Bayesian Optimization guided comparison study of the Jeung-Chetham (JC) model on LiX salts

by

Zhihao Deng

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Science

in

THE FACULTY OF SCIENCE
(Chemistry)

The University of British Columbia
(Vancouver)

April 2020

Abstract

Table of Contents

List of Tables

List of Figures

Abbreviations

JC: Jeung-Chetham
LJ: Lennard-Jones
LiX: Lithium halides
GP: Gaussian process
PDF: Probability density function
BO: Bayesian optimization
MD: Molecular dynamics
DFT: Density functional theory
RDF: Radial distribution function

Acknowledgments

Chapter 1

Introduction

The ability to predict the correct crystal structures and its relevant information using theoretical models has always been what researchers pursue. Nevertheless, sometimes the predicted value does not match with the experimental value, so people add correction such as changing parameters or augmenting terms to guide results manually. During the manual correction, we still want to extract physical information instead of ad-hoc non-physical parameters(statistical fitting). Here, this thesis aims to show how we used Bayesian optimization, a machine learning/tuning procedure, to guide the learning procedure and what we have learned about the commonly-used JC[5] empirical model while fitting to the parameters of the LiX salts.

Consider the problem described in the previous paragraph as an inverse problem. We can split such physical/mathematical problem solving process into three cases: direct problems, reconstruction problems, and identification problems. It is convenient to reformulate these three problems using a simple function. If there exists $y = f(x)$, direct problems aim to find y given f and x, reconstruction problems aim to find x given f and y, and identification problems aim to find f given x and y. This study involves direct problems and reconstruction problems, i.e the ability to predict the output and the ability to find the input given output(targets), the latter is known as the inverse problem. Note that the function in this problem adopt the same framework, which will be provided by the JC model.

We studied 4 LiX salts: LiF, LiCl, LiBr and LiI. The ultimate goal is to reproduce the correct experimental crystal structures, lattice parameters and absolute lattice energy. The subtlety lies in the first objective, where correct experimental crystal structure implies correct stability order with respect to the relative energy difference among all possible crystal structures for each LiX.

1.1 JC Model

Structural determination and the relevant physical properties are based on finding the global minimum on the potential energy surface. The approach this study took was the force field method. We

used JC to model the pairwise potential, which has the following form:

$$u_{ij}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r} + 4\varepsilon_{ij}\left[(\frac{\sigma_{ij}}{r})^{12} - (\frac{\sigma_{ij}}{r})^{6}\right],$$

where r is the pairwise distance, ε_{ij} is the well depth and σ_{ij} is the distance at which the intermolecular potential between 2 particles is 0, i.e. a relative measure to the bonding distance. Hence, the first term of JC potential is the exact expression for coulomb potential, and the second term is commonly known as the LJ potential. The r^{12} term in LJ implies the Coulombic and Pauli repulsion and the r^6 term represents the long-ranged dispersion energy. r12 was chosen for historical reason since it simplifies the calculation. For each LiX salt, the potential energy surface is parametrized by 3 sets of σ and ε since there are 3 unique types of interaction: metal-metal, metal-halide, and halide-halide. Therefore, given the same starting configuration, the most stable crystal structure of the salt (global minimum) depends on 6 parameters. For this study, we also introduce a well known approximation method: Lorentz-Berthelot rules[6][2],

$$\begin{cases} \sigma_{MX} = \frac{\sigma_M + \sigma_X}{2} \\ \varepsilon_{MX} = \sqrt{\varepsilon_M \varepsilon_X} \end{cases},$$

where σ_{MX} and ε_{MX} are parameters for the cross correlation interaction. By imposing the combining rules, the JC model is parametrized by 4 instead of 6.

Energy minimization, or equivalently geometry optimization, is designed by Hayden and it is a gradient descent optimization where the algorithm uses the computational package GROMACS to calculate the single point energy. By varying 6 unit cell parameters (a,b,c,α,β,γ), we can get the optimized geometry and all its physical information. In practice, to speed up the calculation, we fixed 3 angles to known crystal structures, e.g. rock salt, and repeat the geometry optimization for each crystal structure.

In the original paper of JC model[5], the parameters were fit to the solvation free energies, radial distribution functions, ion-water interaction energies, lattice energies, and lattice parameters at 300 K. The reference parameter was taken from the extended simple point charge (SPC/E) water model. In this model (JC,SPC/E), it predicts rock salt as the most stable form for LiF and LiCl, but wurtzite as the stable crystal structure for LiBr and LiI. In nature, the actual stable geometry for all 4 salts is rock salt, but the reference JC model could not get such prediction. The following table with data from Hayden's report shows the problem: (E. diff, energy difference, is defined to be the difference between wurtzite and rock salt.)

Salt	Favors	Rock. E.	Wurtz. E.	E. diff
LiF	Rock.	-1052.6	-1030.57	22.03
LiCl	Rock.	-870.32	-863.97	6.35
LiBr	Wurtz.	-826.03	-827.08	-1.05
LiI	Wurtz.	-761.78	-770.42	-8.64

(a) Predicted physical properties using JC SPC/E model. Energy in kJ/mol

Salt	Rock. E.
LiF	-1054
LiCl	-862
LiBr	-821
LiI	-765

(b) Experimental energies of 4 salts. Energy in kJ/mol

Salts	Rock. a	Wurtz. a	Wurtz. c
LiF Exp.	4.0262	**3.1289**	**4.6916**
LiF JC	4.2860	3.3588	5.4391
LiCl Exp.	5.1399	3.8701	6.1121
LiCl JC	5.2027	4.0666	6.3576
LiBr Exp.	5.5016	4.1510	6.6502
LiBr JC	5.5107	4.2545	6.7196
LiI Exp.	6.0127	4.5141	7.3112
LiI JC	6.0043	4.5782	7.2831

(c) Experimental and JC lattice constants of 4 salts. For rock salt, a=b=c. For wurtzite, a=b. The emboldened LiF wurtzite lattice constants are taken from the best DFT calculation(PW1PW-D3(BJ)). Lattice constants are in *Å*

Table 1.1: Physical properties using the reference JC parameters

Therefore, using 4 sets of reference JC parameters, we cannot guarantee that rock salt is more stable than wurtzite at all time. Furthermore, Hayden's results(DFT) show that the relative energy difference between wurtzite and rock salt is always around 15 kJ/mol, whereas the JC results in the table above either overestimate or underestimate the energy difference. The reason relative stability is crucial is due to the nature of nucleation process, in which we would like to predict the nucleated LiX salts has the natural rock salt structure. Then, the problem lies in finding a JC parameter set that can give the correct energy difference, lattice constants of both rock salt and wurtzite structures and absolute lattice energy. Note that the importance of these 3 features are also in the same order of how there were being mentioned.

1.2 Gaussian Process

There are two main types of learning an unknown function this paper will introduce: linear parametric and non-linear non-parametric models. Readers will understand why BO is based on GP, and this will help readers to better understand how BO guide the solution-finding process..

