than the experimental IOD value and higher than the experimental SFE value, see Figure 6.6. Specifically, our prediction gives +9.3% deviation for SFE and -1.0% for IOD. This agrees with the MD results from the selected two references (+10.7% and -1.0% from Ref.[18], +9.7% and -1.0% from Ref.[36]).

Bromide: JC Br^- locates to the left of both IOD isoline and SFE isoline, with deviation from SFE larger. This indicates the the predicted MD results overestimated SFE target while underestimated IOD target, see Figure 6.6. Specifically, the deviation found for SFE is +9.3%, for IOD is -2.4% in our model, which agrees with the deviations from the references

in Figure 6.4a and 6.4b.

Effect of different target values

From the solvation free energy results of anions in Figure 6.4a, one can see that the deviations for anions F^-, Cl^-, and Br^- are surprisingly much larger ($\sim$ 10%) than that for the cation species. This is because there are two experimental sources that disagree with each other quite a lot[49, 65]. To further quantify this deviation, we have performed optimizations of cations with the Schmid values [65] as the target and the Marcus values [49] as the target. It turns out that JC's parameterization for anions will pass the isoline targeting the Schmid values, see Figure 6.7. This is one of the reasons that the discrepancies for anions are large in Figure 6.6. It should be noted that this deviation is not an accident, as many parameterizations of anions with TIP4P like water models: Studies of Jensen, Smith, and Joung [32, 36, 69] have all obtained SFE results closer to the Schmid values. The reason for targeting different experimental values is that the shift of one target property can gain performance of other properties. In other words, the discrepancy between the IOD and SFE isolines can be narrowed down by shifting to a more reasonable target value.

It is worth noting that here we used the TIP4P/2005 water model while JC calculated their results with TIP4P/Ew water model. This similarity again shows that there is a good transferability between the TIP4P/2005 water and TIP4P-Ew water[18].

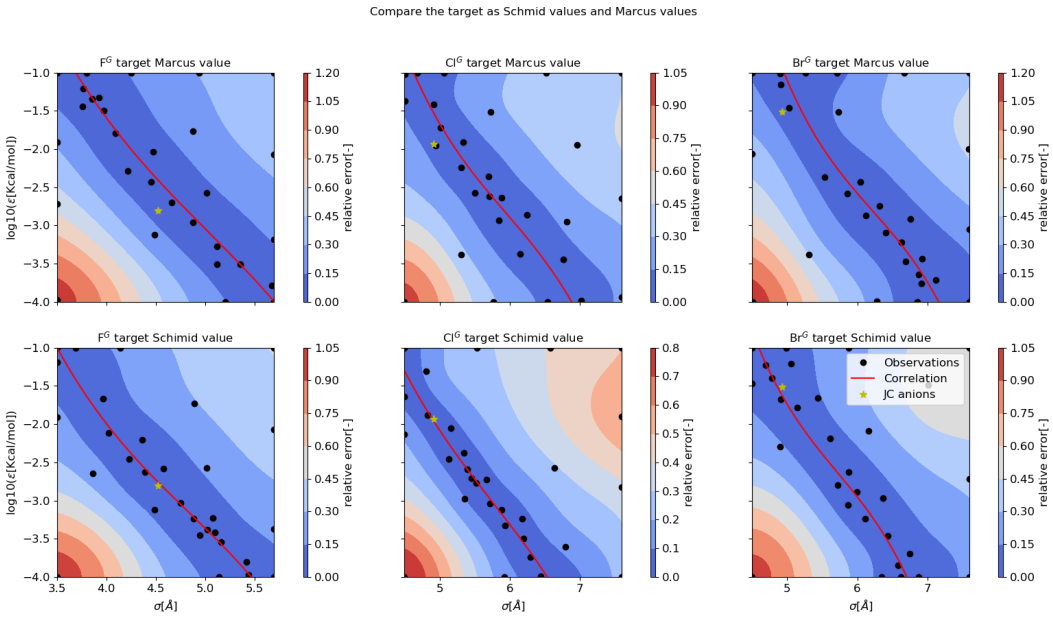


Figure 6.7: Shift of SFE isolines due to change of target values. First row contains search results with Marcus values as the target. Second row contains search results with Schmid values as the target.

6.1.4. Divalent cations

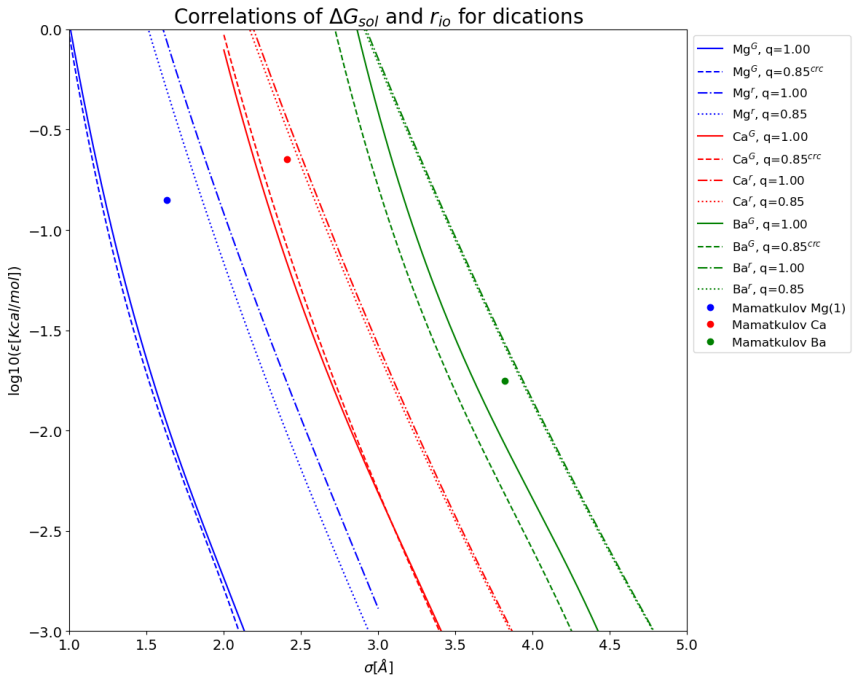


Figure 6.8: Isolines for SFE and IOD of divalent cations. The notations are same as before.

In the case of divalent cations, we use Mamatkulov parameter sets to examine the performance of GP predictions. The MD simulation results of TIP4P/2005&Mamatkulov force field combination have been reported in Ref[18]. The predictions for all the divalent cations studied in this project (Mg^{2+}, Ca^{2+}, Ba^{2+}) yield satisfying results compared with references, see Figure 6.4a and 6.4b.

The Mamatkulov parameter sets locate to the right of SFE isoline and to the left of IOD isoline for all three divalent cation species, Mg^{2+}, Ca^{2+}, and Ba^{2+}. This indicates that all these parameter sets underpredict both SFE and IOD results with TIP4P/2005 water. One can quantitatively compare our GP predictions with MD results from Ref.[18] in Figure 6.4a, 6.4b, and Table B.3.

6.1.5. Conclusions of correlation maps

Figure 6.5, 6.6, and 6.8 give a clear view of force field performance discrepancies between two target properties: SFE and IOD. The monovalent cations (Li^+, Na^+, K^+, Rb^+, Cs^+) have the lowest discrepancies between SFE and IOD isoline, compared with divalent cations (Mg^{2+}, Ca^{2+}, Ba^{2+}) and anions (F^-, Cl^-, Br^-). This is also reflected by that monovalent cations have clear intersections of SFE and IOD isoline, while the other ion species do not. Monovalent anions are parameterized with large size parameters, so they could be easily polarized. Thus a high level of discrepancy between target properties can be found. Moreover, the choice of reference data[49, 65] also contributes to this discrepancy. Small and high charge ions like Mg^{2+} and Ca^{2+} have low polarizability, but they will polarize the water in the system due to high charge density. Therefore the polarizable water model is more advisable for narrowing down the discrepancy of high charge ions. The charge scaling

treatment will reduce SFE results significantly, while the compensation of Döpke et al.[18] shows good performance. This can be reflected in the correlation map that the isolines with ECC and compensation are close to the isolines with unscaled charge, suggesting the SFE results of scaled charge cases are compensated correctly to the unscaled charge cases. The charge scaling treatment will also slightly increase the IOD results. These correlation maps provide a clear idea of how the 4 force field parameters (ϵ, σ, q, m) will influence the ionic target properties. From Figure 6.4a and 6.4b, we can also conclude that the test parameter sets [36, 48] performed very well with SFE. However, due to that they did not consider IOD as a target, their force fields perform very poorly in terms of IOD.

For the sake of easy representations, some studies have approximated the optimum region with mathematical equations[36, 79]. Here we use the following expression :

$$\epsilon = 10^{\sum_{i=0}^{3} c_i * (\sigma)^i} \tag{6.1}$$

The coefficients c_i for different isolines are listed in Appendix B.5. Since the parameter sets on the isoline are all optimums of a specific property, the original 2D multiple objective optimization problem can be simplified through finding this isoline. We can reduce the number of independent variables by one because two L-J parameters are correlated with each other.

The discrepancies between force field simulation results of different properties have also been reported in other studies [23, 48]. These discrepancies are imposed by the empirical formalization of Lennard-Jones 6-12 potential rather than errors from our model. The balance of discrepancies between different ionic properties is a very delicate job, as it is not possible to find a parameter set that gives an agreement between all target properties. There must be some trade-offs. We can combine the isolines of different properties to achieve a better representation of trade-offs. For each ion species, we can use the GP prediction to generate costs of mean for test points in the unsampled domain. After the data normalisation, the costs of different properties are at the same order of magnitude. We add these costs together to get the combined cost:

$$\text{Total costs} = \sum_i \text{cost}_i \tag{6.2}$$

The location where the combined cost is the lowest can be used to locate the mixed isoline. This combined isoline finds its location between the SFE and IOD isoline, so it will not give the perfect SFE and IOD results simultaneously. But the costs of these two properties will not be too high.

6.2. Diffusivity

We use two references to cross-validate the combined isoline of the primary targets: experimental results from Marcus et al.[49] and MD simulation results from Döpke et al.[18] with force field combination as TIP4P/2005&Madrid. We only show the diffusivity results using L-J parameters along the isoline in Figure 6.9, 6.10 and 6.11. The off isoline cases are attached in Appendix A.

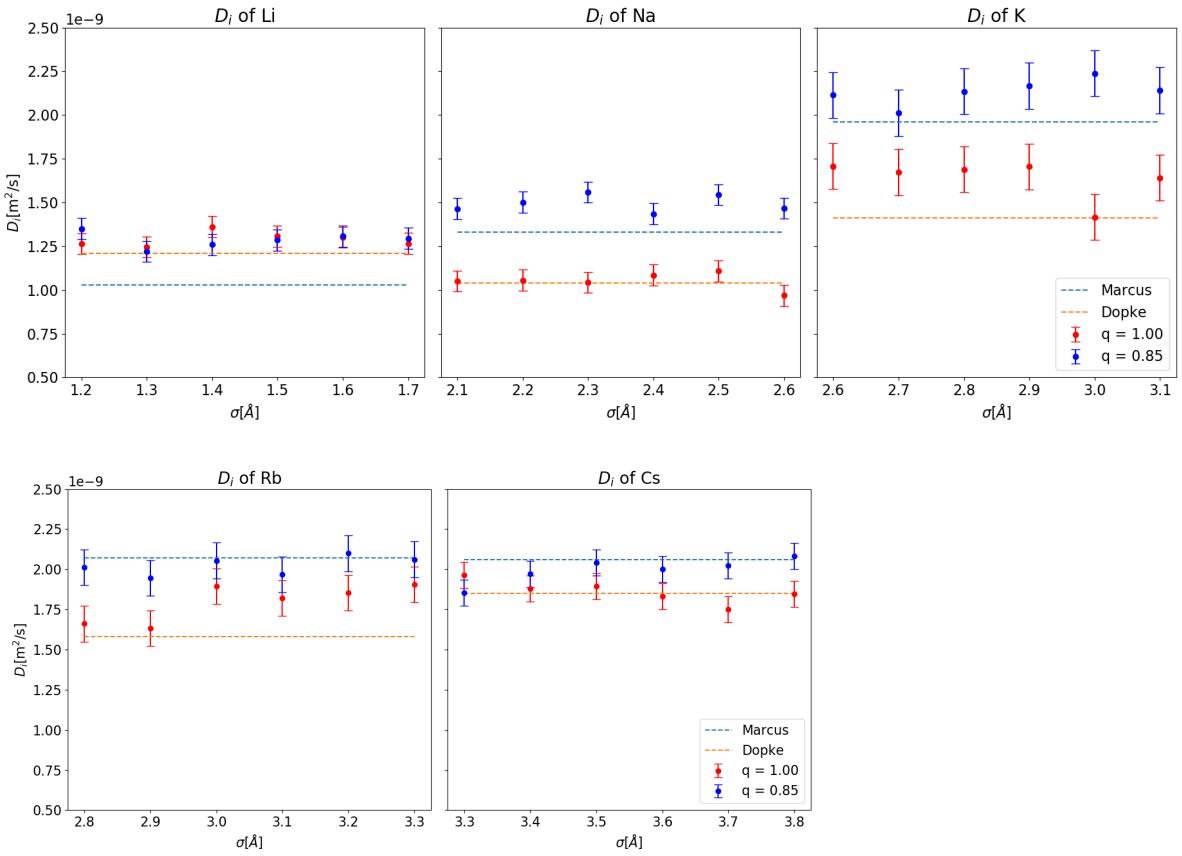


Figure 6.9: Self diffusivity of cations with ionic L-J parameters along their combined isoline.

Monovalent cations: The charge scaling treatment can significantly improve the self diffusivity results, except for Li^+. This is because the unscaled charge cases usually underestimate the self diffusion coefficient of monovalent cations. Consequently, ECC treatment improves the performance of ionic diffusivity results by increasing them, see Figure 6.9. The Li^+ is an exception because MD result of unscaled charge cases already seems to overshoot the experimental value. The variation of self diffusivity result is quite small along the isoline. In terms of parameterization trend off the isolines for monovalent cation ions, we have found that diffusivity will increase as the L-J σ parameter increase, this effect is most pronounced for Na^+ and K^+. While the diffusivity will increase as the L-J ϵ parameter increase, this effect is most pronounced for K^+, Rb^+, and Cs^+, see Figure A.1 and A.2.

Anions: In Figure 6.10, the variation of ionic diffusivity along the combined isoline is also very small. It can be seen that MD results along the isoline all underpredict experimental values. This discrepancy is found to be the largest for Cl^-, which almost reaches 25% of the experimental value. The charge scaling treatment can compensate for these underpredictions, gives nearly the same results as the experimental values. As for the parameterization trend off the isolines, the increase of both the L-J parameters can cause the diffusivity to increase. This effect can be observed for all anions that have been studied, see Figure A.1 and A.2.

Divalent cations: We can see that the diffusivity of all three divalent ion species has increased by about 0.15 [1e-9 m^2/2] after scaling charge from 1.00 to 0.85. However, the performance of Mg^{2+} and Ca^{2+} has not been improved because the unscaled Mg^{2+} and

Ca^{2+} already matches the experimental value well. The performance of Ba^{2+} has been improved after the charge scaling treatment. As for the off isoline region, a slightly increasing trend is found only for Ba^{2+} when we are varying the L-J ϵ parameter. This indicates that L-J parameters are not the dominant factors for determining self diffusivity of divalent cations Mg^{2+}, Ca^{2+}, and Ba^{2+}, see Figure A.1.

