博士論文

Data-driven search for storage battery electrolyte development

（蓄電池開発のためのデータ駆動型電解液探索）

中山　智文

Data-driven search for storage battery electrolyte development
（蓄電池開発のためのデータ駆動型電解液探索）

Tomofumi Nakayama

Supervisor
Masato Okada

University of Tokyo, 2020
Graduate School of Frontier Sciences
Department of Complexity Science and Engineering

Abstract

In the field of materials design, which has a huge diversity, approaches based on computational science and materials informatics, as well as the experience and intuition of research and development personnel, are being actively studied. Computational approaches are designing better functioning materials first-principles from material design parameters, and materials informatics approaches are applying data science to materials science.

A typical target for materials informatics is the search for materials for lithium-ion batteries. The aim of the materials search for lithium-ion batteries is to find materials with superior material properties in multiple aspects such as higher voltage, higher capacity, longer life, higher safety and faster charge/discharge. For these properties, various new "electrode" materials have been reported[1, 2, 3]. On the other hand, no significant progress has been made in the development of new excellent electrolytes for commercial use, especially for lithium salts, since 1991. This is because the electrolyte is a liquid and its structure is more complex than that of the usual solid electrode materials, and therefore, it is difficult to search for electrolytes with various performance requirements[4, 5, 6, 7]. Virtual screening using materials informatics is one way to discover new electrolytes with the required performance. In this screening process, a database of descriptor material features is first constructed using data from first-principles calculations, molecular dynamics simulations and experiments. Next, a prediction model is built to predict a target property from the descriptors in the database with information technology. Finally, the prediction model is used to predict the properties of a large number of candidate materials, and the materials that are predicted to meet the required performance are extracted. Several applications of virtual screening for the search of new lithium-ion battery materials have been reported, but most of them are limited to the study of solid materials[8, 9, 10], while only a few applications of liquid materials have been reported[11, 12, 13].

Motivated by the above situation, in this doctoral thesis, we create a database for virtual screening and prediction models for the search for the electrolyte of lithium-ion batteries. In order to show the versatility of this method, we apply the same method to the prediction for electrolytes of alkali metals other than Li such as Na, K, Rb and Cs. In this study, the features to be predicted are "coordination energy" and "diffusion coefficients", and the prediction models are constructed by applying sparse modeling techniques such as exhaustive search and a sparse linear mixture model.

Acknowledge

Throughout the writing of this dissertation I have received a great deal of support and assistance.

I would first like to thank my supervisor, Professor Masato Okada, whose expertise was invaluable in formulating the research questions and methodology. His advice has helped me not only in the field of research, but also in every situation.

I would like to acknowledge my collaborators, Professor Yasuhiko Igarashi, Dr. Keitaro Sodeyama, Dr. Kenji Nagata and Dr. Atsushi Ishikawa, for their valuable guidance throughout my studies. They provided me with the tools and data that I needed to choose the right direction and successfully complete my dissertation.

I would also like to thank all members in our laboratory. They provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

I would like to express my gratitude to everyone at KARAKURI, Inc. for their understanding of my decision to put my studies first, even though I am CTO.

I would like to acknowledge my friend, Naoki Fujii, for introducing me to this laboratory. Unfortunately, he had passed away before I entered here, but I am proud to have carried on his work. He was the reason why I decided to go to the doctoral course.

In addition, I am grateful to my parents, Kosho Nakayama and Yoko Nakayama, they understood my decision to enter the graduate school, which was a great support for me.

Finally, I could not have completed this dissertation without the support of my wife, Narumi Nakayama. Before entering the graduate school, she has been providing me with emotional (and English) support at all times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to Materials Informatics

In materials science, improvements in computer performance have made it possible to perform accurate simulations using first-principles calculations. The results of the simulations are fed back into experimental science, and new data generated by the experiments are reflected in new simulations (Figure 1.1). Materials science has developed through the interaction of computational and experimental sciences. In recent years, with the development of information technology, data science, which extracts information from various types of data, has been attracting attention. Materials informatics is being applied to accelerate materials exploration by applying data science to conventional computational and experimental science in materials science. In the United States, the Materials Genome Initiative was launched in 2011 with the aim of doubling the speed of materials development by making full use of information science, and has been successful [14]. The competition for materials exploration using materials informatics is fierce worldwide, with similar projects being launched in Europe and China[15, 16, 17, 18, 19].

In this study, we develop methods for materials informatics using data science methods such as sparse modeling and Bayesian inference, specifically targeting computational science.

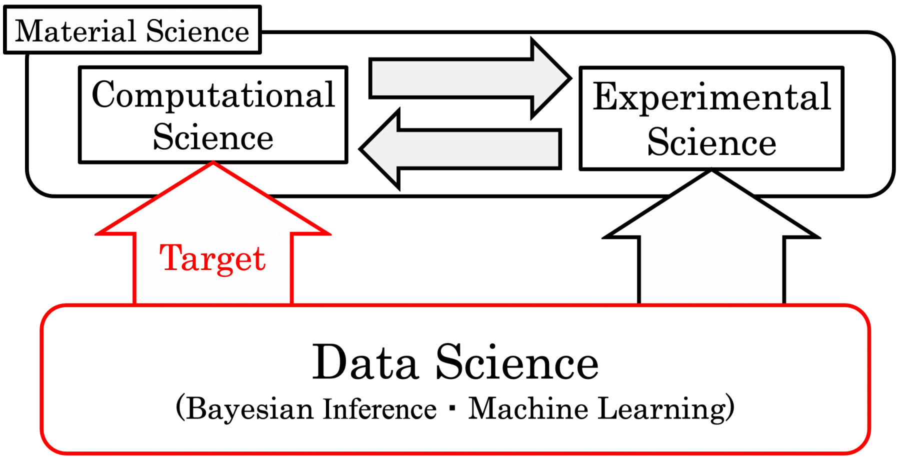


Figure 1.1: Materials informatics overview and our target.

1.2 Li-ion Battery

A typical target in materials informatics is the search for materials for lithium-ion batteries (LIB) and other secondary batteries (storage batteries)[20, 21]. The aim of the materials search for secondary batteries is to find materials with excellent material properties in multiple aspects such as higher voltage, more capacity, longer life, higher safety, faster charge and discharge, etc. With the widespread use of smartphones and electric vehicles, the demand for high performance secondary batteries is increasing rapidly. Efficient material search is strongly demanded by industry. However, the search for solid materials such as cathode and anode for LIB has progressed rapidly[8, 9, 10], while the commercial electrolyte has remained the same since 1991[11, 12, 13]. This is due to the fact that the search for materials for the electrolyte is more difficult than for solid materials such as the cathode and the anode, because the electrolyte is a disordered system. In recent years, large-scale materials science simulations, which were not possible in the past, have become possible using large-scale computers such as the K computer and Fugaku. Computational exploration of materials for lithium-ion battery electrolytes using first-principles molecular dynamics techniques is progressing, which is expected to further improve the performance of LIB. However, even with the use of large computers, first-principles molecular dynamics calculations for macromolecules are computationally expensive, and therefore a method to find materials with better performance is required.

1.3 Mission

Virtual screening using materials informatics is one way to discover new electrolytes with the required performance. An overview of the virtual screening is shown in Figure 1.2. Computational science in materials science is represented by the black boxes. In computational science, the parameters of material design are first defined, and then the parameters are used to calculate the features by first principles. The functions of the material (life, durability, etc.) are designed using the characteristics. This process is repeated in order to advance the material search, but it is often very expensive. The purpose of this study is to apply data science methods such as sparse modeling and Bayesian inference to computational science and to reduce the enormous computational cost. Concretely, we construct databases for virtual screening and examine prediction models for the search of the electrolyte of LIB. In addition, in order to demonstrate the versatility of the method, we apply the same method to alkali metals other than Li such as Na, K, Rb and Cs, and investigate the prediction models.

1.4 Structure of this paper

This paper consists of five chapters - Chapter 1 is this introduction, Chapter 2-4 is main contents, and Chapter 5 is the conclusion.

In Chapter 2, we describe our database. First, as candidate materials for the database, we selected 103 solvent molecules that are commercially available as battery grade materials from KISHIDA CHEMICAL Co.,Ltd.[22]. We also prepared two kinds of features: experimental values (experimental values) and calculated values (calculated

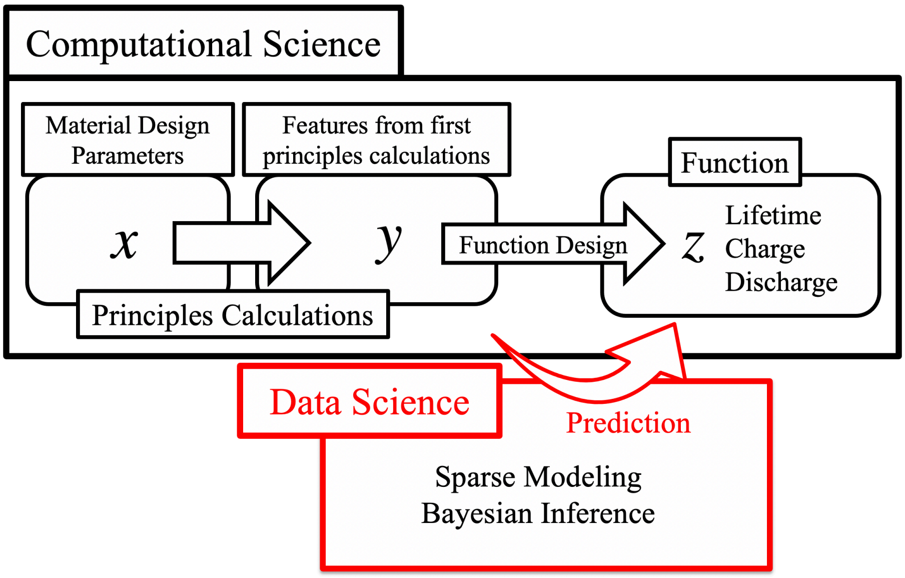


Figure 1.2: Applying data science to computational science

values). As the experimental values, we used following values from the catalogs of commercially available materials: melting point, boiling point, flash point, solvent density, and molecular weight of each solvent molecule. Calculated values were obtained from the Density Functional Theory (DFT) and DFT-Molecular Dynamics (DFT-MD) calculations. For the DFT calculations, Gaussian09 was used to perform cluster model DFT calculations of molecular systems for the 103 solvent molecules described earlier, producing six types of data: coordination energy between Li ions and solvent molecules, mulliken charges of atoms coordinating with Li ions (mainly oxygen atoms), the distance between the Li ion and the coordinating atom, HOMO energy, LUMO energy, dipole moment. DFT calculations were performed by replacing Li ions with other alkali metals (Na, K, Rb, and Cs) as a database for alkali metals other than Li, and the same data were generated. The database includes information on the cations used, such as ionic radius, electronegativity and atomic weight. In the DFT-MD calculations, Li ions are placed in the solvent and DFT-MD calculations are performed under the periodic boundary condition to simulate the diffusion of Li ions in the solvent. The diffusion coefficients of Li ions were used as the database value.

In Chapter 3, we describe the development of prediction models for the coordination energy based on the database[23, 24, 25]. The coordination energy is a quantity related to the diffusion of Li ions and is considered to be related to the rate of charge and discharge of the secondary battery. Therefore, it is an important issue in LIB to find a solvent molecule with a reasonable coordination energy. The coordination energy is obtained by DFT calculations using a cluster model. Since it takes several hours to obtain the coordination energy of a single solvent molecule, the purpose of this chapter is to construct models to accurately predict it. One of the important aspects of building a prediction model is variable selection. The database may contain variables that are noisy in predicting the coordination energy. It is necessary to remove them appropriately and to extract the variables that are important for prediction. Sparse modeling is the modeling that assumes that the important variables are part of a large number of variables. In this study, we perform an exhaustive search (ES) method, which is one of the sparse modeling methods. In ES, prediction models are built for all combinations of variables, and the combinations of variables are evaluated with some criteria. In the prediction of the coordination energy of the electrolyte solvent for LIB, we used exhaustive search with linear regression (ES-LiR) and exhaustive search with Gaussian Process (ES-GP). In both methods, as a criterion, we used Cross Validation Error (CVE), which is one of the measures of predictive performance for unknown data. As a result, CVE was reduced by 14% and 52% for ES-LiR and ES-GP, respectively, compared to linear regression of all variables. We have also visualized the key variables and clarified which ones are important in improving the performance of the model for predicting the coordination energy of the electrolyte solvent of LIB. To demonstrate the versatility of the method, we used a database of four alkali metals (Na, K, Rb, and Cs) for Li ions, and predicted the coordination energy of the alkali metal cations. In the construction of the prediction model, ES-LiR and ES-GP were performed as in the case of Li ions only and compared with the case using all variables. The results showed that CVE decreased by 0.7% and 88% in ES-LiR and ES-GP, respectively, compared to the linear regression of all variables, and while there was no significant difference in ES-LiR, very accurate predictions were obtained in ES-GP.

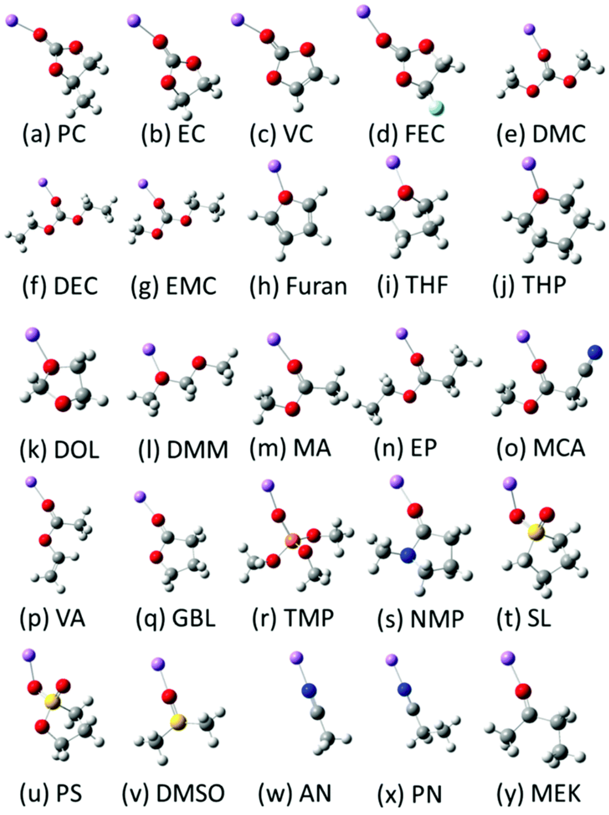
In Chapter 4, we construct prediction models for diffusion coefficients based on the database we created. The diffusion coefficient is a quantity that expresses the speed of diffusion of Li ion in the electrolyte and is considered to be related to the speed of charge/discharge of the secondary battery as well as the coordination energy, so finding a solvent molecule with a reasonable diffusion coefficient is an important issue. The diffusion coefficients are obtained by DFT-MD calculations. Since it usually takes several weeks to months of calculation time to obtain the diffusion coefficient for a single solvent molecule, the purpose of this chapter is to develop a model to predict this. As with the prediction of the coordination energy in Chapter 3, we first performed linear regression with all variables, GP with all variables, ES-LiR and ES-GP and we confirmed the accuracy of their predictions. The results showed that the prediction accuracy was not so high. One possible reason for the low prediction accuracy is that the data for the diffusion coefficients may come from a mixture model. In materials science, data can come from multiple backgrounds. In that case, there is limitations of a single model to explain the data. Thus, we applyed a mixture model to this data. In this study, we used Sparse Linear Mixture Model (SpLMM)[26, 27, 28] which is a method of sparse modeling. SpLMM is a model that assumes that the data originate from multiple linear models and predicts the target variable using spaces of explanatory variables, where each model is not necessarily identical. Since multiple models are used to make predictions, SpLMM do not produce a single prediction value. However, in materials science, even if the model does not have a single prediction, the model can narrow down the candidates for the desired material from a large number of candidates, so it is efficient to perform DFT-MD calculations and experiments to find out the exact value. Therefore, if we can find a more consistent model to explain the data, we can achieve a faster material search. In this study, we used log loss as an indicator to compare the degree of model agreement. The log loss is the mean of the sign-reversed logarithm of the predictive distribution over the test data. The smaller the value, the better the model matches the data. The test was performed with 10-fold CVs and the average of the log loss values was used as the comparison. As a result, the log loss value of SpLMM was reduced by about 80% compared to ES-LiR and ES-GP. This indicates that mixture models that predict multiple models are better at explaining this data than LiR or GP, which make predictions with a single model, and that mixture models such as SpLMM are very effective in predicting data consisting of multiple backgrounds, which often occur in the field of materials science.

Finally, Chapter 5 discusses the potential impact of the series of studies described in this paper on industry and the research community, as well as issues and directions for future research.

Chapter 2

Database

To predict novel LIB liquid electrolytes with desired properties by the information techniques, we constructed a database of known liquid electrolytes. We selected 103 solvent molecules which were commercialized as battery grade materials from KISHIDA Chemical Co., Ltd[22]. We adopted the values of melting point, boiling point, flash point, density of solvent, and molecular weight from the catalogue data. Some of representative solvent molecules are shown in Figure 2.1.



(a) PC (b) EC (c) VC (d) FEC (e) DMC

(f) DEC (g) EMC (h) Furan (i) THF (j) THP

(k) DOL (l) DMM (m) MA (n) EP (o) MCA

(p) VA (q) GBL (r) TMP (s) NMP (t) SL

(u) PS (v) DMSO (w) AN (x) PN (y) MEK

Figure 2.1: Representative 25 solvent molecules for the database (Li, purple; O, red; N, blue; C, grey; F, light blue; S, yellow; P, orange; H, white). The solvent names are referred to in Table 2.1.

Abbreviation	Solvent name	Chemical formula	E_{coord} (kcal/mol)	HOMO (eV)	LUMO (eV)	Dipole moment (Debye)	Mulliken charge	R(Li–O) (Å)
PC	Propylene carbonate	C4H6O3	-57.4	-7.93	0.946	5.255	-0.243	1.747
EC	Ethylene carbonate	C3H4O3	-55.9	-8.017	0.919	5.07	-0.24	1.752
VC	Vinylene carbonate	C3H2O3	-51.7	-6.973	-0.137	4.365	-0.231	1.76
FEC	Fluoro-ethylene carbonate	C3H3O3F	-51.2	-8.468	0.493	4.487	-0.222	1.763
DMC	Dimethyl carbonate	C3H6O3	-50.0	-7.774	1.115	0.342	-0.306	1.747
DEC	Diethyl carbonate	C5H10O3	-52.6	-7.654	1.217	0.613	-0.308	1.74
EMC	Ethyl methyl carbonate	C4H8O3	-51.3	-7.713	1.168	0.514	-0.307	1.744
DAC	Diallyl carbonate	C7H14O3	-31.7	-7.419	-0.238	0.494	-0.306	1.74
Furan	Furan	C4H4O	-48.7	-6.265	0.296	0.511	-0.17	1.866
THF	Tetrahydro-furan	C4H8O	-47.2	-6.832	1.38	1.434	-0.323	1.808
THP	Tetrahydro-pyran	C5H10O	-43.2	-6.711	1.537	1.301	-0.324	1.804
DOL	1,3-Dioxolane	C3H6O2	-64.4	-6.955	1.493	1.324	-0.315	1.818
DMM	Dimethoxy methane	C3H8O2	-52.0	-6.846	1.459	2.165	-0.298	1.905
MA	Methyl acetate	C3H6O2	-53.5	-7.371	0.339	1.733	-0.265	1.755
EP	Ethyl propionate	C5H10O2	-58.6	-7.31	0.414	1.763	-0.269	1.787
GBL	g-Butyro-lactone	C4H6O2	-54.7	-7.269	0.254	4.296	-0.237	1.758
TMP	Trimethyl phosphate	C3H9O4P	-56.8	-7.765	1.112	3.356	-0.467	1.74
NMP	N-Methyl-2-pyrrolidone	C5H9ON	-65.1	-6.421	0.842	3.609	-0.299	1.724
ES	Ethylene sulfite	C2H4O3S	-63.9	-7.725	-0.823	3.123	-0.423	1.758
SL	Sulfolane	C4H8O2S	-63.7	-7.383	0.826	5.087	-0.459	2.014
PS	1,3-Propane sultone	C3H6O3S	-57.3	-7.917	0.549	5.468	-0.426	2.034
DMSO	Dimethyl sulfoxide	C2H6OS	-67.8	-6.01	0.963	3.821	-0.542	1.718
AN	Acetonitrile	C2H3N	-47.0	-8.933	0.898	3.743	-0.181	1.92
PN	Propionitrile	C3H5N	-48.4	-8.802	0.587	3.826	-0.185	1.914
MEK	Methyl ethyl ketone	C4H8O	-53.0	-6.601	-0.386	2.771	-0.225	1.759

Table 2.1: Calculated values of the coordination energy (E_{coord}), the HOMO energy, the LUMO energy, the dipole moment, the Mulliken charge of the oxygen (nitrogen) atom, and the distance between the Li-ion and the oxygen (nitrogen) atom (R(Li–O)) of 25 solvent molecules for the database.

2.1 DFT Calculation

To make the database of the electrolytes more substantial, we added the following values obtained by density functional theory (DFT) calculations of the molecular systems using the Gaussian 09[29], the coordination energy between a Li-ion and a solvent molecule, the Mulliken charge of the atom (typically oxygen atom) that is coordinated to a Li-ion, the distance between a Li-ion and the coordinated atom (typically Li-O distance) (R(Li-O)), the HOMO energy, the LUMO energy, and the dipole moment values of the 103 solvent molecules [23]. The calculated data of the representative solvent molecules are shown in Table 2.1, and the complete data are listed in Table A.2 in the Appendix. The coordination energies (E_{coord}) are evaluated by the difference between the “ total energy of a Li–solvent complex ” and “ the total energies of a solvent molecule and that of a Li-ion ” ($E_{coord} = E(\text{Li-solvent}) - E(\text{solvent}) + E(\text{Li-ion})$). We adopted the B3LYP functional[30] with cc-pVDZ basis sets[31]. The Mulliken charges and the dipole moments are obtained from the DFT calculations of pure solvent molecules without Li-ions. Geometry optimizations of the Li-solvent complexes and the pure solvent molecules were also carried out.

In addition, we changed the cations from Li to other alkali metals (Na, K, Rb, and Cs), performed DFT calculations, and created a similar database[24]. In this DFT calculation, M06-2X was used for the exchange–correlation functional, since this functional is reported to accurately predict the thermodynamic properties of main group elements [32, 33]. The Def2-SVP basis set was used for all the elements, and the pseudo-potential was used for K, Rb, and Cs [34]. Another alkali ion, Fr, is omitted in this work because it is unstable and radioactive, thus not relevant for batteries. Atomic charges were calculated by the natural population analysis method proposed by Weinhold et al., using the NBO 6 program [35]. All the calculations were performed with Gaussian16 [36]. The data of Ionic Radius, Electronegativity, and Atomic Weight are added to the database as cation information. The full list of electrolyte solvents examined is shown in Table A.3 in the Appendix.

2.2 DFT-MD Calculation

In order to develop a prediction model for the diffusion coefficient of Li-ion in each solvent, a database of diffusion coefficients was constructed as validation data. In this study, the behavior of the atoms in the electrolyte is simulated by using first-principles molecular dynamics calculations (DFT-MD). We performed DFT-MD simulations with the Car-Parrinello electronic and ionic dynamics[37] using CPMD code[38]. Total energies were calculated at the G point in a supercell approach by using a PBE generalized gradient-corrected exchange-correlation functional[39, 40]. A fictitious electronic mass of 500 au and a time step of 4 au (0.10 fs) were chosen. The energy cutoff of the plane wave was set to 90 Ry. Goedecker's norm-conserving pseudopotentials[41, 42] for C, H, O, N, S, F and Li were used. Nuclear temperature was controlled using a Nosé thermostat[43] with a target temperature of 298 K. The electronic wave function was quenched to the Born-Oppenheimer surface approximately every 1 ps to maintain adiabaticity. After equilibration with the NVT ensemble at 298 K for at least approximately 50 ps, the Nosé thermostat was switched off, and statistical averages were computed from trajectories of another 50 ps in length with the NVE ensemble at an average kinetic temperature of 298 K[44]. For the statistical average, four to eight different trajectories were calculated in each settings with different initial guess structures.

We calculated diffusion coefficients, from the Mean Square Displacement (MSD) of the DFT-MD simulations of the HC and LC systems with neutral charge. The MSD was calculated as follows. Let $n_k = 5000k(k = 0, \cdots, 80)$. Let $x_{n_k}^t$ be the coordinates of the Li-ion at the $n_k + t(t = 0, \cdots, 100000)$ steps. We calculated MSD for $t = 0, \cdots, 100000$ as follows.

$$\mathrm{MSD}(t) = \frac{1}{80}\sum_{k=0}^{80} |x_{n_k}^t - x_{n_k}^0|^2.$$

That is, MSD(t) is the mean of the square of the distance moved after the t steps. It shows how much Li-ion is diffused in the t steps. Figure 2.2 shows a graph with the time t(fs) = 10t(steps) on the horizontal axis and the MSD on the vertical axis. The relationship between the number of steps and the MSD can be expressed as a linear function in sufficiently large steps. The slope of the linear function is called the diffusion coefficient. In this study, this slope was obtained by linear regression of the data after 1000 steps. Finally, we created a database of diffusion coefficients for 38 out of the 103 samples described earlier. The complete database is listed in Table A.2 in Appendix.

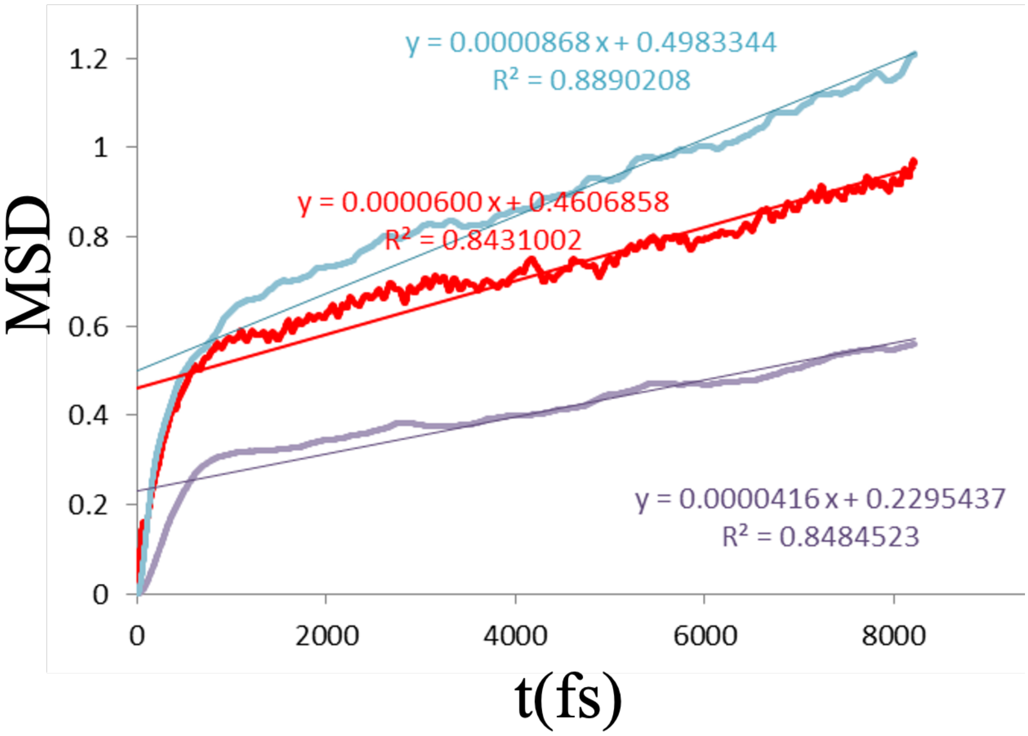


Figure 2.2: MSD at each step.

Chapter 3

Coordination Energy Prediction

3.1 Introduction

In the search for LIB liquid electrolytes, the evaluation of the properties of ion transport and electrochemical stability is indispensable. For the transport, solvation to and desolvation from Li-ions at the electrolyte/electrode interface plays a crucial role, and thus the coordination energy of the solvent to Li-ions is an important measure. For the electrochemical stability, the quantities such as ionization potential and electron affinity are significant. Here, however, we focus on the quantities related to the Li-ion transport as the first target.

In this study, we investigated the estimation accuracy of the MLR, LASSO, ES-LiR, ES-GP techniques in the search for liquid electrolyte materials [23]. We estimated the coordination energies as the required properties of the LIB liquid electrolytes and discussed the extracted descriptors by LASSO, ES with linear regression (ES-LiR) and ES with Gaussian process (ES-GP). As a result, we succeeded in extracting the important descriptors for predicting the coordination energies with high accuracy and clarified the effectiveness of ES.

In order to demonstrate the versatility of these methods, we have also tested the database with alkali metals (Na, K, Rb and Cs) other than Li-ion as cations[24]. The results show that the ES can be used to extract descriptors that can predict the coordination energy with high accuracy.

In this chapter, the ES method is first described in detail in Section 3.2. Next, in Section 3.3, we show that ES can correctly select variables in ideal data using synthetic data. Then, in Section 4, ES is applied to the two simulation data. The first data is the coordination energy of Li-ion, and the second data is Li-ion data plus other alkali metals (Na, K, Rb and Cs). We apply the ES method to these data and discuss the results and accuracy of the variable selection. Finally, we conclude in Section 3.5.

3.2 Method

In this section, we first introduce the exhaustive search (ES) method as a sparse modelling method for variable selection to improve the prediction accuracy of regression [45]. The ES method is available for a wide range of learning tasks with various learning machines because the ES method only requires a change in input variables, evaluates the variables and searches all the combinations of variables. In this section, we introduce exhaustive search with a linear regression model (ES-LiR) [46, 23] and exhaustive search with a Gaussian process (ES-GP)[47, 48].

3.2.1 Exhaustive Search

When there are N explanatory variables, the simplest method for selecting variables is to exhaustively search for all combinations, which requires the combinations of variables to be estimated $2^N = {}_N\mathrm{C}_0 + {}_N\mathrm{C}_1 + \cdots + {}_N\mathrm{C}_N$ times the number of estimations [49, 45].We call this naive method the ES method [50]. Cover and Van Campenhout reported that any exact methods for variable selection come at the expense of a computational complexity of at least $O(2^N)$ [51], and this is also true for the ES method. However, since the size of the data is not large in this study and because of the achieved improvements in computer performance, we can easily apply the ES method for the estimation [23, 46]. Moreover, the ES method is available for a wide range of learning tasks with various learning machines, such as linear regression and Gaussian process, since the ES method only requires a change in input variables, evaluates the variables and searches all the combinations of variables.

To formulate the ES method, we introduce an indicator

$$\mathbf{c} = [c^1, \dots c^\mu, \dots, c^N], \tag{3.1}$$

which represents a set of selected explanatory variables. $c^\mu = 1$ indicates that the μ-th explanatory variable x^μ is selected, and $c^\mu = 0$ indicates that it is not selected. Each realization value represented by $\mathbf{c}$ is called "state" by analogy with statistical mechanics [52]. Using an indicator $\mathbf{c}$, the input-output relationship of the learning machine between the descriptor data $\mathbf{x}_* = [x_*^1, \cdots, x_*^\mu, \cdots, x_*^N]$ of sample "$*$" and the estimated objective variable y_* is formulated as follows:

$$y_* = f_{\mathbf{c}}(\mathbf{x}_*(\mathbf{c})), \tag{3.2}$$

where $f_{\mathbf{c}}$ is the functional relationship learned by using the learning machine for state $\mathbf{c}$ and $\mathbf{x}(\mathbf{c})$ represents a subset of $\mathbf{x}$ excluding unrelated elements.

Cross Validation

In the ES method, we compare and evaluate all states $\mathbf{c}$ under a criterion for evaluating the estimation accuracy. In this research, cross validation (CV) is used to evaluate the estimation accuracy of the descriptors. CV approximately extracts prediction error using only limited data, and the CV procedure consists of three steps, as explained below. First, the complete learning data $(\mathbf{X}, \mathbf{y})$ are divided into training data $(\mathbf{X}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}})$ and test data $(\mathbf{X}_{\mathrm{te}}, \mathbf{y}_{\mathrm{te}})$. Let p_{te} be the number of test data and p_{tr} be the number of training data. Next, $f_{\mathbf{c}}$ is adjusted such that the output of the learning machine $f_{\mathbf{c}}(\mathbf{X}_{\mathrm{tr}}(\mathbf{c}))$ and the value of $\mathbf{y}_{\mathrm{tr}}$ are close. Finally, we evaluate how well the trained learning machine can correctly describe the input-output relation of the test data $(\mathbf{X}_{\mathrm{te}}(\mathbf{c}), \mathbf{y}_{\mathrm{te}})$ by using the loss function

$$\epsilon(\mathbf{X}_{\mathrm{te}}(\mathbf{c}), \mathbf{y}_{\mathrm{te}}) \equiv \frac{1}{p_{\mathrm{te}}} \sum_{i=1}^{p_{\mathrm{te}}} (y_{\mathrm{te},i} - f_{\mathbf{c}}(\mathbf{x}_{\mathrm{te},i}(\mathbf{c})))^2, \tag{3.3}$$

where $(\mathbf{x}_{\mathrm{te},i}(\mathbf{c}), y_{\mathrm{te},i})(i = 1, \cdots, p_{\mathrm{te}})$ is the i-th sample of $(\mathbf{X}_{\mathrm{te}}(\mathbf{c}), \mathbf{y}_{\mathrm{te}})$.

