Computational and Theoretical Investigation of Energy Materials and Biological Systems

Thesis Submitted for the Degree of

Doctor of Philosophy (Science)

of

Jadavpur University

by

Sudipta Mitra

Department of Chemical and Biological Sciences
S. N. Bose National Centre for Basic Sciences
Block-JD, Sector-III, Salt Lake
Kolkata-700106, India

August 2025

S.N. Bose National Centre for Basic Sciences

Sector - III, Block - JD, Salt Lake, Kolkata - 700098

Dr. Ranjit Biswas
Senior Professor
Chemical and Biological Sciences

Phone: 033 2335 5706-08 (O)
email: ranjit@bose.res.in
Fax: 033 2335 3477

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **"Computational and Theoretical Investigation of Energy Materials and Biological Systems"** submitted by Mr. **Sudipta Mitra** (Index no. 106/21/Phys./27), who got his name registered on 23/11/2021 for the award of the Ph.D (Science) degree of Jadavpur University, is absolutely based upon his own work under the supervision of Prof. Ranjit Biswas and that neither his thesis nor any part of it has been submitted for either any degree / diploma or any other academic award anywhere before.

Ranjit Biswas
02.09.25

DR. RANJIT BISWAS
Senior Professor
Dept. of Chemical, Biological & Macromolecular Sciences
S. N. Bose National Centre for Basic Sciences
Block - JD, Sector-III, Salt Lake, Kolkata - 700 106, India

Dedicated to my
Family Members and
Teachers

Acknowledgement

First and foremost, I wish to express my heartfelt gratitude to my supervisor, Prof. Ranjit Biswas, SNBNCBS Kolkata, for granting me the rare privilege of exploring several captivating fields of research during my Ph.D. This experience has deepened my understanding of the fundamental principles of statistical physics while also honing my skills in computer simulation. Under his mentorship, I have cultivated the ability to think critically, approach problems from multiple perspectives, and carry out research with independence and confidence. His guidance has enriched not only my academic journey but also my personal growth, imparting invaluable lessons that will remain with me for a lifetime.

Special thanks to Prof. Suman Chakrabarty, SNBNCBS Kolkata, for always guiding me with the same care and attention like his own students. His generosity in offering me numerous opportunities to learn and grow, coupled with his readiness to help whenever I needed support, has had a profound impact on both my academic progress and personal development.

I truly feel very fortunate and lucky to have Prof. Ranjit Biswas and Prof. Suman Chakrabarty as mentors throughout my Ph.D journey. Their vast knowledge, insightful way of thinking, unwavering dedication to their work, and constant encouragement to explore and experiment with new ideas have been a continual source of inspiration. Equally, their kindness and genuine care have provided me with both comfort and motivation, making my research journey deeply enjoyable.

I am also immensely grateful to Prof. Anirban Bhunia, Bose Institute Kolkata, for the opportunity to collaborate with him on his remarkable projects.

My sincere thanks to all of the members of my thesis committee and research advisory committee (RAC), Prof. Pradip Kumar Ghorai, IISER Kolkata, Prof. Nabin Baran Manik, Jadavpur University and Prof. Manik Pradhan, SNBNCBS Kolkata, for generously sharing their valuable time with me despite their demanding schedules.

I am deeply grateful to our institute, SNBNCBS, for providing financial support throughout my Ph.D. I also extend my sincere appreciation to all those involved in procuring and maintaining the computational facilities at SNBNCBS. Their hard work and dedication have ensured uninterrupted access to the resources essential for carrying out my research effectively.

I wish to express my deepest gratitude to my family for their unwavering faith in me—often believing in my abilities far more than I believed in myself. Their constant encouragement, unconditional love, and steadfast support in every possible way have been the bedrock that allowed me to remain fully focused on my Ph.D. They have been my greatest source of strength, standing beside me through every challenge and celebrating every milestone along the way.

I would like to express my sincere appreciation to all the PCCP lab members—Dr. Jayanta Mondal, Dr. Narayan Chandra Maity, Dr. Dhrubajyoti Maji, Dr. Jayeta Banerjee, Dr. Tonima Nandy, Ms. Amrita Mondal, Mr. Rik Niranjan Mukherjee, and Mr. Abdul Aziz Mondal—for creating a supportive, collaborative, and vibrant atmosphere in the lab. Their camaraderie, encouragement, and willingness to share knowledge have not only made my research experience enriching but also turned the lab into a place I have genuinely enjoyed coming to every day. I am deeply grateful for their unwavering companionship and the countless moments of learning, laughter, and inspiration we have shared.

I would also like to extend my heartfelt appreciation to the Chakrabarty lab members—Dr. Krishnendu Sinha, Dr. Samir Ahamed, Dr. Saheb Dutta, Mr. Dibyendu Maity, Mr. Sutanu Mukhopadhyay, Mr. Sreyan Bhowmick, Ms. Sayari Bhattacharya, and Ms. Shaheerah Shahid—for their constant support and cooperation. Their willingness to assist me in various aspects of my research, whether through sharing expertise, offering constructive feedback, or lending a helping hand in practical matters, have been truly invaluable.

Finally, I will always remain grateful to my dear friends—Anirban, Soumen, Ishita, Koushik, Avik, Shinjini, Soham, Animesh, and Rajdeep—for all the laughter, conversations, and unforgettable moments we shared at SNBNCBS. Their friendship, encouragement, and constant presence turned countless stressful days into memorable ones, making both the Ph.D journey and campus life not only bearable but truly enjoyable.

Abstract

This thesis, entitled ***“Computational and Theoretical Investigation of Energy Materials and Biological Systems”***, primarily focuses on elucidating how molecular-level interactions govern structural, dynamical, and functional properties of diverse systems, spanning from chemical to biological domains. The works employ molecular dynamics simulations in synergy with the theoretical frameworks of physical chemistry, statistical physics, and modern machine learning techniques. The systems studied encompass Li-ion battery electrolytes (LIBs), aqueous Zn-ion battery electrolytes (AZIBs), and an enzyme known as Laccase, which holds significant potential for application in the textile industry. The central theme of this thesis is the in-depth investigation of ion transport in battery electrolyte systems, alongside the exploration of the structure–function relationship of Laccase. In addition, the development of a novel enhanced sampling method for studying the kinetics of biological processes is presented.

We began by exploring the effects of ion-ion, ion-solvent and solvent-solvent interactions on the complex ion transport phenomena in battery electrolytes. In particular, we examined the capabilities and limitations of the Van Hove function (VHF) in capturing ion-ion dynamical correlations and assessing their influence on ionic conductivity within the framework of the Onsager’s transport theory for concentrated electrolyte solutions. Our analysis demonstrates the potential of the VHF to identify different types of ion–ion dynamical correlations; however, we emphasize the need for caution when relying solely on the VHF to fully characterize these correlations and their impact on ion transport in battery electrolytes. Subsequently, to investigate the influence of ion-solvent and solvent-solvent interactions, we explored the role of water structure and dynamics in the presence of a co-solvent on ion transport phenomena in a representative experimental AZIB electrolyte system. Our results reveal that the co-solvent modulated water hydrogen bond (H-bond) network and dynamics around cations are strongly correlated with ion transport properties and provide a microscopic explanation on the experimentally observed mundane viscosity dependence of ionic conductivity in AZIB systems in the presence of co-solvent. Our findings will certainly help to understand ion transport for developing next generation battery electrolytes.

Next, we investigated the structure–function relationship of a fungal laccase, capable of degrading textile industry dye effluents. Our results elucidate the molecular thermodynamic origin of the experimentally observed substrate promiscuity in laccase, revealing the remarkable flexibility of its active site in accommodating dye molecules of diverse shapes and charges. Moreover, using extensive molecular dynamics simulations integrated with advanced machine learning techniques, we comprehensively characterized the free energy landscape of the laccase apo form. Based on our analysis, we hypothesize that dye binding occurs predominantly via a conformational selection mechanism, with inter-residue hydrogen bonds playing a key role in governing the slow kinetics associated with transitions between different dye molecule binding competent metastable states of the apo enzyme. Furthermore, our findings underscore the presence of allosteric regulation in laccase.

Finally, we introduce a new enhanced sampling approach, termed “WeTICA”, built upon the weighted ensemble (WE) path sampling algorithm. Our method employs a low-dimensional linear collective variable space to enhance sampling toward a specified target state. WeTICA has demonstrated success in capturing the unfolding kinetics of several benchmark proteins and holds promise for investigating a broad range of biological processes in a straightforward and user-friendly manner.

সারাংশ

এই থিসিস মূলত ব্যাখ্যা করে যে কীভাবে আণবিক-স্তরের মিথস্ক্রিয়া বিভিন্ন সিস্টেমের গঠনগত, গতিশীল এবং কার্যগত বৈশিষ্ট্য নিয়ন্ত্রণ করে, যা রসায়ন থেকে জীববিজ্ঞানের ক্ষেত্র পর্যন্ত বিস্তৃত। এই কাজটিতে অণু-গতিবিদ্যা সিমুলেশনকে ভৌত রসায়ন, পরিসংখ্যানগত পদার্থবিজ্ঞান এবং আধুনিক মেশিন লার্নিং কৌশলের তাত্ত্বিক কাঠামোর সাথে একত্রে ব্যবহার করা হয়েছে। গবেষণায় অধ্যয়ন করা সিস্টেমগুলির মধ্যে রয়েছে লিথিয়াম-আয়ন ব্যাটারি ইলেক্ট্রোলাইট, জিঙ্ক-আয়ন জলীয় ব্যাটারি ইলেক্ট্রোলাইট এবং ল্যাকেস নামে একটি উত্সেচক, যার বয়নসংক্রান্ত শিল্পে প্রয়োগের জন্য উল্লেখযোগ্য সম্ভাবনা রয়েছে। এই থিসিসের কেন্দ্রীয় বিষয় হল ব্যাটারি ইলেক্ট্রোলাইট সিস্টেমে আয়ন পরিবহনের গভীর অনুসন্ধান, পাশাপাশি ল্যাকেস উত্সেচক -এর গঠন–কার্য সম্পর্কের অধ্যয়ন। এছাড়াও, জৈব প্রক্রিয়ার গতিবিদ্যা অধ্যয়নের জন্য একটি নতুন বর্ধিত নমুনা গ্রহণ পদ্ধতির বিকাশ উপস্থাপন করা হয়েছে।

আমরা প্রথমে ব্যাটারি ইলেক্ট্রোলাইটে জটিল আয়ন পরিবহন ঘটনায় আয়ন-আয়ন, আয়ন-দ্রাবক এবং দ্রাবক-দ্রাবক পারস্পরিক মিথস্ক্রিয়ার প্রভাবগুলি পরীক্ষা করি। বিশেষ করে, আমরা ঘনীভূত ইলেক্ট্রোলাইট দ্রবণের জন্য অনসাগারের পরিবহন তত্ত্বের কাঠামোর মধ্যে, আয়ন-আয়ন গতিশীল সহসম্পর্ক ধারণ এবং তাদের আয়নিক পরিবাহিতায় প্রভাব মূল্যায়নে ভ্যান হোভ অপেক্ষক -এর ক্ষমতা ও সীমাবদ্ধতা বিশ্লেষণ করেছি। আমাদের বিশ্লেষণ দেখায় যে, ভ্যান হোভ অপেক্ষক বিভিন্ন ধরনের আয়ন–আয়ন গতিশীল সহসম্পর্ক চিহ্নিত করার সম্ভাবনা রাখে; তবে, শুধুমাত্র ভ্যান হোভ অপেক্ষক -এর উপর নির্ভর করে এই সহসম্পর্ক এবং তাদের আয়ন পরিবহনে প্রভাব পুরোপুরি চিহ্নিত করার ক্ষেত্রে সতর্কতা প্রয়োজন। পরবর্তীতে, আয়ন–দ্রাবক এবং দ্রাবক–দ্রাবক পারস্পরিক মিথস্ক্রিয়ার প্রভাব তদন্ত করার জন্য, আমরা একটি প্রতিনিধিত্বমূলক পরীক্ষামূলক জিঙ্ক-আয়ন জলীয় ব্যাটারি ইলেক্ট্রোলাইট সিস্টেমে সহ-দ্রাবকের উপস্থিতিতে জলের গঠন এবং গতিবিদ্যার ভূমিকা অন্বেষণ করেছি। আমাদের ফলাফলগুলি প্রকাশ করে যে, সহ-দ্রাবক দ্বারা নিয়ন্ত্রিত জলের হাইড্রোজেন বন্ধন নেটওয়ার্ক এবং ক্যাটায়নের চারপাশের জলের গঠন এবং গতিবিদ্যা আয়ন পরিবহন বৈশিষ্ট্যের সাথে দৃঢ়ভাবে সম্পর্কিত এবং সহ-দ্রাবকের উপস্থিতিতে জিঙ্ক-আয়ন জলীয় ব্যাটারি ইলেক্ট্রোলাইট সিস্টেমে পরীক্ষামূলকভাবে পর্যবেক্ষিত সান্দ্রতার উপর আয়নিক পরিবাহিতার সাধারণ নির্ভরতার একটি অণুবিশ্বীয় ব্যাখ্যা প্রদান করে। আমাদের অনুসন্ধানগুলি ভবিষ্যৎ প্রজন্মের ব্যাটারি ইলেক্ট্রোলাইট উন্নয়নে আয়ন পরিবহন বোঝার ক্ষেত্রে নিঃসন্দেহে সহায়ক হবে।

পরবর্তী পর্যায়ে, আমরা বয়নসংক্রান্ত শিল্পের রঙিন বর্জ্য অবক্ষয় করতে সক্ষম একটি ফাঙ্গাল ল্যাকেসের গঠন–কার্য সম্পর্ক তদন্ত করেছি। আমাদের ফলাফলগুলি ল্যাকেসে পরীক্ষামূলকভাবে পর্যবেক্ষিত সাবস্ট্রেট প্রমিসকুইটির আণবিক তাপগতীয় উৎস ব্যাখ্যা করে, যা ল্যাকেসের সক্রিয় স্থান-এর অসাধারণ নমনীয়তা প্রকাশ করে, যা বিভিন্ন আকার এবং আধান রঞ্জক পদার্থকে গ্রহণ করতে সক্ষম। তদুপরি, বিস্তৃত অণু-গতিবিদ্যা সিমুলেশন এবং উন্নত মেশিন লার্নিং কৌশল একত্রে ব্যবহার করে, আমরা ল্যাকেসের অ্যাপো রূপের মুক্ত শক্তি ভূদৃশ্য সম্পূর্ণভাবে চরিত্রায়িত করেছি। আমাদের বিশ্লেষণের ভিত্তিতে, আমরা অনুমান করি যে রঞ্জক পদার্থ বাঁধাই প্রধানত একটি কনফরমেশনাল নির্বাচনের মাধ্যমে ঘটে, যেখানে অ্যামিনো অ্যাসিড হাইড্রোজেন বন্ধনগুলি ধীর গতিবিদ্যা নিয়ন্ত্রণে গুরুত্বপূর্ণ ভূমিকা পালন করে, যা অ্যাপো উত্সেচক-এর বিভিন্ন রঞ্জক বাঁধাই-সক্ষম মেটাস্টেবল অবস্থার মধ্যে রূপান্তরের সাথে সম্পর্কিত। এছাড়াও, আমাদের অনুসন্ধানগুলি ল্যাকেসে অ্যালোস্টেরিক নিয়ন্ত্রণের উপস্থিতি নির্দেশ করে।

সবশেষে, আমরা একটি নতুন বর্ধিত নমুনা গ্রহণ পদ্ধতি উপস্থাপন করছি। আমাদের পদ্ধতিতে একটি নিম্ন-মাত্রার রৈখিক সমষ্টিগত পরিবর্তনশীল রাশি ক্ষেত্র ব্যবহার করে নির্দিষ্ট লক্ষ্য অবস্থার দিকে নমুনা গ্রহণ বৃদ্ধি করা হয়। আমাদের পদ্ধতি একাধিক প্রোটিনের পাক খুলিয়া দেওয়া গতিবিদ্যা ধরতে সাফল্য দেখিয়েছে এবং সহজ ও ব্যবহার-বান্ধব পদ্ধতিতে বিস্তৃত জৈব প্রক্রিয়া অধ্যয়নের জন্য সম্ভাবনাময়।

List of publications

1) Sudipta Mitra, Ranjit Biswas & Suman Chakrabarty, *"WeTICA: A Directed Search Weighted Ensemble Based Enhanced Sampling Method to Estimate Rare Event Kinetics in Reduced Dimensional Space",* **J. Chem. Phys.**, 162, 034106 (**2025**).

2) Sudipta Mitra & Ranjit Biswas*, "Exploring the Capabilities and Limitations of the Van Hove Function to Understand Directional Correlations in Ion Movements with Li-ion Battery Electrolytes",* **J. Chem. Phys**., 161, 064501 **(2024)**.

3) Sudipta Mitra, Arnab Sil, Ranjit Biswas & Suman Chakrabarty, "*Molecular Thermodynamic Origin of Substrate Promiscuity in the Enzyme Laccase: Toward a Broad-Spectrum Degrader of Dye Effluents*" , **J. Phys. Chem. Lett.,** 14, 1892−1898 **(2023)**.

