UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding chemical reactions through the theoretical lenses: Markov State Model and Gaussian Process Regression for an identification of reaction coordinates and computation of multidimensional free energy surfaces

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Chemistry

by

Wasut Pornpatcharapong

Committee in charge:

Professor John Weare, Chair
Professor Clifford Kubiak
Professor Katja Lindenberg
Professor Francesco Paesani
Professor Ruth Williams

2018

The dissertation of Wasut Pornpatcharapong is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2018

EPIGRAPH

A man is never too old to learn.
- Chinese Proverb

Not all those who wander are lost.
- J. R. R. Tolkien

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere thanks to Professor John H. Weare, who contributed greatly to my current self. His guidance teaches me to look at problems from multiple dimensions, and gives me freedom to explore things, both academically and non-academically. All these allow me to compile bits and pieces of seemingly unrelated knowledge into a distinct package that I today cherish.

For my current standing in the academia, I would like to thank these three people: Professor George C. Schatz of Northwestern University and Dr. Nicolas Winter, for introducing me to the world of theoretical chemistry and my fabulous experience in Chicago, and Dr. Eric J. Bylaska at PNNL for introducing me to the wonderful Python programming language, as well as for tiredlessly helping me with using NWChem.

This thesis would not complete without the help from Professor Gabor Csányi of the University of Cambridge for his immense help during the early stage of my work on Gaussian Process Regression (GPR), as well as Professor Jonathan Weare of the University of Chicago for his advices on free energy computations in general, as well as the Eigenvector Method for Umbrella Sampling (EMUS) algorithm. I would like to also thank Professor Cecilia Clementi of Rice University, and Professor Frank Noé of the Free University of Berlin for a close collaboration and numerous theoretical exchanges on Markov State Model (MSM). They are also extremely helpful and answered all my questions without hesitating. It was my great pleasure working with them. I also would like to thank and acknowledge the committee members, Professor Clifford Kubiak, Professor Katja Lindenberg, Professor Francesco Paesani, and Professor Ruth Williams for their time and guidance.

To all Weare Group members that I knew: Dr. Ying Chen and Duo Song, it has been a pleasure for me to get to know and work with both of you. I wish both of you all the best for your future endeavors.

I would not become a person I am today without the supports from the following people: Sumaetee Tangwancharoen (Por), Pisrut Phummirat (Todd), and Kullachate Muangnapoh (Oath), my past roommates. I appreciate all the time we spent together, no matter happy or sad, and also the fact that all of you are the first refuge of strength in my darkest days. You guys are truly my great friends! Thank you for everything, and let us keep in touch! Kantapon Kaewtip (Jom), for many insightful conversations. Thada Udomprapasup (Kim), for being a good-natured persona that inspires me to be devoted to the causes. Teerapong Pirojsirikul (Tee), and Wiroj Nanthasetpong (Wave), for being my nice and honest friends who I can always turn to every time. Watcharapong Hongjamrassilp (WiN), for your neverending curiosity and your positive energy bursts, and Pichaya Lertvilai (Tonmai), for a place to stay to write this thesis, and all your intellectual capacity. Moreover, I would like to also extend my thanks to all my other friends who might not be named here who contributed bits and pieces at some point of time in my life. No matter how small they are, they all constitute and influence my current self. Thank you for being an integral part of my life!

Also, I would like to thank my parents for their neverending supports, especially during the toughest moments of the study. With the completion of this dissertation and the thesis defense, I vow to make you proud!

Last but not least, I would like to thank UCSD and the surrounding communities. Living in California has shaped a new, different way of thinking in me, and seeing all the new innovations in such a cradle is an honor to me. These truly inspire me to keep innovating and be different.

Chapters 2 and 3, in full, are parts of the material titled “Markov State Modeling for Ion Pairing Dynamics in Aqueous Solutions” by Pornpatcharapong, Wasut, Noé, Frank, Clementi, Cecilia, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material, and all co–authors have approved the use of the material for this dissertation.

Chapters 4 and 5, in full, are parts of the material titled “Efficient Two-dimensional Ion Pairing Free Energy Landscape Calculation with Gaussian Process Regression” by Pornpatcharapong, Wasut, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material

VITA

2010	B.A. in Chemistry	Northwestern University
2012	M.S. in Chemistry	University of California, San Diego
2018	Ph.D. in Chemistry	University of California, San Diego

ABSTRACT OF THE DISSERTATION

Understanding chemical reactions through the theoretical lenses:
Markov State Model and Gaussian Process Regression for an identification of
reaction coordinates and computation of multidimensional free energy surfaces

by

Wasut Pornpatcharapong

Doctor of Philosophy in Chemistry

University of California, San Diego, 2018

Professor John Weare, Chair

Understanding a proper reaction coordinate and the free energy profile of a chemical reaction provides valuable information in elucidating the kinetics and thermodynamics properties, as well as the underlying reaction mechanisms. Nevertheless, identifying a proper reaction coordinate for a specific chemical reaction or computing the free energy landscape are difficult. Hence, any methods that could systematically provide insights into both issues would play important roles in studies of any kind of chemical reactions.

On the reaction coordinates front, Markov State Model (MSM) is a tool that can be used to identify the reaction coordinate of the slowest motion in a simulation. Instead of the deterministic view of the MD simulations, MSM takes a probabilistic view that a configuration has a probability to evolve to another configuration after time τ defined by a transfer operator, which allows us to identify each motion in the system in terms of the information encapsulated in each of the operators corresponding eigenfunctions and eigenvalues. The eigenfunctions of the transfer operator can then be projected onto the collective variable space, and minimal representation of each eigenfunction in the collective variable space could be obtained.

On the free energy front, efficient multidimensional free energy landscape can be reconstructed smoothly from noisy free energy estimators through Gaussian Process Regression (GPR). In this dissertation, we proposed a rigorous GPR workflow that also ensures the consistency through projection of the multidimensional landscape into each individual one-dimensional surface with errors bounded by Eigenvector Method for Umbrella Sampling (EMUS).

This dissertation employed both MSM and GPR to study the dynamics of cation—anion association in aqueous solutions using LiCl and NaCl as a model. With MSM, we have completely identified all significant motions in the association process and in the bulk, and we also identify important contributions to the slowest process of the dynamics. With GPR, we have achieved a smooth reconstruction of a free energy landscape of both systems using 2 collective variables and large efficiency gain relative to traditional two-dimensional windowed simulations.

CHAPTER 1

Historical Backgrounds of Ion Pairing Studies in Aqueous Solutions

1.1 Motivations

Despite being a simple chemical system, studying ion pairing in aqueous solutions could be the key to understanding many other complex chemical reactions of current research interests, for the prescence of the ions in aqueous solution is an integral part of human body, seawater, catalytic environments, or energy-efficient materials. There are numerous evidences of the ions in aqueous solutions that play crucial roles in complex biological processes, such as the effects on macrobiomolecules like proteins or DNA, salt bridge formations in proteins, protein-DNA interactions, or assisting the creation of tertiary and quarternary structures of proteins. [1–5] Moreover, the ions in an oceanic environment also regulates the chemistry of both the ocean and the atmosphere, which could provide a better mechanistic understanding of global warming. [6–8] Recent advances of batteries and energy materials also necessitate a good understanding in ion transport in aqueous solution. [9–12]

With water molecules removed, the ion pairing interactions is easily characterized as mostly electrostatic interactions between the cation and the anion, which does not have any associated barriers due to the inverse dependence on the interionic separation, r_{+-}. However, ions in aqueous solutions behave differently, as polar water molecules could arrange around the ions by the means of charge - dipole interactions, solvating the ions. Thus, the solvation structure around the ions need to change in order for the cation and the anion to come close. The need to overcome the solvation energy of the ions to put two ions closer together; hence, involves a free energy barrier. For this reason, thermodynamics and kinetics information of ionic association in aqueous solutions can be obtained given that the free energy landscape of this chemical process can be extracted, paving the ways to isolate the most likely solvation structures around the ions and elucidate the mechanisms of the reaction in terms of the proper reaction coordinate for this system.

The biggest obstacle for thoroughly understanding not only this kind of process, but also the thermodynamics and the kinetics of all kinds of chemical reactions, is the lack of knowledge of such a reaction coordinate. Throughout the document, the reaction coordinate is defined as a linear combination of functions $\xi_i(\mathbf{x})$, where each ξ_i is called a collective variable. A collective variable is a function of all the Cartesian coordinate in the phase space defining a collective group of atoms in the regions of our interests, which includes, but not limited to r_{+-}, the angle, the dihedral angle, or even the coordination number around an atom.

$$\Psi = \sum_i c_i \xi_i(\mathbf{x}) \tag{1.1}$$

The definition of a proper reaction coordinate Ψ for a chemical reaction is such coordinate should capture the slowest motion across the division between two committor surfaces. However, in most cases, we do not know a proper combination of the collective variables, nor that we know which collective variables should we choose. Therefore, most of the attempts to study the mechanisms of chemical reactions usually based on the intuitive guesses of a few collective variables. As a result, we cannot know with certainty about the true underlying mechanism that actually represents the slowest motion, for we only have a limited picture of all the influences.

1.2 Prior Investigations to this Problem

The accepted concept of ion pairing that is widely cited today came from the work of Fuoss [13] and Winstein [14] during the 1950s. Before their work, the solvent was usually modeled as a part of the continuum model, where the solvent's electrostatic influence was treated as one single entity, which gave rise to a two-state model. One of those two states being the associated states, where the ions are held close together by their electrostatic interactions, and another state being the dissociated state, where the ions are separated far enough to not attracting each other again. However, Fuoss and Winstein showed in their work that there exists another state that lies in between the dissociated and the associated states of the ions. Therefore, the state in between shall not have the ions too separated that they do not see another's attraction, but not too close that the association is mainly the electrostatic interaction between the ions. For this reason, the state in between could be thought of as another associated state with some influence from the solvent. In this document, we will call this state the solvent separated ion pair (SSIP), and the associated state the contacted ion pair (CIP) for convenience. We also provide a summary illustration of the concept below,

$$\underset{\text{free ions}}{\mathrm{M}^{+}+\mathrm{X}^{-}} \iff \underset{\text{SSIP}}{\mathrm{M}\cdots\mathrm{X}} \iff \underset{\text{CIP}}{\mathrm{M}\cdot\mathrm{X}} \tag{1.2}$$

The earliest simulation that attempted to study this type of reaction came from the work of Belch et al. in 1986. [15] Although the work did not compute the free energy, it was among the first attempts to analyze the behavior of the solvent through computer simulations. Belch et al. proposed that for a solution of NaCl in water, the Na^{+} cation tries to maintain the octahedral structure in the CIP state, the SSIP state, and the dissociated state. The work also pointed out that as the dissociation takes place, one water molecule would rotate to form a bridge between the two ions and the hydrogen bond structure between water molecules in the first and the second solvation shells become disrupted. It is evident from this result that the solvent plays a role in the dissociation of CIP. However, how the solvent plays the role in this process still remained a mystery at that point.

The thermodynamics and the kinetics model of this type of reaction would not complete without the free energy landscape and the information of the free energy barrier between the SSIP and the CIP states. For any simulations performed under the canonical ensemble (fixed numbers of particles, volume, and temperature), the relative Helmholtz free energy can be computed by just taking a natural logarithm of the probability of finding a particular configuration of the ensemble,

$$A(\mathbf{x}) = -k_B T \ln p(\mathbf{x}) \tag{1.3}$$

where k_B is the Boltzmann's constant, and T is the temperature of the simulation. Since a collective variable $\xi(\mathbf{x})$ is also a function of the Cartesian coordinates in the phase space, the Helmholtz free energy could also be computed in terms of the collective variable through the following relationship,

$$A\left(\xi(\mathbf{x})\right) = -k_B T \ln \int \delta\left(\xi(\mathbf{x}') - \xi(\mathbf{x})\right) e^{V(\mathbf{x})/k_B T} d\mathbf{x}' \tag{1.4}$$

Initial attempts on free energy landscape computation began from a very simple model of 1 collective variable chosen from an intuitive guess. Since we hope to study the association / dissociation of the ions, the most natural collective variable that would serve the purpose would be r_{+-}. Hence, the relative Helmholtz free energy based on this collective variable would allow us to quantitatively determine the free energy barrier that separates the SSIP state from the CIP state. Numerous works have since published the one-dimensional free energy landscapes in r_{+-}, most of which allowed both qualitative and quantitative distinctions between the SSIP and CIP states as two separate metastable states. [16–19] For group 1 cations, Fennell et al. published detailed comparisons of the free energy trend down the same group. [18]. He found that the CIP structure of LiF has a very steep well of about 8.0 kcal/mol, whereas the CIP free energy barrier of CsF is very shallow, and for CsF, the CIP structure is even less thermodynamically stable than the SSIP structure.

Although the free energy of ion pairing in aqueous solutions in terms of r_{+-} can give us a rough idea of the SSIP and the CIP behaviors of the system, it does not capture the dynamics of the solvent as hinted by Belch et al. [15], which was confirmed by the later work of Geissler et al. [20]. In the work of Geissler et al., they employed transition path sampling to characterize the behavior at the transition state, and found that there are many probable transition pathways where collective motion of solvent molecules play an important role. A subsequent work by Ballard and Dellago also pointed out that using only r_{+-} as a hypothetical reaction coordinate for this process is inherently a bad choice. [21] They also found that the influence of the solvent for the NaCl reaction extends up to the third shell, consistent with what was found by Belch et al. [15]

Subsequent works have contributed a lot of efforts in determining the solvent effects of ion pairing, but a solid consensus on the solvent coordinate is yet to be made. The most popular opinion among recent publications has been imagining the number of water molecules in the first solvation shell of the cation (n_+) as a driving force for the ionic association. This viewpoint was clearly influenced from the ligand exchanging process commonly perceived by the inorganic chemist community, where they have put mechanistic labels of such process as either *associative* (involving the expansion of the solvation shell as another ligand came close to the center cation before ejecting one water molecule) or *dissociative* (involving the ejection of a water molecule from the cation before the association with another ligand), originally proposed by Eigen and Wilkins during the 1950s. [22–24] Therefore, later works involved characterizing n_+ as a solvent coordinate and later computations free energy landscapes of the ionic association in aqueous solution started to include this collective variable as well. [25–28] The most common features in these works are that the free energy minima mostly occurs around the integer values of n_+ and these minima are very narrow in the n_+ dimension. Moreover, recent publications with two-dimensional free energy landscapes in the r_{+-} and n_+ coordinates revealed the fact that for group 1 or group 2 cations, the SSIP - CIP transition is usually characterized by an initial association of the cation and the anion from the SSIP coordinated state to an intermediate structure where the cation and the anion both have the CIP separation but the cation still have the same coordination number as the SSIP state, then a water molecule is ejected from the first solvation shell in a much faster process to form a CIP structure. The results indicate that for group 1 and group 2 cations, the preferred Eigen - Wilkins mechanistic label for the cation - anion association process in aqueous solution is the associative pathway.

Despite recent efforts in incorporating the n_+ collective variable in free energy computations for this type of process, the contribution from solvent molecules in the outer solvation shell still remains unclear as the n_+ coordinate does not include any water molecules in any outer solvation shells, nor does it include the contributions from the solvation shell of the anion. In this regard, a recent work by Mullen et al. found that by maximizing the likelihood of a linear combination of 3 collective variables out of their 73 candidates with respect to the crossing of the committor surfaces, there are two solvent coordinates that play significant roles in the ionic association reaction of NaCl in aqueous solution, which are the number of water molecules that are simultaneously coordinated with both the cation and the anion by forming bridges between them (n_B), and the density of water molecules around the midpoint between the cation and the anion (ρ_{ii}). [29] The interpretation with respect to n_B collective variable is particularly interesting, as visualizing the n_B coordinate in the transition state is consistent with the earlier proposal from the work of Belch et al. [15] If we also added the ρ_{ii} coordinate into the mix, we would have another coordinate that takes the water molecules in the outer solvation shells into an account as well. Hence, the linear combination of the collective variables proposed by Mullen et. al. with maximum likelihood across the isocommittor surfaces [CHECK] represents the most likely reaction coordinate across the slowest motion across the actual free energy barrier between the CIP and the SSIP states, which indicates that having either one or two collective variables are not sufficient in describing the dynamics of the ionic association in aqueous solutions, and an optimum reaction coordinate across the barrier involves a specific linear combination of the collective variables.

1.3 Areas of Improvement and Objectives

The work of Mullen et al. changed the perception of how we should approach this problem in a hugely significant manner by introducing a number of possible candidates to describe the dynamics of the solvent; however, there are still several obstacles for us to fully grasp the whole picture. There are two possible questions that arose from Mullen et al. First, although the proper reaction coordinate can be found by maximizing the likelihood of a linear combination of collective variables that go across the dividing surface between two isocommittor surfaces, there is a possibility that more than 3 collective variables are actually involved and a linear combination of 3 collective variables may still be inadequate to properly representing the slowest motion. Second, the work of Mullen et al. focused mainly in the SSIP - CIP transition, but did not discuss the dynamics of the bulk, so the mechanics of the ionic association from the bulk remain relatively little understood.

There are several ways that can be used to identify proper reaction coordinates for a chemical reaction. The maximum likelihood method used in Mullen et al. is just one method among many other proposals of methods that allow us to study the most important coordinate of a chemical system, such as string method [30–32], markov state model (MSM) [33–37], principal component analysis (PCA), diffusion map, and many other. [38–41] Among these methods, MSM is one of the most widely used tools in biochemical simulations to identify different conformations of large biomolecules such as proteins or DNA. [42–45] Besides being used to identify metastable states in the system, the information from MSM eigenfunctions also directly relates to different motions in a chemical system. Therefore, not only that MSM could give an insight for the slowest motion, MSM can also be useful in a stituation where a system with motions in a very similar timescale and the slowest motion cannot be fully distinguished from the next slowest motion in the system. This would allow us to study any possible secondary or tertiary motions and their possible effects on the primary motion across the free energy barrier. Besides, the information of each MSM eigenfunction can also be projected onto the collective variables space and we could determine each unique linear combination of the collective variables specific to each motion in the system. [46–48]

In accompanying a proper reaction coordinate which may have contributions from several collective variables, a multidimensional free energy landscape is needed to represent those collective variables to give a viable insight into the reaction mechanisms. However, the current capabilities limited the free energy computation to a two-dimensional space due to the high computational cost of multidimensional sampling. As relative free energy between two points in the collective variable space relates to the natural logarithm of the probability of going from one state to another, any free energy barriers are considered as rare event regions that are naturally less sampled for any unbiased molecular dynamics simulations. Hence, we have relatively little information to correctly determine the free energy landscapes around the barriers, which may cause errors in rate calculation and inaccurate description of the transition state - a crucial element for full understandings of the entire mechanism. For this type of problem, Cuendet and Tuckerman has noted the difficulties in the free energy computation of the ionic association reaction of NaCl in aqueous solution which presents challenges for any novel free energy computation methods that usually validate their models with the free energy landscape of alanine dipeptide with respect to their two dihedral angles. Compared to the free energy of alanine dipeptide, the free energy of NaCl's SSIP and CIP states differ only in the order of a few kcal/mol, prompting the need to make sure that we need a free energy calculation scheme that cannot produce a large error which would otherwise cause severe inaccuracies in the result. Moreover, the CIP feature of NaCl's one-dimensional free energy is relatively narrow, which implies that the free energy gradient is relatively rapidly changing and giving rise to statistical noise when averaging the free energy estimators. [49]

The preferred method of free energy computation among the community has always been window-based simulations, where an average of a probability density or average force along a particular collective variable was computed in a restrained simulation environment to ensure adequate amount of sampling, especially in the rare event regions. Despite providing accurate results, the main disadvantage of this class of free energy simulation method has always been the cost, which scales as $\mathcal{O}(n^D)$ for a D-dimensional problem. Although it is also possible to perform a simulation in each window in parallel given that we know a good initial configuration in each window, this does not eliminate the scaling issue for a many-dimension problem. [50] Another main issue in free energy simulation is the inherent statistical noises of the average probability density or an average force along a collective variable, which also produce noisy free energy landscapes that may become an issue for a chemical system with sensitive free energy profile.

In order to circumvent the two issues of free energy computations presented in the above paragraph, we need to first be able to quickly explore the free energy landscape to eliminate the need to perform windowed simulation while also collecting the needed information to compute the free energy landscapes, and then we need to find a reconstruction algorithm that ensures a smooth reconstructed data. The first task could be done using well-tempered metadynamics (WT-MTD) simulation, which has been shown in the work by Mones et al. to have fastest exploration of the free energy landscape compared to many other algorithms. [51] During the WT-MTD simulation, biased instantaneous forces (BIFs) along the collective variables are collected and then later are unbiased using the known information of the deposited WT-MTD Gaussians to obtain unbiased instantaneous forces (UIFs). The UIFs could be treated as a noisy, unaveraged version of the average force along the collective variable commonly used in thermodynamic integration to compute the free energy landscapes. Therefore, a machine learning-based approach such as Gaussian process regression (GPR) could be used to infer the most likely free energy landscape based on our training data of UIFs, providing a smooth reconstruction of the free energy landscape in any number of dimensions. [51, 52]

1.4 Organization of Chapters

This dissertation presents our research which aim to address the problems highlighted in 1.3, and contains 6 chapters including this chapter. Chapter 2 is dedicated to discussing the theories of MSM and the projection of MSM eigenfunctions onto the collective variable spaces using tICA and subsequent dimensional reduction through matching pursuit (MP). Chapter 3 presents our application of MSM to the classical MD simulations of LiCl and NaCl in aqueous solutions and our discussions on the significance of the results in terms of the plausible SSIP - CIP transition mechanism based on the reaction coordinates found using MSM and the dynamics of the bulk. Chapter 4 discusses the theories of free energy computation and the best way to overcome the dimensionality problem using a combination of fast exploration through WT-MTD simulations and a machine learning-based reconstruction of free energy with GPR to ensure a smooth reconstruction of the free energy surfaces in any numbers of dimensions. Chapter 5 features the applications of GPR to compute the two-dimensional free energy landscapes of LiCl and NaCl in the r_{+-} and n_+ coordinates, which is the first application of a two-dimensional GPR calculation beyond the peptide rotation models, and chapter 6 highlights the present challenges and possible future directions of this kind of research,

CHAPTER 2

Identification of Reaction Coordinates and the Kinetic Model of Chemical Reactions

2.1 Introduction

Although intuitions allow us to describe a reaction coordinate that suits the narratives of the researchers' perceptions, the main problem in determining the contributing collective variables for a reaction coordinate from pure intuition is there are no systematic ways to confirm the researchers' beliefs. The fact that n_+ was used as a collective variable in the solvent coordinate demonstrates the intuition-based choice that are relatively accepted in many recent literatures. For example, the recent work of Raiteri et al. uses this fact to construct two-dimensional free energy landscapes with interionic separation (r_{+-}) and n_+ as collective variables of the alkali earth ions pairing interactions with carbonate ions in aqueous solution to validate their empirical potential model for these ions. [26] Roy et al. also recently published a two-dimensional free energy calculation using the same set of C.V.'s for alkali ions. [27, 28] However, the work of Mullen et al. challenged this belief by claiming that n_+ does not play a significant role in the process at all. On the other hand, the solvent coordinate that plays an important role for this process are n_B and ρ_{ii} coordinates mentioned in 1.2. [29]

The significance of the work of Mullen et al. was that it was the first work that performed a systematic analysis of multiple collective variables to determine the best contribution of the collective variables according to the reaction coordinate. However, we have mentioned in 1.3 that there are some possible issues that arose from their results, mainly his crude restriction of the reaction coordinate that should only consist of a linear combination of 3 collective variables, while there are possibly more collective variables than 3 that could actually involve in the process. Moreover, their work mainly focused on the SSIP - CIP transition without thoroughly covering the behavior of the bulk region. Therefore, the association pattern from the bulk still remained relatively little understood. Also, the efficiency of the method outlined in Mullen et al. also rely a lot on the fact that we could accurately find two isocommittor surfaces and the events that the simulation trajectory crosses between these two, which could be cumbersome in the case where systems do not have a well-defined barrier or the system with multiple metastable states that are close to one another with relatively low barrier.

As mentioned in Rohrdanz et al., there are many possible methods that can be used to tackle this problem. [41] Out of the methods highlighted in Rohrdanz et al., the Markov State Model (MSM) is a viable candidate yet it is still not being applied to this type of problem. In our opinion, the fact that any MD simulations that is ergodic are shown to exhibit Markovian properties [53] implies that we could express the dynamics of any MD simulations in terms of their chatacteristic eigenfunctions and eigenvalues corresponding to specific transitions between metastable states. [36, 37, 44, 48, 54] Therefore, MSM allows a characterization of any transition between any metastable states besides the slowest motion across the highest free energy barrier. The MSM eigenfunctions also entail important information on dynamical variables, and could be projected onto the collective variable space to gain valuable insights on all possible variables corresponding to different processes in a chemical system. All of these benefits could be achieved with a relatively short simulation time; hence, MSM is an attractive tool among simulations in biochemical processes. [42, 44, 45, 55–57] Nevertheless, there are no MSM interpretations for processes governing the solution dynamics at all. Therefore, we hope to present the MSM interpretation of two model reactions for ionic association process of LiCl and NaCl in aqueous solutions.