Since the value of the loss varies depending on how the data are divided, the above three steps are repeatedly conducted with different partitions of the data to evaluate the typical value of the loss function. For example, in M-fold CV (in this study, $M = 10$), the learning data are randomly divided into M subsets to have the same size, and the mean of the loss values is calculated by performing the above three steps M times using each subset as test data only once, and we call the mean cross validation error (CVE). Let CVE($\mathbf{c}$) be the CVE for the state $\mathbf{c}$, and it is formulated as follows:

$$\mathrm{CVE}(\mathbf{c}) = \frac{1}{M} \sum_{\mathbf{X}_{\mathrm{te}}, \mathbf{y}_{\mathrm{te}}} \epsilon(\mathbf{X}_{\mathrm{te}}(\mathbf{c}), \mathbf{y}_{\mathrm{te}}). \tag{3.4}$$

3.2.2 Exhaustive Search with Linear Regression, ES-LiR

Assuming that the relationship between descriptors and the objective variable is linear and using all of the descriptors, one can select multiple linear regression (MLR), and the estimation form is the linear sum of descriptors as follows:

$$y_* = f(\mathbf{x}_*; \mathbf{w}) = w^0 + \sum_{\mu=1}^{N} w^\mu x_*^\mu \tag{3.5}$$

where $\mathbf{w} = [w^1, \cdots, w^\mu, \cdots, w^N]$ is the weight of descriptors and w^0 is a constant term. An algorithm that applies ES to MLR is ES-LiR [46]. When the learning machine is in a state $\mathbf{c}$, its input-output relationship is expressed as follows:

$$y_* = f_{\mathbf{c}}(\mathbf{x}_*(\mathbf{c}); \mathbf{w}(\mathbf{c})) = f(\mathbf{x}_*; \mathbf{w} \circ \mathbf{c}) = w^0 + \sum_{\mu=1}^{N} c^\mu w^\mu x_*^\mu. \tag{3.6}$$

Here, the symbol $\circ$ represents the Hadamard product and is defined as $(\mathbf{x} \circ \mathbf{c})^\mu = x^\mu c^\mu$. In ES-LiR, CV is performed for each state, and by searching for the state where CVE is the minimum, it is possible to find the optimal combination of descriptors for estimating the objective variable. Since MLR and ES-LiR have the assumption that the relationship between descriptors and the objective variable is linear, the estimation may fail when it is not satisfied.

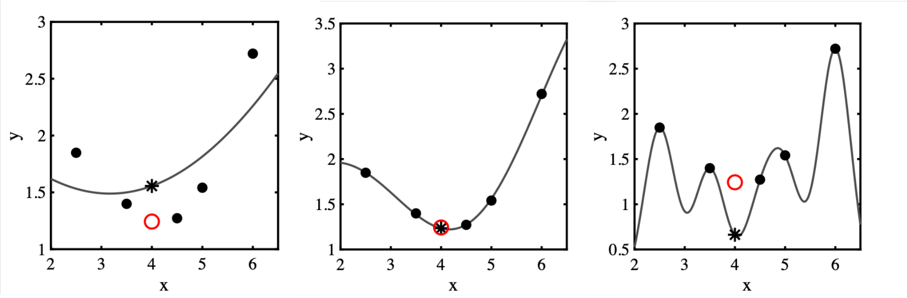
3.2.3 Exhaustive Search with Gaussian Process, ES-GP

To further improve the estimation accuracy by both estimating the nonlinear relationship and variable selection, we conduct ES-GP, which is the method for applying ES to GP.

To learn the nonlinear relationship, we assumed that from two materials that are sufficiently similar in the descriptors, the target features will be similar and can be determined by interpolation. The method of GP regression enables us to interpolate and predict target features [53] using the similarity of descriptors and approximates a nonlinear relationship and provides an unbiased prediction of intermediate values using a few parameters.

In GP regression, the descriptor data $\mathbf{x}_* = [x_*^1, \cdots, x_*^\mu, \cdots, x_*^N]$ of sample "$*$" and the estimated objective variable y_* are given by the indicator $\mathbf{c}$, expressed as follows:

$$\begin{aligned} y_* &= f_{\mathbf{c}}(\mathbf{y}; \mathbf{w}(\mathbf{x}_*(\mathbf{c}), \mathbf{X}(\mathbf{c}))) && (3.7) \\ &= \mathbf{w}(\mathbf{x}_*(\mathbf{c}), \mathbf{X}(\mathbf{c})))^{\mathrm{T}} \mathbf{y}, && (3.8) \end{aligned}$$



(a) Too small β (b) Optimum β (c) Too large β

Figure 3.1: Effects of difference in a parameter of the Gaussian process. The vertical axis is the objective variable y, and the horizontal axis is the value of the parameter. The black circles are observation points, "*" is the true value of the point to be estimated, and the red circle is the estimated value of the point to be estimated. If a parameter that is too small is used, overfitting will occur and the estimation accuracy will decrease. If the parameter is too large, the regression curve will be too smooth.

where $\mathbf{w}(\mathbf{c}) = \mathbf{w} = [w_1, \cdots, w_i, \cdots, w_p]^{\mathrm{T}}$ is a weight vector determined by the distance between $\mathbf{x}_*(\mathbf{c})$ and $\mathbf{X}(\mathbf{c})$ and $\mathbf{y} = [y_1, \cdots, y_p]^{\mathrm{T}}$ is the training data of the objective variable. The distance is treated by applying a similarity function k, which is often called the kernel between $\mathbf{x}_*(\mathbf{c})$ and $\mathbf{X}(\mathbf{c})$. Note that the weight vector $\mathbf{w}(\mathbf{c})$ in ES-LiR is the weight of descriptors and describes the relationship between descriptors and the objective variable. The weight vector $\mathbf{w}(\mathbf{c})$ in ES-GP describes the distance between samples for interpolation, where the researcher has no knowledge regarding the descriptors. The former and the latter are classified in statistics as parametric and nonparametric estimation, respectively.

Assuming that the noise follows a normal distribution of the variance σ^2, $\mathbf{w}$, which minimizes Eq. (3.3), can be strictly calculated [53], and $\mathbf{w}(\mathbf{c})$ is also expressed as follows:

$$\mathbf{w}(\mathbf{x}_*(\mathbf{c}), \mathbf{X}(\mathbf{c}))) = \left[(\mathbf{k}_*(\mathbf{c}))^{\mathrm{T}}(\mathbf{K}(\mathbf{c}) + \sigma^2\mathbf{I})^{-1}\right]^{\mathrm{T}}, \tag{3.9}$$

where $\mathbf{I}$ is a unit matrix of size p, and $\mathbf{K}(\mathbf{c}) = \{k(\mathbf{x}_i(\mathbf{c}), \mathbf{x}_j\mathbf{c}))\}_{i,j}$ is a matrix consisting of kernels between training data. We set

$$\mathbf{k}_*(\mathbf{c}) = (k(\mathbf{x}_1(\mathbf{c}), \mathbf{x}_*(\mathbf{c})), \ldots, k(\mathbf{x}_p(\mathbf{c}), \mathbf{x}_*(\mathbf{c})))^{\mathrm{T}} \tag{3.10}$$

as the Gaussian kernel function as follows:

$$k(\mathbf{x}_i(\mathbf{c}), \mathbf{x}_*(\mathbf{c})) = \exp(-\beta|\mathbf{x}_i(\mathbf{c}) - \mathbf{x}_*(\mathbf{c})|^2), \tag{3.11}$$

where β represents the reciprocal of the scale of the Gaussian kernel.

GP regression results are greatly different by the two parameters β, σ. β in Eq. (3.11) is represented by a grey line in Fig. 3.1. This parameter determines the width of the Gaussian kernel of each learning data. If β is too small, overfitting will occur (Fig. 3.1(a)). If it is too large, the curve of the estimated value becomes too smooth, and the estimation accuracy decreases (Fig. 3.1(c)). σ is the variance of noise. As is the case with β, if σ is too small, overfitting will occur, and if it is too large, the curve of the estimated value becomes smooth, which leads to lower estimation accuracy. As described above, to avoid overfitting while maintaining estimation accuracy, it is necessary to tune these two parameters. In this research, we conduct a grid search of two parameters with CVE as the criterion.

3.3 Synthetic Data Analysis

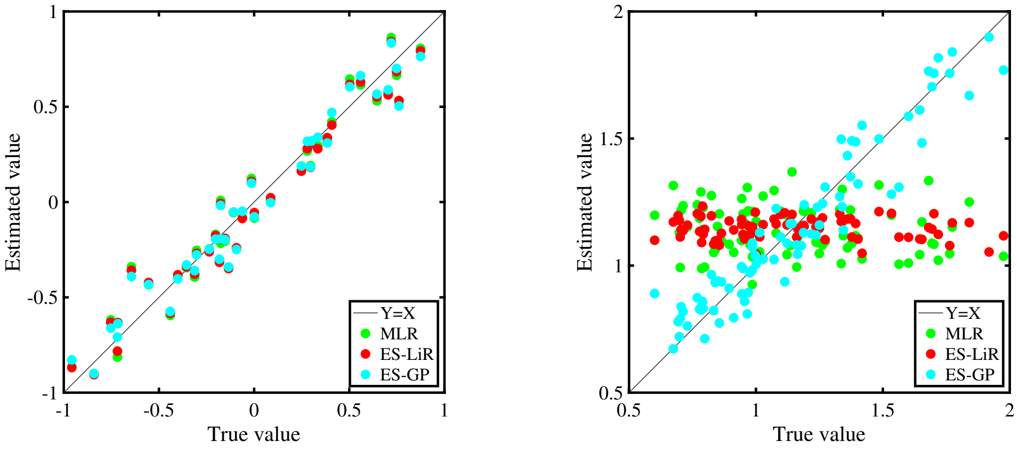
To make the experimental condition the same as that described in Chapter 2, we set the dimension of the descriptor of synthetic data $\mathbf{x} = \{x^1, ..., x^N\}$, $N = 10$, and uniformly sampled from the interval $[-1, 1]^{10}$. We then generated two types of objective variables, $\mathbf{y} = (y_1, \cdots, y_p)^{\mathrm{T}}$. The first one is linear data such that $y_i = x_i^1 + 1.5x_i^2 + \epsilon_i$, where i is the sample index and ϵ_i is sampled from a normal distribution with an average of 0 and a variance of 0.3. The second one is nonlinear data such that $y_i = \cos(x_i^1) + (x_i^2)^2 + \epsilon_i$ generated by $y_i = \cos(x_i^1) + (x_i^2)^2 + \epsilon_i$. The number of samples to be generated, p, is 80 to match the size of the simulation dataset in Chapter 2.

First, let us consider the results of linear synthetic data analysis, as shown in Fig. 3.2(a). The prediction errors of the results of MLR and ES-LiR are 0.11 and 0.10, respectively. The prediction error by GP is 0.11, and linear synthetic data can also be regressed by performing interpolation. Moreover, assuming that the coefficients are sparse, we apply ES-GP to the synthetic data and obtained the best indicator with a prediction error of 0.10, which is almost the same prediction error in ES-LiR.

Similarly, the results of nonlinear artificial data analysis are shown in Fig.3.2(b). In the case of nonlinear data, the prediction errors of MLR and ES-LiR are much larger than those in the linear data analysis since the linear regression model does not match the synthetic data well. Meanwhile, the prediction error of GP for the nonlinear synthetic data was approximately 37% lower than the ES-LiR prediction error, and thus, GP can regress data with more flexibility because there is no parametric shape of the unknown regression function considered in advance. Furthermore, the prediction error was 0.10 for ES-GP, and the prediction error was approximately 70% lower than that of ES-LiR since, if noise variables are included as descriptors, the prediction accuracy may decrease. Thus, selecting a few descriptors of GP and removing the noise variables greatly improve the interpolative prediction.

Figure 3.3 is a diagram that shows the result of variable selection by the ES method. From left to right, 25 CVE's lowest prediction error indicators are lined up, and the vertical axis shows the descriptors. Figures 3.3(a) and 3.3(c) present the result on linear data, and Figures 3.3(b) and 3.3(d) present the result on nonlinear data. The colour of each cell in Figures 3.3(a) and 3.3(b) represents the coefficient (weight) of that descriptor obtained as a result of the linear regression in that indicator. A white cell indicates that a descriptor is not selected in the indicator. Similarly, for Figures 3.3(c) and 3.3(d), white cells indicate that a descriptor is not selected in that indicator. The black cell represents the descriptor being used.

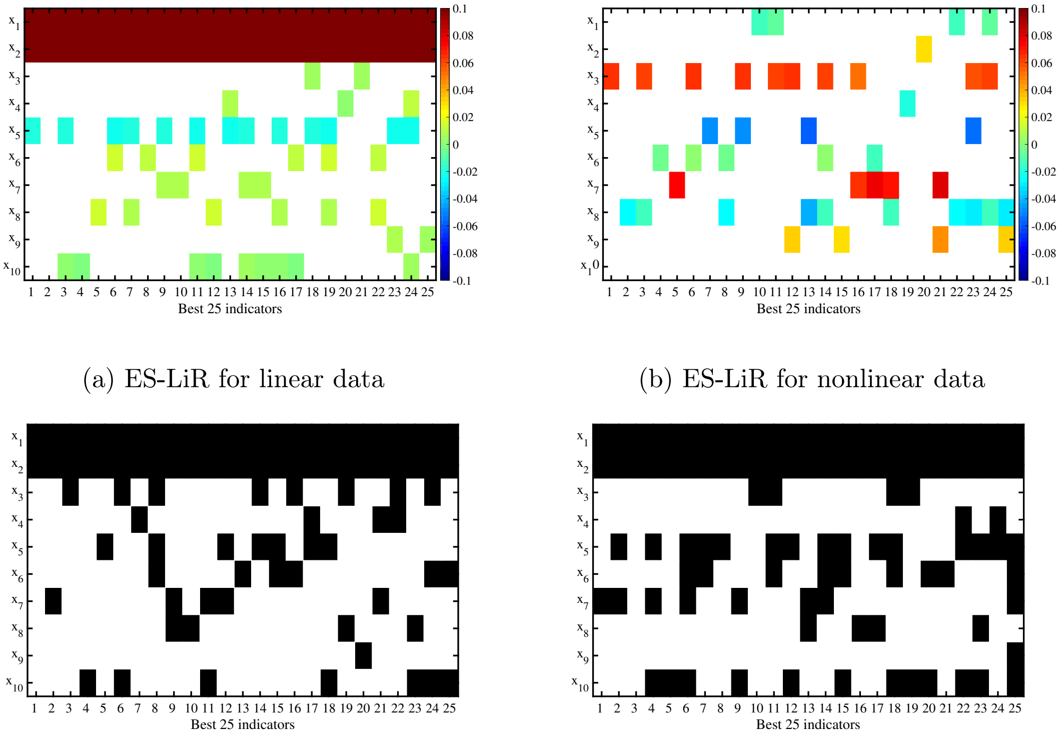
As shown in Figures 3.3(a) and 3.3(c) in linear data, x_1, x_2, which is a truly efficient descriptor for both ES-LiR and ES-GP, is the best 25. It is selected as important descriptors for predicting object variables consistently. However, in nonlinear data, ES-GP is chosen as a truly efficient descriptor, but ES-LiR is not. This result indicates the possibility that ES-LiR cannot select the dimension in nonlinear data.



(a) Linear data estimation

(b) Nonlinear data estimation

Figure 3.2: Generate $\mathbf{x} = \{x^1, \cdots, x^{10}\}$ uniformly from $[-1, 1]^{10}$ and generate $Y = x^1 + 1.5x^2 + \epsilon$ using 80 samples of nonlinear data using the relationship $y = \cos(x^1) + (x^2)^2 + \epsilon$. Here, ϵ is sampled from a normal distribution with average 0 and variance 0.3. (a) shows the result of linear data, and (b) shows the result of regression using MLR, ES-LiR, and ES-GP for nonlinear data. The horizontal axis is the true value, the vertical axis is the predicted value, and the black line is the graph of $y = x$. It can be said that the estimation accuracy is higher as each point is closer to $y = x$.



(a) ES-LiR for linear data

(b) ES-LiR for nonlinear data

(c) ES-GP for linear data

(d) ES-GP for nonlinear data

Figure 3.3: Variable selection by weight diagram. (a) and (c) are linear data, and (b) and (d) are the results of regression on nonlinear data. (a) and (b) show the coefficients of each dimension when ES-LiR is performed, and (c) and (d) show the variables used when performing ES-GP in black cells there. A white cell represents an unused variable. The vertical axis is the descriptor, and the horizontal axis is the order of the indicators in the order of the lowest CVE in order from the left.

3.4 Simulation Data Analysis

3.4.1 Prediction of coordination energies for Li-ion electrolyte

Database

The coordination energy "$\mathrm{E_{coord}}$" is the difference between the energy of the Li-solvent complex "E(Li-solvent)" and the total energy of the solvent molecule "E(solvent)" and the Li-ion "E(Li-ion)". The coordination energy "$\mathrm{E_{coord}}$" is defined by the following formula:

$$\mathrm{E_{coord}} = \mathrm{E(Li\text{-}solvent)} - \{\mathrm{E(solvent)} + \mathrm{E(Li\text{-}ion)}\}. \tag{3.12}$$

We thus constructed a learning database that includes the coordination energy, which expresses "performance", and the five feature values obtained through DFT calculations and the five feature values of physical properties (boiling point, melting point, flash point, density of solvent, and molecular weight) on the catalogue. For data preprocessing, 21 types of data with missing descriptors or the performance are removed, and data of a total of $p = 82$ solvents are used as the data for analysis. In addition, we standardized the data because the scales of the descriptors are different. The standard value is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. For predicting the normalized coordinate energy, we used a database that includes sample size $p = 82$ and $N = 10$ descriptors, as shown in Table 3.1.

Let us introduce the formation of symbols representing this database. We set the normalized coordination energies $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_p]^{\mathrm{T}}$ and the i-th sample's descriptors $\mathbf{x}_i = [x_i^1, \cdots, x_i^N]$ as the target objects and $N = 10$ descriptors defined in Table 3.1, respectively. These formations are summarized as follows:

$$\mathbf{x}_i \equiv [x_i^1, \cdots, x_i^N], \tag{3.13}$$

$$\mathbf{X} \equiv \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_p \end{pmatrix}, \tag{3.14}$$

$$\mathbf{y} \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}, \tag{3.15}$$

where $\mathbf{X}$ represents the descriptor database, which consists of rows of respective sample data $\mathbf{x}_i$.

Results

Figure 3.4 shows the fitting results of MLR, ES-LiR, GP and ES-GP. The vertical axis shows the predicted value and the horizontal axis shows the true value. Compared to MLR, which uses all variables, ES-LiR reduced CVE by 14%. In the case of GP, a nonlinear regression technique, the CVE of ES-GP was reduced by 24% compared to the CVE of regular GP, which uses all variables. Comparing MLR and ES-GP, CVE was reduced by 52%.

Table 3.1: Correspondence between symbols and features.

Descriptor	Feature
x_i^1	boiling point (BoilingP)
x_i^2	density of solvent (Density)
x_i^3	dipole moment (Dipole)
x_i^4	flash point (FlashP)
x_i^5	HOMO
x_i^6	LUMO
x_i^7	melting point (MeltingP)
x_i^8	molecular weight (MolecularW)
x_i^9	Mulliken charge (MullikenO)
x_i^{10}	Li-O distance($\mathrm{R_{LiO}}$)

Figure 3.4 describes the best results of ES-GP and ES-LiR with the minimum prediction error, but the ES-GP and ES-LiR verify the prediction error for all combinations of descriptors and the statistical verification of the prediction errors is available, as shown in Fig. 3.5. Comparing the histogram of ES-LiR and the histogram of ES-GP, we can see that the average value of CVE of ES-GP is 8.30 and the average value of CVE of ES-LiR is 9.92. Therefore, the prediction error of ES-GP is 16% smaller than the prediction error of ES-LiR. These results show that nonlinearity of the regression model is important for predicting coordination energy from the simulation database used in this study even when compared statistically.

As noted in Section 3.3, the ES method achieves high prediction accuracy by extracting some variables that are essentially important and performing regression. In this data, the prediction accuracy was improved by sparsification. This suggests that there are only a few descriptors in the whole that are important for the prediction of the coordination energy.

The ES method not only minimizes the CVE but also derives the CVE in all combinations, so you can see the whole picture of them. Using the whole pictures, the ES-LiR and ES-GP method can be used to construct the weight diagram, which shows the top 25 best combinations of the descriptors, as shown in Figure 3.6. The weight diagram reveals the stability of the important descriptors for the prediction, even if the error is at the same level as the other methods. In the weight diagram of ES-LiR (Figure 3.6(a)), each color represents the fitted coefficient of each descriptor, which shows the importance for the coordination energy prediction. In the weight diagram of ES-GP (Figure 3.6(b)), Because GP does not compute descriptor weights, the cells of used descriptors are colored black. The white-blocks of the map correspond to the descriptorsak which are not adopted for the prediction.

First, let us explain the interpretation of the results of the weight diagram of ES-LiR. From the weight diagram of ES-LiR, the Mulliken charge (MullikenO) is the significant descriptor for the coordination energy prediction and flashing point (FlashP.), and $\mathrm{R_{LiO}}$ can also contribute to it. The coordination energy is highly affected by the Coulomb interaction between the Li cation and the oxygen atom that has a negative electron charge. Thus, the extraction of the Mulliken charge as a good descriptor fits our chemical intuition, even if the Mulliken charge values are sometimes quantitatively not stable with the basis functions. The $\mathrm{R_{LiO}}$ is also a trivial descriptor for the estimation of the solvation energy because the distance corresponds to the strength of the interaction between Li and O. On the other

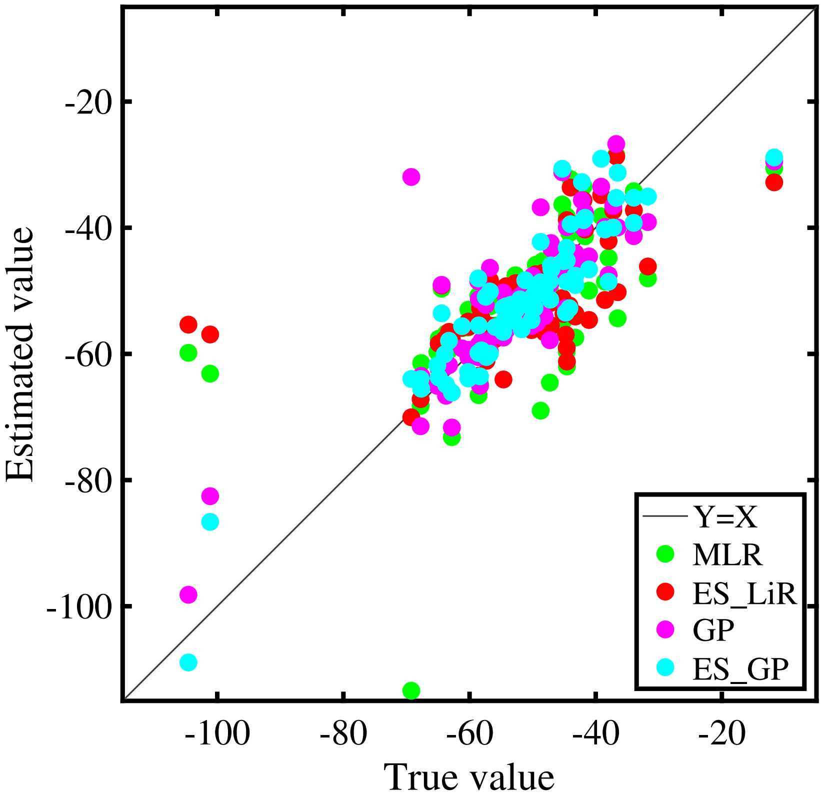


Figure 3.4: A comparison of the prediction results of MLR, ES-LiR, GP and ES-GP. The horizontal axis is a true value, and the vertical axis is an estimated value. The diagonal line represents $y = x$, and the closer to this line, the higher the prediction accuracy is. The prediction error of all variable MLR was 10.20 (kcal/mol), the prediction error of ES-LiR was 8.78 (kcal/mol), the prediction error of GP was 6.42 (kcal/mol), and the prediction error of ES-GP was 4.89 (kcal/mol).

hand, the flashing point is not a trivial descriptor. It might be a weak relationship between "the oxygen radical reaction for burning" and "the Li cation–solvent interaction", though the number of the samples should be increased for such a discussion.

On the other hand, as shown in Figure 3.6(b), ES-GP selected different variables than ES-LiR. For example, the mulliken charge, which is selected for all the top 25 indicators in the ES-LiR, is selected for only 14 out of the top 25 indicators in the ES-GP. The boiling point, dipole moment and HOMO energy are rarely selected for ES-LiR, but they are frequently selected for ES-GP. As shown in Fig. 3.3(a), ES-GP can extract descriptors in linear relations in addition to descriptors in nonlinear relations. From this fact, considering nonlinearity, HOMO energy, boiling point, and dipole moment are more likely to predict coordination energy than Mulliken charge.

Although the selected descriptors of ES-LiR, as shown in Fig. 3.6(a), are easier to use because of simplicity, the physical meaning is also important when extracting features is performed, in ES-GP, it is generally difficult to understand all the physical phenomena

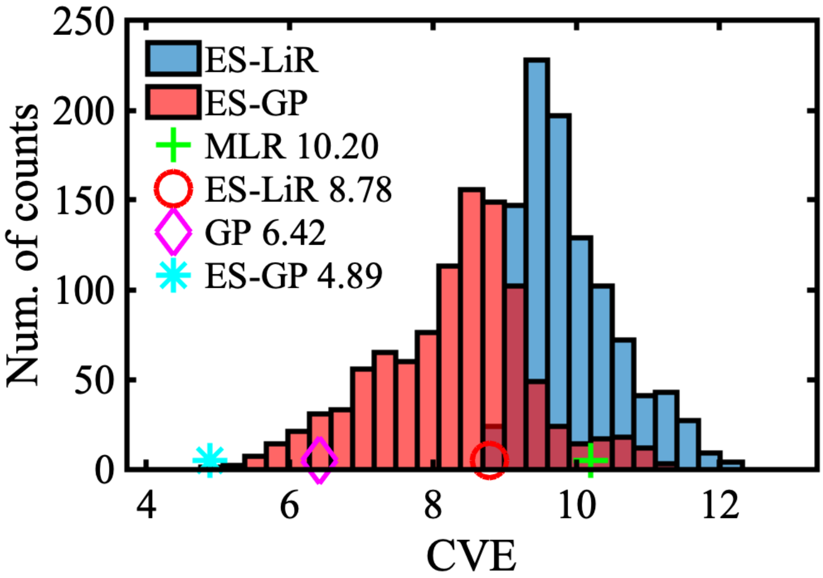
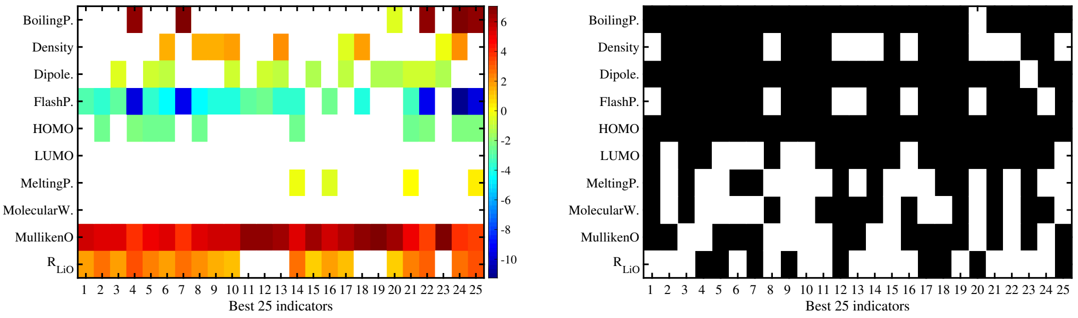


Figure 3.5: Histogram of ES-GP and ES-LiR indicator. The horizontal axis represents CVE, and the vertical axis represents the number of indicators. The circle represents the CVE minimum value of ES-LiR, the diamond represents the CVE value of GP, and the asterisk represents the CVE minimum value of ES-GP.

behind extracted features under the linear relation assumption. This result shows that there are nonlinear relationships behind the selected descriptors and desired properties, and further investigation is need to determine the physical mechanism behind this. Other possibilities of these results indicate that there are several relationships between the selected descriptors and desired properties and that the combination of these phenomena results in the complex relationship.

It is possible that ES-GP is greatly over learning the result of the weight diagram change between ES-LiR and ES-GP. Therefore, to verify the results of this research, we conducted a verification with Y-Scrambling [54]. Y-Scrambling is a method of randomly shuffling the target variable Y in the learning data to learn random training data without changing the distribution of Y to validate the model. In the conventional ES method, validity was not verified, but in this study, we compare each density of state (DoS) with the ES method and confirm the extent to which the ES method can extract information. Figure 3.7 compares their DoS. The average CVE of the normal ES-GP is 8.23 kcal/mol, while the average CVE of Y-scrambling ES-GP is 11.6 kcal/mol. The average CVE of the normal ES-GP is 29% lower than the average CVE of Y-Scrambling ES-GP. Therefore, it is clear that ES-GP can extract information.



(a) ES-LiR

(b) ES-GP

Figure 3.6: Comparison of weight diagrams between ES-LiR (a) and ES-GP (b). There are 25 low indicators of CVE in order from left to right. The vertical axis is each descriptor. The colour of each cell in (a) represents the coefficient of each descriptor when linear regression is performed. A white cell is an unselected descriptor. In (b), the black cell is the selected variable, and the white cell is the unselected variable. The optimal hyper parameters of the best indicator of ES-GP are $\sigma = 8.22 \times 10^{-6}$ and $\beta = 4.35 \times 10^{-6}$.

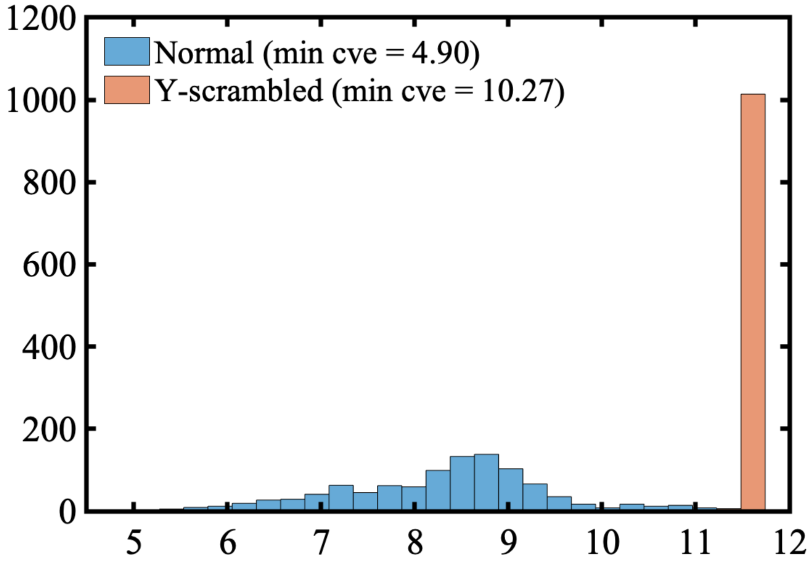


Figure 3.7: Histogram of ES-GP analysis of raw data and shuffled data for verification of ES-GP results by Y-Scrambling. The horizontal axis represents CVE, and the vertical axis represents the number of indicators.

Table 3.2: Correspondence between symbols and features.

Descriptor	Feature	Experimental / Computational	Cation / Solvents
x_i^1	ionic radius	Experimental	Cations
x_i^2	electronegativity	Experimental	Cations
x_i^3	atomic weight	Experimental	Cations
x_i^4	NBO charge (NBOchargeOatom)	Computational	Solvents
x_i^5	HOMO energy	Computational	Solvents
x_i^6	LUMO energy	Computational	Solvents
x_i^7	dipole moment (TotalDipole)	Computational	Solvents
x_i^8	total energy	Computational	Solvents
x_i^9	boiling point	Experimental	Solvents
x_i^{10}	flashing point	Experimental	Solvents
x_i^{11}	melting point	Experimental	Solvents
x_i^{12}	molecular weight	Experimental	Solvents
x_i^{13}	density of solvent (Density)	Experimental	Solvents

3.4.2 Prediction of coordination energies for alkali group elements

Database

For the descriptors or explanatory variables, the following were used as 'computational' descriptors: energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), dipole moment, natural bond orbital (NBO) charge of the O atom that coordinates to the metal ion, total energy (i.e. electronic energy plus nuclear repulsion), and total dipole moment. From an atomic/molecular perspective, the ion-solvent interaction can be understood as an acid–base interaction, since the ion works as a hard acid and the solvent works as a hard or soft Lewis base. Common organic electrolyte solvents have alkoxy or carbonyl groups, and in these cases the O atom works as the Lewis base site. For this reason, we assumed that the ion coordinated to this O atom. Also, the NBO charge on the coordinating O atom was included in the descriptors. For the optimized geometries of the cation-coordinated system, see Figure B.1 in the Appendix. The computational properties of the solvent are obtained by DFT calculation of the pure solvent, i.e. without ions. All the experimental and computational descriptors for the solvent molecules are shown in Table 3.2.

First, we discuss the accuracy of the three methods to estimate the true (i.e. DFT-calculated) E_{coord} values. Here, the data set includes all the E_{coord} data (i.e. coordination of Li, Na, K, Rb, and Cs to solvent molecule). In other words, solvent descriptors and ion descriptors were independently made and combined to form the whole data set. Since we have 70 solvents, the E_{coord} data set consists of $5 \times 70 = 350$ points. Our calculated E_{coord} values for Li, Na, K, Rb, and Cs are summarized in the bar chart in Figure 3.8, and the selected numerical values for E_{coord} are shown in Table 3.3. The range of E_{coord} for the five ions are: Li -1.32 to -2.91 eV (mean value: -2.20 eV), Na -0.88 to -2.18 (-1.60), K -0.61 to -1.73 (-1.20), Rb -0.55 to -1.60 (-1.11), and Cs -0.46 to -1.44 eV (-0.98). Thus, the E_{coord} of metal ions can be ranked as Li $>$ Na $>$ K $\sim$ Rb $>$ Cs.

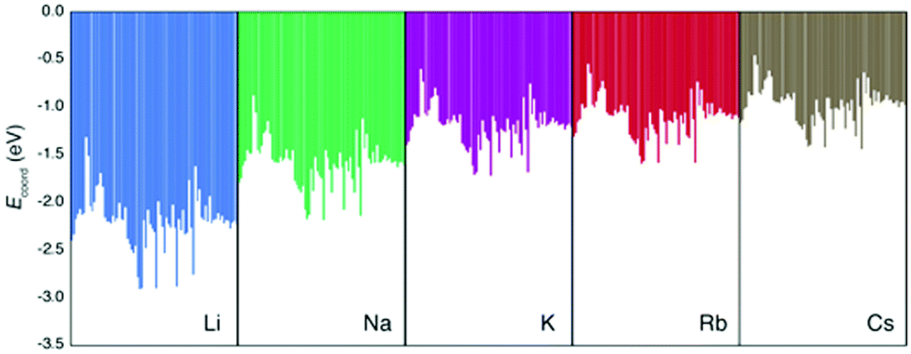


Figure 3.8: E_{coord} values of 70 solvents and five ions (Li, Na, K, Rb, and Cs).