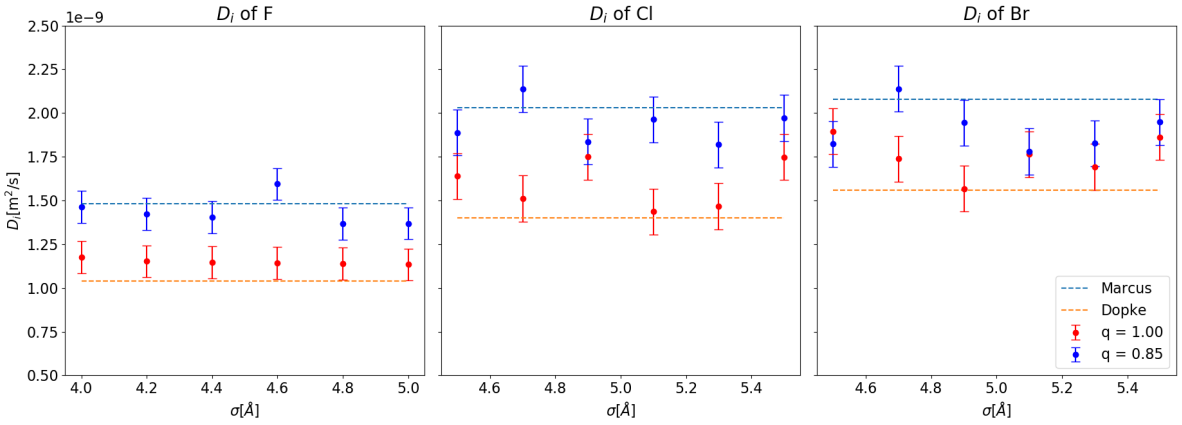


Figure 6.10: Self diffusivity of anions with ionic L-J parameters along their combined isoline.

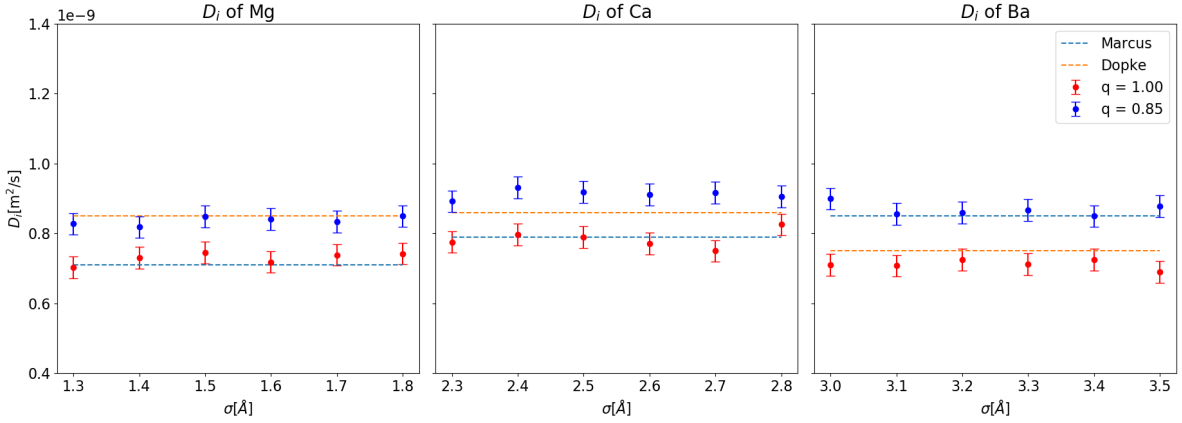


Figure 6.11: Self diffusivity of divalent cations with ionic L-J parameters along their combined isoline.

6.3. Conclusions for diffusivity calculations

Probing along the optimum region of primary targets, we have found the diffusivity results tend to underpredict the experimental values, except for Li^+, Mg^{2+}, and Ca^{2+}. The use of scaling charge can improve the result of obtained ionic diffusivities. This phenomenon has also been reported in the study of Benavides et al.[7], where they performed calculation of concentration dependence of self diffusivities for Na^+ and Cl^-. We have extended this calculation for other common ion species and it shows that scaling charge can increase diffusivity result to a mount of 0.1 - 0.4 [1e-9 m^2/s] for L-J parameters from the optimum

regions of corresponding ion species. Therefore, a force field with scaled charge treatment should be considered when the diffusivity is aimed for a target.

For ions with small size parameters like Li^+, Mg^+, and Ca^{2+}, their diffusivity results already overshoot the experimental values. This is because these ions have high charge densities so the polarization of water molecules is significant.

6.4. Structural Properties

Now we investigate the structural properties. We use the radial distribution function (RDF) to describe the ion-water structure and ion-ion structure in section 6.4.1 and 6.4.2.

6.4.1. Ion-water structure

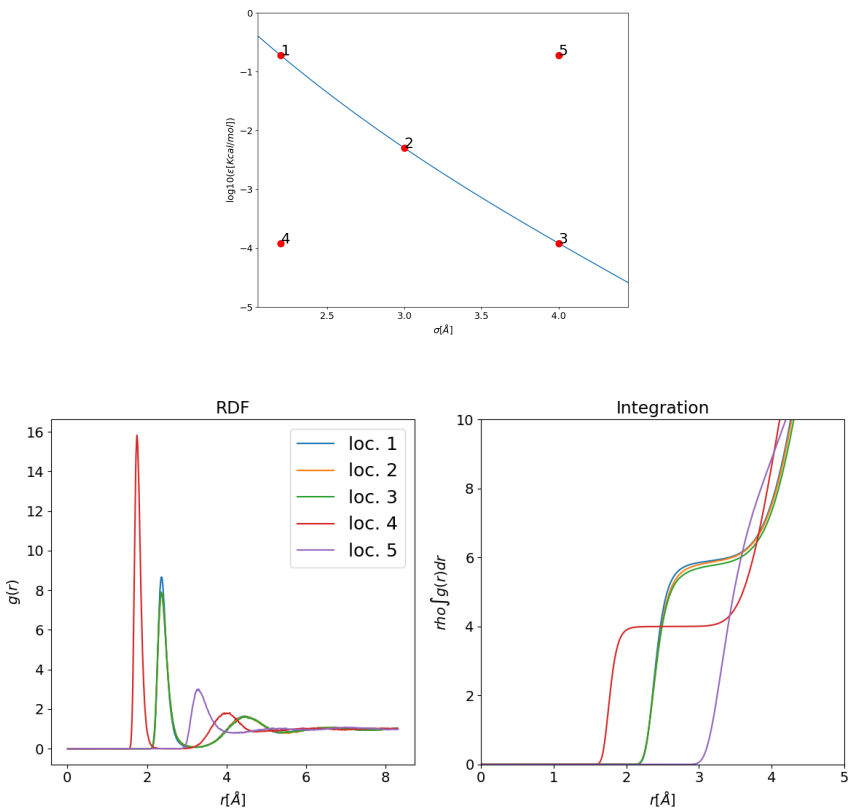


Figure 6.12: Cation oxygen RDF at different locations in the parameter space and their corresponding integration plots. The water model used here is TIP4P/2005

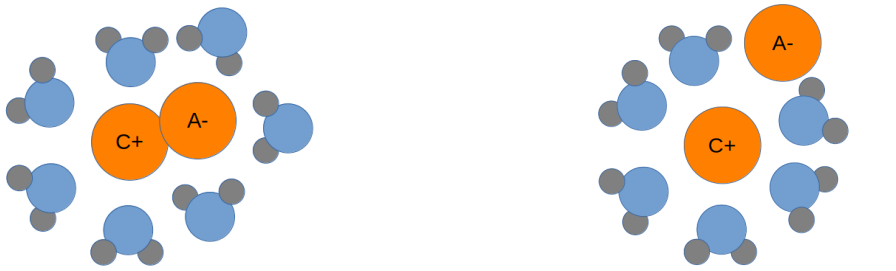
Figure 6.12 shows the RDF plot for 5 different locations, describing the ion-water structure with different force field parameter sets. The water model used here is TIP4P/2005. It reveals that the radii of the first hydration shell are indeed the same for all cases along the isoline, as shown by the peak position of locations 1, 2, and 3 in Figure 6.12. Furthermore, the height of peak will decrease as the σ increase, and the coordination number will also

slightly decrease as σ increase. The location and size of the second hydration shell as well as the other parts of the RDF are nearly the same for simulation results along the isoline. This similarity suggests the local structures for ions along the IOD isoline are nearly identical.

At location 4, where both of the L-J parameters are small, the IOD result is smaller than that from the isoline. The height of the first peak is higher than that along isoline, and yet the integration value is lower. At location 5, the height of the first peak is rather low and the second peak has disappeared, suggesting that the nested shell structure around the center ion has almost vanished.

6.4.2. Ion pairing

Ion pairing is commonly used by researchers to deduce the the solubility limit of electrolyte solution [7, 23, 36, 79, 81]. Figure 6.15 shows the cation-anion RDF. Inside the cut-off range, one can observe two peaks. The peak height indicating the probability of finding a distribution particle at that distance, relative to the average density of the system. The first peak represents the direct ion pairing where the cation and anion directly formed a cluster. The second peak represents the solvent separated ion pairing. Figure 6.13 shows these two types of ion pairs schematically. There is a long standing problem in ionic solution simulation that the direct ion pairing will lead to unphysical ion precipitation even at a low salt concentration[5, 7, 36, 79, 81]. Figure 6.14 gives an example of ion cluster under solubility limit. Once this ion-ion complex is formulated, involved ions will not be separated. This phenomenon can deteriorate the performance of a system and underpredict the solubility limit. Therefore, one should be extra careful with the direct ion pairs when performing force field parameterization [7, 23, 79, 81].



(a) Direct ion pairs.

(b) Solvent separated ion pairs.

Figure 6.13: Two different types of ion pairing

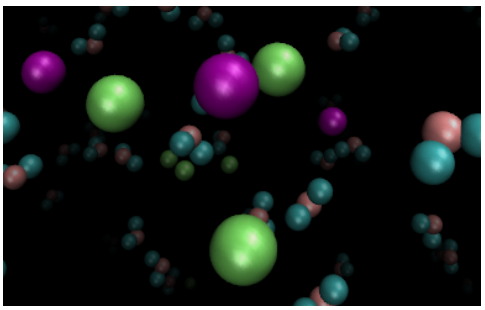


Figure 6.14: A typically ion cluster in MD simulation. One Ba^{2+} attracts three Cl^{-}. Involved ions will not be separated again.

It should be pointed out that the experimental data for the cation-anion contact pairs are generally non-accessible, so we take empirical criteria to measure the level of precipitation in the solution. An approach proposed by Benavides et al. has been accepted by various studies [7, 79, 81]. In this method, the contact ion pairs are employed to judge the level of ion aggregations. This is achieved by computing the CIP results near the salt solubility limit. A system without notable precipitations should have a low CIP value. They also proposed a threshold of 0.5 for checking [7, 81].

Conventional way of determining CIP number requires a large system size and a long simulation length such that the full precipitation can be observed. Moreover, at the solubility limit, the amount of ions could be at the same order of magnitude as the number of water molecules, e.g., LiCl has a solubility of 19.95 M. Such a high amount of additional particles will increase the computation complexity and make the system being hard to equilibrate. Therefore, apart from using the 0.5 criteria from Benavides et al. [7], we also determine the precipitation with the aid of observing the shape of cation-anion RDF. From our investigations, we have found that the shape of cation-anion RDF is very similar for systems of different concentrations. In other words, if the first peak is higher than the second peak at a high concentration, then the height of the first peak will also be higher at a low concentration. This similarity between the RDF shape has also been reported in the study of Fyta et al. [23] and Benavides et al. [7]. This can be explained by the probability of finding distribution atoms around the central atom will be influenced simultaneously by the change of salt concentration. A system without obvious precipitation usually has a low height of first peak, see Figure 6.15 for the first case. Therefore, we use the height comparison and CIP result to determine the level of precipitation.

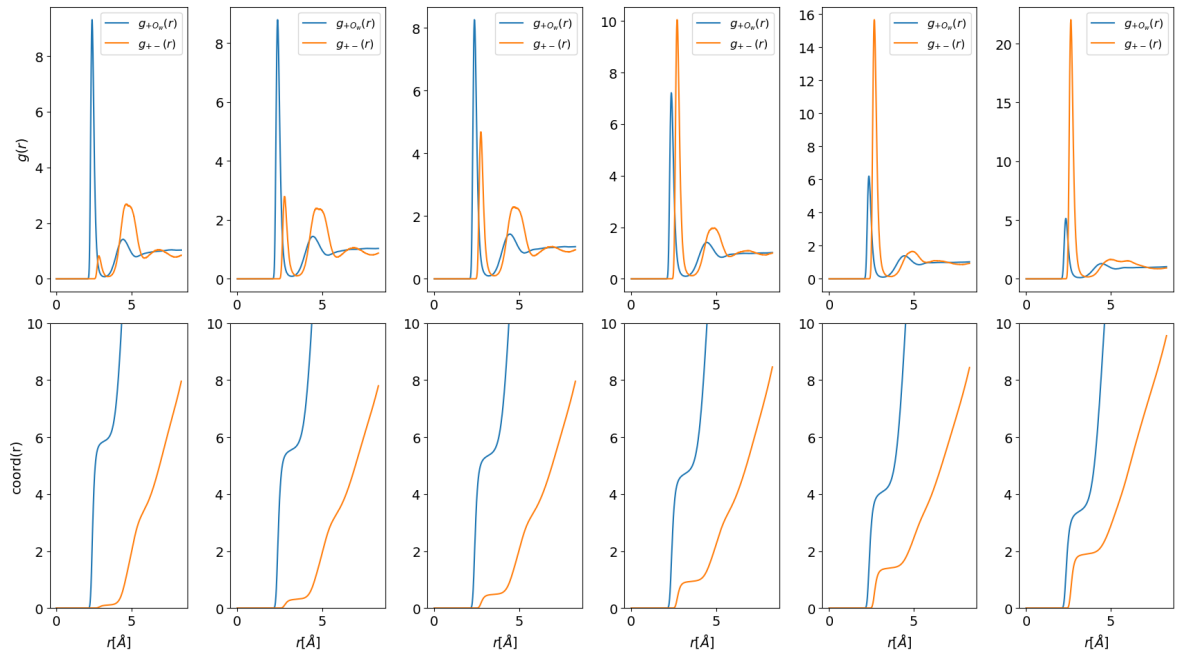


Figure 6.15: Contact ion pairs of NaCl. The counter Cl^- is taken from JCTIP4P-Ew force field. L-J parameters follow the combined isoline, the σ is increasing from left to right. The blue cases represent the ion-oxygen RDFs and orange cases represent the ion-ion RDFs.

Figure 6.15 also shows parameterization trends along the isoline. We only take the

cation as the center atom and determine the cation-anion RDF for verifying the parameterization trends. The JCTIP4P-Ew Cl^- is used as the counterion. Zeron et al.[81] and Benavides et al.[7] have suggested that the location of the upper integration limit of ion-ion RDF (the first minimum) should be very similar to that of ion-water RDF, so we plot them together for comparison. One can see clearly the height of the first peak increasing drastically as the force field parameters probe to the right segment of the isoline (larger σ and small ϵ). This trend has been found for monovalent cations and Ba^{2+}, see Appendix A.2 for other salt species.

Using parameters from the isolines, Mg^{2+} and Ca^{2+} barely see any variation of CIP results, see Appendix A.2. This can be attributed to the charge density for Mg^{2+} and Ca^{2+} is very high as they are divalent ions with small σ parameters. They strongly attract the surrounding anions in simulation so the L-J parameters have no significant effect here.

6.5. Parameterization trends

These 4 force field parameters (σ, ϵ, m, q) control the overall performance of a L-J ion. This section provides reasonings for the parameterization trends of ϵ, σ, q. Mass of ions will not be considered as a variation here since it is fixed for a specific ion species. The general trend for dependence of mass is that heavy ion species from the same group will tend to have large L-J parameters. The significance of understanding the parameterization trends is to have a guide for the final calibrating. This is helpful for the cases where we cannot probe the full landscape of the target properties (D_i, CIP), but we still want to include them into the parameterization. Table 6.1 first summarizes the trend. In the subsequent sections we will provide reasonings for these trends. In this table, "+" stands for the property increases with the force field parameter increases, while "-" stands for the property will decrease with the force field parameter increase. Note for CIP computations, the parameter trends are found for L-J parameters of the center cations.