2.2 Markovianity of MD Simulations

In 1983, Zwanzig proposed that the transition between two metastable configurations in the phase space can be treated as a continuous time random walk problem. [53] In order to understand the Markovianity of a MD trajectory, it is necessary to think about MD simulations in terms of the probability density of possible configurations, where the evolution of configurations over time is governed by a generalized classical master equation for the distribution of the waiting times to change from one configuration to another. The short memory approximation of the transition memory kernels from the master equation can then be approximated, and it ignores the possibility of the later return to the same state. Hence, if the system's dynamics is sufficiently complex and metastable states are chosen sensibly, then this implies the memoryless characteristics of the interstate jumps. [53]

As opposed to the typical view of the MD simulations as composed of distinct trajectories, MSM takes an ensemble approach given that the dynamics is ergodic in the phase space Ω, that is, there always exist a connected configuration to the current configuration in the phase space. Therefore, the evolution of the ensembles of the trajectories take a probabilistic approach in MSM. Since we assert that a transition from one configuration to another configuration after a lag time τ is Markovian; therefore, starting from a configuration $\mathbf{x}$ at time t where $\mathbf{x} \in \Omega$, the probability that a trajectory starting at $\mathbf{x}$ at time t will be in an infinitestimally small region $d\mathbf{y}$ around point $\mathbf{y} \in \Omega$, $p(\mathbf{x}, \mathbf{y}; \tau)$ can be defined as, [36, 37]

$$p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y} = \mathbb{P}[\mathbf{x}(t+\tau) \in \mathbf{y} + d\mathbf{y} | \mathbf{x}(t) = \mathbf{x}] \tag{2.1}$$

Therefore, equation 2.1 implies that for a set of configurations $A \subset \Omega$, the following also holds true for all configurations $\mathbf{y} \in A$,

$$p(\mathbf{x}, A; \tau) = \int_{\mathbf{y} \in A} p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y} \tag{2.2}$$

Since the phase space Ω is assumed to be ergodic, the state $\mathbf{x}$ will be visited infinitely often as $t \to \infty$. Hence, a unique stationary probability density $\mu(\mathbf{x})$ can be written to represent this fact, and $\mu(\mathbf{x})$ represents the ensemble's equilibrium density; for example, for a canonical ensemble (NVT variables are held constant), $\mu(\mathbf{x})$ can be written as,

$$\mu(\mathbf{x}) = \frac{e^{-\beta H(\mathbf{x})}}{Z} \tag{2.3}$$

where $Z = \int \exp(-\beta H(\mathbf{x})) d\mathbf{x}$ is the canonical partition function, $\beta = (k_B T)^{-1}$, and $H(\mathbf{x})$ is a classical Hamiltonian of the system. In order to model reversible reactions, another assumption that the configuration $\mathbf{x}(t)$ is reversible is needed. Therefore, the following detailed balance condition,

$$\mu(\mathbf{x}) p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y}) p(\mathbf{y}, \mathbf{x}; \tau) \tag{2.4}$$

as to be satisfied. Considering a probability density of a configuration $p_t(\mathbf{x})$, the transition probability $p(\mathbf{x}, \mathbf{x}; \tau)$ governs that after some times τ, the probability density of $\mathbf{x}$ at time $t+\tau$is expressed as $p_{t+\tau}(\mathbf{x})$. Hence, one could define a propagator $\mathcal{Q}(\tau)$that satisfies the following properties,

$$p_{t+\tau}(\mathbf{y}) = \mathcal{Q}(\tau) p_t(\mathbf{y}) = \int_{\mathbf{x} \in \Omega} p(\mathbf{x}, \mathbf{y}; \tau) p_t(\mathbf{x}) d\mathbf{x} \tag{2.5}$$

When weighted by the stationary density, another way to look at equation 2.5 is through the transfer operator $\mathcal{T}(\tau)$ which propagates the weighted probability density. Thus, it must follow that,

$$u_{t+\tau}(\mathbf{y}) = \mathcal{T}(\tau)u_t(\mathbf{y}) = \frac{1}{\mu(\mathbf{y})}\int_{\mathbf{x}\in\Omega} p(\mathbf{x},\mathbf{y};\tau)\mu(\mathbf{x})u_t(\mathbf{x})d\mathbf{x} \tag{2.6}$$

In general, $\mathcal{T}(\tau)$ is a preferred operator because of the fact that $\mathcal{T}(\tau)$ operates on the weighted probability density. Therefore, it can be said that $\mathcal{T}(\tau)$ has to conform with the following properties,

1. $\mathcal{T}(\tau)$ fulfills the Chapman-Kolmogorov equation $u_{t+k\tau}(\mathbf{x}) = [\mathcal{T}(\tau)]^k u_t(\mathbf{x})$, where $[\mathcal{T}(\tau)]^k$ represents the k-th power of the transfer operator matrix.
2. There exist eigenfunctions ψ_i that corresponds to the eigenvalue problem $\mathcal{T}(\tau)\psi_i = \lambda_i\psi_i$, where λ_i is the corresponding eigenvalue of an eigenfunction ψ_i
3. ψ_i relates to the i-th eigenfunction of the propagator $\mathcal{Q}(\tau)$ through the stationary distribution; that is, $\mu(\mathbf{x})^{-1}\phi_i(\mathbf{x}) = \psi_i(\mathbf{x})$. Both ϕ_i and ψ_i share the same eigenvalue λ_i.
4. The eigenvalues λ_i are real numbers and $\lambda_i \in (-1, 1]$, and the first eigenvalue λ_1 is always 1 and corresponds to the stationary density $\mu(\mathbf{x})$, and it must follow that $1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$

Thus, the weighted probability density $u_{t+\tau}(\mathbf{x})$ can be written as a sum of all the eigenfunctions of $\mathcal{T}(\tau)$, which represent the spectral decomposition of the dynamics in our system. All motions in the dynamics are thought of the superimposition of independent motions represent by the i-th eigenfunction. Hence, $u_{t+k\tau}(\mathbf{x})$ is now written as,

$$u_{t+k\tau}(\mathbf{x}) = \sum_{i=1}^{m} \lambda_i^k \left\langle u_t, \psi_i \right\rangle_\mu \psi_i(\mathbf{x}) + \sum_{j=m+1}^{\infty} \lambda_j^k \left\langle u_t, \psi_j \right\rangle_\mu \psi_j(\mathbf{x}) \tag{2.7}$$

The first term of equation 2.7 represents m slowest motions which are deemed significant, while the second term represents all the other fast motions that may be considered irrelevant to the rate–determining process. $\left\langle u_t, \psi_i \right\rangle_\mu$ is simply the inner product between u_t and ψ_i weighted by the stationary density $\mu_{\mathbf{x}}$. Equation 2.7 also implies that in the limit where $k \to \infty$, only the first eigenfunction would remain as all other eigenvalues are strictly less than one except the first one, recovering the equilibrium distribution. The patterns of the eigenvalues thus would imply that for any motions where λ_i is less than 1, all the terms where $i > 1$ would decay over time according to the value of λ_i, which also determines the implied relaxation timescale of each motion through the following relationship,

$$t_i = -\frac{\tau}{\ln \lambda_i} \tag{2.8}$$

Equation 2.7 can thus be now written as,

$$u_{t+k\tau}(\mathbf{x}) = 1 + \sum_{i=2}^{m} e^{-\frac{k\tau}{t_i}} \left\langle u_t, \psi_i \right\rangle_\mu \psi_i(\mathbf{x}) + \sum_{j=m+1}^{\infty} \lambda_j^k \left\langle u_t, \psi_j \right\rangle_\mu \psi_j(\mathbf{x}) \tag{2.9}$$

The representation of $u_{t+k\tau}(\mathbf{x})$ outlined in equation 2.7 is now separated into three parts; the stationary distribution, the $m-1$ important process that have distinctly different eigenvalues, from which the implied timescale of the process can now be extracted. However, important information can also be found from the first $m-1$ eigenfunctions as well through the projection onto the collective variable space, which we will discuss in the following section.

2.3 Approximating the Eigenfunctions of the Transfer Operator

2.3.1 Approximation of the Eigenfunctions through Variational Principle

The solution of an eigenvalue problem in equation 2.7 gives us $m-1$ eigenvalues and eigenfunctions that best approximates the probability of arriving at an MD configuration at time τ after the current time. As mentioned in 2.2, the information provided by eigenvalues implies the inherent timescale of their corresponding eigenfunctions. However, the eigenfunctions themselves also contain several important dynamical information of the process of interest. We could see that the inner product $\langle u_t, \psi_i \rangle$ tells us about the projection of the probability density of the configuration at current time; thus, the eigenfunction ψ_i with the highest eigenvalue contains the contributions that give rise to the slowest motion for this process. In a situation where we have $\lambda_2 \gg \lambda_i \; \forall i > 2$, it readily implies that the slowest motion is a dominant motion and the other eigenfunctions decay much faster at lag time τ, and ψ_2 would be able to be taken to dominate the dynamics.

Equations 2.7 and 2.9 provided a spectral decomposition of all different motions in the system, where the information of each motion can be obtained from the eigenvalue problem $\mathcal{T}(\tau)\psi_i = \lambda_i \psi_i$. The implied timescale of each process could also be obtained from the eigenvalues $\lambda_i = e^{-\frac{\tau}{t_i}}$. However, we still do not know how to extract the dynamical information from the eigenfunctions. Nonetheless, equation 2.7 implies that for any function that relates to a configuration $\mathbf{x}$ at time t, the time-autocorrelation function of an arbitrary function f as a function of τ can be written as,

$$\langle f(\mathbf{x}_t) f(\mathbf{x}_{t+\tau}) \rangle_t = \sum_{i=1}^{\infty} e^{-\frac{\tau}{t_i}} \langle \phi_i, f \rangle^2 \tag{2.10}$$

Hence, the time-autocorrelation function of the i-th normalized eigenfunction ψ_i can be used to recover the i-th eigenvalue due to the fact that for a normalized eigenfunction, $\langle\psi_i(\mathbf{x}_t)\psi_i(\mathbf{x}_{t+\tau})\rangle^2 = 1$. Thus, given that the eigenfunctions can be computed, the eigenvalues and the implied timescales can be approximated from the time-autocorrelation function as a function of τ, now called a *lag time*.

$$\tilde{\lambda}_i = \langle\psi_i(\mathbf{x}_t\psi_i(\mathbf{x}_{t+\tau}))\rangle \tag{2.11}$$

Nevertheless, the main question here is how would we know the eigenfunction. It is very likely that we will never know the true form of the eigenfunctions ψ_i. However, since the eigenvalues λ_i can be modeled from the time-autocorrelation of ψ_i, we always know that for any model eigenfunction $\tilde{\psi}_i$,

$$\left\langle\tilde{\psi}_i(\mathbf{x}_t)\tilde{\psi}_i(\mathbf{x}_{t+\tau})\right\rangle \leq e^{-\frac{\tau}{t_i}} \tag{2.12}$$

Hence, the variational principle can be applied to find a good approximation of $\tilde{\psi}_i$ given that we could find such a function that gives $e^{-\frac{\tau}{\tilde{t}_i}} \leq e^{-\frac{\tau}{t_i}}$ with a value of $\tilde{t}_i$, the variational approximation of t_i, as close as possible to t_i. Since the equality between the modeled and the actual implied timescale only holds if $\tilde{\psi}_i = \psi_i$, the modeled eigenfunction $\tilde{\psi}_i$ that best approximates the true eigenfunction ψ_i needs to maximize the modeled implied timescale $\tilde{t}_i$, or equivalently, maximize the modeled eigenvalue $\tilde{\lambda}_i$. In fact, the principles behind the application of the variational principle to approximate the eigenfunctions of MSM is analogous to that of quantum mechanics, where the wavefunctions are best approximated by minimizing the energy.

2.3.2 Time-lagged Independent Component Analysis (TICA)

Suppose that a modeled eigenfunction $\tilde{\psi}_i$ can be modeled as a linear combination of an orthonormal basis of the ansatz χ_k, the reconstruction of $\tilde{\psi}_i$ as a linear combination of basis functions in the set $\chi = \left\{\chi_1, \chi_2, \ldots, \chi_{N_\chi}\right\}$ would take the following form,

$$\tilde{\psi}_i = \sum_{k=1}^{N_\chi} b_{ik}\chi_k \tag{2.13}$$

where the optimal set of coefficients b_{ik} can be determined from solving the following eigenvalue problem,

$$\mathbf{C}^\chi(\tau)\mathbf{b}_i = \mathbf{b}_i\tilde{\lambda}_i(\tau) \tag{2.14}$$

where $\mathbf{C}^\chi(\tau)$ is the autocorrelation matrix of the ansatz functions at time τ with the following form,

$$\begin{aligned} c_{ij}^\chi(\tau) &= \langle\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+\tau})\rangle_t \\ c_{ij}^\chi(0) &= \langle\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_t)\rangle_t \end{aligned} \tag{2.15}$$

under the condition of orthonomal ansatz functions $\langle\chi_i, \chi_j\rangle_\mu = \langle\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_t)\rangle = \delta_{ij}$, where δ_{ij} is the Kronecker's delta. However, for a non-orthonomal basis of the set χ, the basis must be orthonormalized first through generalizations of equation 2.14 by using the autocorrelation of the ansatz function at time 0.

$$\mathbf{C}^{\chi}(\tau)\mathbf{b}_i = \mathbf{C}^{\chi}(0)\mathbf{b}_i\tilde{\lambda}_i(\tau) \tag{2.16}$$

Solving equation 2.16 yields an optimal set of the coefficients $\mathbf{b}_i$ for any non-orthonormal basis set χ. In our case, since we would like to determine the optimal set of collective variables for a particular eigenfunction, we would then construct a basis set consisting of as many collective variables as possible then solve equation 2.16 for optimal coefficients for each collective variable in $\tilde{\psi}_i$. In order to properly obtain the autocorrelation function in the basis set of the collective variables, we need to subtract the mean of each collective variable to create the mean–free input from our data such that,

$$\xi_i^{MF}(\mathbf{x}) = \xi_i(\mathbf{x}) - \langle \xi_i(\mathbf{x}) \rangle_t \tag{2.17}$$

Therefore, we can now write an approximation of the eigenfunction $\tilde{\psi}_i$ as a linear combination of the mean–free collective variables ξ_k^{MF},

$$\tilde{\psi}_i = \sum_{k=1}^{N_\xi} b_{ik}\xi_k^{MF} \tag{2.18}$$

Consequently, we have demonstrated that $\tilde{\psi}_i$ can be approximated from the time–auto-correlation of itself at a lag time τ, and $\tilde{\psi}_i$ can be projected onto the collective variable space. According to equation 2.7, only $m-1$ eigenfunctions are needed to sufficiently describe the dynamics. Since λ_i and t_i are functions of τ, it is appropriate to only select the eigenfunctions such that the implied relaxation timescale is greater than τ in order to take all the slowest motions beyond the lag time into an account. Nonetheless, one problem remains — since $\tilde{\psi}_i$ is a linear combination of all the collective variables in the provided basis set, it can be hard to interpret the physical meaning of $\tilde{\psi}_i$ when it involves a lot of collective variables. Hence, a reduced representation of $\tilde{\psi}_i$ written in terms of only relevant collective variables is desirable for the physical interpretation purposes. In order to achieve this, we use an algorithm called Matching Pursuit (MP), which we will discuss in the following section.

2.3.3 Matching Pursuit (MP): Reduced Representation for Eigenfunctions

If $\tilde{\psi}_i$ is expanded with a non–orthonormal basis, some of the collective variables may not be entirely independent; hence, complicating the situation further by having coefficients in two or more possible collective variables that are not independent, and does not help us in achieving our goal to truly reducing the dimensionality expression of the chemical process of interest. Matching Pursuit algorithm (MP), first proposed by Mallat and Zhang [58], can help us achieve our goal by finding a sparse solution of $\tilde{\psi}_i$ and reassign the coefficients accordingly.

Let us begin by suppose that there are $m - 1$ modeled eigenfunctions $\tilde{\psi}_i$ such that $2 \leq i < m$. These $\tilde{\psi}_i$ are the most important components that we obtained from solving the TICA problem from the previous section. If we initially build such eigenfunctions from our library of N_ξ collective variables, then each $\tilde{\psi}_i$ is simply a linear combination of those N_ξ variables as expressed in equation 2.18. However, as each $\tilde{\psi}_i$ is a function of N_ξ variables, often times, it is very hard to use these variables altogether to discern the physical meaning of each $\tilde{\psi}_i$. Preferably, one would prefer that each $\tilde{\psi}_i$ has a physically meaningful representation with only dominant collective variables represented. The MP algorithm reduced the representation of $\tilde{\psi}_i$ by finding only the productive coefficients that best summarize the behavior of the function based on our data. To better understand the MP algorithm, we would assume an arbitrary function $f(t)$ that can be written as a linear expansion on a basis $\chi = \{\chi_1, \chi_2, \ldots, \chi_k\}$ that could either be orthonormal or not.

$$f(t) = \sum_{i=1}^{k} b_i \chi_i(t) \tag{2.19}$$

The inner product of f and a basis function $\chi_k(t)$, $\langle f, \chi_k(t) \rangle$, can be computed from the following equation,

$$\begin{aligned} \langle f, \chi_k \rangle &= \sum_{i=1}^{N} b_i \langle \chi_i, \chi_k \rangle \\ &= \sum_{i=1}^{N} b_i c_{ik}^{\chi}(0) \\ &= \sum_{i=1}^{N} c_{ki}^{\chi}(0) b_i \\ &= \left(\mathbf{C}^{\chi}(0)\mathbf{b}\right)_k \end{aligned} \tag{2.20}$$

And a residual norm of f, $\left\| f_{res}^2 \right\|$, is computed as follow,

$$\begin{aligned} \left\| f_{res}^2 \right\| &= \left\langle \sum_{i=1}^{N} b_i \chi_i(t) \sum_{j=1}^{N} b_j \chi_j(t) \right\rangle \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} b_i b_j \langle \chi_i(t) \chi_j(t) \rangle \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} b_i c_{ij}^{\chi}(0) b_j = \mathbf{b}^{\top} \mathbf{C}^{\chi}(0) \mathbf{b} \end{aligned} \tag{2.21}$$

The algorithm to compute the productive coefficients, $\mathbf{b}^{\ddagger}$ of f uses $\langle f, \chi_k \rangle$ and $\left\| f_{res}^2 \right\|$ computed from equations 2.20 and 2.21. This presents an iterative method that converges to a set tolerance value of $\left\| f_{res}^2 \right\|$. The result of this algorithm is the reduced representation of f, $f^{\ddagger}$, such that,

$$\begin{aligned} f^{\ddagger}(t) &= \sum_{i=1}^{N_{reduced}} b_i^{\ddagger} \chi_i(t) \\ &\approx f(t) \end{aligned} \tag{2.22}$$

where $N_{reduced} < N$. As shown in 2.3.2, TICA computes $\tilde{\psi}_i$ according to the variational principle to get a best approximation that is reasonably close to ψ_i. Since TICA computes $\tilde{\psi}_i$ as a linear projection onto the basis of collective variables, the MP projection of $\tilde{\psi}_i$, $\psi_i^{\ddagger}$, can be computed in the same fashion as equation 2.18,

$$\begin{aligned} \psi_i^{\ddagger} &= \sum_{k=1}^{N_{reduced}} b_{ik}^{\ddagger} \chi_k^{MF} \\ &\approx \tilde{\psi}_i \end{aligned} \tag{2.23}$$

$\psi_i^{\ddagger}$ is now expressed as a linear combination of $N_{reduced} < N_{\xi}$ variables. Therefore, a reduced representation of ψ_i would allow us to interpret the dynamics of each motion involving only important collective variables that are dominant in each specific ψ_i. This way, one could gain important mechanistic insights into the slowest dynamics in any kinds of systems.

2.4 Markov State Model (MSM)

The application of TICA to a mean–free input data in the collective variable space results in a set of $m - 1$ slowest eigenfunctions $\tilde{\psi}_i$ with corresponding implied relaxation timescale $\tilde{t}_i$ associated with its eigenvalue $\tilde{\lambda}_i$ that is slower than a lag time τ of interest. According to Noé et al., these $\tilde{\psi}_i$ contain enough information to cover the entire kinetic map with cumulative kinetic variance very close to 1. [59, 60] However, these MD input data are often very large, containing millions of data points. This presents a major challenge in the analysis of this data, because many algorithms scale poorly for matrices with size in the order of millions by millions. In order to ease the computational workload to build a kinetic model of an MD trajectory, discretization of a large data set with respect to an appropriate subspace that is thought to completely describe the whole data is a viable strategy that sparsifies a large matrix into a more managable problem.

Once we have proved that our set of $\tilde{\psi}_i$ has a cumulative kinetic variance very close to 1, we can choose to make a discretization within the subspace of $\tilde{\psi}_i$. Usually, the discretization is done through k-means clustering [61, 62]. The discretization generates a Voronoi diagram with a specific n clusters, each of which has a center weighted with respect to the distribution of the original data in the space of $\tilde{\psi}_i$. In order to construct a Markov State Model from these n clusters, the population in each cell of the Voronoi diagram is counted using a step function $\theta_i(\mathbf{x})$ defined as follow,

$$\theta_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in S_i \\ 0 & \mathbf{x} \notin S_i \end{cases} \tag{2.24}$$

where S_i is the i-th cluster of the Voronoi diagram. From this model, a transition probability matrix of the discretization $\mathbf{T}(\tau) \in \mathbb{R}^{n \times n}$ can be computed from θ_i from equation 2.24,

$$T_{ij}(\tau) = \frac{\langle \theta_j, (\mathcal{T}(\tau)\theta_i) \rangle_\mu}{\langle \theta_i, \theta_i \rangle_\mu} \tag{2.25}$$

The matrix element transition matrix $\mathbf{T}(\tau)$, T_{ij}, approximates the transition probability between the states i and j, which represents a discretized version of the transfer operator $\mathcal{T}(\tau)$. Moreover, the eigenvalues and eigenvectors of $\mathbf{T}$ can also be computed, and according to the variational principle, the eigenvalues obtained with MSM is better than those obtained from TICA, representing a better relaxation timescale for each eigenfunction. [46, 63] However, if the size of n is relatively large, the kinetic model of a chemical reaction can still be hard to interpret. Therefore, $\mathbf{T}(\tau)$ can be further coarse–grained into a very small set of states that we hope to describe the key metastable states presented in the dynamics. The coarse–graining process is done using Perron Cluster Cluster Analysis (PCCA) [64] to look at the structure of the eigenfunctions obtained from MSM. The kinetic information, such as the average time spent to move from states i to j, can be computed through the commute map which make use of the scaling of the TICA eigenfunction. [60]

2.5 Summary

A molecular dynamics simulation can be interpreted in the probabilistic view, where the Markovian properties hold in the phase space Ω. Taking an advantage of the Markovian properties, eigenfunctions and eigenvalues of the transfer operator $\mathcal{T}(\tau)$ contain numerous information pertaining the dynamics of any system of interest, where different motions in the system can be viewed as a spectral decomposition of the weighted probability density for a particular configuration $u_{t+\tau}(\mathbf{x})$ after a lag time τ. Each of the motion has a corresponding eigenvalue, which relates to the relaxation timescale, where each motion decays differently. As each eigenfunction strictly corresponds to a specific motion in the system, the terms *eigenfunction of the transfer operator* and *reaction coordinate* can be use synonymously.

Time-lagged Independent Component Analysis (TICA) allows us to approximate the eigenfunctions of $\mathcal{T}(\tau)$ under the condition of variational principle in a similar fashion to quantum mechanics. The approximated eigenfunctions by TICA can be projected onto the collective variable space, allowing us to interpret the physical meaning of different motions in a chemical system. However, if a set of collective variable is large, the eigenfunctions obtained from TICA may be hard to decipher. A reduced representation of TICA eigenfunctions through Matching Pursuit (MP) can be computed, which reduced the number of variables needed to express the eigenfunctions, allowing us to deduce the physical interpretations of each reaction coordinate better.

After the $m - 1$ slowest processes are determined from TICA, a discretization of the input data can be performed in the subspace of those $m-1$ reaction coordinates to form a discrete Markov State Model of the process, which predicts a better relaxation timescale than TICA. The coarse–grained MSM allows us to interpret the transition probabilities between key metastable states in the system. Therefore, the contents presented in this chapter encompass the entire workflow of using the Markovian properties of MD simulations to gain the dynamical insights to a chemical reaction, which is summarized in figure 2.1.

Chapter 2, in full, is a part of the material titled "Markov State Modeling for Ion Pairing Dynamics in Aqueous Solutions" by Pornpatcharapong, Wasut, Noé, Frank, Clementi, Cecilia, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material, and all co–authors have approved the use of the material for this dissertation.

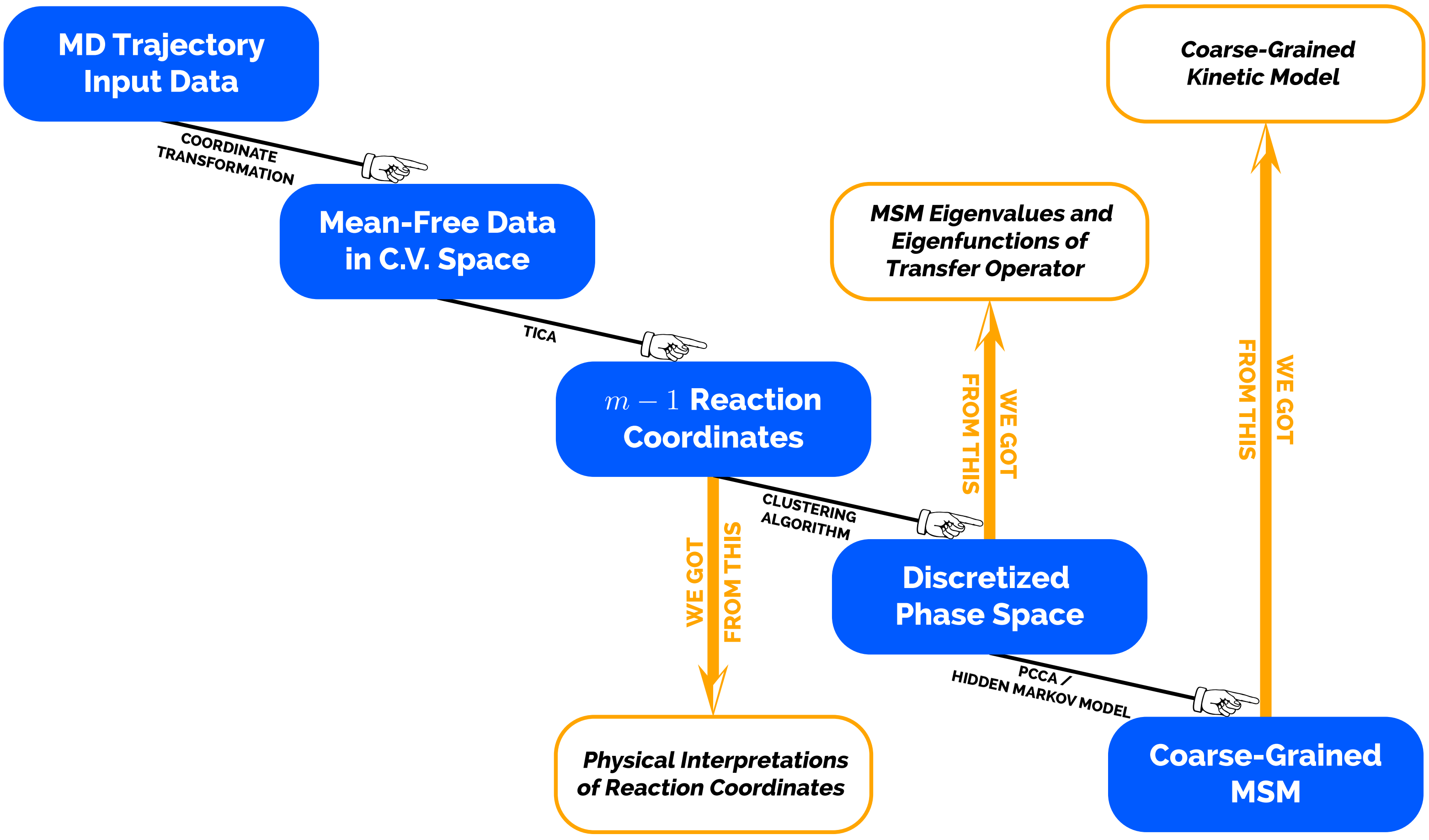
MD Trajectory Input Data
COORDINATE TRANSFORMATION
Mean-Free Data in C.V. Space
TICA
$m-1$ Reaction Coordinates
CLUSTERING ALGORITHM
Discretized Phase Space
PCCA / HIDDEN MARKOV MODEL
Coarse-Grained MSM
WE GOT FROM THIS
Physical Interpretations of Reaction Coordinates
WE GOT FROM THIS
MSM Eigenvalues and Eigenfunctions of Transfer Operator
WE GOT FROM THIS
Coarse-Grained Kinetic Model

Figure 2.1: A workflow of preparation of simulations for an analysis with MSM

CHAPTER 3

Ionic Association of NaCl in Aqueous Solution

3.1 Introduction

The Markovian interpretation of MD simulations opened a wide avenue for calculating many interesting properties from a rigorous mathematical formalisms introduced in Chapter 2. Nevertheless, there are many components in this formalism, such as Time-lagged Independent Component Analysis (TICA), Matching Pursuit (MP), Markov State Model (MSM), or the recently published concepts of commute maps and commute distances. [36, 37, 58–60, 63, 65–67] Therefore, often times, it is difficult for researchers to grasp the whole concept of the Markovianity of MD simulations, preventing them to fully utilize the potentials. Recently, there have been significant breakthroughs in the applications of MSM and TICA in numerous simulations for biological systems, where the metastable states of the biomolecular configurations and the mechanisms and the probability of transition between the metastable states can be computed. [35, 37, 48, 68–70] However, in our opinion, the entire package has rarely been used outside the biomolecuar simulation community. Hence, we believe that the interpretations of the Markovianity of chemical reactions can be applied to solve numerous problems in chemical dynamics, especially for problems that are easier to solve and a large enough amount of data can be obtained to assure that a simulation was performed to conform with an equilibrium under the canonical ensemble.

Therefore, we chose to apply the formulisms of Chapter 2 to study the ionic association of NaCl in aqueous solution. Despite being a relatively simple problem, the mechanistic point of view of ionic association in aqueous solution remains relatively little understood. The work of Mullen et al. has opened new insights to this type of problem through the introduction of the n_B

collective variable describing the number of water molecules simultaneously associating with both the cation and the anion. [29] This finding challenged the prevalent views of this process from the ligand exchange perspective, where the association of the ions is driven by the loss of water molecule from the cation's first solvation shell in order to make room for the CIP configuration. The identification of n_B as an important collective variable; thus, is an important step towards better understandings of this problem. Nevertheless, McGibbon et al. proposed that a true reaction coordinate representing the slowest motion of the dynamics must rigorously follow the following three conditions, [69]

1. It has to be at a reduced dimension from $\Omega \to \mathbb{R}$
2. It needs to be uniquely determined by the dynamics rather than being conditions enforced *a priori*
3. It needs to be a maximally predictive projection

In order to satisfy the three conditions above without any preconditioning of the committor surfaces a priori, we opted for the Markovian formulisms proposed in the previous chapter, where the slowest motion of the dynamics can be readily determined from the eigenfunction of the transfer operator with a corresponding highest eigenvalue that is not 1. From this formulism, the variational principle projection of the slowest eigenfunction can be done onto the basis set of collective variables through TICA, where the linear combination of the collective variables constitute a TICA eigenfunction, or a reaction coordinate. The TICA reaction coordinates can also be further reduced with MP to form a more physically interpretable version of the reaction coordinates focusing only on a few relevant collective variables. Having the information of m slowest reaction coordinates from TICA that covers the majority of the cumulative kinetic variance, a good MSM for this system can then be built based on the coordinates that fully cover the extent of the simulation, from where the transition matrix between the coarsed metastable states can be computed either through the Hidden Markov Model (HMM), or Perron-Cluster Cluster Analysis (PCCA), allowing us to calculate the transition rate between relevant metastable states and the commute map of this system. Thus, here we present the Markovian interpretation of the ionic association of NaCl in an aqueous solution.