	$E_{coord}(eV)$				
Solvent	**Li**	**Na**	**K**	**Rb**	**Cs**
Ethylene carbonate	-2.343	-1.747	-1.365	-1.272	-1.135
Propylene carbonate	-2.399	-1.789	-1.397	-1.307	-1.165
Vinylene carbonate	-2.179	-1.610	-1.246	-1.157	-1.025
Fluoroethylene carbonate	-2.128	-1.569	-1.210	-1.129	-1.001
Dimethyl carbonate	-2.068	-1.454	-1.078	-0.968	-0.842
Diethyl carbonate	-2.130	-1.492	-1.106	-1.010	-0.877
Ethyl methyl carbonate	-2.114	-1.488	-1.108	-1.006	-0.878
Furan	-1.320	-0.884	-0.605	-0.545	-0.461
Tetrahydrofuran	-2.047	-1.454	-1.065	-0.978	-0.851
Ethyl acetate	-2.206	-1.574	-1.185	-1.083	-0.950
Isopropyl acetate	-2.222	-1.585	-1.187	-1.093	-0.958
Methyl propionate	-2.138	-1.524	-1.133	-1.030	-0.896
Methyl formate	-2.011	-1.444	-1.082	-0.981	-0.861
Vinyl acetate	-2.052	-1.454	-1.076	-0.984	-0.857
Sulfolane	-2.481	-1.879	-1.450	-1.350	-1.200
Dimethyl sulfoxide	-2.905	-2.183	-1.725	-1.590	-1.427
Cyclohexanone	-2.259	-1.654	-1.265	-1.158	-1.025
Benzaldehyde	-2.177	-1.570	-1.188	-1.085	-0.958
Benzyl benzoate	-2.758	-2.139	-1.682	-1.591	-1.441
Diphenyl ether	-1.625	-1.120	-0.758	-0.738	-0.638
Acetone	-2.190	-1.600	-1.219	-1.117	-0.987
Chloroacetone	-1.938	-1.399	-1.047	-0.964	-0.845
Methyl acrylate	-2.195	-1.570	-1.178	-1.069	-0.938

Table 3.3: DFT-calculated E_{coord} of Li, Na, K, Rb, and Cs for 23 selected solvents.

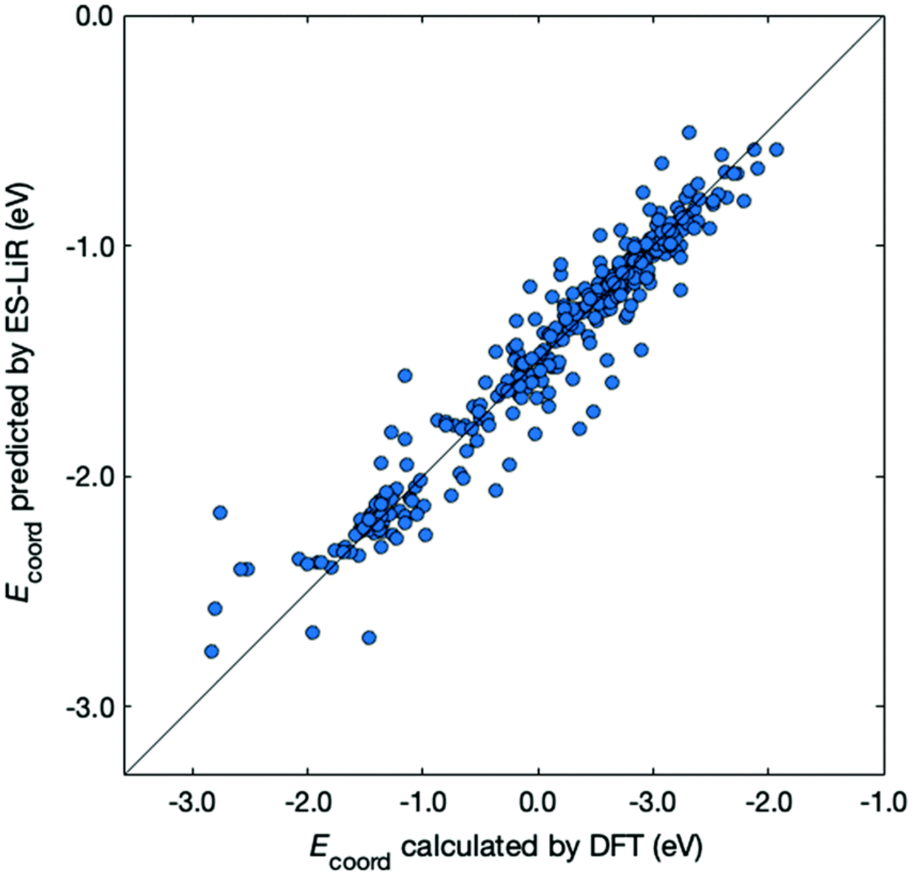


Figure 3.9: Comparison between E_{coord} calculated by DFT (x-axis) and that predicted by ES-LiR (y-axis). The diagonal line corresponds to a perfect match.

Results

Next, we examined the regression of E_{coord} from the solvent and ion properties. Figure 3.9 demonstrates a good correlation between E_{coord} values calculated by DFT and those estimated by ES-LiR. The CV error for ES-LiR in Figure 3.9 was 0.127 eV. This is only 5.7 % for the average Li coordination energy, indicating that the regression formula from ES-LiR gives accurate results. We also observe that the prediction accuracy tends to be lower at $E_{coord} < -2.5$ eV. As we shall see later, the important descriptors are the O charge and the total dipole. The deviation from this regression formula indicates other effects, for example, large distortion of the ion–solvent complex would contribute to large E_{coord} values.

The accuracy of the estimation methods can be evaluated by the CV errors. The smallest CV error calculated with the MLR, LASSO, and ES-LiR methods was 0.1280, 0.1278, and 0.1271 eV, respectively. These values are shown in Table 3.4, together with selected combinations of descriptors. Values in Table 3.4 suggest that ES-LiR gives the smallest CV error and thus the best prediction accuracy, although the differences between the three methods are moderate. It is well known that the CV error is intimately related to the choice of descriptors. Since the ES-LiR examines all combinations of descriptors, it is always guaranteed to choose the best combination. In all three regression formulae, the ionic radius of the metal ion has the largest coefficient and thus it is the most important descriptor. This can be understood in terms of Pearson's hard–soft acid–base rule,

	MLR	**LASSO**	**ES-LiR**
Ionic radius	0.6637	0.6542	0.6637
Electronegativity	0.1612	0.1569	0.1612
Atomic weight	-0.0986	-0.0930	-0.0986
NBO charge of Oatom	0.1832	0.1751	0.186
HOMO energy	0.0121	0.0111	0
LUMO energy	0.026	0.0248	0.0273
Total dipole	-0.1467	-0.1420	-0.1475
Total energy	-0.1384	-0.1261	-0.1476
Boiling point	-0.0956	-0.0941	-0.0977
Flashing point	0.1154	0.1034	0.1182
Melting point	-0.0202	-0.0151	0
Molecular weight	-0.1156	-0.1051	-0.1215
Density	0.0249	0.027	0
CV error	0.128	0.1278	0.1271

Table 3.4: Coefficient of descriptors in the three regression formulae (MLR, LASSO, and ES-LiR) with the smallest CV error, and their CV errors

which states that the smaller ion has hard acid character. The positive coefficient of ionic radius in Table 3.4 indicates that smaller ions give the smaller E_{coord} values (thus the stronger ion–solvent interaction). After the ionic radius, the NBO charge on the O atom coordinating to the ion has the second largest coefficient. Since the ion–solvent interaction mainly has an electrostatic cationic–anionic character, a more negative O charge leads to a stronger interaction and thus a larger E_{coord} value. This conclusion is the same as in Section 3.4.1, in which the O atomic charge is the most important descriptor for the Li coordination on electrolyte solvent molecules. We also found that the total dipole has a relatively large coefficient. This adds to the charge–charge electrostatic interaction via charge–dipole interaction, so this also contributes to the ion–solvent interaction.

Another important difference among the three regression methods is the sparseness of the regression formula. In MLR and LASSO, all descriptors have some non-zero coefficients, and thus these methods are the least sparse among the three. Contrary to these two methods, ES-LiR gives a more sparse regression formula because three descriptors (HOMO energy, melting point, and density) have zero coefficients. This indicates that the regression formula given by ES-LiR is the most accurate of the three methods, and at the same time its physical and chemical meanings are the easiest to interpret.

Up to now, our discussion is based on the optimal combination of descriptors that minimize the CV error. Estimation accuracy for other descriptor combinations can also be found using the ES-LiR, because this method examines all combinations of descriptors. The number of counts in the descriptor combination within a fixed CV error range can be summarized by the histogram in Figure 3.10, where descriptor combinations that reduce CV error to below 0.14 are rather rare. From this, we can infer that the combination of particular descriptors is important for achieving accuracy.

This issue can be analyzed with the linear coefficient of the accurate regression formula. This is another important piece of information obtained by ES-LiR. The plot of linear coefficients for ten descriptor combinations that give low CV errors is shown in weight diagram in Figure 3.11. Since we can find the contribution of descriptors for several combinations of them, the stability of the important descriptors can be found

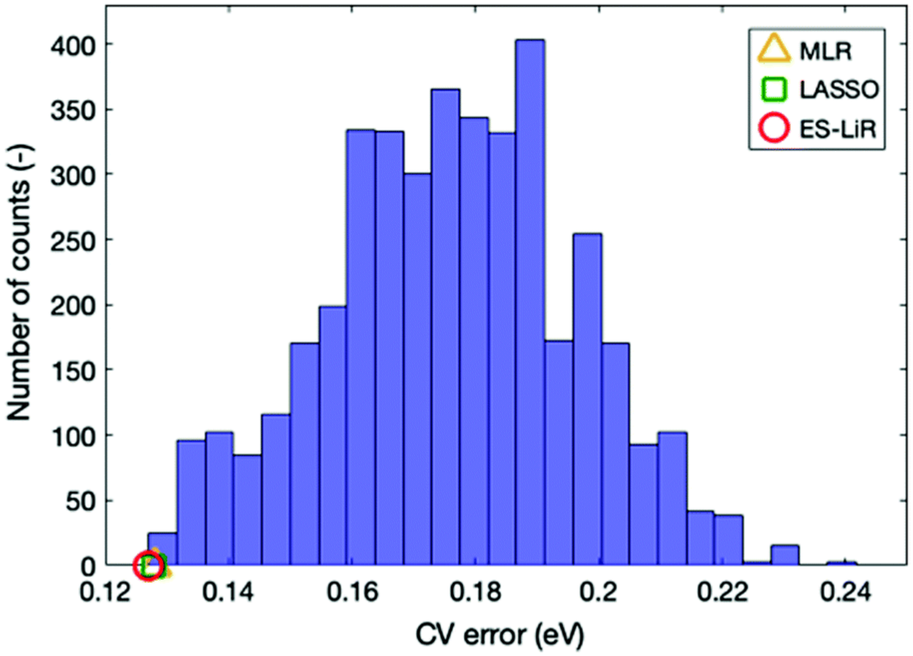


Figure 3.10: The number of counts for the CV error (i.e. histogram) for various descriptor combinations. The orange, green, and red symbols show the smallest CV errors for ES-LiR, LASSO, and MLR, respectively.

from the weight diagram. We consider that analysis with several regression formulae is important, because multicollinearity often occurs in the linear regression model; inspecting the descriptor weights for multiple combinations of regression models is more robust than analysis based on a single regression model.

In the weight diagram, the ionic radius has the largest contribution to the regression formula in all descriptor combinations. Thus, this property is the most important and also most stable descriptor in the E_{coord} prediction, as stated above. Since the ionic radius is the most important descriptor in all top 20 descriptor combinations, it is also the most stable one in the present descriptor set. The next important descriptor is the NBO charge of the coordinating O atom, which is also a stable descriptor among the 20 combinations.

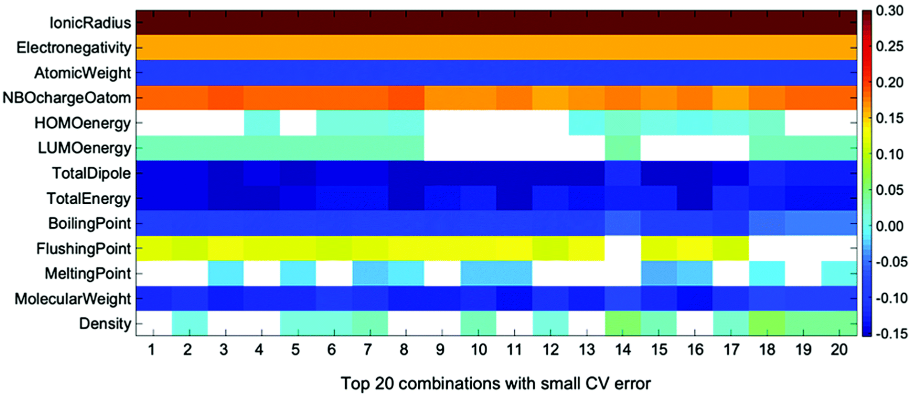


Figure 3.11: Weight diagram for the descriptors of top 20 combinations with small CV error in ES-LiR. Descriptors with coefficients smaller than 10^{-10} shown in white box.

Other descriptors, such as dipole moment, boiling point, and density, are also important, but their stability is not as high as the ionic radius or the solvent O NBO charge.

We also note that the atomic weights of cation species have large weight. The atomic weight works as a secondary factor for the ionic radius, as can be confirmed by carrying out the ES-LiR without the ionic radius; in this case the atomic weights have the largest weight in the regression formula. However, the calculated CV error is considerably higher (0.2807 eV), indicating that the ionic radius does much better in the linear regression model.

Finally, we applied the ES-GP method for E_{coord} prediction. This includes the nonlinear terms of the descriptors, which were not taken into account in the ES-LiR method. According to this feature, we can expect higher prediction accuracy with ES-GP, which was already shown in Section 3.4.1. Here, the same data set used for ES-LiR was used for ES-GP. We used the following seven descriptors in the ES-GP; ionic radius, NBO charge, total dipole moment, total energy, boiling point, melting point, and density. We selected these descriptors as they minimize the CV error of the ES-GP prediction; the dependence of the CV error on the number of descriptors is shown in Figure B.2 in the Appendix.

In Figure 3.12, we compare the E_{coord} values calculated by DFT and predicted by ES-GP. The CV error for ES-GP was 0.016 eV, which is significantly better than that for ES-LiR (0.127 eV). The accuracy of the ES-GP method is 1.54 in kJ mol-1 unit, which is sufficient for most purposes for battery-related study. From these results, we can conclude that the combined use of ES-LiR and ES-GP is advantageous in obtaining good physical or chemical intuition and achieving high prediction accuracy.

3.5 Conclusion

In this study, data-driven science techniques were applied to a database obtained by computational chemistry in order to find new electrolyte materials.

We first developed a prediction model for the coordination energy in the LIB electrolyte. We used MLR, LASSO, GP, ES-LiR, and ES-GP to predict the coordination energy of LIBs. As a result, the coordination energies can be predicted from more computationally inexpensive descriptors. Compared with each method, ES-GP reduced the prediction error by 52% compared to MLR. This demonstrates the importance of sparsification and nonlinearization in predicting the coordination energy. The prediction model based on ES enables us to extract the combinations of important descriptors for the prediction of the coordination energy by using weight diagrams. The weight diagram allows us to choose the balance between descriptor data acquisition cost and prediction accuracy. This feature is general for all the material exploring studies with virtual screening. This treatment can be a key technique to future material searches.

Next, we created a similar database with alkali metals other than Li, and carried out the same verification for the database. It is now possible to predict the coordination energies of various ions using only the properties of the ions and solvents. The CVEs of ES-LiR and ES-GP are 0.127 eV and 0.016 eV respectively, achieving very high prediction accuracy.

This study has shown that combined use of computational chemistry and data-driven science can be an efficient and accurate tool for coordination energy prediction. We succeeded in showing that this approach can be applicable to any alkali metal ion

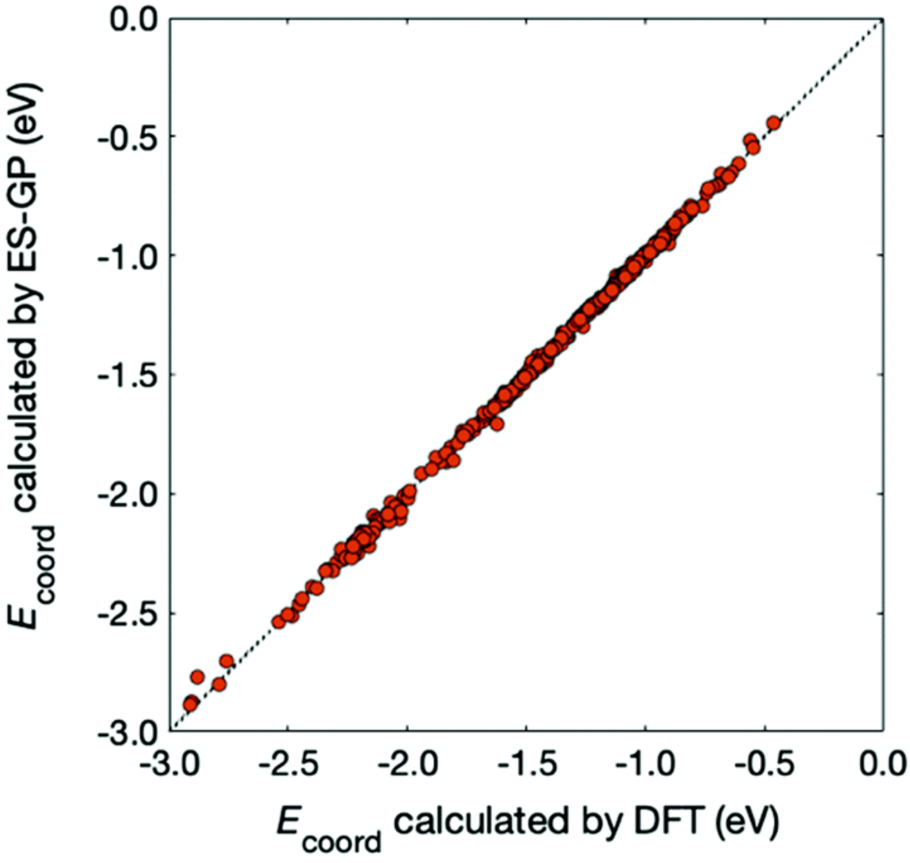


Figure 3.12: Comparison between E_{coord} calculated by DFT (x-axis) and predicted by ES-GP (y-axis). The diagonal line corresponds to a perfect match.

coordination. The constructed regression models are accurate enough for practical use in the search for battery electrolytes. These features will be important in developing post-Li next-generation batteries.

Chapter 4

Diffusion Coefficient Prediction

4.1 Introduction

The speed of charge and discharge is an important factor in the search for the electrolyte in secondary batteries. Faster charging would reduce the amount of time a smartphone being connected to a charging cable and the quicker discharge enable us to develop a vacuum cleaner with stronger suction or a powerful chainsaw. The charging and discharging mechanism of a LIB is as follows. Firstly, during the charging process, Li-ions with a positive charge are attracted to the negative electrode, and potential difference between the positive and negative electrodes occur. During discharge, electrons move from the anode to the cathode through the circuit to compensate for the potential difference, and Li-ions move to the cathode. In other words, charging and discharging are performed by the movement of Li-ions. Therefore, it is necessary to search for an electrolyte that allows Li-ions to diffuse more easily in order to create batteries with high charge/discharge speed. In recent years, improvements in computer performance have made it possible to simulate microscopic phenomena by first-principles calculations and to observe phenomena that cannot be observed by experiments. In the field of electrolyte exploration for LIB, Sodeyama et al. are studying the diffusion of Li-ions in the electrolyte by simulating the diffusion of Li-ions in the electrolyte using the DFT-MD (Density Functional Theory-Molecular Dynamics) method and investigating the diffusion speed (diffusion coefficient)[55]. However, even with large-scale computers, it takes weeks to months to calculate the properties of a single solvent molecule.

In this study, we aim to predict the value of the diffusion coefficient by using a machine learning method. The DFT-MD calculations were performed on a large scale computer and the results of electrolyte simulations with 38 different solvents were obtained (Ref: Section 2.2). We used the 11 descriptors, which were used as explanatory variables in Chapter 2 and the values of the coordination energy as explanatory variables. Using a formula as a database for regression, we can formulate this database as follows. Let D be the number of explanatory variables and N be the number of observed solvents. Here $D = 11, N = 38$. Let $X = (\mathbf{x}_n)_{n=1,\cdots,N}$ be the explanatory variable for the observed data, and $Y = (y_n)_{n=1,\cdots,N}$ be the objective variable (diffusion coefficient) for the observed data. $\mathbf{x}_n$ represents the explanatory variable of the n-th data observed and $\mathbf{x}_n = (x_n^1, \cdots, x_n^D)$. Explanatory variables and features correspond to each other as shown in Table 4.1. The goal of this study is to build a prediction model that explains the relationship between this explanatory variable X and the objective variable Y.

Descriptor	Feature
x_n^1	Molecular weight
x_n^2	Density of solvent
x_n^3	Boiling point
x_n^4	Melting point
x_n^5	Flashing point
x_n^6	LUMO energy
x_n^7	HOMO energy
x_n^8	Dipole moment
x_n^9	Coordination energy
x_n^{10}	Distance between Li and O
x_n^{11}	Mulliken charge

Table 4.1: Correspondence between symbols and features.

We have tried two main models for predicting the diffusion coefficient. The first method is to make predictions with a single model. Specifically, it is a prediction model in which the predictions are output as a single value, such as the linear regression of all variables, GP, ES-LiR and ES-GP, which were discussed in Chapter 3. The second method is to make predictions using mixture models. In materials science, the properties of materials can occur through different mechanisms. Even if we use the same explanatory variables to make predictions in such cases, there is a limit to the accuracy of the predictions. In this study, we used Sparse Linear Mixture Model (SpLMM) [26, 56] to build the model in the different spaces of explanatory variables. The results show that the mixture model is better suited to explain the data, as the log loss is reduced by about 80% in the mixture model compared to the method predicted by a single model.

This chapter is outlined as follows. In Section 4.2, we first describe the Sparse Linear Mixture Model used as a prediction model. Next, we explain the logarithmic loss, which is criterion for comparison with the linear regression, GP, ES-LiR, and ES-GP used in Chapter 3. In Section 4.3, we describes the results, and in Section 4.4, we discusses and concludes the results.

4.2 Method

4.2.1 Sparse Linear Mixture Model

The data of $\mathcal{D} = (\mathbf{x}_n, y_n)_{n=1,\cdots,N}$ is set as the observation data, where $\mathbf{x}_n = (x_n^1, \cdots, x_n^D)(n = 1, \cdots, N)$ is an explanatory variable of the D dimension and $y_n(n = 1, \cdots, N)$ is the objective variable. Assuming that the model has a observed noise ϵ following a normal distribution $\mathcal{N}(0, \sigma^2)$, the relationship between the output y and the explanatory variable $\mathbf{x}$ is represented as follows using a learning machine $f(\mathbf{x}; \mathbf{w})$ with a training parameter $\mathbf{w}$.

$$y = f(\mathbf{x}; \mathbf{w}) + \epsilon.$$

In the linear model, it is expressed as follows.

$$f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x},$$
$$\mathbf{w} = (w_1, \cdots, w_N)^{\mathrm{T}}.$$

Expressing this as the conditional probability of y, we get the following.

$$p(y \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(y \mid f(\mathbf{x}; \mathbf{w}), \sigma^2).$$

In the case of the mixed model, assume that the conditional probability of y is a weighted sum of multiple normal distributions, as shown below.

$$p(y \mid \mathbf{x}, \{\mathbf{w}_k\}_{k=1}^K, \boldsymbol{a}, K) = \sum_{k=1}^K a_k \mathcal{N}(y \mid f(\mathbf{x}; \mathbf{w}_k), \sigma^2).$$

where K is the number of mixtures of the model, $\boldsymbol{a} = \{a_k\}_{k=1,\cdots,K}$ is the mixture ratio, which satisfies $\sum_{k=1}^K a_k = 1, a_k \geq 0 (k = 1, \cdots, K)$, and $\mathbf{w}_k (k = 1, \cdots, K)$ are parameters which are trained in each model. Additionally, in the case of Sparse Linear Mixture Model (SpLMM), the indicator vector $\mathbf{c}_k (k = 1, \cdots, K)$ introduced in Chapter 3 are also estimated. Thus, the trainer in the k-th model is represented as follows.

$$f(\mathbf{x}; \mathbf{w}_k, \mathbf{c}_k) = (\mathbf{w}_k \circ \mathbf{c}_k)^{\mathrm{T}} \mathbf{x}.$$

where $\circ$ is the adamant product. Let Θ be all parameters of SpLMM, which means $\Theta = \{\mathbf{w}_k, \mathbf{c}_k, \boldsymbol{a}_k\}_{k=1,\cdots,K}$. The conditional probability of y is as follows.

$$p(y \mid \mathbf{x}, \Theta) = \sum_{k=1}^K a_k \mathcal{N}(y \mid f(\mathbf{x}; \mathbf{w}_k, \mathbf{c}_k), \sigma^2).$$

In this study, we assume that the prior distributions of $\{\mathbf{w}_k\}_{k=1}^K, \{\mathbf{c}_k\}_{k=1}^K, \boldsymbol{a}$ are independent of each other. That is, $p(\Theta) = p(\boldsymbol{a}) \prod_{k=1}^K p(\mathbf{w}_k) p(\mathbf{c}_k)$. Additionally, we assume that the prior distribution of $\mathbf{w}_k$ is a multivariate normal distribution with a mean vector $\mathbf{0}$, covariance matrix $\boldsymbol{\Sigma}$, the prior distribution of $\mathbf{c}_k$ is a distribution where the probability of the d-th element c_d becoming $0, 1$ is $\mu_d, (1 - \mu_d)$ respectively. The formula is as follows.

$$\begin{aligned} p(\mathbf{w}_k) &= \mathcal{N}(\mathbf{w}_k \mid \mathbf{0}, \boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{w}_k\right), \\ p(\mathbf{c}_k) &= \prod_{d=1}^D (\mu_d)^{c_d} (1 - \mu_d)^{1 - c_d}, \\ p(\boldsymbol{a} \mid K) &= \mathrm{Dir}(\boldsymbol{a} \mid \boldsymbol{\alpha}) \\ &= C(\boldsymbol{\alpha}) \prod_{k=1}^K a_k^{\alpha_k - 1}. \end{aligned}$$

where $C(\boldsymbol{\alpha})$ is a normalized term of the Dirichlet distribution and

$$C(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)}.$$

In this study, Σ was set to I (I is the unit matrix), $\mu_d = 1/2$ and each element of the $\boldsymbol{\alpha}$ is set to 1. The posterior distribution $p(\Theta \mid \mathcal{D}, K)$ is calculated as follows.

$$\begin{aligned}
p(\Theta \mid \mathcal{D}, K) &= p(\Theta \mid (\mathbf{x}_n, y_n)_{n=1,\cdots,N}, K) \\
&\propto \prod_{n=1}^{N} \{p(y_n \mid \mathbf{x}_n, \Theta, K)\}\, p(\Theta \mid K) \\
&= \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} a_k \mathcal{N}(y_n \mid f(\mathbf{x}_n; \mathbf{w}_k, \mathbf{c}_k), \sigma^2) \right\} p(\boldsymbol{a} \mid K) \prod_{k=1}^{K} p(\mathbf{w}_k) p(\mathbf{c}_k) \\
&\propto \prod_{n=1}^{N} \left\{ \sum_{k=1}^{K} a_k \exp\left(-\frac{(y_n - f(\mathbf{x}_n; \mathbf{w}_k, \mathbf{c}_k))^2}{2\sigma^2} \right) \right\} \prod_{k=1}^{K} a_k^{\alpha_k - 1} \exp\left(-\frac{1}{2} \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{w}_k \right).
\end{aligned}$$

In this study, we used Bayesian Free Energy (BFE) to determine the number of mixtures K.

The BFE is a negative logarithmic marginal likelihood $-\log(p(\mathcal{D} \mid K))$, which is defined as

$$-\log(p(\mathcal{D} \mid K)) = -\log\left\{ \int \prod_{n=1}^{N} \{p(y_n \mid \mathbf{x}_n, \Theta, K)\}\, p(\Theta \mid K) d\Theta \right\} + \text{const.}$$

Let $K_{\mathcal{D}}$ be K which minimizes BFE. That is, $K_{\mathcal{D}} = \operatorname{argmin}_K - \log(p(\mathcal{D} \mid K))$. We used exchange MCMC [57] to calculate the posterior distribution and BFE. Detail of the calculation method of the exchange MCMC is shown in Section C.1 in the Appendix.

To compare the degree of model fit with linear regression, GP, ES-LiR and ES-GP, we use log loss as criterion in this study. Log loss is the mean of the negative logarithm of the probability density of the predictive distribution of the test data $p(y_m \mid \mathbf{x}_m, \mathcal{D}^{\text{train}})$. Here, $\mathcal{D}^{\text{train}} = (\mathbf{x}_n, y_n)_{n=1,\cdots,N}$ is the training data and $\mathcal{D}^{\text{test}} = (\mathbf{x}_m, y_m)_{m=1,\cdots,M}$ is the test data. In other words, the formula is as follows.

$$\text{LogLoss} = -\frac{1}{M} \sum_{m=1}^{M} \log p(y_m \mid \mathbf{x}_m, \mathcal{D}^{\text{train}}).$$

In particular, in the case of SpLMM, it is as follows.

$$\begin{aligned}
\text{LogLoss}_{\text{SpLMM}} &= -\frac{1}{M} \sum_{m=1}^{M} \log p(y_m \mid \mathbf{x}_m, \mathcal{D}^{\text{train}}, K_{\mathcal{D}^{\text{train}}}) \\
&= -\frac{1}{M} \sum_{m=1}^{M} \log \int p(y_m \mid \mathbf{x}_m, \Theta, K_{\mathcal{D}^{\text{train}}}) p(\Theta \mid \mathcal{D}^{\text{train}}, K_{\mathcal{D}^{\text{train}}}) d\Theta.
\end{aligned}$$

In this study, 10-fold cross validation was used to calculate the mean of the log loss for each test data. Specifically, we divide $\mathcal{D}$ into 10 datasets $\mathcal{D}_t (t = 1, \cdots, 10)$ and create 10 pairs of train-test data $(\mathcal{D}_t^{\text{train}}, \mathcal{D}_t^{\text{test}}) = (\bigcup_{s \neq t} \mathcal{D}_s, \mathcal{D}_t)(t = 1, \cdots, 10)$. The following values were set as the log loss values.

$$\begin{aligned}
\text{LogLoss}_{\text{SpLMM}} &= -\frac{1}{N} \sum_{n=1}^{N} \log p(y_n \mid \mathbf{x}_n, \mathcal{D}_{t(n)}^{\text{train}}, K_{\mathcal{D}_{t(n)}^{\text{train}}}) \\
&= -\frac{1}{N} \sum_{n=1}^{N} \int p(y_n \mid \mathbf{x}_n, \Theta, K_{\mathcal{D}_{t(n)}^{\text{train}}}) p(\Theta \mid \mathcal{D}_{t(n)}^{\text{train}}, K_{\mathcal{D}_{t(n)}^{\text{train}}}) d\Theta
\end{aligned}$$

where $\mathcal{D}_{t(n)}^{\text{train}}$ denotes $\mathcal{D}_t^{\text{train}}$ such that $(x_n, y_n) \notin \mathcal{D}_t^{\text{train}}$.

4.3 Results & Discussion

In this chapter, we develop prediction models for the diffusion coefficient data and compare with five different methods: linear regression for all variables, GP for all variables, ES-LiR, ES-GP, SpLMM. Since the model selection (selection of K) of SpLMM was done by BFE in this study, model selection (variable selection) in ES-LiR and ES-GP was also done by BFE to maintain consistency.

First, the number K of mixture is estimated using the BFE. Figure 4.1 compares the BFEs for each mixture number $K = 1, 2, 3, 4$. The height of each bar represents the mean of the BFE obtained in each of the 10-fold CVs and the error bars represent the standard deviation. $K = 2$ is the minimum and the value of the BFE at $K \geq 2$ is much smaller than the value of the BFE at $K = 1$. The BFE of $K = 2$ was also minimal in each trial. This result shows that the mixture models of $K = 2$ explains the data best by using multiple models. it is possible that the values of the diffusion coefficients may occur through multiple mechanisms with different descriptor spaces. Detailed results of the BFE for each CV are presented in the Figure C.1 in the Appendix.

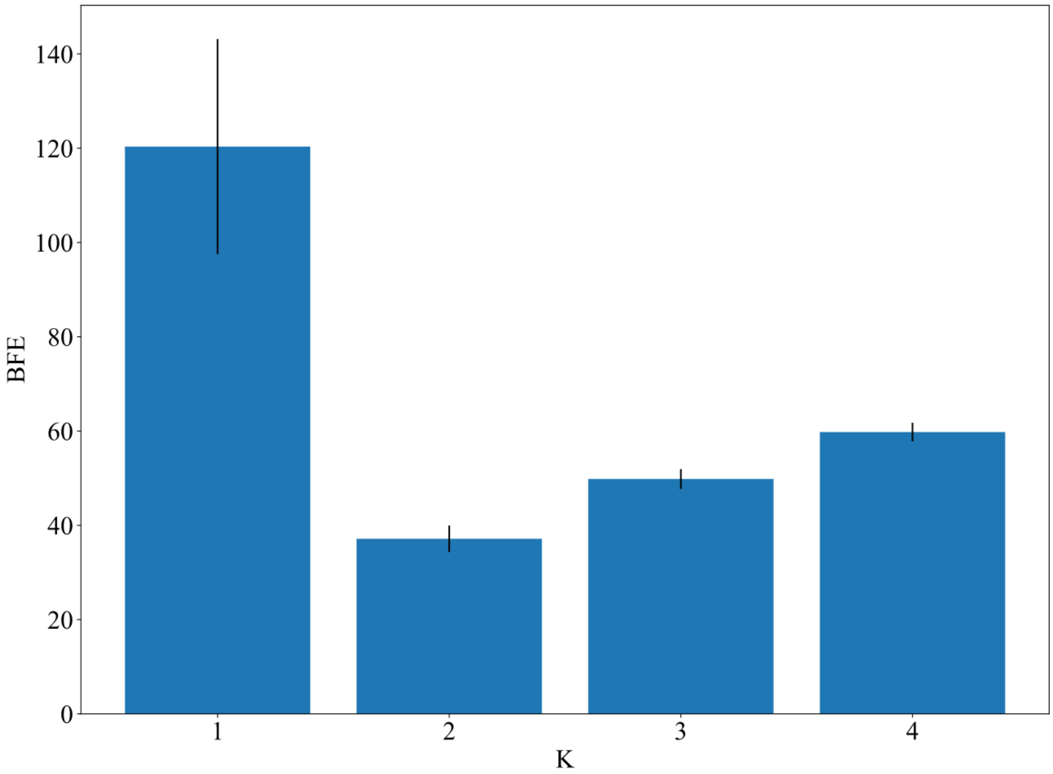


Figure 4.1: Comparison of Bayesian free energy at each K. The height of each bar represents the mean of the BFE obtained in each of the 10-fold CVs and the error bars represent the standard deviation. Compared to the BFE of $K = 1$, the BFE of $K = 2, 3, 4$ is much smaller, indicating that the model of $K \geq 2$ explains the data better than the model of $K = 1$.