	σ	ϵ	q
ΔG_{sol}	-	-	+
r_{io}	+	+	-
D_i	+	+	-
n^{CIP}	+	-	+

Table 6.1: The qualitatively dependence of 4 properties on ions force field parameters.

6.5.1. Solvation Free Energy

The SFE (absolute value) of ion will decrease as the L-J parameters increase. This dependence trend has also been backed in many studies [23, 28, 36, 45, 46, 48, 79]. The charge of an ion will greatly influence its solvation results, because the charge magnitude is the major influence of Coulombic interaction in the two-stage TI.

The reason for the dependence on the L-J σ parameter can be attributed to the shift of ion size, which will result in a change of the hydration shell and cavity size. Hence the Coulombic potential will change accordingly. As the ion-water distance becomes large, the Coulombic potential will become weak. So the magnitude of solvation free energy drops.

As for the dependence on the L-J ϵ parameter, one needs to consider the potential in the system as well as the local environment near the first hydration. The change of the L-J ϵ parameter will indirectly affect the free energy result by altering the electrostatic potential. The strong Coulombic interaction causes the water around the ion to change their orientation to either the hydrogen or oxygen atom towards the central ion. The van der Waals potential here provides a repulsion to balance the Coulombic attraction. An increase in the ϵ parameter will cause strong repulsions so the IOD result will increase. Therefore, the increase in distance in return triggers the Coulombic interaction to become weaker, so the solvation free energy drops. The weak dependence of ϵ parameter also agrees with this indirect triggering effect.

6.5.2. Ion Oxygen Distance

The dependence of Ion-oxygen distance on these 3 force field parameters can be closely related to local structure near the ion. A large L-J σ parameter will give rise to a large IOD result, but the IOD result will not exceed the distance at minimum energy depth: $r_{io} < \sigma_{ij} * 2^{1/6}$ ($\sigma_{ij} = (\sigma_i + \sigma_j)/2$) in our simulations. Figure 6.16 provides a quantitative relation between the r_{io} and the $\sigma_{ij} * 2^{1/6}$. The fact that r_{io} is smaller than $\sigma_{ij} * 2^{1/6}$ also reveals the van der Waals part of water-ion interaction is repulsion.

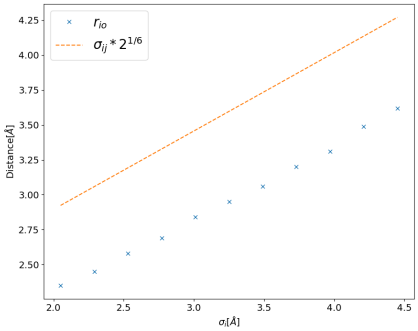


Figure 6.16: The ion-oxygen radii and distance at minimum energy depth vs. the L-J σ parameter of sodium ion. The energy parameter is set be a large constant (0.316 Kcal/mol) to maximize IOD. Water model used is TIP4P/2005.

The dependence of IOD on the L-J ϵ parameters is explained before in the section 6.5.1.

The dependence of IOD on ionic charge can be understood by the Coulombic interaction of ion and water. The scaled charge treatment (reduce the ion charge magnitude) will cause the ion-water Coulombic attractions become weak, so IOD result becomes large. This can be verified in our correlation map, where the scale charge treatment shift the isoline to the left bottom location.

6.5.3. Self Diffusivity

Figure 6.17 shows the local structure of the first hydration shell, which will influence the transport property of ion significantly. A compact hydration shell will hinder the Brownian motion of center ions, whereas a relative loose hydration structure will result in more

notable Brownian motion.

Due to the strong Coulombic attraction, the water around the cations will form a structure in which the oxygen orient towards the ions, see Figure 6.17a. This structure is considered to hinder central ions' motion since the net dipole moment of water are point to the center, and the second hydration shell can also be formulated in a similar orientation. In contrast, the local structure of the anions is very loose. One hydrogen site of water will orient towards the center anion while another hydrogen site will be attracted to the oxygen site from second hydration shell. Compared to the first structure where the total water dipole are point center, this kind of local structure is less stable. Moreover, the size parameterization of monovalent anions (~ 5Å) is much larger than that of monovalent cations (~ 2Å). The local structures of anions will have less constrain of the center ion. As for the influence of ionic charge, ion will diffusive more after reducing the charge magnitude. This is because the Coulombic interaction, which hinders the ions motion through the hydration shell, will be less pronounced after scaling charge.

(a) Cation.

(b) Anion.

Figure 6.17: The orientation of first hydration shell around cations and anions.

The increase of both L-J parameters can give rise to higher ionic self diffusion coefficients. This could be partially attributed to the fact that a large L-J σ parameters causes an increase of IOD, which makes the hydration shell become loose. Large L-J ϵ parameters cause the vdW repulsion stronger so ionic diffusion coefficients are increased. Appendix A contains the parameterization trends of diffusivity. The reason that one does not see monotonically increasing trends for all ion species can be attributed to that self diffusivity is a non-linear black-box function of force field parameters. Only for certain force field parameter combinations one can observe increasing trends.

6.5.4. Contact Ion Pairs

According to our MD results, along the isoline, a very low L-J ϵ parameter or a very high L-J σ parameter for cations will lead to the nonphysical direct ion clustering, see the ion-ion RDF in Figure 6.15 and Appendix A.2. This trend was also reported in the study of Fyta et al.[23]. This tendency can be explained by that a large cation size will attract more anions. Whereas a small energy parameter will expose the Coulombic interaction in the system so the ion-clustering phenomenon will become more significant. This trend makes CIP result along the isoline very distinguishable, along which the parameterization from the left segment (large ϵ) has a small CIP value and parameterization from the right segment has a large CIP

value.

6.6. Design Force Field Parameters

We can now select parameters from model that we have built so far. This section gives our solution for designing the 12-6 L-J parameters. It should be noted that the design of L-J parameters can have different criteria[7, 23, 36, 45, 48, 79, 81] and our design is only based on the single ion and ion-ion properties that we have studied in this project. Here our purpose is to illustrate the procedures of designing ionic force field parameters. Based on the L-J correlation we have found, although we can pick various points along the combined isoline or either one of the single isoline. It is still problematic to determine a specific single final parameter set from these correlations. Even if we introduce more ionic properties to find more optimum regions and parameter correlations, we still need to balance the costs of choosing one parameter set over another. The optimum regions of different targets would not cross for the same set of force field parameters. There are always some discrepancies between optimum regions of different target properties. There is no such perfect parameter set for us to select; trade-offs must be made.

The determination of the final parameters varies from different research groups. Joung and Cheatham[36] used the solid-state properties of ions crystals as a reference for their design. Yagasaki et al. calibrated their parameter set to match the salt solubility according to the crystalline potential energy[79]. Mamatkulov et al. determined their parameter sets by calculating the activity derivatives through the Kirkwood-Buff (KB) solution theory and chose the one that yields the closest result to the experimental values[48]. Li et al. used L-J parameters of noble gas to narrow down the search range[45]. Inspired by Li's work, we believe it is reasonable to select the parameters according to chemical species group, e.g. the alkali ions group, and use L-J parameters of neighboring ion species to narrow down the search range.

The 12th power term in L-J potential describes the Pauli repulsion due to the overlap of electron orbitals in the short range. The 6th power term represents the van der Waals attraction in the long range. Clearly, chemical species from the same group have the electronic structures differ in the number of orbitals. A more dispersive electron cloud means both of the L-J parameters should be large. Take ions from the earth alkali group as an example. In our study, ions from this group should have L-J parameters smaller than the Cs^+, since Cs^+ has the largest number of electrons in this project. Moreover, the parameterization of Rb^+ should be located between the parameterization of K^+ and Cs^+. Hence we could use a curve to connect the parameterization of ions from same group. This curve represents the σ and ϵ relation for the ions with similar electron cloud structures. Essentially it provides stability for the force field parameters of ions. Together with the previously found isolines of SFE and IOD, which essentially provides stability for the force field parameter of water, the force field design problem will be easy to tackle. The dependence study we have carried out also gives us significant assistance for calibrating the force field parameters.

Finally, we calibrate our force field parameters according to the following procedures:

- Near the combined isoline for SFE and IOD.
- Between the existing L-J parameterizations from the same element group.

- A low value of direct contact ion pairs.
- Closest self diffusivity to the experimental value.

The first attempt is to select the parameters near the intersection of the SFE and IOD isoline. This applies to all of the alkali metal cations except for Li^+, in which case there are no clear intersections in the chosen domain. For these cases with no clear intersections, we chose the parameters along the combined isoline. Then the selection is based on tuning parameters for the diffusion coefficient and CIP. Take the sodium as a concrete example, the first choice is ϵ = 0.11630 [Kcal/mol], the CIP near the solubility limit is found to be 0.49, which is quite close to the 0.5 threshold. So we went for a larger energy parameter and calibrated it with the diffusivity trends. The final parameter set is (ϵ, σ) = (0.18425 Kcal/mol, 2.20253 Å).

We also use one metric suggested by Vega et al. [76] to evaluate the L-J parameter of our design.

$$M = \max\left\{\left(10 - \left|\frac{X - X_{\text{exp}}}{P_{\text{tol}} X_{\text{exp}}}\right|\right), 0\right\} \tag{6.3}$$

Where the P_{tol} is the relative tolerance for property X. The score is 10 if the relative error is within 0.5 times the tolerance. The score is 0 if the relative error is 10 times the tolerance. The tolerance is set to be P_{tol} = 1% for SFE and IOD, P_{tol} = 10% for self diffusion coefficient.

Table 6.2: Selected parameters and corresponding obtained ionic property values. The units for values of SFE, IOD, and D_i are Kcal/mol, Å, and 1e-9m^2/s

	ϵ[Kcal/mol]	σ[Å]	value		
			SFE	IOD	Di
Li^+	0.11247	1.55087	110.4	2.00	1.27
Na^+	0.18425	2.20253	87.7	2.37	1.06
K^+	0.83278	2.75247	71.3	2.80	1.68
Rb^+	0.98174	2.94043	67.8	2.92	1.82
Cs^+	0.86889	3.36745	61.7	3.12	1.93
F^-	0.00060	4.84700	121.0	2.68	1.33
Cl^-	0.00618	5.14400	87.7	3.17	1.63
Br^-	0.00874	5.43700	80.8	3.37	1.74
Mg^{2+}	0.44770	1.30519	431.6	1.93	0.73
Ca^{2+}	0.69720	2.16307	351.1	2.36	0.80
Ba^{2+}	0.83068	2.92288	298.3	2.72	0.78

We conclude the evaluation of the final design force field parameters on four properties in Table 6.2. The performances of these parameter sets are compared with the transferred TIP4P-Ew optimized parameter sets from Ref.[36]. Figure 6.18 compares the performance scores between optimized parameters and transferred parameters, evaluated by Equation 6.3. Figure 6.19 summarizes obtained CIP values using the optimized parameter sets under different concentrations of salts. For LiCl, the solubility limit is too high (19 M) so we abandoned calculation of its CIP value near the solubility limit.

The transferred parameters [36, 48] performed well for SFE. But their performances are not so good for IOD and CIP. Therefore, we use the parameterization trends and optimum

regions that have been developed so far to balance the trade-offs. It should be pointed out that we are able to simultaneously optimize SFE, IOD, and diffusivity. Since there are clearly intersections between the optimum regions of SFE and IOD, and diffusivity results near that region are also quite good. However, we still tend to deviate from that parameter region in order to gain performance for CIP. The biggest SFE compromisations are made for Rb^+ and Cs^+ since their ion aggregation phenomena are the most severe among all monovalent ion species. After optimization, the performance of IOD, diffusivity, and CIP see notable improvements. The optimized parameter sets for RbCl and CsCl give CIP values of 0.410 and 0.770 at a moderate concentration of 4.5 M. Using the optimized parameters, the spontaneous ion aggregation issue has been eased. For anions, the performance is very poor for SFE due to the inappropriate target values as it has been discussed in section 6.1.3.

From Figure 6.18 and 6.19, we can see that there is great potential for optimizing better parameter sets for divalent cations. Our optimized parameter sets can simultaneously improve the performance of SFE, IOD, diffusivity, and CIP for divalent ions. Except for IOD result of Mg^{2+} has deteriorated slightly. Although salt with high charge density cation like Mg^{2+} seems hard to achieve a good CIP result. Salts with large size cation can have huge potential for easing the ion clustering issue. The CIP results are reduced to a low level both for $CaCl_2$ and $BaCl_2$ using the optimized parameters.

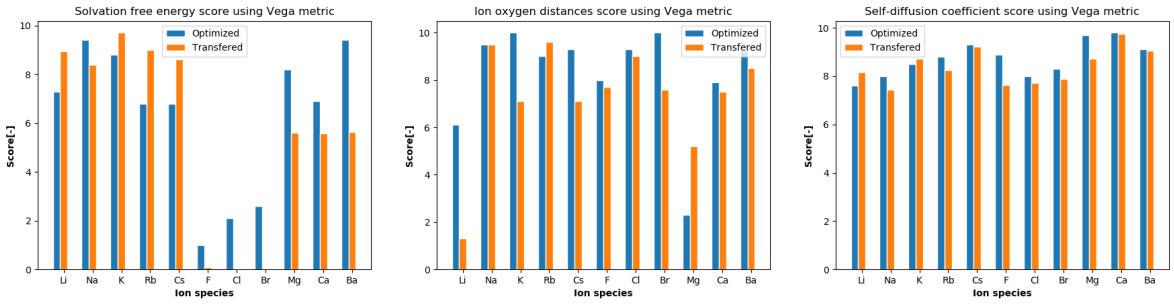


Figure 6.18: Comparsion of scores between the transferred parameters and optimized parameters.

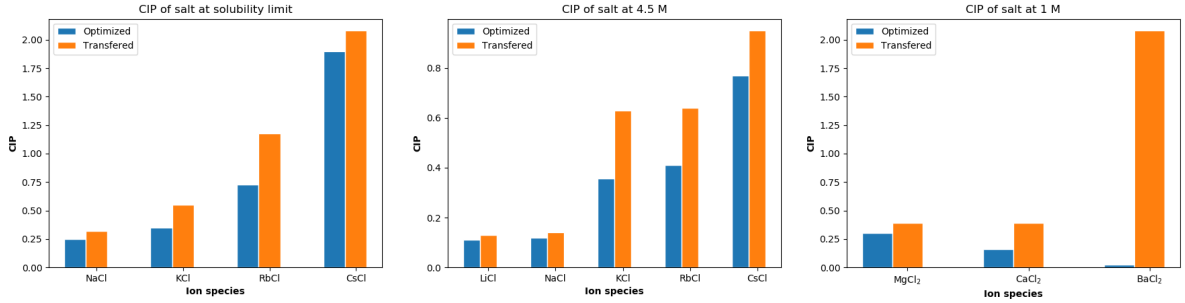


Figure 6.19: Comparsion of CIP results between the transferred parameters and optimized parameters.

7

Conclusion and Outlook

7.1. Conclusion

Parameterization of the ionic force field with nonpolarizable rigid water model TIP4P/2005 has been studied in this project. Now if we look back at the research objectives in the first chapter, we can answer them as the following

- (1) *Formulate an efficient alternative approach to sample the force field parameters with the help of a probabilistic Machine Learning model.*

An alternative efficient parameterization method based on Bayesian inference has been studied. This gradient-free optimization routine uses acquisition functions to guide the sampling and Gaussian Process to estimate the unknown parameter space. With the help of this probabilistic machine learning model, we can find the optimum solution for a force field design problem quickly. The efficiency gain is demonstrated by searching the optimum parameter sets in only a few optimization steps, compared to hundreds of iterations from the conventional grid search method from previous studies [28, 36, 45]. This method has a huge potential in tackling complex force field design problems with only a few iterations, in which case the number of parameters that need to be considered might be dozens.