3.2 Simulation Details

3.2.1 General Settings

The system, $\mathrm{NaCl} + 495\,\mathrm{H_2O}$, contains one Na cation, one Cl anion, and 495 water molecules, was prepared by placing the ions randomly in a box of 30.0 Å, and the water molecules were then placed in the same box using PACKMOL. [71] In order to get an equilibrium box size at 1.0 atm, the simulation in an isothermic-isobaric ensemble was performed for 960 ps using Nosé-Hoover Langevin piston [72, 73] Once the equilibrium box size is obtained, the production simulation was performed under the canonical ensemble using Langevin dynamics with a damping constant of 5.0 ps^{-1} with NAMD [74] with the simulation timestep of 2.0 fs. Water molecules in this simulation are treated as rigid under a TIP3P model [75], and the force field parameters for all atoms are derived from the work by Joung and Cheatam. [76] The electrostatic interactions were modeled by Particle Mesh Ewald [77] algorithm. The temperature of the simulation was controlled at 300 K, and periodic boundary conditions were applied throughout the simulation. The trajectory was saved every 10 steps. The total simulation time for this system is 220 ns.

The analysis part started from reading the entire trajectory with MDTraj [78] using a 10-

frame stride; therefore, each of the snapshot input is 100 steps, or 0.2 picoseconds apart from one another. The analysis started with finding the slowest variational eigenfunctions $\tilde{\psi}_i$ with TICA at different lag times τ, from where we picked a value of τ such that all the $m-1$ eigenfunctions consist of different motions that cover largest extent of the kinetic map possible. In order to form a Markov State Model (MSM) for this reaction, k-means clustering [61, 62] with k-means++ was used to discretize the input data into the space of $m-1$ $\tilde{\psi}_i$ obtained earlier from TICA, from which a coarse kinetic model, transition matrix, and commute maps can be obtained from Perron Cluster Clustering Approach (PCCA) or Hidden Markov Model (HMM) [64]. All the analysis from TICA to coarse-grained MSM was performed with a software package PyEMMA. [79]

3.2.2 Collective Variables

In order to get a best reaction coordinate, a good set of collective variables are needed so that there are more basis functions available for TICA. For this work, we chose two different sets of the collective variables for a comparison purpose. One set (CV Set 1) is a set of 13 collective variables modeled after the work by Mullen et al., and another set (CV Set 2) is a set of 32 intuition-based variables that are thought to describe both the ionic and the solvent coordinates. Table 3.1 shows brief descriptions of the chosen collective variables for this reaction for CV set 1, as well as their corresponding index numbers. Based on our choice for the collective variables, they can be further subdivided into the following classes,

The Interionic Separation Class - This class of the collective variable is based on the interionic separation between the cation and the anion, which is defined to be the Euclidean distance between the two ions in the simulation,

Table 3.1: A list of collective variables used in CV Set 1 inspired by the previous work in Mullen et al.

Index	**Feature**	**Feature Class**	**Remarks**
0	$n_+^{(1)}$	Ion Coordination	
1	$n_-^{(1)}$	Ion Coordination	
2	n_B	Water Between Ions	
3	$n_+^{(2)}$	Ion Coordination	Including first shell
4	$n_-^{(2)}$	Ion Coordination	Including first shell
5	$n_+^{(2)}$	Ion Coordination	Excluding first shell
6	$n_-^{(2)}$	Ion Coordination	Excluding first shell
7	r_{+-}	Interionic Separation	
8	ρ_{ii}	Water Density	$\sigma = 3.57$ Å
9	ρ_{ii}	Water Density	$\sigma = r_{+-}/4$
10	ρ_{ii}	Water Density	$\sigma = r_{+-}/3$
11	ρ_{ii}	Water Density	$\sigma = r_{+-}/2$
12	ρ_{ii}	Water Density	$\sigma = r_{+-}$

$$r_{+-} = \|\mathbf{r}_{\mathrm{Na}} - \mathbf{r}_{\mathrm{Cl}}\| \tag{3.1}$$

The collective variables belonging to this class are r_{+-} itself, as well as its derivatives, such as $\frac{1}{r_{+-}}$ or $\frac{1}{r_{+-}^2}$, which are defined in the case that the derivatives can describe the dynamics better than the original variable.

The Ionic Coordination Class - This class of the collective variables aim to describe the behavior relating to the first and the second solvation shells of the cation and the anion to monitor the possibility of the ligand exchange type of reactions involved in the ion pairing. For CV Set 1, the definition of the ionic coordination class collective variables are taken straight from the work of Mullen et al., with slight modification of parameters to suit our simulation better,

$$
\begin{aligned}
n_{+}^{(i)} &= \sum_{j=1}^{N_{wat}} \frac{1 - \tanh\left[\alpha\left(\left\|\mathbf{r}_{\mathrm{Na-O}_j}\right\| - b_{\mathrm{Na}}^{(i)}\right)\right]}{2} \\
n_{-}^{(i)} &= \sum_{j=1}^{N_{wat}} \frac{1 - \tanh\left[\alpha\left(\left\|\mathbf{r}_{\mathrm{Cl-H}_j}\right\| - b_{\mathrm{Cl}}^{(i)}\right)\right]}{2}
\end{aligned}
\tag{3.2}
$$

where $n^{(i)}$ represents the i-th solvation shell of a corresponding ion, $\mathbf{r}_{\mathrm{Na-O}_j}$ is simply $\mathbf{r}_{\mathrm{Na}} - \mathbf{r}_{\mathrm{O}}$, $\mathbf{r}_{\mathrm{Cl-H}_j} = \mathbf{r}_{\mathrm{Cl}} - \mathbf{r}_{\mathrm{H}}$ where $\mathbf{r}_{\mathrm{H}}$ is the vector that points to the nearest hydrogen atom of each water molecule to Cl^{-}, and $b^{(i)}$ is the i-th minimum of the radial distribution function between the ions and the water molecules. For this simulation, $b_{\mathrm{Na}}^{(1)} = 3.18$ Å, $b_{\mathrm{Na}}^{(2)} = 5.88$ Å, $b_{\mathrm{Cl}}^{(1)} = 3.98$ Å, and $b_{\mathrm{Cl}}^{(2)} = 6.28$ Å. The constant α also varies with the ions and the solvation shell, where $\alpha_{\mathrm{Na}}^{(1)} = 3$, $\alpha_{\mathrm{Na}}^{(2)} = 12$, $\alpha_{\mathrm{Cl}}^{(1)} = 7$, and $\alpha_{\mathrm{Cl}}^{(2)} = 15$.

The Water Between Ions Class - This class of the collective variables was first conceived in Mullen et al. [29], which aims to describe the simultaneous association of a water molecule with respect to either of the ion. Therefore, for higher r_{+-}, there should be no water molecules associating with both of the ions. However, for small r_{+-}, more water molecules can associate with the ions, bridging the ions together. The definition of n_B, the number of water molecules associating with both of the ions, is defined as follow,

$$
n_B = \sum_{j=1}^{N_{wat}} \max\left(\frac{1 - \tanh\left[\alpha\left(\left\|\mathbf{r}_{\mathrm{Na-O}_j}\right\| - b_{\mathrm{Na}}^{(1)}\right)\right]}{2}, \frac{1 - \tanh\left[\alpha\left(\left\|\mathbf{r}_{\mathrm{Cl-H}_j}\right\| - b_{\mathrm{Cl}}^{(1)}\right)\right]}{2}\right) \tag{3.3}
$$

where all the parameters, α, $b_{\mathrm{Na}}^{(1)}$, $b_{\mathrm{Cl}}^{(1)}$ are the same as the ones defined for the ionic coordination class. This variable, n_B, is used in CV Set 1.

The Water Density Class - This class of collective variables describes the density of water molecules around the midpoint between the two ions. In Mullen et al., this variable is defined as follow,

$$\rho_{ii} = \frac{1}{(2\pi\sigma^2)^{3/2}} \sum_{j=1}^{N_{wat}} \exp\left[-\frac{\left\| \mathbf{r}_{\mathrm{O}_j} - \mathbf{r}_{mid} \right\|^2}{2\sigma^2} \right] \tag{3.4}$$

where $\mathbf{r}_{mid}$ is the vector that points to the midpoint between the two ions. This variable has a unit of length^{-3}, and there are 5 variables in CV Set 1 belonging to this class, each of which has a different value of σ, which indicates how rapidly the density would vary around the midpoint between the ions. The smaller the value of σ, the density varies more abruptly. Another way to look at this variable is that it represents the number of water molecules inside the volume $V_{ii} = \left(2\pi\sigma^2\right)^{3/2}$. In this work, we used the values of σ at 3.54 Å, $r_{+-}/4$, $r_{+-}/3$, $r_{+-}/2$, and r_{+-}.

3.3 Results and Discussion

3.3.1 Reaction Coordinates of NaCl Ionic Association in Aqueous Solutions from TICA

Although $\tilde{\psi}_i$ approximated by TICA will not equal to the actual eigenfunction ψ_i of the transfer operator $\mathcal{T}(\tau)$, the main advantage of TICA is that $\tilde{\psi}_i$ can be computed directly from the correlation of the meanfree collective variables input transformed from any MD trajectory. As each $\tilde{\psi}_i$ is written as a linear combination of the basis function in the collective variable space, $\tilde{\psi}_i$, the contribution from each collective variable to a particular $\tilde{\psi}_i$ can be determined. Figure 3.1 shows the value of the implied relaxation timescale $\tilde{t}_i$, which directly relates to its corresponding $\tilde{\psi}_i$, where

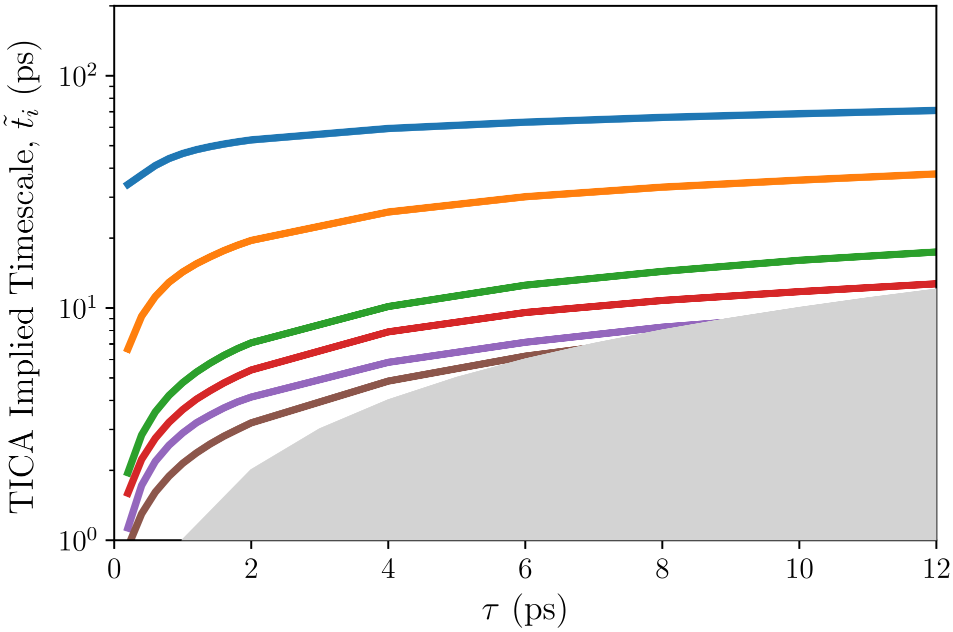

Figure 3.1: Different implied relaxation timescales ($\tilde{t}_i$) corresponding to each specific $\tilde{\psi}_i$ at different values of τ.

the slowest, nonstationary, relaxation timescale ($\tilde{t}_2$) ends at around 60 ps at $\tau = 12$ ps. In this figure, the second slowest relaxation timescale ($\tilde{t}_3$) has the same order of magnitude as $\tilde{t}_2$. Here one could readily see that the first two nonstationary eigenfunctions, $\tilde{\psi}_2$ and $\tilde{\psi}_3$ dominate the dynamics of NaCl in aqueous solution, with a relaxation timescale generally one order of magnitude greater than the next slower eigenfunctions. The shaded area of figure 3.1 represents the area where the implied timescale is less than the lag time; therefore, any eigenfunctions falling within this shaded area are treated as fast processes. Therefore, at $\tau = 4$ ps, there are 6 slowest reaction coordinates according to TICA, 2 of which are dominant and 4 others are auxiliary. Similarly, if one consider a long limit of τ (e.g. at 12 ps), the number of slowest reaction coordinates reduces to 4. Since we are interested in obtaining a better representation of the dynamics, picking the slowest reaction coordinates at $\tau = 4$ ps covers more processes than picking at $\tau = 12$ ps, allowing us to encode the information from these 6 coordinates during the discretization step for building MSM afterwards.

Now that we obtained 6 slowest reaction coordinates from the dynamics with TICA, and we have verified that these 6 reaction coordinates sufficiently describe the slowest dynamics by looking at a cumulative kinetic variance of these 6 coordinates. For this set of $\tilde{\psi}_i$, the cumulative kinetic variance is found to be 0.997, which is reasonably close to 1. The next question would be how do we describe these coordinates. In particular, we are interested in the first two slowest reaction coordinates as they are very close in relaxation timescales; thus, understanding both of these coordinates would be beneficial for understanding the reaction mechanism for this process. In order to determine the best correlated collective variable with respect to a reaction coordinate, a correlation matrix between the meanfree collective variable inputs and $\tilde{\psi}_i$ can be computed following an equation below,

$$C(\xi_i^{MF}, \tilde{\psi}_j) = \frac{1}{\sigma_{\xi_i^{MF}}} \sum_{k=1}^{N_\xi} \left[\mathbf{Cov}(\xi_i^{MF}, \xi_k^{MF})\right]^{1/2} \mathbf{U}_{ki} \tag{3.5}$$

Table 3.2 summarizes the correlation between the collective variables to $\tilde{\psi}_2$ and $\tilde{\psi}_3$ computed according to equation 3.5. The result for $\tilde{\psi}_2$ suggests that the 3 features that play important role in this reaction coordinate are r_{+-}, ρ_{ii} with $\sigma = r_{+-}$, and n_B. The interpretation of this reaction coordinate would have to involve these three collective variables, where the association of the ions would drive the number of water molecules simultaneously associated with the two ions up, changing the water density around the midpoint between the two ions, which is defined from the largest volume possible when the two ions are far apart. However, $\tilde{\psi}_3$, slightly faster in the relaxation timescale than $\tilde{\psi}_2$, mostly involve the changes in water density around the midpoint between the ions significantly more than the association / de-association of the ions.

Figures 3.2 and 3.3 also show a graphical correlation between $\tilde{\psi}_2$ and $\tilde{\psi}_3$ with respect to

Table 3.2: Collective variables' correlation with $\tilde{\psi}_2$ and $\tilde{\psi}_3$

Feature	**Correlation with $\tilde{\psi}_2$**
r_{+-}	8.22
ρ_{ii} ($\sigma = r_{+-}$)	-7.70
n_B	-6.02
ρ_{ii} ($\sigma = r_{+-}/4$)	3.47
ρ_{ii} ($\sigma = 3.57$ Å)	3.21
ρ_{ii} ($\sigma = r_{+-}/2$)	2.45

Feature	**Correlation with $\tilde{\psi}_3$**
ρ_{ii} ($\sigma = r_{+-}/3$)	-3.30
ρ_{ii} ($\sigma = r_{+-}/2$)	-3.19
ρ_{ii} ($\sigma = r_{+-}/4$)	-2.82
n_B	1.94
r_{+-}	1.71
$n_{+}^{(1)}$	-1.65

their 3 dominant collective variables. The time series plots for both $\tilde{\psi}_2$ and $\tilde{\psi}_3$ are consistent with the results we obtained in table 3.2. The slowest motion, $\tilde{\psi}_2$, contains about 20 transitions to CIP regions (where r_{+-} is minimum) and several transitions to the SSIP regions (where $r_{+-} \approx 5.0$ Å), and each of these transitions in the inter ionic distance has a pattern in the time series that matches with the evolution of $\tilde{\psi}_2$ over time. Moreover, the evolution of n_B peaks around the same point where r_{+-} is at minimum, while remaining mostly zero throughout the course of the simulation. This result indicates that the CIP configuration of NaCl has to occur in tandem with at least two water molecules simultaneously coordinating with both ions, whereas the region where $n_B = 0$ indicates the bulk region. The SSIP region is usually identified when $r_{+-} \approx 5.0$ Å, where numerous previous literature has consistently found this value from the onedimensional free energy landscape computation of NaCl in aqueous solutions [20, 21, 29, 80], occurs in sync with the region of $n_B \approx 1$, indicating that there is only 1 water molecule bridging between the two ions in the SSIP structure.

According to figure 3.3, the key transition in this reaction coordinate is observed with collective variables 10, 11, and 9 (water density around the midpoint between the ions), all of which correlate strongly with this reaction coordinate. However, the minimum r_{+-} from figure 3.2 does not correlate very well with the high jumps in $\tilde{\psi}_3$, indicating that the association between the ions does not play a key role in this reaction coordinate, which is dominated by the solvent rearrangement around the midpoint of the ions. As this is a slightly faster process than the reaction

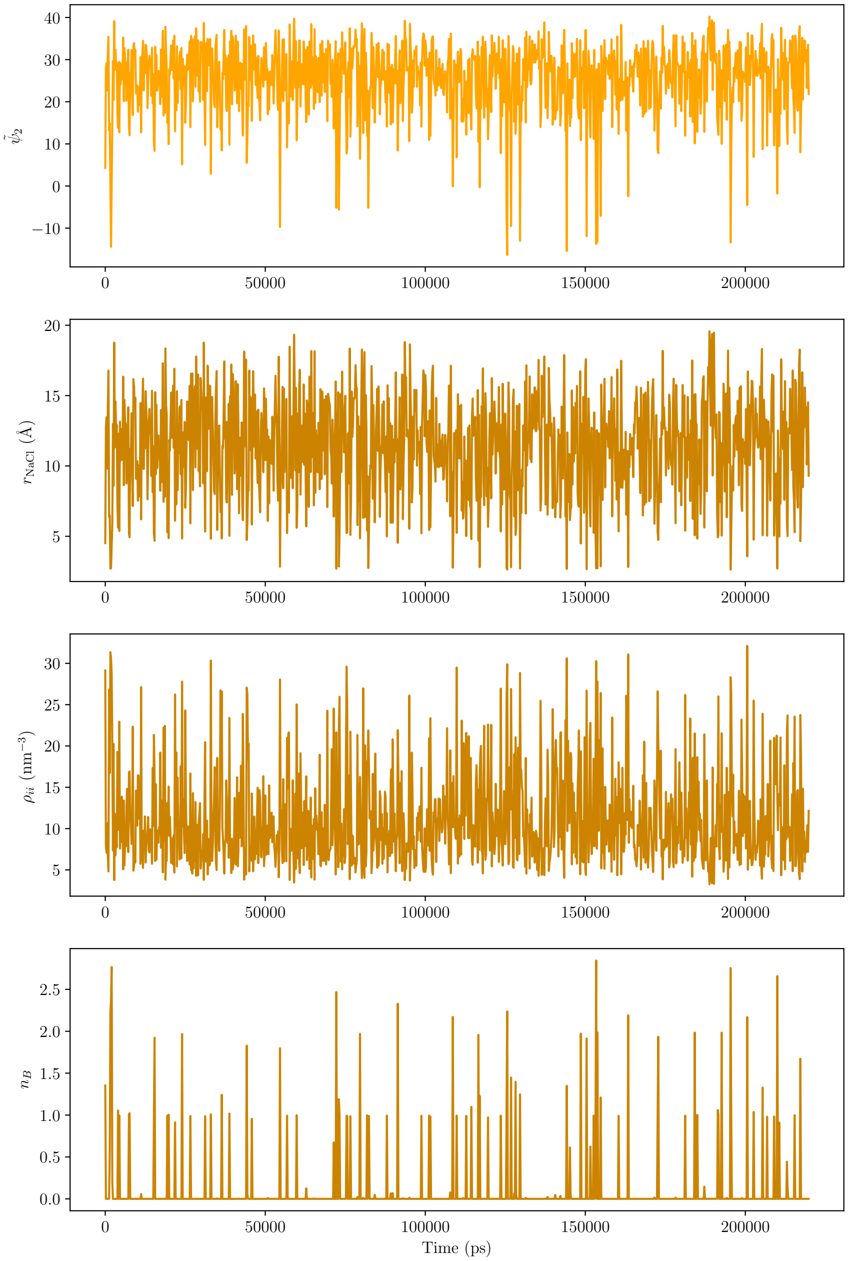

Figure 3.2: Evolution of $\tilde{\psi}_2$ over time as well as r_{+-}, $\rho_i i$ ($\sigma = r_{+-}$), and n_B

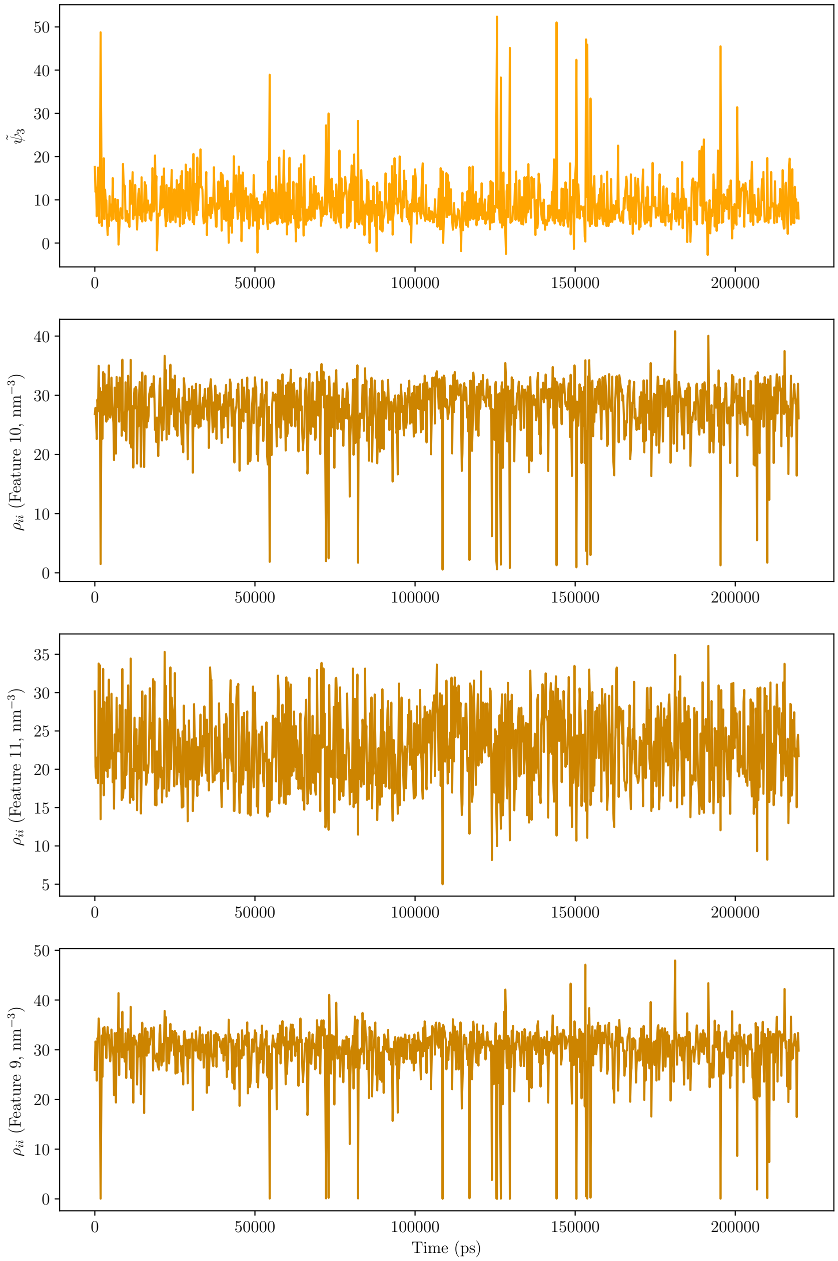

Figure 3.3: Evolution of $\tilde{\psi}_3$ over time as well as ρ_{ii} (Features 10, 11, 9 according to table 3.2)

coordinate $\tilde{\psi}_2$, the rearrangement of the solvent molecules should occur before the association of the ions. It is also interesting to note that in this reaction coordinate, the $n_+^{(1)}$ coordinate also has a slight contribution to this reaction coordinate as well, although not as important as the solvent arrangement between the ions. This finding is consistent with the work of Mullen et al., where they also found that by optimizing a set of three collective variables, the set with maximum likelihood of crossing the two dividing committor surfaces contains either the solvent variables from the water in between or the water density classes, but not from the ion coordination class. The result also indicates that ligand exchange-type reaction is less likely to play an assisted role in the association between the two ions, contrary to the previous hypotheses.

3.3.2 MSM and Kinetic Model of NaCl Ionic Association in Aqueous Solution

The discretization of the MD trajectory under the $\tilde{\psi}_i$ subspace obtained from TICA allows us to formulate the Markov State Model (MSM) for this system, where the transition probability matrix can be computed for a discrete data set. In this settings, the number of k-means clusters were 1,000, and the eigenvalues were calculated from the discretized data. Table 3.3 compares the eigenvalues obtained from MSM ($\lambda_i^\ddagger$) to the eigenvalues obtained from TICA ($\tilde{\lambda}_i$), which we generally find that $\lambda_i^\ddagger > \tilde{\lambda}_i$, where, according to the variational principle, implies that $\lambda_i^\ddagger$ is a better approximation to the actual eigenvalue of the transfer operator than $\tilde{\lambda}_i$. Figure 3.4 also shows the comparison between the implied relaxation timescales obtained from MSM (dashed line) with respect to those obtained from TICA, where we also observed the same trend as the estimation of the eigenvalues.

The better eigenvalues from MSM would also imply that the eigenfunctions from MSM

Table 3.3: Comparison between the values of the eigenvalues obtained from MSM ($\lambda_i^{\ddagger}$) and from TICA ($\tilde{\lambda}_i$) from $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system. Note that ψ_1 represents the stationary distribution $\mu(\mathbf{x})$ with no exponential decays.

Approximated Eigenfunction	**MSM Eigenvalues**	**TICA Eigenvalues**
ψ_2	0.952	0.935
ψ_3	0.907	0.857
ψ_4	0.816	0.674
ψ_5	0.784	0.602
ψ_6	0.700	0.438

should approximate the real eigenfunctions of the transfer operator better than TICA as well. However, there are crucial differences between the eigenfunctions obtained from MSM and from TICA — the MSM eigenfunctions are computed in terms of the clusters of the Voronoi diagram obtained once we had a discretized trajectory, while TICA eigenfunctions is a projection onto the collective variable space based on the continuous trajectory. Thus, despite giving a worse approximation, TICA eigenfunctions are better suited for interpretations of the physical behavior of distinct reaction coordinates in the system. On the other hand, MSM eigenfunctions have some uses as well, as the coarse-graining of the MSM transition probability matrix requires the MSM eigenfunctions to estimate the key metastable states in the system.