4) Sudipta Mitra & Ranjit Biswas, "*Correlations Between Ionic Conductivity and Co-solvent Modulated Water Structure and Dynamics in Aqueous Zn-ion Battery Electrolytes*", **(Under review).**

5) Sudipta Mitra, Ranjit Biswas & Suman Chakrabarty, "*Unveiling Dye Effluent Binding Mechanism and Allostery in the Substrate Promiscuous Enzyme Laccase using Molecular Dynamics and Machine Learning Approaches*", **(Under preparation).**

6) ***Karishma Biswas, **Sudipta Mitra**, Dibakar Roy, Sanhita Roy, Dibakar Sarkar, DeokHyun Son, Rohit Das, Anuradha Roy, Dulal Senapati, Humaira Ilyas, A. Harikishore, Ranjit Biswas, Suman Chakrabarty, DongKuk Lee, Indranil Biswas, Sudipto Saha, Pallob Kundu, Anirban Bhunia, "*Transgenic tobacco plants expressing synthetic peptides: functional and structural analysis for pathogen resistance*", **(Accepted in the Plant Biotechnology Journal, 2025).**

7) ***Jayanta Mondal, Dhrubajyoti Maji, **Sudipta Mitra** & Ranjit Biswas*, "Temperature Dependent Dielectric Relaxation Measurements of (Betaine + Urea + Water) Deep Eutectic Solvent in Hz-GHz Frequency Window: Microscopic Insights into Constituent Contributions and Relaxation Mechanisms",* **J. Phys. Chem. B**, 128, 6567−65806568 **(2024)**.

8) ***Narayan Chandra Maity, **Sudipta Mitra** & Ranjit Biswas, "*What Dictates the Optimal Concentration of Li-Based Battery Electrolytes? A Combined Experimental and Simulation Study",* **(Under preparation)**.

*** Not included in this thesis.

Contents

Chapter 4: Correlations between ionic conductivity and co-solvent modulated water structure and dynamics in aqueous Zn-ion battery electrolytes……...53

Chapter 5: Molecular thermodynamic origin of substrate promiscuity in the enzyme laccase: Toward a broad spectrum degrader of dye effluents……77

1

Introduction

Molecular level interactions, that is, the forces (for example, electrostatic interactions, van der Waals forces, covalent bonding etc) acting between the atoms and molecules are the foundation of all natural phenomena[1,2]. These interactions govern the intricate dynamics and reactivity of atoms and molecules, ultimately shape the macroscopic behaviours of complex physical, chemical and biological systems. Various experimental techniques are designed and employed to quantify macroscopic phenomena in an accessible, human-readable format[3,4]; however, experimental results alone are typically insufficient to completely elucidate the underlying molecular level mechanisms of a system's observed macroscopic behaviours. Nevertheless, by connecting the molecular level insights to experimental observations—— one can validate theoretical models, and design innovative solutions across fields like materials science, chemistry, and biology.

Computer simulation is a virtual microscope used to describe the microscopic evolution of atoms and molecules of a system, provided a model that mimics the interactions between them[5,6]. Therefore, computer simulation provides molecular level insights behind macroscopic experimental observations. Moreover, computer simulation also acts as a bridge between theory and experiment[7]. We can validate a theoretical model by simulating it and subsequently compare the results with experiment. There are two main classes of computer simulation techniques: (i) Molecular dynamics (MD)[8] and (ii) Monte Carlo (MC)[9]. However, there are lots of hybrid techniques as well that combine the both[10–12]. In this thesis, we will mostly talk about various applications of classical MD in understanding molecular level interactions that govern the dynamics of various systems. Classical MD works by numerically integrating Newton's equations of motion for the atoms. Therefore MD provides atomistic insights into the structure and dynamics of complex systems[8], making it a powerful tool across fields like material science, chemistry and biology[13–15]. Moreover, it is important to note that computer simulation is not a historical phenomenon, but a rapidly developing field of science.

The potential of MD in providing molecular level insights behind experimental observations was first demonstrated in the pioneering work of A. Rahman in 1964, where the author simulated liquid argon using Lennard-Jones potential[16]. The first MD simulation study on biological system was conducted in 1975 by Levitt and Warshel[17]. Later, in 1976, the first MD simulation study to understand a biological process, the first step in the vision process, was conducted[18]. This study provided molecular level descriptions of the events that happen after absorbing a photon by rhodopsin, thereby complemented the observations from Raman and picosecond laser experiments. Subsequently, in 1977, MD simulation study of folding of the bovine pancreatic trypsin inhibitor (BPTI) was performed by solving the equations of motion for the atoms using an empirical potential energy function[19]. The results showed that proteins were not rigid and their internal motions play a functional role. Although, all these initial studies showed the ability of MD in providing in-depth understanding of various biological processes, the unavailability of optimize algorithms and powerful computers prohibited simulations of long time scale processes. Now a days, development of numerous advanced simulation algorithms and availability of supercomputers have immensely contributed to the development of this field[20–25]. Today, some of the major applications of MD in biology are sampling the conformational space of biomolecules[26–28], computing thermodynamics of biological processes[29–31], studying enzymatic reactions[32–34] and quantifying protein-ligand binding/unbinding for drug discovery[35–39] etc. Moreover, the application of MD is not restricted only to biology. MD has widespread applications in other fields like material science and chemistry[13,14,40].

On the other hand, the late 1990s witnessed the rise of research on electrochemical energy storage devices, such as rechargeable batteries, employing MD simulations[41–43]. The main components of a rechargeable battery cells are electrolyte, electrodes and separator[44]. The electrolyte is an extremely important part of a battery cell, which plays the role of an ion conducting medium as well as an electronic insulator[45,46]. Electrode materials determine energy density, cyclic efficiency and lifetime of a battery cell. MD simulations are extremely advantageous in understanding the structure and dynamics of ions and solvents in electrolyte and examining chemical reactions at the electrode-electrolyte interface to gain molecular level insights for improving battery design[47]. For example, MD simulations can be used to compute macroscopic observables, such as the self-diffusion coefficients of ions[48]. Thus, MD not only provides a molecular level picture of ion transport, but also makes it possible to predict several electrolyte properties without the need of doing sophisticated and expensive experiments, such as nuclear magnetic resonance (NMR)[49]. Moreover, Ab initio molecular dynamics (AIMD) has also made it possible to study the decomposition of solvents and ions on the electrode interface to understand the molecular level mechanism responsible for the formation of solid electrolyte interfaces (SEIs)[50–52] and cathode electrolyte interfaces (CEIs)[53–55], which are crucial for the long lifetime and efficiency of a battery cell.

While battery electrolytes and biological systems serve distinct purposes, they share a fundamental commonality; precise molecular interactions lead to solution structure and functionality. In battery electrolytes, ion-ion and ion-solvent interactions dictate performance[56,57], much like inter-residue, inter-atomic and protein-bound water interactions dictate protein structure and drive function[58,59]. Moreover, both fields face challenges; battery electrolytes require improved safety and energy density[60–62], while biological systems demand deeper insights into the structure-function relationship to understand and prevent complex diseases[63,64].

Interdisciplinary approach, like integrating chemistry, physics, and biology together is driving the development of next-generation theories[65,66]. Together, these domains exemplify how molecular-level understanding fuels transformative technologies for a sustainable and healthier future. Therefore, in this thesis, we have explored how the molecular level interactions govern the functionality of battery electrolytes as well as an enzyme with industrial relevance using MD simulations complemented by several theories from physical chemistry, statistical physics and machine learning (ML) approaches.

This thesis is comprised of a total of 8 chapters, with this introductory chapter being the first one. In **Chapter 2**, we have discussed about the theories and methodologies used in this thesis work.

In **Chapter 3**, we have discussed the role of microscopic directional correlations in the movement of ions arising due to the electrostatic interaction between them, in dictating ionic conductivity and investigated the potential of a pair correlation function, known as Van Hove function (VHF)[67], in order to capture the ion dynamical correlations from MD simulation data. When ions of same or opposite type move in the same direction, the directional correlation between their movement is said to be positive, while if they move in opposite direction, their movement is anti-correlated. Onsager transport theory and the corresponding transport coefficients are widely used to understand these correlations[68,69]. However, computing these coefficients from MD simulation is not a trivial task. On the other hand, the VHF is also capable of determining correlated motions and computing VHFs from MD simulation data is much simpler than computing the Onsager transport coefficients. However, identifying various types of ion correlated motions in battery electrolytes using VHF is not well explored. Thus, we have conducted MD simulations of a representative experimental lithium-ion battery (LIB) electrolyte system[70] - lithium hexafluorophosphate ($LiPF_6$) at different concentrations in (9:1 wt%) mixture of ethyl methyl carbonate (EMC) and fluoroethylene carbonate (FEC) using general Amber force filed (GAFF2)[71] at room temperature in order to explore the capabilities and limitations of using VHF to investigate different types of ion dynamical correlations. We have concluded that the analysis of VHF can qualitatively describe both the positive correlation between the movement of cations and anions at different salt concentrations and the negative correlation between cation-cation and anion-anion present at high salt concentration, but it cannot foretell which correlation is dominating at any given electrolyte concentration. This type of quantitative information can be obtained only via Onsager's approach. This could be seen as a limitation of relying solely on VHF to fully understand ion correlation in electrolyte media.

Next, we have investigated how ion-solvent interactions and solvent dynamics affect ion transport in battery electrolyte solutions in aqueous zinc ion batteries. Currently, the non-aqueous LIBs are dominating the electrochemical energy storage industry because of their high energy density and long life cycle. However, increasing concern about limited lithium resources, high cost and safety issues have seriously hindered the continuous large scale manufacturing of LIBs[72]. Sodium-ion batteries (SIBs)[73] and potassium-ion batteries (KIBs)[74] are plausible alternatives to LIBs because of the relative abundant of sodium (potassium) over lithium, but suffer from low energy density, high operating cost and security issues. These drawbacks of SIBs and KIBs have motivated the scientists to explore alternative battery chemistry. Aqueous zinc-ion (Zn^{2+}) batteries (AZIBs) are gaining popularities due to several factors over LIBs, but the corrosion of the zinc metal anode through reaction with the Zn^{2+} ion solvated water molecules hinder its practical usage[75,76]. The use of various co-solvents as AZIB electrolyte additives has been proven to be very effective in treating anode corrosion[77].

In **Chapter 4**, we have investigated correlations between ionic conductivity, that characterizes ion transport and co-solvent modulated water structure and dynamics in AZIB electrolyte solutions. We have performed all-atom classical MD simulation of an experimental AZIB system[78]; 2M $\mathrm{Zn(OTf)_2}$ aqueous solution with tetramethylurea (TMU) as a co-solvent at room temperature. We have found that the presence of TMU increases the residence time of water molecules inside Zn^{2+} solvation shell by promoting prolonged hydrogen bonding (H-bond) between solvated waters and their immediate neighbouring water molecules. Moreover, we have observed that ionic conductivity is positively correlated with the number of water-water H-bonds around Zn^{2+} ions, but anti-correlated with both the Zn^{2+} solvation shell water residence time and aforementioned water-water H-bond lifetime. The increase in water-water H-bond lifetime around Zn^{2+} ions enforces the vehicular movement of Zn^{2+}, thus slows down ion transport and decreases ionic conductivity. Macroscopically, this is reflected in the overall rise in solution viscosity with increasing TMU concentration. Furthermore, we have also investigated the role of ion-ion correlations using the Onsager's transport coefficients in dictating ionic conductivity and the transference number of AZIB electrolytes.

Upon discussing the role of molecular level interactions, e.g ion-ion and ion-solvent interactions in dictating the macroscopic transport properties, such as ionic conductivity and transference number of the battery electrolyte solutions using MD simulations and several theoretical approaches, we will now shift gear and discuss the role of molecular level interactions to the function of an enzyme, known as Laccases (EC 1.10.3.2)[79,80], which is capable of degrading industrial dye effluents. Laccases belong to the family of multicopper oxidoreductases (MCOs)[81] found mostly in white-rot fungi, have been experimentally observed to decolorize dye wastewater to different extents. Laccases can oxidize a variety of phenols, polyphenols, aromatic amines and non-phenolic organic compounds using oxygen as a reactant and release water as the only by-product[82].

In **Chapter 5**, we have explored the molecular thermodynamic origin of the substrate promiscuity[83] in a fungal laccase (PDB: 1KYA) using molecular docking[84] and MD simulation studies. We choose five dye molecules with varying charge, size and shape: brilliant blue, coumarin 343, methyl green, crystal violet and thioflavin T. We have discovered the presence of various distinct conformations of a loop (resids: 159-164) in the protein active site that can accommodate the wide range of dye molecules. We have also observed that the diverse selection of dye molecules exhibits surprisingly similar binding affinity due to cancellation of different thermodynamic factors.

In **Chapter 6**, we have identified and characterized different metastable apo conformations of the laccase using MD simulations coupled with various machine learning (ML) techniques to unveil the binding mechanism of the aforementioned dye effluents to the laccase. We have identified key residue pairs using the Random forests classifier[85] that can distinguish protein apo and bound states, and compared two different classes of dimensionality reduction methods; time-lagged Independent Component Analysis (TICA, a linear method)[86,87] and variational autoencoder (a non-linear method)[88], to obtain the best low dimensional representation for the laccase apo conformational landscape. We then performed kinetic clustering using the neural network implementation of variational approach for Markov processes (VAMPnets)[89] to identify metastable apo conformations of laccase. Finally, we have found dye effluent binding competent protein conformations that could possibly lead to ligand binding via conformational selection mechanism[90]. Hydrogen bond occupancy analysis showed significant difference in inter-residue hydrogen bond network among these metastable conformations, results in sluggish kinetics as revealed by Markov State Model[91]. These observations have shown the vital role of inter-residue interactions in controlling protein conformational changes. Moreover, we have also proposed an allosteric connection[92] between the active site loop (resids:159-264) and a distal loop (resids: 332-337) in laccase. Therefore, these two studies provide a complete molecular level understanding on the dye degradation mechanism by laccase.

Despite the remarkable advances in MD software and hardware that enable us to access millisecond time scales at the atomistic resolution, normal MD simulations usually struggle to overcome the barriers associated with different processes making such events rare and difficult to capture. A plethora of methods, known as enhanced sampling techniques, have been developed as a solution for sampling rare events[93–96]. Weighted Ensemble (WE) simulation, a special class of enhanced sampling technique, offers a way to directly calculate kinetic rate constants from biased trajectories without the need to modify the underlying energy landscape using bias potentials[97,98]. Conventional WE algorithms use different binning schemes to partition the collective variable (CV) space separating the two metastable states of interest.

In **Chapter 7,** we have discussed our attempt in developing a new "binless" WE simulation algorithm to bypass the hurdles of optimizing binning procedures. Our proposed protocol called "WeTICA" uses a low dimensional CV space to drive the WE simulation towards the specified target state. We have applied this new algorithm to recover the unfolding kinetics of three proteins: A) TC5b Trp-cage mutant, B) TC10b Trp-cage mutant and C) Protein G with unfolding times[20] spanning the range between 3 μs − 40 μs using projections along predefined fixed Time-lagged Independent Component Analysis (TICA)[86,87] eigenvectors as CVs. Calculated unfolding times converge to the reported values with good accuracy with more than one order of magnitude less cumulative WE simulation time than the unfolding time scales with or without a priori knowledge of the CVs that can capture unfolding. Our algorithm can be used with other linear CVs, not limited to TICA. Moreover, the new walker selection criteria for resampling employed in this algorithm can be used on more sophisticated nonlinear CV space for further improvements of binless WE methods.

The final chapter of this thesis is **Chapter 8**, which includes concluding remarks and possible areas of future research.

References:

1 G. N. Lewis, *J Am Chem Soc*, 1916, 38, 762–785.

2 D. Günther, R. A. Boto, J. Contreras-Garcia, J. P. Piquemal and J. Tierny, *IEEE Trans Vis Comput Graph*, 2014, 20, 2476–2485.

3 Y. Seo and W. Jhe, *Reports on Progress in Physics*, 2007, 71, 016101.

4 H. Hölscher, 2002, 349–369.

5 M. Allen, D. Frenkel, J. T.-C. physics reports and undefined 1989.

6 B. Edmonds and D. Hales, *Journal of Mathematical Sociology*, 2005, 29, 209–232.

7 J. Lenhard, *Philos Sci*, 2007, 74, 176–194.

8 M *Tuckerman, GJ MartynaThe Journal of Physical Chemistry B,* 2000, 104, 159–178.

9 R. L. Harrison, *AIP Conf Proc*, 2010, 1204, 17–21.

10 M. E. Tuckerman, B. J. Berne, G. J. Martyna and M. L. Klein, *J Chem Phys*, 1993, 99, 2796–2808.

11 S. W. Chiu, E. Jakobsson, S. Subramaniam and H. L. Scott, *Biophys J*, 1999, 77, 2462–2469.

12 E. C. Neyts and A. Bogaerts, 2014, 277–288.

13 W. F. van Gunsteren and H. J. C. Berendsen, *Angewandte Chemie International Edition in English*, 1990, 29, 992–1023.

14 U. Landman, 1988, 108–123.

15 S. Krishna, I. Sreedhar and C. M. Patel, *Comput Mater Sci*, 2021, 200, 110853.

16 A. Rahman, *Physical Review*, 1964, 136, A405.

17 M. Levitt and A. Warshel, *Nature*, 1975, 253, 694–698.

18 A. Warshel, *Nature*, 1976, 260, 679–683.

19 J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, 1977, 267, 585–590.

20 K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, *Science (1979)*, 2011, 334, 517–520.

21 R. Lazim, D. Suh and S. Choi, *International Journal of Molecular Sciences 2020, Vol. 21, Page 6339*, 2020, 21, 6339.

22 D. Jones, J. E. Allen, Y. Yang, W. F. Drew Bennett, M. Gokhale, N. Moshiri and T. S. Rosing, *J Chem Theory Comput*, 2022, 18, 4047–4069.

23 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindah, *SoftwareX*, 2015, 1–2, 19–25.

24 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput Biol*, 2017, 13, e1005659.

25 J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy,

Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *Journal of Chemical Physics*.

26 J. R. Allison, *Biochem Soc Trans*, 2020, 48, 1707–1724.

27 B. T. Kaynak, J. M. Krieger, B. Dudas, Z. L. Dahmani, M. G. S. Costa, E. Balog, A. L. Scott, P. Doruker, D. Perahia and I. Bahar, *Front Mol Biosci*, 2022, 9, 832847.

28 M. P. D. Hatfield and S. Lovas, *Curr Pharm Des*, 2014, 20, 3303–3313.

29 C. Chipot, *Annu Rev Biophys*, 2023, 52, 113–138.

30 A. K. Padhi, M. Janežič and K. Y. J. Zhang, *Advances in Protein Molecular and Structural Biology Methods*, 2022, 439–454.