- (2) *Explore the ionic force field parameter space and estimate the parameterization trends of target ionic properties.*

Four ionic properties have been investigated for force field parameter design, they are solvation free energy, ion oxygen distance, self diffusivity, and contact ion pairs. The corresponding MD simulations are carried out in the LAMMPS. We have found the optimum regions of SFE and IOD, which match the Marcus experimental values[49], through Bayesian Optimization. We have struggled in finding optimum regions for some other properties because accurately acquiring these data is not a trivial task. Specifically, the determination of ionic self diffusivity and CIP requires a certain level of human interaction. For diffusivity, the high statistical uncertainty causes the final result is not precisely described. For CIP, the height of the first peak varies for different simulation configurations, and one has to read the result via human interactions. To properly including diffusivity and CIP into optimization framework, we have used MD simulations to verify their parameterization

trends, which gives us a direction for further calibration. The dependence study is focused on the combined isoline of the primary targets. For ion species that have been studied, it has been found variations of self diffusivity are small within the common parameterization range along the isoline. Whereas the CIP results vary drastically, a small L-J σ parameter or a large energy parameter can significantly ease the direct ion-pairing phenomenon. The discussion of dependence has been reported in section 6.5.

We studied three ionic force field parameters in this project, namely, two L-J parameters (ϵ, σ) and the charge of the ion, q. It shows the σ parameter has a more significant influence than the ϵ parameter on the ionic properties here. In other words, the effective dimensionality of the ϵ parameter is low, so we used a base 10 logarithm to scale this dimension. The optimization formation after this scaling is much better. Scaling charge of 0.85 is an educated design for TIP4P/2005 water to include the electronic continuum contribution into consideration, so we computed ionic properties with scaled charge of 0.85 to investigate the influence of ionic charge. Our primary targets of SFE and IOD perform very well for unscaled charge cases. The scaled charge treatment only improves the performance of secondary targets. However, applying the charge scaling naively will damage the performance of the solvation property. This problem is subsequently tackled by a study of Döpke et al[18]. in which they derived theoretically that the correction of solvation property can be compensated through multiplying the raw solvation results by ϵ_{el} (which is $1/0.85^2$ adopted by Zeron). We have used MD simulation to prove this correction shows excellent performance, and it can be used as a solid correction in further study where the scaling of ion charge is involved.

- (3) *Identify the possible cheap simulation setup for speeding up the parameterization process.*

To balance the efficiency vs. accuracy dilemma for our force field optimization problem, we have explored different configuration setups and have chosen the cheapest configuration while guaranteed the accuracy, as it was discussed in Chapter 5. We have also used an order-n like algorithm implemented by Jamali et al[31]. to speed up the diffusion calculation. It shows clearly that we can exploit the configuration setup to gain efficiency. This is especially helpful for speeding up parameter selection process. The configurations were chosen differently for each specific property.

- (4) *Propose the major procedures for ionic force field parameters design. Balance the trade-offs for selecting different properties.*

First, one should select representative properties as the targets. The ideal properties should require no human interactions for an accurate description. Second, one should balance the costs of choosing different properties. Systematical parameterizations of more ion species should be carried out in the future. Finally, one needs to calibrate the model by cross-validation.

The conventional force field parameter sets are biased on specific target properties, e.g., solvation free energy. To address these misbalances, we could use the correlation map as a guidance to find better parameter sets. Our calculation has shown that there is still potential for finding better parameter sets for monovalent cations in terms of SFE and IOD. Since these optimum regions intersect in the parameter space, and the transferred parameters

clearly deviate from the intersected optimum region. However, the problem with heavy ions Rb^+ and Cs^+ is that the direct ion-pairing phenomenon is severe, even we have chosen a larger L-J ϵ parameter. The most pronounced improvements have been found for divalent cations, in which case there are no transferred parameter sets that are optimized for TIP4P water model. [36, 81].

7.2. Outlook

Abandon the mixing rules:
The conventional way of innocent applying mixing rules should be reevaluated. The interaction between ion-ion, ion-water, and water-water should be explicitly parameterized. For example, we can tackle the ion aggregation problem while keeping a good performance of single ion properties through parameterizing the ion-ion interaction and ion-water interaction separately. As a result of abandoning the conventional mixing rules, we can narrow down the discrepancies of different target properties.

Addition of more ionic properties:
More ionic properties can be considered in the future. Many recent parameterization studies have revealed limitations of existing ionic force field parameters with some exotic ionic properties[7, 36, 48, 81]. Besides, choosing an appropriate set of target properties is very important in the force field parameterization. Select feasible target properties according to the computing budget will be helpful for more productive parameterizations.

Perform systematic ionic force field parameterization:
More systematic optimizations should be carried out in the future. The conventional way of mixing force field parameters from different sources should be avoided. Due to computing power limitations, this work only performed optimization with TIP4P/2005 water model and 11 ion species. A more thorough parameterization of different ion species can be carried with an efficient Bayesian search.

Feature selection:
In this work, we have two correlated features, ϵ and σ. After scaling ϵ, the optimum regions are identified in the parameter space. A more interesting study could be carried out using implicit features like the forces and energies for force field optimization. This can be done via neural network learning QM data.

Use one cost function for multi-objective optimization:
Complex loss function can be involved for parameterization. Suppose we can acquire all the target properties in a precise way. In that case, we can build a loss function containing all the costs in terms of different properties for a parameter combination.

$$L(\epsilon, \sigma, q) = \sum_i c_i * X_i \tag{7.1}$$

Where the c_i is the importance assigned to property i and X_i is the corresponding relative error has been found under force field parameters (ϵ, σ, q). By directly minimizing Equation 7.1, one can automatically design the force field parameter through Bayesian Optimization.

Further development of ECC and correction terms :
The charge scaling treatment is a hot topic in this community. More ECC studies can be conducted for different ionic properties in the future. Besides, another good way to balance costs of different properties is to develop more reasonable compensation terms. The correction of MD results can narrow the discrepancies between different properties using one force field parameter set. Appropriate compensation terms could make one parameter set represents all the target properties.

Use a more direct calibration form:
Instead of the conventional $\epsilon - \sigma$ form, a more direct calibration form like the AB form can be used for specific target properties, e.g. transport properties, which might not be popular for previous parameter studies. This is not only because high uncertainty of self diffusivity causes quantifying the performance of a force field difficult, but also due to that conventional L-J parameters cannot represent the balance of attraction and repulsion. The ϵ has roles to play both in the attraction and repulsion term in the L-J potential expression. Hence variation of this parameter has little effect on the overall balance. Instead, the AB form can balance the repulsion and attraction more easily. See Equation 7.2.

$$\begin{aligned} u_{ij} =& 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] = \frac{A}{r^{12}} - \frac{B}{r^{6}} \\ \text{With} \quad & A = 4\epsilon\sigma^{12}, \quad B = 4\epsilon\sigma^{6} \end{aligned} \tag{7.2}$$

Where the A and B term represent the repulsion and attraction in the L-J potential. This collective description of potential performance makes the force field parameterization for more straightforward. Besides, once A and B are determined, the ϵ and σ can also be determined.

A

Additional Figures

A.1. Self diffusivity cut plane

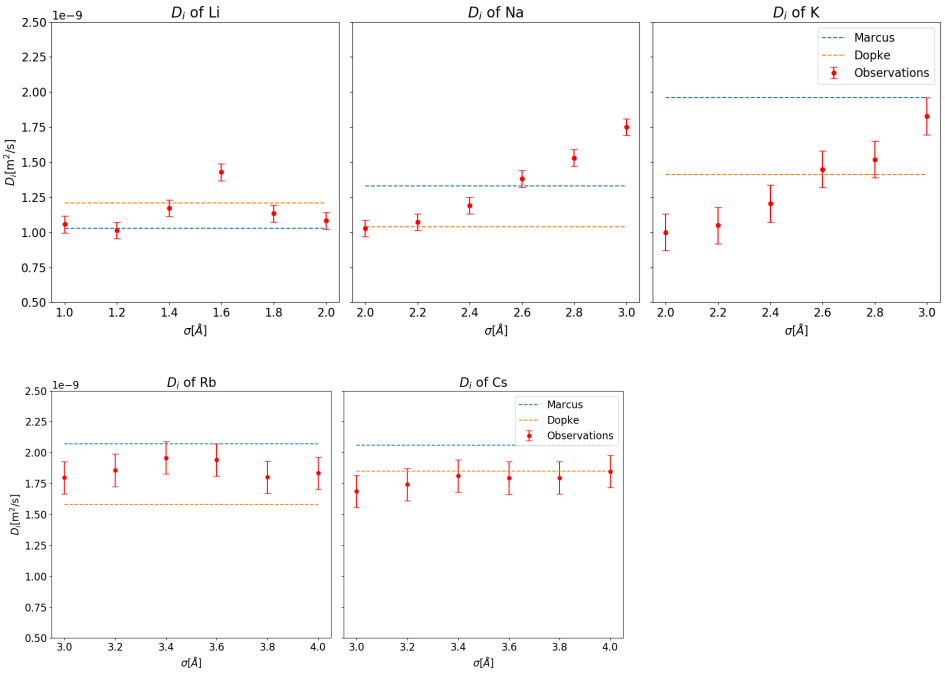


Figure A.1: Monovalent cations self diffusivity result of varying σ while holding ϵ as the JC parameters[36].

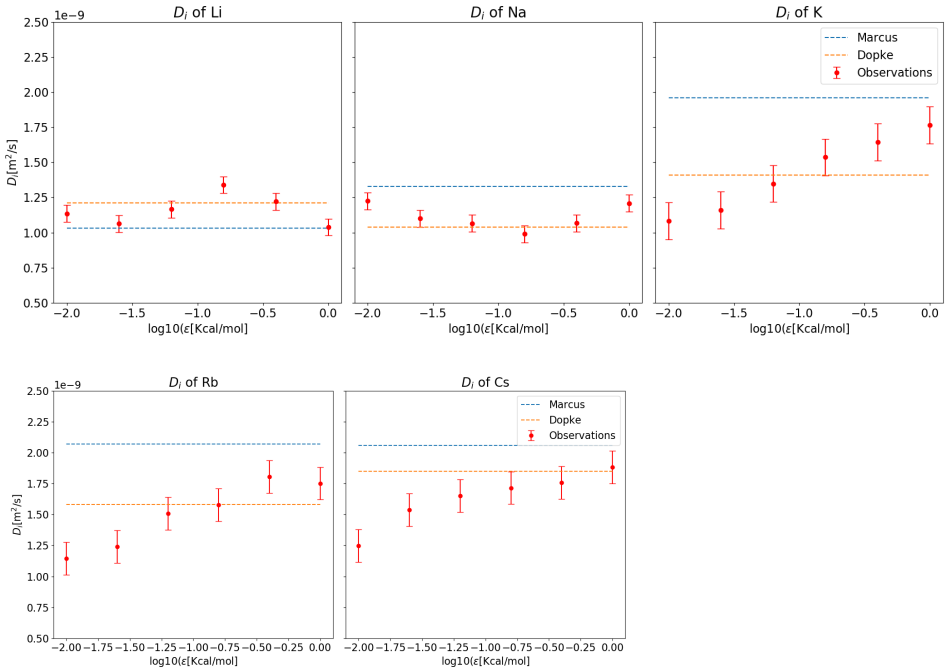


Figure A.2: Monovalent cations self diffusivity result of varying ϵ while holding σ as the JC parameters[36].

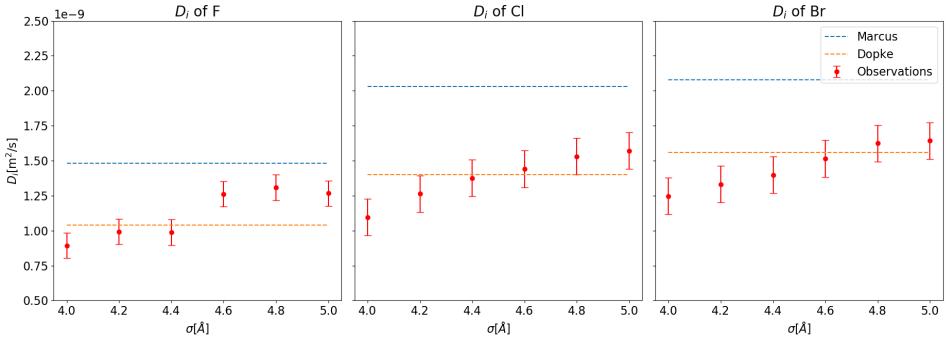


Figure A.3: Anions self diffusivity result of varying σ while holding ϵ as the Joung and Cheatham parameters[36]

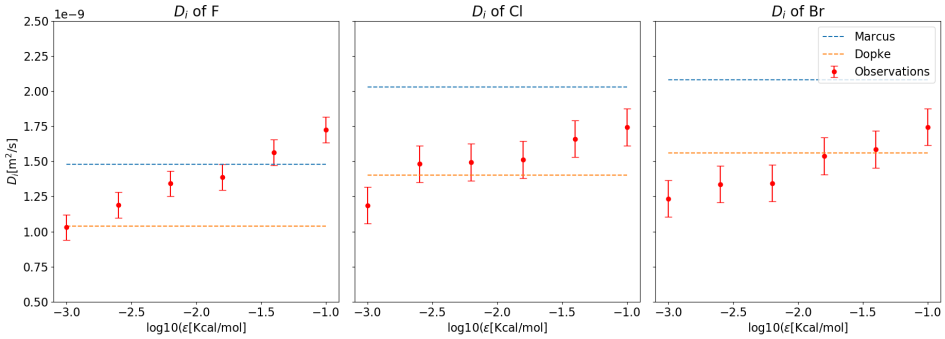


Figure A.4: Anions self diffusivity result of varying ϵ while holding σ as the Joung and Cheatham parameters[36].

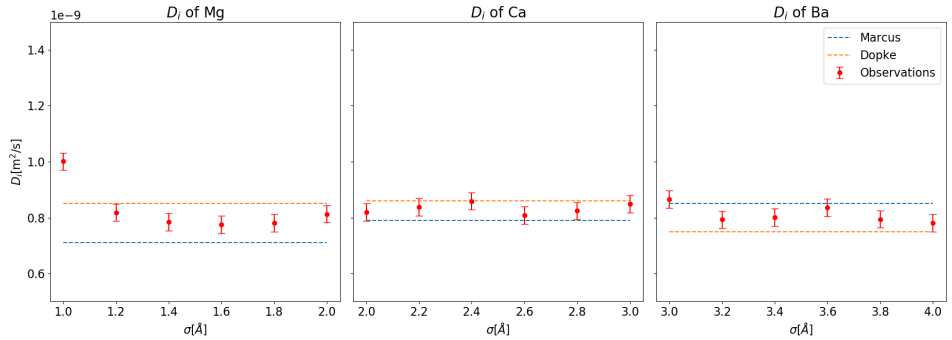


Figure A.5: Dications self diffusivity result of varying σ while holding ϵ as the Mamatkulov parameterization[48]

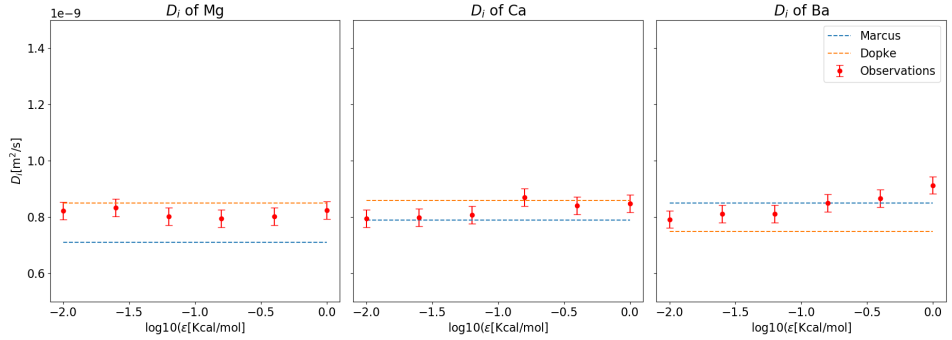


Figure A.6: Dications self diffusivity result of varying ϵ while holding σ as the Mamatkulov parameterization[48].