1. Linear parametric regression: A linear regression model is widely used in interpolating experimental results. For example, given a dataset, $\mathscr{D} = \{(x_i, y_i) | i = 1, 2, ..., n\}$ with n observations, since the goal is to predict what y_j is, where $x_{i-1} < x_j < x_i$, the problem lies in finding what the unknown underlying function is. The better it approximates the real unknown function, the less error we are going to get in predicting y_j. In linear parametric model, we start with assuming the unknown function can be broken down to specific classes of function, e.g. polynomials. Hence, by including only upto power k polynomials, the unknown function $h(x)$ is approximated as

$$h(x) \approx c_0 + c_1 x + c_2 x^2 + c_3 x^3 + ... + c_k x^k.$$

Then, the numerical goal is to find a parameter set $\theta = \{c_i | i = 0, 1, 2, ..., k\}$ such that we can minimize some unknown objectives. We can either introduce an artificial cost function or solve the linear system of equations.

Solving a high dimensional linear system of equations can be very computationally expensive because inverting a matrix is a $\mathscr{O}^3$ operation.[3] Hence, it is better to introduce an artificial cost function during the numerical procedure so that we have a proper mathematical treatment. A common cost function, denoted $\mathscr{L}$, is root mean square error (RMSE), where given $\mathscr{D}$ with n data points and parameter set θ that contains $k+1$ parameters, the RMSE is

$$\mathscr{L}(x) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [(\sum_{j=0}^{k} c_j x_i^j) - y_i]^2}.$$

Then, we can use any minimization algorithm, e.g. gradient descent, to get an optimized parameter set.

Though linear parametric regression is easy to implement, it has several serious disadvantages. First, we need to choose numbers of basis functions, e.g. polynomials. Then number of basis functions depends on size of $\mathscr{D}$. Suppose, we choose $|k| < |i|$, there won't be any exact solution if we solve the system of linear equations, and cost function approach will give a rough estimate of the parameter set. Finally, this method tends to over fit the dataset.[4]

2. Non-linear non-parametric regression: A linear parametric regression we discussed above has many restrictions. To overcome these problems, GP is the solution.[11] If we consider linear regression as a bottom up approach, i.e. use the data pool to construct an explanatory

model, GP is the opposite. Here, we start by allowing infinite numbers of functions, then the question becomes which one explains the dataset the best. In other words, after fitting all possible functions to the data, we need to find the one that gives the highest marginal probability. Any Gaussian has the following properties:

(a) Any marginal of a Gaussian distribution is Gaussian

(b) Any conditional of a Gaussian distribution is Gaussian

(c) There is an analytical formula to integrate Gaussian function

A visual representation is given below:

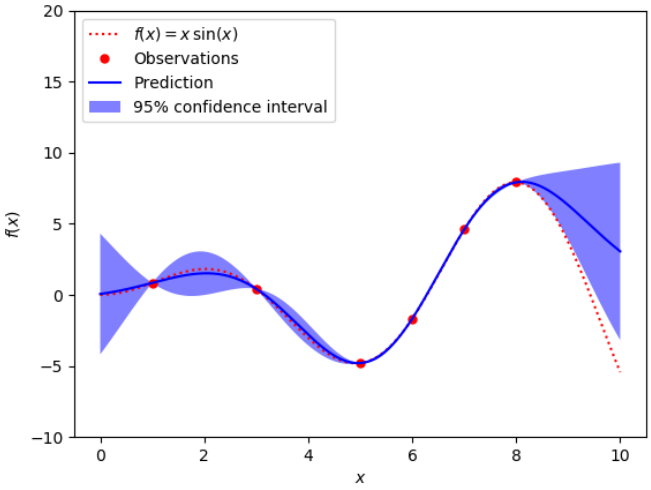

Figure 1.1: A visual representation of GP. [7]

A multivariate Gaussian distribution is defined by two parameters: μ and Σ, where μ is the mean vector and Σ is the covariance matrix. μ gives the expected value of the distribution, and each element gives the mean of the corresponding random variable. Σ models the variance for each random variable and the covariance between different random variables. With properties of multivariate Gaussian distribution, we assume a function is represented as a vector and the unknown vector is drawn from multivariate Gaussian distribution. Specifically, each input along x can be considered as a random variable where the mean represents mean of the function evaluated at the location, and covariance represents correlation among different input locations.

For a univariate normal variable, $x \sim \mathcal{N}(\mu, \sigma^2)$. Similarly, for two normally distributed random variables, $\vec{x} = [x_1, x_2]^T$, we assume $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, where $\mu = [\mu_1, \mu_2]^T$ and $\Sigma =$

$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Σ_{ij} represents variance if $i = j$ or covariance if $i \neq j$.

In general, a bivariate normal distribution includes our prior guess of the underlying function and it allows us to predict unseen points based on $\mathcal{N}(\vec{\mu}, \Sigma)$. We can generalize this to multivariate, and this will be the formulation of GP. Given a black box function h(x) and dataset $\mathscr{D} = (X, y)$ with n training points and their corresponding results, we put a normal distribution on each training point, thus we end up with a joint normal distribution description $p(h(x_1), h(x_2), ..., h(x_n))$. To simplify the notation, we represent the joint distribution as

$$h(x) \sim \mathscr{GP}(m(x), k(x_i, x_j)),$$

where $m(x) = E[h]$ is the mean function and $k(x_i, x_j)$ is the kernel function that models the correlation between input x_i and x_j. In this study, we use the Matern kernel[7].

This surrogate model GP enables us to predict an unseen point of the black box function, $h^* = h(x^*)$. We form a joint normal distribution of $p(h|D) = \mathcal{N}(\mu, \Sigma)$ and $p(h(x^*)|x^*) = \mathcal{N}(\mu^*, \Sigma_{**})$:

$$\begin{bmatrix} h \\ h^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu^* \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{bmatrix} \right),$$

then by conditioning on available information $\mathscr{D}$ and x^*, we get the PDF of the unseen point using GP:

$$p(h(x^*)|\mathscr{D}, x^*) = \mathcal{N}(e, \sigma^2)$$

where $e = \mu(x^*) + \Sigma_*^T \Sigma^{-1}(h - \mu(X))$ and $\sigma^2 = \Sigma_{**} - \Sigma_*^T \Sigma^{-1} \Sigma_*$.

Prediction allows us to estimate value of unseen data point, but we also need to update the prior model given new information. The posterior distribution will be equivalent as first calculating $P(y|h) = \frac{P(y,h)}{P(h)}$, where y is the actual observed value at x^*.(Assume noiseless observation) Then, we use Bayes' rule to get the updated GP model:

$$P(h|y) = \frac{P(y|h)P(h)}{P(y)}.$$

1.2.1 Training GP

Building such GP requires updating the joint normal distribution as more data points are available. Since the goal is to fit GP parameters (mean vector and covariance matrix), training GP means to find two objects such that we can maximize the accuracy. In general, people commonly use the loss function $\mathscr{L}$ introduced previously. In other words, maximizing the accuracy is equivalent as minimizing the loss function.

In GP, mean vectors($\{y_i|i = 1,2,...,n\}$) are taken from training data, so we tune kernel parameters to maximize the accuracy. However, GP does not use loss function. In stead, people use the marginal likelihood as an indicator of accuracy. Hence, let θ be kernel parameters, we can maximize the accuracy of GP by finding the correct θ with respect to $p(y|X,\theta,k)$, where y and X are from training set and k is the choice of kernel function. For example, Matern kernel has 3 parameters, one to control the normalization of y, one to control scale of changes of covariates (how fast the variable changes) and the final one to control the smoothness. Note that the smoothness parameter was chosen in advanced, i.e. it is a constant.