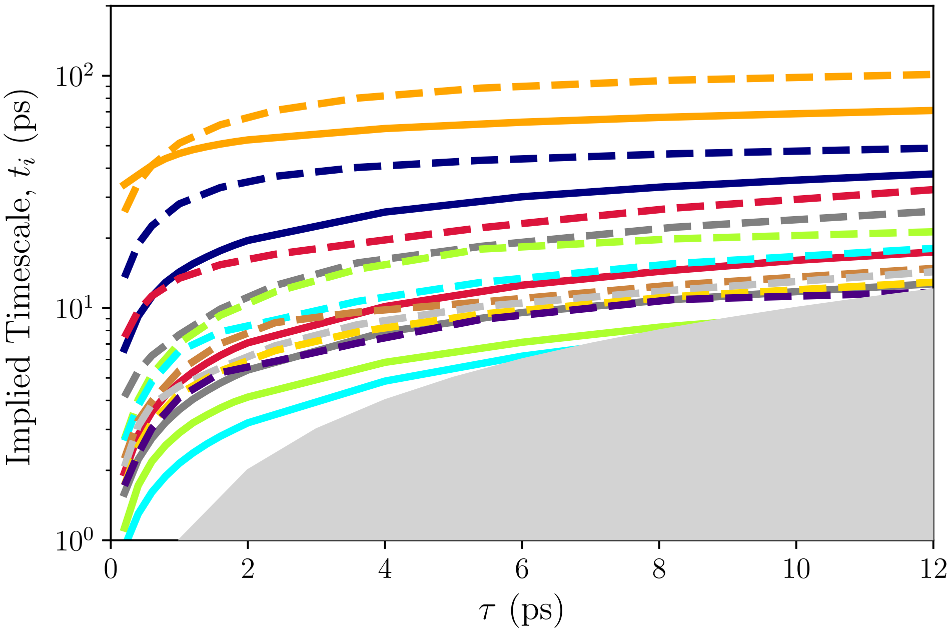

Figure 3.4: A comparison plot between the implied relaxation timescales for MSM (dashed lines) and TICA (solid lines) for the $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system

In reality, we would like to study the dynamics involving the number of metastable states far less than this number. Hence, further coarse–graining with the PCCA is desirable to reduce the number of metastable states to a number that would fit the narratives of the ionic association process of NaCl. As mentioned before in chapter 1, the dynamics of ion pairing process can be divided into three parts: CIP, SSIP, and bulk. Therefore, proposing a good mechanism involve proper identification of the reaction coordinates, as well as identification of the metastable states in this system. After the discretization was performed on the trajectory, we used the PCCA clustering to make a kinetic model of NaCl ionic association with 6 metastable states to compute the coarse transition probability matrix for this model, where the connectivity between each metastable states is identified by forming a Hidden Markov Model (HMM) between these 6 states. The 6-state kinetic model of this process is highlighted in figure 3.5, where the larger dot implies higher probability that the system would remains in that metastable state, and the numbers on or below the arrows indicate the transition probability between metastable states.

Table 3.4: Average collective variable values for each metastable state predicted by 6-state Hidden Markov Model in figure 3.5

HMM State Index	r_{+-} (Å)	n_B	$n_+^{(1)}$	ρ_{ii} (nm^{-3}, $\sigma = r_{+-}/4$)
0	2.85	1.89	4.61	4.86
1	6.24	0.52	5.83	29.59
2	15.80	0.00	5.83	29.68
3	13.31	0.00	5.83	31.10
4	9.42	0.00	5.84	30.35
5	12.29	0.00	5.83	29.92

The indication of the CIP, SSIP, and bulk states are often judged by looking at the features of the one–dimensional free energy surface of the r_{+-} variable [29], which we have computed in chapter 5. According to the results in chapter 5, the leftmost free energy minima in figure 5.1 represents the CIP configuration, which occurs where $r_{+-} \approx 2.7$ Å, while the SSIP state occurs at the second minima pass the main CIP – SSIP barrier (3.7 Å) at 5.2 Å. Anything beyond the second minima could be taken to be the bulk region where the ions do not associate. Our results from table 3.4 and figure 3.5 indicate that the dynamics spend far greater time in the bulk region, which is represented by 4 out of 6 states in figure 3.5. Moreover, the transition probability is also biased towards the bulk from the CIP and the SSIP state, indicating that the stationary density is heavily biased towards the bulk. The SSIP state, represented by a circle number 1 in figure 3.5, only connects to one bulk state and does not have any connections to three other bulk states at all, indicating that the SSIP - bulk transition mostly occurs from the leftmost boundary of the bulk state around 9 Å separation of Na^+ and Cl^- ions, and any other transitions between the SSIP state to the bulk regions with higher r_{+-} than 9 Å is not likely to happen. A small CIP state dot of figure 3.5 implies that the dynamics spends far less time in the CIP region than other regions, with the interstate transition probability biased towards the SSIP state from the CIP state.

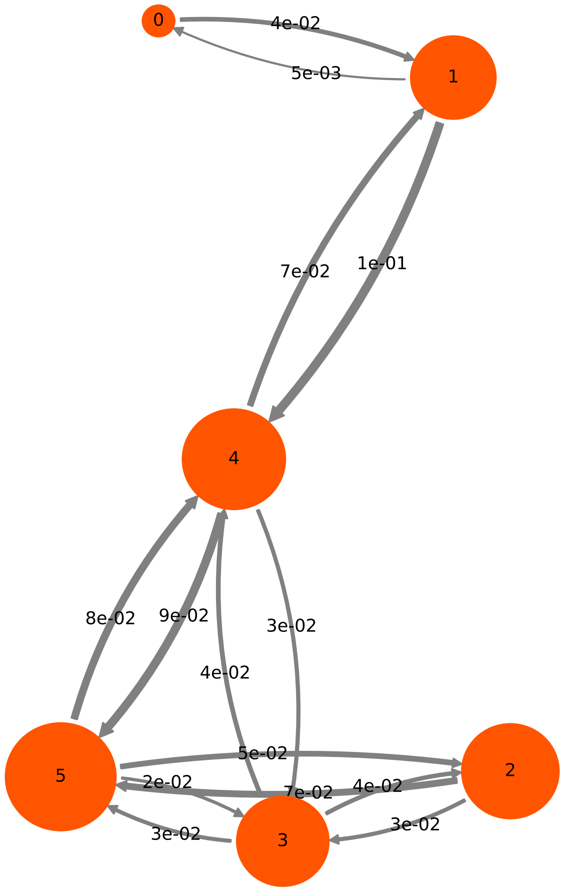

Figure 3.5: The kinetic map of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system using a 6-state Hidden Markov Model highlighting the transition probability between each state. The average position of each index point is listed in table 3.4

The information from the 6-state HMM can also be interpreted in terms of the average value of the important collective variables to elucidate important information on how each configuration is arranged in the collective variable space. In terms of the coordination number of the cation, it is interesting to note that the average cation coordination number for the CIP state is 4.61, which indicates that the CIP structure of NaCl can either have the 4-fold coordination or the 5-fold coordination, with a slight tendency toward a 5-fold coordination, while for both SSIP and bulk states, the average cation coordination number remains stable at 5.8, suggesting that both the SSIP and the bulk states prefer a 6-fold coordination number, with slight possibility of the formation of a 5-fold coordinated complex. In terms of the water density around the midpoint between the ions, ρ_{ii} value of the CIP state is very low, because as the two ions are contacted, there is simply not enough space for the water molecules to distribute around such a point in a very small volume $V_{ii} = (2\pi\sigma^2)^{3/2}$ except the water molecules of the first salvation shells of the ions. As the ions become more separated, V_{ii} becomes larger, allowing more water molecules to be distributed around a point, and we could see that ρ_{ii} values are about the same for both the SSIP and the CIP states. Another feature that distinguishes the CIP, SSIP, and the bulk state is n_B, where the average value of n_B is 1.89 for the CIP metastable states, indicating that there are likely to be 2 water molecules simultaneously coordinating both the ions at the same time, which is possible due to the close contact between the ions. For the SSIP, the number is likely to be 1 due to the further separation between the two ions, where a water molecule between the two ions need to exist in a bridge formation spanning the entire length profile of the water molecules to accommodate both ions at a distance around 5 to 6 Å. For the bulk states, the number of n_B is consistently zero due to the fact that the ions are now further apart, so no water molecules can be simultaneously coordinating with both the ions at the same time.

Although the results from figure 3.5 and table 3.4 gave a good idea of the relative probability between each metastable states as well as how they are arranged in the collective variable space, it does not give a good idea about the mechanistic point of view for this process. In order to do this, we need free energy surfaces to assist the interpretation of the metastable states generated with HMM. According to table 3.4, the metastable states in figure 3.5 still lacks the information from the region where $r_{+-} \approx 5.0$ Å, which is a region where we expect the SSIP state for this potential. Hence, we increased the metastable state approximation by HMM from 6 to 20 states to observe a better metastable state assignment. With the new 20-state HMM metastable states, we could project these points obtained from HMM onto any spaces we wish and superimpose them with the free energy landscapes. Figure 3.6 represents the projection of those 20 metastable states from HMM onto the space of $\tilde{\psi}_2$ and $\tilde{\psi}_3$.

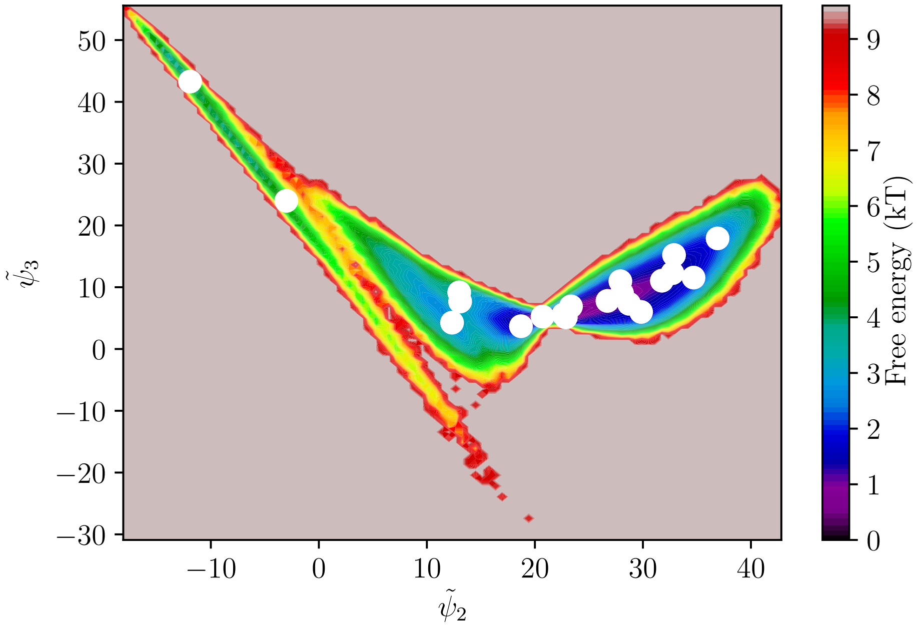

Figure 3.6: Two-dimensional free energy landscape projection onto the space of $\tilde{\psi}_2$ and $\tilde{\psi}_3$, with the 20 metastable states from HMM superimposed as white dots

The projection in figure 3.6 was done with two of the slowest reaction coordinates obtained earlier in section 3.3.1 using TICA, where the relative free energy is computed directly from the probability of a particular configuration with respect to all configurations in the space of $\tilde{\psi}_2$ and $\tilde{\psi}_3$ according to the following equation,

$$A_i(\tilde{\psi}_2, \tilde{\psi}_3) = -\beta^{-1} \ln p_i(\tilde{\psi}_2, \tilde{\psi}_3) \tag{3.6}$$

where $p_i(\tilde{\psi}_2, \tilde{\psi}_3)$ is calculated from the histogram of the bins in the $\tilde{\psi}_2$ and $\tilde{\psi}_3$ space, and $\beta^{-1} = k_BT$. The free energy projection in this space shows three distinct minima; one large minimum to the right, one small purple minimum to the middle, and one narrow green minimum to the left. The arrangement of these 3 main minima supports the view of the ionic association as being classified into CIP, SSIP, and the bulk, where the large blue minimum corresponds to the bulk, the small purple minimum corresponds to the SSIP structure, and the narrow green minimum to the left corresponds to the CIP region. However, the mechanistic interpretation of this process would rely on how well do we understand how each reaction coordinate changes from the bulk states to the SSIP, and eventually, to the CIP state. In order to make such interpretations, two more free energy projections are performed — the projection onto the r_{+-} and the n_B space to describe the behavior of $\tilde{\psi}_2$, and the projection onto the r_{+-} and the ρ_{ii} (feature 10, $\sigma = r_{+-}/3$) space to describe the behavior of $\tilde{\psi}_3$, as n_B and ρ_{ii} are the collective variables that correlate well with $\tilde{\psi}_2$ and $\tilde{\psi}_3$, respectively. Both of these free energy projections are illustrated in figures 3.7 and 3.8.

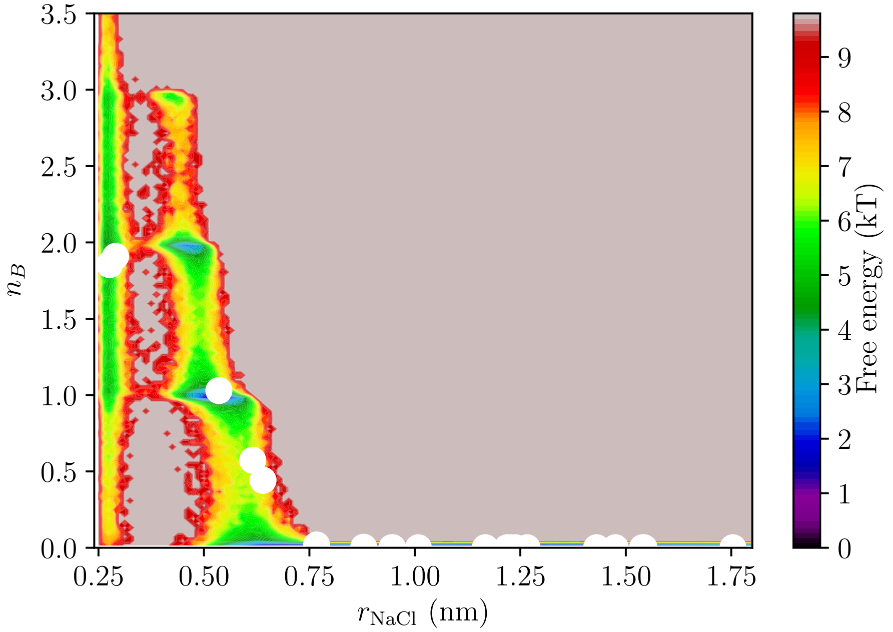

Figure 3.7: Two-dimensional free energy landscape projection onto the two collective variables space: r_{+-} and n_B, with the 20 metastable states from HMM superimposed as white dots

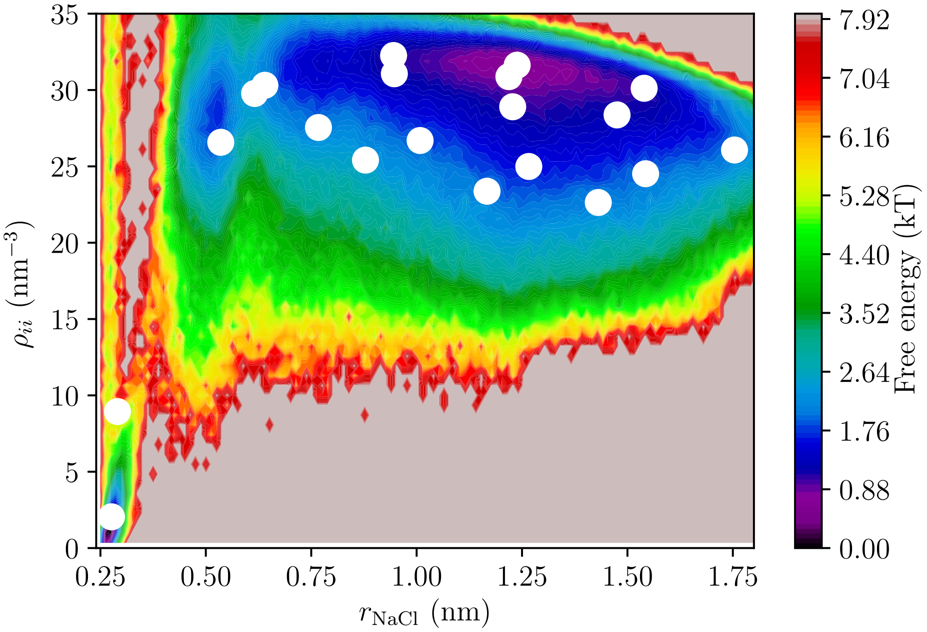

Figure 3.8: Two-dimensional free energy landscape projection onto the two collective variables space: r_{+-} and ρ_{ii} (feature 10, $\sigma = r_{+-}/3$), with the 20 metastable states from HMM superimposed as white dots

The free energy projection in the r_{+-} and the n_B space allows us to interpret the mechanistic picture of $\tilde{\psi}_2$, which is a coordinate that is influenced mostly by these two collective variables. Figure 3.7 implies that the transition from bulk into the CIP state in the $\tilde{\psi}_2$ coordinate involves the association of the ion, with one water molecule associating both of the ion in the SSIP state. The SSIP - CIP transition, according to figure 3.7, is likely driven by the water molecule in the first salvation shell of either the cation and the anion associating with both the ions first to form the second bridge, and then the ions come into a close contact. Nevertheless, this should not be the only possible pathway for this process, as another possible pathway from the SSIP to the CIP state can also undergo the association of the ions first before the water molecules forming a bridge. Thus, the process that governs the reaction coordinate $\tilde{\psi}_2$ is the driving force from the solvent bridge formation between the ions. For the reaction coordinate $\tilde{\psi}_3$; however, the SSIP - CIP transition is mostly driven by the expulsion of water molecules from the region between the ions to reduce the water density. As the two ions come into a close contact, the large size of the ions compared to water molecules would prohibit water molecules to stay between these ions, and the nature of $\tilde{\psi}_3$ affects far more solvent molecules than the coordinate $\tilde{\psi}_2$, which exerts local effect.

Using the information from figures 3.7 and 3.8 allows us to elucidate the structures of the CIP, SSIP, and all the relevant transition states. Figure 3.7 suggested that there are two possible transition pathways between the SSIP and the CIP states, labeled in figure 3.9 as TS1 and TS2. According to figure 3.9, there are two water molecules that are simultaneously coordinated with both the ions, and the overall structure of the simultaneously coordinated water molecules form a near T-shape structure together with Na^+ (purple) and Cl^- (green) ions. Through the first pathway from CIP to SSIP, one water molecule lost contact with the Cl^- ion as both ions separate further until reaching the SSIP state. The second pathway, however, involves an intermediate (labeled INT in figure 3.9) which is observed in figure 3.7. The mechanism of transition from

CIP to SSIP through the second pathway involves the distortion of the T-shaped CIP structure towards a prism-like structure of TS2 that is no longer planar. The prism distorts further in the INT structure so that both water molecules can orient in a way that the hydrogen atoms can have an interaction with the Cl^- ion, and then one water molecules loses contact with both of the ion afterwards, forming the SSIP structure. Both pathways have a very similar maximum free energy barrier, so we do not know for certain which pathway is actually preferred. In order to resolve this question, one would need a smooth two-dimensional free energy landscape in this dimension so that a minimum free energy path algorithm such as the zero-temperature string method can be applied and the free energy gradients can be reasonably calculated without the noises that arose from inadequate sampling that is natural for any unbiased MD simulations. The SSIP - bulk transition state; however, has only one probable pathway, and a HMM metastable state indicates that the structure shown in figure 3.9 should indicate the transition state. In this case, the transition state involves the long-range interaction between the Cl^- ion and the hydrogen atom of a water molecule that is slowly breaking apart as the ions separate further into the bulk.

3.4 Summary

This chapter presents the Markovian interpretation of the dynamics of the ionic association of NaCl in aqueous solutions, where slowest reaction coordinates for this process is determined using TICA. We found that there are 6 coordinates that sums up the cumulative variance for up to 0.997, while the dynamics are mostly dominated by two slowest motions with similar relaxation timescales. These 6 reaction coordinates are then subsequently used for building the MSM of this process using k-means discretization. Further coarse-graining of the MSM transition probability matrix was done through PCCA, and the kinetic profile of this reaction was then reconstructed

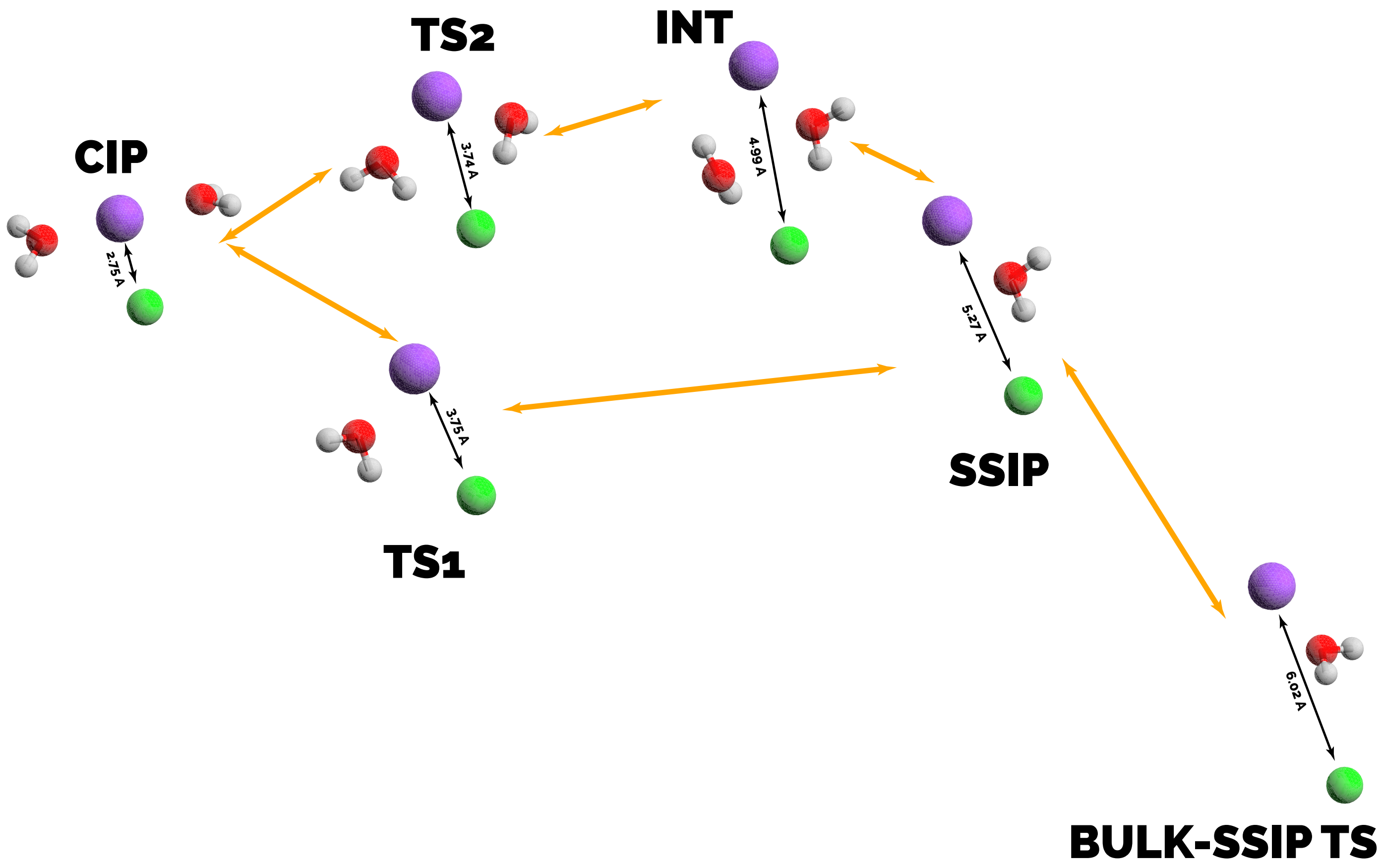

Figure 3.9: Suggested possible transition pathways between the SSIP structure and the CIP structure of the NaCl + 495 H_2O system from 20-state HMM.

with a Hidden Markov Model.

The interpretation of a 6-state HMM allows us to deduce the arrangement of the CIP, SSIP, and the bulk states in the collective variable space, as well as the probability of the dynamics staying in each specific state. We found that for NaCl, the dynamics greatly prefers the bulk, and the contacted ion pair is very rarely formed. The 20-state HMM allows a better assignment of the metastable states, and with the projected free energy landscapes in appropriate coordinates, more information of the dynamics and the mechanistic point of view for this reaction can be determined. The projection of the 20-state HMM onto the free energy landscape in two of the slowest reaction coordinates found from TICA verifies the existence of the CIP, SSIP, and the bulk states, and further projection onto the appropriate collective variables allow us to interpret the physical and mechanistic meaning of the two slowest motions in this system, where the reaction coordinate $\tilde{\psi}_2$ drives the ionic association through the local effect of bridge formation of water molecules that simultaneously coordinate with both the ions, while the reaction coordinate $\tilde{\psi}_3$ drives the ionic association through a more global effect by the expulsion of solvent molecules away from the region between the two ions as the two ions come into a closer contact.

Despite the rigor of the theory, the main drawback here is the computation of the free energy from the histograms of the molecular dynamics trajectory. The areas around the barrier can never be thoroughly sampled without a guided scheme, and the information around the transition state can deviate, causing the incorrect inference of the mechanism and the rate. The next two chapters will present a scheme that can be used to calculate multidimensional free energy landscape with less computational resources, while producing quantitatively plausible results through the application of Gaussian Process Regression (GPR).

Chapter 3, in full, is a part of the material titled “Markov State Modeling for Ion Pairing

Dynamics in Aqueous Solutions" by Pornpatcharapong, Wasut, Noé, Frank, Clementi, Cecilia, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material, and all co–authors have approved the use of the material for this dissertation.

CHAPTER 4

Multidimensional Free Energy Computation

4.1 Importance of Free Energy Computation

The free energy of a chemical system governs the macroscopic behavior along with its thermal fluctuation including both the energetic and entropic influences, and the free energy landscape may contain one or more local minima representing configurations at the metastable states, which are the locations where the system spends a significant amount of time. The change in the system from the reactant to the product states, therefore, is represented in the free energy landscape as a transition from one metastable state to another. During the course of the reaction, the system needs to pass through the free energy barrier, where the probability of barrier crossing with respect to a metastable state labeled as state 1 is defined as,

$$p(1 \to *) \propto e^{-\beta \Delta A_{1 \to *}} \tag{4.1}$$

Equation 4.1 can be further rearranged such that for any configuration $\mathbf{x} \in \Omega$, the relative free energy of the configuration $\mathbf{x}$ is related to the equilibrium probability of finding the configuration $\mathbf{x}$ in the phase space,

$$A(\mathbf{x}) = -\beta^{-1} \ln p(\mathbf{x}) \tag{4.2}$$

Hypothetically, if a simulation is performed for an infinitely large amount of time, one could easily compute $p(\mathbf{x})$ for any configurations and then the configurational free energy landscape can be mapped. Nevertheless, as governed by equation 4.1, the higher the value $\Delta A_{1\to*}$ is, the longer timescale it would take to go from state 1 across the barrier. With limited computational resources, standard molecular dynamics (MD) simulation can only go up to the millisecond timescale, while ab initio molecular dynamics (AIMD) fares far worse in timescale, only in the picosecond timescale due to a much higher cost for expensive first principles calculations. Hence, rare barrier crossing events typically are not well-sampled due to the issue of far less timescale affordable by current computational capabilities, and an attempt to compute the free energy landscape from any unbiased MD simulations would result in a noisy representation in the barrier region, hindering our insights on the dynamics of the transition state.