31 C. L. Brooks and D. A. Case, *Chem Rev*, 1993, 93, 2487–2502.

32 V. Vennelakanti, A. Nazemi, R. Mehmood, A. H. Steeves and H. J. Kulik, *Curr Opin Struct Biol*, 2022, 72, 9–17.

33 R. P. Magalhães, H. S. Fernandes and S. F. Sousa, *Isr J Chem*, 2020, 60, 655–666.

34 A. Warshel, *Acc Chem Res*, 2002, 35, 385–395.

35 S. Decherchi and A. Cavalli, *Chem Rev*, 2020, 120, 12788–12833.

36 S. R. Zia, A. Coricello and G. Bottegoni, *Curr Opin Struct Biol*, 2024, 87, 102871.

37 S. Wolf, *J Chem Inf Model*, 2023, 63, 2902–2910.

38 V. Salmaso and S. Moro, *Front Pharmacol*, 2018, 9, 393738.

39 H. Grubmüller, B. Heymann and P. Tavan, *Science (1979)*, 1996, 271, 997–999.

40 H. Yao, J. Liu, M. Xu, J. Ji, Q. Dai and Z. You, *Adv Colloid Interface Sci*, 2022, 299, 102565.

41 F. Müller-Plathe and W. F. Van Gunsteren, *J Chem Phys*, 1995, 103, 4745–4756.

42 G. Nuspl, M. Nagaoka, K. Yoshizawa, F. Mohri and T. Yamabe, *Bull Chem Soc Jpn*, 1998, 71, 2259–2265.

43 O. Ito, M. Mukaide and M. Yoshikawa, *Solid State Ion*, 1995, 80, 181–187.

44 H. Zhang, H. Zhao, M. A. Khan, W. Zou, J. Xu, L. Zhang and J. Zhang, *J Mater Chem A Mater*, 2018, 6, 20564–20620.

45 K. Xu, *Chem Rev*, 2014, 114, 11503–11618.

46 Y. S. Meng, V. Srinivasan and K. Xu, *Science (1979)*,

47 N. Yao, X. Chen, Z. H. Fu and Q. Zhang, *Chem Rev*, 2022, 122, 10970–11021.

48 J. Wang and T. Hou, *J Comput Chem*, 2011, 32, 3505–3519.

49 M. Holz, S. R. Heil and A. Sacco, *Physical Chemistry Chemical Physics*, 2000, 2, 4740–4742.

50 G. Bouder, H. Bouhani, H. Martinez and P. Carbonniere, *Journal of Physical Chemistry C*, 2025, 129, 8602–8613.

51 A. Wang, S. Kadam, H. Li, S. Shi and Y. Qi, *npj Computational Materials 2018 4:1*, 2018, 4, 1–26.

52 K. Xu, Y. Lam, S. S. Zhang, T. R. Jow and T. B. Curtis, *Journal of Physical Chemistry C*, 2007, 111, 7411–7421.

53 I. Takahashi, H. Kiuchi, A. Ohma, T. Fukunaga and E. Matsubara, *Journal of Physical Chemistry C*, 2020, 124, 9243–9248.

54 A. von Cresce and K. Xu, *J Electrochem Soc*, 2011, 158, A337.

55 D. R. Gallus, R. Wagner, S. Wiemers-Meyer, M. Winter and I. Cekic-Laskovic, *Electrochim Acta*, 2015, 184, 410–416.

56 X. Chen, N. Yao, B. S. Zeng and Q. Zhang, *Fundamental Research*, 2021, 1, 393–398.

57 X. Chen and Q. Zhang, *Acc Chem Res*, 2020, 53, 1992–2002.

58 D. H. Chou and C. V. Morr, *J Am Oil Chem Soc*, 1979, 56, A53–A62.

59 K. Gerwert, E. Freier and S. Wolf, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2014, 1837, 606–613.

60 T. Ould Ely, D. Kamzabek and D. Chakraborty, *Front Energy Res*, 2019, 7, 452480.

61 J. Janek and W. G. Zeier, *Nat Energy*.

62 Y. Yamada, J. Wang, S. Ko, E. Watanabe and A. Yamada, *Nature Energy 2019 4:4*, 2019, 4, 269–280.

63 R. G. Parton, *Annu Rev Cell Dev Biol*, 2018, 34, 111–136.

64 F. Rahimi, A. Shanmugam and G. Bitan, *Curr Alzheimer Res*, 2008, 5, 319–341.

65 F. H. Thaheld, *Biosystems*, 2005, 80, 41–56.

66 G. R. Van Hecke, K. K. Karukstis, R. C. Haskell, C. S. McFadden and F. S. Wettack, *J Chem Educ*, 2002, 79, 837.

67 L. Van Hove, *Physical Review*, 1954, 95, 249.

68 K. D. Fong, J. Self, B. D. McCloskey and K. A. Persson, *Macromolecules*, 2020, 53, 9503–9512.

69 N. M. Vargas-Barbosa and B. Roling, *ChemElectroChem*, 2020, 7, 367–385.

70 D. J. Xiong, M. Bauer, L. D. Ellis, T. Hynes, S. Hyatt, D. S. Hall and J. R. Dahn, *J Electrochem Soc*, 2018, 165, A126.

71 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J Comput Chem*, 2004, 25, 1157–1174.

72 T. Kim, W. Song, D. Y. Son, L. K. Ono and Y. Qi, *J Mater Chem A Mater*, 2019, 7, 2942–2964.

73 S. W. Kim, D. H. Seo, X. Ma, G. Ceder and K. Kang, *Adv Energy Mater*, 2012, 2, 710–721.

74 J. C. Pramudita, D. Sehrawat, D. Goonetilleke and N. Sharma, *Adv Energy Mater*.

75 S. Huang, J. Zhu, J. Tian and Z. Niu, *Chemistry - A European Journal*, 2019, 25, 14480–14494.

76 B. Tang, L. Shan, S. Liang and J. Zhou, *Energy Environ Sci*, 2019, 12, 3288–3304.

77 Y. Geng, L. Pan, Z. Peng, Z. Sun, H. Lin, C. Mao, L. Wang, L. Dai, H. Liu, K. Pan, X. Wu, Q. Zhang and Z. He, *Energy Storage Mater*, 2022, 51, 733–755.

78 Z. Li, Y. Liao, Y. Wang, J. Cong, H. Ji, Z. Huang and Y. Huang, *Energy Storage Mater*, 2023, 56, 174–182.

79 S. Riva, *Trends Biotechnol*, 2006, 24, 219–226.

80 D. M. Mate and M. Alcalde, *John Wiley & Sons, Ltd*, 2017, preprint.

81 E. I. Solomon, U. M. Sundaram and T. E. Machonkin, *Chem Rev*, 1996, 96, 2563–2605.

82 L. Arregui, M. Ayala, X. Gómez-Gil, G. Gutiérrez-Soto, C. E. Hernández-Luna, M. de Los Santos, L. Levin, A. Rojo-Dominguez, D. Romero-Martinez, M. C. N. N. Saparrat, others, M. Herrera De Los Santos, L. Levin, A. Rojo-Domínguez, D. Romero-Martínez, M. C. N. N. Saparrat, M. A. Trujillo-Roldán, N. A. Valdez-Cruz, M. de Los Santos, L. Levin, A. Rojo-Dominguez, D. Romero-Martinez, M. C. N. N. Saparrat and others, *Laccases: structure, function, and potential application in water bioremediation*, 2019, vol. 18.

83 K. Hult and P. Berglund, *Trends Biotechnol*, 2007, 25, 231–238.

84 S. S. Butt, Y. Badshah, M. Shabbir and M. Rafiq, *JMIR Bioinform Biotech 2020;1(1):e14232*, 2020, 1, e14232.

85 L. Breiman, *Mach Learn*, 2001, 45, 5–32.

86 C. R. Schwantes and V. S. Pande, *J Chem Theory Comput*, 2013, 9, 2000–2009.

87 L. Molgedey and H. G. Schuster, *Phys Rev Lett*, 1994, 72, 3634.

88 D. P. Kingma and M. Welling, *Foundations and Trends in Machine Learning*, 2019, 12, 307–392.

89 A. Mardt, L. Pasquali, H. Wu and F. Noé, *Nature Communications 2018 9:1*, 2018, 9, 1–11.

90 A. D. Vogt and E. Di Cera, *Biochemistry*, 2012, 51, 5894–5902.

91 B. E. Husic and V. S. Pande, *J Am Chem Soc*, 2018, 140, 2386–2396.

92 J. Liu and R. Nussinov, *PLoS Comput Biol*, 2016, 12, e1004966.

93 Y. Miao, V. A. Feher and J. A. McCammon, *J Chem Theory Comput*, 2015, 11, 3584–3595.

94 P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, *Annu Rev Phys Chem*, 2002, 53, 291–318.

95 A. Barducci, G. Bussi and M. Parrinello, *Phys Rev Lett*, 2008, 100, 020603.

96 A. F. Voter, *Phys Rev Lett*, 1997, 78, 3908.

97 G. A. Huber and S. Kim, *Biophys J*, 1996, 70, 97–110.

98 D. M. Zuckerman and L. T. Chong, *Annu Rev Biophys*, 2017, 46, 43–57.

2

Theory & methodology

2.1 Onsager's theory of ion transport in concentrated electrolytes

Onsager transport equation[1,2] characterizing ion transport in concentrated electrolyte is shown in Eq. (2.1)

$$\vec{J}_i = -\sum_j L^{ij}\,\vec{\nabla}\mu_j, \tag{2.1}$$

where $\vec{J}_i$ is the flux of species i (cations or anions), $\vec{\nabla}\mu_j$ denotes the gradient in electrochemical potential acting on species j and the summation is over all the species (including species i) in the system. L^{ij}s are called Onsager transport coefficients. These coefficients satisfy Onsager reciprocal relation[2]: $L^{ij} = L^{ji}$, and thus form a symmetric matrix. One way of computing these coefficients from MD simulation is using the Einstein's relation as follows[3,4]:

$$L^{ij} = \frac{1}{6k_BTV} \lim_{t\to\infty} \frac{d}{dt} \left\langle \sum_\alpha [\vec{r}_i^{\,\alpha}(t) - \vec{r}_i^{\,\alpha}(0)] \bullet \sum_\beta \left[\vec{r}_j^{\,\beta}(t) - \vec{r}_j^{\,\beta}(0)\right] \right\rangle, \tag{2.2}$$

and

$$L^{ii} = \frac{1}{6k_BTV} \lim_{t\to\infty} \frac{d}{dt} \sum_\alpha \sum_\beta \left\langle [\vec{r}_i^{\,\alpha}(t) - \vec{r}_i^{\,\alpha}(0)] \bullet \left[\vec{r}_i^{\,\beta}(t) - \vec{r}_i^{\,\beta}(0)\right] \right\rangle. \tag{2.3}$$

In Eqs. (2.2) and (2.3), $\vec{r}_i^{\,\alpha}$ denotes the position of the α^{th} ion of species i relative to the center-of-mass position of the entire system, V is the volume of the simulation box, T the temperature and k_B is the

Boltzmann constant. The mathematical expression of L^{ii} (in Eq. (2.3)) can be divided into a self-part and a distinct part. In Eq. (2.3), summation for $\alpha = \beta$ corresponds to the self-part and $\alpha \neq \beta$ to the distinct part. Therefore, self-part of L^{ii} can be calculated from the following relation[3,4],

$$L^{ii}_{self} = \frac{1}{6k_BTV} \lim_{t \to \infty} \frac{d}{dt} \sum_\alpha \langle [\vec{r}^{\alpha}_i(t) - \vec{r}^{\alpha}_i(0)]^2 \rangle , \qquad (2.4)$$

and the distinct part can be computed indirectly as $L^{ii}_{distinct} = L^{ii} - L^{ii}_{self}$.

Electrostatic interactions between ions give rise to directional correlations between the movement of ions. If ions of same and/or opposite type move preferentially in the same direction, directional correlation between their movement is called positive[5]. If they move in opposite direction, their relative motion is anti-correlated. The Onsager coefficients capture the effects of these ion dynamical correlations on the transport properties of concentrated electrolyte solutions, for example, ionic conductivity and transference number.

Ionic conductivity (σ), ion self-diffusion coefficient (D_i) and cation transference number (t_+) can be computed from these coefficients as follows[3]:

$$\sigma = F^2 \sum_i \sum_j z_i \, z_j \, L^{ij} , \qquad (2.5)$$

$$D_i = \frac{RT}{c_i} L^{ii}_{self} \qquad (2.6)$$

and

$$t_+ = \frac{\sum_j z_i \, z_j \, L^{ij}}{\sum_k \sum_l z_k \, z_l \, L^{kl}} , \qquad (2.7)$$

where F is the Faraday's constant, $z_+ = -z_-$ are the partial charges of ions, R is the real gas constant and c_i is the density of ions of type i.

For a binary electrolyte, only three transport coefficients L^{++} , L^{--} and L^{+-} are required to completely characterize ion transport. Ionic conductivity (σ) can be partitioned into five components as follows[3]:

$$\begin{aligned}\sigma &= F^2(z_+^2 L^{++}_{self} + z_-^2 L^{--}_{self} + z_+^2 L^{++}_{distinct} + z_-^2 L^{--}_{distinct} + 2z_+z_-L^{+-}) \\ &= \sigma^+_{self} + \sigma^-_{self} + \sigma^{++}_{distinct} + \sigma^{--}_{distinct} - 2\sigma^{+-} \qquad (2.8)\end{aligned}$$

For an infinitely dilute ideal solution of non-interacting ions, the sum of all the distinct cross terms in Eq. (2.8) is zero. Thus, for an infinitely dilute solution, the resulting conductivity follows the Nernst-Einstein (NE) value: $\sigma_{\mathrm{NE}} = \sigma^{+}_{\mathrm{self}} + \sigma^{-}_{\mathrm{self}}$.

For positive correlation between the movement of cations and anions, $\sigma^{+-} > 0$. Conversely, for anti-correlated cation-anion motion, $\sigma^{+-} < 0$. Similarly, for positive correlation between the movement of like-charged ions, distinct components ($\sigma^{++}_{\mathrm{distinct}}$ and $\sigma^{--}_{\mathrm{distinct}}$) become positive. Therefore, by observing the sign of these different conductivity components, one can also identify the nature of the directional correlations in ion movement that determines the net ionic conductivity of a given electrolyte system. Moreover, these distinct terms typically decrease the net conductivity from the ideal NE value in concentrated electrolyte solutions.

Furthermore, for binary electrolyte solutions, the expression for cation transference number (t_+) in Eq. (2.7) and its ideal solution value $\left(t_+^{\mathrm{NE}}\right)$ become[3]

$$t_+ = \frac{z_+^2 L^{++} + z_+ z_- L^{+-}}{z_+^2 L^{++} + z_-^2 L^{--} + 2 z_+ z_- L^{+-}} \tag{2.9}$$

and

$$t_+^{\mathrm{NE}} = \frac{z_+^2 L^{++}_{\mathrm{self}}}{z_+^2 L^{++}_{\mathrm{self}} + z_-^2 L^{--}_{\mathrm{self}}} \tag{2.10}$$

2.2 Van Hove function: pair correlation function in space and time

Correlated motions can also be captured using the Van Hove function (VHF)[6,7], which is a pair correlation function in real-space and time. The VHF is defined as[8,9]

$$G\,(r,\ t) = \frac{1}{N 4\pi r^2 dr} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \delta\left(r - \left|\vec{r}_i(t) - \vec{r}_j(0)\right|\right)\right\rangle\ , \tag{2.11}$$

where N is the number of molecules, $\vec{r}_i(t)$ is the position of the i^th^ molecule at time t, $\delta(r)$ is the Dirac delta function. G (r, t) is composed of a self-part $G_s\,(r,t)$ and a distinct part $G_d\,(r,t)$. The self-part represents the probability of finding a molecule at a distance r at a time t given that it was at r = 0 at t = 0. Meanwhile, the distinct part represents the probability of finding a molecule at a distance r at time t given that another molecule was at r = 0 at t = 0. $G_d\,(r,0)$ is identical to the radial distribution function

(RDF). In this thesis, the self-part and the distinct part of the VHF, denoted as $G_s^{\alpha}(r,\ t)$ and $G_d^{\alpha,\beta}(r,\ t)$ respectively, are calculated using the following equations from MD simulation data[10,11]

$$G_s^{\alpha}(r,\ t) = \frac{1}{N_\alpha 4\pi r^2 dr} \sum_{i=1}^{N_\alpha} \langle \delta(r - |\vec{r}_i(t) - \vec{r}_i(0)|) \rangle\ , \tag{2.12}$$

and

$$G_d^{\alpha,\beta}(r,\ t) = \frac{V}{N_\alpha N_\beta 4\pi r^2 dr} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{N_\beta} \langle \delta\big(r - \big|\vec{r}_j(t) - \vec{r}_i(0)\big|\big) \rangle\, , \tag{2.13}$$

where N_α and N_β are the number of ions of species (cations or anions) α and β respectively, V is the volume of the simulation box. To make $G_d^{\alpha,\beta}(r,\ t)\ \approx 1$ at large r and large t, the original definition in Eq. (2.11) is modified by the division through $\rho_\beta = \frac{N_\beta}{V}$. To calculate the intra-species ($\alpha\ =\ \beta$) distinct part of the VHF, one should consider the distinct ions of the same type, that is, the sum in Eq. (2.13) should run for $i\ \neq j$.