Figure 4.2 shows a comparison of the log loss for each model. From left to right: linear regression of all variables, GP of all variables, best model of ES-LiR, best model of ES-GP, and the results of SpLMM for $K = 1, 2, 3, 4$. The linear regression of all variables, GP of all variables, ES-LiR, and ES-GP had log loss of $2.2 \sim 2.6$, while the value of SpLMM's log loss for $K \geq 2$ are $0.3 \sim 0.6$. As with Figure 4.1, this result shows that the model in $K \geq 2$ is a better fit to explain the data than the other models.

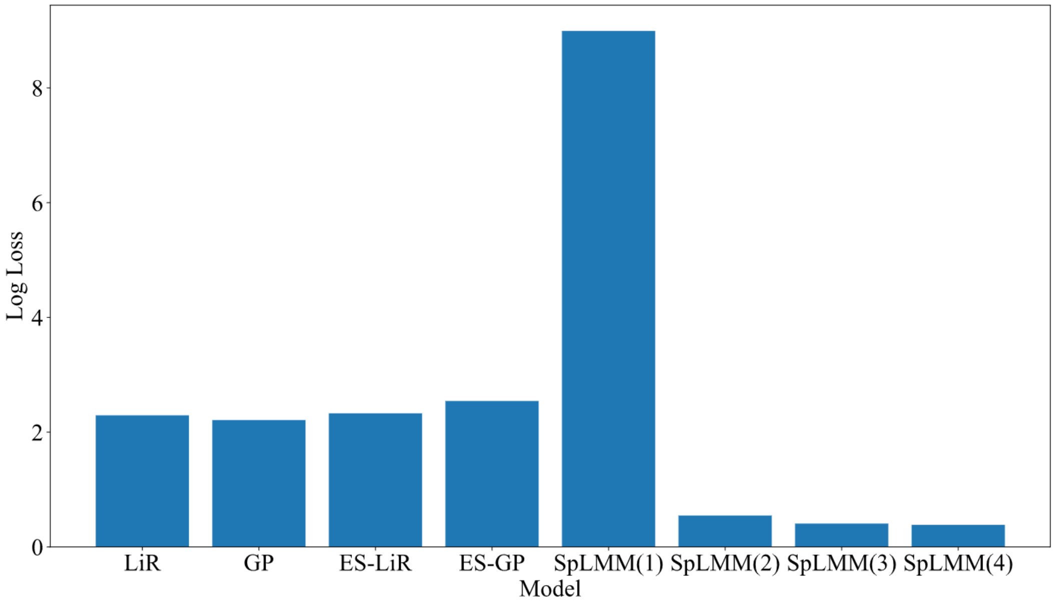


Figure 4.2: Log loss of each models.

From here on, we will discuss why SpLMM was more successful than other methods. Figure 4.3 shows the log loss of each CV trial for linear regression and SpLMM($K = 2$). As can be seen from this figure, the log loss of the fourth trial in linear regression is large, whereas the log loss of the fourth trial in SpLMM is small. Figure 4.4 shows a comparison of the true and predicted values for ES-LiR and ES-GP. The further away from the diagonal line, the more prediction error is. The red circles in the lower right corner of this figure show that the predictions have a significantly large prediction error for both ES-LiR and ES-GP. This point is the molecule Furan, which has the second highest diffusion coefficient of 8.34 among all the data, but the predicted values for ES-LiR and ES-GP are -0.20 and 0.16 respectively. Furan is included in the test data in the fourth trial. This is the reason why the log loss of the fourth trial is high in ES-LiR and ES-GP. Therefore, we will focus on Furan and see how SpLMM differs from other methods.

Figure 4.5 shows comparing the predictive distributions of Furan for ES-LiR, ES-GP, and SpLMM ($K = 2$). The horizontal axis represents the predicted diffusion coefficient and the vertical axis represents the probability density of the predicted value. The dotted line is the true value of the diffusion coefficient. The predicted distributions of ES-LiR and ES-GP are generally located at low diffusion coefficients. On the other hand, the distribution is wider in SpLMM, with peaks at higher diffusion coefficients. Since log loss is a sign inversion of the logarithm of the probability density of the true value in the predictive distribution, the higher the probability density of the dotted line, the lower the value of log loss. This is why the log loss of the fourth trial is smaller for SpLMM and higher for ES-LiR and ES-GP.

We compare the selected descriptors to see why the predictions for ES-LiR and ES-GP are so far off. Figure 4.6 shows the descriptors selected by the ES-LiR, ES-GP, and SpLMM ($K = 2$) methods in the fourth trial. The top two diagrams represent ES-LiR

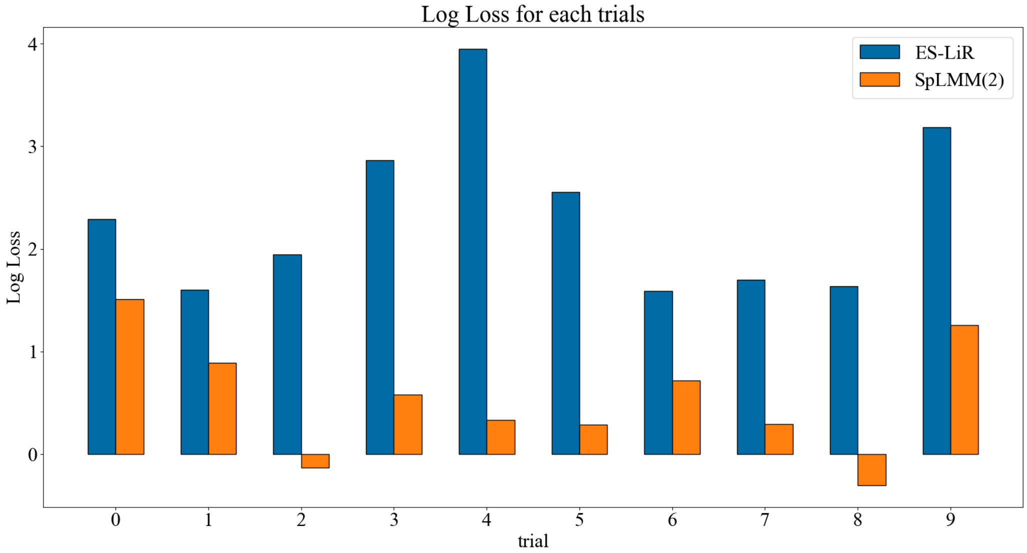


Figure 4.3: Comparison of log loss between ES-LiR and SpLMM ($K = 2$). The horizontal axis shows each trial in CV and the vertical axis shows the log loss in the trial.

and ES-GP, respectively, and the bottom two represent two mixture models of SpLMM. Each diagram has descriptors on the vertical axis and indicators on the horizontal axis. The indicators in the ES are arranged from left to right in order of decreasing BFE. To see the top results, the top 100 out of the 2^{11} items are displayed here. The indicators in the SpLMM are arranged from left to right in order of posterior probability. The top 1000 are shown here. The black cells represent selected descriptors, and the white cells represent non-selected descriptors. In the above three diagrams, x_1 (Molecular weight), x_7 (HOMO energy) and x_9 (Coordination energy) are selected relatively frequently. On the other hand, in the bottom diagram, the only descriptor commonly selected is x_1 (Molecular weight). Therefore, SpLMM is considered to be a mixture of two models as follows: a model that makes predictions similar to ES-LiR and ES-GP and a model that only emphasizes molecular weight. This is very important for the correct prediction of Furan. The absolute value of coordination energy indicates the strength of coordination between Li-ion and solvent molecules, i.e., the strength of attraction, and basically, the stronger the force, the faster the diffusion of Li-ion (i.e., the larger the diffusion coefficient). In fact, the results of linear regression take a negative value for the coefficient of coordination energy (where coordination energy is negative value). However, Furan differs from that trend. The absolute value of Furan's coordination energy is the smallest among all the data. Therefore, choosing the coordination energy as a descriptor does not predict the Furan diffusion coefficient correctly. On the other hand, molecular weight which is only frequently chosen in the second model of SpLMM, is negatively correlated, as lighter molecules are expected to diffuse faster. In fact, the linear regression results show that the coefficient of Molecular weight is negative. Furan's molecular weight is the third lowest among all the data, and if only molecular weight is used as a descriptor, the diffusion coefficient can be expected to increase. In other words, although many molecules can be better predicted using not only molecular weight, but also coordination energy, on the other hand, there are some molecules, such as Furan, that may cause the low accuracy when coordination energy is used. This means that only SpLMM can correctly predict the data generated by different mechanisms.

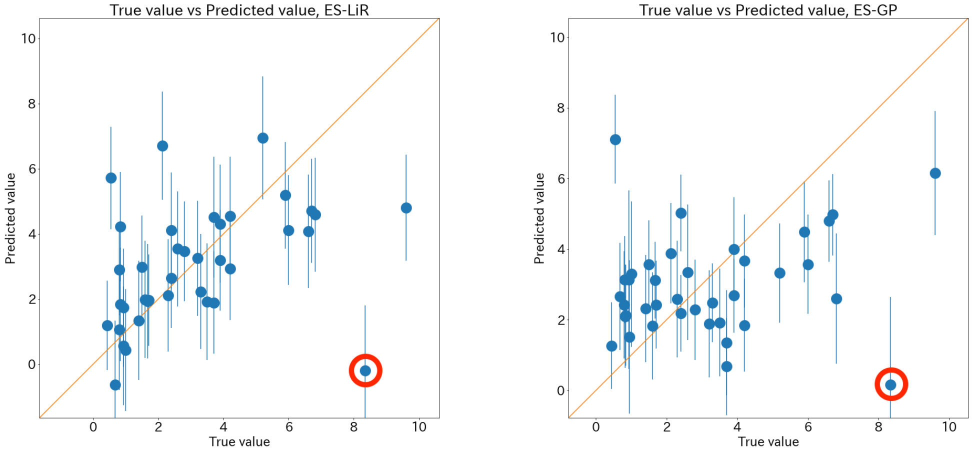


Figure 4.4: Predictions for each solvent molecule in the best models of ES-LiR (left) and ES-GP (right). The horizontal axis represents the true value and the vertical axis represents the predicted value. The error bars are the standard deviation of the predicted distribution for each point. The diagonal line indicates that the true value is equal to the prediction, and the further away the diagonal line is, the more the prediction error is. The dots circled in red represent the Furan.

In order to investigate how the mechanism of diffusion of Li-ion in Furan diffuses differently from that of other solvents, we carefully observed the results of the simulations. As a result, the migration of Li-ion tended to be different from that of other molecules. Basically, Li-ion coordinates with four solvent molecules to move in the electrolyte. It is called "vehicle-type" because Li-ion uses solvents like a vehicle. Since the vehicle-type solvents must be strongly bound to Li-ion, the higher the coordination energy, the faster the diffusion. However, Furan is not a vehicle-type migrant and tends to migrate by changing the coordination of its molecules. Although further research is needed to conclude this, we believe that it is noteworthy that such a different mechanism can be extracted by SpLMM.

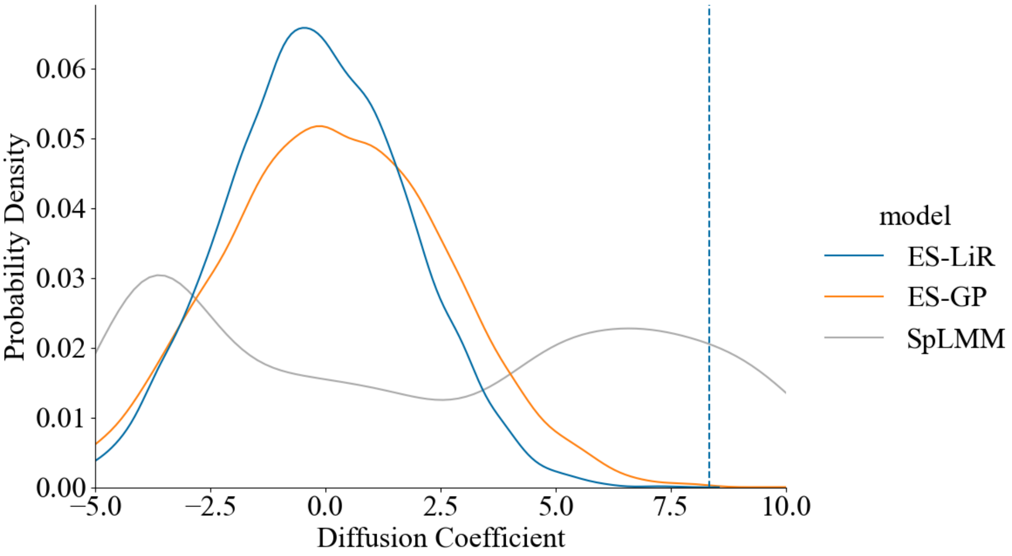
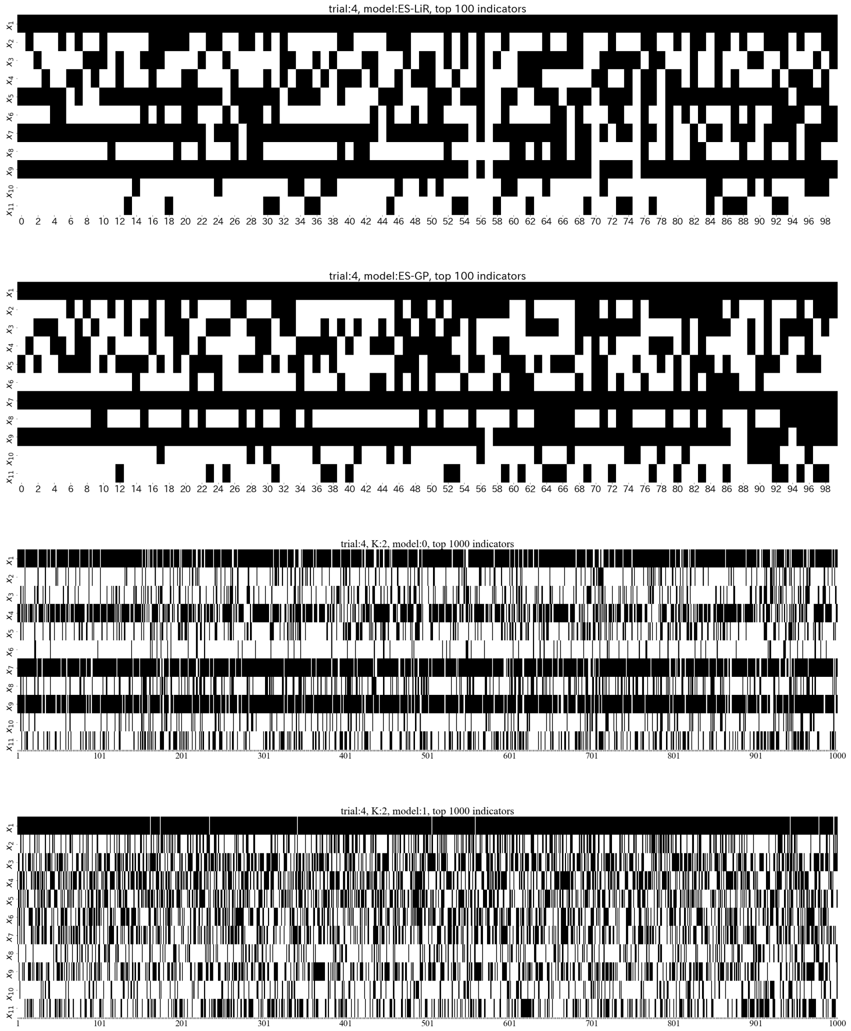


Figure 4.5: Comparison of the predictive distributions of Furan in ES-LiR, ES-GP, and SpLMM ($K \geq 2$). The horizontal axis represents the predicted diffusion coefficient and the vertical axis represents the probability density of the predicted value. The dotted line is the true value of the diffusion coefficient. Complete figures for all molecules are presented in Section C.4 in the Appendix.



trial:4, model:ES-LiR, top 100 indicators

trial:4, model:ES-GP, top 100 indicators

trial:4, K:2, model:0, top 1000 indicators

trial:4, K:2, model:1, top 1000 indicators

Figure 4.6: Comparison of the indicators chosen for each model in the fourth trial of the CV. The first model of ES-LiR, ES-GP, and SpLMM($K = 2$), the second model in order from the top. Each diagram has descriptors on the vertical axis and indicators on the horizontal axis. The indicators in the ES are arranged from left to right in order of decreasing BFE. Here we took the top 100 pieces out of the 2^{11} pieces to see the top results. The indicators in the SpLMM are arranged from left to right in order of posterior probability. In this case, we took out 1000 pieces. The black cells represent used descriptors and the white cells represent unused descriptors. The descriptors on the vertical axis are defined as follows. x_1: Molecular Weight, x_2: Density, x_3: Boiling Point, x_4: Melting Point, x_5: Flashing Point, x_6: LUMO Energy, x_7: HOMO Energy, x_8: Dipole Moment, x_9: Coordination Energy, x_{10}: Distance between Li and O, x_{11}: Mulliken Charge. The results of the other trials are available on Appendix C.3.

4.4 Conclusion

In this chapter, we compared the values of diffusion coefficients using SpLMM in addition to linear regression with all variables, GP with all variables, ES-LiR, and ES-GP to predict the values of diffusion coefficients. The results suggest that the data of diffusion coefficients generated from more than one mechanism, indicating the limitations of a single-value prediction method for such data. In particular, we focused on Furan, where the results of SpLMM and the other methods differed greatly, and we conducted a detailed study, and we showed that some molecules are better predicted without the coordination energy, which has been thought to be correlated with the diffusion coefficient.

This indicates that predicting with a single model during virtual screening may not find minorities with different mechanisms than the majority, and on the other hand, using SpLMM that has multiple spaces of descriptors, such minorities can be searched appropriately and desired materials can be found quickly. In addition, SpLMM allows us to compare combinations of mixed descriptors and obtain information on the data generation mechanism of each model by visualizing the indicators as shown in Figure 4.6. In some cases, the information can be viewed by materials science experts to infer the hidden descriptors that determine each model. This is very important for materials science because new discoveries are more likely to be made in the unexplored mechanisms of minorities than in the well-studied mechanisms of majorities.

Finally, we describe the future development of this research. There are two directions for future research. First, we will further explore the different tendencies of the molecules we discovered in this study to understand why they are different from other molecules. Secondly, the SpLMM, which was validated in this study, will be applied to a larger scale database of candidate materials to perform virtual screening. Through these studies, we hope to dramatically advance the search for electrolytes, which has not made significant progress for decades, and contribute to the discovery of new high-performance electrolyte materials.

Chapter 5

Conclusion & Discussion

In recent decades, there have been no significant developments in the search for electrolyte materials for commercial LIB. Solid materials such as cathode and anode can be easily treated by static calculations at absolute zero point because their atomic structure does not change significantly. However, in the case of electrolyte, the dynamics of the atomic structure at finite temperatures is required. It makes handling electrolyte difficult and increases the computational cost. That is one reason why there have been no significant developments in the search for electrolyte materials for commercial LIB. To address this issue, we have developed a new framework that combines computational chemistry and data science throughout this paper. In Chapter 2, we created a database using computational chemistry, and in Chapters 3 and 4, we discussed data science methods using the database. In this Chapter, we briefly discuss those conclusions and their effects not only for the search for electrolytes in secondary batteries, but also for materials science and information science.

In Chapter 2, we create the database for the search of the electrolyte of secondary batteries. In this study, we selected commercially available materials for the electrolyte solvent of secondary batteries to investigate our methods. The LIB electrolyte database contains five experimental values (boiling point, density, flash point, melting point, and molecular weight) and seven calculated values (dipole moment, HOMO energy, LUMO energy, mulliken charge, Li-O distance). We used cluster model DFT calculation to get following six values: dipole moments, HOMO energy, LUMO energy, mulliken charge, the distance between the cation and the oxygen atom and the coordination energy. Finally we obtained a database consisting of 103 electrolyte solvents. For the diffusion coefficient, we executed a long time DFT-MD calculation and obtained 38 values of diffusion coefficient. In addition, we also created a database of the electrolytes of secondary batteries including not only Li but also alkali metals (Na, K, Rb, and Cs). We added the ionic radius, atomic weights, and electronegativity to the database as descriptors of the cations.

In Chapter 3, we selected the coordination energy as a value related to the function of the secondary battery, and built the prediction models of the coordination energy. The coordination energy is one of the most computationally expensive values in our database, and it is worthwhile to predict it using low-cost descriptors. In this study, we applied exhaustive search, which is one of the methods of sparse modeling, to build a predictive model and select variables. Firstly, we tested the effectiveness of the method on the LIB data set. As a result, the prediction errors of ES-LiR and ES-GP are smaller than those of linear regression with all variables, 14% and 52%, respectively, making it possible to accurately predict the coordination energy from descriptors with small computational costs. In addition, by visualizing the exhaustive search results using weight diagrams, we have identified which combinations of variables are important in predicting the coordination energy in each model. We found following descriptors are important to predict coordination energy. In the linear model (ES-LiR), important descriptors are mulliken charge of the O atom, the distance between Li-ion and the O atom, flashing point. In the nonlinear model (ES-GP), the important descriptors are HOMO energy, dipole moment and the boiling point.

In addition, we used the database of not only Li but also other alkali metals (Na, K, Rb, and Cs) to predict the coordination energies. The CVEs of ES-LiR and ES-GP were 0.127 eV and 0.016 eV, respectively, which showed a very high prediction accuracy. This makes it possible to predict the coordination energies of not only Li but also various other ions using only the properties of ions and solvents.

Throughout Chapter 3, we have achieved a practical level of accuracy in predicting the coordination energy. We also used ES to physically interpret the results and discuss the relationships between the descriptors and coordination energy. We believe that these results will greatly advance the search for new electrolytes for secondary batteries. For reference, the results of using these methods to predict other data are shown in Appendix D,E.

In Chapter 4, we chose the diffusion coefficient as a value related to the function of the LIB and built a prediction model for it. Although the diffusion coefficient is an important value that relates to the speed of charge and discharge of the LIB, a long DFT-MD calculation is required to calculate it and it takes several weeks to several months. Therefore, it is of great value to predict diffusion coefficients from other descriptors that are less computationally expensive. In this study, in addition to MLR, GP, ES-LiR and ES-GP used in Chapter 3, we constructed a prediction model using the Sparse Linear Mixture Model (SpLMM), which makes predictions using multiple models. The results showed that the log loss of SpLMM was significantly lower than those of the single-model prediction methods, which means SpLMM can best explain the data. In order to investigate how SpLMM improves prediction accuracy, we focused on the case of Furan, where the single model approach failed to predict correctly. We found that SpLMM is able to automatically extract these different mechanisms from the data and accurately predict minority solvents with different mechanisms than the majority. In the field of materials science, as well as the search for LIB electrolytes, the variables we want to predict are not always generated by the same mechanism. In particular, the potential for new scientific progress lies not in the well-studied majority data, but in the minorities that are not well-studied. In this study, we were able to show that a mixture model with sparsity, such as SpLMM, is a method for exploring not only majorities but also minorities.

Through this paper, we proposed a framework for exploring the electrolyte in secondary batteries. It is the following framework. First, create a database as much as you can using computational chemistry approach. Next, extract important descriptors and build a prediction model using the ES method. If the prediction model is enough for practical use, prepare new candidate database to perform virtual screening. Decide which values should be included in the database based on a balance between computational cost and prediction accuracy using the result of the ES method. If the prediction accuracy is not practically sufficient, build mixture models using SpLMM and perform virtual screening. This framework allows us to search for useful electrolyte materials while interpreting the generated data.

Finally, we will describe the impact and future development of this study based on the current situation in information science and materials informatics. Recently, many researches on Deep Learning-based methods have been conducted in the field of information science. However, given the high cost of database construction in the field of materials science, machine learning methods that require large amounts of data, such as Deep Learning, cannot be applied to this field. Moreover, methods that output predictions as a single value cannot be learned when the data generated from multiple mechanisms. In this regard, SpLMM treats multiple models by Bayesian inference, which enables us to extract multiple mechanisms while preventing over-training even with small amounts of data. It can be said that materials science and Bayesian inference are very compatible with each other. In the future, Bayesian inference will become more active in material exploration. One of the methods in materials informatics for handling data with Bayesian inference is materials search by Bayesian optimization. It is possible to apply Bayesian optimization to this database (ref: Appendix F). However, most of the existing applications of Bayesian optimization to MI use a single model to make predictions. Using the SpLMM, Bayesian optimization can also be extended to sparse mixture model. It is expected that Bayesian optimization with SpLMM can captures not only the majority tendency but also the minority tendency and can be performed quickly. In future research, we will utilize such a sparse mixture model for Bayesian optimization, and contribute to the development of materials science by accelerating the search for the electrolyte of secondary batteries.

Appendix A

Supporting Information for Chapter 2

A.1 Experimental value database of Li-ion electrolytes

Table A.1: A feature database of experimental values constructed using the catalog[22].

name	molecularWeight	density	boilingPoint	meltingPoint	flashingPoint
PC	102.09	1.21	242	-49	132
EC	88.06	1.32	244	38	152
VC	86.05	1.255	162	22	72
VEC	114	1.188	237		96.7
FEC	106.05	1.497	210	17.3	122
DMC	90.08	1.07	90	4	17
DEC	118.13	0.975	126	-43	25
EMC	104.1	1.015	107	-55	23
DAC	142.15	1.069	217		86
Dimethyl2,5-dioxahexanedioate	178.14	1.24	220		
Diethyl 2,5-dioxahexanedioate	206.19	1.15	227		
Furan	68.07	0.94	31	-85.6	-35
2,5-Dimethyl furane	96.13	0.89	94	-63	16
THF	72.11	0.889	66	-108	-17
2-MeTHF	86.13	0.855	78	-136	-11
THP	86.13	0.881	88	-45	-15
DOL	74.08	1.065	75	-95	-4
DIOX	88.11	1.034	101	11.5	12
12-Crown 4-ether	176.21	1.11	65	16	109
18-Crown 6-ether	264.32	1.175	116	39.5	109
DMM	76.09	0.86	42	-105	-32
DME	90.12	0.867	85	-58	0
DEE	118.17	0.842	121	-74	35
Diglyme	134.17	0.945	162	-64	56
Triglyme	178.18	0.9862	216	-45	113
Tetraglyme	222.28	1.013	275.5	-30	143.5
MA	74.08	0.932	57.5	-98	-9
EA	88.11	0.902	77	-83	-4.4
PA	102.13	0.886	101.67	-92	14
iPA	102.13	0.872	88	-73	4
BA	116.16	0.883	125.5		27
MFA	110.06	1.272	85.5		24

name	molecularWeight	density	boilingPoint	meltingPoint	flashingPoint
EFA	142.08	1.194	61		-1
MP	88.11	0.915	79	-88	6
EP	102.13	0.891	99	-73	12
PP	116.16	0.88	122	-76	27
MF	60.05	0.974	34	-100	-26
EF	74.08	0.917	54	-80	-18
EB	116.16	0.875	120	-93	25
iPB	130.18	0.859	130		30
MiB	102.13	0.89	92	-85	13
MCA	99.09	1.13	200	-23	105
VA	86.09	0.934	72.5	-93	-8
GBL	86.09	1.13	204	-44	101
GVL	100.12	1.057	207	-31	81
d-Valero lactone	100.12	1.079	245	-13	112
e-Capro lactone	114.14	1.03	253	-1	109
g-Hexano lactone	114.14	1.03	220	-18	98
g-Undeca lactone	184.27	0.943	286		113
TMP	140.07	1.197	197	-46	107
Tri n-propyl phosphate	224.23	1.012	121	113	
TPhP	326.28		224	49	200
NMP	99.13	1.028	204	-23	99
DMF	73.09	0.944	153	-61	60
DMI	114.15	1.04	226	8.2	93
DMAC	87.12	0.936	165	-20	66
3-Methyl-2-oxazolidinone	101.1	1.17	88	17	110
Ethylene diamine	60.1	0.898	117.3	11.3	34
Pyridine	79.1	0.9827	115.2	-41.6	16
N-Methyl imidazole	82.1	1.04	106	-6	92
Dimethyl sulfate	126.13	1.333	188	-32	83
Dimethyl sulfite	110.13	1.21	126		30
Dipropyl sulfite	166.24	1.03	192		110
ES	108.12	1.426	68		79
Dimethyl sulfone	94.13		238	110	143
ethylmethyl sulfone	108.16			34	163

name	molecularWeight	density	boilingPoint	meltingPoint	flashingPoint
Diphenyl sulfone	218.27		379	127	184
Bis(4-Fluoro phenyl sulfone)	254.25				98.5
SL	120.17	1.26	285	28	165
3-MeSL	134.2	1.2	274		163
Methanesulfonic acid methyl ester	110.13	1.3	202.5	20	140
Benzen sulfonic acid methyl ester	172.19	1.27	280	-4	143
Trifluoromethane sulfonic acid methyl ester	164.1	1.5	99		38
PS	122.14	1.51	180	32	158
BS	136.17	1.331	153	14	154
DMSO	78.13	1.1	189	18	95
Diphenly disulfide	218.33	1.22	310	59	178
Dimethyl sulfide	62.13	0.846	38	-98	-34
Diethyl sulfide	90.19	0.837	91	-100	-10
AN	41.05	0.783	82	-45	2
PN	55.08	0.772	97	-93	6
Adiponitrile	108.14	0.96	295	1	93
Valeronitrile	83.13	0.8	140	-96	40
Glutaronitrile	94.12	0.989	286	-26	109
Malononitrile	66.06	1.049	220	32	112
Succinonitrile	80.09	1.02	267	56	158
Pimelonitrile	122.17	0.95	175		112
Suberonitrile	136.19	0.93	172	-3.5	161
Isobutyronitrile	69.11	0.77	107.5	-72	8
Succinic Anhydride	100.07	1.5	261	119.5	157
MEK	72.11	0.805	79	-87	-3

A.2 Calculated value database of Li-ion electrolytes

Table A.2: A database constructed by performing cluster model DFT calculations using Li as a cation. The diffusion coefficients (diffCoeff) were obtained by a long time DFT-MD calculation.