Being able to compute relatively noise-free free energy landscapes, thus, opens up new insights into several relevant chemical processes, as well as their underlying mechanisms, as these free energy landscapes offer us insights into the relationship between all the metastable states and the dynamics around the transition state of amy chemical reactions. Nevertheless, as a chemical system may contain up to thousands of atoms, determining the free energy as a function of thousands of phase space variables is highly complex and expensive. However, *relative free energies* can be computed by constraining the absolute free energies into the collective variable space. Consequently, the relative free energy $A(\xi_1, \xi_2, \ldots, \xi_D)$ can be written as a function of D collective variables, where D is the dimensionality of the problem which relates to the number of collective variables hypothesized to involve in the process where the reaction coordinate is a hypothetical linear combination of these D variables that correlates with the slowest motion of the system across the free energy barrier. [29, 41, 48] When written in terms of the collective variables, the free energy at $\Xi = (\xi_1^*, \xi_2^*, \ldots, \xi_D^*)$ is related to the natural logarithm marginal probability density,

$$A(\xi_1^*, \xi_2^*, \ldots, \xi_D^*) = -\beta^{-1} \ln \int \prod_{i=1}^{D} \delta\left(\xi_i(\mathbf{x}) - \xi_i^*\right) \exp\left(-\beta V(\mathbf{x})\right) d\mathbf{x} \tag{4.3}$$

where $\mathbf{x}$ is a configuration snapshot. Therefore, $A(\xi_1^*, \xi_2^*, \ldots, \xi_D^*)$ is computed by integrating over all the possible snapshots in the simulation. By taking the derivative of equation 4.3, the gradient of $A(\xi_1^*, \xi_2^*, \ldots, \xi_D^*)$ is found to be related to the statistical average of the gradient of the potential $V(\mathbf{x})$,

$$\nabla_{\mathbf{x}} A = \langle \nabla_{\mathbf{x}} V \rangle \tag{4.4}$$

Equations 4.3 and 4.4 imply that there exists free energy estimator, that is, a monomer unit that can be used to deduce the bigger picture of the free energy landscape. For equation 4.3, the free energy estimator is a local probability density $p_i(\mathbf{x})$, whereas for equation 4.4, the free energy estimator is a mean force $f_i = -\langle \nabla_{\mathbf{x}} V(\mathbf{x}_i) \rangle$. In order to get the free energy estimators for each part of the phase space, extensive sampling of that part is required to ensure statistical viability of the free energy estimators. Hence, the earliest approaches to compute free energy estimators came from constraining the simulations into several windows, where the local probability density or the mean force for each window can be computed using techniques such as Umbrella Sampling (US) [81, 82] for biased local probability densities, Thermodynamic Integration (TI) [83] for the mean forces to be integrated into the free energy either through fixing the atoms involving in the reaction coordinates using SHAKE algorithm, or using a stiff harmonic restraint to limit sampling around the center of the windowwindow, which is better known as Umbrella Integration (UI) technique. [84] Another approach to sample the CV space is through the holonomic constraints, which is used to confine the system to the collective variable hypersurface such as Blue Moon Sampling. [85, 86]

However, the major disadvantage of the aforementioned techniques is the computational cost. To minimize possible statistical uncertainties of the estimators, a long simulation in each window is necessary for adequate sampling of any part of the collective variable space we desire to explore, including the rare event regions. Therefore, these kinds of simulations are costly due to the need to minimize sampling errors. There are two main kinds of error associating with the process. First of which is the statistical error due to the number of samples, and second of which is the error due to step size (i.e. the width of the window). In order to minimize these errors, one needs to reduce the width of window to reduce the error from the large step sizes, as well as performing very long simulation in each of the window to reduce the error from inaqequate sampling. While this is doable in one dimension, the cost to obtain samples in a D-dimensional problem usually scales as $\mathcal{O}(N_{window}^D)$, where N_{window} is the number of windows usually required to obtain an acceptable result in one dimension. Moreover, as the dimensionality of the problem increases, the sampling area becomes larger, which necessitates even longer simulation per sampling area, further increasing total computational cost. Therefore, with limited computational resources, windowed simulations for multidimensional problems are inherently expensive and takes very long time even for classical MD simulations.

In the previous decade, numbers of methods were introduced to selectively apply biases along the collective variable spaces in order to force the exploration away from free energy minima. Adaptive Biasing Force (ABF) [87, 88] makes use of sampling local free energy gradients and use it as the biasing force that pushes the dynamics away from the minima. Temperature-Accelerated Molecular Dynamics (TAMD) [89, 90] manipulates the dynamics in the collective variable space to make it faster than the actual coordinate space, and let the faster dynamics of the collective variable space pull the actual dynamics away from the minima, and Metadynamics (MTD) [91, 92] periodically adds repulsive bias potential along the visited regions in the collective variable space, thereby enhancing the rare event sampling frequency. These three methods can also be used to directly approximate the free energy in the collective variable space at very long time limit. While using these methods to compute the free energy sounds attractive, the main drawback is that the free energy can only be recovered at very long time limit. The work of Raiteri et al. which performed a similar simulation reported a total simulation time of 250 ns for their alkali earth carbonate simulations to obtain acceptable two-dimensional free energy landscapes. [26] Moreover, methods such as MTD or TAMD requires parameters adjustment, where a bad set of parameters may never give a good answer to the problems. Nevertheless, these methods show potential uses as tools to quickly explore the CV space, as shown in the work by Maragliano and Vanden-Eijnden, [90] as well as Cuendet and Tuckerman. [49]

In order to efficiently compute free energy landscapes, the key challenges of the methods mentioned in the above paragraphs, such as the need to compute local probability densities or the mean forces in sampling cases or the need to let the dynamics asymtotically converge to the free energy at very long time presented by the adaptive methods, need to be addressed. Recently, machine learning has become a buzzword in a scientific and engineering community due to active fundings by large enterprises with huge computational resources driven by the need to predict underlying patterns in large amount of available data. In chemistry, it has been used in various applications; for example, approximating *ab initio* energies, deriving newer force field models, [93, 94] or structural characterization of biomolecules or advanced materials. [95–97] In free energy computation, Gaussian Process Regression (GPR) and Artificial Neural Networks (ANN) have successfully been applied for various polypeptide computational models with up to 8 dimensions using dihedral angles as the collective variables. [51, 52, 98] However, GPR has been around for a significantly longer time, and it has recently been used in a one-dimensional free energy computation from an expensive AIMD simulation as the first example beyond polypeptide [99] computational models. Nevertheless, its recent application in free energy computations implies that there is no established simulation protocols to be based on, especially for multidimensional problems in chemically reactive systems.

GPR addresses the key challenges mentioned earlier by entirely eliminating the need to obtain statistical averages for free energy estimators by assuming statistical noises associated with each estimator is part of the procedure, and the conditional expectation of the free energy based on the available information of the estimators and the explored CV can be computed from the instantaneous force free energy estimators without the need to perform simulations for very long time, provided that the CV space of interest is adequately explored. Recent work by Mones et al. [51] shows that the fastest exploration of the CV space can be achieved by using well-tempered metadynamics (WT-MTD) [100–102] simulation with relatively long Gaussian deposition rate to ensure the quasi-equilibrium condition and a high bias factor to quickly encourage the system to quickly overcome the free energy barriers, which has an effect of giving a noisy reconstruction of the free energy landscape. As a result, one can construct multidimensional free energy landscapes with much less efforts due to the elimination of the need to sample N_{window}^{D} areas in the collective variable space and the entire simulation was transformed into a single biased simulation that seeks to explore the collective variable space as quickly as possible, while a reconstruction of multidimensional free energy surface was fitted according to the maximum likelihood of the non-averaged instantaneous forces in the collective variable space as free energy estimators to obtain the best fit to the simulation data. In the next section, the theoretical aspect of GPR will be discussed in detail, as well as our proposed protocol for effective free energy reconstruction with GPR while ensuring a good agreement with traditional free energy computation methods while keeping the computational cost minimal. With our protocol, there are huge future implications for any kinds of computationally expensive problems.

4.2 Gaussian Process Regression (GPR)

4.2.1 Training Data as a Gaussian Process

Gaussian process regression (GPR) is a class of machine learning algorithms, where one aims to make a prediction of the relationship between one set of input values to another set of output values based on the provided *training data* $\mathcal{D}$, which is a set of data obtained from observations or experimental evidence. A set of the training data $\mathcal{D}$ with N training examples is written as $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is a $1 \times D$ row vector called a *feature vector* containing D features. The main objective of any machine learning algorithm is to recover an underlying relationship $f : \mathbb{R}^D \to \mathbb{R}$ that maps $\mathbf{X} = [\mathbf{x}_1\, \mathbf{x}_2\, \ldots\, \mathbf{x}_N]^\top$ to $\mathbf{y} = [y_1\, y_2\, \ldots\, y_N]^\top$. The predicted form of f would then be used to make a prediction in a *test set* with $N_\mathcal{T}$ entries $\mathcal{T} = \{(\mathbf{x}_i^*, y_i^*)\}_{i=1}^{N_\mathcal{T}}$ with an assertion that for any $\mathbf{x}_i^*, y_i^* \in \mathcal{T}$, each value of y_i^* relates to $f(\mathbf{x}_i^*)$. Among the available machine learning algorithms, GPR is a subset of a class of algorithms called *Bayesian Concept Learning* [62], where the main concept is to compute a conditional expectation of f given a set of $\mathcal{D}$ based on Bayesian inference. The key idea behind GPR lies in the model that is used to interpret $\mathcal{D}$, where each y_i is not a perfect mapping of $\mathbf{x}_i$ by f, and each mapping $f(\mathbf{x}_i)$ differs from each y_i by ϵ_i,

$$y_i = f(\mathbf{x}_i) + \epsilon_i \tag{4.5}$$

Equation 4.5; therefore, represents the imperfection in the data collection procedures or inherent uncertainties in any algorithms or experiments, because each ϵ_i can be thought of as a statistical error of the i-th mapping of f. In GPR, the deviation ϵ_i is assumed to have a normal distribution,

$$\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon_i}^2) \tag{4.6}$$

To determine the underlying f that maps $\mathbf{X}$ to $\mathbf{y}$, let us define a vector $\mathbf{f}$ where each element represents the evaluation of some arbitrary function f that we hope to represent the actual relationship we are looking for each $\mathbf{x}_i$,

$$\mathbf{f} = [f(\mathbf{x}_1)\, f(\mathbf{x}_2)\, \ldots\, f(\mathbf{x}_N)]^\top \tag{4.7}$$

If the elements of $\mathbf{f}$ in equation 4.7 have a joint Gaussian distribution, then the function f that gives rise to the aforementioned property is said to be a Gaussian Process,

$$f \sim \mathcal{GP}\left(m(\cdot), k(\cdot, \cdot)\right) \tag{4.8}$$

4.2.2 The Covariance Matrix

Equation 4.8 states that f is a Gaussian Process with mean m and covariance k. For the purpose of approximating the conditional expectation of f, the mean of this Gaussian Process is set to zero. The covariance of the Gaussian Process can be approximated by a kernel where the key property of the covariance represents the correlation between two points in the feature space $\mathbf{x}_i$ and $\mathbf{x}_j$. If the two points are identical, then the correlation should be at its maximum. If the two points are very far apart, then the correlation is expected to decay towards zero, or no correlations at all. To satisfy this property, the most common choice of the covariance approximation kernel is the squared exponential kernel $\mathbf{K}_{SE}$, which is a $ND \times ND$ matrix whose elements are defined as follow,

$$k_{ij} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \chi^2 \exp\left(-\frac{1}{2} \sum_{a=1}^{D} \frac{\left(\mathbf{x}_i^{(a)} - \mathbf{x}_j^{(a)} \right)^2}{\left(\lambda^{(a)} \right)^2} \right) \tag{4.9}$$

where χ represents the overall deviation of the function value in the region of interest, and $\lambda^{(a)}$ is called the *length scale* in the a-th dimension. For a training data with D dimensions, there are D values of the length scales, each of which controls how abrupt the correlation between the two points in the feature space should be in a specific dimension. Thus, if a dimension has a small value of $\lambda^{(a)}$, it means that the value of the function should vary more rapidly in that dimension, and if a dimension has a large value of $\lambda^{(a)}$, it means that the value of the function would vary slowly in that dimension. Equations 4.5 and 4.9 are, therefore, key ingredients to the GPR modeling of the test set, and the success of such modeling would require a good set of $2D + 1$ parameters containing D values of $\sigma_{\epsilon_i}^2$, D values of $\lambda^{(a)}$, and one value for χ. These parameters are called the *hyperparameters* in machine learning literatures, and an optimum set of hyperparameters should result in a maximum likelihood $p(\mathcal{D}|\mathbf{y}^*, \mathbf{X}^*)$ that represents the best fit to the training data. However, Stecher et al. suggested that hyperparameters optimization is a computationally expensive task, but one could make a good choice of the hyperparameters from the *a priori* knowledge of our training data. Moreover, in the application for free energy surface reconstruction, the range of good hyperparameters can vary at a large range while does not result in a significantly different reconstructed free energy landscape. [52]

4.2.3 Inference of the Conditional Expectation

Due to the fact that f satisfies the conditions of the Gaussian Process and elements of $\mathbf{f}$ from equation 4.7 have a joint Gaussian distribution, if we apply the same function f on the inputs of the test set $\mathcal{T}$, we would have a vector $\mathbf{f}^* = \left[f(\mathbf{x}_1^*)\, f(\mathbf{x}_2^*)\, \ldots\, f(\mathbf{x}_{N_\mathcal{T}}^*)\right]^\top$ whose elements also have a joint Gaussian distribution. Therefore, [103]

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \Bigg| \mathbf{X}, \mathbf{X}^* \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*) \\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*) \end{bmatrix}\right) \tag{4.10}$$

Let $\epsilon = [\epsilon_1\, \epsilon_2\, \ldots\, \epsilon_N]^\top$, the vectorized form of equation 4.5 can be written as follow,

$$\begin{aligned} \mathbf{y} &= \mathbf{f} + \epsilon \\ \mathbf{y}^* &= \mathbf{f}^* + \epsilon^* \end{aligned} \tag{4.11}$$

Since we have defined that each element of $\mathbf{f}$ and $\mathbf{f}^*$ is a Gaussian random variable, and the elements of ϵ and ϵ^* is also a Gaussian random variable as well, each element of $\mathbf{y}$ and $\mathbf{y}^*$ must also be a Gaussian random variable. Hence, similar to equation 4.10, we could also express the joint Gaussian distribution between $\mathbf{y}$ and $\mathbf{y}^*$ as,

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \Bigg| \mathbf{X}, \mathbf{X}^* \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X},\mathbf{X}) + \Sigma^2 I & \mathbf{K}(\mathbf{X},\mathbf{X}^*) \\ \mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*) + (\Sigma^*)^2 I \end{bmatrix}\right) \tag{4.12}$$

where $\Sigma^2 I$ and $(\Sigma^*)^2 I$ are the diagonal matrices of the individual variance of each point in the training data and the test data, respectively. Since both $\mathbf{y}$ and $\mathbf{y}^*$ have joint Gaussian distribution, we could compute the conditional expectation of $\mathbf{y}^*$ in the test set using the rules of the joint Gaussian distribution of two Gaussian random variables, [104]

$$\mathbb{E}\left[\mathbf{y}^*|\mathbf{y},\mathbf{X},\mathbf{X}^*\right] = \mathbf{K}(\mathbf{X}^*,\mathbf{X})\left[\mathbf{K}(\mathbf{X},\mathbf{X}) + \Sigma^2 I\right]^{-1}\mathbf{y} \tag{4.13}$$

4.2.4 Learning from Derivative Training Data

Besides being able to predict the expectation of the functions in the test set from the training data from the function itself, GPR is also able to predict the expectation of the functions in the test set from the training data of its partial derivatives with respect to each of the individual features while also including the inherent statistical errors in all dimensions. As the derivative of Gaussian Process is still a Gaussian Process, we could write a vector $\mathbf{f}' = [f'(\mathbf{x}_1)\, f'(\mathbf{x}_2)\, \ldots\, f'(\mathbf{x}_{N_\mathcal{T}})]^\top$ such that each element of $\mathbf{f}'$ is also a Gaussian random variable. Thus, it is possible to substitute $\mathbf{f}$ in equation 4.10 with $\mathbf{f}'$,

$$\begin{bmatrix}\mathbf{f}' \\ \mathbf{f}^*\end{bmatrix}\Bigg|\, \mathbf{X},\mathbf{X}^* \sim \mathcal{N}\left(0, \begin{bmatrix}\mathbf{K}_{f'f'}(\mathbf{X},\mathbf{X}) & \mathbf{K}_{f'f}(\mathbf{X},\mathbf{X}^*) \\ \mathbf{K}_{ff'}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}_{ff}(\mathbf{X}^*,\mathbf{X}^*)\end{bmatrix}\right) \tag{4.14}$$

where $\mathbf{K}_{ff}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{K}(\mathbf{x}_i,\mathbf{x}_j)$ from equation 4.9, and

$$\begin{aligned}\mathbf{K}_{f'f'}(\mathbf{x}_i,\mathbf{x}_j) &= \nabla_i\nabla_j\mathbf{K}_{ff}(\mathbf{x}_i,\mathbf{x}_j) \\ \mathbf{K}_{ff'}(\mathbf{x}_i^*,\mathbf{x}_j) &= \nabla_j\mathbf{K}_{ff}(\mathbf{x}_i^*,\mathbf{x}_j)\end{aligned} \tag{4.15}$$

Therefore, for a noisy derivative training data $\mathbf{y}' = [\mathbf{y}'_1\, \mathbf{y}'_2\, \ldots\, \mathbf{y}'_N]^\top$ and $\mathbf{y}'_i = \left[\mathbf{y}'_{i,1}\, \mathbf{y}'_{i,2}\, \ldots\, \mathbf{y}'_{i,D}\right]$, the conditional expectation of $\mathbf{y}^*$ given the training data can be computed in a similar fashion to equation 4.13,

$$\mathbb{E}\left[\mathbf{y}^*|\mathbf{y}', \mathbf{X}, \mathbf{X}^*\right] = \mathbf{K}_{ff'}(\mathbf{X}^*, \mathbf{X})\left[\mathbf{K}_{f'f'}(\mathbf{X}, \mathbf{X}) + (\Sigma')^2 I\right]^{-1} \mathbf{y}' \tag{4.16}$$

where $(\Sigma')^2$ is the associated Gaussian variance of the derivative training data.

4.3 GPR and Free Energy Computation

4.3.1 Fast Exploration of the Collective Variable Space

Since the inherent statistical noises of the free energy estimators are already included in the formulation of GPR, we do not need to perform expensive multidimensional windowed simulations to minimize the statistical noise as in any regular US or TI variants. As mentioned earlier, the free energy estimator can either be local probability density $p_i(\xi_1, \xi_2, \ldots, \xi_D)$ or the mean force $f_i(\xi_j) = -\left\langle \nabla_{\xi_j} V \right\rangle$. However, GPR already took care of the noise in its algorithm; therefore, in principle, the mean force does not need to be averaged despite an individual expression of the force along a particular collective variable would imply a relatively large amount of noise, and the unbiased instantaneous forces (UIFs), ϕ can thus be used as a noisy free energy estimator for GPR,

$$\phi(\xi_j) = -\nabla_{\xi_j} V(\xi_1, \xi_2, \ldots, \xi_D) \tag{4.17}$$

Removing the need to perform windowed simulations is the main source of the improved efficiency for using GPR to reconstruct free energy landscapes. Consequently, in order to get the free energy landscape that covers a specific area in the collective variable space, it is logical to devise a strategy that would allow us to achieve fastest sampling of the configuration space to save the simulation efforts. Cuendet and Tuckerman [49] suggested that MTD, TAMD, or adiabatic free energy dynamics can be used for guiding the exploration of the configuration space, and Mones et al. [51] suggested that using well-tempered metadynamics (WT-MTD), a variant of metadynamics, resulted in the fastest exploration of the regions of interest in their two and four-dimensional free energy landscapes of alanine peptides.

4.3.2 Well-Tempered Metadynamics as an Exploration Tool

To understand why WT-MTD allows a speedy exploration of the configuration space, it is important to first understand the underlying theoretical aspects behind regular metadynamics. Originally proposed by Laio and Parrinello in 2002, regular metadynamics adds repulsive Gaussians V_{MTD} of the following form over a long period of simulation, [91]

$$V_{MTD}(\xi_1, \xi_2, \ldots, \xi_D, t) = \sum_{t=0,\, \Delta t,\, 2\Delta t,\, \ldots,\, N_G \Delta t} h \exp\left(-\frac{1}{2} \sum_{i=1}^{D} \frac{(\xi_i - \xi_i(t))^2}{\sigma_i^2} \right) \tag{4.18}$$

where h represents the height of each Gaussian, which is constant over time, N_G is the number of deposited Gaussians, and σ_i^2 is the width of the Gaussian in the i-th dimension. If more Gaussians are deposited in a metadynamics simulation, the barrier would become more accessible due to less difference in the potential energy. According to equation 4.1, this would mean that the probability of accessing the barrier would exponentially increase, and if any metadynamics simulations are performed for a long period of time, the deposited Gaussians would converge to the negative value of the free energy.

$$V_{MTD}(\xi_1, \xi_2, \ldots, \xi_D, t \to \infty) = -A(\xi_1, \xi_2, \ldots, \xi_D) + C \tag{4.19}$$

The rate of the convergence of metadynamics depend on our choice of the parameters h and σ_i^2. While σ_i^2 governs the rate of exploration by dictating how wide of the area in its collective space domain that the Gaussians shall cover, h governs how high the Gaussians would be. If h is not chosen carefully, then it is possible that the converged V_{MTD} may become noisy and not smooth. In order to ameliorate this issue, well-tempered metadynamics (WT-MTD) was introduced by Barducci et al. to ensure that over time, h would steadily be decreasing until it asymtotically converges to zero. The bias potential form of WT-MTD takes a similar form with equation 4.18, with a slight difference.

$$V_{WT-MTD}(\xi_1, \xi_2, \ldots, \xi_D, t) = \sum_{t=0,\, \Delta t,\, 2\Delta t,\, \ldots,\, N_G \Delta t} h(t) \exp\left(-\frac{1}{2} \sum_{i=1}^{D} \frac{(\xi_i - \xi_i(t))^2}{\sigma_i^2} \right) \tag{4.20}$$

Instead of being a constant, the Gaussian height in WT-MTD simulations, $h(t)$, decays over time according to the following equation,

$$h(t) \propto \exp\left(-\frac{V_{WT-MTD}(\xi_1, \xi_2, \ldots, \xi_D, t)}{k_B \Delta T} \right) \tag{4.21}$$

The exponential term in equation 4.21 tells us that as the Gaussians are deposited, $h(t)$

would be modified by the exponential factor of the negative value of all the deposited Gaussians at that time. Hence, as $t \to \infty$, $h(t)$ would asymptotically converge to zero, which indicates the point where the deposited Gaussians are appropriately smoothened, eliminating the noises that may present from using the regular variant of metadynamics. The rate at which $h(t)$ decays over time is governed by an additional parameter, ΔT, which is called the well-tempered temperature. In WT-MTD simulations, ΔT can be thought of as a high temperature value that adds up to the regular, thermostatted simulation temperature that modifies the probability of finding a particular configuration in equation 4.1 to the following,

$$p(1 \to *) \propto e^{-\beta^* \Delta A_{t \to *}} \tag{4.22}$$

where $\beta^* = [k_B(T + \Delta T)]^{-1}$. This means when ΔT is high, the probability of sampling the rare event is significantly greater, while when $\Delta T \to 0$, the WT-MTD simulation would converge to a normal MD simulation in a canonical ensemble. A value that is commonly used in literature to indicate the strength of the WT-MTD bias is called a *bias factor*, γ, which is defined as follow, [105]

$$\gamma = \frac{T + \Delta T}{T} \tag{4.23}$$

However, the drawback of WT-MTD is that the Gaussians do not converge to the free energy as in regular metadynamics. Rather, it converges to a factor of the free energy, and the factor is determined by the choice of ΔT, or γ.

$$\begin{aligned} V_{WT-MTD}(\xi_1, \xi_2, \ldots, \xi_D, t \to \infty) &= -\frac{\Delta T}{T + \Delta T} A(\xi_1, \xi_2, \ldots, \xi_D) + C \\ &= -\frac{\gamma - 1}{\gamma} A(\xi_1, \xi_2, \ldots, \xi_D) + C \end{aligned} \tag{4.24}$$

Nevertheless, WT-MTD still retains the same advantage as regular metadynamics that the Gaussian deposition is implicitly guided by the original probability of finding a configuration (equation 4.1). Thus, the Gaussian deposition would still be aimed more toward any regions with deep free energy minima, and will still leveling out those deep wells so that the simulation would move across the free energy barriers more frequently. The exploration is further aided by our choice of γ. For the purpose of the free energy computations, the accepted range of γ for WT-MTD simulation is between 10 to 15. Using a significantly higher γ than this range would still result in a noisy reconstruction of the free energy landscape, an issue that usually presents with the regular variant of metadynamics. However, using a very high value of γ has a benefit for exploration of the configuration space of any systems with deep minima and high free energy barriers, because the exploration can be done far more quickly while the Gaussian heights would also be partially decayed. Mones et al. reported their exploration of the free energy landscapes of alanine peptides with a value of $\gamma = 33.3$ ($T = 300\,\mathrm{K}$ and $\Delta T = 10,000\,\mathrm{K}$), which is unsuitable for recovering the free energy, but hugely aids the exploration efforts. [51]

4.3.3 Computation of the Unbiased Instantaneous Forces (UIFs) from a Biased WT-MTD Simulation

We have mentioned earlier in section 4.1 that a normal, unbiased MD simulation does not sample the rare event regions of the free energy landscape very well due to low probability.

Therefore, we would usually get a good information in the free energy basins, but the free energy of the rare event regions obtained from unbiased MD simulations tend to be noisy due to poor sampling. In order to obtain enough information in the rare event regions without performing windowed simulations, guided sampling in the collective variable spaces can be done efficiently with WT-MTD simulations, where an additive biasing potential $V_{WT-MTD}(\xi_1, \xi_2, \ldots, \xi_D, t)$ is added to the Hamiltonian over time.

For any variants of metadynamics, the forms of the potentials similar to equations 4.18 and 4.20 are added to the normal potential that is the part of the original Hamiltonian of the system. Therefore, the simulation with these added potential are biased, and the biased total potential for a WT-MTD simulation would have the following form,

$$V_{\text{biased}}(\xi_1, \xi_2, \ldots, \xi_D, t) = V(\xi_1, \xi_2, \ldots, \xi_D) + V_{WT-MTD}(\xi_1, \xi_2, \ldots, \xi_D, t) \tag{4.25}$$

where $V(\xi_1, \xi_2, \ldots, \xi_D)$ is the potential that is a part of the original Hamiltonian projected onto the collective variable space, and is invariant over time. V_{WT-MTD} is defined in equation 4.20.

Equation 4.17 outlines the expression of the UIF, which is the negative of the gradient of the original potential $V(\xi_1, \xi_2, \ldots, \xi_D)$. By taking the gradients of equation 4.25, we can immediately see that

$$-\nabla_{\xi_j} V_{\text{biased}} = -\nabla_{\xi_j} V - \nabla_{\xi_j} V_{WT-MTD} \tag{4.26}$$

Let $\phi_{\text{biased}}(\xi_j) = -\nabla_{\xi_j} V_{\text{biased}}$ and $\phi_{WT-MTD}(\xi_j) = -\nabla_{\xi_j} V_{WT-MTD}$, then equation 4.26 becomes,

$$\begin{aligned}\phi_{\text{biased}}(\xi_j) &= \phi(\xi_j) + \phi_{WT-MTD}(\xi_j) \\ \phi(\xi_j) &= \phi_{\text{biased}}(\xi_j) - \phi_{WT-MTD}(\xi_j)\end{aligned} \tag{4.27}$$

Since the computation of the free energy requires $\phi(\xi_j)$ as a free energy estimator for the ξ_j dimension, we need to compute $\phi_{\text{biased}}(\xi_j)$ and $\phi_{WT-MTD}(\xi_j)$ from a biased WT-MTD simulation. $\phi_{\text{biased}}(\xi_j)$ is computed from V_{biased} by a transformation of V_{biased} from a Cartesian coordinate system usually common in frequently used MD software packages into the collective variable space using the following equation, [106]

$$\mathbf{f}_{\text{biased}} = -(\mathbf{G}_W^{-1}\mathbf{W}) \cdot \nabla_{\mathbf{X}} V(\mathbf{X}) + \beta^{-1}\nabla_{\mathbf{X}} \cdot (\mathbf{G}_W^{-1}\mathbf{W}) \tag{4.28}$$

where $\mathbf{f}_{\text{biased}} = [\phi_{\text{biased}}(\xi_1)\, \phi_{\text{biased}}(\xi_2)\, \ldots\, \phi_{\text{biased}}(\xi_D)]^\top$ is a $D \times 1$ vector of the biased instantaneous forces (BIFs), $\mathbf{G}_W^{-1} = \mathbf{W}\nabla_{\mathbf{X}}\xi$ is the generalized Gram matrix, $\nabla_{\mathbf{X}}\xi$ is the Jacobian of the collective variables, which is a $3N_{atoms} \times D$ matrix, and $\mathbf{W} = \nabla_{\mathbf{X}}\xi^\top \mu^{-1}$, where $\mu^{-1} = \delta_{ij} m_i^{-1}$ is a $3N_{atoms} \times 3N_{atoms}$ matrix whose diagonal elements represent masses of each atom in the system. However, one could immediately see that the second term of equation 4.28 requires a computation of a Hessian matrix, which can be expensive in case a collective variable involves many atoms. Nevertheless, Darve et al. has pointed out the relationship between equation 4.28 and the average of the potentials gradient. [88] Therefore, equation 4.28 can be simplified as follow,

$$\mathbf{f}_{\text{biased}} = \frac{d}{dt}(\mathbf{W}\nabla_{\mathbf{X}}\xi)^{-1}\frac{d\xi}{dt} \tag{4.29}$$

Equation 4.29 can be computed numerically provided that the biased trajectory is evenly spaced by a small enough timestep Δt to minimize the numerical error. Therefore, a vector $\mathbf{f} = [\phi(\xi_1)\,\phi(\xi_2)\,\ldots\,\phi(\xi_D)]^\top$ of the UIFs can be computed by subtracting $\mathbf{f}_{\text{biased}}$ with $\mathbf{f}_{WT-MTD} = [\phi_{WT-MTD}(\xi_1)\,\phi_{WT-MTD}(\xi_2)\,\ldots\,\phi_{WT-MTD}(\xi_D)]^\top$, which could be easily computed by simply taking the gradient of V_{WT-MTD}, already expressed in terms of the collective variables of interests.

$$\mathbf{f}_{WT-MTD} = -\nabla_\xi V_{WT-MTD} \tag{4.30}$$

With $\mathbf{f}$ determined, a training set of the collective variable points and the unbiased instantaneous forces can be built for GPR reconstruction of a D-dimensional free energy landscape.