2.3 Autocorrelation functions

2.3.1 Residence time autocorrelation function

Residence time of a molecule is the time it spends in a specific region of space[12]. In solvation dynamics study, the residence time of molecules inside the solvation shell of another molecules is one of the important properties. Residence time can be computed using the residence time autocorrelation function[13] from MD simulation data as defined in Eq. (2.14):

$$C(t) = \frac{\langle H_{ij}(0) H_{ij}(t) \rangle}{\langle H_{ij}(0) H_{ij}(0) \rangle}, \tag{2.14}$$

where $H_{ij}(t) = 1$ if the i^th^ molecule lies inside the solvation shell of the j^th^ molecule at time t or $H_{ij}(t) = 0$ otherwise and $\langle \dots \rangle$ indicates ensemble average over different time origins. The simulated $C(t)$ can be fitted with a stretched exponential function $C(t) =\ e^{-\left({t}/{\tau}\right)^{\beta}}$, where τ denotes the residence time and the parameter β determines the nature of the exponential decay.

2.3.2 Structural hydrogen bond autocorrelation function

Hydrogen bond (H-bond) plays an important role in various chemical and biological processes. The structural H-bond relaxation[14,15], influenced by both translational and orientational molecular diffusion is quantified by the autocorrelation function defined in Eq. (2.15):

$$C_{HB}(t) = \frac{\langle h_{ij}(0)h_{ij}(t)\rangle}{\langle h_{ij}(0)h_{ij}(0)\rangle}, \tag{2.15}$$

Where the variable $h_{ij}(t)$ denotes the presence or absence of a H-bond between two molecules i and j at time t. More specifically, if a H-bond exists between two molecules at time t, then $h_{ij}(t) = 1$, otherwise $h_{ij}(t) = 0$. In simulation, the criteria for the formation of H-bond employed are: (i) $r_{DA} \leq$ 3.5 Å and (ii) $150^0 \leq \theta_{DHA} \leq 180^0$, where r_{DA} and θ_{DHA} are the donor (D) - acceptor (A) distance and D—hydrogen (H) - A angle, respectively.

The simulated $C_{HB}(t)$ is usually fitted with the sum of three exponential functions and time integrated analytically to obtain the structural H-bond relaxation time τ_{HB} as follows:

$$\tau_{HB} = \int_0^\infty dt\, C_{HB}(t) = \int_0^\infty dt \sum_{i=1}^3 a_i\, e^{-t/\tau_i} = \sum_{i=1}^3 a_i\, \tau_i\,, \tag{2.16}$$

where a_i are the fit parameters such that $\sum_{i=1}^3 a_i = 1$.

2.3.3 Pressure autocorrelation function and shear viscosity

Shear viscosity η can be obtained from the Green-Kubo integral of the pressure autocorrelation function defined in Eq. (2.17) as follows[16,17],

$$\eta = \frac{V}{6K_BT} \int_0^\infty dt\, \langle P_{\alpha\beta}(0)P_{\alpha\beta}(t)\rangle\,, \tag{2.17}$$

where V, K_B and T represent the box volume, the Boltzmann constant and temperature of the system, respectively. $P_{\alpha\beta}$ denotes the cross-diagonal terms of the pressure tensor. Although the pressure autocorrelation function theoretically decays to zero in the long time, leading to the convergence of Eq. (2.17), in practical application fluctuations prevent the integral from converging to a constant value[18]. To tackle this issue, instead of taking a single off-diagonal pressure component to calculate pressure autocorrelation function, the autocorrelation functions are computed for five independent pressure components $P_{xy}, P_{xz}, P_{yz}, (P_{xx} - P_{yy})/2$ and $(P_{yy} - P_{zz})/2$ and the average of them are calculated.

2.4 Dimensionality reduction techniques

Typically, the raw MD data contains the positions of all the atoms in the system (say, N number of atoms) and the systems under study are usually composed of thousands of atoms. Thus, the raw MD data is very high dimensional (3N dimensional). Therefore, in studying biological systems, it is very useful to visualize the different metastable conformational states of a biomolecule using a lower dimensional representation of the actual high dimensional data. There are several ways to reduce the dimensionality of a dataset as follows:

2.4.1 Principal component analysis (PCA)

Principal Component Analysis (PCA)[19,20] is a widely used linear dimensionality reduction technique that aims to reduce the dimensionality of a dataset while preserving as much variance in the original data as possible. Mathematically, PCA works by first constructing a covariance matrix C_{ij} using the high dimensional input time series dataset $\mathbf{X} = \{[x]_N^{t_1},\ [x]_N^{t_2},\ [x]_N^{t_3}, \dots\}$, where $[x]_N^{t_i}$ is the feature vector with N elements at time t_i as follows:

$$C_{ij} = \langle (x_i - \langle x_i \rangle)\big(x_j - \langle x_j \rangle\big) \rangle, \tag{2.18}$$

where x_i denotes the i^th feature. Then the orthogonal eigenvectors of this matrix are calculated and sorted in descending order according to the corresponding eigenvalues. Subsequently, the input time series dataset **X** is projected on the top d ($\ll$ N) eigenvectors corresponding to the largest d eigenvalues, known as the principal component (PC) eigenvectors. The eigenvector with the largest eigenvalue captures the direction in which the variance in the input dataset is maximum and the corresponding projection of the input dataset on this eigenvector is called the first principal component (PC1).

2.4.2 Time-lagged independent component analysis (TICA)

Time-lagged Independent Component Analysis (TICA)[21–24] is another linear dimensionality reduction technique. But unlike PCA, TICA tries to capture the slow processes from the input time series data **X**. It is particularly useful in the context of MD simulation where understanding the slow processes reveals information about the conformational changes and functional mechanisms of biomolecules. Mathematically, TICA works by first constructing an instantaneous covariance matrix C_{ij}^{0} like PCA (Eq. (2.18)) and an additional time-lagged covariance matrix C_{ij}^{τ} for a pre-defined lag-time τ as follows:

$$C_{ij}^{\tau} = \langle (x_i(t_0) - \langle x_i \rangle)\big(x_j(t_0 + \tau) - \langle x_j \rangle\big) \rangle_{t_0}, \tag{2.19}$$

where τ denotes the lag-time and $\langle \dots \rangle_{t_0}$ represents ensemble average over different time origins t_0. Then, the general eigenvalue problem $C^{\tau} v_i = \lambda_i C^0 v_i$ is solved to find the orthogonal eigenvectors v_i. Subsequently, the input time series dataset **X** is projected on the top d ($\ll$ N) eigenvectors corresponding to the largest d eigenvalues to obtain the time-lagged independent components. The eigenvector corresponding to the largest eigenvalue captures the direction of the slowest process.

2.4.3 Variational autoencoder (VAE)

Variational autoencoder (VAE)[25,26] is a neural network based deep learning model used to generate low dimensional latent space from high dimensional data. VAE comprises an encoder and a decoder as shown in Figure 2.1. Unlike normal autoencoder, VAE encodes the input data as a distribution over the latent space instead of encoding to deterministic single points to avoid overfitting or memorization. The loss function of VAE consisted of two parts: a reconstruction loss term as shown in Eq. (2.20) is calculated as the mean squared error (MSE) between the original and reconstructed data, and the regularization term defined as the Kullback-Leibler (KL) divergence[27] between the distribution produced by the encoder and a standard prior gaussian distribution as shown in Eq. (2.21). VAE loss function is mathematically expressed in Eq. (2.22) as a sum of these two terms.

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - f_{\varphi}(g_{\theta}(x_i))\right)^2, \tag{2.20}$$

$$\mathcal{L}_{KL} = - \int q_{\theta}(z|x_i) \log\left(\frac{p(z)}{q_{\theta}(z|x_i)}\right) dz\,, \tag{2.21}$$

$$\mathcal{L}_{VAE} = \mathcal{L}_{MSE} + \mathcal{L}_{KL}\ , \tag{2.22}$$

where x_i is the original input data, z is the latent variable, N is the number of samples, f_{φ} is the decoder function, g_{θ} is the encoder function, p(z) is the latent prior distribution and $q_{\theta}(z|x_i)$ is the latent posterior distribution generated by the encoder. If we want both of the distributions to be gaussian as well as the prior gaussian distribution with mean = 0 and variance = 1, then the KL-divergence ($\mathcal{L}_{KL}$) term in Eq. (2.21) becomes,

$$\mathcal{L}_{KL} = -\frac{1}{2}\left[\sum_{j=1}^{n} 1 + \log\left(\sigma_j^2\right) - \sigma_j^2 - \mu_j^2\right] \tag{2.23}$$

In Eq. (2.23), n is the number of latent variables, μ and σ^2 are the mean and variance of the latent variables, respectively.

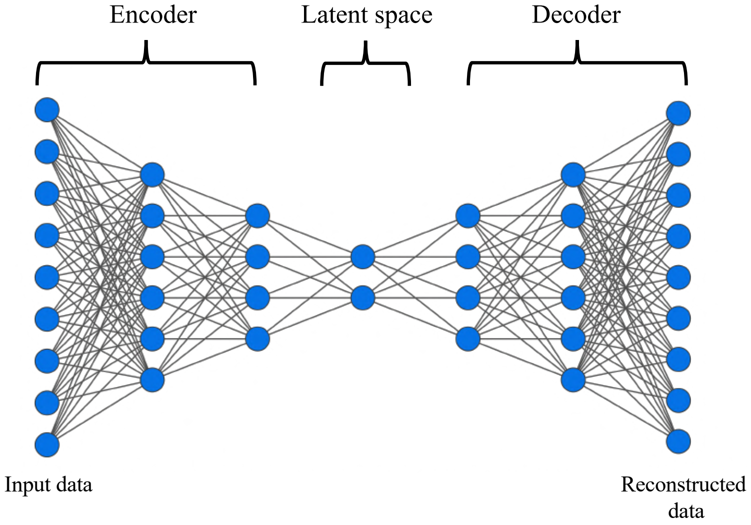


Figure 2.1 A basic autoencoder architecture.

2.5 Markov State Model

Markov State Model (MSM)[28–31] has made it possible to study the long time scale processes in biomolecular dynamics using short MD simulation trajectories. MSM discretizes MD simulation trajectories into a set of discrete metastable states, where transitions between the states are Markovian i.e memoryless. The dynamics is modelled using a transition probability matrix $\mathrm{T}(\tau)$ for a given lag-time τ, where $\mathrm{T_{ij}}(\tau)$ represents the conditional probability of finding the system in state j at time $\mathrm{t}+\tau$ provided the system was in state i at time t. However, choosing the correct lag-time τ for which the transitions show Markovianity is not trivial. To find the optimal lag-time, implied time scales $(\mathrm{t_i})$ for lag-time τ, $\mathrm{t_i} = -\frac{\tau}{\ln\lambda}$, where $\lambda < 1$ are the eigenvalues of the respective transition matrix $\mathrm{T}(\tau)$, are calculated. If the transitions are Markovian, $\mathrm{t_i}$ does not depend on the choice of τ. Therefore, convergence to Markovianity is tested by calculating $\mathrm{t_i}$ for various lag-times $\{\tau\}$.

The conventional pipeline[32,33] of building a MSM starts with featurization from molecular coordinates, followed by dimensionality reduction, grouping related conformations into hundreds of microstates using dimension-reduced data, estimate a MSM and finally coarse-grain the estimated MSM down to a few metastable states using Robust Perron Cluster Analysis (PCCA+)[34] to obtain an easily interpretable model. However, recent developments in neural network based deep learning methods have made it possible to construct a single end-to-end framework for building MSM[35].

2.5.1 VAMPnets

Employing the variational approach for Markov processes using neural networks (VAMPnets)[35] for molecular kinetics is a deep learning model that aims to combine the whole MSM building pipeline in a single end-to-end framework and provides a fuzzy kinetic clustering directly from molecular coordinates by maximizing a VAMP variational score e.g VAMP2 score[36]. VAMPnets works by first transforming the molecular coordinates $\{\mathbf{x}_t\}$ to features or latent variables; $\boldsymbol{\chi}_0(\mathbf{x}) = (\boldsymbol{\chi}_{01}(\mathbf{x}), \boldsymbol{\chi}_{02}(\mathbf{x}), \boldsymbol{\chi}_{03}(\mathbf{x}), \ldots . \boldsymbol{\chi}_{0m}(\mathbf{x}))^{\mathrm{T}}$ and $\boldsymbol{\chi}_1(\mathbf{x}) = (\boldsymbol{\chi}_{11}(\mathbf{x}), \boldsymbol{\chi}_{12}(\mathbf{x}), \boldsymbol{\chi}_{13}(\mathbf{x}), \ldots . \boldsymbol{\chi}_{1m}(\mathbf{x}))^{\mathrm{T}}$ such that the dynamics in these variables are approximately governed by a linear operator matrix **K** as;

$$\mathbb{E}[\boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})] \approx \mathbf{K}^{\mathrm{T}}\, \mathbb{E}[\boldsymbol{\chi}_0(\mathbf{x}_t)], \tag{2.24}$$

where $\mathbb{E}$ stands for expectation values and τ is the lag-time. Given feature transformations $\boldsymbol{\chi}_o$ and $\boldsymbol{\chi}_1$, the following covariance metrices are constructed,

$$\mathrm{C}_{00} = \mathbb{E}_t\, [\boldsymbol{\chi}_0(\mathbf{x}_t)\, \boldsymbol{\chi}_0(\mathbf{x}_t)^{\mathrm{T}}] \tag{2.25}$$

$$\mathrm{C}_{01} = \mathbb{E}_t\, [\boldsymbol{\chi}_0(\mathbf{x}_t)\, \boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})^{\mathrm{T}}] \tag{2.26}$$

$$\mathrm{C}_{11} = \mathbb{E}_{t+\tau}\, [\boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})\, \boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})^{\mathrm{T}}] \tag{2.27}$$

The optimal **K** that minimizes the least square error $\mathbb{E}_t\, [\left\|\boldsymbol{\chi}_1(\mathbf{x}_{t+\tau}) - \mathbf{K}^{\mathrm{T}}\boldsymbol{\chi}_0(\mathbf{x}_t)\right\|^2]$ is $\mathbf{K} = \mathrm{C}_{00}^{-1}\, \mathrm{C}_{01}$. Now the remaining problem is to find the suitable transformations $\boldsymbol{\chi}_o$ and $\boldsymbol{\chi}_1$ employing the VAMP theorem. For any two linearly independent functions $\boldsymbol{\chi}_o(\mathbf{x})$ and $\boldsymbol{\chi}_1(\mathbf{x})$, the VAMP2 score is defined as the Frobenius norm of the matrix $\mathrm{C}_{00}^{-\frac{1}{2}}\, \mathrm{C}_{01}\, \mathrm{C}_{11}^{-\frac{1}{2}}$,

$$\text{VAMP2 score} = \left\|\mathrm{C}_{00}^{-\frac{1}{2}}\, \mathrm{C}_{01}\, \mathrm{C}_{11}^{-\frac{1}{2}}\right\|_{\mathrm{F}}^{2} \tag{2.28}$$

To find the optimize $\boldsymbol{\chi}_o$ and $\boldsymbol{\chi}_1$, the VAMP2 score is maximized using two neural network lobes in VAMPnets.

2.6 Protein-ligand binding energy calculation

In this thesis work, we have used two methods to calculate protein-ligand binding energy from MD simulation data[37–39]: **(A)** Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) and **(B)** Thermodynamic integration (TI).

2.6.1 Molecular Mechanics/Generalized Born Surface Area (MM/GBSA)

The binding free energy for a protein-ligand complex can be estimated as[40–44]

$$\Delta G_{\mathrm{bind}} = \langle G_{\mathrm{COM}} \rangle - \langle G_{\mathrm{REC}} \rangle - \langle G_{\mathrm{LIG}} \rangle \tag{2.29}$$

where COM = Complex (receptor + ligand), REC = Receptor, LIG = Ligand.

Each term to the right in Eq. (2.29) is given by

$$\langle G_{\mathrm{x}} \rangle = \langle E_{\mathrm{MM}} \rangle - \langle G_{\mathrm{SOL}} \rangle - \langle \mathrm{TS} \rangle \tag{2.30}$$

where SOL = Solvation, MM = Molecular Mechanics.

ΔG_{bind} can also be represented as,

$$\Delta G_{\mathrm{bind}} = \Delta H - T\Delta S. \tag{2.31}$$

In Eq. (2.31), ΔH corresponds to the enthalpy of binding, and $-T\Delta S$ to the change in conformational entropy associated with ligand binding. When the entropic term is dismissed, the computed value is the energetic part of free energy, which is usually sufficient for comparing relative binding free energies. We have not considered the entropic component of binding free energy in our works. Within the MM/GBSA framework, the binding energy is estimated as a difference between the effective solvation energy of the ligand in the complex with the receptor as compared to the ligand in aqueous solution environment. The solvation energy is computed using Generalized Born(GB) continuum electrostatic method for the polar part and surface area (SA) method for the non-polar part as follows:

$$\Delta H = \Delta E_{\mathrm{MM}} + \Delta G_{\mathrm{sol}} \tag{2.32}$$

$$\Delta G_{\mathrm{sol}} = \Delta G_{\mathrm{polar}} + \Delta G_{\mathrm{non-polar}} = \Delta G_{\mathrm{GB}} + \Delta G_{\mathrm{non-polar}} \tag{2.33}$$

$$\Delta G_{\text{non-polar}} = NP_{\text{TENSION}} \times \Delta SASA + NP_{\text{OFFSET}} \tag{2.34}$$

ΔE_{MM} corresponds to the molecular mechanical (MM) energy changes in the gas phase, and the non-polar component of the solvation energy is usually assumed to be proportional to the molecule's total solvent accessible surface area (SASA) with a proportionality constant derived from experimental solvation energies of small non-polar molecules.