name	LUMO	HOMO	dipoleMoment	coordinationEnergy	dist	mulliken	diffCoeff
PC	0.946	-7.93	5.255	-0.0915301	1.7472	-0.242515	1.7
EC	0.919	-8.017	5.07	-0.0890956	1.7523	-0.240074	3.7
VC	-0.137	-6.973	4.365	-0.082463	1.76	-0.23108	
VEC	-0.72	-7.829	5.28	-0.0917539	1.7461	-0.239975	
FEC	0.493	-8.468	4.487	-0.0815472	1.7625	-0.221583	0.93
DMC	1.115	-7.774	0.342	-0.0796927	1.7469	-0.306358	3.7
DEC	1.217	-7.654	0.613	-0.0838614	1.7403	-0.308275	0.83
EMC	1.168	-7.713	0.514	-0.08183	1.7438	-0.307262	
DAC	-0.238	-7.419	0.494	-0.0827409	1.7398	-0.306046	
Dimethyl2,5-dioxahexanedioate	0.922	-7.933		-0.0793365	1.7491	-0.299555	
Diethyl 2,5-dioxahexanedioate	0.99	-7.856		-0.0818361	1.745	-0.300971	
Furan	0.296	-6.265	0.511	-0.050559	1.8659	-0.170214	8.34
2,5-Dimethyl furane	0.514	-5.621	0.033	-0.0581926	1.8484	-0.216976	
THF	1.38	-6.832	1.434	-0.0776553	1.8076	-0.322937	
2-MeTHF	1.454	-6.535	1.535	-0.0796459	1.8068	-0.347471	3.9
THP	1.537	-6.711	1.301	-0.0752215	1.8044	-0.323574	
DOL	1.493	-6.955	1.324	-0.068876	1.8177	-0.31521	0.55
DIOX	1.773	-6.483		-0.0688849	1.8103	-0.325603	0.41
12-Crown 4-ether	1.965	-6.655		-0.1611528	1.8893	-0.331037	
18-Crown 6-ether	1.674	-6.491		-0.1666453	2.0207	-0.318812	
DMM	1.459	-6.846	2.165	-0.1026767	1.9045	-0.297578	6.81
DME	1.669	-6.863	0.101	-0.0719001	1.8258	-0.333908	
DEE	1.665	-6.839		-0.0750624	1.8178	-0.339686	
Diglyme	1.616	-6.898	1.131	-0.071306	1.8271	-0.333589	
Triglyme	1.61	-6.907		-0.0710693	1.8269	-0.333696	
Tetraglyme	1.575	-6.915	1.129	-0.0710173	1.8269	-0.333576	
MA	0.339	-7.371	1.733	-0.0828816	1.7553	-0.264722	6.7
EA	0.38	-7.304	1.88	-0.0851344	1.7514	-0.266523	2.4
PA	0.385	-7.299	1.941	-0.0857991	1.7497	-0.266057	2.8
iPA	0.38	-7.28	1.794	-0.0858845	1.7488	-0.269185	4.2
BA	0.391	-7.29	1.925	-0.0862446	1.7492	-0.266279	

name	LUMO	HOMO	dipoleMoment	coordinationEnergy	dist	mulliken	diffCoeff
MFA	-0.734	-8.032	1.521	-0.0882974	1.8879	-0.249354	1.1
EFA	-0.83	-8.259	3.09	-0.0821895	1.8813	-0.230915	3.79
MP	0.372	-7.376	1.612	-0.0831532	1.7846	-0.267276	6
EP	0.414	-7.31	1.763	-0.0853054	1.7865	-0.269171	3.9
PP	0.418	-7.304	1.827	-0.0858786	1.7868	-0.268821	6
MF	0.089	-7.688	3.899	-0.0840117	1.7682	-0.231004	9.6
EF	0.105	-7.603	3.997	-0.0863572	1.7626	-0.234361	6.6
EB	0.414	-7.303	1.682	-0.0865553	1.7954	-0.269889	1.6
iPB	0.414	-7.279	1.605	-0.0876426	1.7933	-0.27273	
MiB	0.294	-7.299	1.671	-0.0846759	1.7936	-0.266754	1.5
MCA	-0.371	-8.134	5.219	-0.0934029	1.9086	-0.231463	3.5
VA	-0.496	-6.941	3.658	-0.0871123	1.7608	-0.217018	5.9
GBL	0.254	-7.269	4.296	-0.0904884	1.7582	-0.236546	3.2
GVL	0.304	-7.208	4.327	-0.0926031	1.753	-0.239196	3.3
d-Valero lactone	0.155	-7.092	4.459	-0.0958987	1.7512	-0.239822	2.3
e-Capro lactone	0.288	-7.074	4.449	-0.0959784	1.7531	-0.23859	0.44
g-Hexano lactone	0.314	-7.186	4.394	-0.0933717	1.7517	-0.23961	0.8
g-Undeca lactone	0.335	-7.159	4.56	-0.0945197	1.7516	-0.240109	
TMP	1.112	-7.765	3.356	-0.1037461	1.7399	-0.467	0.68
Tri n-propyl phosphate	1.253	-7.633	3.427	-0.1090916	1.7327	-0.469684	
TPhP	-0.566	-6.649	3.157	-0.1096657	1.8305	-0.43594	
NMP	0.842	-6.421	3.609	-0.1018721	1.7244	-0.29909	2.4
DMF	0.731	-6.623	3.698	-0.0975107	1.7362	-0.296931	4.2
DMI	1.337	-6.216	3.603	-0.1034159	1.7122	-0.324571	1.4
DMAC	0.8	-6.411	3.554	-0.1008166	1.723	-0.301032	2.6
3-Methyl-2-oxazolidinone	1.066	-6.805	4.764	-0.1000823	1.748	-0.283473	
Ethylene diamine	1.469	-6.244	10.804	-0.0776468	1.9686	-0.183361	
Pyridine	-0.796	-6.985	8.718	-0.0786491	1.9406	-0.223084	0.81
N-Methyl imidazole	0.64	-6.135	7.989	-0.0929765	1.9122	-0.237705	0.84
Dimethyl sulfate	0.341	-8.195	3.401	-0.0799501	2.0285	-0.392499	0.94
Dimethyl sulfite	-0.327	-7.416	1.406	-0.0903177	1.9009	-0.456224	
Dipropyl sulfite	-0.248	-7.292	1.55	-0.096901	1.9011	-0.456483	
ES	-0.823	-7.725	3.123	-0.0844953	1.7577	-0.422981	
Dimethyl sulfone	0.905	-7.683	4.465	-0.0972333	2.0233	-0.459806	

name	LUMO	HOMO	dipoleMoment	coordinationEnergy	dist	mulliken	diffCoeff
ethylmethyl sulfone	0.889	-7.567	4.381	-0.0997257	2.0133	-0.463961	
Diphenyl sulfone	-1.538	-7.222	5.078	-0.1054381	2.0009	-0.454931	
Bis(4-Fluoro phenyl sulfone)	-1.61	-7.204	3.493	-0.0998239	2.0046	-0.452934	
SL	0.826	-7.383	5.087	-0.1015673	2.0137	-0.459225	1
3-MeSL	0.737	-7.366	5.086	-0.1024981	2.0117	-0.461402	
Methanesulfonic acid methyl ester	0.696	-8.252	4.127	-0.0870525	2.0637	-0.438264	
Benzen sulfonic acid methyl ester	-1.442	-7.618	5.014	-0.0932842	2.0499	-0.435605	
Trifluoromethane sulfonic acid methyl ester	0.016	-8.981	3.209	-0.0691796	1.8116	-0.407552	
PS	0.549	-7.917	5.468	-0.091351	2.034	-0.426438	
BS	0.41	-8.017	5.384	-0.0933286	2.033	-0.430998	
DMSO	0.963	-6.01	3.821	-0.1079828	1.7176	-0.542095	
Diphenly disulfide	-1.108	-6.111	1.597	-0.0672389	2.4133	0.003611	
Dimethyl sulfide	1.107	-5.895	1.485	-0.0541073	2.3891	-0.07032	
Diethyl sulfide	0.833	-5.809	1.499	-0.059315	2.369	-0.055851	
AN	0.898	-8.933	3.743	-0.074943	1.9204	-0.180717	5.2
PN	0.587	-8.802	3.826	-0.0771706	1.9139	-0.184843	
Adiponitrile	0.312	-8.976		-0.0721845	1.9164	-0.175813	
Valeronitrile	0.728	-8.704	4.056	-0.0792032	1.9084	-0.183419	
Glutaronitrile	-0.136	-9.081	3.716	-0.0703936	1.9212	-0.170109	
Malononitrile	-0.51	-9.544	3.623	-0.0614103	1.9446	-0.15267	
Succinonitrile	-0.097	-9.31	0.001	-0.0654351	1.929	-0.164515	
Pimelonitrile	0.346	-8.858	3.734	-0.0744282	1.9139	-0.177488	
Suberonitrile	0.497	-8.789	0.001	-0.0753287	1.9115	-0.17968	
Isobutyronitrile	0.642	-8.727	3.847	-0.0788726	1.9734	-0.188931	
Succinic Anhydride	-0.866	-7.794	4.234	-0.0766437	1.7835	-0.199212	
MEK	-0.386	-6.601	2.771	-0.084393	1.7592	-0.225071	2.12

A.3 Database of alkali-ion electrolytes

Table A.3: A database constructed by performing cluster model DFT calculations using Li, Na, K, Rb and Cs as a cation.

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
PROPYLENE CARBONATE	-2.399	0.9	0.97	6.94	-0.59344	-0.3719	0.07808	2.16718404	-10375.104
Ethylene carbonate	-2.343	0.9	0.97	6.94	-0.59131	-0.37523	0.07999	2.0916744	-9306.486
VINYLENE CARBONATE	-2.179	0.9	0.97	6.94	-0.59085	-0.31578	0.03559	1.83682274	-9273.2915
Fluoroethylene carbonate	-2.128	0.9	0.97	6.94	-0.57386	-0.39223	0.06581	1.91022747	-12004.26
Dimethyl carbonate	-2.068	0.9	0.97	6.94	-0.67204	-0.36873	0.09098	0.07937243	-9339.1429
DIETHYL CARBONATE	-2.13	0.9	0.97	6.94	-0.6787	-0.36679	0.09279	0.15446412	-11476.157
Ethyl methyl carbonate	-2.114	0.9	0.97	6.94	-0.67569	-0.36782	0.0896	0.29941679	-10407.644
FURAN	-1.32	0.9	0.97	6.94	-0.48625	-0.28532	0.05396	0.2187243	-6251.9579
2,5-DIMETHYLFURAN	-1.515	0.9	0.97	6.94	-0.50904	-0.259	0.06162	0.0179921	-8389.1973
TETRAHYDROFURAN	-2.047	0.9	0.97	6.94	-0.62428	-0.31168	0.09161	0.70868651	-6317.4253
2-METHYLTETRAHYDROFURAN	-2.096	0.9	0.97	6.94	-0.63458	-0.31027	0.09209	0.65453619	-7386.0098
TETRAHYDROPYRAN	-1.996	0.9	0.97	6.94	-0.62007	-0.3143	0.09446	0.56267776	-7386.0963
1,3-DIOXOLANE	-1.825	0.9	0.97	6.94	-0.62864	-0.32401	0.10229	0.46742258	-7293.3309
1,4-DIOXANE	-1.808	0.9	0.97	6.94	-0.61676	-0.30415	0.1073	1.04E-05	-8361.789
Ethoxymethoxymethane	-1.837	0.9	0.97	6.94	-0.63798	-0.34033	0.09866	0.14969769	-8394.449
ETHYL ACETATE	-2.206	0.9	0.97	6.94	-0.62731	-0.34685	0.05997	0.64684699	-8362.9517
ISOPROPYL ACETATE	-2.222	0.9	0.97	6.94	-0.62661	-0.34444	0.0612	0.68800247	-9431.5003
METHYL PROPIONATE	-2.138	0.9	0.97	6.94	-0.62146	-0.34869	0.06332	0.63677806	-8362.8859
ETHYL PROPIONATE	-2.202	0.9	0.97	6.94	-0.62261	-0.34726	0.06069	0.68798682	-9431.3541
METHYL FORMATE	-2.011	0.9	0.97	6.94	-0.60582	-0.36295	0.05505	0.68750544	-6225.7198
ETHYL BUTYRATE	-2.18	0.9	0.97	6.94	-0.63193	-0.34311	0.05741	0.62912016	-10499.797
METHYL ISOBUTYRATE	-2.183	0.9	0.97	6.94	-0.61939	-0.34384	0.05919	0.65885942	-9431.3249
VINYL ACETATE	-2.052	0.9	0.97	6.94	-0.60014	-0.31114	0.02966	0.64433921	-8329.5868
Gamma-Butyrolactone	-2.381	0.9	0.97	6.94	-0.58246	-0.34261	0.05711	1.77086298	-8330.2469
Gamma-Valerolactone	-2.438	0.9	0.97	6.94	-0.58352	-0.34066	0.05797	1.78595247	-9398.8408
Delta-Valerolactone	-2.502	0.9	0.97	6.94	-0.58226	-0.33582	0.04675	1.79851237	-9398.6211
Epsilon-caprolactone	-2.536	0.9	0.97	6.94	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Gamma-Hexanolactone	-2.454	0.9	0.97	6.94	-0.58392	-0.33992	0.0582	1.81361705	-10467.253
TRIMETHYL PHOSPHATE	-2.79	0.9	0.97	6.94	-1.08448	-0.3538	0.09285	0.47819921	-20719.462
TRIETHYL PHOSPHATE	-2.914	0.9	0.97	6.94	-1.08833	-0.34517	0.08308	0.94269476	-23924.985
SULFOLANE	-2.481	0.9	0.97	6.94	-0.95081	-0.33695	0.06799	2.06428499	-19192.049

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
METHYL METHANESULFONATE	-2.08	0.9	0.97	6.94	-0.78189	-0.36367	0.09127	1.2015052	-19132.267
METHYL BENZENESULFONATE	-2.232	0.9	0.97	6.94	-0.92765	-0.333	-0.01708	1.34600337	-24343.813
1,3-Propanesultone	-2.275	0.9	0.97	6.94	-0.77035	-0.3632	0.06727	2.31386941	-20168.047
1,4-BUTANE SULTONE	-2.314	0.9	0.97	6.94	-0.78202	-0.36768	0.07208	2.14753013	-21236.86
Dimethyl sulfoxide	-2.905	0.9	0.97	6.94	-0.97186	-0.28391	0.08537	1.61570745	-15043.485
SUCCINIC ANHYDRIDE	-1.99	0.9	0.97	6.94	-0.5436	-0.36492	0.01241	1.71813593	-10342.727
CYCLOHEXANONE	-2.259	0.9	0.97	6.94	-0.55268	-0.30724	0.02926	1.19994939	-8422.2456
CAPROLACTONE	-2.536	0.9	0.97	6.94	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Propiolactone	-2.162	0.9	0.97	6.94	-0.56703	-0.35549	0.05287	1.63151562	-7261.067
CYCLOPENTANONE	-2.23	0.9	0.97	6.94	-0.55019	-0.30855	0.02666	1.12642153	-7353.6424
Diketene	-2.023	0.9	0.97	6.94	-0.54443	-0.32081	0.03875	1.37862436	-8296.2676
ACETOPHENONE	-2.272	0.9	0.97	6.94	-0.55641	-0.31593	-0.02263	1.1340558	-10461.15
Guaiacol	-2.882	0.9	0.97	6.94	-0.5759	-0.26469	0.03188	0.818535	-11469.446
Benzaldehyde	-2.177	0.9	0.97	6.94	-0.5358	-0.32133	-0.03058	1.20204662	-9392.4809
2-METHYLCYCLOHEXANONE	-2.293	0.9	0.97	6.94	-0.55411	-0.30368	0.03108	1.16980232	-9490.6559
METHYL METHANESULFONATE	-2.08	0.9	0.97	6.94	-0.78188	-0.36366	0.09128	1.20127447	-19132.267
DIETHYLSULFATE	-2.335	0.9	0.97	6.94	-0.92686	-0.36695	0.08888	0.5422029	-23313.549
2,3-butanedione	-1.768	0.9	0.97	6.94	-0.52628	-0.31341	-0.04212	0.05681995	-8329.6147
ACETOPHENONE	-2.272	0.9	0.97	6.94	-0.55644	-0.31592	-0.02264	1.13390682	-10461.15
BENZYL BENZOATE	-2.758	0.9	0.97	6.94	-0.62251	-0.3091	-0.01378	0.75661917	-18786.228
DIPHENYL ETHER	-1.625	0.9	0.97	6.94	-0.55198	-0.27253	0.01595	0.32082062	-14636.695
PENTANAL	-1.992	0.9	0.97	6.94	-0.54101	-0.32531	0.02944	0.99136767	-7385.9481
2-Methoxyethyl acetate	-2.159	0.9	0.97	6.94	-0.60959	-0.33082	0.05696	0.79694746	-11475.4
Acetone	-2.19	0.9	0.97	6.94	-0.55717	-0.31808	0.03473	1.09041901	-5249.3638
DIETHYL ETHER	-2.031	0.9	0.97	6.94	-0.62564	-0.32133	0.09976	0.43419498	-6350.0684
METHYL METHACRYLATE	-2.208	0.9	0.97	6.94	-0.6154	-0.32758	-0.00119	0.64353794	-9398.0624
Chloroacetone	-1.938	0.9	0.97	6.94	-0.5315	-0.33056	0.00616	0.87229665	-17751.431
N-BUTYL ACETATE	-2.219	0.9	0.97	6.94	-0.62731	-0.34637	0.06014	0.62318974	-10499.754
2-HEPTANONE	-2.188	0.9	0.97	6.94	-0.56542	-0.31564	0.03547	0.99533591	-9523.0342
4-Heptanone	-2.176	0.9	0.97	6.94	-0.56146	-0.31039	0.03512	1.06729309	-9523.0054
6-METHYL-5-HEPTEN-2-ONE	-2.114	0.9	0.97	6.94	-0.56584	-0.28215	0.0352	1.05658336	-10558.161
3-PENTANONE	-2.143	0.9	0.97	6.94	-0.56638	-0.31673	0.03708	1.01191668	-7386.3028
ISOPROPYL ACETATE	-2.225	0.9	0.97	6.94	-0.62658	-0.34444	0.06121	0.68826134	-9431.5003

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
METHYL VINYL KETONE	-2.19	0.9	0.97	6.94	-0.55668	-0.32424	-0.01687	1.03493657	-6284.4618
METHYL ACRYLATE	-2.195	0.9	0.97	6.94	-0.62028	-0.34645	-0.00786	0.57305282	-8329.4875
ETHYL ACRYLATE	-2.273	0.9	0.97	6.94	-0.62125	-0.34579	-0.00804	0.64677909	-9397.9583
Butyl butyrate	-2.214	0.9	0.97	6.94	-0.6194	-0.3408	0.05816	0.68955821	-12636.606
Butyl ethyl ketone	-2.197	0.9	0.97	6.94	-0.56693	-0.31606	0.03662	0.98805446	-9523.1054
ISOAMYL ACETATE	-2.23	0.9	0.97	6.94	-0.62714	-0.346	0.06056	0.61117322	-11568.188
PROPYLENE CARBONATE	-1.789	1.16	1.01	22.9	-0.59344	-0.3719	0.07808	2.16718404	-10375.104
Ethylene carbonate	-1.747	1.16	1.01	22.9	-0.59131	-0.37523	0.07999	2.0916744	-9306.486
VINYLENE CARBONATE	-1.61	1.16	1.01	22.9	-0.59085	-0.31578	0.03559	1.83682274	-9273.2915
Fluoroethylene carbonate	-1.569	1.16	1.01	22.9	-0.57386	-0.39223	0.06581	1.91022747	-12004.26
Dimethyl carbonate	-1.454	1.16	1.01	22.9	-0.67204	-0.36873	0.09098	0.07937243	-9339.1429
DIETHYL CARBONATE	-1.492	1.16	1.01	22.9	-0.6787	-0.36679	0.09279	0.15446412	-11476.157
Ethyl methyl carbonate	-1.488	1.16	1.01	22.9	-0.67569	-0.36782	0.0896	0.29941679	-10407.644
FURAN	-0.884	1.16	1.01	22.9	-0.48625	-0.28532	0.05396	0.2187243	-6251.9579
2,5-DIMETHYLFURAN	-1.059	1.16	1.01	22.9	-0.50904	-0.259	0.06162	0.0179921	-8389.1973
TETRAHYDROFURAN	-1.454	1.16	1.01	22.9	-0.62428	-0.31168	0.09161	0.70868651	-6317.4253
2-METHYLTETRAHYDROFURAN	-1.492	1.16	1.01	22.9	-0.63458	-0.31027	0.09209	0.65453619	-7386.0098
TETRAHYDROPYRAN	-1.411	1.16	1.01	22.9	-0.62007	-0.3143	0.09446	0.56267776	-7386.0963
1,3-DIOXOLANE	-1.273	1.16	1.01	22.9	-0.62864	-0.32401	0.10229	0.46742258	-7293.3309
1,4-DIOXANE	-1.25	1.16	1.01	22.9	-0.61676	-0.30415	0.1073	1.04E-05	-8361.789
Ethoxymethoxymethane	-1.283	1.16	1.01	22.9	-0.63798	-0.34033	0.09866	0.14969769	-8394.449
ETHYL ACETATE	-1.574	1.16	1.01	22.9	-0.62731	-0.34685	0.05997	0.64684699	-8362.9517
ISOPROPYL ACETATE	-1.585	1.16	1.01	22.9	-0.62661	-0.34444	0.0612	0.68800247	-9431.5003
METHYL PROPIONATE	-1.524	1.16	1.01	22.9	-0.62146	-0.34869	0.06332	0.63677806	-8362.8859
ETHYL PROPIONATE	-1.571	1.16	1.01	22.9	-0.62261	-0.34726	0.06069	0.68798682	-9431.3541
METHYL FORMATE	-1.444	1.16	1.01	22.9	-0.60582	-0.36295	0.05505	0.68750544	-6225.7198
ETHYL BUTYRATE	-1.542	1.16	1.01	22.9	-0.63193	-0.34311	0.05741	0.62912016	-10499.797
METHYL ISOBUTYRATE	-1.564	1.16	1.01	22.9	-0.61939	-0.34384	0.05919	0.65885942	-9431.3249
VINYL ACETATE	-1.454	1.16	1.01	22.9	-0.60014	-0.31114	0.02966	0.64433921	-8329.5868
Gamma-Butyrolactone	-1.777	1.16	1.01	22.9	-0.58246	-0.34261	0.05711	1.77086298	-8330.2469
Gamma-Valerolactone	-1.82	1.16	1.01	22.9	-0.58352	-0.34066	0.05797	1.78595247	-9398.8408
Delta-Valerolactone	-1.863	1.16	1.01	22.9	-0.58226	-0.33582	0.04675	1.79851237	-9398.6211
Epsilon-caprolactone	-1.9	1.16	1.01	22.9	-0.5888	-0.33631	0.05707	1.81728849	-10467.061

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
Gamma-Hexanolactone	-1.833	1.16	1.01	22.9	-0.58392	-0.33992	0.0582	1.81361705	-10467.253
TRIMETHYL PHOSPHATE	-2.081	1.16	1.01	22.9	-1.08448	-0.3538	0.09285	0.47819921	-20719.462
TRIETHYL PHOSPHATE	-2.177	1.16	1.01	22.9	-1.08833	-0.34517	0.08308	0.94269476	-23924.985
SULFOLANE	-1.879	1.16	1.01	22.9	-0.95081	-0.33695	0.06799	2.06428499	-19192.049
METHYL METHANESULFONATE	-1.527	1.16	1.01	22.9	-0.78189	-0.36367	0.09127	1.2015052	-19132.267
METHYL BENZENESULFONATE	-1.68	1.16	1.01	22.9	-0.92765	-0.333	-0.01708	1.34600337	-24343.813
1,3-Propanesultone	-1.721	1.16	1.01	22.9	-0.77035	-0.3632	0.06727	2.31386941	-20168.047
1,4-BUTANE SULTONE	-1.752	1.16	1.01	22.9	-0.78202	-0.36768	0.07208	2.14753013	-21236.86
Dimethyl sulfoxide	-2.183	1.16	1.01	22.9	-0.97186	-0.28391	0.08537	1.61570745	-15043.485
SUCCINIC ANHYDRIDE	-1.45	1.16	1.01	22.9	-0.5436	-0.36492	0.01241	1.71813593	-10342.727
CYCLOHEXANONE	-1.654	1.16	1.01	22.9	-0.55268	-0.30724	0.02926	1.19994939	-8422.2456
CAPROLACTONE	-1.9	1.16	1.01	22.9	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Propiolactone	-1.592	1.16	1.01	22.9	-0.56703	-0.35549	0.05287	1.63151562	-7261.067
CYCLOPENTANONE	-1.632	1.16	1.01	22.9	-0.55019	-0.30855	0.02666	1.12642153	-7353.6424
Diketene	-1.481	1.16	1.01	22.9	-0.54443	-0.32081	0.03875	1.37862436	-8296.2676
ACETOPHENONE	-1.635	1.16	1.01	22.9	-0.55641	-0.31593	-0.02263	1.1340558	-10461.15
Guaiacol	-2.077	1.16	1.01	22.9	-0.5759	-0.26469	0.03188	0.818535	-11469.446
Benzaldehyde	-1.57	1.16	1.01	22.9	-0.5358	-0.32133	-0.03058	1.20204662	-9392.4809
2-METHYLCYCLOHEXANONE	-1.679	1.16	1.01	22.9	-0.55411	-0.30368	0.03108	1.16980232	-9490.6559
METHYL METHANESULFONATE	-1.527	1.16	1.01	22.9	-0.78188	-0.36366	0.09128	1.20127447	-19132.267
DIETHYLSULFATE	-1.76	1.16	1.01	22.9	-0.92686	-0.36695	0.08888	0.5422029	-23313.549
2,3-butanedione	-1.235	1.16	1.01	22.9	-0.52628	-0.31341	-0.04212	0.05681995	-8329.6147
ACETOPHENONE	-1.635	1.16	1.01	22.9	-0.55644	-0.31592	-0.02264	1.13390682	-10461.15
BENZYL BENZOATE	-2.139	1.16	1.01	22.9	-0.62251	-0.3091	-0.01378	0.75661917	-18786.228
DIPHENYL ETHER	-1.12	1.16	1.01	22.9	-0.55198	-0.27253	0.01595	0.32082062	-14636.695
PENTANAL	-1.438	1.16	1.01	22.9	-0.54101	-0.32531	0.02944	0.99136767	-7385.9481
2-Methoxyethyl acetate	-1.54	1.16	1.01	22.9	-0.60959	-0.33082	0.05696	0.79694746	-11475.4
Acetone	-1.6	1.16	1.01	22.9	-0.55717	-0.31808	0.03473	1.09041901	-5249.3638
DIETHYL ETHER	-1.473	1.16	1.01	22.9	-0.62564	-0.32133	0.09976	0.43419498	-6350.0684
METHYL METHACRYLATE	-1.583	1.16	1.01	22.9	-0.6154	-0.32758	-0.00119	0.64353794	-9398.0624
Chloroacetone	-1.399	1.16	1.01	22.9	-0.5315	-0.33056	0.00616	0.87229665	-17751.431
N-BUTYL ACETATE	-1.589	1.16	1.01	22.9	-0.62731	-0.34637	0.06014	0.62318974	-10499.754
2-HEPTANONE	-1.58	1.16	1.01	22.9	-0.56542	-0.31564	0.03547	0.99533591	-9523.0342

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
4-Heptanone	-1.553	1.16	1.01	22.9	-0.56146	-0.31039	0.03512	1.06729309	-9523.0054
6-METHYL-5-HEPTEN-2-ONE	-1.504	1.16	1.01	22.9	-0.56584	-0.28215	0.0352	1.05658336	-10558.161
3-PENTANONE	-1.541	1.16	1.01	22.9	-0.56638	-0.31673	0.03708	1.01191668	-7386.3028
ISOPROPYL ACETATE	-1.587	1.16	1.01	22.9	-0.62658	-0.34444	0.06121	0.68826134	-9431.5003
METHYL VINYL KETONE	-1.58	1.16	1.01	22.9	-0.55668	-0.32424	-0.01687	1.03493657	-6284.4618
METHYL ACRYLATE	-1.57	1.16	1.01	22.9	-0.62028	-0.34645	-0.00786	0.57305282	-8329.4875
ETHYL ACRYLATE	-1.632	1.16	1.01	22.9	-0.62125	-0.34579	-0.00804	0.64677909	-9397.9583
Butyl butyrate	-1.578	1.16	1.01	22.9	-0.6194	-0.3408	0.05816	0.68955821	-12636.606
Butyl ethyl ketone	-1.582	1.16	1.01	22.9	-0.56693	-0.31606	0.03662	0.98805446	-9523.1054
ISOAMYL ACETATE	-1.601	1.16	1.01	22.9	-0.62714	-0.346	0.06056	0.61117322	-11568.188
PROPYLENE CARBONATE	-1.397	1.52	0.91	39.1	-0.59344	-0.3719	0.07808	2.16718404	-10375.104
Ethylene carbonate	-1.365	1.52	0.91	39.1	-0.59131	-0.37523	0.07999	2.0916744	-9306.486
VINYLENE CARBONATE	-1.246	1.52	0.91	39.1	-0.59085	-0.31578	0.03559	1.83682274	-9273.2915
Fluoroethylene carbonate	-1.21	1.52	0.91	39.1	-0.57386	-0.39223	0.06581	1.91022747	-12004.26
Dimethyl carbonate	-1.078	1.52	0.91	39.1	-0.67204	-0.36873	0.09098	0.07937243	-9339.1429
DIETHYL CARBONATE	-1.106	1.52	0.91	39.1	-0.6787	-0.36679	0.09279	0.15446412	-11476.157
Ethyl methyl carbonate	-1.108	1.52	0.91	39.1	-0.67569	-0.36782	0.0896	0.29941679	-10407.644
FURAN	-0.605	1.52	0.91	39.1	-0.48625	-0.28532	0.05396	0.2187243	-6251.9579
2,5-DIMETHYLFURAN	-0.741	1.52	0.91	39.1	-0.50904	-0.259	0.06162	0.0179921	-8389.1973
TETRAHYDROFURAN	-1.065	1.52	0.91	39.1	-0.62428	-0.31168	0.09161	0.70868651	-6317.4253
2-METHYLTETRAHYDROFURAN	-1.095	1.52	0.91	39.1	-0.63458	-0.31027	0.09209	0.65453619	-7386.0098
TETRAHYDROPYRAN	-1.028	1.52	0.91	39.1	-0.62007	-0.3143	0.09446	0.56267776	-7386.0963
1,3-DIOXOLANE	-0.911	1.52	0.91	39.1	-0.62864	-0.32401	0.10229	0.46742258	-7293.3309
1,4-DIOXANE	-0.888	1.52	0.91	39.1	-0.61676	-0.30415	0.1073	1.04E-05	-8361.789
Ethoxymethoxymethane	-0.881	1.52	0.91	39.1	-0.63798	-0.34033	0.09866	0.14969769	-8394.449
ETHYL ACETATE	-1.185	1.52	0.91	39.1	-0.62731	-0.34685	0.05997	0.64684699	-8362.9517
ISOPROPYL ACETATE	-1.187	1.52	0.91	39.1	-0.62661	-0.34444	0.0612	0.68800247	-9431.5003
METHYL PROPIONATE	-1.133	1.52	0.91	39.1	-0.62146	-0.34869	0.06332	0.63677806	-8362.8859
ETHYL PROPIONATE	-1.172	1.52	0.91	39.1	-0.62261	-0.34726	0.06069	0.68798682	-9431.3541
METHYL FORMATE	-1.082	1.52	0.91	39.1	-0.60582	-0.36295	0.05505	0.68750544	-6225.7198
ETHYL BUTYRATE	-1.142	1.52	0.91	39.1	-0.63193	-0.34311	0.05741	0.62912016	-10499.797
METHYL ISOBUTYRATE	-1.165	1.52	0.91	39.1	-0.61939	-0.34384	0.05919	0.65885942	-9431.3249
VINYL ACETATE	-1.076	1.52	0.91	39.1	-0.60014	-0.31114	0.02966	0.64433921	-8329.5868

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
Gamma-Butyrolactone	-1.377	1.52	0.91	39.1	-0.58246	-0.34261	0.05711	1.77086298	-8330.2469
Gamma-Valerolactone	-1.412	1.52	0.91	39.1	-0.58352	-0.34066	0.05797	1.78595247	-9398.8408
Delta-Valerolactone	-1.456	1.52	0.91	39.1	-0.58226	-0.33582	0.04675	1.79851237	-9398.6211
Epsilon-caprolactone	-1.478	1.52	0.91	39.1	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Gamma-Hexanolactone	-1.426	1.52	0.91	39.1	-0.58392	-0.33992	0.0582	1.81361705	-10467.253
TRIMETHYL PHOSPHATE	-1.611	1.52	0.91	39.1	-1.08448	-0.3538	0.09285	0.47819921	-20719.462
TRIETHYL PHOSPHATE	-1.71	1.52	0.91	39.1	-1.08833	-0.34517	0.08308	0.94269476	-23924.985
SULFOLANE	-1.45	1.52	0.91	39.1	-0.95081	-0.33695	0.06799	2.06428499	-19192.049
METHYL METHANESULFONATE	-1.141	1.52	0.91	39.1	-0.78189	-0.36367	0.09127	1.2015052	-19132.267
METHYL BENZENESULFONATE	-1.263	1.52	0.91	39.1	-0.92765	-0.333	-0.01708	1.34600337	-24343.813
1,3-Propanesultone	-1.327	1.52	0.91	39.1	-0.77035	-0.3632	0.06727	2.31386941	-20168.047
1,4-BUTANE SULTONE	-1.356	1.52	0.91	39.1	-0.78202	-0.36768	0.07208	2.14753013	-21236.86
Dimethyl sulfoxide	-1.725	1.52	0.91	39.1	-0.97186	-0.28391	0.08537	1.61570745	-15043.485
SUCCINIC ANHYDRIDE	-1.095	1.52	0.91	39.1	-0.5436	-0.36492	0.01241	1.71813593	-10342.727
CYCLOHEXANONE	-1.265	1.52	0.91	39.1	-0.55268	-0.30724	0.02926	1.19994939	-8422.2456
CAPROLACTONE	-1.478	1.52	0.91	39.1	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Propiolactone	-1.228	1.52	0.91	39.1	-0.56703	-0.35549	0.05287	1.63151562	-7261.067
CYCLOPENTANONE	-1.247	1.52	0.91	39.1	-0.55019	-0.30855	0.02666	1.12642153	-7353.6424
Diketene	-1.126	1.52	0.91	39.1	-0.54443	-0.32081	0.03875	1.37862436	-8296.2676
ACETOPHENONE	-1.235	1.52	0.91	39.1	-0.55641	-0.31593	-0.02263	1.1340558	-10461.15
Guaiacol	-1.535	1.52	0.91	39.1	-0.5759	-0.26469	0.03188	0.818535	-11469.446
Benzaldehyde	-1.188	1.52	0.91	39.1	-0.5358	-0.32133	-0.03058	1.20204662	-9392.4809
2-METHYLCYCLOHEXANONE	-1.286	1.52	0.91	39.1	-0.55411	-0.30368	0.03108	1.16980232	-9490.6559
METHYL METHANESULFONATE	-1.141	1.52	0.91	39.1	-0.78188	-0.36366	0.09128	1.20127447	-19132.267
DIETHYLSULFATE	-1.384	1.52	0.91	39.1	-0.92686	-0.36695	0.08888	0.5422029	-23313.549
2,3-butanedione	-0.899	1.52	0.91	39.1	-0.52628	-0.31341	-0.04212	0.05681995	-8329.6147
ACETOPHENONE	-1.235	1.52	0.91	39.1	-0.55644	-0.31592	-0.02264	1.13390682	-10461.15
BENZYL BENZOATE	-1.682	1.52	0.91	39.1	-0.62251	-0.3091	-0.01378	0.75661917	-18786.228
DIPHENYL ETHER	-0.758	1.52	0.91	39.1	-0.55198	-0.27253	0.01595	0.32082062	-14636.695
PENTANAL	-1.079	1.52	0.91	39.1	-0.54101	-0.32531	0.02944	0.99136767	-7385.9481
2-Methoxyethyl acetate	-1.139	1.52	0.91	39.1	-0.60959	-0.33082	0.05696	0.79694746	-11475.4
Acetone	-1.219	1.52	0.91	39.1	-0.55717	-0.31808	0.03473	1.09041901	-5249.3638
DIETHYL ETHER	-1.056	1.52	0.91	39.1	-0.62564	-0.32133	0.09976	0.43419498	-6350.0684