A.2. Cation-anion RDF alone the isoline

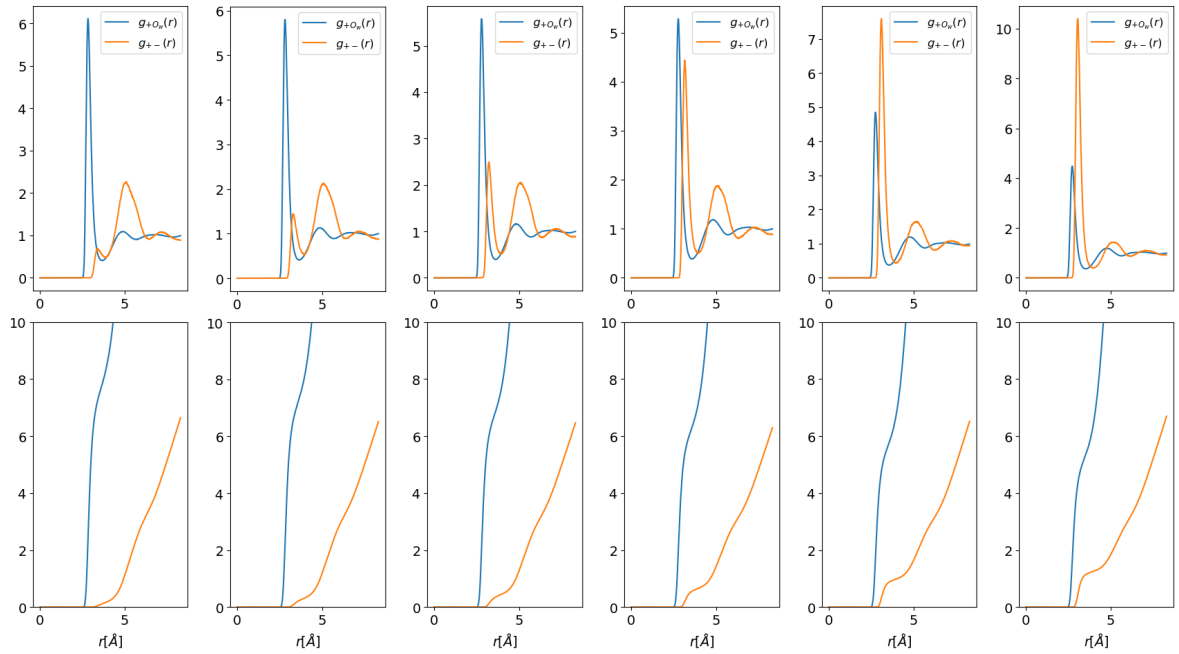


Figure A.7: Cation-anion RDF along the combined isoline of K^+.

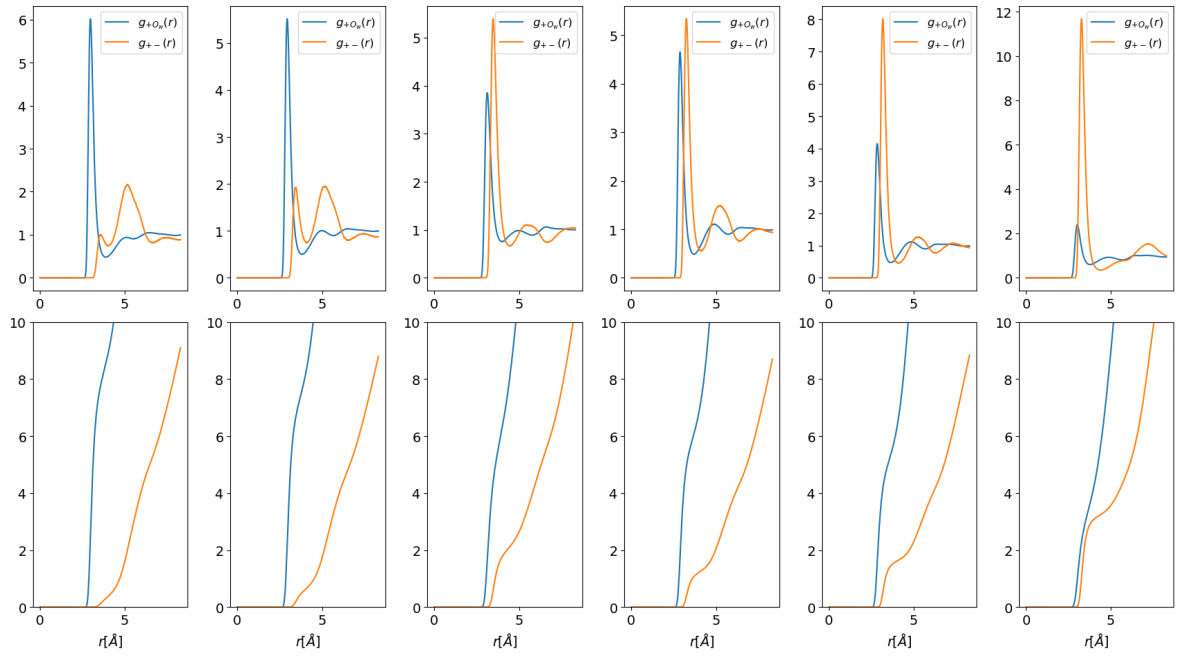


Figure A.8: Cation-anion RDF along the combined isoline of Rb^+.

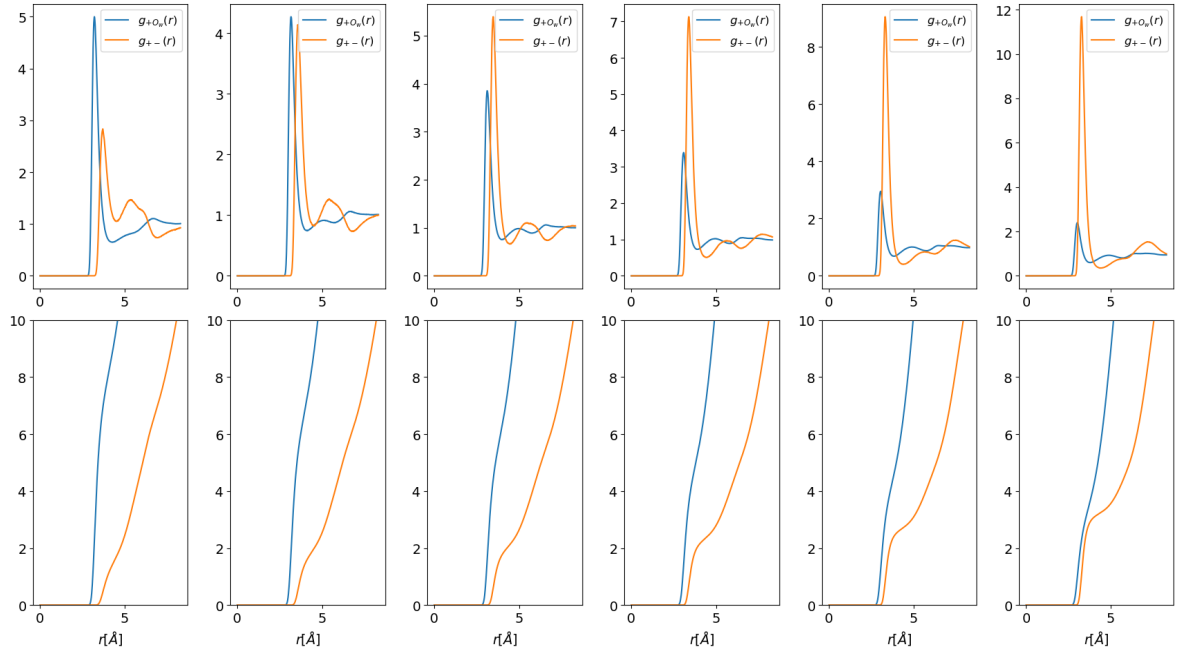


Figure A.9: Cation-anion RDF along the combined isoline of Cs^+.

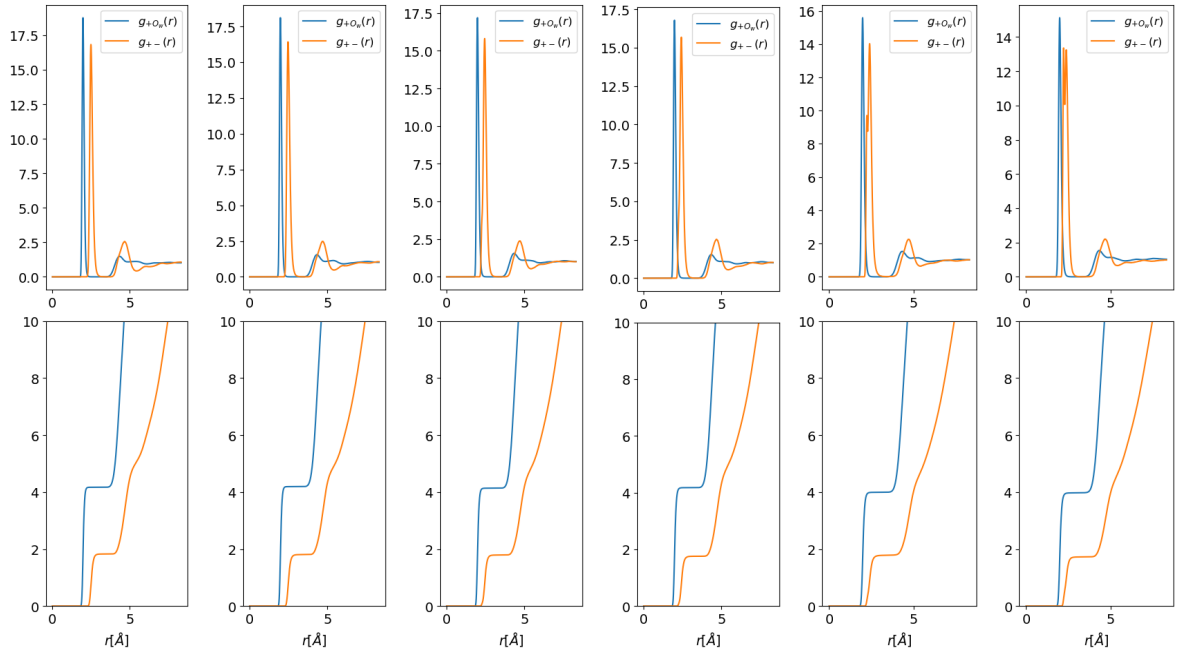


Figure A.10: Cation-anion RDF along the combined isoline of Mg^{2+}.

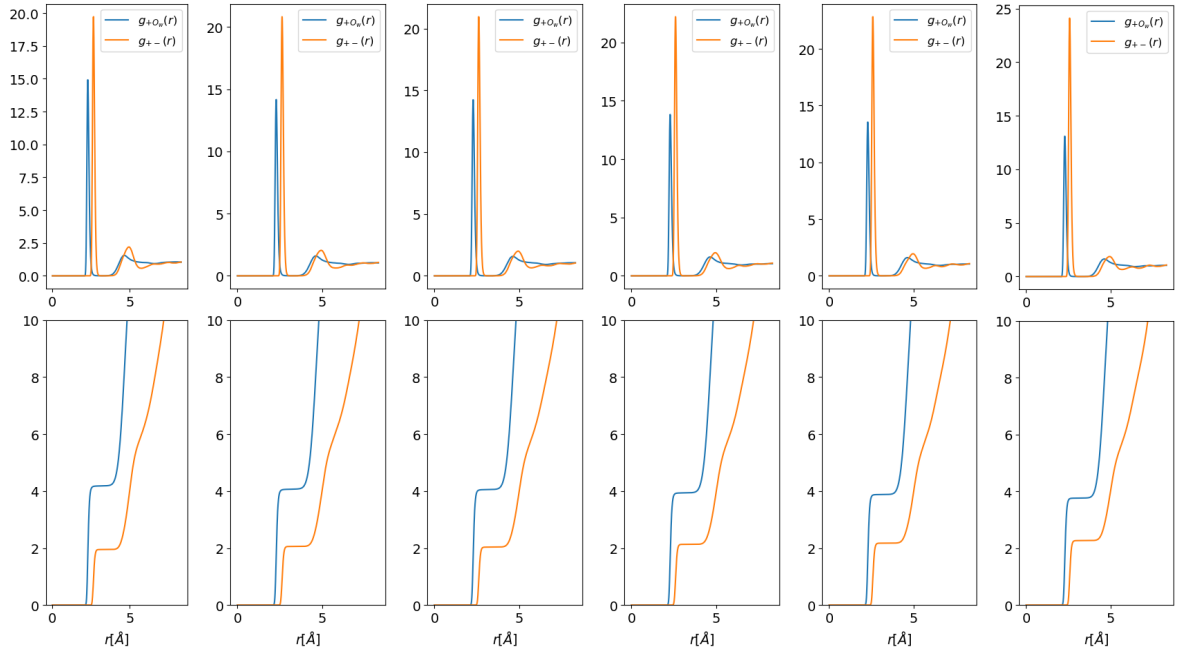


Figure A.11: Cation-anion RDF along the combined isoline of Ca^{2+}.

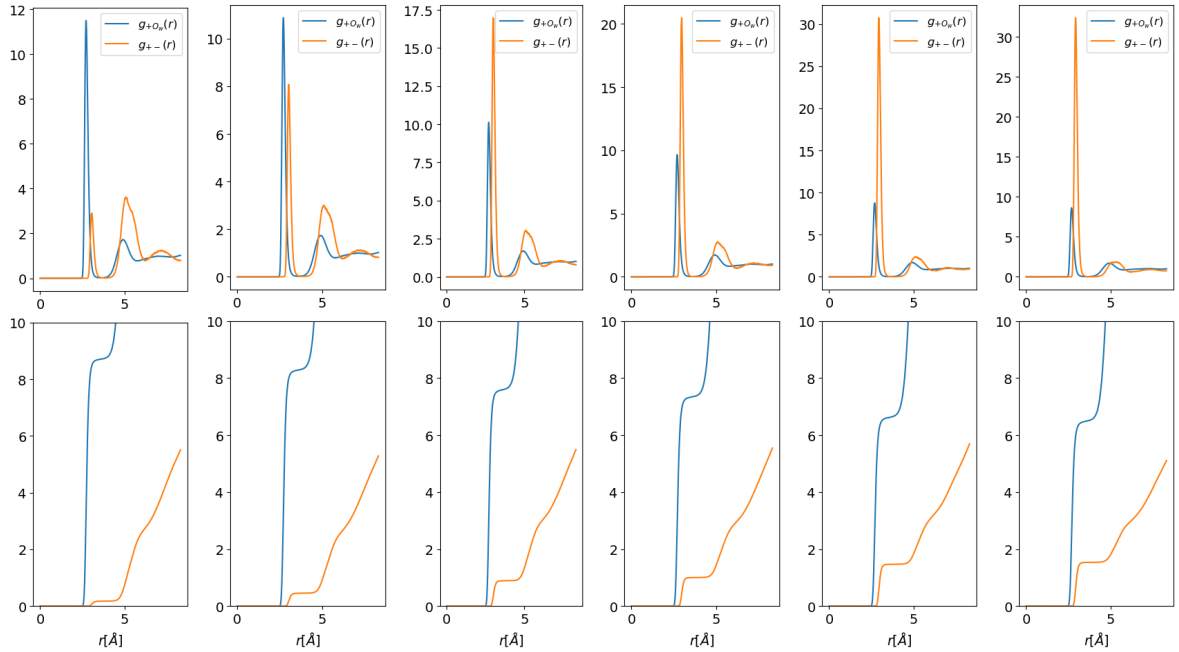


Figure A.12: Cation-anion RDF along the combined isoline of Ba^{2+}.