4.4 Validation of GPR Free Energy Surfaces

The previous sections have listed all the theoretical requirements for free energy landscape reconstruction for any D-dimensional problems using GPR. However, in order to ensure that the GPR free energy landscapes are sound, the errors of the GPR free energy surfaces are needed to be computed quantitatively. In fact, as part of the characteristic joint distribution between vectors $\mathbf{f}'$ and $\mathbf{f}$ in equation 4.16, it is also possible to compute the variance of all the points in the test data. [104] However, since the computation of variance involves the $(\Sigma')^2 I$ matrix of the training data's variance, it is not a representative variance of the actual free energy landscape computed from a more traditional method with minimum statistical noises of the free energy estimators like US or TI.

In order to validate the GPR free energy landscape, GPR results can be compared with a reference surface computed using methods that are more widely accepted among the community, which implies that the best choice of a reference surface must come from windowed simulations. Although using a reference surface from windowed simulation is not usually a problem in a one-dimensional case, it becomes more problematic for the two-dimensional space and beyond due to the scaling issue, which defeats the main purpose of needing a method for free energy computation that is fast and resource-efficient.

As Mones et al. have highlighted that the one-dimensional windowed simulations are computationally cheap enough to perform, one could easily get D one-dimensional free energy landscapes for each individual dimension in our problem. A simple thought experiment would validate that it is more beneficial to perform D one-dimensional windowed simulation than to perform one D-dimensional windowed simulation if $D > 1$. Let N_{1D} be the number of windows required for a good one-dimensional free energy landscape from windowed simulations, for a D-dimensional problem, we would need to run a reference calculation that involve only DN_{1D} windows, whereas one D-dimensional reference free energy landscape computed using windowed simulation would require $\mathcal{O}(N_{1D}^D)$ windows. While this may not be a big problem for classical MD simulations, the scaling issue can be a big consideration once expensive simulation protocols such as AIMD are involved.

4.4.1 Bounding 1D Umbrella Sampling Errors with EMUS

Although umbrella sampling (US) simulations have existed for a long time since the original proposal by Torrie and Valleau, most of the work of free energy computations using US rarely published the errors of their surfaces. [18, 50] In order to get the free energy out of the windowed simulation, the most commonly used method among researchers is the Weighted-Histogram Analysis Method (WHAM). [107] When the error of the US free energy landscape constructed from WHAM is needed, a method called Monte Carlo bootstraping analysis is needed, which involves the resampling from generated fake data. [108] This causes a very long time to sample when the resampling size is large.

There have also been recent developments in the US simulation as well, and the recent work by Thiede et al. proposed a reformulation of the free energy computation from US data into an eigenproblem, which they called the Eigenvector Method for Umbrella Sampling (EMUS). [109] They also claimed that EMUS also allowed a computation of the asymptotic variance of the free energy landscape without the need to resample any fake data.

Key to this theory is the expression of the i-th normalization constant, z_i, as a vector. The definition of z_i is the following,

$$z_i = \frac{\int V_i^b(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}}{\sum_{i=1}^{N_{window}} \int V_i^b(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}} \tag{4.31}$$

where $V_i^b(\mathbf{x})$ is the i-th biased potential in each window, and $\pi(\mathbf{x})$ is the biased probability density of finding the configuration $\mathbf{x}$. The free energy in each window is, thus, related to $_i$ as,

$$A = -\beta^{-1} \ln z_i \tag{4.32}$$

It then can be show that z_i can also be written as a left eigenvector $\mathbf{z}$ of the operator $\mathbf{F}$, where

$$\begin{aligned} z_j &= \sum_{i=1}^{N_{window}} z_i F_{ij} \\ F_{ij} &= \left\langle \frac{V_j^b}{\sum_{k=1}^{N_{window}} V_j^b} \right\rangle_i \end{aligned} \tag{4.33}$$

Equation 4.33 indicates that $\mathbf{z}$ can be solved as an eigenvalue problem, and the average free energy for each window can then be computed using the knowledge of $\mathbf{z}$. Not only that, the information of $\mathbf{z}$ and $\mathbf{F}$ can also be used to estimate the asymptotic variance of EMUS as well according to the central limit theorem. Therefore, for any US simulations, one can computed the free energy along with its associated asymptotic variance using EMUS, which perfectly serves as good references for a D-dimensional free energy landscape computed using GPR.

4.4.2 Projection of a Two-dimensional Free Energy Landscape into One-dimensional Free Energy Landscape

If our D-dimensional free energy landscape is quantitatively sound, then its projection into each individual dimension needs to quantitatively agree with the results we obtained from one-dimensional US simulations. In a two-dimensional scenario, the projection of the free energy into each individual variable is relatively easy. Define a free energy as a function of two collective variables $A(\xi_1, \xi_2)$. Therefore, we can write $A(\xi_1, \xi_2)$ in terms of the marginal probability density $p(\xi_1, \xi_2)$,

$$\begin{aligned} A(\xi_1, \xi_2) &= -\beta^{-1} \ln p(\xi_1, \xi_2) \\ &= -\beta^{-1} \ln \left[p(\xi_1) p(\xi_2|\xi_1) \right] \\ &= -\beta^{-1} \ln \left[e^{(-\beta A(\xi_1, \xi_2))} p(\xi_2|\xi_1) \right] \end{aligned} \tag{4.34}$$

The conditional probability $p(\xi_2|\xi_1)$ in the canonical ensemble can be written as,

$$p(\xi_2|\xi_1) = \frac{e^{-\beta A(\xi_1,\xi_2)}}{\int e^{-\beta A(\xi_1,\xi_2)} d\xi_2} \tag{4.35}$$

By plugging the expression for $p(\xi_2|\xi_1)$ from equation 4.35 into equation 4.34, we can now express $A(\xi_1)$ by integrating out the elements in the ξ_2 dimension according to the following equation,

$$A(\xi_1) = -\beta^{-1} \ln \int e^{-\beta A(\xi_1,\xi_2)} d\xi_2 \tag{4.36}$$

We could also repeat the similar procedures in equations 4.34 and 4.35 to get $A(\xi_2)$, which is now written as,

$$A(\xi_2) = -\beta^{-1} \ln \int e^{-\beta A(\xi_1,\xi_2)} d\xi_1 \tag{4.37}$$

Hence, both $A(\xi_1)$ and $A(\xi_2)$ from equations 4.36 and 4.37 can be used to compared with the US results to determine the quantitative agreement of GPR with respect to reference US free energy surfaces.

4.5 Summary

Windowed simulation techniques for free energy computation such as Umbrella Sampling (US) or Thermodynamic Integration (TI) suffered from the performance issue. While any one-dimensional problem is relatively inexpensive, this class of free energy computation methods does not scale very well when the problem has higher dimensionality, which is known as the *curse of* *the dimensionality*. This issue discouraged many potential great simulations for chemical systems that require more than one variable to describe the thermodynamics behaviors of the process due to prohibitively expensive computational cost for larger systems such as biomolecules, or expensive simulation schemes such as *ab initio* molecular dynamics (AIMD) simulations. The high cost of this class of simulation arose from the fact that one needs to perform enough simulations in each window to obtain good enough free energy estimators with least amount of statistical errors as possible, especially around the rare event regions.

Recent advances has shown that the *curse of the dimensionality* can be mitigated by performing a fast exploration of the configurational space for a computation of biased free energy estimators, then one can unbias these estimators given that the bias information is in a mathematically convenient form, and then use a machine learning method to learn the free energy landscape from the information of the unbiased free energy estimators. In this regard, we used well-tempered metadynamics (WT-MTD) simulations to quickly sweep the configurational space and then use Gaussian Process Regression (GPR) to reconstruct smooth multidimensional free energy landscapes. The benefits of having a smooth multidimensional free energy landscape are immense, as the thermodynamical properties of the systems of interested can be projected into more than one configurational variable, allowing us to deduce the metastable states of the system as well as the minimum free energy pathway that links two metastable states of interest. The smoothness of the free energy landscape also assists algorithms such as nudged elastic band (NEB) or zero temperature string method (ZTS) through better computations of the free energy gradients.

The success of using GPR to reconstruct multidimensional free energy landscapes depends on an *a priori* information of the system. Thus, windowed simulations for each dimension are needed as references for comparing each individual one-dimensional GPR reconstruction to obtain optimum hyperparameters in each dimension through the mean of minimizing the error between the GPR-constructed free energy with respect to those obtained from windowed simulations. Performing individual one-dimensional windowed simulations are also much cheaper than performing one multidimensional windowed simulations, and the efficiency gains become much more evident as the problem contains more dimensions. As WT-MTD and GPR simulations usually cost less than windowed simulations even in one-dimension, this scheme offers a significant speedup that we sought after. After optimum hyperparameters in each dimension are obtained, they are used to construct the multidimensional free energy landscape with GPR from free energy estimators obtained from a multidimensional WT-MTD simulation. The entire scheme for multidimensional computation of free energy landscapes with WT-MTD and GPR are shown in figures 4.1 and 4.2.

One problem of using windowed simulations to compute the free energy landscape is also its high cost of error computation; thus, error analysis of multidimensional free energy landscapes from windowed simulations are rarely seen. In order to get around this issue, we proposed that a multidimensional free energy landscape be projected into individual one-dimensional free energy landscapes for each corresponding variable, where the Eigenvector Method for Umbrella Sampling (EMUS) can compute the asymptotic variance of the free energy landscape. Therefore, the error of a multidimensional free energy landscape obtained from GPR can be both qualitatively and quantitatively bound with trusted results from windowed simulations with minimal efforts without the need to perform expensive multidimensional reference calculations from windowed simulations.

Chapter 4, in full, is a part of the material titled “Efficient Two-dimensional Ion Pairing Free Energy Landscape Calculation with Gaussian Process Regression” by Pornpatcharapong, Wasut, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material

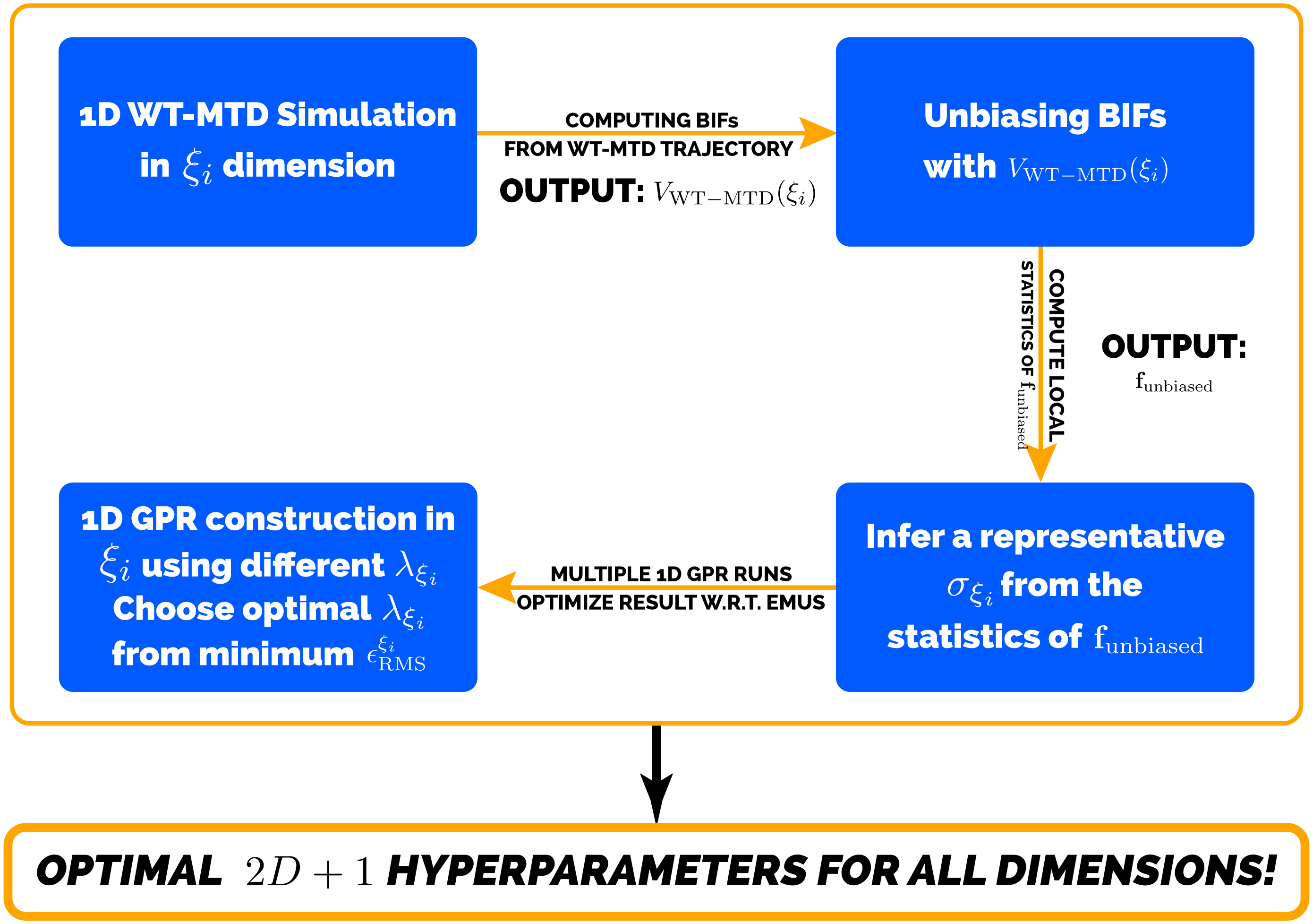
1D WT-MTD Simulation in ξ_i dimension
COMPUTING BIFs FROM WT-MTD TRAJECTORY
OUTPUT: $V_{\mathrm{WT-MTD}}(\xi_i)$
Unbiasing BIFs with $V_{\mathrm{WT-MTD}}(\xi_i)$
COMPUTE LOCAL STATISTICS OF $\mathrm{f}_{\mathrm{unbiased}}$
OUTPUT: $\mathrm{f}_{\mathrm{unbiased}}$
Infer a representative σ_{ξ_i} from the statistics of $\mathrm{f}_{\mathrm{unbiased}}$
MULTIPLE 1D GPR RUNS OPTIMIZE RESULT W.R.T. EMUS
1D GPR construction in ξ_i using different λ_{ξ_i} Choose optimal λ_{ξ_i} from minimum $\epsilon_{\mathrm{RMS}}^{\xi_i}$
OPTIMAL $2D+1$ HYPERPARAMETERS FOR ALL DIMENSIONS!

Figure 4.1: A workflow for one-dimensional GPR computations to obtain optimum hyperparameters for each dimension for use in multidimensional simulations

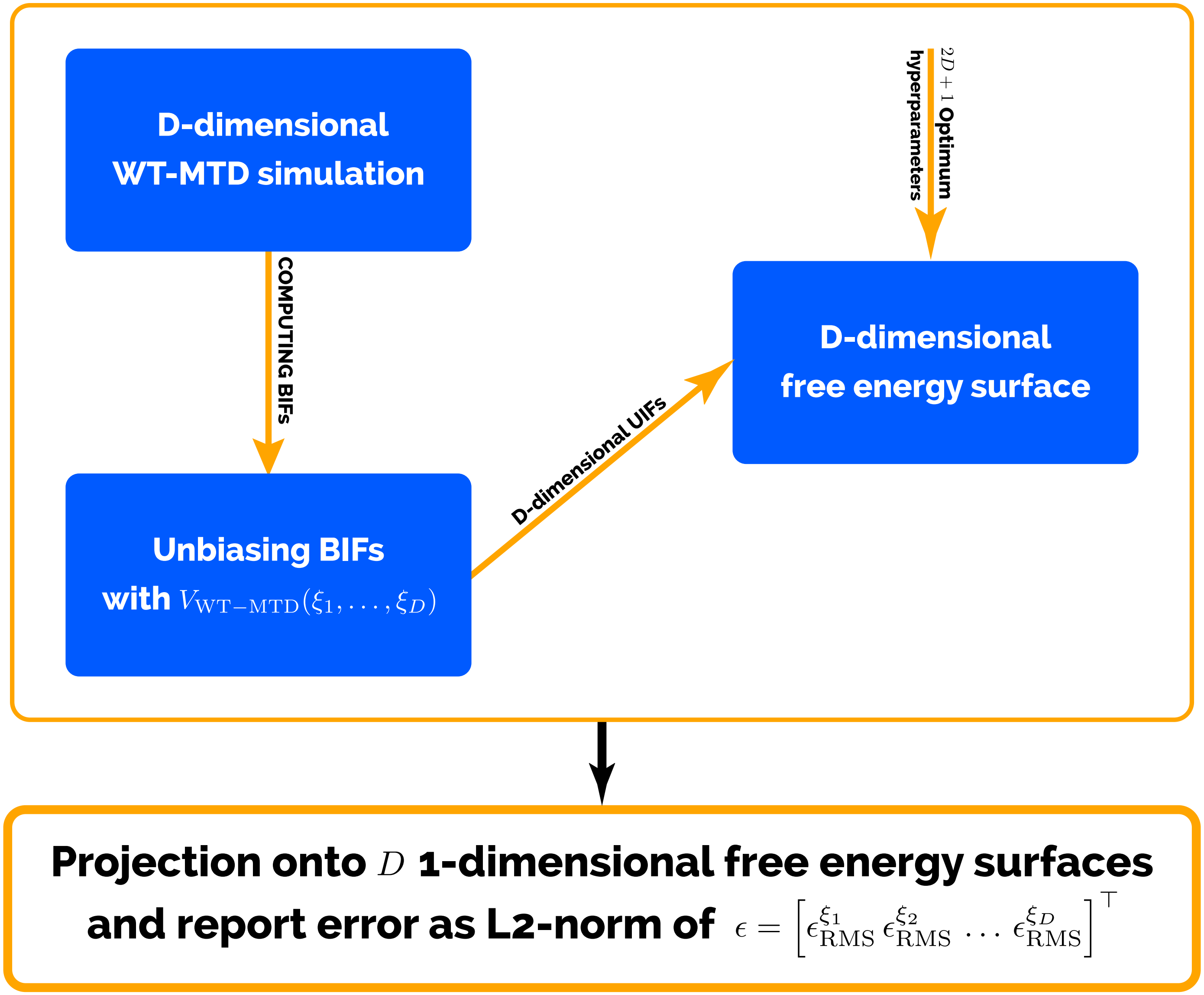
D-dimensional
WT-MTD simulation
COMPUTING BIFs
Unbiasing BIFs
with $V_{\mathrm{WT-MTD}}(\xi_1, \ldots, \xi_D)$
D-dimensional UIFs
$2D+1$ Optimum
hyperparameters
D-dimensional
free energy surface
Projection onto D 1-dimensional free energy surfaces
and report error as L2-norm of $\epsilon = \left[\epsilon^{\xi_1}_{\mathrm{RMS}}\, \epsilon^{\xi_2}_{\mathrm{RMS}} \,\ldots\, \epsilon^{\xi_D}_{\mathrm{RMS}}\right]^\top$

Figure 4.2: A workflow for multidimensional free energy computation with GPR using optimum hyperparameters obtained after performing all the works in figure 4.1

CHAPTER 5

GPR Computation of Two-dimensional Free Energy Landscapes of NaCl in Aqueous Solutions

5.1 Introduction

Chapters 2 and 3 took advantages of the Markovian properties of MD simulations to interpret many aspects of the dynamical information of a chemical reaction; however, as shown in chapter 3, in order to get a full mechanistic picture, a good free energy landscape greatly assists the analysis of the metastable states and the kinetic models obtained from MSM. However, obtaining a free energy landscape with a good quantitative result in the barrier region is no easy task, as that implies adequate sampling of the rare event regions. However, the free energy computation method used in chapter 3, where the histograms of the distribution in the variable space obtained directly from unbiased simulations are often noisy and lack adequate resolution. Moreover, the relative free energy in the barrier region computed using this method have very high statistical uncertainties, since the MD trajectory would rarely visit this region. In some cases, the histogram registers zero visit in some particular bin, causing a gap in the free energy data.

A good free energy landscape with acceptable statistical uncertainties would greatly enhance the results from TICA and MSM by helping us calculating a good rate constant across the free energy barrier, thereby giving a complete information of the mechanisms and the rate of any chemical reactions. While during earlier attempts, the free energy landscapes were often computed in the one–dimensional space due to relative simplicity, as we have demonstrated in chapter 3, it is possible for a chemical reaction to contain more than one relevant reaction coordinates to adequately describe the reaction. Moreover, each reaction coordinate may be described by more than one dominant collective variables. Consequently, an efficient and quantitively robust method to compute multidimensional free energy surfaces is highly beneficial.

This chapter presents an application of Gaussian Process Regression (GPR) for a calculation of a two–dimensional free energy landscapes for the reaction in $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system that we studied in chapter 3. Here, we propose a protocol for multidimensional free energy surfaces calculation by combining the efficiency of fast sampling with Well-Tempered Metadynamics (WT-MTD) simulation with the GPR reconstruction of the free energy landscapes from noisy free energy estimators to recreate smooth results. The quantitative agreement of the GPR–constructed free energy landscapes is then verified by the Eigenvector Method for Umbrella Sampling (EMUS), where a statistical error of the free energy landscapes in each dimension can be easily computed using one–dimensional umbrella sampling (US) simulations. The error of a two–dimensional free energy landscape can thus be compared efficiently with reference free energy landscapes in each dimension through the projection of the two–dimensional result into a one–dimensional result without the need to perform an expensive US reference simulation in two dimensions.

5.2 Simulation Details

5.2.1 General Settings

The system of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ was initialized in the same fashion as done in chapter 3, where one $\mathrm{Na^+}$ ion, one $\mathrm{Cl^-}$ ion, and 495 water molecules were initialized in a 30.0 Å cubic box. The ions were placed in the box first, then water molecules are solvated in the box using PACKMOL. [71] The equilibrium box size was determined using a 960 ps simulation in the isothermalisobaric ensemble using Nosé–Hoover Langevin piston [72, 73] to control the pressure at 1.0 atm. The system was then equilibrated under the canonical ensemble for 30.0 ns using Langevin dynamics with a damping constant of 5.0 ps^{-1} at 300 K. The simulations were performed with NAMD [74] with a timestep of 0.75 fs in order to minimize the error of numerical time derivatives computation in equation 4.29. Water molecules in this simulation are treated as rigid with TIP3P model [75], and the force field which describes the interactions between the ions and water molecules is taken from the parameters from Joung and Cheatam. [76] The electrostatic interactions in the simulation were calculated using Particle Mesh Ewald [77], and periodic boundary conditions were enforced for the entire simulation.

The biased simulations were then calculated using well-tempered metadynamics (WT-MTD) simulations in one and two dimensions under the canonical ensemble with the same basic simulation settings as above. The WT-MTD simulation employs $\gamma = 24.33$, corresponding to the well-tempered temperature parameter ΔT of 7,000 K. The collective variables for the WT-MTD simulation are r_{+-} and $n_{+}^{(1)}$, where our choice of both variables will be discussed in more details in the section below. The widths of the WT-MTD Gaussians in each dimension are $\sigma_r = 0.063$ Å, and $\sigma_n = 0.063$, respectively. The initial Gaussian height parameter, h, was set to 0.4 kcal/mol, and the WT-MTD Gaussians are deposited at every 1,000 steps, with the deposition rate gradually adjusted in case the previous batch of the simulation became too unstable. A half-harmonic potential is applied when the simulation goes out of the set collective variables boundaries, which are set at 2.2 and 7.0 Å for the r_{+-} coordinate, and at 2.5 and 6.5 for the $n_{+}^{(1)}$ coordinate. The force constants for the half–harmonic potentials were 5.0 kcal/mol $\cdot$ Å^2 in the r_{+-} coordinate, and 5.0 kcal/mol in the $n_{+}^{(1)}$ coordinate to slightly encourage more sampling in the desired region in the collective variable space.

The biased instantaneous forces (BIF) data collection were then collected from the WT-

MTD simulations at every 100th steps of the WT-MTD simulation, from where equation 4.29 was used to transform the raw information in the Cartesian coordinate system of the MD trajectories into the BIFs in the desired collective variable space, where the Jacobian of transformation was computed analytically from the definitions of the collective variables in 5.2.2. For each training set, we aim to collect from 500,000 to 1,000,000 training data of UIFs, which were computed by subtracting the BIFs with the biased forces in the collective variable space from the knowledge of deposited Gaussians in the WT-MTD simulations. The GPR free energy reconstruction process employs data clustering using k-means / k-means++ with 200 clusters for the one-dimensional problem, and 1500 clusters for the two-dimensional problem. Trajectory processing was done using MDTraj [78], while the GPR was performed using the code provided in the work of Mones et al. [51] and the Jacobian matrices were computed using a code written by our group.

Quantitative analysis of the GPR results were compared against onedimensional free energy landscapes in each individual dimension using Eigenvector Method for Umbrella Sampling (EMUS) [109], where the US simulations were performed in windows with 0.1 Å width in the r_{+-} dimensions, and 0.1 in the $n_{+}^{(1)}$ dimension. The collective variable range for each onedimensional US simulation was set at between 2.2 to 7.0 Å in the r_{+-} dimension, and between 2.5 to 6.5 in the $n_{+}^{(1)}$ dimension. For proper comparison between the GPR and the US results, any UIFs for GPR calculation beyond the collective variable ranges used in US simulations are discarded in order to ensure that the resulting free energy landscapes come from the same probability measure for both WT-MTD / GPR set of results and for the US set of results. The US simulation settings were also similar with the canonical ensemble equilibration, except that for each simulation window, the simulations were performed for a total time of 3.0 ns with harmonic biased potential present in each window. The force constant for the harmonic potential was 750.0 $kcal/mol \cdot \text{Å}^2$ in the r_{+-} dimension, and 750.0 kcal/mol in the $n_{+}^{(1)}$ dimension. The EMUS code was taken from the work of Thiede et al., where the one-dimensional free energy landscapes as well as their associated asymptotic deviations were obtained.

5.2.2 Collective Variables

As the purpose of this chapter intends to demonstrate the usage of GPR for multidimensional free energy landscapes of an actual chemical process, we opt to use a simple set of collective variables for simplicity in calculation in order to gain more expertise in the relatively unknown area. Our choice is, therefore, set for the interionic separation (r_{+-}) and the number of water molecules in the first solvation shell of the cation ($n_{+}^{(1)}$). r_{+-} was chosen to model the association between the ions, while $n_{+}^{(1)}$ was chosen to model the Eigen–Wilkins type of a ligand exchange problem. Although chapter 3 has proved that the $n_{+}^{(1)}$ collective variable does not play a key role in this process, we made this choice here because the definition of $n_{+}^{(1)}$ does not involve picking a maximum between two functions for each water molecule, making the computation of the Jacobian matrix ($\nabla_{\mathbf{X}}\xi$) for GPR more streamlined for our purposes.