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
METHYL METHACRYLATE	-1.186	1.52	0.91	39.1	-0.6154	-0.32758	-0.00119	0.64353794	-9398.0624
Chloroacetone	-1.047	1.52	0.91	39.1	-0.5315	-0.33056	0.00616	0.87229665	-17751.431
N-BUTYL ACETATE	-1.202	1.52	0.91	39.1	-0.62731	-0.34637	0.06014	0.62318974	-10499.754
2-HEPTANONE	-1.187	1.52	0.91	39.1	-0.56542	-0.31564	0.03547	0.99533591	-9523.0342
4-Heptanone	-1.159	1.52	0.91	39.1	-0.56146	-0.31039	0.03512	1.06729309	-9523.0054
6-METHYL-5-HEPTEN-2-ONE	-1.111	1.52	0.91	39.1	-0.56584	-0.28215	0.0352	1.05658336	-10558.161
3-PENTANONE	-1.152	1.52	0.91	39.1	-0.56638	-0.31673	0.03708	1.01191668	-7386.3028
ISOPROPYL ACETATE	-1.19	1.52	0.91	39.1	-0.62658	-0.34444	0.06121	0.68826134	-9431.5003
METHYL VINYL KETONE	-1.187	1.52	0.91	39.1	-0.55668	-0.32424	-0.01687	1.03493657	-6284.4618
METHYL ACRYLATE	-1.178	1.52	0.91	39.1	-0.62028	-0.34645	-0.00786	0.57305282	-8329.4875
ETHYL ACRYLATE	-1.237	1.52	0.91	39.1	-0.62125	-0.34579	-0.00804	0.64677909	-9397.9583
Butyl butyrate	-1.189	1.52	0.91	39.1	-0.6194	-0.3408	0.05816	0.68955821	-12636.606
Butyl ethyl ketone	-1.174	1.52	0.91	39.1	-0.56693	-0.31606	0.03662	0.98805446	-9523.1054
ISOAMYL ACETATE	-1.215	1.52	0.91	39.1	-0.62714	-0.346	0.06056	0.61117322	-11568.188
PROPYLENE CARBONATE	-1.307	1.66	0.89	85.47	-0.59344	-0.3719	0.07808	2.16718404	-10375.104
Ethylene carbonate	-1.272	1.66	0.89	85.47	-0.59131	-0.37523	0.07999	2.0916744	-9306.486
VINYLENE CARBONATE	-1.157	1.66	0.89	85.47	-0.59085	-0.31578	0.03559	1.83682274	-9273.2915
Fluoroethylene carbonate	-1.129	1.66	0.89	85.47	-0.57386	-0.39223	0.06581	1.91022747	-12004.26
Dimethyl carbonate	-0.968	1.66	0.89	85.47	-0.67204	-0.36873	0.09098	0.07937243	-9339.1429
DIETHYL CARBONATE	-1.01	1.66	0.89	85.47	-0.6787	-0.36679	0.09279	0.15446412	-11476.157
Ethyl methyl carbonate	-1.006	1.66	0.89	85.47	-0.67569	-0.36782	0.0896	0.29941679	-10407.644
FURAN	-0.545	1.66	0.89	85.47	-0.48625	-0.28532	0.05396	0.2187243	-6251.9579
2,5-DIMETHYLFURAN	-0.649	1.66	0.89	85.47	-0.50904	-0.259	0.06162	0.0179921	-8389.1973
TETRAHYDROFURAN	-0.978	1.66	0.89	85.47	-0.62428	-0.31168	0.09161	0.70868651	-6317.4253
2-METHYLTETRAHYDROFURAN	-1.007	1.66	0.89	85.47	-0.63458	-0.31027	0.09209	0.65453619	-7386.0098
TETRAHYDROPYRAN	-0.94	1.66	0.89	85.47	-0.62007	-0.3143	0.09446	0.56267776	-7386.0963
1,3-DIOXOLANE	-0.836	1.66	0.89	85.47	-0.62864	-0.32401	0.10229	0.46742258	-7293.3309
1,4-DIOXANE	-0.801	1.66	0.89	85.47	-0.61676	-0.30415	0.1073	1.04E-05	-8361.789
Ethoxymethoxymethane	-0.809	1.66	0.89	85.47	-0.63798	-0.34033	0.09866	0.14969769	-8394.449
ETHYL ACETATE	-1.083	1.66	0.89	85.47	-0.62731	-0.34685	0.05997	0.64684699	-8362.9517
ISOPROPYL ACETATE	-1.093	1.66	0.89	85.47	-0.62661	-0.34444	0.0612	0.68800247	-9431.5003
METHYL PROPIONATE	-1.03	1.66	0.89	85.47	-0.62146	-0.34869	0.06332	0.63677806	-8362.8859
ETHYL PROPIONATE	-1.072	1.66	0.89	85.47	-0.62261	-0.34726	0.06069	0.68798682	-9431.3541

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
METHYL FORMATE	-0.981	1.66	0.89	85.47	-0.60582	-0.36295	0.05505	0.68750544	-6225.7198
ETHYL BUTYRATE	-1.04	1.66	0.89	85.47	-0.63193	-0.34311	0.05741	0.62912016	-10499.797
METHYL ISOBUTYRATE	-1.063	1.66	0.89	85.47	-0.61939	-0.34384	0.05919	0.65885942	-9431.3249
VINYL ACETATE	-0.984	1.66	0.89	85.47	-0.60014	-0.31114	0.02966	0.64433921	-8329.5868
Gamma-Butyrolactone	-1.292	1.66	0.89	85.47	-0.58246	-0.34261	0.05711	1.77086298	-8330.2469
Gamma-Valerolactone	-1.328	1.66	0.89	85.47	-0.58352	-0.34066	0.05797	1.78595247	-9398.8408
Delta-Valerolactone	-1.351	1.66	0.89	85.47	-0.58226	-0.33582	0.04675	1.79851237	-9398.6211
Epsilon-caprolactone	-1.389	1.66	0.89	85.47	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Gamma-Hexanolactone	-1.333	1.66	0.89	85.47	-0.58392	-0.33992	0.0582	1.81361705	-10467.253
TRIMETHYL PHOSPHATE	-1.512	1.66	0.89	85.47	-1.08448	-0.3538	0.09285	0.47819921	-20719.462
TRIETHYL PHOSPHATE	-1.595	1.66	0.89	85.47	-1.08833	-0.34517	0.08308	0.94269476	-23924.985
SULFOLANE	-1.35	1.66	0.89	85.47	-0.95081	-0.33695	0.06799	2.06428499	-19192.049
METHYL METHANESULFONATE	-1.056	1.66	0.89	85.47	-0.78189	-0.36367	0.09127	1.2015052	-19132.267
METHYL BENZENESULFONATE	-1.18	1.66	0.89	85.47	-0.92765	-0.333	-0.01708	1.34600337	-24343.813
1,3-Propanesultone	-1.24	1.66	0.89	85.47	-0.77035	-0.3632	0.06727	2.31386941	-20168.047
1,4-BUTANE SULTONE	-1.281	1.66	0.89	85.47	-0.78202	-0.36768	0.07208	2.14753013	-21236.86
Dimethyl sulfoxide	-1.59	1.66	0.89	85.47	-0.97186	-0.28391	0.08537	1.61570745	-15043.485
SUCCINIC ANHYDRIDE	-1.025	1.66	0.89	85.47	-0.5436	-0.36492	0.01241	1.71813593	-10342.727
CYCLOHEXANONE	-1.158	1.66	0.89	85.47	-0.55268	-0.30724	0.02926	1.19994939	-8422.2456
CAPROLACTONE	-1.389	1.66	0.89	85.47	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Propiolactone	-1.14	1.66	0.89	85.47	-0.56703	-0.35549	0.05287	1.63151562	-7261.067
CYCLOPENTANONE	-1.144	1.66	0.89	85.47	-0.55019	-0.30855	0.02666	1.12642153	-7353.6424
Diketene	-1.049	1.66	0.89	85.47	-0.54443	-0.32081	0.03875	1.37862436	-8296.2676
ACETOPHENONE	-1.131	1.66	0.89	85.47	-0.55641	-0.31593	-0.02263	1.1340558	-10461.15
Guaiacol	-1.398	1.66	0.89	85.47	-0.5759	-0.26469	0.03188	0.818535	-11469.446
Benzaldehyde	-1.085	1.66	0.89	85.47	-0.5358	-0.32133	-0.03058	1.20204662	-9392.4809
2-METHYLCYCLOHEXANONE	-1.178	1.66	0.89	85.47	-0.55411	-0.30368	0.03108	1.16980232	-9490.6559
METHYL METHANESULFONATE	-1.056	1.66	0.89	85.47	-0.78188	-0.36366	0.09128	1.20127447	-19132.267
DIETHYLSULFATE	-1.299	1.66	0.89	85.47	-0.92686	-0.36695	0.08888	0.5422029	-23313.549
2,3-butanedione	-0.807	1.66	0.89	85.47	-0.52628	-0.31341	-0.04212	0.05681995	-8329.6147
ACETOPHENONE	-1.132	1.66	0.89	85.47	-0.55644	-0.31592	-0.02264	1.13390682	-10461.15
BENZYL BENZOATE	-1.591	1.66	0.89	85.47	-0.62251	-0.3091	-0.01378	0.75661917	-18786.228
DIPHENYL ETHER	-0.738	1.66	0.89	85.47	-0.55198	-0.27253	0.01595	0.32082062	-14636.695

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
PENTANAL	-0.984	1.66	0.89	85.47	-0.54101	-0.32531	0.02944	0.99136767	-7385.9481
2-Methoxyethyl acetate	-1.04	1.66	0.89	85.47	-0.60959	-0.33082	0.05696	0.79694746	-11475.4
Acetone	-1.117	1.66	0.89	85.47	-0.55717	-0.31808	0.03473	1.09041901	-5249.3638
DIETHYL ETHER	-0.96	1.66	0.89	85.47	-0.62564	-0.32133	0.09976	0.43419498	-6350.0684
METHYL METHACRYLATE	-1.078	1.66	0.89	85.47	-0.6154	-0.32758	-0.00119	0.64353794	-9398.0624
Chloroacetone	-0.964	1.66	0.89	85.47	-0.5315	-0.33056	0.00616	0.87229665	-17751.431
N-BUTYL ACETATE	-1.098	1.66	0.89	85.47	-0.62731	-0.34637	0.06014	0.62318974	-10499.754
2-HEPTANONE	-1.081	1.66	0.89	85.47	-0.56542	-0.31564	0.03547	0.99533591	-9523.0342
4-Heptanone	-1.06	1.66	0.89	85.47	-0.56146	-0.31039	0.03512	1.06729309	-9523.0054
6-METHYL-5-HEPTEN-2-ONE	-1.015	1.66	0.89	85.47	-0.56584	-0.28215	0.0352	1.05658336	-10558.161
3-PENTANONE	-1.055	1.66	0.89	85.47	-0.56638	-0.31673	0.03708	1.01191668	-7386.3028
ISOPROPYL ACETATE	-1.095	1.66	0.89	85.47	-0.62658	-0.34444	0.06121	0.68826134	-9431.5003
METHYL VINYL KETONE	-1.084	1.66	0.89	85.47	-0.55668	-0.32424	-0.01687	1.03493657	-6284.4618
METHYL ACRYLATE	-1.069	1.66	0.89	85.47	-0.62028	-0.34645	-0.00786	0.57305282	-8329.4875
ETHYL ACRYLATE	-1.134	1.66	0.89	85.47	-0.62125	-0.34579	-0.00804	0.64677909	-9397.9583
Butyl butyrate	-1.087	1.66	0.89	85.47	-0.6194	-0.3408	0.05816	0.68955821	-12636.606
Butyl ethyl ketone	-1.089	1.66	0.89	85.47	-0.56693	-0.31606	0.03662	0.98805446	-9523.1054
ISOAMYL ACETATE	-1.112	1.66	0.89	85.47	-0.62714	-0.346	0.06056	0.61117322	-11568.188
PROPYLENE CARBONATE	-1.165	1.81	0.86	132.91	-0.59344	-0.3719	0.07808	2.16718404	-10375.104
Ethylene carbonate	-1.135	1.81	0.86	132.91	-0.59131	-0.37523	0.07999	2.0916744	-9306.486
VINYLENE CARBONATE	-1.025	1.81	0.86	132.91	-0.59085	-0.31578	0.03559	1.83682274	-9273.2915
Fluoroethylene carbonate	-1.001	1.81	0.86	132.91	-0.57386	-0.39223	0.06581	1.91022747	-12004.26
Dimethyl carbonate	-0.842	1.81	0.86	132.91	-0.67204	-0.36873	0.09098	0.07937243	-9339.1429
DIETHYL CARBONATE	-0.877	1.81	0.86	132.91	-0.6787	-0.36679	0.09279	0.15446412	-11476.157
Ethyl methyl carbonate	-0.878	1.81	0.86	132.91	-0.67569	-0.36782	0.0896	0.29941679	-10407.644
FURAN	-0.461	1.81	0.86	132.91	-0.48625	-0.28532	0.05396	0.2187243	-6251.9579
2,5-DIMETHYLFURAN	-0.56	1.81	0.86	132.91	-0.50904	-0.259	0.06162	0.0179921	-8389.1973
TETRAHYDROFURAN	-0.851	1.81	0.86	132.91	-0.62428	-0.31168	0.09161	0.70868651	-6317.4253
2-METHYLTETRAHYDROFURAN	-0.874	1.81	0.86	132.91	-0.63458	-0.31027	0.09209	0.65453619	-7386.0098
TETRAHYDROPYRAN	-0.821	1.81	0.86	132.91	-0.62007	-0.3143	0.09446	0.56267776	-7386.0963
1,3-DIOXOLANE	-0.715	1.81	0.86	132.91	-0.62864	-0.32401	0.10229	0.46742258	-7293.3309
1,4-DIOXANE	-0.688	1.81	0.86	132.91	-0.61676	-0.30415	0.1073	1.04E-05	-8361.789
Ethoxymethoxymethane	-0.682	1.81	0.86	132.91	-0.63798	-0.34033	0.09866	0.14969769	-8394.449

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
ETHYL ACETATE	-0.95	1.81	0.86	132.91	-0.62731	-0.34685	0.05997	0.64684699	-8362.9517
ISOPROPYL ACETATE	-0.958	1.81	0.86	132.91	-0.62661	-0.34444	0.0612	0.68800247	-9431.5003
METHYL PROPIONATE	-0.896	1.81	0.86	132.91	-0.62146	-0.34869	0.06332	0.63677806	-8362.8859
ETHYL PROPIONATE	-0.935	1.81	0.86	132.91	-0.62261	-0.34726	0.06069	0.68798682	-9431.3541
METHYL FORMATE	-0.861	1.81	0.86	132.91	-0.60582	-0.36295	0.05505	0.68750544	-6225.7198
ETHYL BUTYRATE	-0.902	1.81	0.86	132.91	-0.63193	-0.34311	0.05741	0.62912016	-10499.797
METHYL ISOBUTYRATE	-0.926	1.81	0.86	132.91	-0.61939	-0.34384	0.05919	0.65885942	-9431.3249
VINYL ACETATE	-0.857	1.81	0.86	132.91	-0.60014	-0.31114	0.02966	0.64433921	-8329.5868
Gamma-Butyrolactone	-1.148	1.81	0.86	132.91	-0.58246	-0.34261	0.05711	1.77086298	-8330.2469
Gamma-Valerolactone	-1.18	1.81	0.86	132.91	-0.58352	-0.34066	0.05797	1.78595247	-9398.8408
Delta-Valerolactone	-1.203	1.81	0.86	132.91	-0.58226	-0.33582	0.04675	1.79851237	-9398.6211
Epsilon-caprolactone	-1.238	1.81	0.86	132.91	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Gamma-Hexanolactone	-1.188	1.81	0.86	132.91	-0.58392	-0.33992	0.0582	1.81361705	-10467.253
TRIMETHYL PHOSPHATE	-1.346	1.81	0.86	132.91	-1.08448	-0.3538	0.09285	0.47819921	-20719.462
TRIETHYL PHOSPHATE	-1.415	1.81	0.86	132.91	-1.08833	-0.34517	0.08308	0.94269476	-23924.985
SULFOLANE	-1.2	1.81	0.86	132.91	-0.95081	-0.33695	0.06799	2.06428499	-19192.049
METHYL METHANESULFONATE	-0.928	1.81	0.86	132.91	-0.78189	-0.36367	0.09127	1.2015052	-19132.267
METHYL BENZENESULFONATE	-1.055	1.81	0.86	132.91	-0.92765	-0.333	-0.01708	1.34600337	-24343.813
1,3-Propanesultone	-1.102	1.81	0.86	132.91	-0.77035	-0.3632	0.06727	2.31386941	-20168.047
1,4-BUTANE SULTONE	-1.138	1.81	0.86	132.91	-0.78202	-0.36768	0.07208	2.14753013	-21236.86
Dimethyl sulfoxide	-1.427	1.81	0.86	132.91	-0.97186	-0.28391	0.08537	1.61570745	-15043.485
SUCCINIC ANHYDRIDE	-0.904	1.81	0.86	132.91	-0.5436	-0.36492	0.01241	1.71813593	-10342.727
CYCLOHEXANONE	-1.025	1.81	0.86	132.91	-0.55268	-0.30724	0.02926	1.19994939	-8422.2456
CAPROLACTONE	-1.238	1.81	0.86	132.91	-0.5888	-0.33631	0.05707	1.81728849	-10467.061
Propiolactone	-1.011	1.81	0.86	132.91	-0.56703	-0.35549	0.05287	1.63151562	-7261.067
CYCLOPENTANONE	-1.012	1.81	0.86	132.91	-0.55019	-0.30855	0.02666	1.12642153	-7353.6424
Diketene	-0.928	1.81	0.86	132.91	-0.54443	-0.32081	0.03875	1.37862436	-8296.2676
ACETOPHENONE	-0.997	1.81	0.86	132.91	-0.55641	-0.31593	-0.02263	1.1340558	-10461.15
Guaiacol	-1.227	1.81	0.86	132.91	-0.5759	-0.26469	0.03188	0.818535	-11469.446
Benzaldehyde	-0.958	1.81	0.86	132.91	-0.5358	-0.32133	-0.03058	1.20204662	-9392.4809
2-METHYLCYCLOHEXANONE	-1.043	1.81	0.86	132.91	-0.55411	-0.30368	0.03108	1.16980232	-9490.6559
METHYL METHANESULFONATE	-0.928	1.81	0.86	132.91	-0.78188	-0.36366	0.09128	1.20127447	-19132.267
DIETHYLSULFATE	-1.148	1.81	0.86	132.91	-0.92686	-0.36695	0.08888	0.5422029	-23313.549

name	E_{coord}	Ionic Radius	Electro-negativity	Atomic Weight	NBO Charge	HOMO	LUMO	Dipole Moment	Total Energy
2,3-butanedione	-0.7	1.81	0.86	132.91	-0.52628	-0.31341	-0.04212	0.05681995	-8329.6147
ACETOPHENONE	-0.997	1.81	0.86	132.91	-0.55644	-0.31592	-0.02264	1.13390682	-10461.15
BENZYL BENZOATE	-1.441	1.81	0.86	132.91	-0.62251	-0.3091	-0.01378	0.75661917	-18786.228
DIPHENYL ETHER	-0.638	1.81	0.86	132.91	-0.55198	-0.27253	0.01595	0.32082062	-14636.695
PENTANAL	-0.859	1.81	0.86	132.91	-0.54101	-0.32531	0.02944	0.99136767	-7385.9481
2-Methoxyethyl acetate	-0.907	1.81	0.86	132.91	-0.60959	-0.33082	0.05696	0.79694746	-11475.4
Acetone	-0.987	1.81	0.86	132.91	-0.55717	-0.31808	0.03473	1.09041901	-5249.3638
DIETHYL ETHER	-0.825	1.81	0.86	132.91	-0.62564	-0.32133	0.09976	0.43419498	-6350.0684
METHYL METHACRYLATE	-0.944	1.81	0.86	132.91	-0.6154	-0.32758	-0.00119	0.64353794	-9398.0624
Chloroacetone	-0.845	1.81	0.86	132.91	-0.5315	-0.33056	0.00616	0.87229665	-17751.431
N-BUTYL ACETATE	-0.964	1.81	0.86	132.91	-0.62731	-0.34637	0.06014	0.62318974	-10499.754
2-HEPTANONE	-0.947	1.81	0.86	132.91	-0.56542	-0.31564	0.03547	0.99533591	-9523.0342
4-Heptanone	-0.923	1.81	0.86	132.91	-0.56146	-0.31039	0.03512	1.06729309	-9523.0054
6-METHYL-5-HEPTEN-2-ONE	-0.879	1.81	0.86	132.91	-0.56584	-0.28215	0.0352	1.05658336	-10558.161
3-PENTANONE	-0.92	1.81	0.86	132.91	-0.56638	-0.31673	0.03708	1.01191668	-7386.3028
ISOPROPYL ACETATE	-0.96	1.81	0.86	132.91	-0.62658	-0.34444	0.06121	0.68826134	-9431.5003
METHYL VINYL KETONE	-0.959	1.81	0.86	132.91	-0.55668	-0.32424	-0.01687	1.03493657	-6284.4618
METHYL ACRYLATE	-0.938	1.81	0.86	132.91	-0.62028	-0.34645	-0.00786	0.57305282	-8329.4875
ETHYL ACRYLATE	-1.001	1.81	0.86	132.91	-0.62125	-0.34579	-0.00804	0.64677909	-9397.9583
Butyl butyrate	-0.95	1.81	0.86	132.91	-0.6194	-0.3408	0.05816	0.68955821	-12636.606
Butyl ethyl ketone	-0.948	1.81	0.86	132.91	-0.56693	-0.31606	0.03662	0.98805446	-9523.1054
ISOAMYL ACETATE	-0.977	1.81	0.86	132.91	-0.62714	-0.346	0.06056	0.61117322	-11568.188

Appendix B

Supporting Information for Chapter 3

B.1 Optimized structures

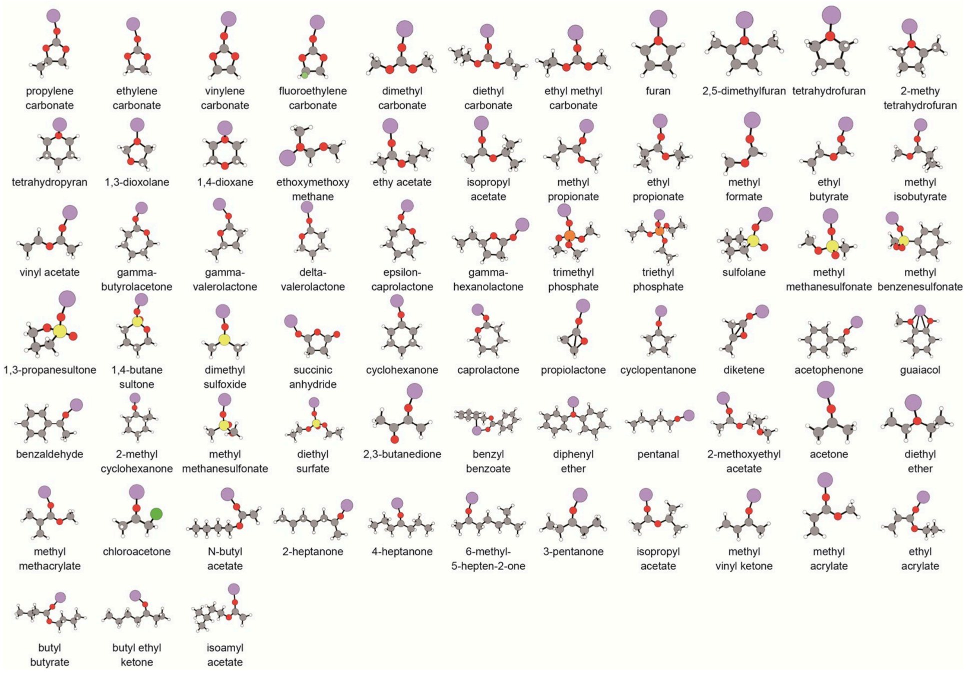


Figure B.1: Optimized structures of 70 Li-coordinated solvent systems. The gray, red, yellow, orange, light green, green, purple and white sphere corresponds to C, O, S, P, F, Cl, Li, and H atom, respectively. M06-2X with Def2-SVP basis set was used in the DFT calculations

B.2 Cross Validation Error for each K

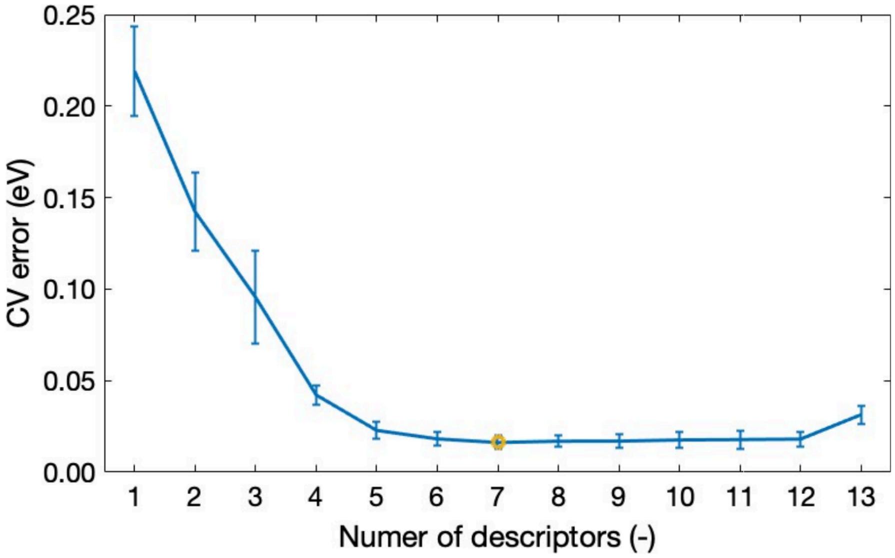


Figure B.2: The CV error (in eV) dependence on the number of descriptors when the ES-GP regression model was used. The yellow circle shows the minimum in the CV error.

Appendix C

Supporting Information for Chapter 4

C.1 Posterior distribution and free energy calculation using Exchange MCMC

When performing Bayesian inference, it is often difficult to calculate the posterior distribution. In such cases, we may use MCMC to sample from the posterior distribution and use the sampled results to perform approximate calculations. In this study, we selected exchange MCMC[57]. The posterior distributions $p(\Theta \mid \mathcal{D}, K)$ in the main content are calculated using exchange MCMC using the following procedure.

1. Using L reverse temperatures, define $\{\beta_l\}_{l=1}^{L}, 0 = \beta_1 < \beta_2 < \cdots < \beta_{L-1} < \beta_L = 1$.
2. Let $p_{\beta_l}(\Theta_l \mid \mathcal{D}, K) \propto \exp(-\beta_l E(\Theta_l))\, p(\Theta_l)$ be the posterior distribution with inverse temperature.
3. For each temperature distribution, update $\{\Theta_l\}_{l=1}^{L}$ using the Metropolis algorithm.
4. Create a pair of neighboring temperatures and compute the following probability r.

$$r = \frac{p_{\beta_l+1}(\Theta_l)\, p_{\beta_l}(\Theta_{l+1})}{p_{\beta_l}(\Theta_l)\, p_{\beta_l+1}(\Theta_{l+1})}$$

5. Generate a uniform random number R in $0 \leq R < 1$ and, if $R < r$, exchange Θ_{l+1} and Θ_l.
6. Return to Step 3.

The posterior distribution obtained by this procedure is the one obtained by p_{β_L}.

In order to perform the integration of the free energy with respect to Θ by sampling the MCMC method, we consider the following $f_\beta(K)$ with the introduction of the inverse temperature β into the free energy.

$$f_\beta(K) = -\log \int d\Theta \exp(-\beta E(\Theta)) p(\Theta \mid \mathcal{D}, K)$$

Introducing $f_\beta(K)$, Free energy $F(K)$ can be written as follows.

$$\begin{aligned}
F(K) &= f_1(K) \\
&= \int_0^1 d\beta \frac{\partial f_\beta}{\partial \beta} \\
&= \int_0^1 d\beta \frac{\int d\Theta E(\Theta) \exp(-\beta E(\Theta)) p(\Theta \mid \mathcal{D}, K)}{\int d\Theta \exp(-\beta E(\Theta)) p(\Theta \mid \mathcal{D}, K)} \\
&= \int_0^1 d\beta \int d\Theta E(\Theta) p_\beta(\Theta \mid \mathcal{D}, K)
\end{aligned}$$

The integral with respect to Θ is the expectation of $E(\Theta)$ according to a probability distribution $p_\beta(\Theta \mid \mathcal{D}, K)$, which can be calculated using the sampling results of the MCMC method. We can perform Bayesian estimation for the number K of mixture by computing the free energy.

C.2 Bayesian Free Energy for each K

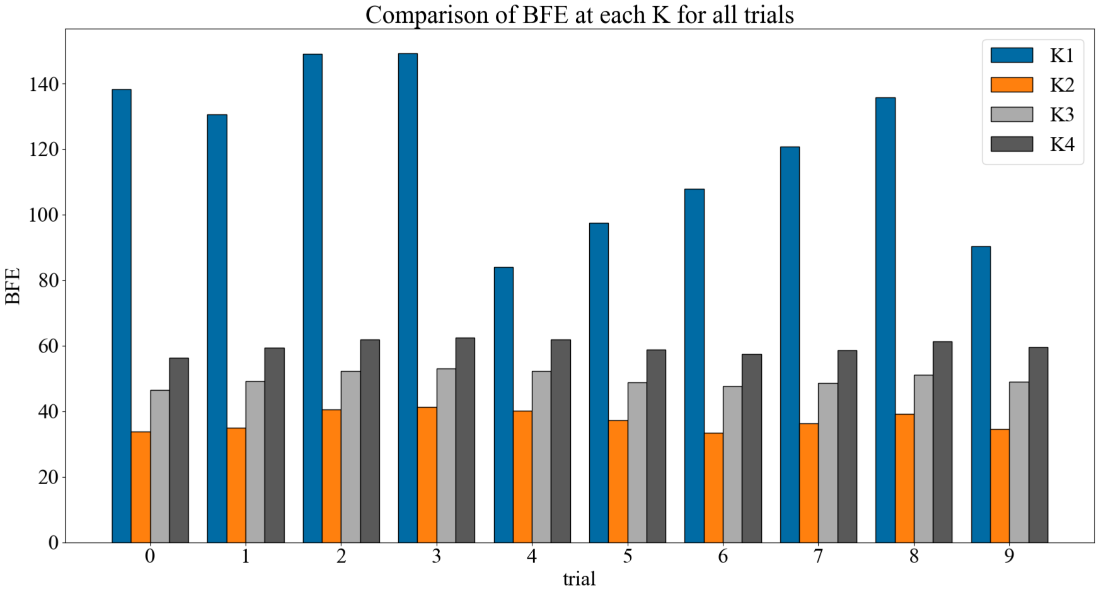


Figure C.1: Comparison of BFE at each K for all trials. The horizontal axis shows each trial of the CV and the vertical axis is the BFE. The BFE of K=2 is the smallest for all trials.

C.3 Indicator Ranking

The following figures show the top 1000 indicators of the posterior distribution of the indicators in the prediction of diffusion coefficients by SpLMM. When sampling from the posterior distribution, the number of unique indicators can be below 1000. In such a case, all indicators are displayed. The black cells represent used descriptors and the white cells represent unused descriptors.