A.3. The search history

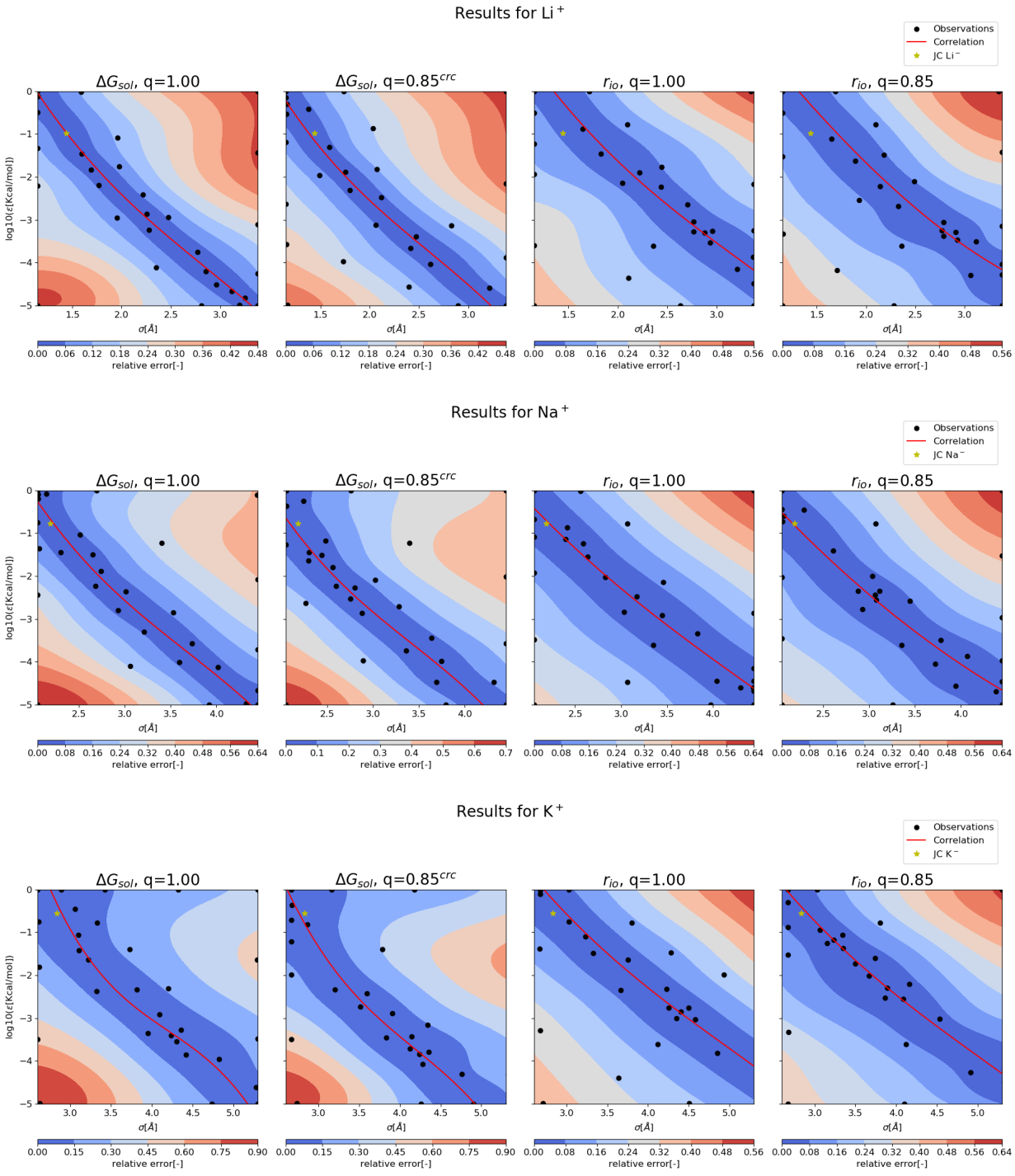


Figure A.13: The SFE and IOD isolines found under different charge conditions for Li, Na, and K. Experimental values from Marcus[49].

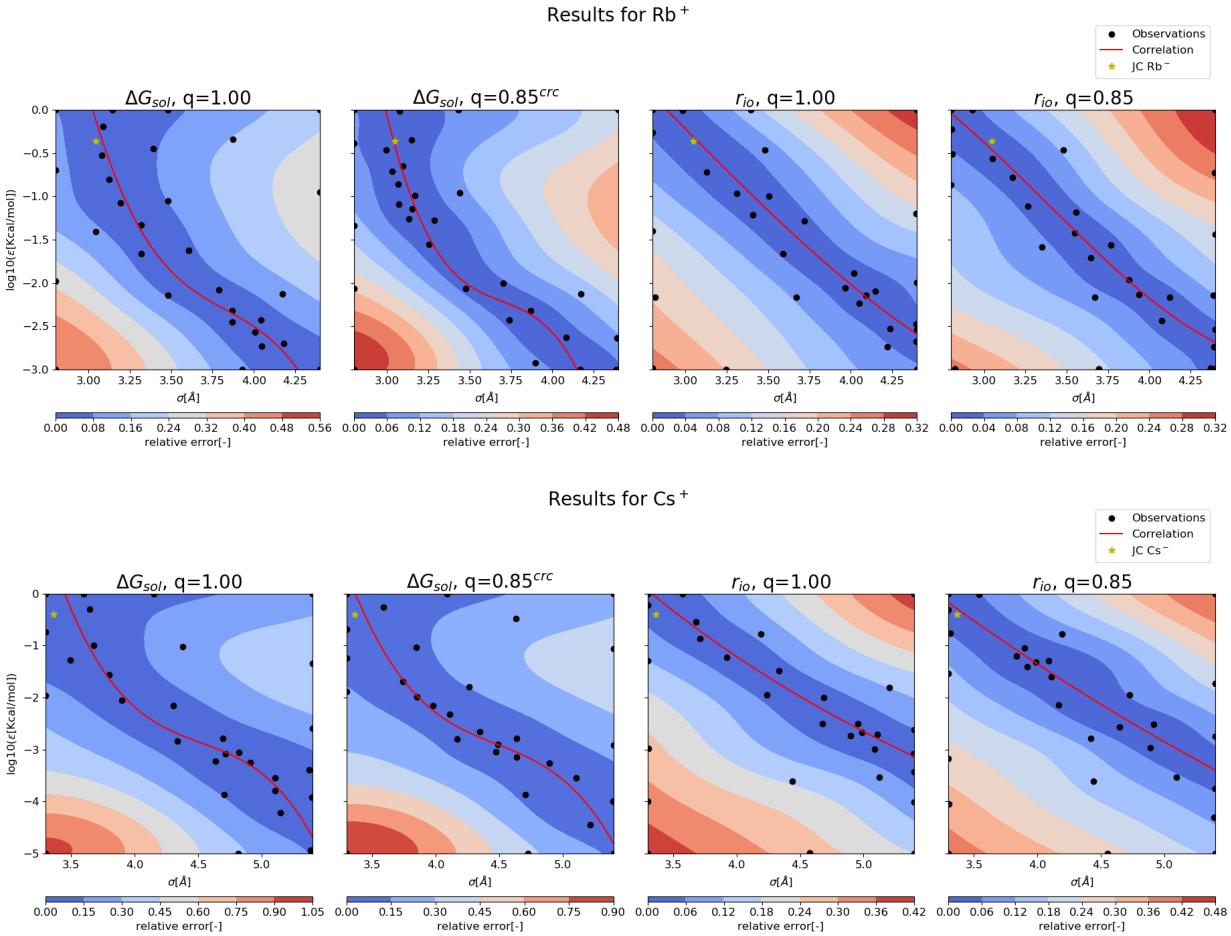


Figure A.14: The SFE and IOD isolines found under different charge conditions for Rb and Cs. Experimental values from Marcus[49].

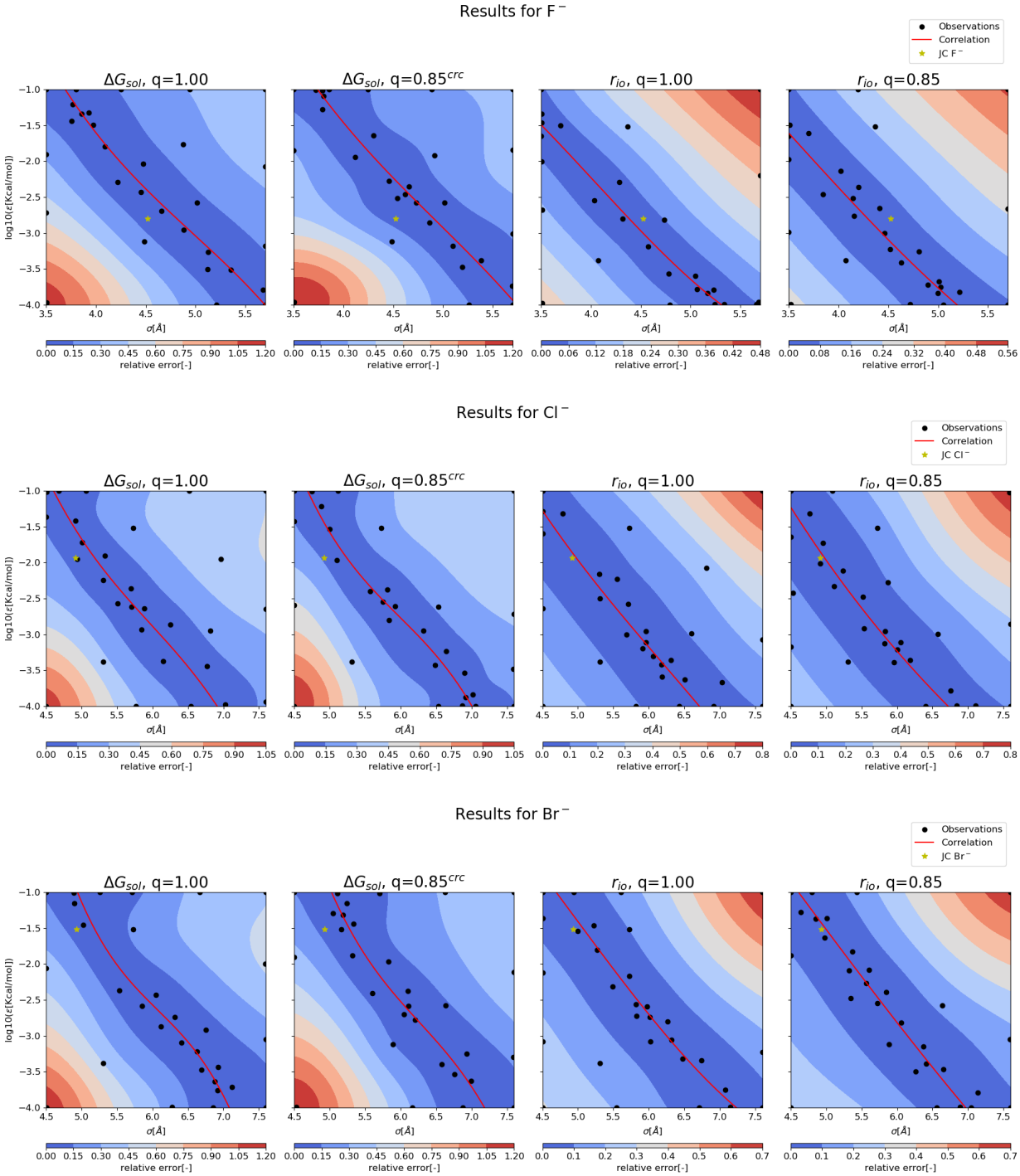


Figure A.15: Search results of anions with TIP4P/2005 water. The $q = 0.85^{\text{crc}}$ stands for applying correction of scaled charge proposed by Döpke et al[18]. Experimental values from Marcus[49].

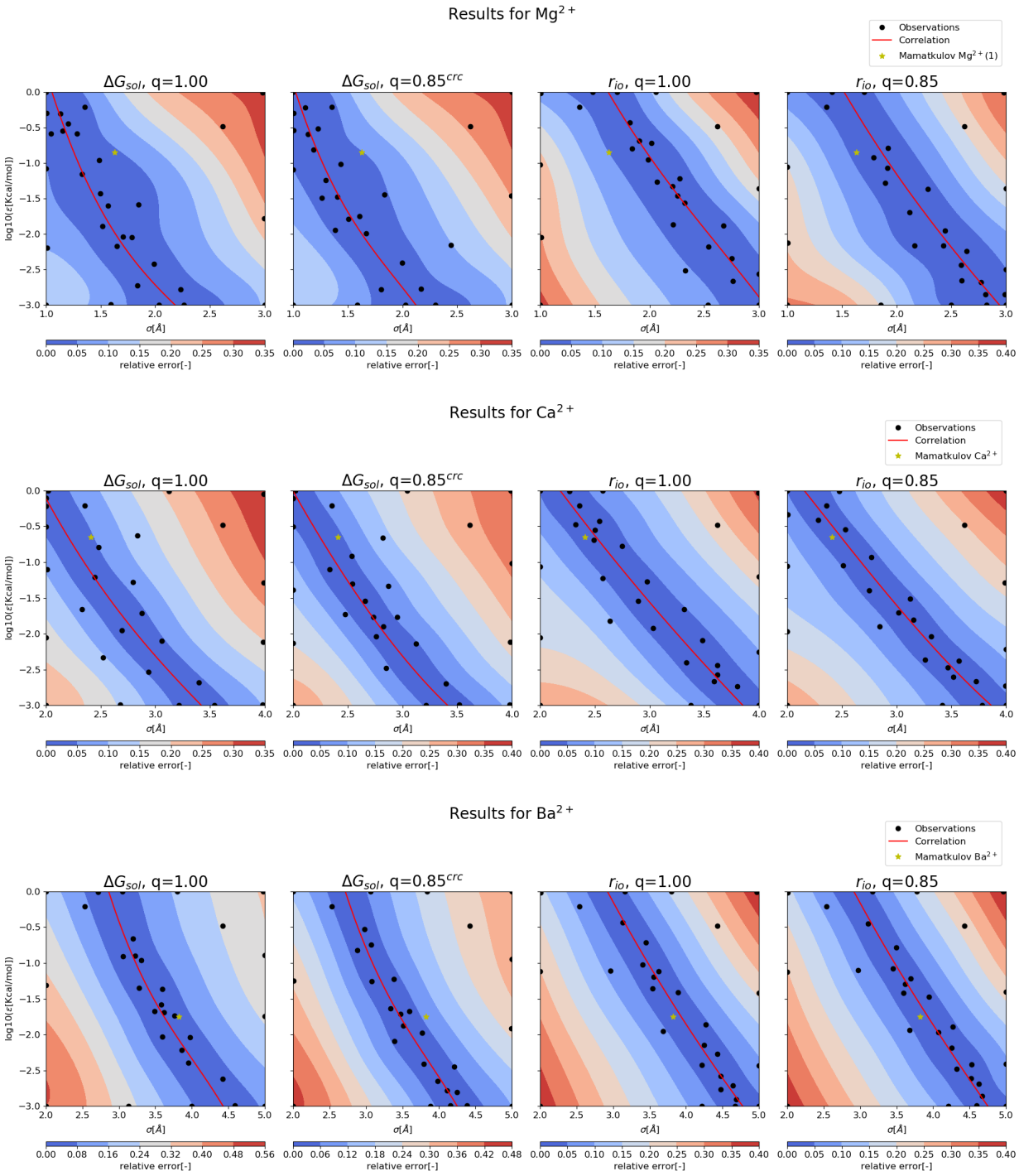


Figure A.16: The SFE and IOD isolines found under different charge conditions for dications. Experimental values from Marcus[49].

B

Additional Tables

B.1. Results of test parameter sets compared with different sources

Table B.1: Compare the cation results of JCTIP4P-Ew parameter set from different sources. The percentage value inside round brackets are the deviation from the Marcus experimental values [49]. The force field combination is TIP4P/2005&JCTIP4P-Ew for our predictions and Ref [18], TIP4P-Ew&JCTIP4P-Ew for Ref [36]. Units for ΔG, r_{io} are [Kcal/mol], [Å].