1.3 Bayesian Optimization

If we can get infinite numbers of evaluations of the underlying unknown function, GP will give us the exact solution. However, the geometry optimization calculation can be time consuming, thus we would like to get only the optimal information instead of the general information. In this case, we want to find if there is a JC parameter set that matches with selected experimental values. The optimization we are considering here can be formulated as following:

$$x^* = \underset{x \in \mathscr{X}}{\operatorname{argmin}} f(x),$$

where x^* is the potential solution and $f(x)$ is a cost function that wraps around the functional form of the JC model, i.e. the black box function. Let $h(x) : \mathbb{R}^4 \to \mathbb{R}^4$ be the black-box function where the input $x \in \mathbb{R}^4$, i.e. 4 JC parameters, and the output is a vector containing 4 elements: lattice energy of the rock salt, lattice constant a of rock salt, and lattice constants a & c of wurtzite. Using RMSE as our cost function, we can put a GP prior on f, where f is a scalar function. Before getting to the optimization process, we need another component for BO: that is the acquisition function. Given a surrogate model, which is GP in this case, we can predict unseen data much faster than actually evaluating it. The natural question is how do we find the next point to evaluate since computation time is valuable. To answer this question, BO introduces acquisition function. In this study, we used upper/lower confidence bound (UCB/LCB) as our acquisition functions. These two functions have the following form:

$$\alpha_{UCB} = \mu(x^*) + \kappa\sigma(x^*),$$

where x^* represents unobserved new points and κ is the exploration-exploitation constant. LCB has the similar form with minus sign replacing the plus sign in UCB. The interpretation is straightforward: based on what we have observed so far, we use the acquisition function to indicate the chance that we can improve current estimate of the optimum state. If κ is small, we focus on what we have observed so far, i.e. neighbors around the current optimum point. If κ is large, we believe the true optimum lies in the area where we have not explored yet. Therefore, we turned the original

minimization problem into another optimization problem with respect to acquisition functions. The following figure [8] gives the visual representation of roles of acquisition functions.

We want to maximize the acquisition function since the corresponding parameter is the next best point to evaluate. The green line in each frame is plot for sampled input parameters given GP. The red triangle indicates the maximum value and corresponding input parameters. Red dots following the previous frame is the actual evaluation using the parameter of red triangles.

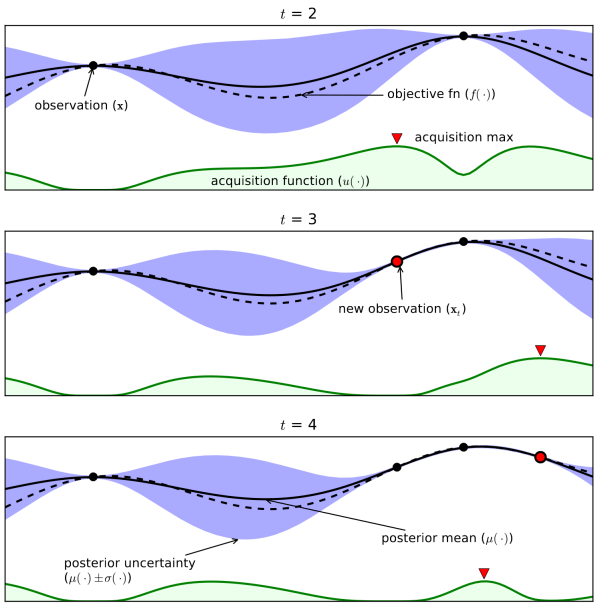

Figure 1.2: A visual representation of BO. t is the step of iteration. The red triangle shows the position where the acquisition function achieves its maximum, and we evaluate the black box function at the red triangle position for next iteration. BO is a process of repeating these two steps. Note that choice of acquisition function affect the optimization.

The BO algorithm is summarized below:

Bayesian Optimization Algorithm

The goal is to optimize a cost function $f(x)$ given dataset $\mathscr{D}$.

for n iterations:

$\rightarrow$ Select new x_{n+1} by maximizing the acquisition function using GP prior.

$\rightarrow$ Evaluate $y_{n+1} = f(x_{n+1})$.

$\rightarrow$ Augment data $\mathscr{D}_{n+1} = \{\mathscr{D}, (x_{n+1}, y_{n+1})\}$.

$\rightarrow$ Update the GP model.

end for

1.3.1 Modification of the single objective BO

The single objective BO can be time consuming since we have thrown away some essential information by put only one prior (one GP) on the RMSE, which we concatenate the output vector given the target vector into a single scalar value. This method is adapted from the paper.[10] Let the black box function be $y = h(x)$, where $x \in \mathbb{R}^4$ and y is our general output vector. Denote y^* as our target output vector for 4 different salts and dataset as $\mathscr{D}$. Our goal is to find a better probability description of the error, $P(d|x,\mathscr{D})$, where $d = ||y - y^*||^2$.
Consider if all individual components are independent, each element y_i in

$$y = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$$

follows a normal distribution where $y_i \sim \mathscr{N}(\mu_i, \sigma_i^2)$. Then

$$y \sim \mathscr{N}(\mu, diag(\sigma^2)),$$

where $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$, and $\sigma = [\sigma_1, \sigma_2, ..., \sigma_n]^T$.
Given the target vector $y^* = [y_1^*, y_2^*, ..., y_n^*]^T$:

$$d = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} d_i^2.$$

Note that since each $y_i \sim \mathscr{N}(\mu_i, \sigma_i^2)$, $(y_i - y_i^*) \sim \mathscr{N}(\mu_i - y_i^*, \sigma_i^2)$. Then

$$d_i \sim \mathscr{N}(\mu_i - y_i^*, \sigma^2).$$

Hence, summing over d_i and let $\sigma_i = 1$, then

$$d \sim NC_{\chi^2}(n, \lambda),$$

where n(degree of freedom) is number of components in the output vector and non-centrality term $\lambda = \sum_{i=1}^{n} (\mu_i - y_i^*)^2$.
Now the non-central chi squared distribution captures the real distribution of $P(d|X)$ since the non-central chi squared distribution does not have negative values in its support, where as a normal distribution random variable does. After re-parametrization, the reference paper shows the acquisition function LCB using the error distribution derived above.

1.4 Damping functions of the geometry optimization

In this study, we introduced two types of damping functions: close range (CR) damping and D3-like rational damping. The CR damping function was imposed at all times to ensure successful output.

D3-like rational damping was imposed to study whether this type of damping function helps the prediction of absolute lattice energy.

1. CR damping:
   One of the essential assumption to model the data this ways is that the unknown function must be well-behaved. A function which produces undefined value cannot be artificially defined since it will change the posterior distribution of the multivariate normal distribution(GP). Nevertheless, while collecting initial data points, there were several times when the geometry optimization codes gave error outputs. The reason of such problem arises from ill-defined geometry where lattice constants are too small such that the coulomb attraction dominates the potential and the unit cell collapses as the result. Hence, to prevent such thing happen, an artificial close range damping function was added to make sure the geometry optimization code can at least give some results, whether it is physical or not.

   The damping function applies only to the close range coulomb term. Here, an additional sigmoid function was put in front of the coulomb term such that we do not have negative infinity when inter-atomic is small. The sigmoid function had the following form:

   $$\frac{1}{1+e^{-c*(x-r_d)}},$$

   where c indicates how steep the function is and r_d is the mid-point of the sigmoid function, i.e. when the function is evaluated to be 0.5. Based on physical insights and time needed for numerical computation, we chose c to be 150 and r_d to be 0.1. In general, none of calculations will get to the point where the inter-atomic distance is 0.1 $\AA$.