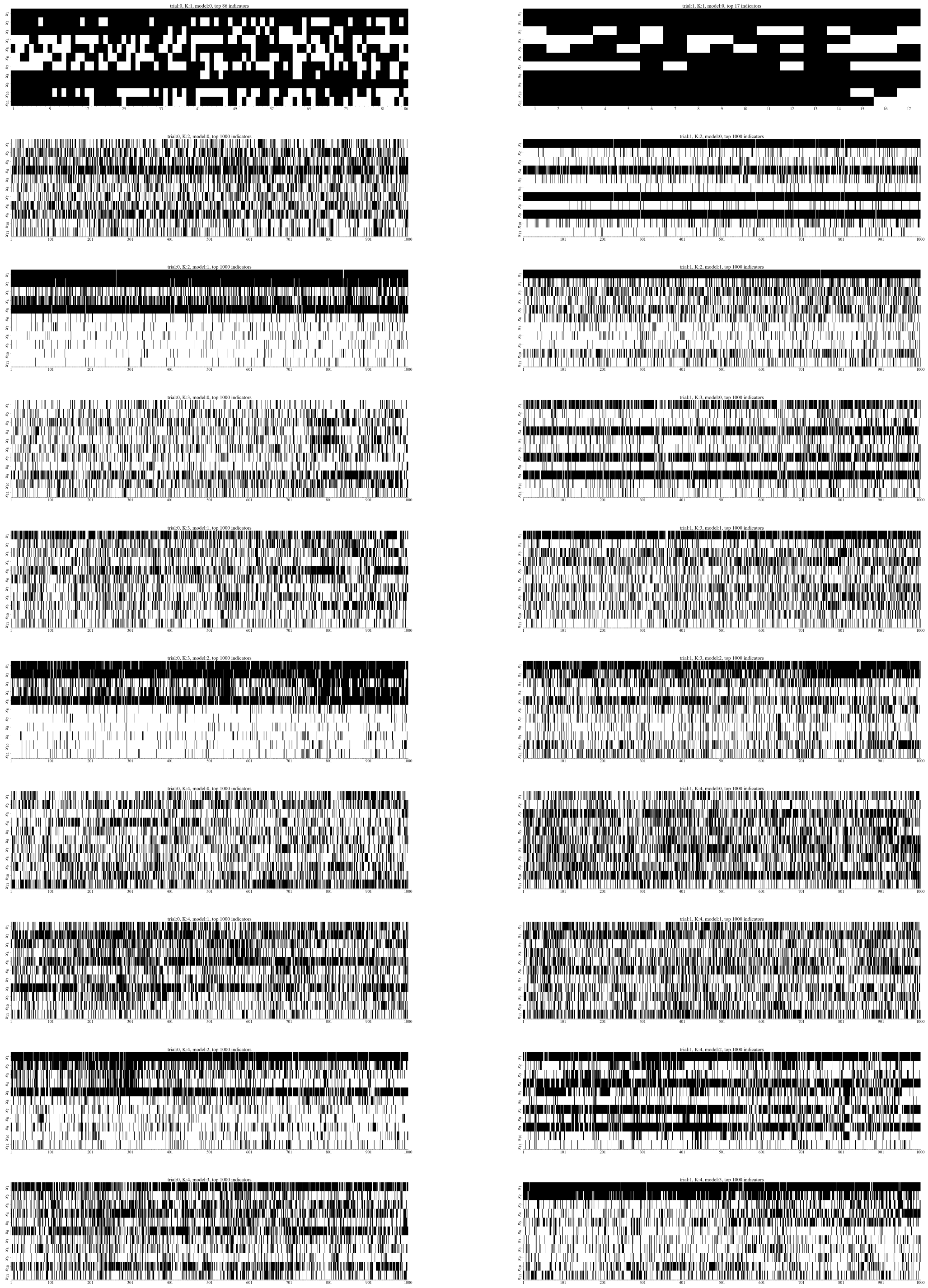


Figure C.2: Indicator Ranking (Trial=0,1)

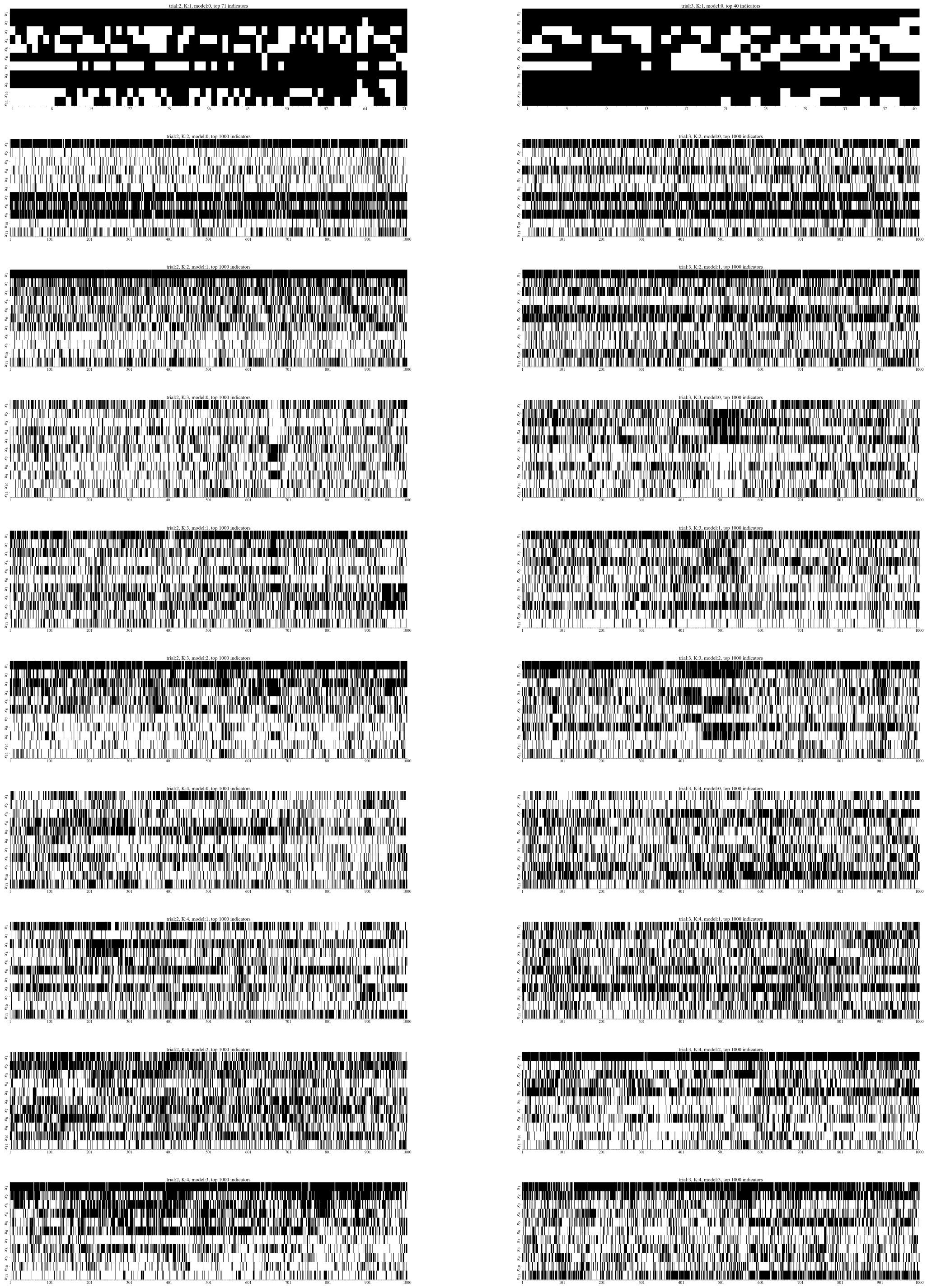


Figure C.3: Indicator Ranking (Trial=2,3)

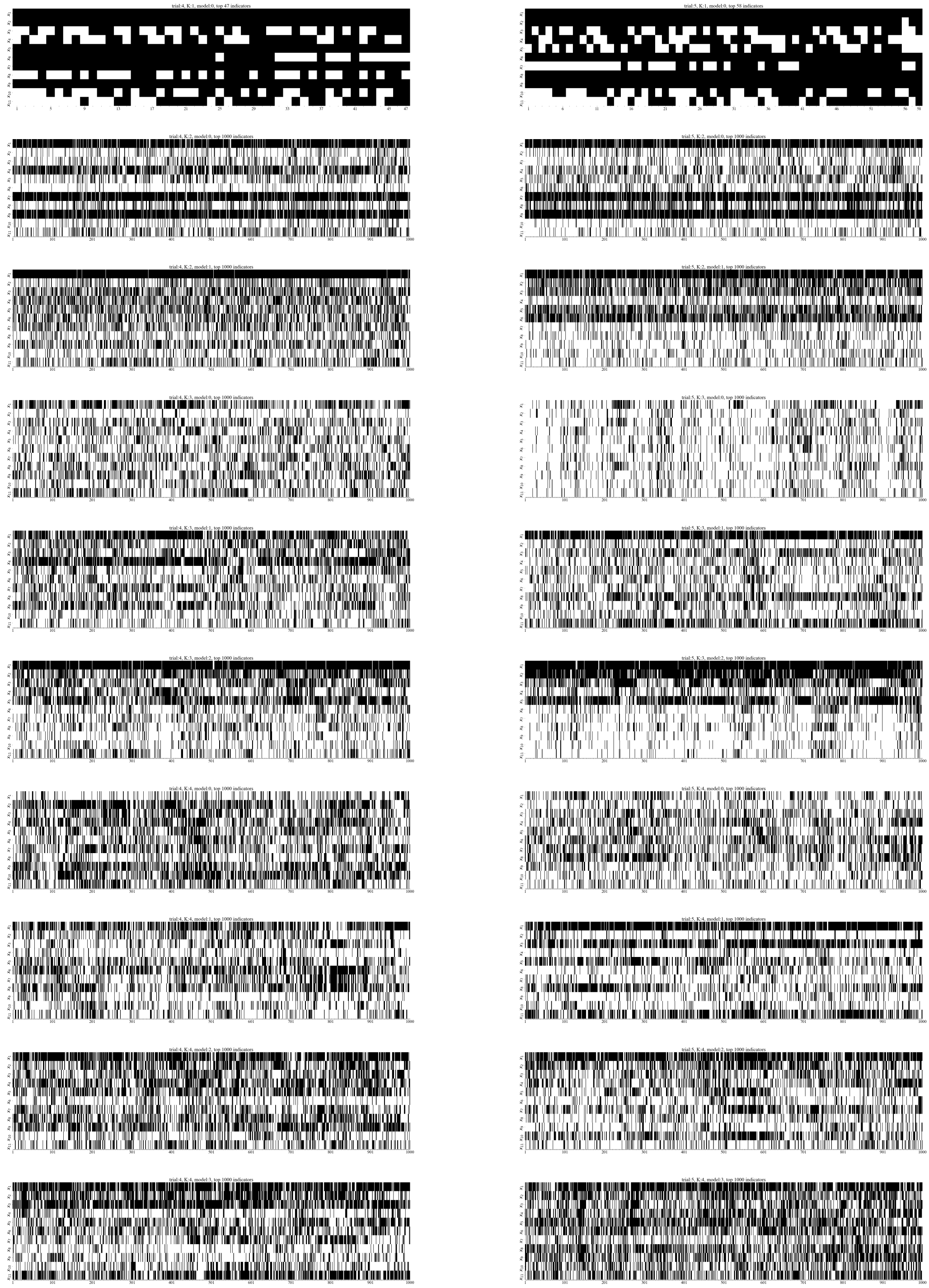
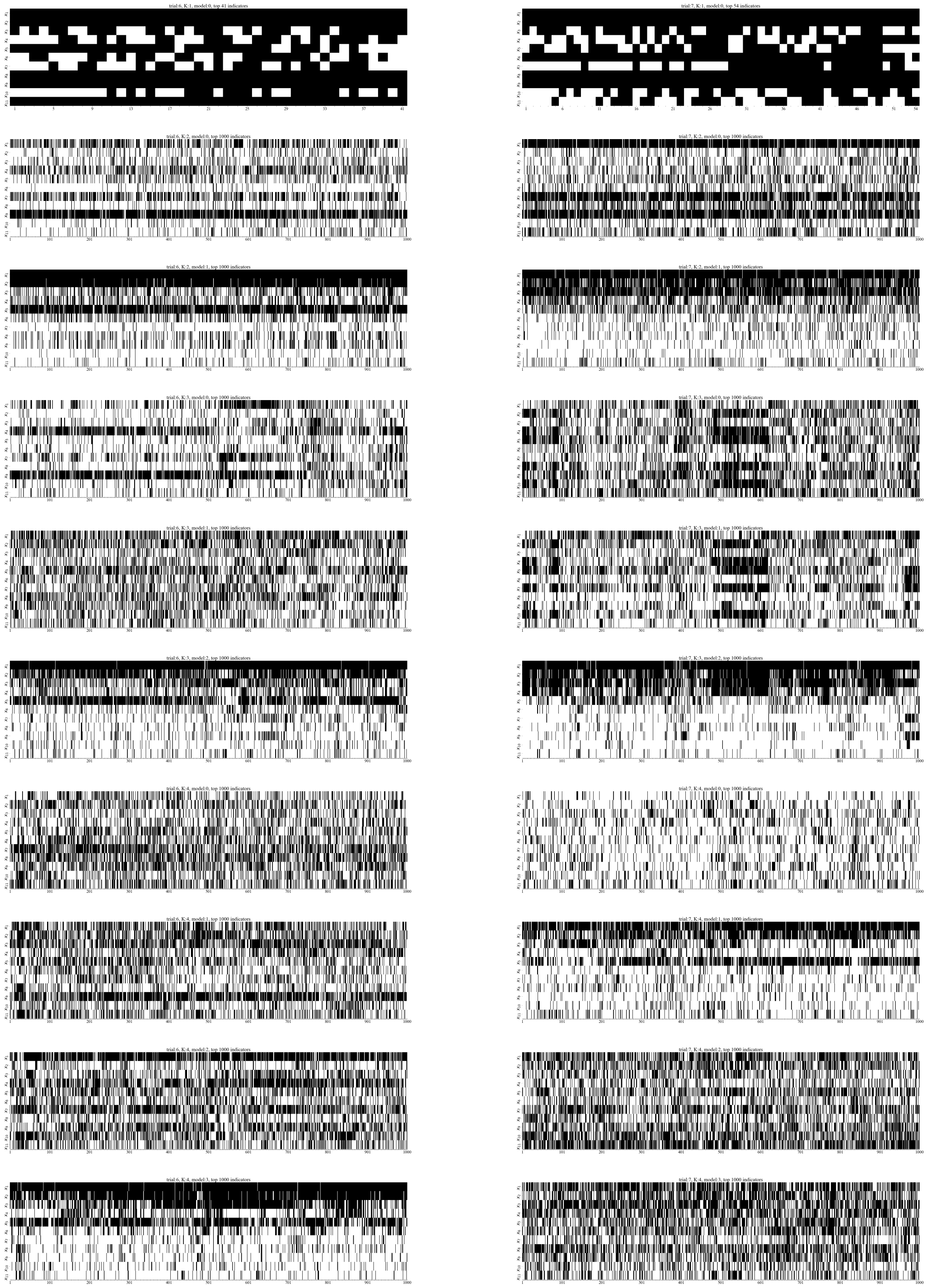


Figure C.4: Indicator Ranking (Trial=4,5)



trial:6, K:1, model:0, top 41 indicators

trial:7, K:1, model:0, top 54 indicators

trial:6, K:2, model:0, top 1000 indicators

trial:7, K:2, model:0, top 1000 indicators

trial:6, K:2, model:1, top 1000 indicators

trial:7, K:2, model:1, top 1000 indicators

trial:6, K:3, model:0, top 1000 indicators

trial:7, K:3, model:0, top 1000 indicators

trial:6, K:3, model:1, top 1000 indicators

trial:7, K:3, model:1, top 1000 indicators

trial:6, K:3, model:2, top 1000 indicators

trial:7, K:3, model:2, top 1000 indicators

trial:6, K:4, model:0, top 1000 indicators

trial:7, K:4, model:0, top 1000 indicators

trial:6, K:4, model:1, top 1000 indicators

trial:7, K:4, model:1, top 1000 indicators

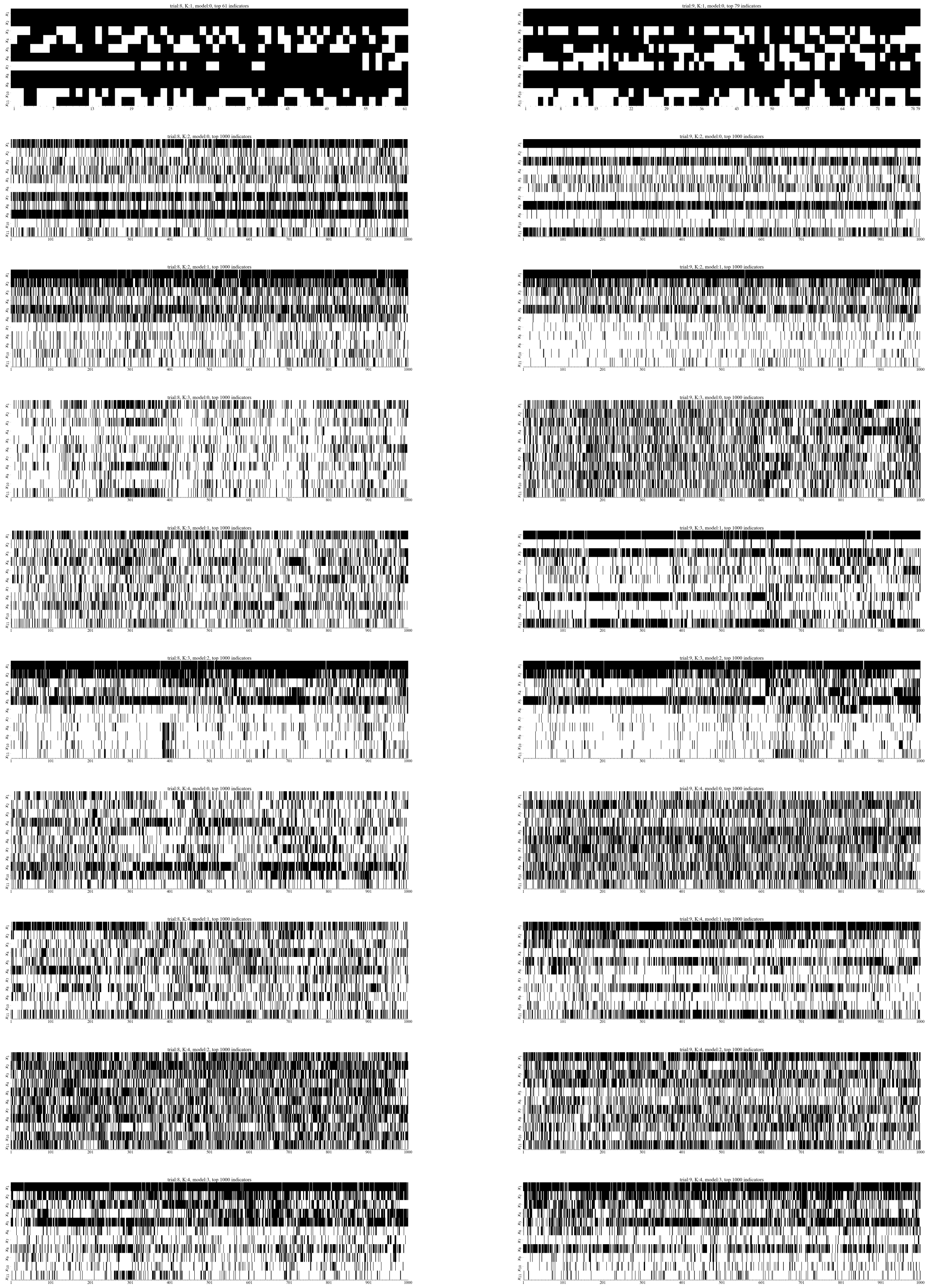
trial:6, K:4, model:2, top 1000 indicators

trial:7, K:4, model:2, top 1000 indicators

trial:6, K:4, model:3, top 1000 indicators

trial:7, K:4, model:3, top 1000 indicators

Figure C.5: Indicator Ranking (Trial=6,7)



trial:8, K:1, model:0, top 61 indicators

trial:9, K:1, model:0, top 79 indicators

trial:8, K:2, model:0, top 1000 indicators

trial:9, K:2, model:0, top 1000 indicators

trial:8, K:2, model:1, top 1000 indicators

trial:9, K:2, model:1, top 1000 indicators

trial:8, K:3, model:0, top 1000 indicators

trial:9, K:3, model:0, top 1000 indicators

trial:8, K:3, model:1, top 1000 indicators

trial:9, K:3, model:1, top 1000 indicators

trial:8, K:3, model:2, top 1000 indicators

trial:9, K:3, model:2, top 1000 indicators

trial:8, K:4, model:0, top 1000 indicators

trial:9, K:4, model:0, top 1000 indicators

trial:8, K:4, model:1, top 1000 indicators

trial:9, K:4, model:1, top 1000 indicators

trial:8, K:4, model:2, top 1000 indicators

trial:9, K:4, model:2, top 1000 indicators

trial:8, K:4, model:3, top 1000 indicators

trial:9, K:4, model:3, top 1000 indicators

Figure C.6: Indicator Ranking (Trial=8,9)

C.4 Predicted distribution of each solvent molecule.

The following figure shows the predicted distributions of the diffusion coefficients of each solvent molecule in ES-LiR, ES-GP and SpLMM. The results for each solvent molecule in each Trial are side-by-side, and the predicted distributions for $K = 1, 2, 3, 4$ are side-by-side vertically. The horizontal axis represents the diffusion coefficient and the vertical axis represents the probability density. The dotted line is the true value of the diffusion coefficient.

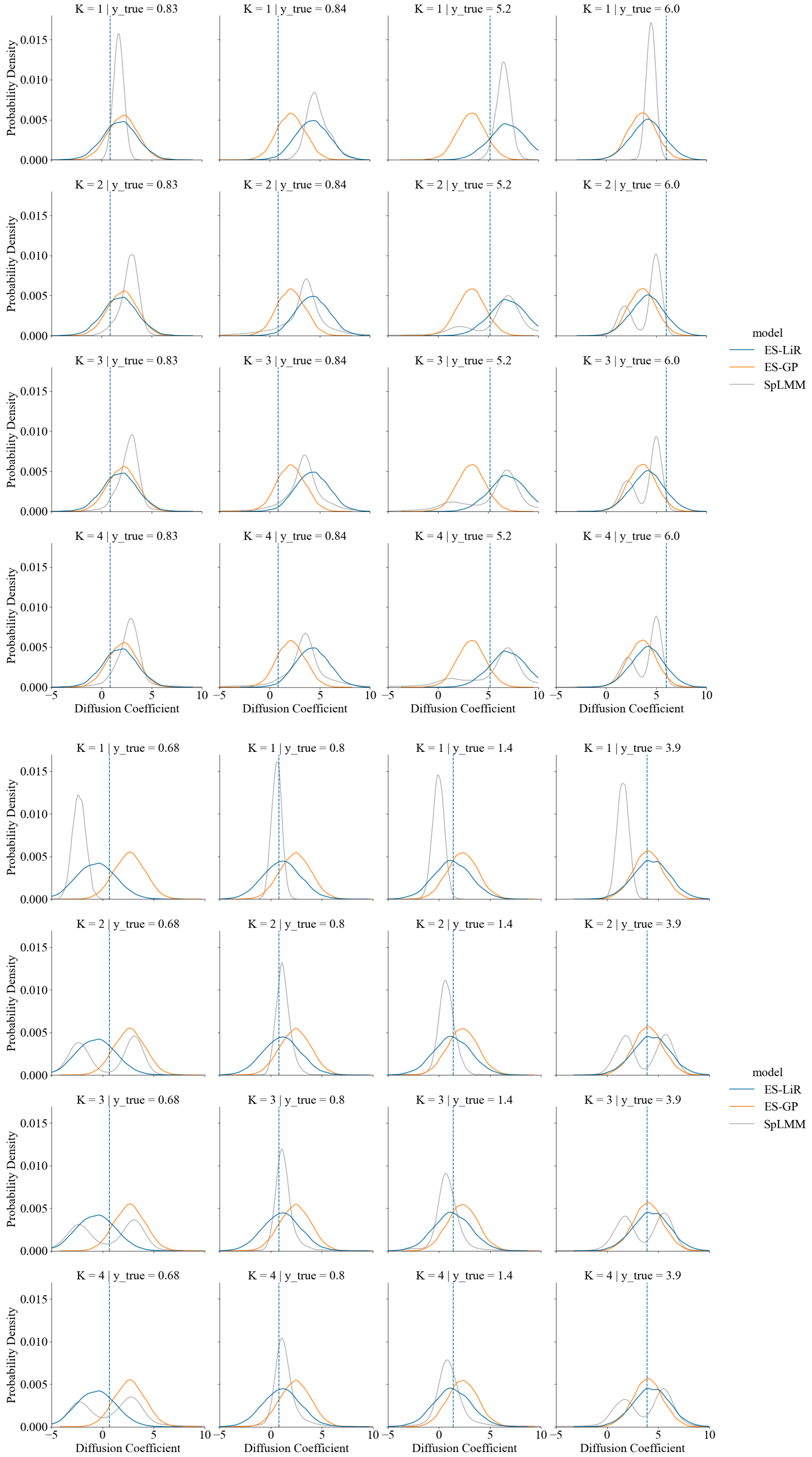


Figure C.7: Predicted Distribution of Diffusion Coefficients of Solvent Molecules (Trial=0, 1)

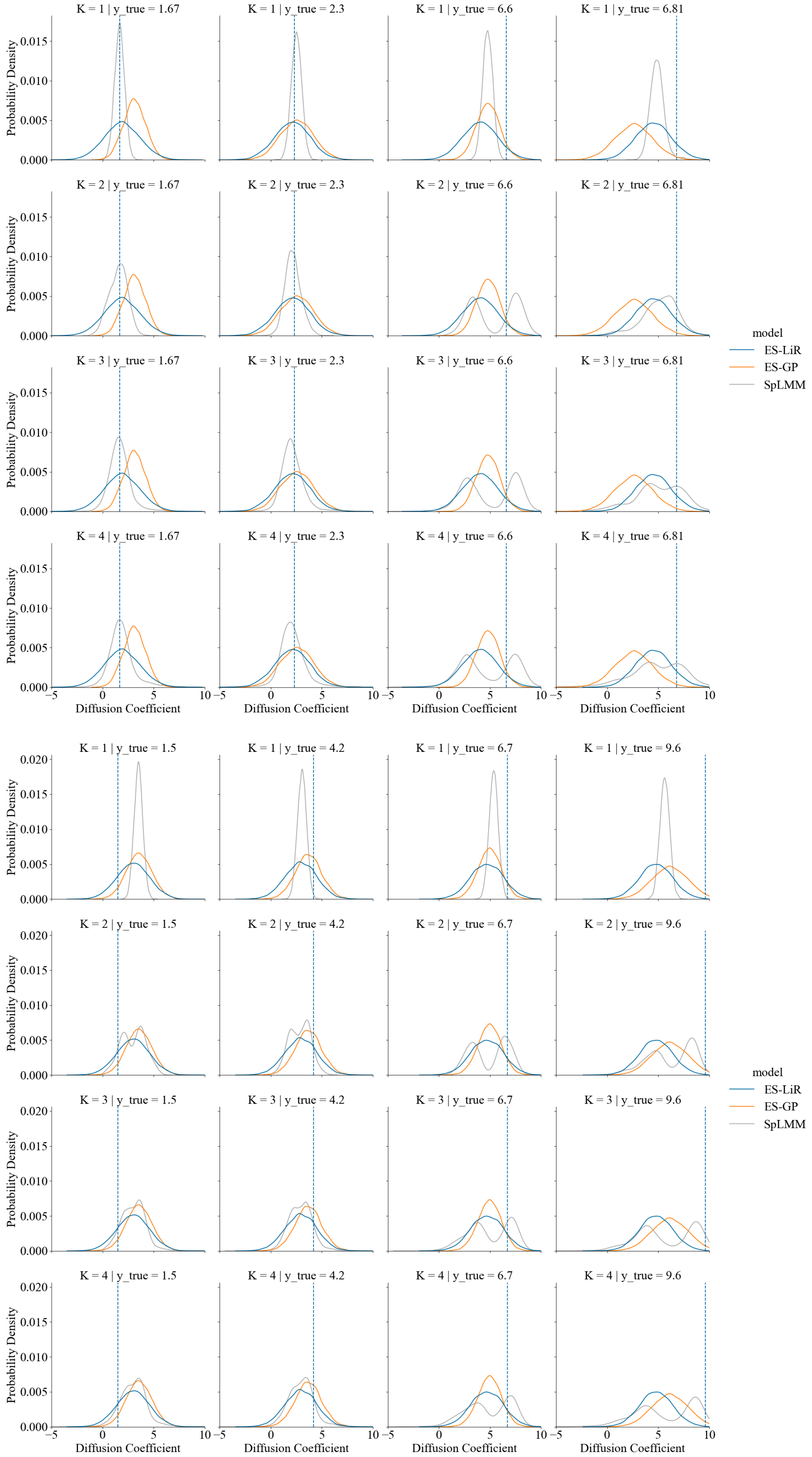


Figure C.8: Predicted Distribution of Diffusion Coefficients of Solvent Molecules (Trial=2, 3)

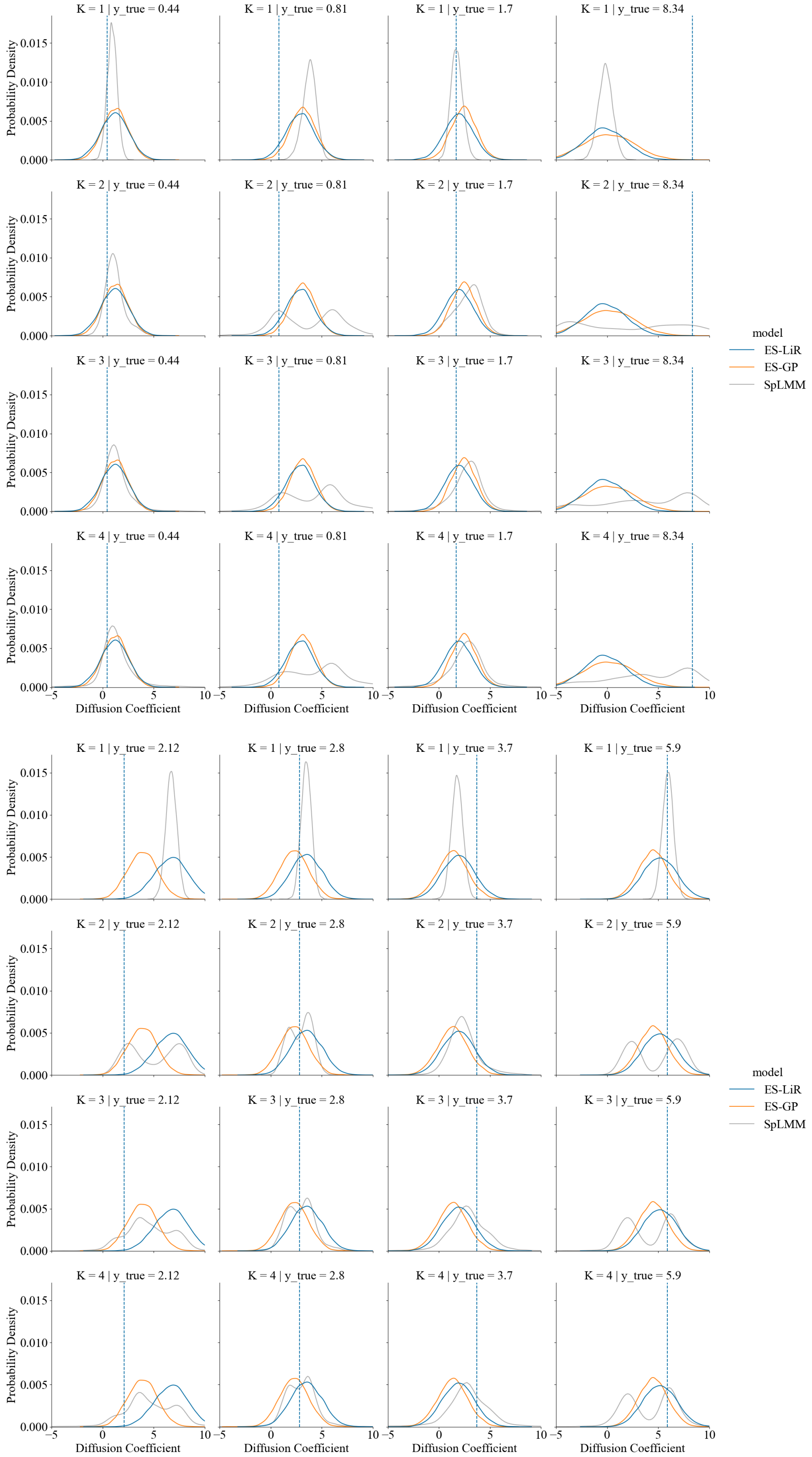


Figure C.9: Predicted Distribution of Diffusion Coefficients of Solvent Molecules (Trial=4, 5)

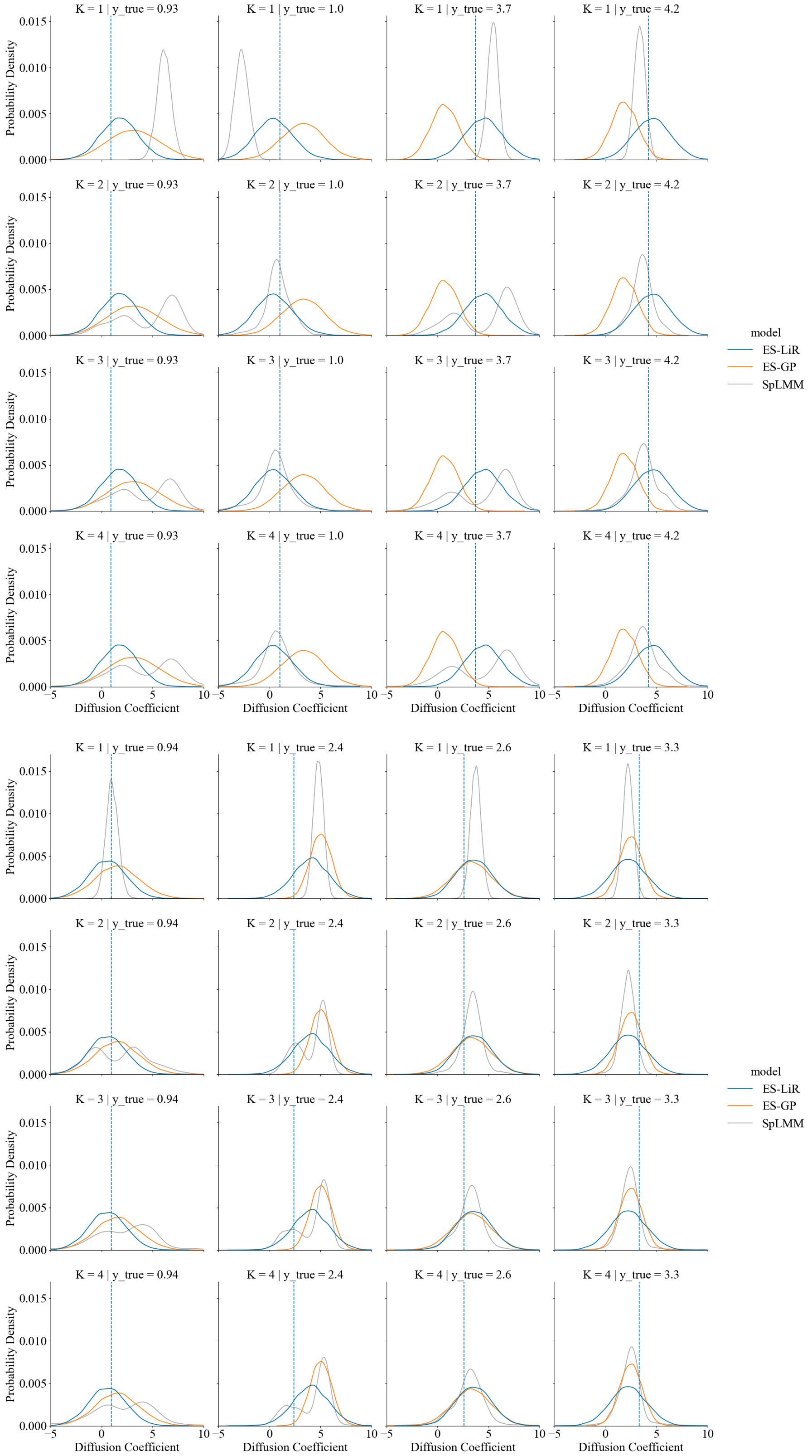


Figure C.10: Predicted Distribution of Diffusion Coefficients of Solvent Molecules (Trial=6, 7)

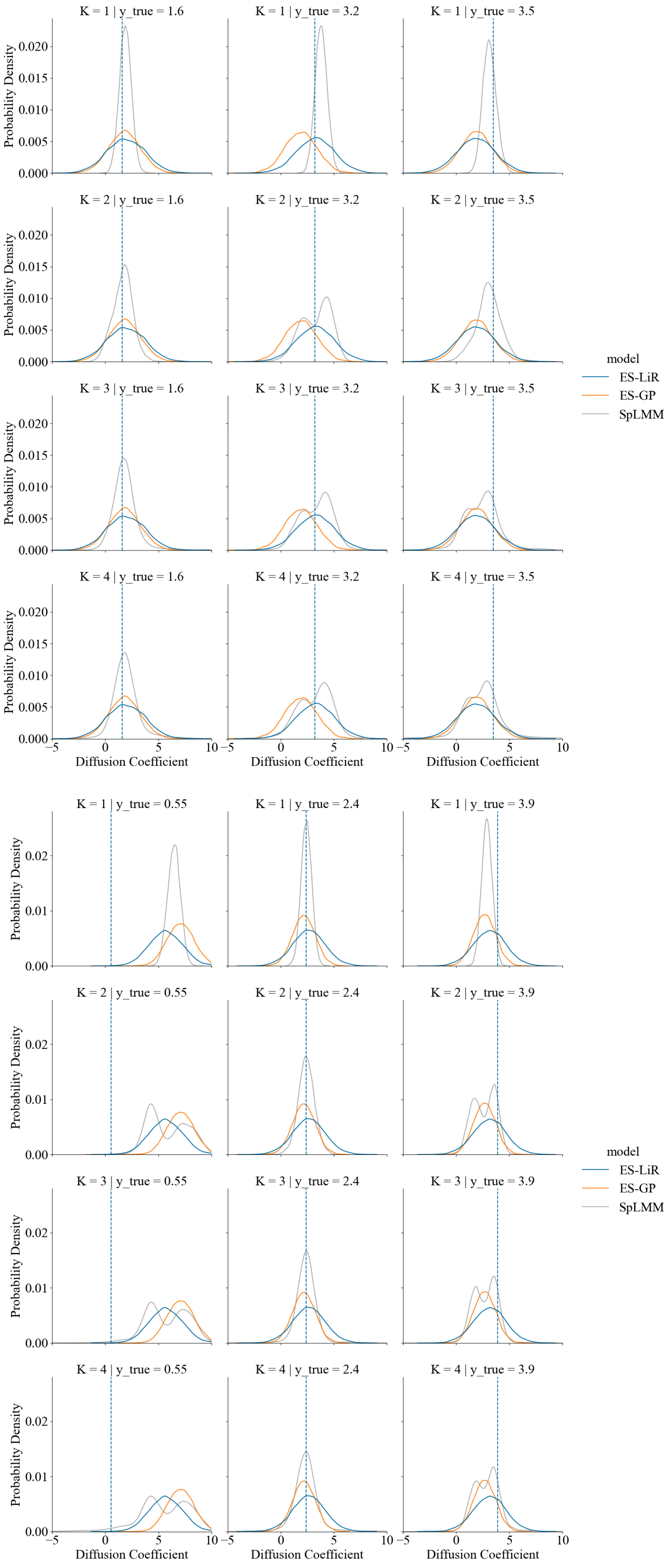
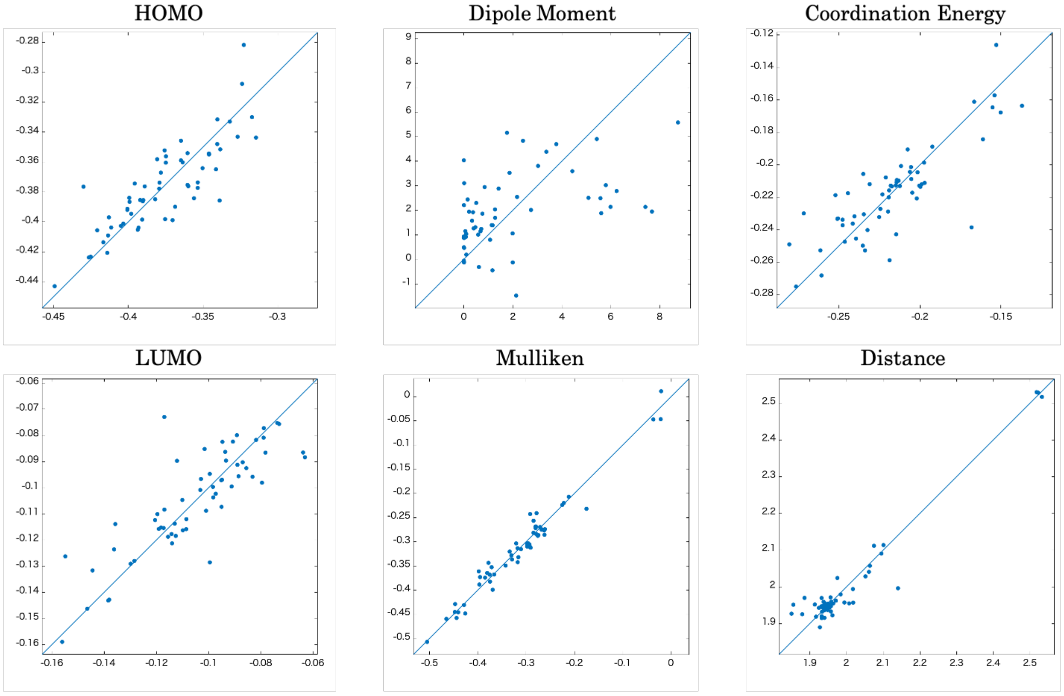


Figure C.11: Predicted Distribution of Diffusion Coefficients of Solvent Molecules (Trial=8, 9)

Appendix D

Database of multiple coordinated molecules and prediction

In calculating the calculated values in the database created in Chapter 2, we assumed that one solvent molecule is coordinated to Li-ion. However, in reality, several (typically four) solvent molecules are coordinated to Li-ion in the electrolyte. Therefore, we performed DFT calculations in a multiple-molecule coordination situation and created a database. While this database is useful, this calculation is more expensive than the DFT calculation with a single coordinated molecule and takes several days to several weeks. If the results of a DFT calculation with multiple solvents can be predicted from the results of a DFT calculation with one molecule coordinating, we do not need to calculate for every molecule. Therefore, it is worthwhile to build a predictive model for it. Figure D.1 shows the predicted values from the DFT calculations with one molecule coordinating to the DFT calculations with multiple molecules coordinating using ES-LiR. This result shows that the prediction accuracy is high for all calculated values except dipole moment.