2.6.2 Thermodynamic integration (TI)

In this method a coupling parameter λ (values lies between 0 and 1) is used to monitor the transition from state-1 with potential energy function U_1 (for example, ligand fully interacting with the surrounding environment) to state-2 with potential energy function U_2 (e.g no non-bonded interactions between the ligand and the surrounding environment i.e decoupled) according to Eq. (2.35)[45]

$$U(\lambda) = (1 - \lambda)U_1 + \lambda U_2 \ . \tag{2.35}$$

When λ = 0, the system is in state-1 and for λ = 1 the system is in state-2. A large number of intermediate non-physical states are created at different λ values. Then, for each intermediate state, ensemble average of the derivative of potential energy with respect to the coupling parameter λ i.e $\langle \frac{dU}{d\lambda} \rangle_\lambda$ is computed, where the brackets denote ensemble average. Free energy difference between the two end states is computed as follows[45–51]:

$$\Delta G = G(\lambda = 1) - G(\lambda = 0) = \int_0^1 \langle \frac{dU}{d\lambda} \rangle_\lambda \ d\lambda \ . \tag{2.36}$$

Therefore, the free energy difference is the area under the curve of $\langle \frac{dU}{d\lambda} \rangle_\lambda$ vs λ plot.

Ligand binding free energy is the difference between the solvation energy of ligand in water without any receptor and the solvation energy of ligand bound to the receptor. Final binding free energy is calculated as follows[52]:

$$\Delta G_{\text{bind}} = -\Delta G^{\text{prot}}_{\text{elec+vdw+rest}} + \Delta G^{\text{solv}}_{\text{elec+vdw}} + \Delta G^{\text{solv}}_{\text{rest\_on}} \tag{2.37}$$

where $\Delta G^{\text{prot}}_{\text{elec+vdw+rest}}$ = free energy change for ligand decoupling form receptor in the presence of restraints, $\Delta G^{\text{solv}}_{\text{elec+vdw}}$ = free energy change for ligand decoupling in water without the presence of receptor and $\Delta G^{\text{solv}}_{\text{rest\_on}}$ = free energy change to restraint the decoupled ligand in water without receptor.

$\Delta G_{\text{rest\_on}}^{\text{solv}}$ is calculated using the following equation[53],

$$\Delta G_{\text{rest\_on}}^{\text{solv}} = \text{RTln}\left[\frac{8\pi^2 V^0}{r_0^2 \sin\theta_A \sin\theta_B} \frac{\left(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C}\right)^{1/2}}{(2\pi kT)^3}\right]. \tag{2.38}$$

In Eq. (2.38), R = Ideal gas constant, T = Temperature (300K), r_0 = reference distance for restraints applied to ligand, θ_A and θ_B are reference angles for restraints , K_x are force constants for one distance (r_0), two angles (θ_A and θ_B) and three dihedral (ϕ_A , ϕ_B, ϕ_C) restraints and V^0 = volume corresponding to one molar standard state = 1.66 nm^3.

2.7 Enhanced sampling methods

MD is plagued by a time scale problem. The integration time step in MD, a few femtoseconds, is many orders of magnitude small than the timescale of the biophysical processes of interest. Usually relevant conformations or metastable states of a biomolecule belong to different free energy basins separated by high energy barriers. Therefore, in practice, normal MD simulation remains stuck in one of the energy basins for a long time. Several enhanced sampling methods have been developed over the years for the quick exploration of the free energy surface as well as enable the calculations of thermodynamic and kinetic properties of the system[54–57]. In this thesis work, we have used two different classes of enhanced sampling methods: (A) **Metadynamics**[58–60] and (B) **Weighted Ensemble**[61,62] path sampling. In the following sections, we provide a brief theoretical description of these methods.

2.7.1 Metadynamics

Metadynamics belongs to a class of enhanced sampling methods where a bias potential or force is applied along some predefined degrees of freedom called collective variables (CVs) to enhance the sampling. In the earlier versions of metadynamics, a history-dependent repulsive bias as a function of the CVs is added to the Hamiltonian of the system. Let's define a set of n CVs, $S(\mathbf{R})$ as a function of molecular coordinates $\mathbf{R}$ as

$$S(\mathbf{R}) = (S_1(\mathbf{R}), S_2(\mathbf{R}), S_3(\mathbf{R}), \ldots S_n(\mathbf{R})). \tag{2.39}$$

At time t, the metadynamics bias potential $V_G(S, t)$ is represented by a sum of gaussians deposited along the CVs and can be written as[63]

$$V_G(S,t) = \int_0^t dt' \, \omega \exp\left(-\sum_{i=1}^{n} \frac{\left(S_i(\mathbf{R}) - S_i\left(\mathbf{R}(t')\right)\right)^2}{2\sigma_i^2} \right). \qquad (2.40)$$

In Eq. (2.40), σ_i is the width of the Gaussian for the ith CV, ω is defined as $\omega = \frac{W}{\tau_G}$, W is the Gaussian height and τ_G is the stride rate. This repulsive bias potential discourages the system from revisiting the conformations that have been already sampled.

Recently, on-the-fly probability enhanced sampling (OPES)[64,65] method has been developed as a successor of metadynamics.

2.7.2 Weighted ensemble path sampling

Weighted ensemble (WE) path sampling[62,66,67] is a special kind of enhanced sampling methods where no external bias potential has been deposited along the CVs like metadynamics, instead an ensemble of trajectories are run simultaneously to enhance the sampling. Therefore, WE is unbiased in the sense that this method does not modify the underlying equilibrium distribution. In the conventional WE protocol, the configurational space is mapped onto a low dimensional CV space that describes the transition of interest, followed by dividing it into bins (say, M bins) with a target allocation of n trajectories per bin. Each trajectory is assigned with a weight or probability w_t^k at any time t such that the weights are normalized: $\sum_k w_t^k = 1$. Initially n trajectories are started from a single bin pre-assigned with weights $w_0^k = \frac{1}{n}$. Next the trajectories are propagated forward in time for a time interval τ according to the natural dynamics of the system. Trajectories that reach new bin are "cloned" or "split" to reach the maximum allocation of n trajectories in that bin. Conversely, if the number of trajectories in each bin exceeds the maximum allocation, then some trajectories are pruned with their survival probability proportional to their weights; this process is called "merging". The sum of the weights in each bin is always maintained in the trajectory resampling (cloning + merging) processes. In this way, the simulations evolve under the natural dynamics of the system with trajectory resampling for many iterations to enhance the sampling of rare events. Furthermore, WE setup also uses a "recycling" boundary condition at the target state to achieve non-equilibrium steady state and computation of rate constant of any processes.

To calculate the first order rate constant of transition between the two states A and B, the Hill relation between the flux and mean first passage time (MFPT) is used as follows[62]:

$$\mathrm{MFPT}_{\mathrm{A}\rightarrow\mathrm{B}}[\rho_{\mathrm{A}}] = \frac{1}{\mathrm{Flux}(\mathrm{A}\rightarrow\mathrm{B};\mathrm{SS})} \tag{2.41}$$

where $\mathrm{Flux}(\mathrm{A} \rightarrow \mathrm{B}; \mathrm{SS})$ is the probability per unit time to arrive at state B in A-to-B non-equilibrium steady state condition and ρ_{A} is the distribution of trajectory starting points within state A. Rate constant is equal to the inverse of MFPT of the process.

References:

1 K. D. Fong, H. K. Bergstrom, B. D. McCloskey and K. K. Mandadapu, *AIChE Journal*, 2020, **66**, e17091.

2 L. Onsager, *Physical Review*, 1931, **37**, 405.

3 K. D. Fong, J. Self, B. D. McCloskey and K. A. Persson, *Macromolecules*, 2020, **53**, 9503–9512.

4 K. D. Fong, J. Self, B. D. McCloskey and K. A. Persson, *Macromolecules*, 2021, **54**, 2575–2591.

5 N. M. Vargas-Barbosa and B. Roling, *ChemElectroChem*, 2020, **7**, 367–385.

6 L. Van Hove, *Physical Review*, 1954, **95**, 249.

7 C. Donati, J. F. Douglas, W. Kob, S. J. Plimpton, P. H. Poole and S. C. Glotzer, *Phys Rev Lett*, 1998, **80**, 2338.

8 B. Wu, T. Iwashita and T. Egami, *Phys Rev Lett*, 2018, **120**, 135502.

9 Y. Shinohara, A. S. Ivanov, D. Maltsev, G. E. Granroth, D. L. Abernathy, S. Dai and T. Egami, *Journal of Physical Chemistry Letters*, 2022, **13**, 5956–5962.

10 Y. Shinohara, R. Matsumoto, M. W. Thompson, C. W. Ryu, W. Dmowski, T. Iwashita, D. Ishikawa, A. Q. R. Baron, P. T. Cummings and T. Egami, *Journal of Physical Chemistry Letters*, 2019, **10**, 7119–7125.

11 R. A. Matsumoto, M. W. Thompson, V. Q. Vuong, W. Zhang, Y. Shinohara, A. C. T. van Duin, P. R. C. Kent, S. Irle, T. Egami and P. T. Cummings, *J Chem Theory Comput*, 2021, **17**, 5992–6005.

12 H. R. Sánchez, *Journal of Physical Chemistry B*, 2022, **126**, 8804–8812.

13 P. Kubisiak, P. Wróbel and A. Eilmes, *Journal of Physical Chemistry B*, 2020, **124**, 413–421.

14 D. C. Rapaport, *Mol Phys*, 1983, **50**, 1151–1162.

15 K. Chahara, T. Ohno, M. Kasai, Y. Kozono, R. Von Helmolt, J. Wecker, B. Holzapfel, L. Schultz, K. Samwer and S. Jin, *Nature 1996 379:6560*, 1996, **379**, 55–57.

16 E. M. Kirova and G. E. Norman, *J Phys Conf Ser*, 2015, **653**, 012106.

17 Y. Zhang, A. Otani and E. J. Maginn, *J Chem Theory Comput*, 2015, **11**, 3537–3546.

18 B. Hess, *J Chem Phys*, 2002, **116**, 209–217.

19 M. A. Balsera, W. Wriggers, Y. Oono and K. Schulten, *Journal of Physical Chemistry*, 1996, **100**, 2567–2572.

20 C. C. David and D. J. Jacobs, *Methods in Molecular Biology*, 2014, **1084**, 193–226.

21 L. Molgedey and H. G. Schuster, *Phys Rev Lett*, 1994, **72**, 3634.

22 G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, *Journal of Chemical Physics*.

23 S. Schultze and H. Grubmüller, *J Chem Theory Comput*, 2021, **17**, 5766–5776.

24 C. R. Schwantes and V. S. Pande, *J Chem Theory Comput*, 2013, **9**, 2000–2009.

25 D. P. Kingma and M. Welling, *Foundations and Trends in Machine Learning*, 2019, **12**, 307–392.

26 Z. Belkacemi, M. Bianciotto, H. Minoux, T. Lelièvre, G. Stoltz and P. Gkeka, *Journal of Chemical Physics*, 2023, **159**, 24122.

27 D. I. Belov and R. D. Armstrong, *British Journal of Mathematical and Statistical Psychology*, 2011, **64**, 291–309.

28 B. E. Husic and V. S. Pande, *J Am Chem Soc*, 2018, **140**, 2386–2396.

29 V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, **52**, 99–105.

30 J. D. Chodera and F. Noé, *Curr Opin Struct Biol*, 2014, **25**, 135–144.

31 R. D. Malmstrom, C. T. Lee, A. T. Van Wart and R. E. Amaro, *J Chem Theory Comput*, 2014, **10**, 2648–2657.

32 G. R. Bowman, 2014, 7–22.

33 W. Wang, S. Cao, L. Zhu and X. Huang, *Wiley Interdiscip Rev Comput Mol Sci*.

34 S. Röblitz and M. Weber, *Adv Data Anal Classif*, 2013, **7**, 147–179.

35 A. Mardt, L. Pasquali, H. Wu and F. Noé, *Nature Communications 2018 9:1*, 2018, **9**, 1–11.

36 H. Wu and F. Noé, *J Nonlinear Sci*, 2020, **30**, 23–66.

37 H. J. Woo and B. Roux, *Proc Natl Acad Sci U S A*, 2005, **102**, 6825–6830.

38 M. K. Gilson and H. X. Zhou, *Annu Rev Biophys Biomol Struct*, 2007, **36**, 21–42.

39 H. Gouda, I. D. Kuntz, D. A. Case and P. A. Kollman, *Biopolymers*, 2003, **68**, 16–34.

40 M. S. E. Valdés-Tresanco, M. S. E. Valdés-Tresanco, P. A. Valiente and E. Moreno, *J Chem Theory Comput*, 2021, **17**, 6281–6291.

41 T. Hou, J. Wang, Y. Li and W. Wang, *J Chem Inf Model*, 2011, **51**, 69–82.

42 J. Srinivasan, J. Miller, P. A. Kollman and D. A. Case, *J Biomol Struct Dyn*, 1998, **16**, 671–682.

43 E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. H. Zhang and T. Hou, *Chem Rev*, 2019, **119**, 9478–9508.

44 P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Acc Chem Res*, 2000, **33**, 889–897.

45 T. P. Straatsma and H. J. C. Berendsen, *J Chem Phys*, 1988, **89**, 5876–5886.

46 S. Bruckner and S. Boresch, *J Comput Chem*, 2011, **32**, 1320–1333.

47 G. Hummer, *J Chem Phys*, 2001, **114**, 7330–7337.

48 M. J. Mitchell and J. A. McCammon, *J Comput Chem*, 1991, **12**, 271–275.

49 C. X. Wang, Y. Y. Shi, F. Zhou and L. Wang, *Proteins: Structure, Function, and Bioinformatics*, 1993, **15**, 5–9.

50 M. Lawrenz, R. Baron, Y. Wang and J. A. McCammon, *Methods in Molecular Biology*, 2012, **819**, 469–486.

51 M. Aldeghi, A. Heifetz, M. J. Bodkin, S. Knapp and P. C. Biggin, *Chem Sci*, 2015, **7**, 207–218.

52 P. V. Klimovich, M. R. Shirts and D. L. Mobley, *J Comput Aided Mol Des*, 2015, **29**, 397–411.

53 S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *Journal of Physical Chemistry B*, 2003, **107**, 9535–9551.

54 R. C. Bernardi, M. C. R. Melo and K. Schulten, *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2015, **1850**, 872–877.

55 Y. I. Yang, Q. Shao, J. Zhang, L. Yang and Y. Q. Gao, *Journal of Chemical Physics*, 2019, **151**, 70902.

56 F. Sohraby and A. Nunes-Alves, *Trends Biochem Sci*, 2023, **48**, 437–449.

57 S. Mehdi, Z. Smith, L. Herron, Z. Zou and P. Tiwary, *Annu Rev Phys Chem*, 2024, **75**, 347–370.

58 A. Laio and F. L. Gervasio, *Reports on Progress in Physics*, DOI:10.1088/0034-4885/71/12/126601.

59 G. Bussi and A. Laio, *Nature Reviews Physics*, 2020, **2**, 200–212.

60 A. Barducci, G. Bussi and M. Parrinello, *Phys Rev Lett*, 2008, **100**, 020603.

61 G. A. Huber and S. Kim, *Biophys J*, 1996, **70**, 97–110.

62 D. M. Zuckerman and L. T. Chong, *Annu Rev Biophys*, 2017, **46**, 43–57.

63 A. Barducci, M. Bonomi and M. Parrinello, *Wiley Interdiscip Rev Comput Mol Sci*, 2011, **1**, 826–843.

64 M. Invernizzi and M. Parrinello, *Journal of Physical Chemistry Letters*, 2020, **11**, 2731–2736.

65 V. Rizzi, S. Aureli, N. Ansari and F. L. Gervasio, *J Chem Theory Comput*, 2023, **19**, 5731–5742.

66 D. Aristoff, J. Copperman, G. Simpson, R. J. Webber and D. M. Zuckerman, *Journal of Chemical Physics*.

67 B. W. Zhang, D. Jasnow and D. M. Zuckerman, *Journal of Chemical Physics*.

3

Exploring the capabilities and limitations of the Van Hove function to understand directional correlations in ion movements within Li-ion battery electrolytes

3.1 Introduction

Lithium-ion batteries (LIB) are powering our modern society with its widespread use in electric devices and vehicles because of their high energy density and long life cycle[1,2]. One of the components in a LIB cell, the electrolyte, plays the critical role in proper functioning of LIB by enabling ion transport between the electrodes[3,4]. Usually, LIB energy materials contain lithium salt at 1 M concentration[5]. Interestingly, a few experimental studies[6,7] have shown good thermal and reactive stabilities of different electrolytes at higher salt concentrations (> 1 M), fueling research on such concentrated electrolytes[8,9]. Ionic conductivity of an electrolyte is the most important property that one aims at tuning for a better battery designing purpose. Ionic conductivity can be calculated from the Nernst-Einstein (NE) equation using the self-diffusion coefficients of cation and anion derived from either NMR experiments[10] or MD simulations[11]. However, the NE conductivity is only a good approximation at very dilute salt concentrations. This is because the NE formulation does not contain the effects of directional correlation between the movement of ions[12]. At moderate to higher salt concentrations, motion of a particular ion is affected by the movement of other ions because of the strong electrostatic interactions between them[13]. As discussed in section 2.1 of Chapter 2, if ions of the same or opposite charges move into the same direction, then correlation between their movement is said to be positive. If ions move in opposite directions, their relative motions become anti-correlated[14]. In those scenarios, different correlations between the movement of ions have significant effect on the mechanism of ion transport in electrolyte solutions and the final ionic conductivities. This necessitates the need of a molecular level understanding of the directional correlation in ion transport. Fortunately, MD simulations provide a way to build microscopic understanding of various phenomena employing the well-established theories of statistical mechanics[15].