The definition of r_{+-} was directly taken from equation 3.1. However, the definition of $n_{+}^{(1)}$ is different from that defined in chapter 3. Our definition of $n_{+}^{(1)}$ is the following,

$$n_{+}^{(1)} = \sum_{j=1}^{N_{wat}} \frac{1 - \left(\frac{\left\|\mathbf{r}_{\mathrm{Na-O}_j}\right\|}{r_0}\right)^{n}}{1 - \left(\frac{\left\|\mathbf{r}_{\mathrm{Na-O}_j}\right\|}{r_0}\right)^{m}} \tag{5.1}$$

where $r_0 = 3.20$ Å, n and m are positive integers where $n \ll m$. In this work, we chose $n = 8$ and $m = 36$. Our choice of n and m; however, will not result in a good representation of the step function, but this choice is chosen because it avoids the cusp in the distribution of the $n_+^{(1)}$ variable. Had our choice of n and m represents the step function that counts the coordination number exactly, the distribution of $n_+^{(1)}$ would have a cusp at every integer value. According to equations 4.15 and 4.16 in chapter 4, GPR reconstructs the free energy landscape using a Gaussian basis where every point in the space is infinitely smooth and differentiable, meaning that we can take the derivative of this function to any order. However, with the existence of the cusp in the function, the points around the cusp are no longer smooth, and the Gaussian basis functions would not be suitable in this scenario. Nevertheless, the free energy landscape with modified $n_+^{(1)}$ in this work can easily be mapped to another distribution of $n_+^{(1)}$ that better represents the step function, so that we can deduce the information in the actual coordination number space from the free energy.

5.3 Results and Discussion

5.3.1 One–dimensional Free Energy Landscapes

Figures 5.1 and 5.2 show the one–dimensional free energy surfaces computed with EMUS. For figure 5.1, the free energy was computed in the r_{+-} dimension, where two distinct minima are located at r_{+-} = 2.7 Å and r_{+-} = 5.2 Å. According to the mechanistic label proposed earlier by Fuoss and Winstein [13, 14], we could readily imply that the minimum at r_{+-} = 2.7 Å corresponds to the CIP structure, while the minimum at r_{+-} = 5.2 Å corresponds to the SSIP structure. At r_{+-} = 2.7 Å, it is impossible to have any water molecules between these ions, as this distance typically lies between the first and the second solvation shells of the ions, leaving no available space for a water molecule to insert between. Therefore, at this separation, the ion has to be in a contacted position where the electrostatic interactions between the two ions contribute to the free energy minimum. However, at $r_{+-} = 5.2$ Å, the interionic separation is large enough for a water molecule to locate between the two ions, which justifies the assignment of the SSIP label to this minimum. When looking at the free energy difference between the CIP minimum and the SSIP minimum, the EMUS result computed the free energy difference to be 0.11 kcal/mol, with the CIP state slightly lower in the relative free energy than the SSIP state. This amount of the free energy difference, however, is very close to zero, meaning that at equilibrium, the CIP and the SSIP states are almost equally likely to exist and either the CIP or the SSIP state is not significantly more stable than the other state. The barrier between the CIP and the SSIP state locates at the free energy maximum between the two local minima for both states, where the free energy barrier to cross from one state to another is in the order of 2.7 to 2.8 kcal/mol. Another feature in this free energy landscape is the free energy maximum that occurs at $r_{+-} = 6.0$ Å, which is about 0.5 kcal/mol higher than the SSIP barrier. According to our finding in figure 3.7, this barrier corresponds to the transition from SSIP to the bulk state. The fact that the SSIP - bulk free energy barrier is far less than the SSIP - CIP free energy barrier indicates that the transition from SSIP to bulk is far more likely than the transition from SSIP to CIP. Therefore, this result agrees with our findings in chapter 3 that the overall dynamics spent far longer time in the bulk than in the associated states. The asymptotic deviation computed from EMUS indicate that on average, the standard deviation for this free energy surface is 0.046 kcal/mol, where the highest deviation is 0.627 kcal/mol, which occurs at $r_{+-} \approx 2.2$ Å to the left of the CIP minimum. However, on average, the CIP region has a free energy error in the range of 0.04 kcal/mol, and the error becomes lower gradually towards the SSIP region, where the free energy error is in the range of 0.02 kcal/mol. Comparing the order of magnitude of our error to the free energy barriers, the free energy landscape in r_{+-} dimension for $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system is remarkably accurate, and the prediction of the free energy barrier has a very small error which is one to two orders of magnitude lower.

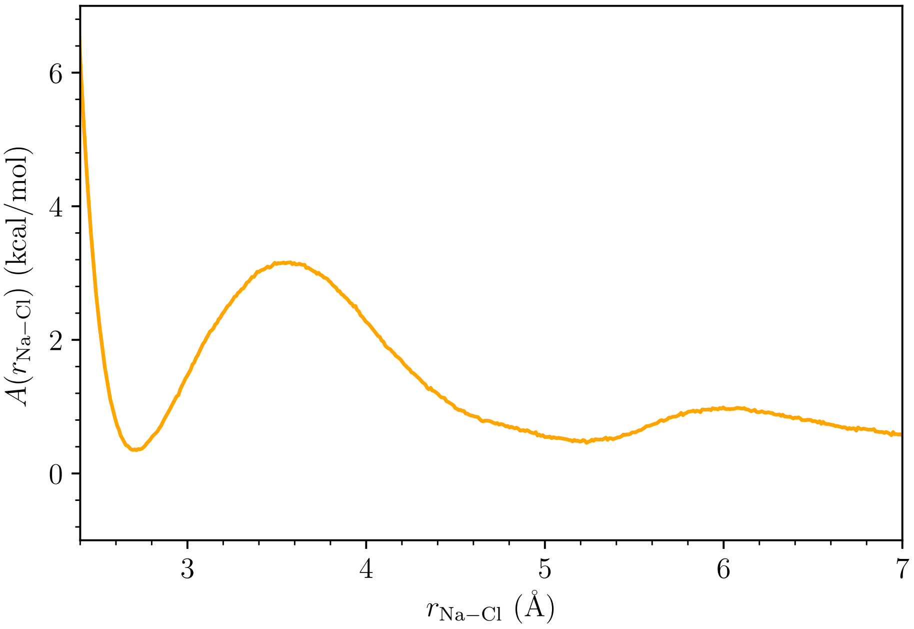

Figure 5.1: A onedimensional free energy surface of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system computed using EMUS in the r_{+-} dimension

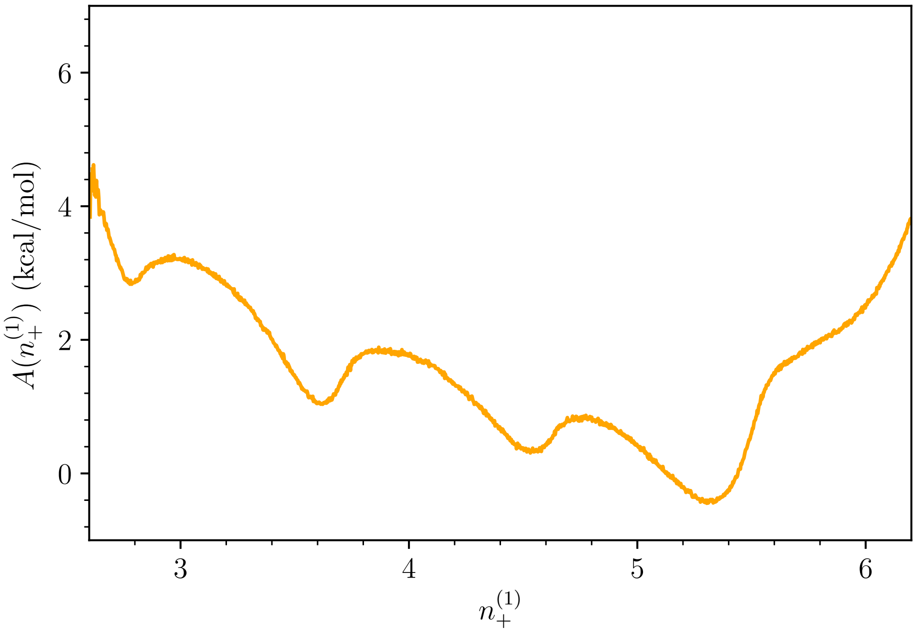

Figure 5.2: A onedimensional free energy surface of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system computed using EMUS in the $n_+^{(1)}$ dimension defined in section 5.2.2

The one-dimensional free energy landscape in the $n_+^{(1)}$ space has 4 minima representing the number of water molecules surrounding the Na^+ ion in the first salvation shell. Although the numbers shown in figure 5.2 are not perfect integer due to the definition of the $n_+^{(1)}$ in this chapter, the lowest minimum corresponds to the 6-fold coordination of the Na^+ ion, and each minimum to the left e corresponds to one successively lower number of water molecules surrounding the cation. As expected from the choice of parameters used to define this variable in section 5.2.2, the free energy surface of the $n_+^{(1)}$ contains no cusp, where a GPR reconstruction using Gaussian basis functions should have no issues approximating the shape of this result. According to our results, the free energy differences between the successive minima for the 4-fold, 5-fold, and 6-fold coordinated Na^+ ion in this system is about 0.5 kcal/mol, while the free energy barrier to take out one water molecule from the first solvation shell of the cation when it is at 4-fold, 5-fold, and 6-fold coordination state is around 1.0 kcal/mol. By comparing the free energy of removing one water molecule from the first solvation shell of the cation to the free energy of the CIP - SSIP transition from figure 5.1, it is obvious that the the lost of a water molecule from the cations first solvation shell occurs far more quickly than the association of the two ions, which is a point of view that is reinforced by our results in chapter 3 which indicated that the cations first solvation shell does not play a significant role in the slowest motion of the dynamics of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system. Towards the right side of figure 5.2, we observed an inflection point occurring at $n_+^{(1)} \approx 6.0$, corresponding to the 7-fold coordinated state of Na^+, indicating that the 7-fold coordinated state cannot be isolated as it will quickly lose a water molecule to come back to the 6-fold coordinated state. The 3-fold coordinated state is also a metastable state in figure 5.2. However, the relative free energy of the 3-fold coordinated states minimum is much higher than those for the 4-fold, 5-fold, and 6-fold states. Thus, we can put our interest in the dynamics of the SSIP - CIP transition when the cation is at the 4-fold, 5-fold, or the 6-fold coordinated states as the cation is far more likely to have those

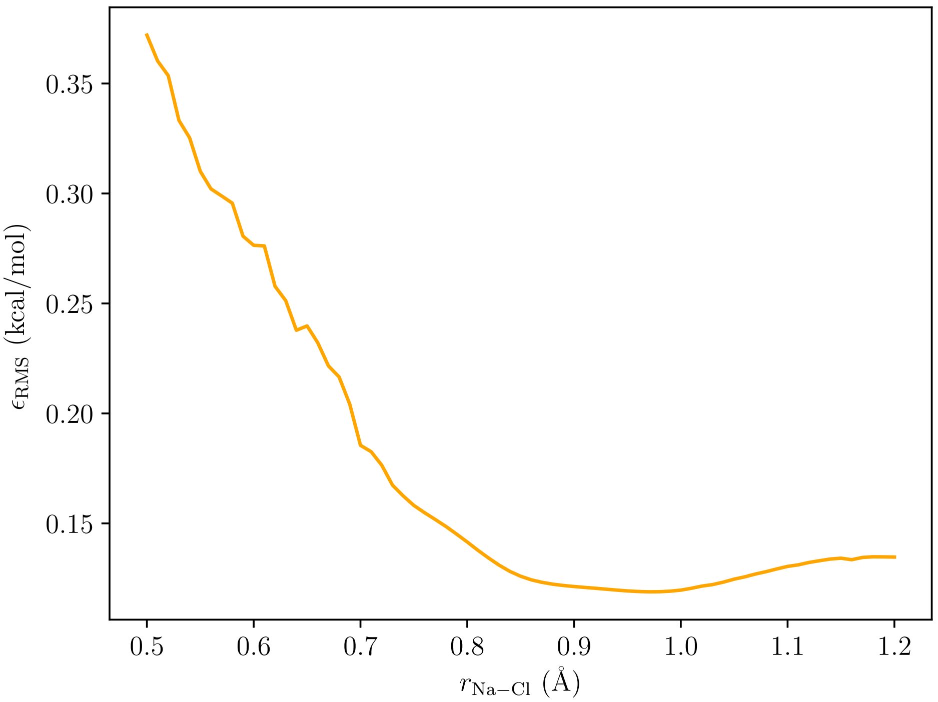

Figure 5.3: RMS error, ϵ_{RMS}, between the GPR-constructed free energy landscape in r_{+-} dimension as a function of the length scale $\lambda^{(r)}$ from a training data of roughly 900,000 UIFs with fixed σ_r^2 and χ

numbers of water molecules surrounding itself in its first solvation shell.

Free energy reconstruction from UIF data with GPR requires an optimum set of hyperparameters, which is a set of $2D + 1$ for a D-dimensional problem that creates the best fit to the EMUS results in each individual dimension. For this problem, the $2D+1$ hyperparameters contains D values of the length scale, λ, $\lambda^{(r)}$ and $\lambda^{(n)}$, D values of the variance of the UIFs, σ_r^2 and σ_n^2, and one value of the function deviation χ, where the subscripts / superscripts r and n denote the r_{+-} and the $n_+^{(1)}$ dimensions, respectively. According to Stecher et al. and Mones et al., although it is possible to perform an optimization calculation to obtain the best hyperparameters for this problem, in reality, optimization requires the GPR calculation, which has $\mathcal{O}(N_{sp} \cdot N^2)$ complexity for every step, where N_{sp} is the number of sparse points that represents the data, and $N \gg N_{sp}$

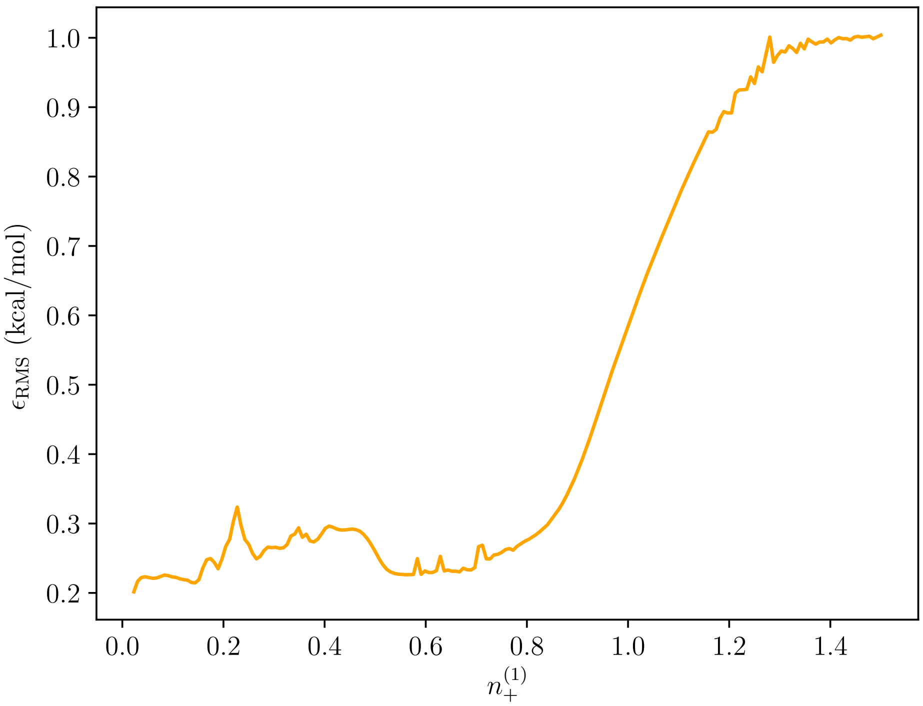

Figure 5.4: RMS error, ϵ_{RMS}, between the GPR-constructed free energy landscape in $n_+^{(1)}$ dimension as a function of the length scale $\lambda^{(n)}$ from a training data of roughly 900,000 UIFs with fixed σ_n^2 and χ

is the number of training data. This would make the overall process computationally expensive. However, given that we can know some information of the system *a priori*, it is possible that we could put constraints on other hyperparameters to find the optimum length scales, which are the most difficult hyperparameters to infer their optimum value. [51, 52] With this information, the optimum values for the hyperparameters σ_r^2 and σ_n^2 can be estimated from the statistical information of the UIFs in the calculation themselves, which is shown in figures 5.5 and 5.6. Our results suggest that σ_r can be in the order of 10.0 to 30.0 kcal/mol $\cdot$ Å in order to represent the variance of the UIFs in the r_{+-} dimension, and in the order between 10.0 to 20.0 kcal/mol in the $n_+^{(1)}$ dimension. We picked two values for each dimension for testing purposes: $\sigma_r = 25.2$ and 14.5 kcal/mol $\cdot$ Å in the r_{+-} dimension, and $\sigma_n = 19.8$ and 14.0 kcal/mol in the $n_+^{(1)}$ dimension. In this work, we chose the value of the function deviation χ to be 2.0 kcal/mol, implying that the expected value of the test set shall deviate in the range of χ. It is also possible to choose a higher value of χ; however, the higher value for χ can introduce artifacts in our reconstructed result due to lack of regularization, causing the reconstructed function to wildly fluctuate in values. With these three hyperparameters known *a priori*, the optimum length scales in each dimension can be found by minimizing the RMS error of the onedimensional GPR free energy landscapes with respect to EMUS results through the following equation,

$$\epsilon_{\mathrm{RMS}}^{\xi_i} = \left\{ \frac{1}{N_{ref}} \sum_{j=1}^{N_{ref}} \left[A_{\mathrm{GPR}}(\xi_i^{(j)}) - A_{\mathrm{EMUS}}(xi_i^{(j)}) \right]^2 \right\}^{1/2} \tag{5.2}$$

where N_{ref} is the number of reference points used in the ϵ_{RMS} calculation. The optimum length scales for both dimensions can be found in figures 5.3 and 5.4 where the ϵ_{RMS} in each dimension is at the local minimum. We found that the optimum length scale in the r_{+-} dimension, $\lambda^{(r)}$, is 0.97 Å with $\epsilon_{\mathrm{RMS}}^{(r)} = 0.11$ kcal/mol, and for the $n_+^{(1)}$ dimension, the optimum value of $\lambda^{(n)}$ is 0.56 with $\epsilon_{\mathrm{RMS}}^{(n)} = 0.22$ kcal/mol. Figure 5.4 also suggests that there is another minimum at a very small value of $\lambda^{(n)}$. However, picking the smaller value of the length scale also causes the results to be less regularized in the same fashion as taking too high value of χ. Hence, we chose the optimum value for $\lambda^{(n)}$ from the minimum with as large value of $\lambda^{(n)}$ as possible to avoid the issues of overfitting.

In order to assess the performance and viability of GPR, comparisons between the relative simulation efforts used for GPR and EMUS is also necessary besides a comparison of ϵ_{RMS}. Figures 5.7 and 5.8 shows ϵ_{RMS} of GPR-constructed free energy landscapes from different length and size of the UIF training data in each dimension with respect to EMUS results. The results suggest that the ϵ_{RMS} of the GPR-constructed free energy surfaces start to converge to the range of 0.10

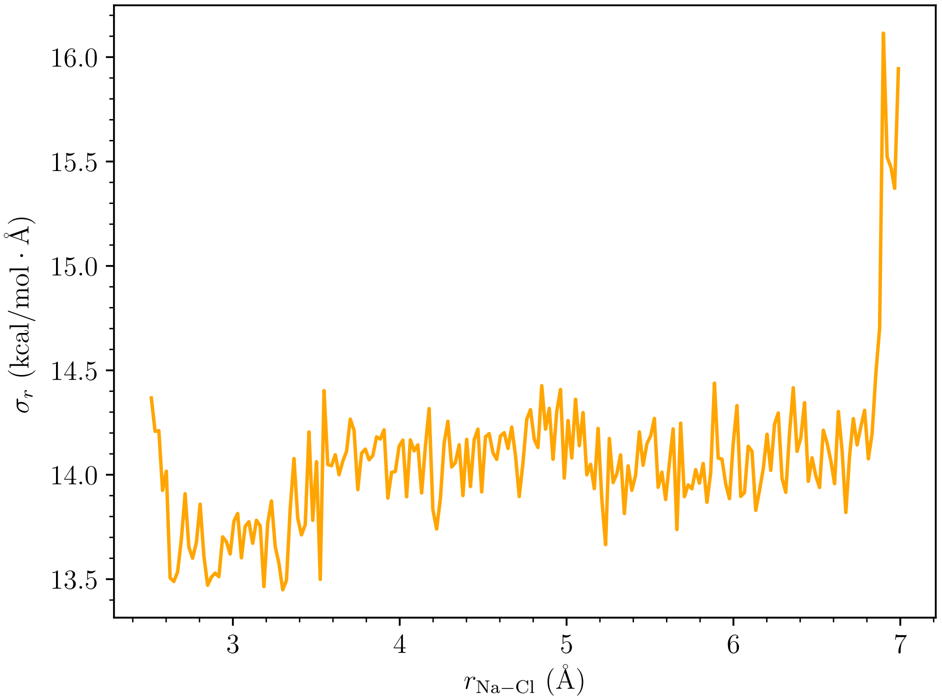

Figure 5.5: Local standard deviations of the UIFs from the onedimensional training data in the r_{+-} dimension

to 0.20 kcal/mol around the point where the size of the training data is about 100,000. As the size of the training data also roughly scales with the simulation time, it is possible to say that we need a WT-MTD simulation in the order of 10.0 ns worth of total simulation time to get a result that is close enough to the EMUS result, while the trend of the ϵ_{RMS} also converge beyond that point. Thus, running more simulations beyond that point does not greatly improve the GPR result. In fact, we observed greater ϵ_{RMS} once the total simulation time went beyond the range of 10.0 ns in the $n_+^{(1)}$ dimension. Looking closer to the results in figure 5.6 gave a clue that the pattern of the local standard deviation of the UIFs in the $n_+^{(1)}$ dimension always peak corresponding to the free energy minima positions in this dimension as well. Coupled with our findings from chapter 3 that the cations first solvation shell does not play crucial roles in the slowest reaction coordinates, the changes in this variable shall occur far more rapidly than in the r_{+-} dimension, causing high

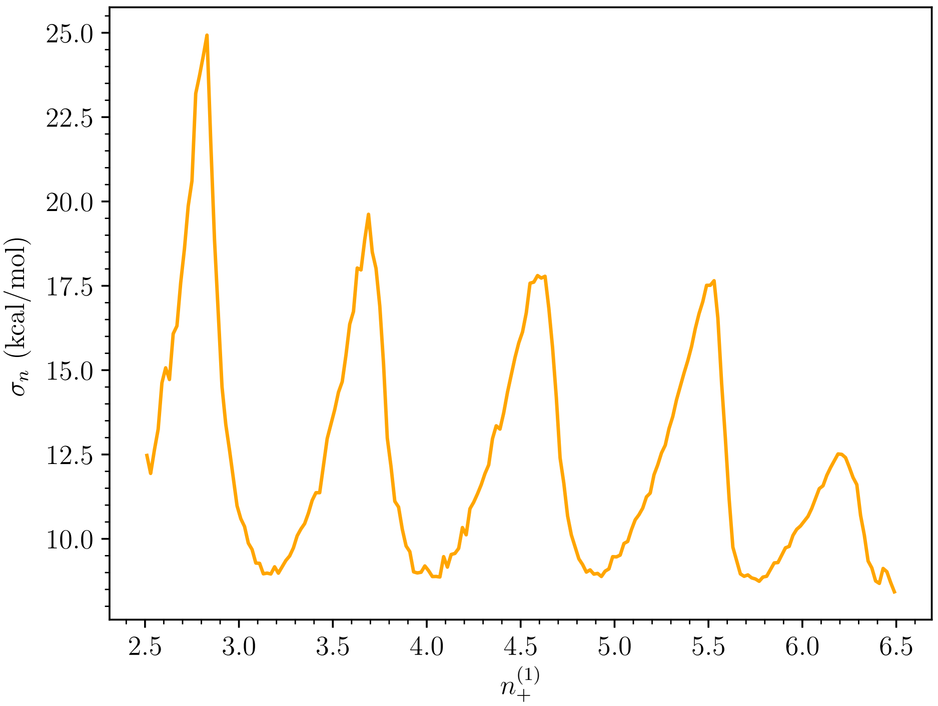

Figure 5.6: Local standard deviations of the UIFs from the onedimensional training data in the $n_+^{(1)}$ dimension

variance of the UIFs in this dimension at the free energy minima. By introducing more UIF training data in this dimension, it is also possible that the statistical uncertainties of the UIF in the $n_+^{(1)}$ dimension becomes larger, potentially causing worse GPR reconstruction with higher amount of training data. Nevertheless, we still observe the same converging trend of $\epsilon_{\text{RMS}}^{(n)}$ as in the r_{+-} dimension.

Although figures 5.7 and 5.8 suggest that GPR can rarely be as accurate as EMUS, there are a lot of benefits from using GPR as a free energy reconstruction tool. First, the total simulation time for WT-MTD / GPR simulations to achieve similar result as EMUS is far less costly, enabling us to save precious computational resources in the long run. As demonstrated above, it took only in the order of 10.0 ns total simulation time for WT-MTD / GPR for the trend of ϵ_{RMS} to converge.

To achieve this result with EMUS, we used 40 simulation windows, where the simulations were performed for 3.0 ns in each window to conform with recent US simulation protocols. [27–29] Therefore, the total simulation time for EMUS is at least one order of magnitude higher than those achieved with GPR. While this EMUS can potentially be performed faster by taking advantage of running each window in parallel [110], the total computational cost of EMUS can never be cheaper than GPR, even in one-dimensional problems. Another important benefit from GPR is a smooth reconstruction of the free energy landscape due to the usage of the squared exponential kernel in this work. Since any US variants of simulation requires the knowledge of local probability densities from the simulation data, the US results are often not smooth due to the fact that one could rarely obtain a perfect distribution from such a short simulation time per window. Having a smooth free energy landscape means that the minimum free energy paths can be more easily computed from methods such as Nudged Elastic Band (NEB) [111], or Zero-Temperature String Method (ZTS) [30, 32, 112], where the gradients of the free energy landscapes can even be computed analytically. Despite the non-optimal performance in the fast coordinate like $n_+^{(1)}$, our ϵ_{RMS} in the r_{+-} dimension is acceptable for describing the slowest transition between the CIP and SSIP states, as the average error of 0.11 kcal/mol is far less than 3.0 kcal/mol, which is the barrier height between CIP and SSIP states in this dimension, which actually also corresponds to the slowest motions of the dynamics of this system as we have found in chapter 3.