2. D3-like damping:
   Dispersion forces (C_6 terms) and the corresponding dispersion energy is the central problem of predicting correct physical properties.[9] Since we will fix the part of targets that involve lattice constants, we ignore the coulomb interactions and focus on LJ part of the effective potential description. Notice that in general the wurtzite lattice constant a is smaller than rock salt lattice constant. However, the magnitude of wurtzite a cannot be on the wall part of LJ because it favors rock salt, which is not the predictions of the JC using reference parameters. Hence, by damping the long range dispersion, we lower the attractive interactions among neighbor ions in wurtzite, i.e. we make sure the rock salt crystal form is more stable with increasing relative energy difference.

   The LJ description can be transformed into the following form:

   $$4\varepsilon_{ij}\left[(\frac{\sigma_{ij}}{r_{ij}})^{12}-\frac{\sigma_{ij}}{r_{ij}})^{6}\right]=\frac{C_{12}}{r_{ij}^{12}}-\frac{C_6}{r_{ij}^{6}}.$$

Here, the C_6 is the focus. Once we see the LJ expression in this representation, a natural problem associated with it is the singularity. To prevent singularity when r_{ij} is small, we introduced D3-like Becke-Johnson rational damping function [1]:

$$\frac{C_6}{r_{ij}^6 + f(r_{ij})^6},$$

where $f(r_{ij})^6 = a_1 r_{ij} + a_2$. In the comparison study, we took a_1 and a_2 as constants.

1.5 Optimization procedure

The general optimization procedure can be summarized with the following flowchart:

Optimization parameters settings: $\kappa = 0.8$, 20 repetitions, 4 salts, 40 BO iterations
Optimization targets:

1. Relative lattice energy: Lattice energy of wurtzite - lattice energy of rock salt
2. Rock salt lattice constant: a=b=c
3. Wurtzite lattice constant: a=b, c

---

for 20 repetitions of each salt:
→ If with combing rules: start with 40 training points. Else: 60 training points.
→ Use BO(40/60 iterations) to get optimized JC parameters
→ Pass all BO optimized parameters through the tolerance filter to pick JC parameters
end for

1. Tolerance:
   The filter we imposed was based on results of BO and the necessity of this setting will be mentioned in the next section. As reader will find out LiF has the largest deviation compared to experimental lattice constants and lattice energy, so the upper error bound of LiF was set as the tolerance.

2. Cases studied:
   This study involves cross-comparison of following cases in the table:

   Though the focus is to compare two potential lattice structures of each salt, i.e. rock salt and wurtzite, we also checked all other 5 possible structures to make sure that rock salt structure is the most stable one. These 5 structures are Sphalerite, CsCl, NiAs, BetaBeo and FiveFive.

	15kJ/mol add.	15kJ/mol ~~add.~~	5kJ/mol	0kJ/mol	15kJ/mol D3-like damped
LiF	Analyse deviation of lattice constants and lattice energy of				
LiCl	rock salt and wurtzite with respect to the experimental values.				
LiBr	Tolerance of deviation: energy difference deviation < 0.5kJ/mol				
LiI	Lattice constants deviation $< 0.2\mathring{A}$.				

Table 1.3: ~~add.~~ means no combing rules being imposed.

Chapter 2

Results

In this section, we first present comparison between optimized JC parameters with and without combining rules. The result is suffice to answer whether the combining rule has its physical meaning. Then we will do the comparison study among JC optimizations while gradually decreasing energy difference between rock salt and wurtzite from 15 kJ/mol to 0 kJ/mol. This should give us insight of why using reference JC parameters gave rise to correct lattice energy and lattice constants but the incorrect stability. In other words, we can study how lowering relative energy difference starts to favor theoretical prediction of the wurtzite form. Finally, we will use comparison data between standard case(15 kJ/mol add.) and standard with D3-like damping function (BJ rational damping) to understand whether the damping function is the missing part of the JC model. If so, we will see if adding the damping function will improve the prediction of the absolute lattice energy.

2.1 First case

The reference JC parameters cannot predict all 4 targets correctly, so the first step was to try using BO to get a better parameter set such that we have smaller deviation with respect to all 4 targets. Also, since the combing rule is based on hard sphere model[6], we speculated releasing the non-physical constraint imposed by the combining rule can have a larger parameter space. A higher degree of freedom should lead to the hypothetical optimized results if the optimization with combining rules cannot.

The immediate observation was that for both types of JC, there is no unique solution which can predict all targets correctly. Hence, we used tolerance setting mentioned in the previous section and selected 20 that had the least deviation in terms of absolute energy. The table below shows the comparison of JC parameters between two types of JC model.

	JC params.	LiF	LiCl	LiBr	LiI
add	σ_M	0.1258±0.002	0.1611±0.006	0.1791± 0.005	0.1921±0.006
	σ_X	0.3157±0.005	0.3846±0.002	0.4120±0.002	0.4452±0.002
	ε_M	1.3686±0.1209	1.3862±0.325	1.1922±0.232	1.3431±0.272
	ε_X	0.3731±0.0822	0.6066±0.049	0.5670±0.053	0.6416±0.043
~~add~~	σ_M	0.1508±0.0409	0.2007±0.0276	0.1973±0.0112	0.1810±0.0270
	σ_X	0.3221±0.0090	0.4046±0.0162	0.4248±0.0063	0.4574±0.0129
	ε_M	1.2836±0.1147	1.4689±0.1197	1.2690±0.1140	1.4676±0.1449
	ε_X	0.2906±0.0977	0.3577±0.1807	0.3818±0.0733	0.4611±0.1592

(a) σ_M,σ_M,ε_M,ε_X

	JC params.	LiF	LiCl	LiBr	LiI
add	σ_{MX}	0.2207±0.0028	0.2729±0.0029	0.2955±0.0022	0.3186±0.0027
	ε_{MX}	0.7113±0.1054	0.9100±0.1155	0.8167±0.0786	0.9229±0.0941
~~add~~	σ_{MX}	0.2195±0.0029	0.2702±0.0024	0.2918±0.0016	0.3161±0.0020
	ε_{MX}	0.7416±0.0960	1.0456±0.1289	0.9663±0.0680	1.0185±0.0764

(b) For JC with additivity, cross terms σ_{MX} and ε_{MX} were calculated using combining rules. For JC without additivity, all shown data were directly from BO.

Table 2.1: Comparison of distribution of JC parameters(average ± 1 standard deviation). ~~add~~ means no additivity, i.e. combining rules.

2.2 Second case

As readers have noticed, the filter we used in the first case involves picking 20 parameters that have low absolute energy deviation. Again, the reason we could do this was due to not having a unique solution. If all 4 targets are within the tolerance, the absolute energy is naturally the next hypothetical target being imposed. However, we observed as the anion becomes heavier, the absolute energy deviation becomes smaller. Though original JC parameters give the correct absolute energy, one should notice that the relative energy difference decreases as the anion becomes heavier. Hence, it shows that we can either get the absolute energy or the relative energy difference. Since we care much more about the stability, we put it as out top priority. From this observation, we believe the relative energy difference is a restrictive target imposed during the optimization. Nevertheless, we still would like to see if we can get a better prediction of absolute energy, so we relax the target of relative energy difference gradually, from 15 kJ/mol to 5 kJ/mol, then to 0 kJ/mol.