Onsager transport equation (Eq. (2.1)) provides a powerful framework for analysing transport in concentrated electrolyte solutions from the relevant MD simulation data[16]. This equation contains Onsager transport coefficients, L^{ij}, which capture the aforementioned directional correlations in ion movements. These coefficients can be used to split the total conductivity into individual components arising from a variety of correlated motions between ions, and the signs of these conductivity components reveal the nature of correlations and their effects on the final conductivity. These transport coefficients can also be derived from experimentally measured quantities (e.g self-diffusion coefficients)[17]. This framework was utilised earlier to understand transport in polymer based electrolytes, polyelectrolyte solutions and polymerized ionic liquids.[18,19] A few nonintuitive phenomena, such as, negative cation transference number and anti-correlated cation-anion motion were discussed in those works. Ion correlations in BMIM/BF_4 based electrolytes were also studied.[20] This work showed that ion dynamics in highly concentrated electrolyte solutions were dominated by anti-correlated motions of ions (same or opposite charge types) and the nature of correlation changes from dilute electrolyte solution to the pure ionic liquid limit. This approach was also used to explore correlations associated with ion transport in LiTFSI/EMIM-TFSI and NaTFSI/EMIM-TFSI based electrolytes,[21] and the authors found that positive correlation between cations and anions decreased conductivity. Furthermore, this approach was used to describe transport in aqueous electrolyte solutions[22]. A somewhat similar approach was employed[23] to separate the net conductivity into different individual components where velocity correlation functions between distinct ions of the same or opposite charges were examined to explain ion transport in ionic liquids and electrolyte solutions. Several other studies have also employed the Onsager transport coefficient framework to understand ion correlation in a variety of systems[24–27]. A recent study discusses good practices of computing ionic conductivity from MD simulation data[28]. In spite of the success of this theory to elucidate directional correlations associated with ion movement, computing Onsager transport coefficients from MD simulation data is a non-trivial exercise. This is mainly because of the following two reasons: (i) extraction of the Onsager transport coefficients from simulations requires large scale simulation data [29] and (ii) chance of getting large errors in the reported values is high[29].

On the other hand, the Van Hove function[30] (VHF) expressed in Eq. (2.11), is also capable of capturing correlated motion between particles. Because VHF is a time dependent probability distribution function, it is far easier to calculate than the Onsager transport coefficients, L^{ij} , from MD simulation data. Interestingly, collective dynamics of ions in molten inorganic salts[31] and water-ion/ion-ion interactions in aqueous solutions[32,33] have already been studied through the Van Hove function. In addition, the correlated motion involving different anions in polymerized ionic liquids has also been studied via analysing the simulated distinct part of the VHF.[34]

In this chapter, we have discussed the capabilities and limitations of using the VHF to identify correlations in ion movement and validate the predictions against the results obtained from the Onsager's theory by using all atom classical MD simulations of lithium hexafluorophosphate ($LiPF_6$) solutions in the mixture of ethyl methyl carbonate (EMC) and fluoroethylene carbonate (FEC) at different $LiPF_6$ concentrations at 313 K. The solvent mixture was prepared with 9:1 weight-percentage (wt%) ratio of EMC and FEC respectively. These choices of this solvent mixture and the electrolyte ($LiPF_6$) were motivated by the existing experimental study[35]. Figure 3.1 represents the schematic diagrams of the ions and the solvent species considered in this work.

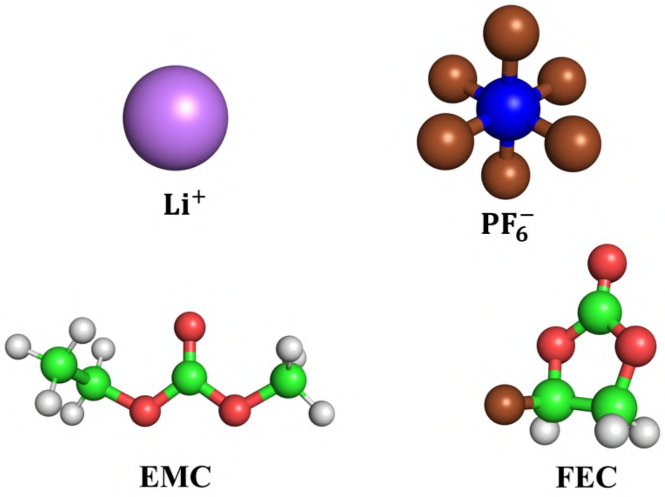


Figure 3.1: Li^+ , PF_6^- ions and solvent molecules are shown in ball-stick representation. O (oxygen) atoms are in red, C (carbon) atoms in green, P (phosphorus) atom in blue, F (fluorine) atoms in brown, and H (hydrogen) atoms in silver colour.

3.2 Computational details

All atom classical MD simulations were performed for electrolyte solutions with varying salt concentration. The number of cations, anions and solvent molecules used are given in Table 3.1. We used General Amber Force Field (GAFF)[36] to model the interactions in our systems. This is because GAFF demonstrated a good performance in describing transport properties of electrolyte solutions[37]. Force field parameters for Li^+ ions and non-bonded parameters for PF_6^- anions were taken from an existing work[38]. Bonded parameters for PF_6^- ions were also taken from an work reported earlier[39]. All partial charges of the ions were scaled using an existing methodology[40] in order to include the effects of electronic polarizability. Atomic partial charges of the solvent molecules were derived by first optimizing their structure using the Becke's three-parameter exchange function combined with the Lee-Yang-Parr correlational functional (B3LYP)[41–43] at the level of aug-cc-pvdz theory using the Gaussian 16 package[44] and then fitting the electrostatic potential surface employing the RESP method[45]. Topology files for the solvent molecules were generated using the LEaP program in the AmberTools2022[46] and converted to GROMACS format (.gro & .top) via the parmed utility of the AmberTools2022.

All the simulations were performed using the GROMACS v2021.5 software package[47]. The initial configurations of molecules in the box of 45 Å × 45 Å × 45 Å dimensions were generated via the PACKMOL software[48]. These initial configurations were first energy minimized using the steepest descent method and then equilibrated for 3ns in the NPT ensemble at a temperature of 313K and pressure of 1 atm. Temperature and pressure were controlled through the velocity rescaling method[49] with a time constant of 0.1 ps and the Berendsen barostat[50] with a time constant of 1 ps. All simulations were performed under the periodic boundary conditions, and the long range electrostatic interactions were dealt with by using the particle mesh Ewald (PME)[51] summation method. The cut-off distances for the electrostatic and the van der Waals interactions were set to 12 Å. Bonds containing hydrogen atoms were constrained with LINCS algorithm[52]. Newton's equations of motion were numerically integrated using the leap-frog algorithm. Electrolyte concentrations were calculated based on the average box volumes after the initial equilibration in the NPT ensemble (see Table 3.1). Final production run of 405 ns for each of the electrolyte concentrations studied was conducted in the NVT ensemble starting from these 'NPT-equilibrated' solution structures. Frames were saved at every 0.5 ps interval. First 5 ns of each trajectory was considered as further equilibration of the system in the NVT ensemble and as a result, only the last 400 ns of each trajectory was used for data analysis. All of the trajectories were processed using the MDAnalysis python package[53,54] for further analysis.

Table 3.1: Number of cations (Li^+), anions (PF_6^-) and solvent molecules (EMC and FEC) used in this study are given below. Salt concentrations (in molarity (M) and molality (m) units) for each simulation box were calculated after initial NPT equilibration.

Number of Li^+	Number of PF_6^-	Number of EMC	Number of FEC	Salt concentration	
				Molarity (M)	Molality (m)
11	11	260	28	0.36	0.37
25	25	260	28	0.80	0.83
37	37	260	28	1.15	1.23
47	47	260	28	1.45	1.57
60	60	260	28	1.80	2.00
80	80	260	28	2.32	2.67
100	100	260	28	2.80	3.33

3.3 Results and discussions

3.3.1 Validation of the force field parameters

Before exploring the capabilities and limitations of the VHF in describing the correlations in ion dynamics, it is essential to validate the forcefield parameters employed in this study. As a first step, we compared our simulated $LiPF_6$ concentration dependent ionic conductivities at 313 K with those from experiments[35]. The comparison is shown in Figure 3.2. Predicted ionic conductivity values were calculated using the Onsager transport coefficients from our simulation data using Eq. (2.8). Notice in Figure 3.2 that the simulated conductivities agree well with those from experiments at low $LiPF_6$ concentrations, but gradually become lower (than experiments) upon the increase of the salt concentration. This is the demerit of using non-polarizable force field[55]. Nevertheless, this non-polarizable force field (GAFF) with scaled charges capture qualitatively correctly the experimental non-monotonic $LiPF_6$ concentration dependence of the solution conductivity. Total conductivity increases initially, reaches a maximum at around 1.5 M salt concentration and then decreases. This type of non-monotonic behaviour of conductivity with varying salt concentrations has also been reported in several previous studies[56–59]. Since the use of polarizable force field is computationally highly expensive[60] and the chosen parameters can capture the trend (concentration dependence of conductivity) correctly, we used these trajectories for further analysis.

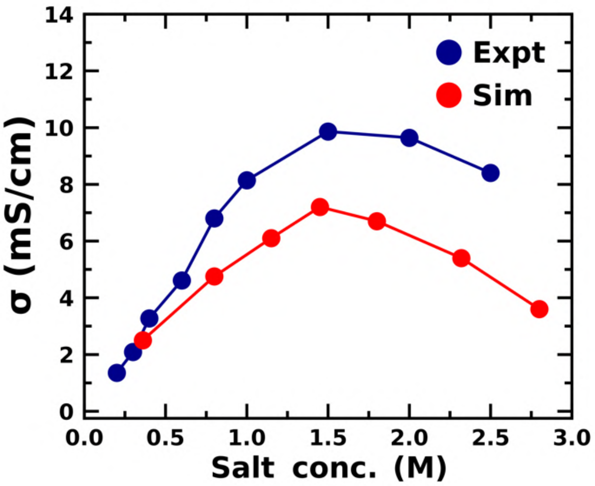


Figure 3.2: Comparison between the simulated (“Sim”) and experimental (“Expt”) ionic conductivities (σ) at different $LiPF_6$ concentrations at 313K temperature.

3.3.2 Capturing ion-ion correlations using the Onsager transport coefficients

In the previous section 3.3.1, we have examined the utility of the force filed parameters employed by comparing the simulated ionic conductivities with those from experiments. As we focus on investigating the capabilities and limitations of the VHF to study correlations in ion movements in electrolytes, we would examine both positively correlated and anti-correlated ion movements in our electrolyte solutions. In this section we explore these correlations by using the Onsager transport coefficient framework.

We calculated the individual components of the total conductivity using the Onsager transport coefficients, L^{++} , L^{--} and L^{+-} from Eqs. (2.2), (2.3) and (2.4). We first divided each 400 ns long trajectories into 40 equal parts of 10 ns. All these individual segments were treated as independent trajectories. We then calculated the displacements of ionic species for each of these independent trajectories. In order to capture the true diffusive transport, slopes should be calculated when term in the angular bracket of Eqs. (2.2), (2.3) and (2.4) are linear with respect to time[61]; that is, $\sum_\alpha \sum_\beta \left\langle [\vec{r}_i^{\alpha}(t) - \vec{r}_i^{\alpha}(0)] \bullet \left[\vec{r}_i^{\beta}(t) - \vec{r}_i^{\beta}(0)\right]\right\rangle \propto t^\delta$ and $\sum_\alpha \langle [\vec{r}_i^{\alpha}(t) - \vec{r}_i^{\alpha}(0)]^2 \rangle \propto t^\delta$ with $\delta = 1$. As shown in the representative plot in Figure 3.A.1, MSD curves for each of the independent short trajectories were not well-behaved. Thus, it was very difficult to find a significant portion of the linear regime for fitting. Therefore, instead of calculating slopes for each of these individual MSD curves, we performed averaging over all the MSD curves (obtained from independent trajectories) and the averaged results are shown in Figure 3.A.2. This exercise generated a relatively smoother MSD curves for each of the salt concentration studied here. Slope of the lines after averaging were used to calculate L^{++} , L^{--}, L^{+-}, L^{++}_{self} and L^{--}_{self} by using Eqs. (2.2), (2.3) and (2.4). For all the simulations in this work, Onsager coefficients were calculated for time windows when δ was between 0.95 to 1.02. This ensured fitting in the linear regime of the MSD plots. After calculating the coefficients, different components of the correlated ionic conductivity (σ) were determined using Eq. (2.8). The results are shown in Figure 3.3a. Both $\sigma^{++}_{distinct}$ and $\sigma^{--}_{distinct}$ share the same sign at every salt concentration. Thus, we choose to plot the ratios of different components of the conductivity to the net conductivity, $(\sigma^{++}_{distinct} + \sigma^{--}_{distinct})/\sigma$ and $2\sigma^{+-}/\sigma$ in Figure 3.3b at low (0.36 M), medium (1.45 M) and high (2.80 M) $LiPF_6$ concentrations. Figure 3.3b shows that at low electrolyte concentrations, the signs of the components arising from the correlations of ions of the same charge-type $(\sigma^{++}_{distinct} + \sigma^{--}_{distinct})$, and from cation-anion correlation $(2\sigma^{+-})$ are positive. As discussed in section 2.1 of Chapter 2, positive sign of the conductivity components indicates a positively correlated ionic motions between cation-cation, anion-anion and cation-anion. Therefore, at low $LiPF_6$ concentration, all of the three types of motions are positively correlated. It can also be noted that cation-anion positive correlation renders the strongest impact among all in modulating the net conductivity at low $LiPF_6$ concentrations.

We further calculated the inverse Haven ratios $H^{-1} = \frac{\sigma}{\sigma_{NE}}$, where σ is the correlated ionic conductivity and σ_{NE} is the ideal Nernst-Einstein ionic conductivity. Thus, when the effects of different ion correlations are mutually cancelled, that is, when $\sigma \approx \sigma_{NE}$, H^{-1} approaches unity. In Figure 3.3c, we plot the inverse Haven ratios H^{-1} and net ionic conductivities at different salt concentrations. Notice that the net ionic conductivity is maximum when H^{-1} is nearest to unity, that is, when the combined effect of all the different ion correlations is the least.

Positive correlation in cation-anion dynamics at low $LiPF_6$ concentration indicates the formation of contact ion pairs (CIPs) even in dilute electrolyte solutions. But why? At low electrolyte concentrations, electrolyte-dissociation is supposed to be complete and the resultant ions are expected to be properly solvated by the solvent molecules in solutions. To investigate this aspect, we defined three types of Li^+ association with anions in its first solvation shell (for example, zero anion, one anion and ≥ 2 anions) based on $Li^+ - P(PF_6^-)$ distance[62,63]. If the distances between a Li^+ ion and P atoms in all PF_6^- anions are greater than the position of the first minima in the corresponding radial distribution functions (RDFs) of $Li^+ - P(PF_6^-)$, which is 4.1 Å (see Figure 3.A.3), then no anions are present in the first solvation shell of this particular Li^+ ion. If the distance between a Li^+ ion and P atom of exactly one PF_6^- ion is less than 4.1 Å , then only one anion is present in the first solvation shell of this Li^+ ion. This cation is then considered to form CIP. Cations with more than one anion in its first solvation shell is part of an aggregate (cluster of cation and anions). The results for the lowest $LiPF_6$ concentration (0.36 M) are shown in Figure 3.3d and those for solutions with varying LiPF_6 concentration is shown in Figure 3.A.4. It is shown that at this lowest electrolyte concentration, most of the cations formed CIPs (aggregates ~ 5%). It is worth mentioning that, unlike dilute electrolyte solutions, ions in concentrated solutions form aggregates. Thus, it is somewhat confusing to identify CIPs in the presence of aggregates by using this simple distance cutoff-based definition[64]. However, this kind of CIP identification method works properly when the percentage of aggregates are negligible in the solution. In this study, at this lowest salt concentration, population of aggregates in the system is ~ 5%. At this low salt concentration, the distance-based criterion to identify CIPs worked correctly. However, with increasing salt concentration, aggregate population increases rapidly (see Figure 3.A.4). Therefore, in the presence of significant number of aggregates, this algorithm may face problem to distinguish between CIPs and aggregates.

EMC has low dielectric constant (~ 3 at 25°C)[65] and FEC has lower Li-ion solvation ability than EMC due to the presence of F atom[66]. Thus, choosing EMC and FEC at 9:1 ratio causes poor Li-ion solvation and the presence of CIPs even at the lowest LiPF_6 concentration studied here. Electrostatic attraction between the ions of opposite charges in CIP are the reason for their positively correlated motion at low LiPF_6 concentrations.

With increasing LiPF_6 concentration, sign of these different components changes. Sign of $(\sigma^{++}_{\text{distinct}} + \sigma^{--}_{\text{distinct}})\,/\sigma$ are changing from positive to negative and the impact of positive correlation between cation-anion diminishes with increasing LiPF_6 concentration. This indicates that the nature of the correlated ion motions is changing upon increasing LiPF_6 concentration. Negative signs of the components arising from the correlations between ions of same charge-type indicate anti-correlated motion between cation-cation, anion-anion at high LiPF_6 concentrations. Cation-anion correlation is still positive at high salt concentration although it is changing gradually to negative correlation. Therefore, we observe here that the positive motional correlations between ions of the same and different charge-types that characterize the solutions at low LiPF_6 concentrations vanish upon increasing LiPF_6 concentration and eventually become anti-correlated at high concentrations[20]. In addition, we also note that at high LiPF_6 concentrations, the anti-correlated dynamics of ions of the same charge-type (cation-cation and anion-anion) imparts a stronger impact on the net conductivity than the cation-anion correlation. It is therefore clear the system being studied here is dynamically complex and offers an opportunity to study the impact of electrolyte concentration dependent positive and anti-correlated ion dynamics on net conductivity. Consequently, this system is suitable to explore the capabilities and limitations of the VHF to identify the directional correlations in ion movements.