	Prediction of GP		Dopke[18]		JC[36]		Marcus[49]	
	SFE	IOD	SFE	IOD	SFE	IOD	SFE	IOD
Li	114.6 (+1.0%)	1.95 (-6.2%)	112.6 (-0.8%)	1.92 (-7.7%)	113.7 (+0.2%)	1.92 (-7.7%)	113.5 (0%)	2.08 (0%)
Na	87.9 (+0.8%)	2.35 (-0.4%)	88.4 (+1.4%)	2.35 (-0.4%)	89.0 (+2.1%)	2.35 (-0.4%)	87.2 (0%)	2.36 (0%)
K	72.4 (+2.7%)	2.72 (-2.8%)	70.4 (0%)	2.72 (-2.9%)	70.7 (+0.3%)	2.72 (-2.9%)	70.5 (0%)	2.80 (0%)
Rb	67.1 (+2.1%)	2.88 (-0.3%)	65.6 (0%)	2.87 (-0.7%)	65.6 (-0.2%)	2.87 (-0.7%)	65.7 (0%)	2.89 (0%)
Cs	62.3 (+4.1%)	3.05 (-2.9%)	60.2 (+0.6%)	3.04 (-3.2%)	60.1 (+0.5%)	3.04 (-3.2%)	59.8 (0%)	3.14 (0%)

Table B.2: Compare the anion results of JCTIP4P-Ew parameter set from different sources. The percentage value inside round brackets are the deviation from the Marcus experimental value [49]. The force field combination is TIP4P/2005&JCTIP4P-Ew for our simulation and Ref. [18], TIP4P-Ew&JCTIP4P-Ew for Ref. [36]. Values in the round brackets are the percentage deviation from Marcus values. Units for SFE, and IOD are [Kcal/mol], and [Å].

	Prediction of GP		Döpke[18]		JC[36]		Marcus[49]	
	SFE	IOD	SFE	IOD	SFE	IOD	SFE	IOD
F	121.9 (+9.7%)	2.69 (+2.3%)	121.2 (+9%)	2.69 (+2.3%)	119.8 (+8.0%)	2.69 (+2.3%)	111.1 (+0%)	2.63 (+0%)
Cl	88.9 (+9.3%)	3.16 (-1.0%)	90.0 (+10.7%)	3.16 (-1.0%)	89.2 (+9.7%)	3.16 (-1.0%)	81.3 (+0%)	3.19 (+0%)
Br	82.3 (+9.3%)	3.29 (-2.4%)	83.8 (+11.2%)	3.30 (-2.0%)	82.8 (+10.0%)	3.31 (-2.0%)	75.3 (+0%)	3.37 (+0%)

Table B.3: Compare the anion results of Mamatkulov parameter set from different sources. The percentage value inside round brackets are the deviation from the Marcus experimental value [49]. The force field combination is TIP4P/2005&Mamatkulov for our simulation and Ref. [18], SPC/E&Mamatkulov for Ref. [48]. Values in the round brackets are the percentage deviation from Marcus values. Units for SFE, and IOD are [Kcal/mol], and [Å].

	Prediction of GP		Döpke[18]		Mamatkulov[48]		Marcus[49]	
	SFE	IOD	SFE	IOD	SFE	IOD	SFE	IOD
Mg	420.0 (-4.4%)	1.99 (-4.8%)	415.0 (-5.5%)	1.99 (-4.8%)	439.1 (0%)	1.96 (-6.2%)	439.3 (0%)	2.09 (0%)
Ca	346.1 (-4.4%)	2.35 (-2.5%)	344.0 (-5.0%)	2.36 (-2.1%)	362.1 (0%)	2.38 (-1.2%)	362.1 (0%)	2.41 (0%)
Ba	287.6 (-4.4%)	2.71 (-1.5%)	289.4 (-3.8%)	2.70 (-1.8%)	300.7 (0%)	2.69 (-2.2%)	300.7 (0%)	2.75 (0%)

B.2. Obtained CIP along isoline

Table B.4: Calculated CIP near solubility limit. Total 6 samples are made along the combined isoline evenly according to σ parameter.

	0	1	2	3	4	5	Extent(σ)
Na	0.11	0.31	0.48	0.94	1.41	1.91	[2.1, 2.6]
K	0.12	0.30	0.45	0.66	0.98	1.25	[2.6, 3.1]
Rb	0.40	0.65	0.93	1.29	1.64	1.94	[2.8, 3.3]
Cs	1.63	1.98	2.13	2.37	2.71	3.27	[3.3, 3.8]
Mg	1.83	1.80	1.80	1.78	1.77	1.77	[1.3, 1.8]
Ca	1.95	2.05	2.04	2.13	2.17	2.27	[2.3, 2.8]
Ba	0.17	0.45	0.89	1.00	1.47	1.53	[3.0, 3.5]

B.3. Regression coefficients for different properties with Equation 6.1

Table B.5: The regression coefficients

		c_3	c_2	c_1	c_0
Li	SFE	-0.240207	1.922921	-6.999021	5.797023
	IOD	-0.078759	0.800124	-4.395506	4.562202
	Comb.	-0.159483	1.361522	-5.697264	5.179613
Na	SFE	-0.202580	2.247030	-9.952319	12.539170
	IOD	-0.044321	0.599527	-4.176320	6.032138
	Comb.	-0.123450	1.423279	-7.064319	9.285654
K	SFE	-0.484852	6.084543	-26.688975	37.397897
	IOD	-0.028188	0.497136	-4.158313	8.185990
	Comb.	-0.256520	3.290840	-15.423644	22.791943
Rb	SFE	-1.295657	15.349315	-61.920486	82.453647
	IOD	-0.115113	1.422862	-7.446152	12.418612
	Comb.	-0.705385	8.386089	-34.683319	47.436130
Cs	SFE	-1.344221	18.261828	-83.634556	126.198375
	IOD	0.006775	0.128691	-3.023407	8.388813
	Comb.	-0.668722	9.195260	-43.328981	67.293594
F	SFE	-0.148382	2.231751	-12.485548	22.140828
	IOD	0.022549	-0.200914	-0.964690	3.383862
	Comb.	-0.062916	1.015418	-6.725119	12.762345
Cl	SFE	-0.191372	3.409747	-21.320186	43.613033
	IOD	-0.014158	0.300453	-3.262071	8.629494
	Comb.	-0.102765	1.855100	-12.291128	26.121263
Br	SFE	-0.249607	4.650339	-29.916454	63.421296
	IOD	0.004773	0.018573	-1.939185	7.238161
	Comb.	-0.122416	2.334456	-15.927819	35.329729
Mg	SFE	-0.747950	4.531075	-11.120492	7.364099
	IOD	-0.126790	1.098908	-5.054092	5.808132
	Comb.	-0.437370	2.814992	-8.087292	6.586116
Ca	SFE	-0.281717	2.741611	-10.563583	12.311047
	IOD	-0.008574	0.260896	-3.129491	5.696672
	Comb.	-0.145146	1.501253	-6.846537	9.003860
Ba	SFE	-0.438271	5.266515	-22.575666	31.748379
	IOD	-0.015047	0.310779	-3.321552	7.426878
	Comb.	-0.226659	2.788647	-12.948609	19.587629

C

Mathematical Background

C.1. PDF and CDF

The probability density function specify the probability of random variable fall into particle range of values, the probability of that range can be obtained by integrating over PDF for that range. Suppose a random variable X follows normal distributions with μ as mean and σ^2 as the variance, $X \sim \mathcal{N}(\mu, \sigma^2)$, it's PDF has an explicit expression of probability density function as:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{C.1}$$

Cumulative distribution function of variable X at x is defined as the probability that X take value less than x. Conventionally it is written as $\Phi(x)$. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, then the CDF can be obtained as:

$$\Phi(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \tag{C.2}$$

C.2. Probability improvement

One famous acquisition is the *probability of improvement* or PI[11, 41]. The utility function is defined as:

$$u(x) = \begin{cases} 1 & f(x) \geq f(x^+) \\ 0 & f(x) \leq f(x^+) \end{cases} \tag{C.3}$$

The utility is 1 if the probed function value is larger than the best function value than have been found. Otherwise the uility is 0. Then the acquisition function is the mathematical expectation of the utility given the observations :

$$\begin{aligned} a(x) = \mathbb{E}[u(x) \mid \mathcal{D}_{1:t}] &= \int_{f(x^+)}^{\infty} \mathcal{N}(\mu(x), \sigma(x)) df(x) \\ &= \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) \end{aligned} \tag{C.4}$$

where $x^+ = \underset{x_i}{\operatorname{argmax}} f(x_i), x_i \in x_{1:t}$, $\Phi()$ is the normal cumulative distribution function, $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the normal distribution at the test points. The interpretation of this function is to model the probability of getting better results than the previous observations. This is why cumulative distribution function is used in Equation C.4. A better understanding can be found in Figure 4.5, where the algorithm sampled three points in the parameter space. The maximum of the observation found at this step is located at x^+. Therefore, the probability of improvement is found by integrating the normal cumulative distribution function from the maximum function value $f(x^+)$, which can be represented by the green shaded area in Figure 4.5.

C.3. Matern class of kernel functions

Some researchers have suggested that the infinitely differentiable property of squared exponential (SE) kernel makes the results sometimes unreasonable[72]. Thus, a more realistic Matern class is proposed for engineering applications. The Matern kernels class is a generalization of SE kernel. It has the from like:

$$k\left(x_i, x_j\right) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l} d\left(x_i, x_j\right)\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{l} d\left(x_i, x_j\right)\right) \tag{C.5}$$

where $d()$ is the Euclidean distance, $K_{\nu}()$ is the modified Bessel function, $\Gamma(\nu)$ is the Gamma function. If $\nu \rightarrow \infty$ then this expression becomes the normal SE kernel. The most notable intermediate values are $\nu = 3/2$ (MA3 kernel) and $\nu = 5/2$ (MA5 kernel).

C.4. R2 score

In statistics, the Coefficient of Determination or the R2 score can evaluate level of regression accuracy. If a data set $y_i = [y_1, y_2, , , , , y_n]$ and the corresponding given references (fitted model) are $f_i = [f_1, f_2, , , , , f_n]$. Then the R2 score is defined as:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \\ \text{with} \quad SS_{\text{res}} &= \sum_i \left(y_i - f_i\right)^2, SS_{\text{tot}} = \sum_i \left(y_i - \bar{y}\right)^2 \end{aligned} \tag{C.6}$$

SS_{res} is the sum squared total errors of data and their means. SS_{tot} is the sum squared regression errors of data and their corresponding fitted values. The R2 score measures how good is a regression compared to a simple horizontal line. If the R2 score is 1, the regression matches the data perfectly. If the R2 score is 0, the regression is the same as using a horizontal line to interpret data. If the R2 score is less than 0, the regression is worse than a horizontal line.

D

Physics Derivation

D.1. Electronic Charge Correction

Since 2009, Leontyev published a series of papers about the ECC, dedicating to account the electronic polarizability effect with a non-polarizable model[42–44]. The key argument is that scaling the charge of ions can include the electronic continuum contribution.

The dielectric constants of water in MD simulation is less than that of real measurement. Specifically, the dielectric constant of water at 298 K and 1 atm is $\epsilon_0 = 78.4$, while the dielectric constant of TIP4P/2005 water at same temperature and pressure is $\epsilon_{sim} = 60$[14]. This gives $\epsilon_{el} = \epsilon_0 / \epsilon_{sim} = 1.31$. Therefore, the electronic continuum contribution could be included by scaling the effective charge:

$$q_{eff} = q / \sqrt{\epsilon_{el}} \tag{D.1}$$

where ϵ_{el} is the high frequency dielectric constant of water. After the scaling charge, the Coulombic energy for two charged sites at distance r can be calculated by

$$q_{eff}^2 \Big/ \epsilon_{sim} r = q^2 \Big/ \epsilon_{sim} \epsilon_{el} r = q^2 \Big/ \epsilon_0 r \tag{D.2}$$

, this form automatically matches the Coulombic energy with dielectric constant of real water molecules. The scaling factor is therefore $1/\sqrt{\epsilon_{el}} = 0.87$. Some sources reported $\epsilon_{sim} = 57$[37], thus the scaling factor becomes $1/\sqrt{\epsilon_{el}} = 0.85$.

D.2. Equivalence between MSD and VACF Approach

The Velocity Auto-correlation Function Approach

The Green-Kubo expression gives the the diffusion coefficient as the infinite time integral of an equilibrium correlation function of form:

$$D = \frac{1}{3} \int_0^\infty \langle v(t) \cdot v(0) \rangle \mathrm{d}t \tag{D.3}$$

Green and Kubo[26, 40] initially formalized this expression. It reveals the relation between the diffusion coefficient and the velocity autocorrelation function (VACF). The VACF essentially measures how strong one property of an atom is correlated to itself over time. VACF tells us the temporal pattern by taking advantage of the time series sample. Same as the MSD approach, VACF only applies to the equilibrium molecular dynamics. By integrating the VACF from 0 to infinity, one can find the diffusivity.

However, the Green-Kubo relation is rarely used. The formation of the integral of VACF is a bad idea to be carried out in computer simulations. Unlike in the MSD approach, in which we can store dispersed displacement data directly, treatment of the integral term is usually accompanied with loss of accuracy. Another issue with the VACF approach is that the tail part is very slowly converging towards zero despite the simulation length[31], which makes it challenging to identify the integral's upper bound.

Equivalence between Two Approaches

The Einstein relation and Green-Kubo equation are in equivalence. More specifically, the formation of diffusion coefficient from MSD approach is:

$$D(t) \equiv \frac{D_x(t) + D_y(t) + D_z(t)}{3} = \frac{1}{6}\frac{d}{dt}\left[\Delta x^2(t) + \Delta y^2(t) + \Delta z^2(t)\right] \tag{D.4}$$

$$D(t) \equiv \frac{D_x(t) + D_y(t) + D_z(t)}{3} = \frac{1}{3}\int_0^t \left[C_{xx}(\tau) + C_{yy}(\tau) + C_{zz}(\tau)\right] d\tau \tag{D.5}$$

Here, $D(t)$ represents the diffusion coefficient at time t, D_x represents the diffusion at x direction, $\Delta x(t)$ represents the MSD at x direction at time instance t, $C_{xx}(\tau)$ is the VACF at x direction. To proof two approaches are in equivalence:

$$D \equiv \frac{1}{6}\frac{d}{dt}\left[\Delta x^2(t) + \Delta y^2(t) + \Delta z^2(t)\right] = \frac{1}{3}\int_0^t \left[C_{xx}(\tau) + C_{yy}(\tau) + C_{zz}(\tau)\right] d\tau \tag{D.6}$$

$$\begin{aligned} D_x(t) &= \int_0^t C_{xx}(\tau) d\tau \\ D_y(t) &= \int_0^t C_{yy}(\tau) d\tau \\ D_z(t) &= \int_0^t C_{zz}(\tau) d\tau \end{aligned} \tag{D.7}$$

The displacement between two time instance with time origins at $t_0 = 0$ for a single particle in one direction e.g. x is:

$$x_i(t) - x_i(0) = \int_0^t u_i(\tau_1) d\tau_1 \tag{D.8}$$

Mean square displacement can be formulated as:

$$[x_i(t) - x_i(0)]^2 = \int_0^t u_i(\tau_1) d\tau_1 \int_0^t u_i(\tau_2) d\tau_2 \tag{D.9}$$

Subscript i represents for ith particle in the system, u represents the velocity projection at x direction. Here t is the variation and $t_0 = 0$ is the time origin and is regarded as fixed. Mean square displacement at x direction can be formulated as:

$$\Delta x^2(t) = \left\langle [x_i(t) - x_i(0)]^2 \right\rangle \tag{D.10}$$

The angle bracket here represents the ensemble average, meaning it averages over all different particles i and all different time origins. The ensemble average and the time integral are exchangeable in calculations, thus Equations D.9 and D.10 yield:

$$\Delta x^2(t) = \int_0^t d\tau_1 \int_0^t d\tau_2 \langle u_i(\tau_1) u_i(\tau_2) \rangle \tag{D.11}$$

Thus,

$$\begin{aligned} D_x(t) &= \frac{1}{2}\frac{d}{dt}\Delta x^2(t) \\ &= \int_0^t d\tau_1 \langle u_i(\tau_1) u_i(t) \rangle \\ &= \int_0^t d\tau_1 \langle u_i(\tau_1 - t) u_i(0) \rangle \\ &= \int_{-t}^0 \langle u_i(\tau_3) u_i(0) \rangle d\tau_3 \\ &= -\int_t^0 \langle u_i(\tau_4) u_i(0) \rangle d\tau_4 \\ &= \int_0^t \langle u_i(\tau_4) u_i(0) \rangle d\tau_4 \end{aligned} \tag{D.12}$$

The third equality in Equation D.12 use the fact that VACF only measures the time difference between two velocities, thus $\langle u_i(\tau_1) u_i(t) \rangle = \langle u_i(\tau_1 - t) u_i(0) \rangle$. The fourth equality in Equation D.12 used integration by substitution $\tau_3 = \tau_1 - t$. The fifth equality used another integration by substitution $\tau_4 = -\tau_3$. Following the same procedure, one can proof:

$$D_y(t) = \int_0^t \langle v_i(\tau) v_i(0) \rangle d\tau \tag{D.13}$$

$$D_z(t) = \int_0^t \langle w_i(\tau) w_i(0) \rangle d\tau \tag{D.14}$$

Hence, two formations of MSD approach and VACF approach are in equivalence.