	15 kJ/mol	5 kJ/mol	0 kJ/mol
LiF	-136	-116	-110
LiCl	-74	-53	-52
LiBr	-50	-29	-25
LiI	-35	-14	-11

Table 2.3: Average absolute energy deviation in kJ/mol.

2.3 Third case

As arguments shown in the introduction section, we speculate that damped dispersion not only favors rock salt, it also increases the absolute lattice energies. Therefore, we carry out the comparison study of adding D3-like damping function to understand this effect.

In all following tables, we repeated the BO 20 times and selected 20 best JC parameters that are the least deviated from the absolute lattice energy. Each table only shows two main structures: rock salt and wurtzite. We checked other 5 structures to make sure rock salt has the minimum lattice energy. In the table, some abbreviations are used, and only important values are shown (all other values are left as blank)

1. Average: Average of 20 observables
2. Stdev: Standard deviation of 20 observables
3. Exp. : Experimental values of each observable
4. DFT-PW1PW-D3: DFT calculation of corresponding observable
5. Avg. dev. : Average deviation is the deviation compared to the DFT calculation

1. JC 15kJ/mol:

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	3.8507	-1189.0603	2.9829	4.8320	-1174.0712
Stdev	0.0145	5.5282	0.0105	0.0169	5.5697
Exp.	4.0262	-1054.0000			
DFT-PW1PW-D3	3.9909		3.1289	4.6916	
Avg. dev.	-0.1402		-0.1460	0.1404	

Table 2.4: LiF, enegy difference set to be 15 kJ/mol.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	4.9262	-925.4299	3.8285	6.2025	-911.4621
Stdev	0.0076	3.0078	0.0058	0.0090	3.0037
Exp.	5.1399	-865.0000	3.8521	6.1181	
DFT-PW1PW-D3	5.0354		3.8701	6.1121	
Avg. dev.	-0.1092		-0.0416	0.0904	

Table 2.5: LiCl, enegy difference set to be 15 kJ/mol.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	5.3126	-860.6971	4.1360	6.7014	-845.5855
Stdev	0.0134	4.1698	0.0109	0.0173	4.2047
Exp.	5.5016	-821.0000	4.1510	6.6502	
DFT-PW1PW-D3	5.3546		4.1323	6.5862	
Avg. dev.	-0.0420		0.0037	0.1152	

Table 2.6: LiBr, enegy difference set to be 15 kJ/mol.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	5.8198	-786.9849	4.5362	7.3492	-772.0132
Stdev	0.0298	2.7434	0.0204	0.0315	2.7019
Exp.	6.0127	-765	4.5141	7.3112	
DFT-PW1PW-D3	5.8691		4.4737	7.271	
Avg. dev.	-0.0493		0.0625	0.0782	

Table 2.7: LiI, enegy difference set to be 15 kJ/mol.

2. D3-like rational damped:

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	3.8499	-1167.3243	2.9896	4.8403	-1152.3624
Stdev	0.0057	1.3422	0.0025	0.0040	1.4242
Exp.	4.0262	-1054.0000			
DFT-PW1PW-D3	3.9909		3.1289	4.6916	
Avg. dev.	-0.1410		-0.1393	0.1487	

Table 2.8: LiF, enegy difference set to be 15 kJ/mol, D3-like damped.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	4.9295	-909.6725	3.8418	6.2200	-895.7043
Stdev	0.0128	1.5079	0.0084	0.0135	1.5867
Exp.	5.1399	-865.0000	3.8521	6.1181	
DFT-PW1PW-D3	5.0354		3.8701	6.1121	
Avg. dev.	-0.1059		-0.0283	0.1079	

Table 2.9: LiCl, enegy difference set to be 15 kJ/mol, D3-like damped.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	5.3073	-843.4844	4.1492	6.7189	-828.5170
Stdev	0.0144	2.3636	0.0116	0.0183	2.4360
Exp.	5.5016	-821.0000	4.1510	6.6502	
DFT-PW1PW-D3	5.3546		4.1323	6.5862	
Avg. dev.	-0.0473		0.0169	0.1327	

Table 2.10: LiBr, enegy difference set to be 15 kJ/mol, D3-like damped.

Structures	Rock salt		Wurtzite		
Observables	a=b=c	Lattice Energy	a=b	c	Lattice Energy
Average	5.8037	-771.3198	4.5454	7.3588	-756.3789
Stdev	0.0257	2.7386	0.0197	0.0301	2.7902
Exp.	6.0127	-765	4.5141	7.3112	
DFT-PW1PW-D3	5.8691		4.4737	7.2710	
Avg. dev.	-0.0654		0.0717	0.0878	

Table 2.11: LiI, enegy difference set to be 15 kJ/mol, D3-like damped.

Chapter 3

Discussion

3.1 Importance of combining rules

Using Table 2.1, two separate plots for JC parameters are as follows.

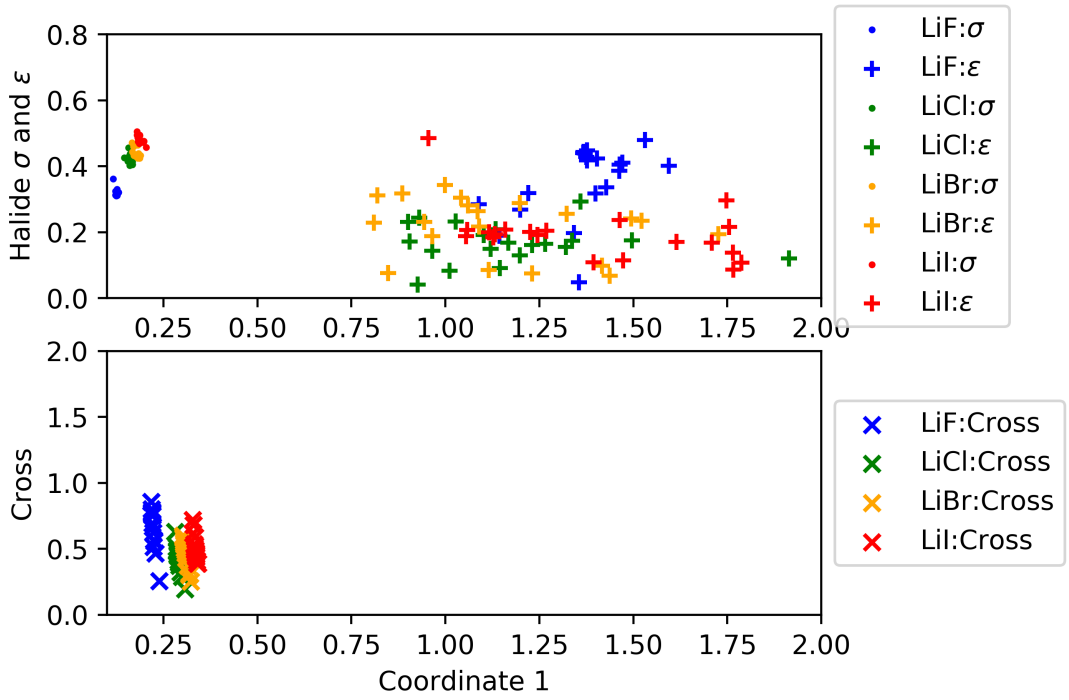

Figure 3.1: Optimized JC parameters for 4 salts. Coordinate 1 is a place holder. For example, in the top sub-figure, the dotted points are (σ_M, σ_X) and plus points are $(\varepsilon_M, \varepsilon_X)$. In the bottom figure, cross points are $(\sigma_{MX}, \varepsilon_{MX})$. Here, cross terms are calculated using combining rules.