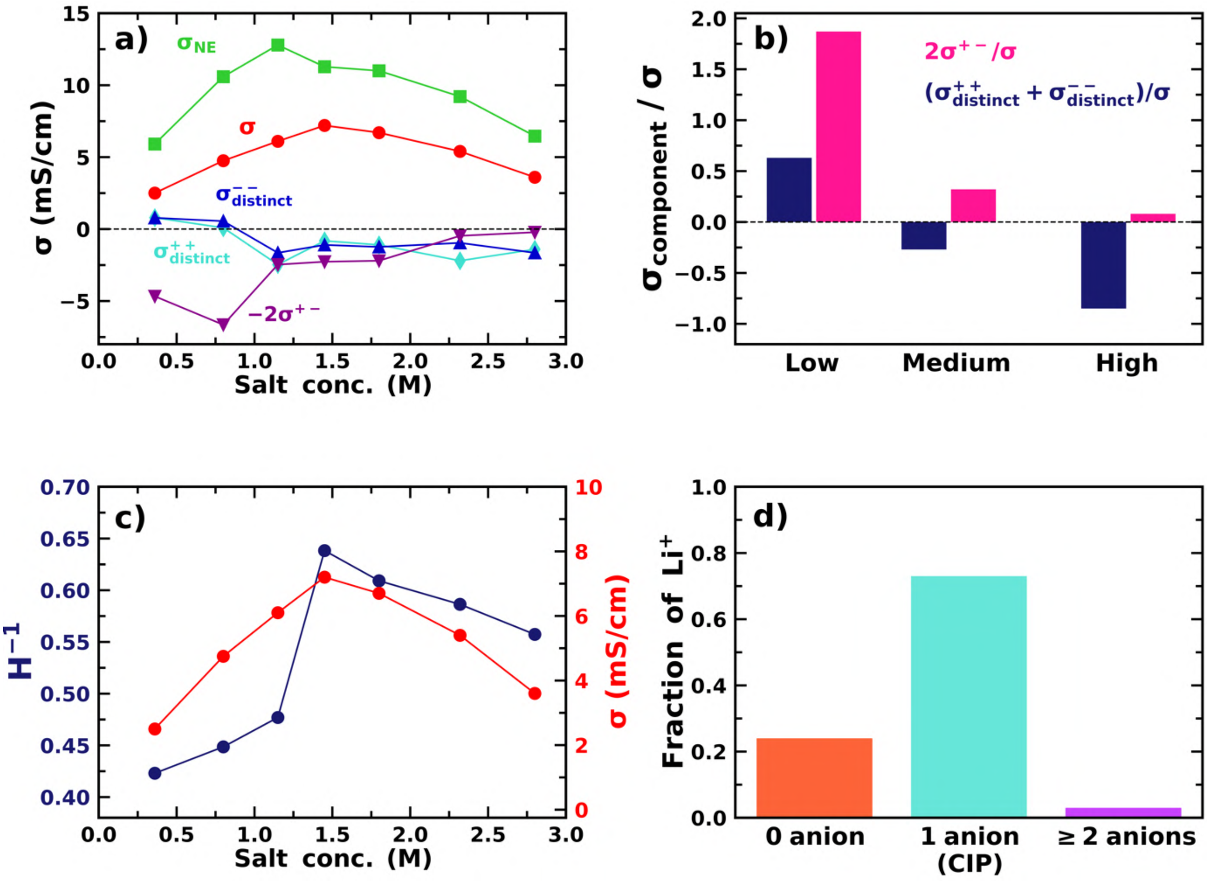


Figure 3.3: **a)** Variation of the net ionic conductivity (σ) and its different components with salt concentration. **b)** Ratios of different components to the net ionic conductivity are shown at low (0.36 M), medium (1.45 M) and high (2.80 M) salt concentrations. **c)** Inverse Haven ratio $H^{-1} = \frac{\sigma}{\sigma_{NE}}$ and the net ionic conductivities σ are shown as a function of salt concentration. Notice that σ is the maximum when H^{-1} is closest to unity . **d)** Fraction of Li^+ ions with various number of anions in its first solvation shell at the lowest (0.36 M) salt concentration. Most of the Li^+ ions are part of CIPs even at this lowest salt concentration.

3.3.3 Residence time of anions in the first solvation shell of Li^+ ions

We used the positions of the Li^+ ions and P atoms of PF_6^- ions to calculate the residence time autocorrelation function $C(t)$ using Eq. (2.14) and fitted with multi-exponential function to calculate the residence time τ as follows: $\tau = \int_0^\infty C(t)dt = \int_0^\infty \sum a_i \, e^{-t/b_i} dt$, where a_i and b_i are the fit parameters, and $\sum a_i$ =1. The threshold value was taken as 4.1 Å which corresponded to the position of the first minimum in the radial distribution function between Li^+- $P(PF_6^-)$ pair (see Figure 3.A.3). The autocorrelation functions were calculated for three independent sets of trajectories to ensure better averaging. The simulated correlation functions at different salt concentrations for one set of trajectories are shown in Figure 3.4a (see Figure 3.A.5 for the other two sets of trajectories). Tail portions of these correlation functions were wavy in nature, thus excluded during multi-exponential fits. The fit parameters are provided in Table 3.A.1. In Figure 3.4b, we have shown the ideal ionic conductivities σ_{NE} (conductivity calculated via the Nernst-Einstein equation) and the average anion residence times τ in the solvation shell of Li^+ ions as a function of salt concentration. Data in this figure indicate that as τ increases, σ_{NE} decreases and σ_{NE} becomes the maximum when τ is the minimum. A previous study[64] reported a linear relationship between the self-diffusivities of ions and the inverse of τ. Thus when τ becomes shorter, ions diffuse faster. As σ_{NE} values were calculated from the self-diffusivities of cations and anions, σ_{NE} increased when τ decreased. This explains the observed relationship between σ_{NE} and τ.

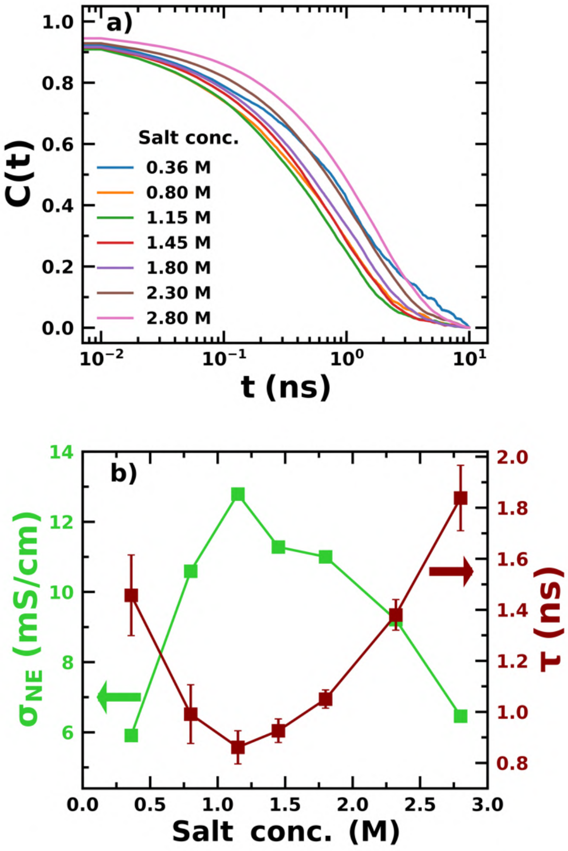


Figure 3.4: a) Residence time autocorrelation functions for anion exchange in the solvation shell of Li^+ ions are shown as a function of salt concentration. **b)** Ideal ionic conductivities, σ_{NE}, and the anion residence times, τ, are shown a function of salt concentration. Error bars represent standard error of mean.

3.3.4 Capturing ion-ion correlations using the Van Hove functions:

3.3.4.1 Cation-anion correlation

First, we will look at the correlation between the movement of cations and anions. We calculated the distinct parts of VHFs, $\mathrm{G_d^{Li^+-PF_6^-}(r,\ t)}$, using Eq. (2.13) at three representative salt concentrations: 0.36 M, 1.45 M and 2.80 M using the positions of $\mathrm{Li^+}$ ions and P-atoms of $\mathrm{PF_6^-}$ ions at different times. In the upper row of the Figure 3.5, the results, $\mathrm{G_d^{Li^+-PF_6^-}(r,\ t) - 1}$, are shown in a series of different times. It can be seen that as the time increases, not only the height of the nearest neighbour peaks around $\mathrm{r} =$ 2.9 Å decreases, but also the peak position slightly moves towards small r with time at all the three salt concentrations. This shows positive correlation between the movement of cations and anions. However, the rate of this peak height decrease becomes slower at higher salt concentration. The decay rate of peak heights in the distinct part of the VHF represents the correlation time scale between the neighboring atoms with the central atom[32,33,67]. If this correlation time scale exceeds the random diffusive time scale of the neighboring atoms, then the movement of the peak position towards small r values with time will not be associated with the positive correlation, rather it would be purely because of random diffusion. As the rate of the peak height decrease becomes slower at higher salt concentration, that is, correlation time increases, we need to compare these two different time scales to draw accurate conclusions. To calculate the correlation time scale between $\mathrm{Li^+}$ and $\mathrm{PF_6^-}$ ions, we fitted the decay of the nearest neighbour peak height (normalized) around r = 2.9 Å of $\mathrm{G_d^{Li^+-PF_6^-}(r,\ t) - 1}$ curves with a two-step relaxation function: $\mathrm{A\,e^{-\left(\frac{t}{\tau_{d1}}\right)^2} + (1-A)\,e^{-\left(\frac{t}{\tau_{d2}}\right)^\beta}}$ with $\tau_{d1} < \tau_{d2}$, where τ_{d1} represents the time scale associated with the ballistic motion of $\mathrm{PF_6^-}$ ions caged by the surrounding $\mathrm{Li^+}$ ions and τ_{d2} represents the correlation time scale of nearest neighbour $\mathrm{PF_6^-}$ ions with the central $\mathrm{Li^+}$ ion[32]. The fitting parameters are provided in Table 3.A.2. In the middle row of the Figure 3.5, we have shown the self-parts of the VHFs, $\mathrm{G_s^{PF_6^-}(r,t)}$ calculated using Eq. (2.12) for $\mathrm{PF_6^-}$ ions at different times and at different salt concentrations mentioned above. Since $\mathrm{G_s^{PF_6^-}(r,t)}$ represents the displacement of $\mathrm{PF_6^-}$ ions from its initial positions, the position of the maximum of $\mathrm{G_s^{PF_6^-}(r,t)}$ represents the most probable distance that $\mathrm{PF_6^-}$ ions can travel during the time interval t. The decay rate of the peak of $\mathrm{G_s^{PF_6^-}(r,t)}$ is associated with the random diffusive time scale of $\mathrm{PF_6^-}$ ions. The decay behavior of this peak height (normalized) of $\mathrm{G_s^{PF_6^-}(r,t)}$ was fitted using sum of two stretched exponential functions: $\mathrm{A\,e^{-\left(\frac{t}{\tau_{s1}}\right)^{\beta_1}} + (1-A)\,e^{-\left(\frac{t}{\tau_{s2}}\right)^{\beta_2}}}$ with $\tau_{s1} < \tau_{s2}$, where τ_{s1} represents the time scale associated with the ballistic motion of ions and τ_{s2} represents the random diffusive time scale of $\mathrm{PF_6^-}$ ions. The fitting parameters are provided in Table 3.A.3.

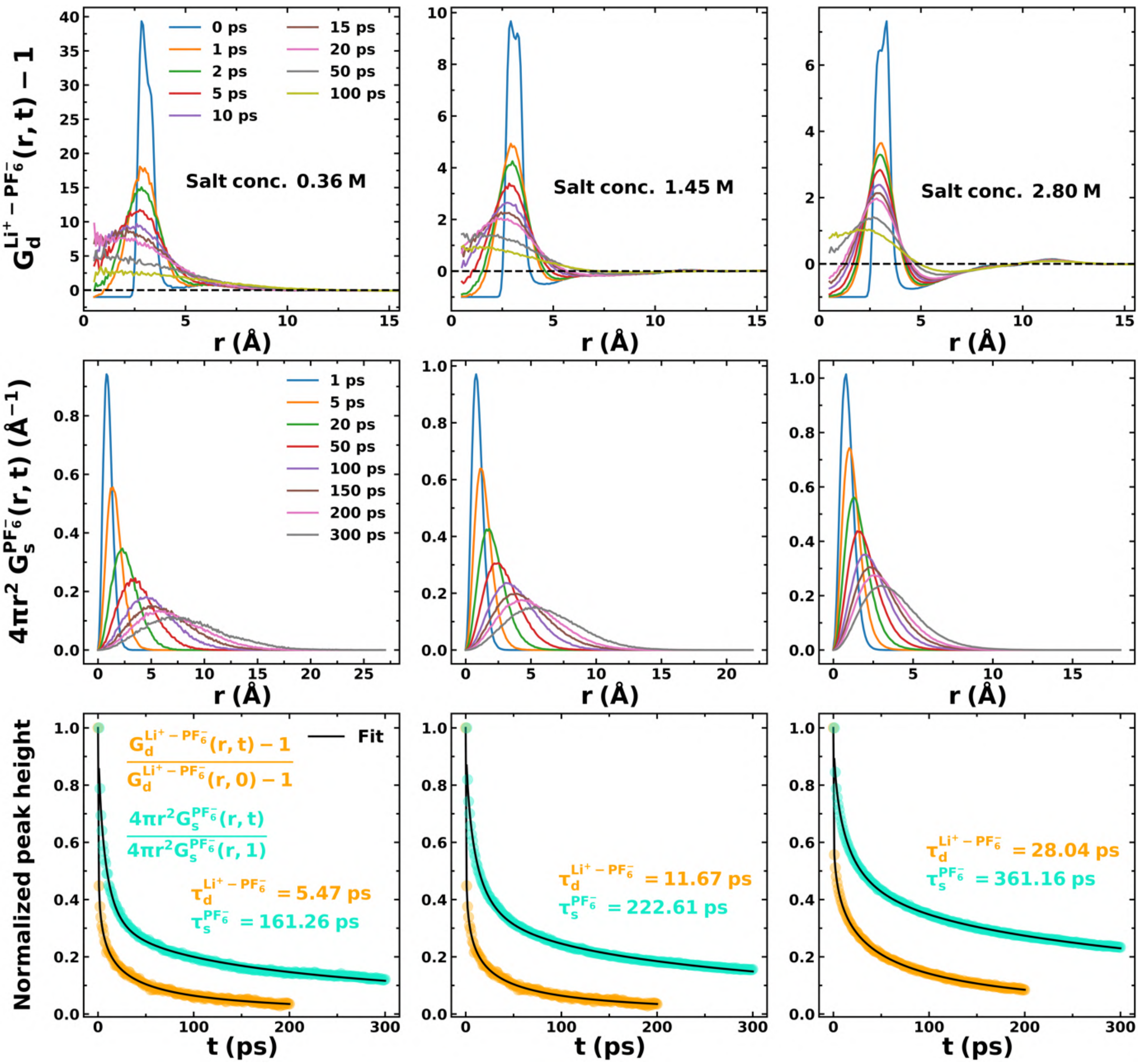


Figure 3.5: **Upper row,** the distinct parts of the VHFs between of Li$^+$ ions and PF$_6^-$ ions, $G_d^{Li^+-PF_6^-}(r,\ t)-1$, are plotted at different times for three representative salt concentrations. The data show that the nearest neighbour peak around $r = 2.9$ Å slightly moves towards small r values with time. This indicates positive correlation between cations and anions. **Middle row,** the self-parts of VHFs for PF$_6^-$ ions, $G_s^{PF_6^-}(r,t)$, are plotted at different times at those salt concentrations. **Lower row,** the decay of the normalized peak height of $G_s^{PF_6^-}(r,t)$ and the nearest neighbour peak height of $G_d^{Li^+-PF_6^-}(r,\ t)-1$ are plotted. The calculated diffusive time scales (denoted as $\tau_s^{PF_6^-}$) of PF$_6^-$ ions and the Li$^+$ − PF$_6^-$ correlation time scales (denoted as $\tau_d^{Li^+-PF_6^-}$) are also shown.

In the lower row of the Figure 3.5, we have shown the decay of the normalized peak heights for $\mathrm{G_s^{PF_6^-}(r,t)}$ and the nearest neighbour peak height around $\mathrm{r = 2.9\ Å}$ of $\mathrm{G_d^{Li^+-PF_6^-}(r,\ t) - 1}$ as well as provide the diffusive time scales (τ_{s2} denoted as $\mathrm{\tau_s^{PF_6^-}}$) of $\mathrm{PF_6^-}$ ions and the $\mathrm{Li^+ - PF_6^-}$ correlation time scales (τ_{d2} denoted as $\mathrm{\tau_d^{Li^+-PF_6^-}}$) at the three salt concentrations. We can see that $\mathrm{\tau_d^{Li^+-PF_6^-} < \tau_s^{PF_6^-}}$ at both the low and high salt concentrations. Therefore, the movement of the nearest neighbour peak positions of $\mathrm{G_d^{Li^+-PF_6^-}(r,\ t) - 1}$ towards small r with time indeed reflects positive correlation between cations and anions at every salt concentration. Interestingly, we noted that the Onsager's approach highlighted the same positive correlations between cations and anions. Thus, the predictions from both these theoretical frameworks at this concentration regime are consistent with each other. Thus, in this way the VHF can be analyzed to capture positive correlations in electrolytes.