5.3.2 Two–dimensional Free Energy Landscapes

Once optimum hyperparameters in all required dimensions are determined from onedimensional simulations, it is possible for us to use these optimum hyperparameters to construct the

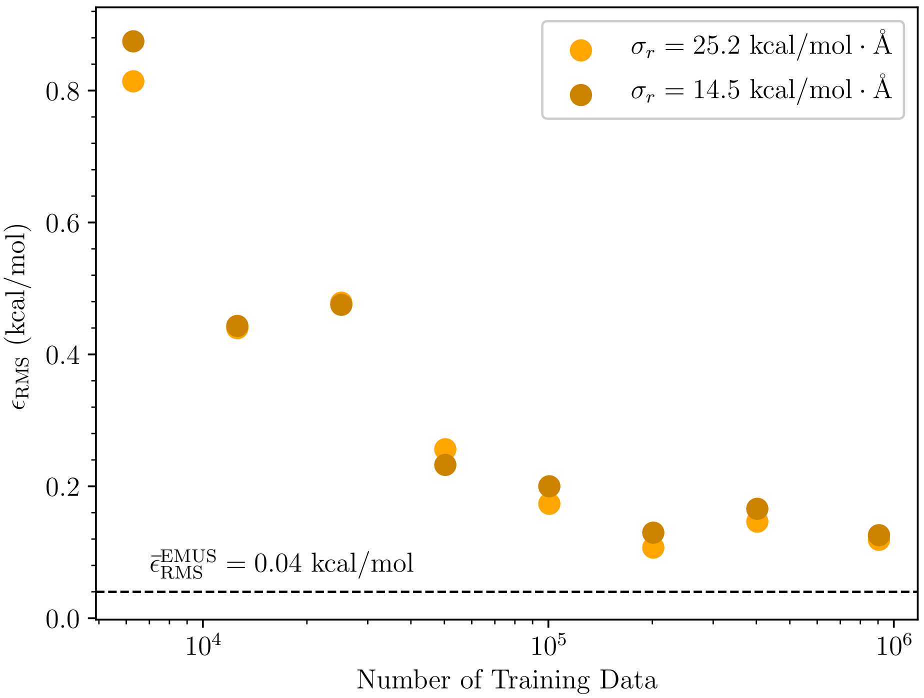

Figure 5.7: The effect of the size of UIF training data and simulation time on $\epsilon_{\text{RMS}}^{(r)}$ with respect to the EMUS result. The less amount of UIF training data also implies shorter simulation time, where the 900,000-UIF training data came from a total simulation time of 75.0 ns.

multidimensional free energy landscapes, as the projection of the marginal probability density of the multidimensional free energy landscape onto each individual dimension would correspond to the one-dimensional marginal probability density computed from one-dimensional simulations. Therefore, we carried the hyperparameters of $\lambda^{(r)} = 0.97$ Å, $\lambda^{(n)} = 0.57$, $\sigma_r = 25.2$ kcal/mol·Å, $\sigma_n = 19.8$ kcal/mol, and $\chi = 2.0$ kcal/mol over for a twodimensional GPR reconstruction of the free energy surface. The size of the training data contains roughly 650,000 pairs of UIFs in both dimensions despite coming from the same 75.0 total simulation time as our one-dimensional simulations and original 1,000,000 pairs of UIFs. 350,000 points of UIFs were discarded due to being out of the simulation range of our US simulations to ensure that our two-dimensional free energy landscape and its one-dimensional projections maintain the same probability measure as our US simulation to

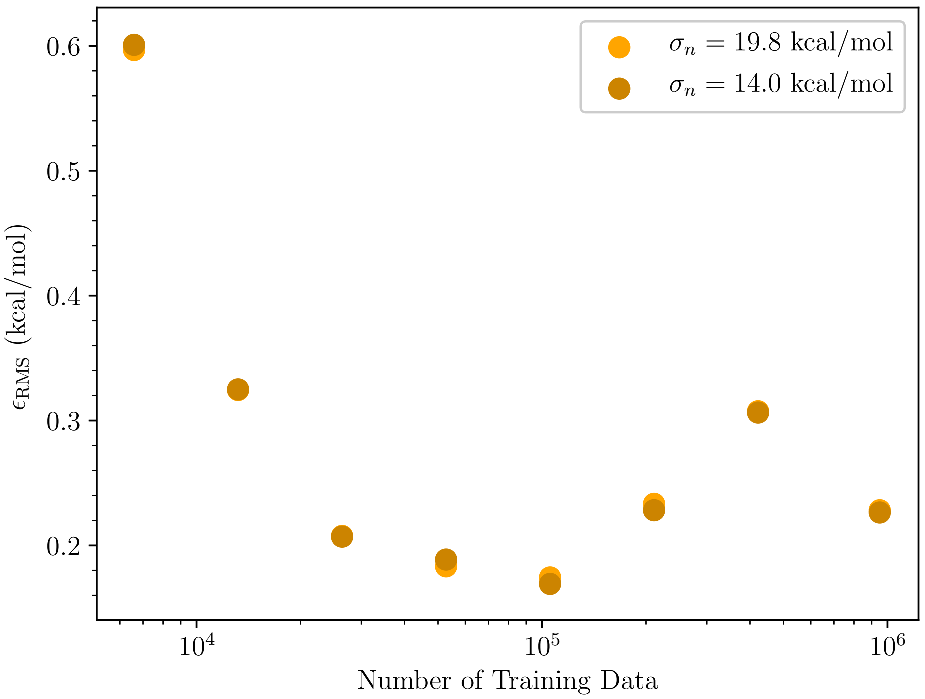

Figure 5.8: The effect of the size of UIF training data and simulation time on $\epsilon^{(n)}_{\mathrm{RMS}}$ with respect to the EMUS result. The less amount of UIF training data also implies shorter simulation time, where the 900,000-UIF training data came from a total simulation time of 75.0 ns.

enable direct comparison of results. Although the size of the training data for the twodimensional reconstruction is less than the one used for one-dimensional reconstructions, the results from the previous section suggests that the size of 650,000 is adequate. The two-dimensional free energy surface is shown in figure 5.9.

The higher dimensionality of the problem presents numerous challenges for accurately computing the free energy landscapes. The first major issue is the scaling problem. Although windowed simulations are widely used for many onedimensional problems, using windowed simulations for problems with more than one dimension unavoidably scale the computational cost. Although it may be possible to attempt such simulations using classical potentials, for expensive simulations such as AIMD, multidimensional windowed simulation is practically impossible due to a prohibitive

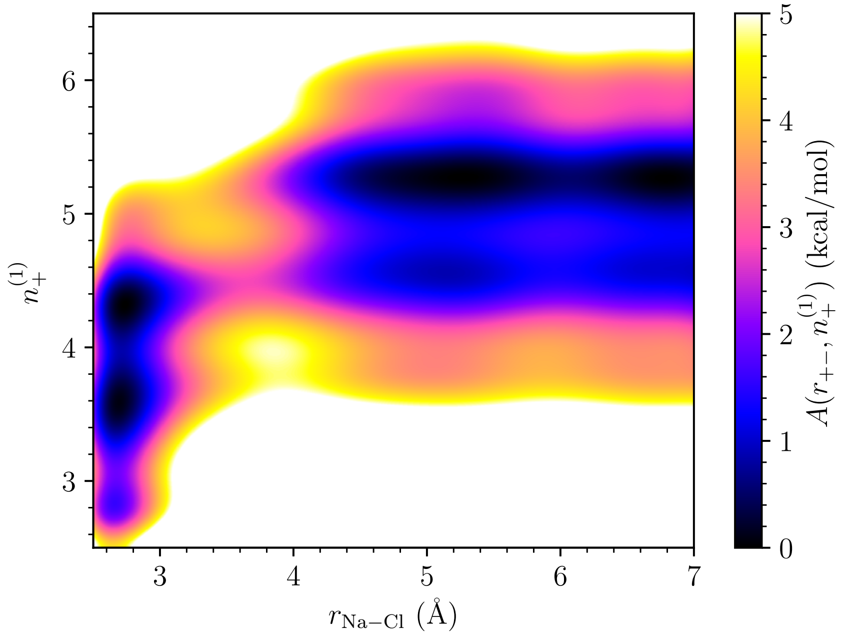

Figure 5.9: A two-dimensional free energy landscape of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system in the r_{+-} and $n_+^{(1)}$ collective variables constructed from GPR

cost of the AIMD simulation. Hence, GPR presents an attractive alternative to tackle this issue due to its robust efficiency even in onedimensional problems, as we have demonstrated in the previous section. However, there is also another important issue with using GPR to compute the multidimensional free energy landscapes, which is the consistency of our results due to the fact that it is very expensive to come up with a quantitatively accurate free energy landscape in many dimensions to use as a reference. In order to circumvent this issue, we based our reference on the onedimensional free energy landscapes in each individual dimension, which is far less expensive to compute than one multidimensional free energy surface from a multidimensional windowed simulation. For the surface in figure 5.9, we used equations 4.36 and 4.37 to project two-dimensional marginal probability densities onto each individual collective variables own space. Comparison in the r_{+-} dimension shows that our projected ϵ_{RMS} from the twodimensional surface is 0.13 kcal/-mol, which is in the same range as the number from one-dimensional GPR in this dimension. On the $n_+^{(1)}$ dimension; however, the projected ϵ_{RMS} in this dimension is 0.30 kcal/mol. This relatively high ϵ_{RMS} in the $n_+^{(1)}$ is due to the fact that we also included the rarely appeared coordinated states of the cation into the calculation as well, such as the 3-fold and the 7-fold states. Nevertheless, if we are concentrated only in the regions from 4-fold to 6-fold states, the ϵ_{RMS} of the projection now lowers to 0.11 kcal/mol, which is also an acceptable value of error comparable to ϵ_{RMS} from the projection into the r_{+-} space.

Although the $n_+^{(1)}$ does not involve in the slowest reaction coordinates for the $\mathrm{NaCl}+495\,\mathrm{H_2O}$ system, the two-dimensional free energy landscape from figure 5.9 allows us to most probable cations coordination in the SSIP and the CIP states. Here we can clearly see that the SSIP state favors the 6-fold coordination of the cation. Moreover, the SSIP state is the global minimum in this surface as well. The two minima at the region where $r_{+-} \approx 2.7$ Å indicates that there are two possible coordinated states for the cations, which are the 4-fold and 5-fold coordination states. The 5-fold CIP is lower in free energy than the 4-fold CIP state; however, both minima are very similar in free energy, meaning that the cation in the CIP structure can both likely exist in the 4-fold and the 5-fold coordinated states. The minimum free energy path from the minimum SSIP state to the minimum CIP state involves the loss of one water molecule from $\mathrm{Na^+}$ ion's first solvation shell before the two ions associate, which classifies this process as the dissociative mechanism according to the Eigen-Wilkins mechanistic label, which is illustrated by figure 5.10 where the minimum free energy path was overlaid onto the free energy landscape with ZTS. The projection of the minimum free energy path onto the reaction coordinate consisting of the collective variables r_{+-} and $n_+^{(1)}$ shows that instead of having one clear barrier between the SSIP and the CIP states as suggested by all the previous results of one-dimensional free energy landscapes in r_{+-} [18, 20, 29], the transformation from SSIP to CIP in the Eigen–Wilkins formulation involves two barrier, where a smaller barrier of about 1.3 kcal/mol corresponds to the faster process where $\mathrm{Na^+}$ ion loses one water molecule from its first solvation shell, forming an intermediate structure where the Na^+ ion has a 5-fold coordination while retaining the same SSIP interionic separation. The main barrier and the rate determining step involves the association of both ions, with the barrier height of about 2.8 kcal/mol from the minimum, as suggested in figure 5.11. Therefore, having a free energy description in more than one dimension allows us to resolve features that may once be obscured from having only one–dimensional descriptions of the process, which, in this case, is consistent with the suggestions in the work of Ballard and Dellago that the ion association process should not involve only the interionic separation and solvent molecules also play significant roles. [21]

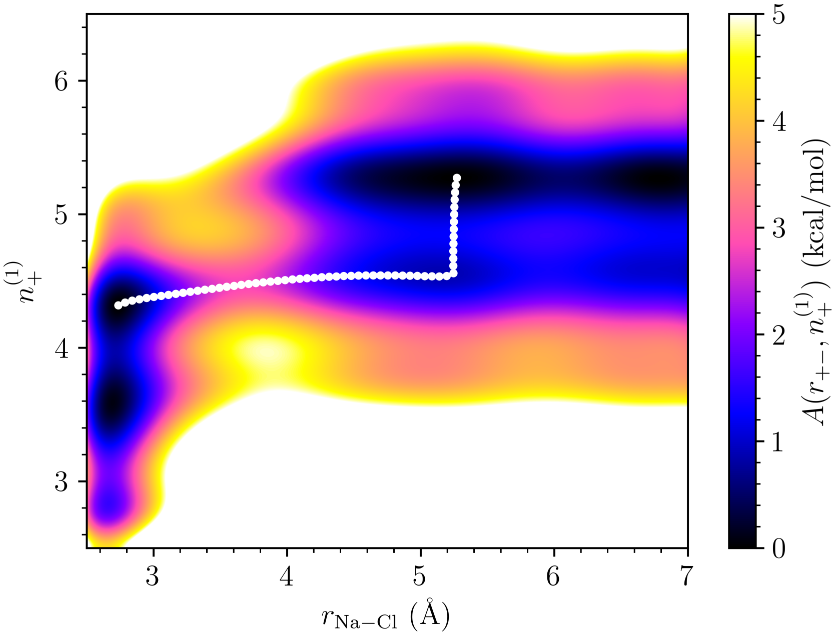

Figure 5.10: Minimum free energy path obtained using Zero Temperature String Method overlaid onto the two-dimensional free energy landscape of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system.

5.4 Summary

Being able to efficiently and accurately compute the free energy landscapes is crucial for better understandings of the reaction mechanisms and the behaviors of chemical systems at the

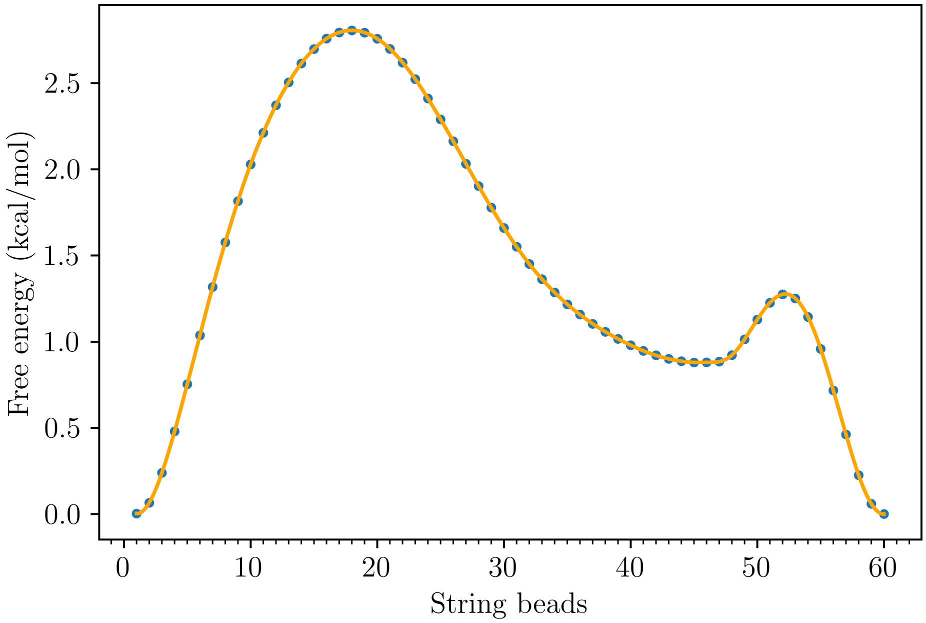

Figure 5.11: The projection of the minimum free energy path of the $\mathrm{NaCl} + 495\,\mathrm{H_2O}$ system onto a reaction coordinate involving two collective variables used in GPR free energy construction in this chapter: r_{+-} and $n_{+}^{(1)}$

transition states. Nevertheless, free energy computations is not an easy task. Although the free energy surfaces can be computed directly from any regular MD simulations by taking the probability density of a specific configuration, the information of the rare event regions is often not very well-sampled. Therefore, in order to get a better information around the rare-event regions, one needs to adequately sample the area. The most common approach for rare-event sampling employs windowed simulations, where the sampling of the rare events are done within a confined region in the configurational space. The main drawbacks of this class of approach; however, are high computational costs that also scale as $\mathcal{O}(N^D)$ for a D-dimensional problem. Therefore, a theoretical assessment of using WT-MTD in conjunction with GPR to construct multidimensional free energy landscapes provided an attractive choice for multidimensional free energy computations.

In this chapter, we have proved that using both WT-MTD / GPR is a robust scheme for free energy computation, which offers at least one order of magnitude in reduction in total computational cost even in one-dimensional problems. Although this combination cannot offer the same accuracy at the same level as windowed simulations, a significant gain in efficiency makes WT-MTD / GPR a viable candidate for free energy computations from expensive simulations such as AIMD. We also provided a guide to infer the optimum hyperparameters for each dimension by directly comparing our results with EMUS, which is an algorithm that provides asymptotic deviations for free energy landscapes computed using US simulations. Although this is the first time that WT-MTD / GPR is attempted for the real chemical process with more than one variables, our findings have been consistent with the model problems performed earlier in the works of Mones et al.

Once optimum hyperparameters for each individual dimension are determined, it is easy to use them to directly construct the desired multidimensional free energy landscapes with GPR for a good quantitative agreement by the means of projection of the marginal probability density onto each individual dimension. With the two-dimensional free energy surfaces, one can uncover more information pertaining to a chemical process of interests by resolving more features that are otherwise hidden in the one-dimensional landscapes. For this system of $\mathrm{NaCl} + 495\,\mathrm{H_2O}$, we are able to classify this reaction as dissociative according to the Eigen-Wilkins mechanistic label for inorganic ligand-exchange type of reaction.

Chapter 5, in full, is a part of the material titled “Efficient Two-dimensional Ion Pairing Free Energy Landscape Calculation with Gaussian Process Regression” by Pornpatcharapong, Wasut, and Weare, John H. The material is currently being prepared for submission. The dissertation author is the primary author of this material

REFERENCES

(1) Xie, W.; Gao, Y. *J. Phys. Chem. Lett.* **2013**, *4*, 4247–4252.

(2) Shamsi, M. H.; Kraatz, H.-B. *J. Inorg. Organomet. Polym. Mater.* **2012**, *23*, 4–23.

(3) Honig, B. H.; Hubbell, W. L. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 5412–5416.

(4) Anderson, K. M.; Esadze, A.; Manoharan, M.; Brüschweiler, R.; Gorenstein, D. G.; Iwahara, J. *J. Am. Chem. Soc.* **2013**, *135*, 3613–3619.

(5) Gaiduk, A. P.; Galli, G. *J. Phys. Chem. Lett.* **2017**, *8*, 1496–1502.

(6) Rycroft, M.; Israelsson, S.; Price, C. *J. Atmos. Sol. Terr. Phys.* **2000**, *62*, 1563–1576.

(7) Hoegh-Guldberg, O.; Mumby, P. J.; Hooten, A. J.; Steneck, R. S.; Greenfield, P.; Gomez, E.; Harvell, C. D.; Sale, P. F.; Edwards, A. J.; Caldeira, K.; Knowlton, N.; Eakin, C. M.; Iglesias-Prieto, R.; Muthiga, N.; Bradbury, R. H.; Dubi, A.; Hatziolos, M. E. *Science* **2007**, *318*, 1737–1742.

(8) Seinfeld, J. H.; Pandis, S. N., *Atmospheric chemistry and physics: from air pollution to climate change*; John Wiley & Sons: 2016.

(9) Bedrov, D.; Borodin, O.; Li, Z.; Smith, G. D. *J. Phys. Chem. B* **2010**, *114*, 4984–4997.

(10) Gélinas, B.; Natali, M.; Bibienne, T.; Li, Q. P.; Dollé, M.; Rochefort, D. *J. Phys. Chem. C* **2016**, *120*, 5315–5325.

(11) Ueno, K.; Murai, J.; Ikeda, K.; Tsuzuki, S.; Tsuchiya, M.; Tatara, R.; Mandai, T.; Umebayashi, Y.; Dokko, K.; Watanabe, M. *J. Phys. Chem. C* **2016**, *120*, 15792–15802.

(12) Chaudhari, M. I.; Nair, J. R.; Pratt, L. R.; Soto, F. A.; Balbuena, P. B.; Rempe, S. B. *J. Chem. Theory Comput.* **2016**, *12*, 5709–5718.

(13) Sadek, H.; Fuoss, R. M. *J. Am. Chem. Soc.* **1954**, *76*, 5905–5909.

(14) Winstein, S.; Clippinger, E.; Fainberg, A.; Heck, R.; Robinson, G. *J. Am. Chem. Soc.* **1956**, *78*, 328–335.

(15) Belch, A. C.; Berkowitz, M.; McCammon, J. *J. Am. Chem. Soc.* **1986**, *108*, 1755–1761.

(16) Zhang, Z.; Duan, Z. *Chem. Phys.* **2004**, *297*, 221–233.

(17) Hess, B.; Holm, C.; van der Vegt, N. *Phys. Rev. Lett.* **2006**, *96*, 147801.

(18) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 6782–6791.

(19) Cauët, E.; Bogatko, S. A.; Bylaska, E. J.; Weare, J. H. *Inorg. Chem.* **2012**, *51*, 10856–10869.

(20) Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.

(21) Ballard, A. J.; Dellago, C. *J. Phys. Chem. B* **2012**, *116*, 13490.

(22) Richens, D. T. *Chem. Rev.* **2005**, *105*, 1961–2002.

(23) Eigen, M. *Pure Appl. Chem.* **1963**, *6*, 97–116.

(24) Eigen, M.; Wilkins, R. G. In *Mechanisms of Inorganic Reactions*; ACS Publications: 1965; Chapter 3, pp 55–80.

(25) Ikeda, T.; Boero, M.; Terakura, K. *J. Chem. Phys.* **2007**, *127*, 074503.

(26) Raiteri, P.; Demichelis, R.; Gale, J. D. *J. Phys. Chem. C* **2015**, *119*, 24447–24458.

(27) Roy, S.; Baer, M. D.; Mundy, C. J.; Schenter, G. K. *J. Phys. Chem. C* **2016**, *120*, 7597–7605.

(28) Roy, S.; Baer, M. D.; Mundy, C. J.; Schenter, G. K. *J. Chem. Theory Comput.* **2017**, *13*, 3470–3477.

(29) Mullen, R. G.; Shea, J.-E.; Peters, B. *J. Chem. Theory Comput.* **2014**, *10*, 659–667.

(30) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301.

(31) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Phys. Chem. B* **2005**, *109*, PMID: 16851751, 6688–6693.

(32) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Chem. Phys.* **2007**, *126*, 164103.

(33) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415.

(34) Pan, A. C.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107.

(35) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99–105.

(36) Prinz, J.-H.; Keller, B.; Noé, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912–16927.

(37) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(38) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, PMID: 11972010, 291–318.

(39) Hummer, G. *J. Chem. Phys.* **2004**, *120*, 516–523.

(40) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732–6737.

(41) Rohrdanz, M. A.; Zheng, W.; Clementi, C. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295–316.

(42) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526.

(43) Malmstrom, R. D.; Lee, C. T.; Wart, A. T. V.; Amaro, R. E. *J. Chem. Theory Comput.* **2014**, *10*, 2648–2657.

(44) Klippenstein, S. J.; Pande, V. S.; Truhlar, D. G. *J. Am. Chem. Soc.* **2014**, *136*, 528–546.

(45) Shukla, D.; Hernández, C. X.; Weber, J. K.; Pande, V. S. *Acc. Chem. Res.* **2015**, *48*, 414–422.

(46) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 07B604.

(47) Wu, H.; Nüske, F.; Paul, F.; Klus, S.; Koltai, P.; Noé, F. *J. Chem. Phys.* **2017**, *146*, 154104.

(48) Noé, F.; Clementi, C. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141–147.

(49) Cuendet, M. A.; Tuckerman, M. E. *J. Chem. Theory Comput.* **2014**, *10*, 2975–2986.

(50) Wojtas-Niziurski, W.; Meng, Y.; Roux, B.; Bernèche, S. *J. Chem. Theory Comput.* **2013**, *9*, 1885–1895.

(51) Mones, L.; Bernstein, N.; Csányi, G. *J. Chem. Theory Comput.* **2016**, *12*, 5100–5110.

(52) Stecher, T.; Bernstein, N.; Csányi, G. *J. Chem. Theory Comput.* **2014**, *10*, 4079–4097.

(53) Zwanzig, R. *J. Stat. Phys.* **1983**, *30*, 255–262.

(54) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.

(55) Weber, J. K.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3405–3411.

(56) Vitalini, F.; Noé, F.; Keller, B. G. *J. Chem. Theory Comput.* **2015**, *11*, 3992–4004.

(57) Husic, B. E.; Pande, V. S. *J. Chem. Theory Comput.* **2017**, *13*, 963–967.

(58) Mallat, S.; Zhang, Z. *IEEE Trans. Signal Process.* **1993**, *41*, 3397–3415.

(59) Noé, F.; Clementi, C. *J. Chem. Theory Comput.* **2015**, *11*, 5002–5011.

(60) Noé, F.; Banisch, R.; Clementi, C. *J. Chem. Theory Comput.* **2016**, *12*, 5620–5630.

(61) Flach, P., *Machine Learning*; Cambridge University Press: 2009.

(62) Murphy, K., *Machine Learning: A Probabilistic Perspective*; Adaptive computation and machine learning; MIT Press: 2012.

(63) Noé, F.; Nüske, F. *Multiscale Model. Simul.* **2013**, *11*, 635–655.

(64) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

(65) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comput. Phys.* **1999**, *151*, 146–168.

(66) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39–59.

(67) Nüske, F.; Keller, B.; Pérez-Hernández, G.; Mey, A. S.; Noé, F. *J. Chem. Theory Comput.* **2014**, *10*, 1739–1752.

(68) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.

(69) McGibbon, R. T.; Husic, B. E.; Pande, V. S. *J. Chem. Phys.* **2017**, *146*, 044109.

(70) Husic, B. E.; Pande, V. S. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

(71) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(72) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(73) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(74) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(75) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(76) Joung, I. S.; Cheatham III, T. E. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(77) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(78) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. *Biophys. J.* **2015**, *109*, 1528–1532.

(79) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542.

(80) Yonetani, Y. *J. Chem. Phys.* **2015**, *143*, 044506.

(81) Torrie, G.; Valleau, J. *J. Comput. Phys.* **1977**, *23*, 187–199.

(82) Kästner, J. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.

(83) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.

(84) Kästner, J.; Thiel, W. *J. Chem. Phys.* **2005**, *123*, 144104.

(85) Carter, E.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472–477.

(86) Ciccotti, G.; Kapral, R.; Vanden-Eijnden, E. *ChemPhysChem* **2005**, *6*, 1809–1814.

(87) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–9183.

(88) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. *J. Chem. Phys.* **2008**, *128*, 144120.

(89) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2006**, *426*, 168–175.

(90) Maragliano, L.; Vanden-Eijnden, E. *J. Chem. Phys.* **2008**, *128*, 184110.

(91) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(92) Barducci, A.; Bonomi, M.; Parrinello, M. *WIREs Comput. Mol. Sci.* **2011**, *1*, 826–843.

(93) Kolb, B.; Luo, X.; Zhou, X.; Jiang, B.; Guo, H. *J. Phys. Chem. Lett.* **2017**, *8*, 666–672.

(94) Glielmo, A.; Sollich, P.; De Vita, A. *Phys. Rev. B* **2017**, *95*, 214302.

(95) Long, A. W.; Phillips, C. L.; Jankowksi, E.; Ferguson, A. L. *Soft Matter* **2016**, *12*, 7119–7135.

(96) Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. *Nat. Commun.* **2016**, *7*, DOI: `10.1038/ncomms11241`.

(97) Khuntawee, W.; Kunaseth, M.; Rungnim, C.; Intagorn, S.; Wolschann, P.; Kungwan, N.; Rungrotmongkol, T.; Hannongbua, S. *J. Chem. Inf. Model.* **2017**, *57*, 778–786.

(98) Galvelis, R.; Sugita, Y. *J. Chem. Theory Comput.* **2017**, *13*, 2489–2500.

(99) Stecher, T.; Reuter, K.; Oberhofer, H. *Phys. Rev. Lett.* **2016**, *117*, 276001.

(100) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(101) Barducci, A.; Bonomi, M.; Parrinello, M. *Biophys. J.* **2010**, *98*, L44–L46.

(102) Dama, J. F.; Parrinello, M.; Voth, G. A. *Phys. Rev. Lett.* **2014**, *112*, 240602.

(103) Rasmussen, C.; Williams, C., *Gaussian Processes for Machine Learning*; Adaptative computation and machine learning series; University Press Group Limited: 2006.

(104) Do, C. B. Gaussian processes., `http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf`, Accessed: 2017-06-17.

(105) Sun, R.; Dama, J. F.; Tan, J. S.; Rose, J. P.; Voth, G. A. *J. Chem. Theory Comput.* **2016**, *12*, 5157–5169.

(106) Lelièvre, T.; Rousset, M.; Stoltz, G., *Free Energy Computations: A Mathematical Perspective*; Imperial College Press: London, 2010.

(107) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(108) Grossfield, A. WHAM: the weighted histogram analysis method, version 2.0.9., `http://membrane.urmc.rochester.edu/content/wham`, Accessed: 2017-08-20.

(109) Thiede, E. H.; Koten, B. V.; Weare, J.; Dinner, A. R. *J. Chem. Phys.* **2016**, *145*, 084115.

(110) Meng, Y.; Roux, B. *J. Chem. Theory Comput.* **2015**, *11*, PMID: 26574437, 3523–3529.

(111) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, ed. by Berne, B. J.; Ciccotti, G.; Coker, D. F., 1998, pp 385–404.

(112) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.