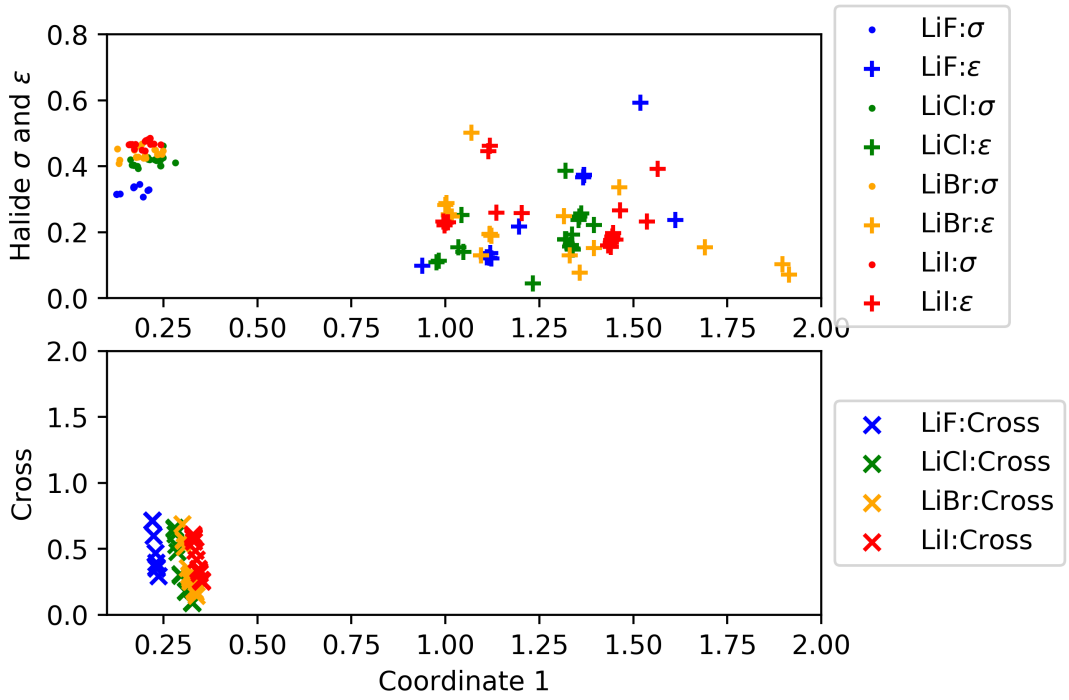

Figure 3.2: Optimized JC parameters for 4 salts. Coordinate 1 is a place holder. For example, in the top sub-figure, the dotted points are (σ_M, σ_X) and plus points are $(\varepsilon_M, \varepsilon_X)$. In the bottom figure, cross points are $(\sigma_{MX}, \varepsilon_{MX})$. Here, cross terms are optimized using BOm so no combining rules are imposed.

From Figure 3.1, there are two main observations: a general increasing trend for both σ_M and σ_X as the anion becomes heavier, and both ε's scatter everywhere. Based on settings of 4 targets (3 lattice constants and stability order), the strong correlation between σ and targets is not surprising since σ of the JC model, by construction, indicates the distance where inter-particle potential is zero. However, the trend shown by Figure 3.1 does not prove the physical restriction of the combining rules.

By alleviating the combining rules, there are 2 more parameters. With higher degree of freedom, we expected an improvement of the optimization: that either we can find a unique JC parameters set, or we can have a better represented trend of ε. However, Figure 3.2 shows a worse situation than optimizations with imposed combining rules. The increasing trend for both σ_M and σ_X is now only an increasing trend of σ_X.(Refer to Table 2.1) Nevertheless, while comparing the cross terms, σ_{MX} is fairly consistent in both cases. Therefore, in terms of predicting σ, this comparison test shows that the combining rules is effective, i.e. the physical interpretation is transferable to the ionic salt structure using the combining rules.

3.2 Behaviours of JC parameters and the damping function

Using results from the second and the third case, an observation is that by decreasing the energy difference as one of the targets, we were able to improve the prediction of the absolute lattice energy. (Table 2.3) Since we only need to set the energy difference being positive so that rock salt will nucleate, the next step is to loose the restriction on the relative energy difference such that we can improve the absolute lattice energy. In other words, after knowing there is no unique solution of the optimization problem, we want to find a better, if not the best, JC potential. Tests in these two cases allow us to both find a way to improve JC potential and set the criteria to find the optimum solution among all JC parameters optimized using BO.

Nevertheless, though decreasing the relative energy difference can improve the prediction of the absolute energy, we still want to keep it the same according to all DFT calculations, which suggests the 15 kJ/mol. Hence, a weak damping function that tunes the dispersion energy might work in this case.

	JC	JC with D3-like	JC with WY
LiF	-136	-113	-118
LiCl	-74	-44	-50
LiBr	-50	-22	-27
LiI	-35	-6	-10

Table 3.1: Average absolute energy deviation for JC with or without D3-like damping functions. Target values are same for both cases. Energies are in kJ/mol.

By adding the D3-like damping function, all targets were kept as original, and table 3.1 shows that predictions of absolute energies improved. Therefore, the D3-like damping function does improve the JC model with respect to getting the absolute energy correct. However, a question still remains: that the improvement of prediction of lattice energy shows a trend, suggesting the damping function works better for heavier anions, e.g. using the same damping function setting, absolute lattice energy deviation of LiI is smaller than LiF. The speculation is that heavier anions are more polarized than the lighter one, e.g. I- ions are more polarized than F- ions, and since the damping function aims to correct the dispersion energy at long distances, the construction of salts with heavier anions (larger inter-nuclei distances) are better suited in this improved JC models.

D3-like damping function is a weak damping function, so the next step was to see whether increasing strength of damping functions can improve predictions of absolute energy even more. Here, we chose the WY function, where it has the form:

$$f_6(r) = \frac{1}{1+e^{-a(\frac{r}{R_0}-1)}},$$

where R_0 is the sum of the atomic vdW radii. However, from the table 3.1, a stronger damping function makes it worse. Therefore, damping function is not the only missing factor of the empirical JC potential.

3.3 MD simulations

The study is based on the premise that JC parameters sets from BO can resolve the trade-off between stability order and poor predictions of lattice constants. Hence, several MD simulations for each salt were carried out to make sure nucleated structures match with the rock salt structure. Furthermore, these simulations allow us to find the difference among JC potentials, i.e. to find the optimum one. From all comparison tests done in previous sections, we are not able to eliminate any JC potentials since all of them give the correct prediction within the tolerance around targets.

All MD simulations start with geometry optimization using JC potential with optimized parameters, hence minimized geometry. Here, a constraint was imposed, where initial crystal structures were restricted to be the rock salt. Hence, the starting topology will always be the preferred rock salt structure. Then, an expanded NVT(pseudo NPT) simulated annealing with different temperature ramps were carried out. The extra space prevent the explosion of the simulation box, but a problem also arises while visualizing the radial distribution function.