E

L-J Parameterizations and Experimental Values

E.1. Common L-J parameter sets in literature

Table E.1: L-J parameter sets

		TIP4P/2005&Madrid	TIP4P-Ew&JC	Mamatkulov
Li^+	σ[Å]	1.43970	1.43969	-
	ϵ[Kcal/mol]	0.10398	0.10398	-
Na^+	σ[Å]	2.21737	2.18448	-
	ϵ[Kcal/mol]	0.35190	0.16843	-
K^+	σ[Å]	2.30140	2.83305	-
	ϵ[Kcal/mol]	0.47460	0.27946	-
Rb^+	σ[Å]	-	3.04509	-
	ϵ[Kcal/mol]	-	0.43314	-
Cs^+	σ[Å]	-	3.36403	-
	ϵ[Kcal/mol]	-	0.39443	-
F^-	σ[Å]	-	4.52220	-
	ϵ[Kcal/mol]	-	0.00157	-
Cl^-	σ[Å]	4.69906	4.91776	-
	ϵ[Kcal/mol]	0.01838	0.01166	-
Br^-	σ[Å]	-	4.93201	-
	ϵ[Kcal/mol]	-	0.03037	-
I^-	σ[Å]	-	5.25986	-
	ϵ[Kcal/mol]	-	0.04170	-
Mg^{2+}	σ[Å]	1.1629	-	1.63(1), 2.63(2)
	ϵ[Kcal/mol]	0.87282	-	0.141013(1), 0.000956(2)
Ca^{2+}	σ[Å]	2.6656	-	2.41
	ϵ[Kcal/mol]	0.12122	-	0.224665
Ba^{2+}	σ[Å]	-	-	3.820
	ϵ[Kcal/mol]	-	-	0.017686

Table E.2: Experimental values, units for ΔG, r_{io}, and, D_i are [Kcal/mol], [Å], and [1e-9m^2/2] respectively.

	Marcus			Schmid		
	ΔG	r_{io}	D_i	ΔG	r_{io}	D_i
Li^+	113.5	2.08	1.03	113.8	-	-
Na^+	87.2	2.36	1.33	88.7	-	-
K^+	70.5	2.80	1.96	71.2	-	-
Rb^+	65.7	2.89	2.07	66.0	-	-
Cs^+	59.8	3.14	2.06	60.5	-	-
F^-	111.1	2.63	1.48	119.7	-	-
Cl^-	81.3	3.19	2.03	89.1	-	-
Br^-	75.3	3.37	2.08	82.7	-	-
I^-	65.7	3.65	2.04	74.3	-	-
Mg^{2+}	439.3	2.09	0.71	-	-	-
Ca^{2+}	362.1	2.41	0.79	-	-	-
Ba^{2+}	300.7	2.75	0.85	-	-	-

Bibliography

[1] Jose LF Abascal and Carlos Vega. A general purpose model for the condensed phases of water: Tip4p/2005. *The Journal of chemical physics*, 123(23):234505, 2005.

[2] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[3] Johan Aqvist. Ion-water interaction potentials derived from free energy perturbation simulations. *The Journal of Physical Chemistry*, 94(21):8021–8024, 1990.

[4] Distill article. "a visual exploration of gaussian processes". *URL https://distill.pub/2019/visual-exploration-gaussian-processes#Multivariate*, 2019.

[5] Pascal Auffinger, Thomas E Cheatham, and Andrea C Vaiana. Spontaneous formation of kcl aggregates in biomolecular simulations: a force field issue? *Journal of chemical theory and computation*, 3(5):1851–1859, 2007.

[6] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.

[7] AL Benavides, MA Portillo, VC Chamorro, JR Espinosa, JLF Abascal, and C Vega. A potential model for sodium chloride solutions based on the tip4p/2005 water model. *The Journal of chemical physics*, 147(10):104501, 2017.

[8] HJC Berendsen, JR Grigera, and TP Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[9] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[10] Thomas C Beutler, Alan E Mark, René C van Schaik, Paul R Gerber, and Wilfred F Van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical physics letters*, 222(6):529–539, 1994.

[11] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[12] Bernard R Brooks, Robert E Bruccoleri, Barry D Olafson, David J States, S a Swaminathan, and Martin Karplus. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry*, 4(2):187–217, 1983.

[13] David A Case, Thomas E Cheatham III, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz Jr, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[14] Martin Chaplin. 2020. http://www1.lsbu.ac.uk/water/water_models.html.

[15] Liem X Dang and Bruce C Garrett. Photoelectron spectra of the hydrated iodine anion from molecular dynamics simulations. *The Journal of chemical physics*, 99(4):2972–2977, 1993.

[16] Telma Woerle de Lima, Antonio Caliri, Fernando Luís Barroso da Silva, Renato Tinós, Gonzalo Travieso, Ivan Nunes da Silva, Paulo Sergio Lopes, Eduardo Marques de Souza, Alexandre Cláudio Botazzo Delbem, Vanderlei Bonatto, et al. Some modeling issues for protein structure prediction using evolutionary algorithms. *Evolutionary Computation*, page 153, 2009.

[17] Alain Dequidt and Jose G Solano Canchaya. Bayesian parametrization of coarse-grain dissipative dynamics models. *The Journal of chemical physics*, 143(8):084122, 2015.

[18] Max F Döpke, Othonas A Moultos, and Remco Hartkamp. On the transferability of ion parameters to the tip4p/2005 water model using molecular dynamics simulations. *The Journal of Chemical Physics*, 152(2):024501, 2020.

[19] David Dubbeldam, Denise C Ford, Donald E Ellis, and Randall Q Snurr. A new perspective on the order-n algorithm for computing correlation functions. *Molecular Simulation*, 35(12-13):1084–1097, 2009.

[20] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[21] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

[22] Jun Fujiki, Shotaro Akaho, Hideitsu Hino, and Noboru Murata. Robust hypersurface fitting based on random sampling approximations. In *International Conference on Neural Information Processing*, pages 520–527. Springer, 2012.

[23] Maria Fyta, Immanuel Kalcher, Joachim Dzubiella, Luboš Vrbka, and Roland R Netz. Ionic force field optimization based on single-ion and ion-pair solvation properties. *The Journal of chemical physics*, 132(2):024911, 2010.

[24] MA González. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique*, 12:169–200, 2011.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[26] Melville S Green. Markoff random processes and the statistical mechanics of time-dependent phenomena. ii. irreversible processes in fluids. *The Journal of Chemical Physics*, 22(3):398–413, 1954.

[27] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O Anatole Von Lilienfeld, Klaus-Robert Muller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 6(12):2326–2331, 2015.

[28] Dominik Horinek, Shavkat I Mamatkulov, and Roland R Netz. Rational design of ion force fields based on thermodynamic solvation properties. *The Journal of chemical physics*, 130(12):124507, 2009.

[29] Hans W Horn, William C Swope, Jed W Pitera, Jeffry D Madura, Thomas J Dick, Greg L Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of chemical physics*, 120(20): 9665–9678, 2004.

[30] Philip Ilten, Mike Williams, and Yunjie Yang. Event generator tuning using bayesian optimization. *Journal of Instrumentation*, 12(04):P04028, 2017.

[31] Seyed Hossein Jamali, Ludger Wolff, Tim M Becker, Mariette de Groen, Mahinder Ramdin, Remco Hartkamp, Andre Bardow, Thijs JH Vlugt, and Othonas A Moultos. Octp: A tool for on-the-fly calculation of transport properties of fluids with the order-n algorithm in lammps. *Journal of chemical information and modeling*, 59(4):1290–1294, 2019.

[32] Kasper P Jensen and William L Jorgensen. Halide, ammonium, and alkali metal ion parameters for modeling aqueous solutions. *Journal of Chemical Theory and Computation*, 2(6):1499–1509, 2006.

[33] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[34] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[35] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[36] In Suk Joung and Thomas E Cheatham III. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The journal of physical chemistry B*, 112(30):9020–9041, 2008.

[37] ZR Kann and JL Skinner. A scaled-ionic-charge simulation model that reproduces enhanced and suppressed water diffusion in aqueous salt solutions. *The Journal of chemical physics*, 141(10):104507, 2014.

[38] Christoph Kirse, Moritz Kindlein, Frederik Luxenburger, Ekaterina Elts, and Heiko Briesen. Analysis of two common algorithms to compute self-diffusion coefficients in infinite dilution from molecular dynamics simulations and application to n-alkanes (c1 to c35) in water. *Fluid Phase Equilibria*, 485:211–219, 2019.

[39] Pavel V Klimovich, Michael R Shirts, and David L Mobley. Guidelines for the analysis of free energy calculations. *Journal of computer-aided molecular design*, 29(5):397–411, 2015.

[40] Ryogo Kubo. Statistical-mechanical theory of irreversible processes. i. general theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6):570–586, 1957.

[41] Harold J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. 1964.

[42] Igor Leontyev and Alexei Stuchebrukhov. Accounting for electronic polarization in non-polarizable force fields. *Physical Chemistry Chemical Physics*, 13(7):2613–2626, 2011.

[43] IV Leontyev and AA Stuchebrukhov. Electronic continuum model for molecular dynamics simulations. *The Journal of chemical physics*, 130(8):02B609, 2009.

[44] IV Leontyev and AA Stuchebrukhov. Electronic polarizability and the effective pair potentials of water. *Journal of chemical theory and computation*, 6(10):3153–3161, 2010.

[45] Pengfei Li, Benjamin P Roberts, Dhruva K Chakravorty, and Kenneth M Merz Jr. Rational design of particle mesh ewald compatible lennard-jones parameters for+ 2 metal cations in explicit solvent. *Journal of chemical theory and computation*, 9(6):2733–2748, 2013.

[46] Pengfei Li, Lin Frank Song, and Kenneth M Merz Jr. Parameterization of highly charged metal ions using the 12-6-4 lj-type nonbonded model in explicit water. *The Journal of Physical Chemistry B*, 119(3):883–895, 2015.

[47] Pu Liu, Qiang Shi, Hal Daumé III, and Gregory A Voth. A bayesian statistics approach to multiscale coarse graining. *The Journal of chemical physics*, 129(21):12B605, 2008.

[48] Shavkat Mamatkulov, Maria Fyta, and Roland R Netz. Force fields for divalent cations based on single-ion and ion-pair properties. *The Journal of chemical physics*, 138(2): 024505, 2013.

[49] Yitzhak Marcus. *Ion properties*. CRC Press, 1997.

[50] Ruben Martinez-Cantin, Nando De Freitas, Eric Brochu, José Castellanos, and Arnaud Doucet. A bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*, 27(2):93–103, 2009.

[51] James L McDonagh, Arnaldo F Silva, Mark A Vincent, and Paul LA Popelier. Machine learning of dynamic electron correlation energies from topological atoms. *Journal of chemical theory and computation*, 14(1):216–224, 2017.

[52] James L McDonagh, Ardita Shkurti, David J Bray, Richard L Anderson, and Edward O Pyzer-Knapp. Utilizing machine learning for efficient parameterization of coarse grained molecular force fields. *Journal of chemical information and modeling*, 59(10): 4278–4288, 2019.

[53] MDAnalysis. 2020. https://www.mdanalysis.org/.

[54] M Mezei. The finite difference thermodynamic integration, tested on calculating the hydration free energy difference between acetone and dimethylamine in water. *The Journal of chemical physics*, 86(12):7084–7088, 1987.

[55] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[56] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

[57] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. Toward global optimization, volume 2, chapter bayesian methods for seeking the extremum, 1978.

[58] Othonas A Moultos, Ioannis N Tsimpanogiannis, Athanassios Z Panagiotopoulos, and Ioannis G Economou. Atomistic molecular dynamics simulations of co2 diffusivity in h2o for a wide range of temperatures and pressures. *The Journal of Physical Chemistry B*, 118(20):5532–5541, 2014.

[59] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[60] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[61] Andrew Ng. Machine learning yearning. *URL: http://www. mlyearning. org/(96)*, 2017.

[62] F Nogueira. Bayesian optimization: Open source constrained global optimization tool for python. *URL: https://github. com/fmfn/BayesianOptimization*, 2014.

[63] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[64] Maria M Reif and Philippe H Hünenberger. Computation of methodology-independent single-ion solvation properties from molecular simulations. iv. optimized lennard-jones interaction parameter sets for the alkali and halide ions in water. *The Journal of chemical physics*, 134(14):144104, 2011.

[65] Roland Schmid, Arzu M Miah, and Valentin N Sapunov. A new table of the thermodynamic quantities of ionic hydration: values and some applications (enthalpy–entropy compensation and born radii). *Physical Chemistry Chemical Physics*, 2(1): 97–102, 2000.

[66] Fatih G Sen, Badri Narayanan, Jeffrey Larson, Alper Kinaci, Kiran Sasikumar, Michael J Davis, Stefan M Wild, Stephen K Gray, Subramanian KRS Sankaranarayanan, and Maria KY Chan. Comparing optimization strategies for force field parameterization. *arXiv preprint arXiv:1812.00326*, 2018.

[67] LAMMPS Molecular Dynamics Simulator. 2020. http://lammps.sandia.gov.

[68] sklogWiki. 2020. http://www.sklogwiki.org/.

[69] David E Smith and Liem X Dang. Computer simulations of nacl association in polarizable water. *The Journal of Chemical Physics*, 100(5):3757–3766, 1994.

[70] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[71] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[72] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

[73] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[74] Søren Toxvaerd and Jeppe C Dyre. Communication: Shifted forces in molecular dynamics, 2011.

[75] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[76] Carlos Vega and Jose LF Abascal. Simulating water with rigid non-polarizable models: a general perspective. *Physical Chemistry Chemical Physics*, 13(44):19663–19688, 2011.

[77] Han Wang, Linfeng Zhang, Jiequn Han, and E Weinan. Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228:178–184, 2018.

[78] Junmei Wang and Tingjun Hou. Application of molecular dynamics simulations in molecular property prediction ii: diffusion coefficient. *Journal of computational chemistry*, 32(16):3505–3519, 2011.

[79] Takuma Yagasaki, Masakazu Matsumoto, and Hideki Tanaka. Lennard-jones parameters determined to reproduce the solubility of nacl and kcl in spc/e, tip3p, and tip4p/2005 water. *Journal of Chemical Theory and Computation*, 16(4):2460–2473, 2020.

[80] In-Chul Yeh and Gerhard Hummer. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *The Journal of Physical Chemistry B*, 108(40):15873–15879, 2004.

[81] IM Zeron, JLF Abascal, and C Vega. A force field of li+, na+, k+, mg2+, ca2+, cl-, and so 4 2- in aqueous solution based on the tip4p/2005 water model and scaled charges for the ions. *The Journal of chemical physics*, 151(13):134504, 2019.

[82] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deepcg: Constructing coarse-grained models via deep neural networks. *The Journal of chemical physics*, 149(3):034101, 2018.

[83] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and E Weinan. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[84] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and E Weinan. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In *Advances in Neural Information Processing Systems*, pages 4436–4446, 2018.