HOMO

Dipole Moment

Coordination Energy

LUMO

Mulliken

Distance

Figure D.1: Prediction of the computed value from the DFT calculation for a single molecule to the computed value from the DFT calculation for multiple coordinated molecules.

Appendix E

Experimental Value Prediction

E.1 Experimental value prediction from calculated values

We tried to predict the experimental values from the calculated values using the database of alkali metal electrolytes built in Chapter 2. In this study, the acquisition cost of the experimental values was not a problem because the experimental data were already available, but in practice, it is often difficult to obtain the experimental data because of the human cost. Figure E.1 shows the prediction of the experimental values from the calculated values only using the ES-LiR introduced in Chapter 3. In this study, the boiling point, density, flashing point and melting point were predicted. For the solvents other than the melting point, it was found that the prediction of solvents other than the melting point is highly accurate even from the calculated values.

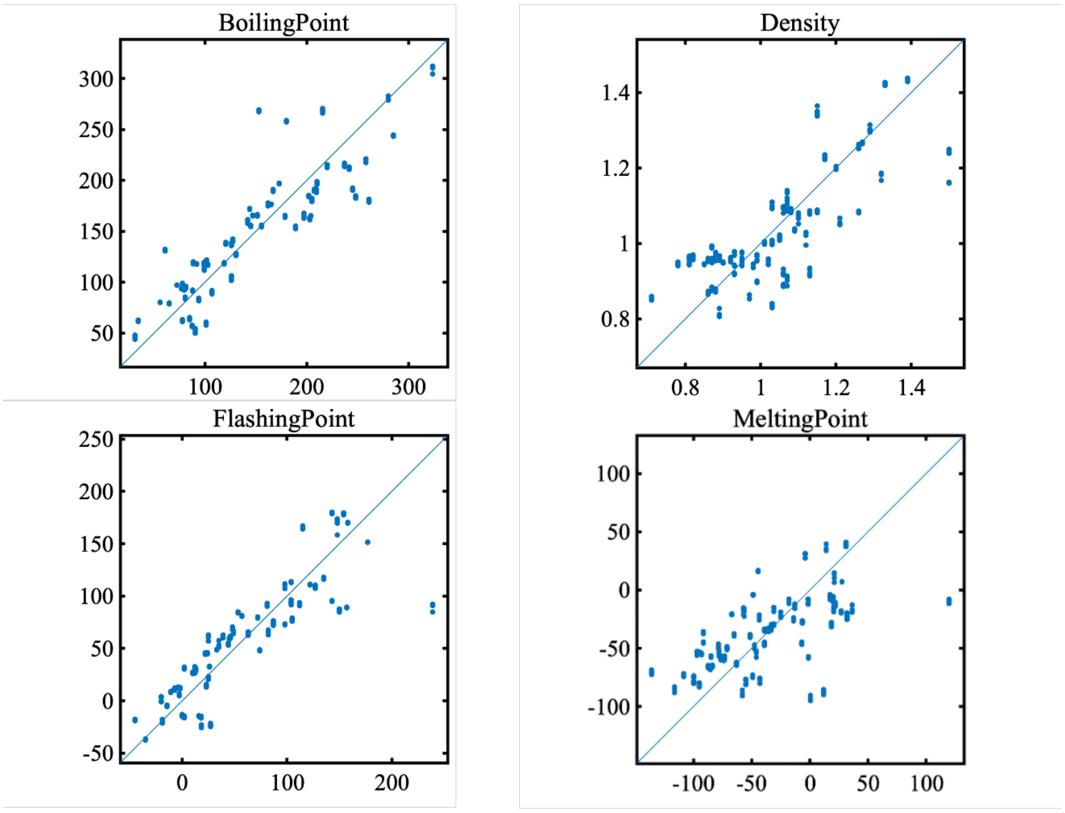


Figure E.1: Predictions of experimental values from ES-LiR. The horizontal axis represents the true value and the vertical axis represents the predicted value.

Appendix F

Application of Bayesian optimization to coordination energy

In Chapter 3, GP was used for function estimation (regression). One of the features of GP is that it can give not only predictions but also predictive variance. Bayesian Optimization (BO) is an optimization method that uses both predictions and predictive variance. In BO, we perform GP estimation from the current results and determine the next point to investigate based on the value and variance. In this study, we used COMmon Bayesian Optimization Library (COMBO) [58], which is a Bayesian optimization Python library that has attracted much attention in the field of materials science calculations, to verify the effectiveness of Bayesian optimization in electrolyte search using the database created in Chapter 2. We also verify sparsification of variables in Bayesian optimization.

F.1 BO for coordination energy

We measured the number of searches to find the sample with the highest coordination energy in the database. BO requires a few randomly obtained data at the beginning, and the results vary depending on which samples are selected first. Therefore, we ran 100 trials with different randomly selected samples and calculated the average. Figure 1 shows the mean and variance for a completely random search and a Bayesian optimization case. In the case of BO, the maximum value was found with less than half the number of searches than random searches.

F.2 Validation of Sparsified BO using synthesis data

When searching for the maximum value of the y objective variable by Bayesian optimization, descriptors that are uncorrelated with the objective variable y is expected to be noisy and the number of searches for the maximum value will increase. We tested the effectiveness of sparsification in BO by examining how much the number of searches for the maximum value is reduced by sparsification. For verification, we generated the following synthetic data. We first generated 100 random sample data in 10 dimensions and created a database of 100 samples of 10 descriptors. We then used two specific descriptors from among the 10 descriptors to generate the objective variable y by a linear sum of

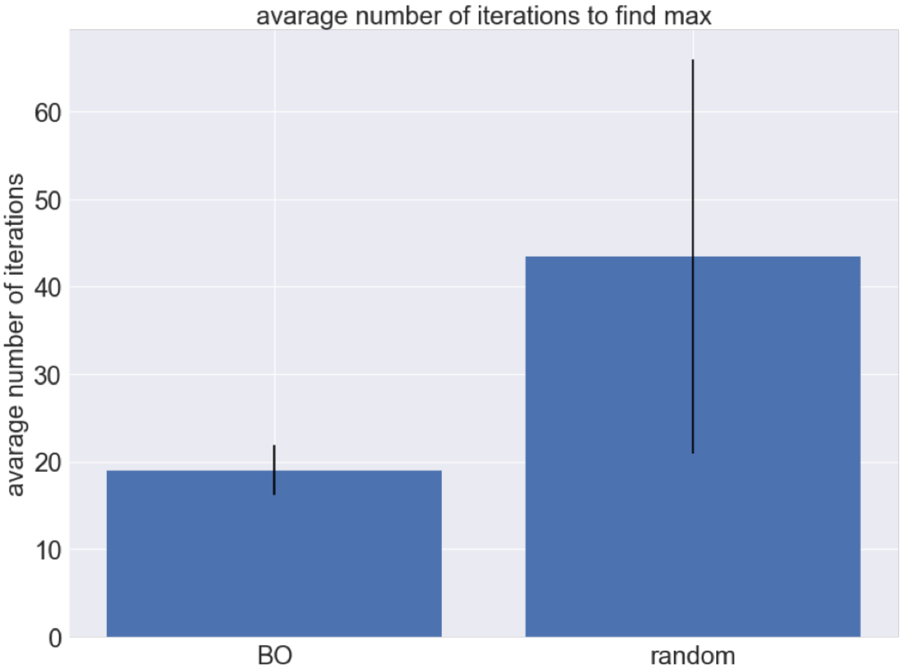
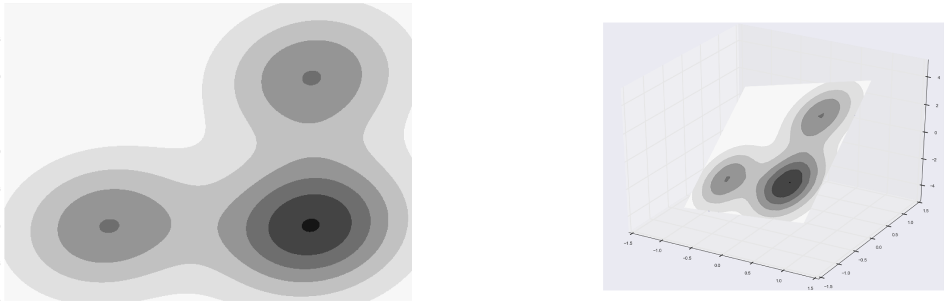


Figure F.1: Average search times to find the maximum value.

three two-dimensional Gaussian distributions, as shown in Figure F.2a. In addition, to examine the case where some descriptors have a linear dependency, as shown in Figure F.2b, we also tested the situation where Figure F.2a is embedded in 3D space and one of the ten descriptors is represented as a linear sum of the two descriptors used to generate the function.



(a) Linear summation of the 2-D Gaussian distribution

(b) Figure (a) embedded in 3D space

Figure F.2: Synthesis data

As in Section F.1, we measured the number of searches to find the maximum sample in the synthesis data. Figure F.3 is a validation of the data shown in Figure F.2a. This is a comparison between BO in 2 dimensions with no noise and BO in 10 dimensions with 8 noises. In the case of sparsification, we confirmed that the optimization can be performed with less than half the number of searches on average.

Figure F.4 is a validation of the data shown in Figure F.2b. "2d" denotes the case with only two descriptors without noise, "3d" denotes the case with a third descriptor that

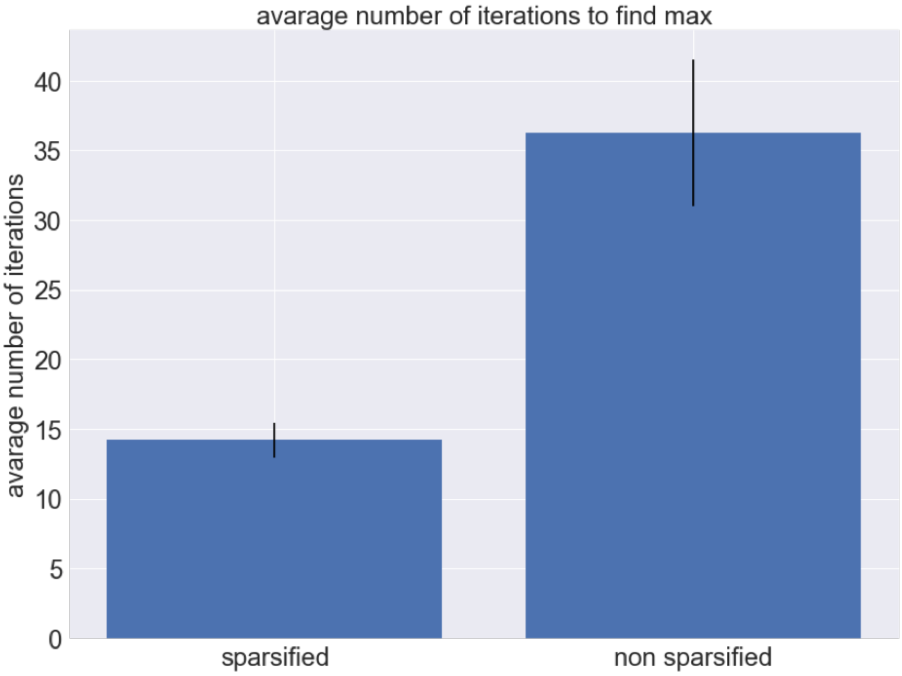


Figure F.3: Average number of searches to find the maximum value. "sparsified" indicates the case of BO with noisy descriptors pruned and sparsified, and "non sparsified" indicates the case where all the descriptors are used. In the case of sparsification, the number of searches was less than half the number of times compared to the case where all descriptors were used.

is linearly dependent, and the rest with noise. The addition of a third linearly dependent descriptor did not differ significantly from the case with only two noise-free descriptors, but we confirmed that the number of search times increased as the number of noise descriptors increased.

F.3 Sparsified BO

In Section F.2, we verified that sparsification works well in BO using synthetic data. In this section, we discuss the results of applying it to the optimization of the coordination energy. In real data, we do not know which descriptors are noise and which are effective, so we need a framework to decide which descriptors are effective. In this study, we used a method called the ES-K method to extract effective descriptors. The ES-K method first determines the number of dimensions K to choose, then trains only the indicators that have only K descriptors, compares the results, and extracts the descriptors. Since the results of Chapter 3 already show results for all indicators, the indicator with the highest prediction accuracy for each K can be selected. In Chapter 3, we used ES-LiR and ES-GP, but since BO uses GP, we used the descriptors extraction results by ES-GP to prune the noise for each K.

Figure F.5 shows the results of BO for each K. The horizontal axis shows the number of iterations and the vertical axis shows the maximum coordination energy found up to that point. The one with $K = 10$ is the one that has not been sparsified. This result

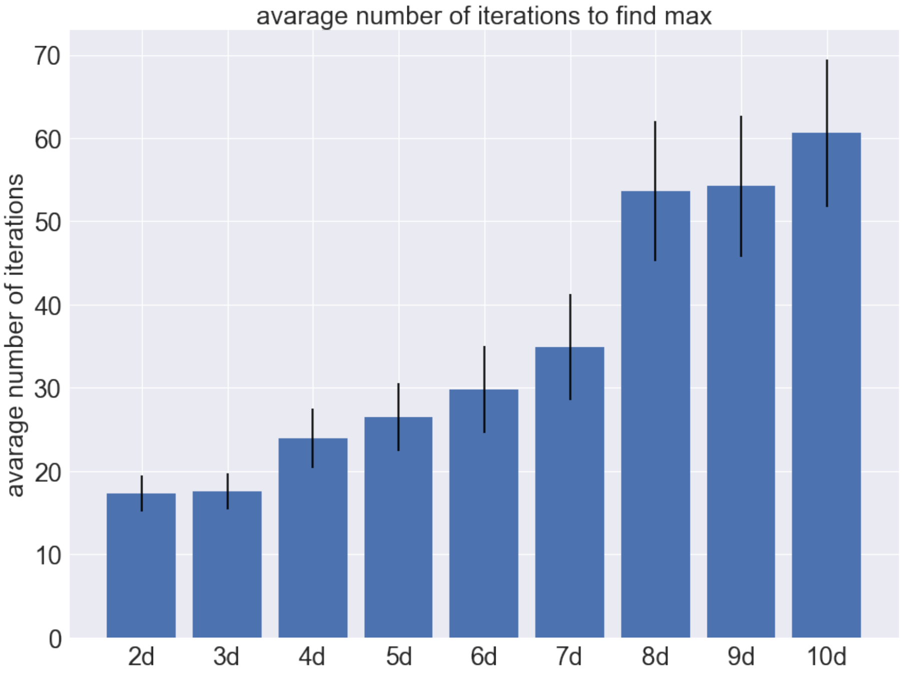


Figure F.4: Average number of searches to find the maximum value in the synthetic data. We see that increasing the number of correlated descriptors does not significantly change the number of searches.

shows that sparsification does not make a significant difference. Figure 1 shows the number of times it took to find the maximum value of the coordination energy. These results also showed that sparsification did not make a significant difference. Considering the results of Section F.2, this may be because there is a strong correlation between the descriptors, or because there are few unrelated descriptors.

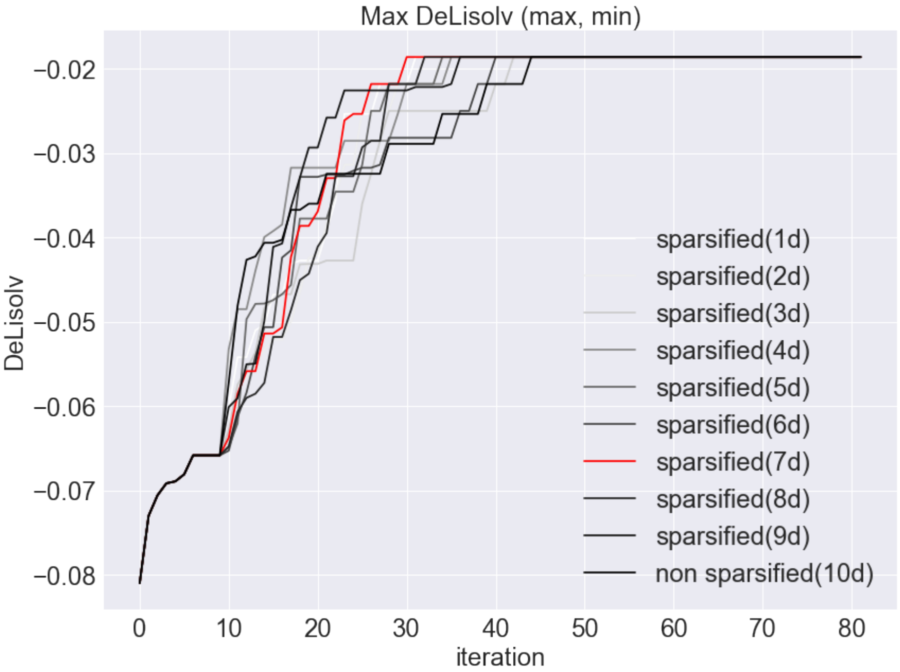


Figure F.5: Results of sparsified BO for each dimension K. The horizontal axis shows the number of iterations and the vertical axis shows the maximum coordination energy found up to that point. The red color is the result of BO for $K = 7$, which is the number of dimensions with the lowest prediction error in the ES-GP.

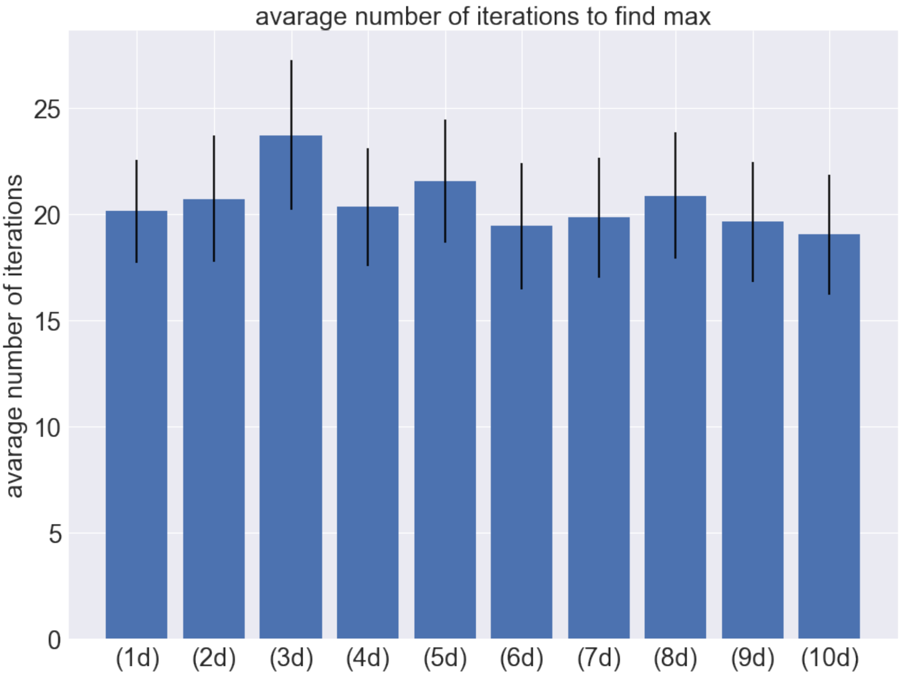


Figure F.6: Number of searches to find the maximum value in sparsified BO for every K of dimension.

Bibliography

[1] H-J Peng, S Urbonaite, Claire Villevieille, H Wolf, K Leitner, and P Novák. Consequences of electrolyte degradation for the electrochemical performance of li1+ x (niacobmn1-ab) 1-xo2. *Journal of The Electrochemical Society*, 162(13):A7072, 2015.

[2] Naoaki Yabuuchi, Mitsue Takeuchi, Masanobu Nakayama, Hiromasa Shiiba, Masahiro Ogawa, Keisuke Nakayama, Toshiaki Ohta, Daisuke Endo, Tetsuya Ozaki, Tokuo Inamasu, et al. High-capacity electrode materials for rechargeable lithium batteries: Li3nbo4-based system with cation-disordered rocksalt structure. *Proceedings of the National Academy of Sciences*, 112(25):7650–7655, 2015.

[3] Fei Luo, Bonan Liu, Jieyun Zheng, Geng Chu, Kaifu Zhong, Hong Li, Xuejie Huang, and Liquan Chen. Nano-silicon/carbon composite anode materials towards practical application for next generation li-ion batteries. *Journal of The Electrochemical Society*, 162(14):A2509, 2015.

[4] Yuki Yamada, Keizo Furukawa, Keitaro Sodeyama, Keisuke Kikuchi, Makoto Yaegashi, Yoshitaka Tateyama, and Atsuo Yamada. Unusual stability of acetonitrile-based superconcentrated electrolytes for fast-charging lithium-ion batteries. *Journal of the American Chemical Society*, 136(13):5039–5046, 2014.

[5] Keitaro Sodeyama, Yuki Yamada, Koharu Aikawa, Atsuo Yamada, and Yoshitaka Tateyama. Sacrificial anion reduction mechanism for electrochemical stability improvement in highly concentrated li-salt electrolyte. *The Journal of Physical Chemistry C*, 118(26):14091–14097, 2014.

[6] Jun Haruyama, Keitaro Sodeyama, Liyuan Han, Kazunori Takada, and Yoshitaka Tateyama. Space–charge layer effect at interface between oxide cathode and sulfide electrolyte in all-solid-state lithium-ion battery. *Chemistry of Materials*, 26(14):4248–4255, 2014.

[7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *Apl Materials*, 1(1):011002, 2013.

[8] Motoaki Nishijima, Takuya Ootani, Yuichi Kamimura, Toshitsugu Sueki, Shogo Esaki, Shunsuke Murai, Koji Fujita, Katsuhisa Tanaka, Koji Ohira, Yukinori Koyama, et al. Accelerated discovery of cathode materials with prolonged cycle life for lithium-ion battery. *Nature communications*, 5(1):1–7, 2014.

[9] Randy Jalem, Takahiro Aoyama, Masanobu Nakayama, and Masayuki Nogami. Multivariate method-assisted ab initio study of olivine-type limxo4 (main group m2+–x5+

and m3+–x4+) compositions as potential solid electrolytes. *Chemistry of Materials*, 24(7):1357–1364, 2012.

[10] Randy Jalem, Mayumi Kimura, Masanobu Nakayama, and Toshihiro Kasuga. Informatics-aided density functional theory study on the li ion transport of tavorite-type limto4f (m3+–t5+, m2+–t6+). *Journal of Chemical Information and Modeling*, 55(6):1158–1168, 2015.

[11] Martin Korth. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. *Physical Chemistry Chemical Physics*, 16(17):7919–7926, 2014.

[12] Tamara Husch, Nusret Duygu Yilmazer, Andrea Balducci, and Martin Korth. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: computing infrastructure and collective properties. *Physical Chemistry Chemical Physics*, 17(5):3394–3401, 2015.

[13] Nav Nidhi Rajput, Xiaohui Qu, Niya Sa, Anthony K Burrell, and Kristin A Persson. The coupling between stability and ion pair formation in magnesium electrolytes from first-principles quantum mechanics and classical molecular dynamics. *Journal of the American Chemical Society*, 137(9):3411–3420, 2015.

[14] Materials genome initiative, https://www.mgi.gov/ (accessed 2020-11-03).

[15] Integrated computational materials engineering expert group, `https://cordis.europa.eu/project/id/606711` (accessed 2020-11-03).

[16] The european materials modelling council, https://emmc.info/ (accessed 2020-11-03).

[17] Non-destructive evaluation (nde) system for the inspection of operation-induced material degradation in nuclear power plants | nomad, https://www.nomad-horizon2020.eu/ (accessed 2020-11-03).

[18] Nccr marvel: nccr-marvel.ch, https://nccr-marvel.ch/ (accessed 2020-11-03).

[19] Materials genome institute of shanghai university, http://en.mgi.shu.edu.cn/ (accessed 2020-11-03).

[20] John B Goodenough and Youngsik Kim. Challenges for rechargeable li batteries. *Chemistry of materials*, 22(3):587–603, 2010.

[21] Kang Xu. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chemical reviews*, 104(10):4303–4418, 2004.

[22] Ltd. KISHIDA CHEMICAL Co. Lbg-溶媒. `http://www.kishida.co.jp/product/battery/lbg/lbg02.html` (accessed 2020-11-03).

[23] Keitaro Sodeyama, Yasuhiko Igarashi, Tomofumi Nakayama, Yoshitaka Tateyama, and Masato Okada. Liquid electrolyte informatics using an exhaustive search with linear regression. *Physical Chemistry Chemical Physics*, 20:22585–22591, 2018.

[24] Atsushi Ishikawa, Keitaro Sodeyama, Yasuhiko Igarashi, Tomofumi Nakayama, Yoshitaka Tateyama, and Masato Okada. Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents. *Physical Chemistry Chemical Physics*, 21(48):26399–26405, 2019.

[25] Tomofumi Nakayama, Yasuhiko Igarashi, Keitaro Sodeyama, and Masato Okada. Material search for li-ion battery electrolytes through an exhaustive search with a gaussian process. *Chemical Physics Letters*, 731:136622, 2019.

[26] Konstantinos Blekas and Aristidis Likas. Sparse regression mixture modeling with the multi-kernel relevance vector machine. *Knowledge and information systems*, 39(2):241–264, 2014.

[27] John Darzentas, George Vouros, Spyros Vosinakis, and Argyris Arnellos. *Artificial Intelligence: Theories, Models and Applications: 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008, Proceedings*, volume 5138. Springer, 2008.

[28] Wei Liu, Bo Zhang, Zhiwei Zhang, Jian Tao, and Adam Branscum. Model selection in finite mixture of regression models: a bayesian approach with innovative weighted g priors and reversible jump markov chain monte carlo implementation. *Journal of Statistical Computation and Simulation*, 85:1–23, 12 2014.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian∼09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[30] Axel D Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of chemical physics*, 98:5648–5652, 1993.

[31] Thom H Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *The Journal of chemical physics*, 90(2):1007–1023, 1989.

[32] Narbe Mardirossian and Martin Head-Gordon. How accurate are the minnesota density functionals for noncovalent interactions, isomerization energies, thermochemistry, and barrier heights involving molecules composed of main-group elements? *Journal of chemical theory and computation*, 12(9):4303–4325, 2016.

[33] Yan Zhao and Donald G Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(1-3):215–241, 2008.

[34] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

[35] Alan E Reed, Robert B Weinstock, and Frank Weinhold. Natural population analysis. *The Journal of Chemical Physics*, 83(2):735–746, 1985.

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

[37] Arthur T Howe and Mark G Shilton. Studies of layered uranium (vi) compounds. iv. proton conductivity in single-crystal hydrogen uranyl phosphate tetrahydrate (hup) and in polycrystalline hydrogen uranyl arsenate tetrahydrate (huas). *Journal of Solid State Chemistry*, 34(2):149–155, 1980.

[38] Richard Car and Mark Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55(22):2471, 1985.

[39] CPMD, http://www.cpmd.org/, Copyright IBM Corp 1990-2008, Copyright MPI für Festkörperforschung Stuttgart 1997-2001.

[40] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996.

[41] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 78(7):1396, 1997.

[42] Stefan Goedecker, Michael Teter, and Jürg Hutter. Separable dual-space gaussian pseudopotentials. *Physical Review B*, 54(3):1703, 1996.

[43] Matthias Krack. Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts*, 114(1-3):145–152, 2005.

[44] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[45] Yasuhiko Igarashi, Hiroko Ichikawa, Yoshinori Nakanishi-Ohno, Hikaru Takenaka, Daiki Kawabata, Satoshi Eifuku, Ryoi Tamura, Kenji Nagata, and Masato Okada. Es-dos: Exhaustive search and density-of-states estimation as a general framework for sparse variable selection. In *J Phys Conf Ser*, volume 1036, page 012001, 2018.

[46] Yasuhiko Igarashi, Hikaru Takenaka, Yoshinori Nakanishi-Ohno, Makoto Uemura, Shiro Ikeda, and Masato Okada. Exhaustive search for sparse variable selection in linear regression. *Journal of the Physical Society of Japan*, 87(4):044802, 2018.

[47] Kota Shiba, Ryo Tamura, Gaku Imamura, and Genki Yoshikawa. Data-driven nanomechanical sensing: Specific information extraction from a complex system. *Scientific reports*, 7(1):3661, 2017.

[48] Tien Lam Pham, Hiori Kino, Kiyoyuki Terakura, Takashi Miyake, and Hieu Chi Dam. Novel mixture model for the representation of potential energy surfaces. *The Journal of Chemical Physics*, 145(15):154103, 2016.

[49] MJ Garside. The best sub-set in multiple regression analysis. *Applied Statistics*, pages 196–200, 1965.

[50] Yasuhiko Igarashi, Kenji Nagata, Tatsu Kuwatani, Toshiaki Omori, Yoshinori Nakanishi-Ohno, and Masato Okada. Three levels of data-driven science. In *Journal of Physics: Conference Series*, volume 699, page 012001. IOP Publishing, 2016.

[51] Thomas M Cover and Jan M Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE Trans. Systems, Man, and Cybernetics*, 7(9):657–661, 1977.

[52] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[53] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian process for machine learning.* MIT press, 2006.

[54] Roberto Todeschini, Viviana Consonni, Andrea Mauri, and Manuela Pavan. Detecting “ bad ” regression models: multicriteria fitness functions in regression analysis. *Analytica Chimica Acta*, 515(1):199–208, 2004.

[55] Sodeyama Keitaro and et al. in preparation.

[56] Kuo-Jung Lee, Ray-Bing Chen, and Ying Nian Wu. Bayesian variable selection for finite mixture model of linear regressions. *Computational Statistics & Data Analysis*, 95:1–16, 2016.

[57] Koji Hukushima and Koji Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[58] Tsuyoshi Ueno, Trevor David Rhone, Zhufeng Hou, Teruyasu Mizoguchi, and Koji Tsuda. Combo: an efficient bayesian optimization library for materials science. *Materials discovery*, 4:18–21, 2016.