3.3.1 Checking the nucleated structure

According the procedure mentioned above, it is convenient to compare the RDF plots of the salt before melting and after nucleation. If peaks of RDF plots match relatively well, we can conclude the JC potential with optimized parameters solved the stability order while giving lattice constants within tolerance. The figure below is a representative plot of the RDF of the before state and the after state. From the figure, since the first three peaks match, the local arrangement upto three outer shells are similar for the salt's before and after state. Lower peaks of the nucleated structure can be explained by the expanded simulation box. Higher volume of the simulation box creates pocket of holes of nucleated structures, and the trajectory file proves such irregular shaped holes, so throughout the averaging process of RDF, lower peaks were observed.

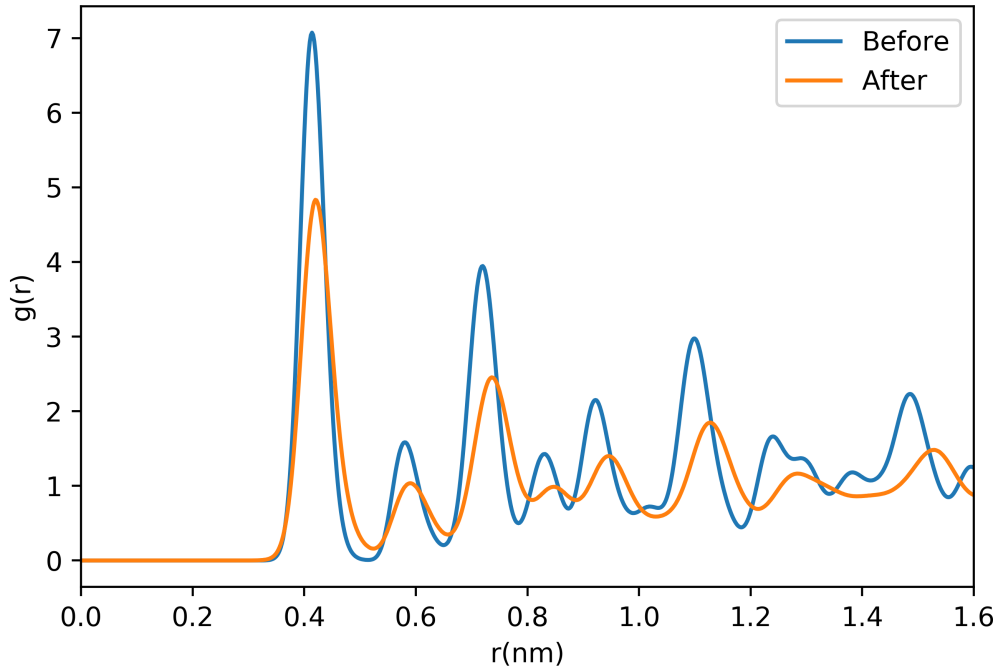

Figure 3.3: The comparison plot of RDF for LiI before melting(before) and for salt after nucleation(after).

This is the essential observation that allows us to conclude the correct nucleation structure if the salt has been through the melting/nucleation process. Hence, it is necessary to also look at the RDF plot of the molten salt. Here, such RDF for LiF is ploted. If LiF shows the typical fluid-like RDF, we can conclude other 3 salts are melted as well.

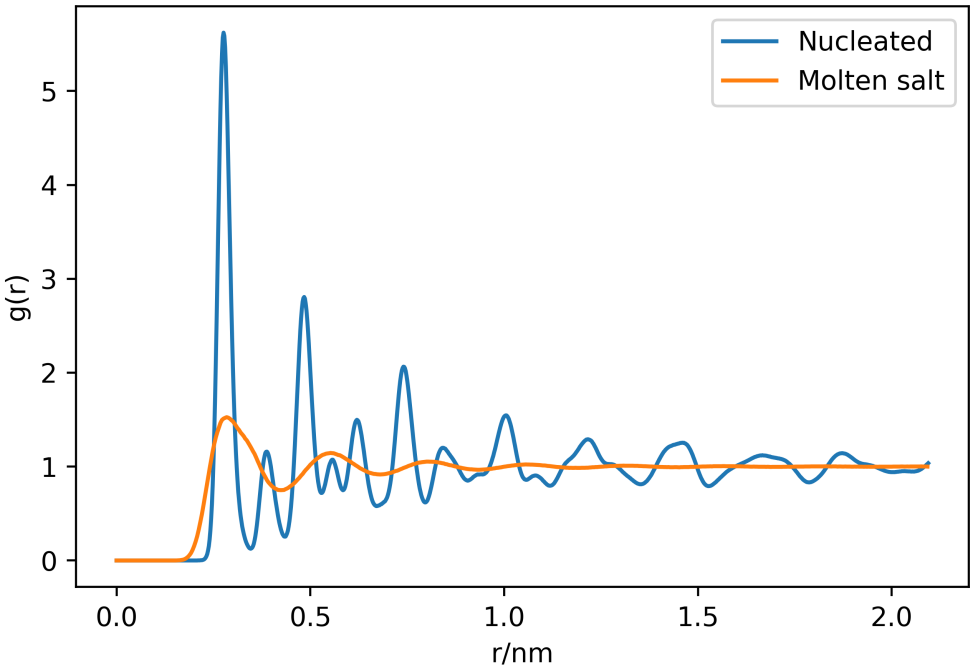

Figure 3.4: The comparison plot of RDF for LiF during melting (molten salt) and for salt after nucleation(nucleated). The molten salt RDF agrees with the typical fluid RDF.

From these two plots, it is possible to conclude that selected JC parameters solve the stability order while giving the correct lattice constants.

3.3.2 Melting/nucleation point

As mentioned before, epsilon parameters of the JC empirical potential model are responsible for determining well-depth, but with results from Section 3.1, one question remains: how can we explain the scattering phenomenon of epsilons. In other words, which JC potential is better, or is there another physical observable that allows us to decide it. Based on the physical interpretation of epsilons, melting point might be able to distinguish different JC potentials, where all have similar sigmas but different epsilons. Therefore, a simulating annealing tests were done on two salts, LiF and LiI, with selected Bayesian optimized JC parameters. These parameters were selected from 20 JC parameter sets mentioned in the section 2.1-2.3, and by ordering them based on the deviation from the absolute lattice energy, 3 parameter sets were chosen, which represented the least, medium, and the most deviated.

To ensure simulations were reproducible, we set cooling rates slowly enough to avoid hysteresis. The indication of nucleation/melting is based on sudden change of potential, i.e. re-arrangement of

the structure from liquid to solid or vice versa. The potential versus temperature plots are the best form for this purpose, and we show such plot for two salts below.

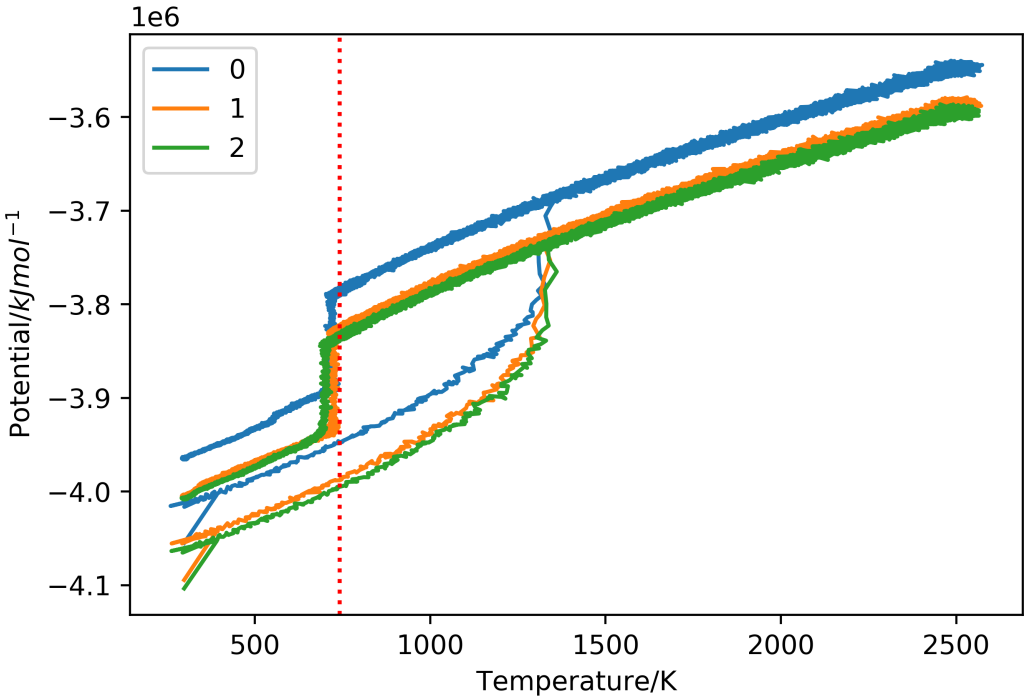

Figure 3.5: Potential(in kJ/mol) versus temperature(in K) plots of LiI with selected JC parameters from BO **with D3-like damping function**. Legend for plots(0,1,2) indicates the order of deviation from the absolute lattice energy. The red dashed vertical line indicates the experimental melting point, which is 742 K.

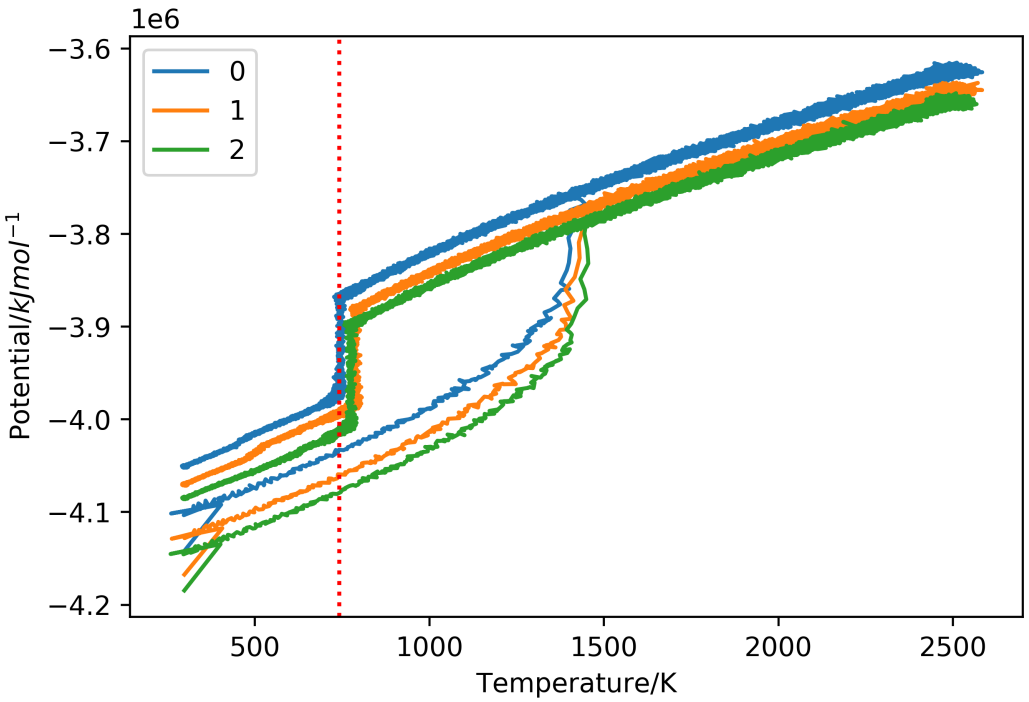

Figure 3.6: Potential(in kJ/mol) versus temperature(in K) plots of LiI with selected JC parameters from BO **without D3-like damping function**. Legend for plots(0,1,2) indicates the order of deviation from the absolute lattice energy. The red dashed vertical line indicates the experimental melting point, which is 742 K.

Placeholder for LiF plots!!!

From these 4 figures, there are two observations. First, the melting point can help setting the criteria in determining the optimum JC potential, but it does not explain or show the relationship between epsilons and the nucleation point. Second, damping function always underestimate the melting point, i.e. it over-stabilize the liquid phase of the salt. Even with small correction to the long range dispersion energy curve, this effect can be magnified with interactions among 6 neighbors. However, the reason why the damping function stabilize the liquid phase more than the solid phase is still not clear.

Chapter 4

Conclusion

We are interested in the nucleation process of LiX, but the original JC potential cannot grantee the most stable crystal structure for LiX salts as being the rock salt. We tried to find another JC potential by changing JC parameters, but optimizing the JC empirical potential with respect to relative energy difference and lattice constants is not a well defined problem, i.e. there is no unique solution. With tolerance set based on LiF, where it gave the largest deviation, Bayesian optimization is able to find solutions to the inverse problem. We studied the optimized JC parameter sets and found a well established conclusion that sigmas of the JC is the most crucial determining factor. Also, the comparison between optimization with and without combining rules imposed show a consistent result of the cross sigma(σ_{MX}), which suggest the physical significance of this simplification has its place in the problem. Furthermore, since all optimized JC potentials do not predict the correct absolute lattice energy, we introduce a weak D3-like damping function, and this does improve all predictions of the absolute lattice energy to a different extent. After BO and analysis of BO results, we carried out several MD simulation to show the nucleated structures for all 4 salts are indeed rock salt. Moreover, we found the melting point is a physical observable which helps to find the optimum JC potential among all optimized potentials. However, damping function over-stabilize the liquid phase for all 4 salts, and with current results, we cannot argue a satisfactory result. This can be a topic for the future study.

Bibliography

[1] A. D. Becke and E. R. Johnson. A density-functional model of the dispersion interaction. *The Journal of chemical physics*, 123(15):154101, 2005. → page 11

[2] D. Berthelot. Sur le mélange des gaz. *Compt. Rendus*, 126:1703–1706, 1898. → page 2

[3] J. B. Fraleigh and R. A. Beauregard. Linear algebra 2nd edition, p 246, 1990. → page 4

[4] Jim. Overfitting regression models: Problems, detection, and avoidance. https://statisticsbyjim.com/regression/overfitting-regression-models/. Accessed: 2020-02-24. → page 4

[5] I. S. Joung and T. 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. → pages 1, 2

[6] H. Lorentz. Ueber die anwendung des satzes vom virial in der kinetischen theorie der gase. *Annalen der physik*, 248(1):127–136, 1881. → pages 2, 13

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. → pages vi, 5, 6

[8] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015. → page 8

[9] D. G. Truhlar. Dispersion forces: Neither fluctuating nor dispersing. *Journal of Chemical Education*, 96(8):1671–1675, 2019. → page 10

[10] A. K. Uhrenholt and B. S. Jensen. Efficient bayesian optimization for target vector estimation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2661–2670, 2019. → page 9

[11] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006. → page 4