3.3.4.2 Cation-cation and anion-anion correlations:

Upon describing the ability of the VHF to capture the positive correlation, now we need to look at its ability to capture negative correlation in ion movements. Onsager's approach highlighted negative correlations between cation-cation and anion-anion at medium (1.45 M) and high (2.80 M) salt concentrations. In Figure 3.6, we have shown the distinct parts of the VHFs for cation-cation $\mathrm{G_d^{Li^+-\,Li^+}(r,\ t) - 1}$ and anion-anion $\mathrm{G_d^{PF_6^-\,-PF_6^-}(r,\ t) - 1}$ at different times for medium and high salt concentrations. The results for low salt concentration (0.36 M) are shown in Figure 3.A.6. It can be seen that for both $\mathrm{Li^+ - Li^+}$ and $\mathrm{PF_6^- - PF_6^-}$ correlations, the first peak decays very quickly, whereas the second peak around $\mathrm{r = 9.8\ Å}$ in $\mathrm{G_d^{Li^+-\,Li^+}(r,\ t) - 1}$ moves away from r = 0. Following the same fitting procedures as mentioned earlier, we calculated the $\mathrm{Li^+ - Li^+}$ correlation time scale after fitting the decay of the peak height around $\mathrm{r = 9.8\ Å}$ in $\mathrm{G_d^{Li^+-\,Li^+}(r,\ t) - 1}$. The fitting parameters are provided in Table 3.A.4. The diffusive time scale of $\mathrm{Li^+}$ ions were obtained by fitting the decay of the normalized peak height of the self-parts of the VHFs of $\mathrm{Li^+}$ ions, $\mathrm{G_s^{Li^+}(r,t)}$. Fit parameters are provided in Table 3.A.3. $\mathrm{G_s^{Li^+}(r,t)}$ at different times at medium and high salt concentrations are shown in Figure 3.A.7. The $\mathrm{Li^+ - Li^+}$ correlation time scales are smaller than the diffusive time scales of $\mathrm{Li^+}$ ions at both medium and high salt concentrations (see Figure 3.A.8). This represents negative correlation between the movement of $\mathrm{Li^+}$ ions at medium to high salt concentrations. For $\mathrm{PF_6^- - PF_6^-}$ correlation, the correlation peaks decay at a much faster rate than $\mathrm{Li^+ - Li^+}$ correlation. So, the same is true also for the $\mathrm{PF_6^- - PF_6^-}$ correlation. Due to the choice of very less number of ions (11 pairs of ions, see Table 3.1) in the simulation box at 0.36 M concentration, $\mathrm{G_d^{Li^+-\,Li^+}(r,\ t) - 1}$ and $\mathrm{G_d^{PF_6^-\,-PF_6^-}(r,\ t) - 1}$ curves at this low concentration are not smooth enough (Figure 3.A.6) to calculate correlation time scales from the decay of the peak height. Note that this result also matches with the predictions from Onsager's theory (Figure 3.3b). Thus, the VHF is capable of capturing both positive and negative correlations between the movement of ions in electrolytes.

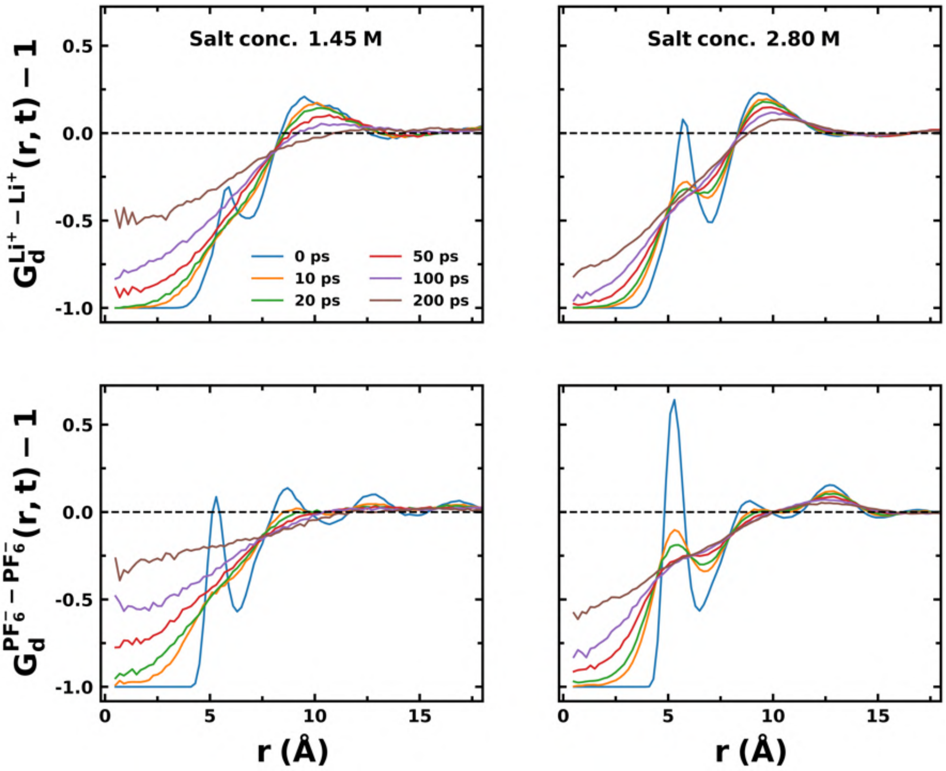


Figure 3.6: The distinct parts of VHFs for cation-cation $G_d^{Li^+-Li^+}(r,\ t)-1$ and anion-anion $G_d^{PF_6^- -PF_6^-}(r,\ t)-1$ are plotted at different times for medium (1.45 M) and high (2.80 M) salt concentrations.

3.4 Conclusions

In this work we systematically investigated directional correlations in ion movements using the VHFs in a mixture of EMC and FEC with varying $LiPF_6$ salt concentration at 313 K. We first employed the Onsager transport coefficients to understand the nature of correlations in ion movements in this solution at different $LiPF_6$ concentrations. We observed that at low electrolyte concentration, positive correlation exists between cation-cation, anion-anion and cation-anion motions. This positive correlation diminishes upon increasing the concentration of the electrolyte employed and anti-correlation appears for cation-cation and anion-anion motions at high concentration.

The VHFs successfully capture both the positive correlation between cation-anion motion at different salt concentrations and the anti-correlation between ions of same charge-type at high salt concentration. All these predictions using the VHFs are in good agreement with those from the Onsager's theory. On the other hand, VHFs cannot foretell which correlation is dominating at any given electrolyte concentration. This type of quantitative information can be obtained from Onsager's approach only. This may be considered as a limitation of relying solely on the VHFs for a complete understanding of ion correlations in electrolyte media. Our study presented here, therefore, shows a systematic protocol to identify different types of correlated motions between ions using the VHF and may help in choosing the suitable approach from this pair of the VHF and the Onsager's framework while studying correlated ion dynamics in LIB electrolytes.

Appendix 3.A

Table 3.A.1: Fit parameters in the equation: $C(t) = \sum_{i=1}^{3} a_i \, e^{-t/b_i}$, used to calculate anion residence times in the first solvation shell of cations using the residence time correlation functions. Residence times were calculated using the equation: $\tau = \sum_{i=1}^{3} a_i \, b_i$. Goodness-of-fit was measured in terms of coefficient of determination, $\chi^2 = 1 - \sum_i \left(\tau_i^{original} - \tau_i^{fit}\right)^2 / \sum_i \left(\tau_i^{original} - \overline{\tau^{original}}\right)^2$. Three independent set of trajectories (Set-1,2,3) were used to calculate residence times. Fitting was performed up to 2.5 ns to exclude the kinky tail portions.

Set-1								
Salt conc. (M)	a_1	b_1 (ps)	a_2	b_2 (ps)	a_3	b_3 (ps)	χ^2	τ (ps)
0.36	0.09	10.53	0.24	224.52	0.67	2697.32	0.99	1841.08
0.80	0.08	4.97	0.16	95.18	0.76	1122.32	0.98	874.77
1.15	0.14	18.04	0.29	337.85	0.57	1482.39	0.99	937.83
1.45	0.13	16.99	0.30	451.81	0.57	1435.35	0.99	957.67
1.80	0.11	14.73	0.33	413.79	0.56	1739.53	0.97	1104.13
2.32	0.09	13.23	0.26	475.66	0.65	2025.74	0.98	1438.13
2.80	0.08	11.37	0.17	500.75	0.75	2423.41	0.99	1905.41
Set-2								
0.36	0.09	10.56	0.14	142.03	0.75	1587.59	0.99	1221.19
0.80	0.17	31.12	0.56	704.43	0.27	3222.81	0.98	1272.43
1.15	0.09	7.29	0.21	158.10	0.70	1301.61	0.97	942.42
1.45	0.11	9.18	0.31	276.46	0.58	1573.49	0.97	1003.05
1.80	0.10	12.52	0.21	304.22	0.69	1484.02	0.98	1084.65
2.32	0.10	17.66	0.28	606.94	0.62	2108.63	0.98	1469.13
2.80	0.06	8.28	0.09	234.51	0.85	2416.89	0.99	2070.19
Set-3								
0.36	0.08	0.39	0.15	148.96	0.77	1682.20	0.98	1310.66
0.80	0.12	13.99	0.19	181.17	0.69	1144.80	0.99	827.00
1.15	0.09	7.70	0.19	149.87	0.72	943.08	0.97	703.17
1.45	0.10	11.95	0.19	243.23	0.71	1089.63	0.98	815.75
1.80	0.09	7.16	0.19	196.94	0.72	1286.47	0.96	963.10
2.32	0.08	9.39	0.17	294.76	0.75	1588.39	0.98	1233.92
2.80	0.06	8.01	0.10	253.50	0.84	1813.09	0.99	1540.90

Table 3.A.2: Fit parameters in the equation: $A\,e^{-\left(\frac{t}{\tau_{d1}}\right)^{2}} + (1-A)\,e^{-\left(\frac{t}{\tau_{d2}}\right)^{\beta}}$, used to fit the decay of the normalized nearest neighbour peak height around $r = 2.9$ Å of the distinct parts of the VHFs for cation-anion, $G_d^{Li^+-PF_6^-}(r,\ t) - 1$.

Salt conc. (M)	A	τ_{d1} (ps)	τ_{d2} (ps)	β
0.36	0.23	0.65	5.47	0.31
1.45	0.29	0.68	11.67	0.37
2.80	0.27	0.15	28.04	0.39

Table 3.A.3: Fit parameters in the equation: $A\,e^{-\left(\frac{t}{\tau_{s1}}\right)^{\beta_1}} + (1-A)\,e^{-\left(\frac{t}{\tau_{s2}}\right)^{\beta_2}}$, used to fit the decay of the normalized peak height of the self-parts of the Van Hove functions for cations and anions, $G_s^{Li^+}(r,t)$ and $G_s^{PF_6^-}(r,t)$ respectively.

For $G_s^{PF_6^-}(r,t)$ peak height decay					
Salt conc. (M)	A	τ_{s1} (ps)	τ_{s2} (ps)	β_1	β_2
0.36	0.56	7.02	161.26	0.80	0.47
1.45	0.51	8.64	222.61	0.71	0.51
2.80	0.44	10.69	361.16	0.64	0.54
For $G_s^{Li^+}(r,t)$ peak height decay					
0.36	0.52	6.83	132.11	0.82	0.48
1.45	0.53	8.86	223.39	0.70	0.51
2.80	0.48	11.51	449.96	0.62	0.62

Table 3.A.4: Fit parameters in the equation: $A\,e^{-\left(\frac{t}{\tau_{d1}}\right)^{2}} + (1-A)\,e^{-\left(\frac{t}{\tau_{d2}}\right)^{\beta}}$, used to fit the decay of the normalized peak height around $r = 9.8$ Å of the distinct parts of the Van Hove functions for cation-cation, $G_d^{Li^+-\,Li^+}(r,\ t) - 1$.

Salt conc. (M)	A	τ_{d1} (ps)	τ_{d2} (ps)	β
1.45	0.25	55.14	83.88	0.64
2.80	0.06	83.83	206.64	0.58

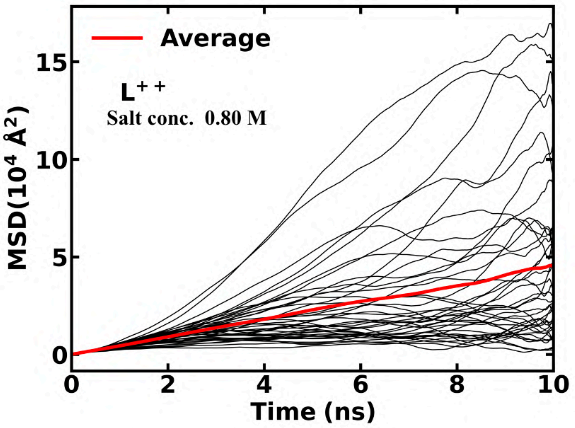


Figure 3.A.1: Collective mean squared displacements (MSD in general term) of all 40 independent trajectories are shown alongside the average curve (red line). Each of the MSD curve of the independent trajectories is very kinky in nature. Slope of the averaged curve (red line) is used to calculate L^{++} at 0.80 M salt concentration. Same procedure was applied to calculate all the other Onsager transport coefficients for all salt concentrations.

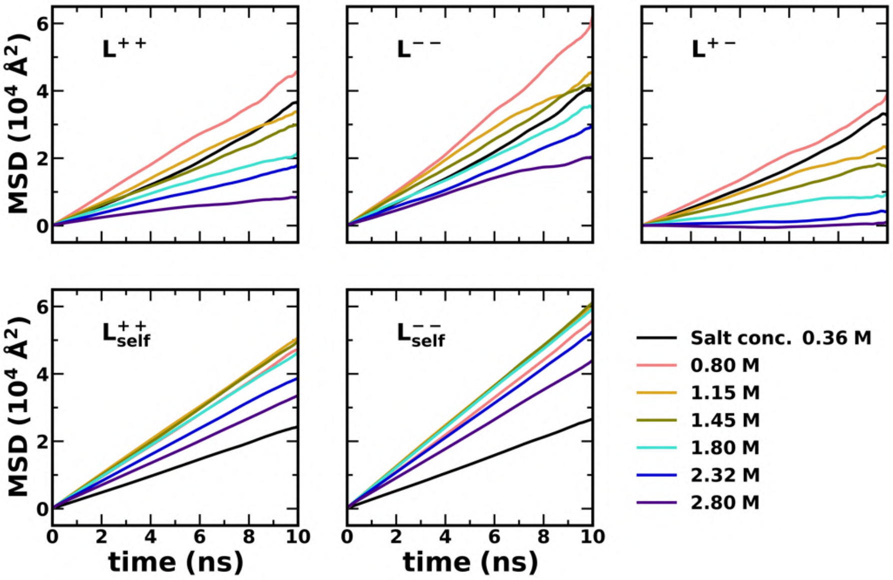


Figure 3.A.2: Each collective mean squared displacement lines (MSD in general term) represents the average over 40 independent trajectories. Slope of these lines were used to calculate corresponding Onsager transport coefficients at different salt concentrations.

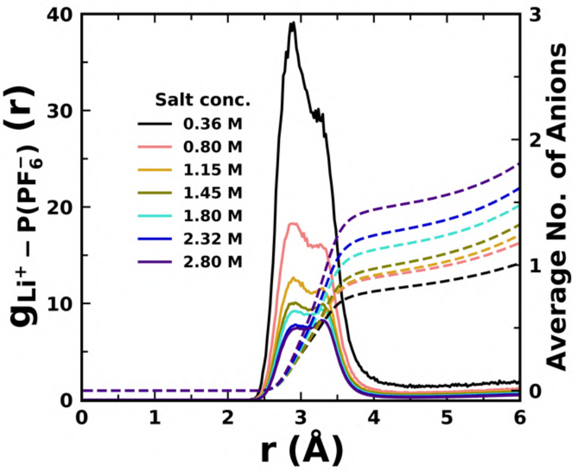


Figure 3.A.3: Radial distribution function (RDF) between Li^+-P(PF_6^-). Position of the first minima is at 4.1 Å. This is the extent of first solvation shell of Li^+ . Cumulative number or the average number of anions within a distance r are also shown by the dashed lines. Average number of anions in the first solvation shell of cations increases with increasing salt concentration.

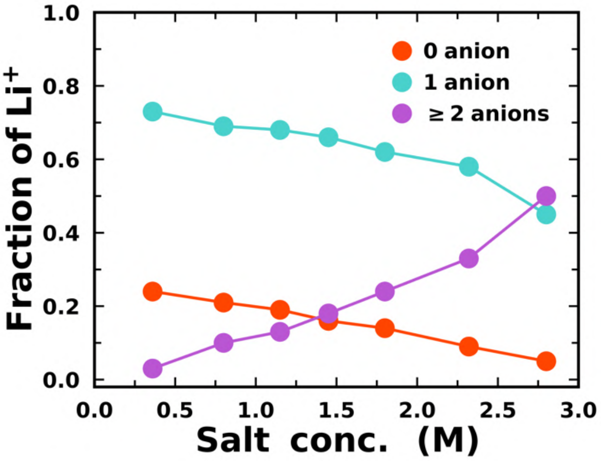


Figure 3.A.4: Fraction of Li^+ ions with different number of anions in its first solvation shell is plotted as a function of salt concentration.

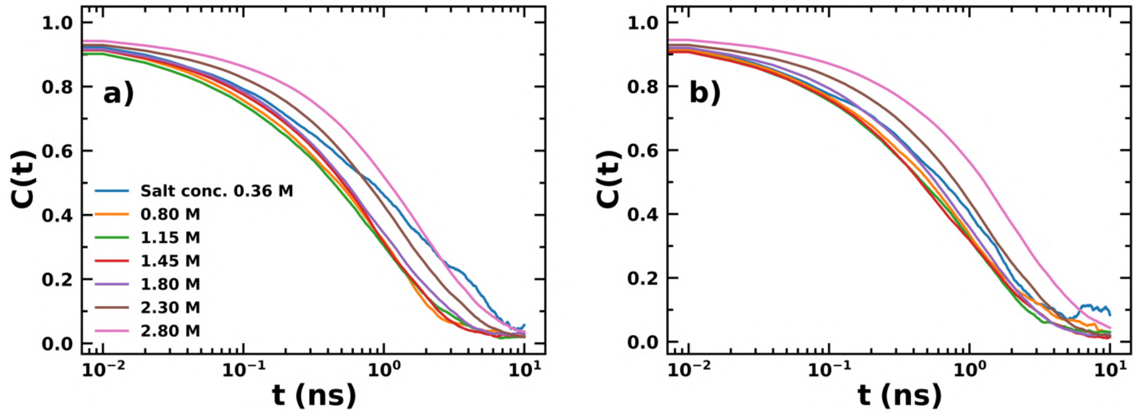


Figure 3.A.5. Residence time autocorrelation functions between $\mathrm{Li^+} - \mathrm{P}(\mathrm{PF_6^-})$ are plotted at different salt concentrations for the other two independent set of trajectories (**a, b**).

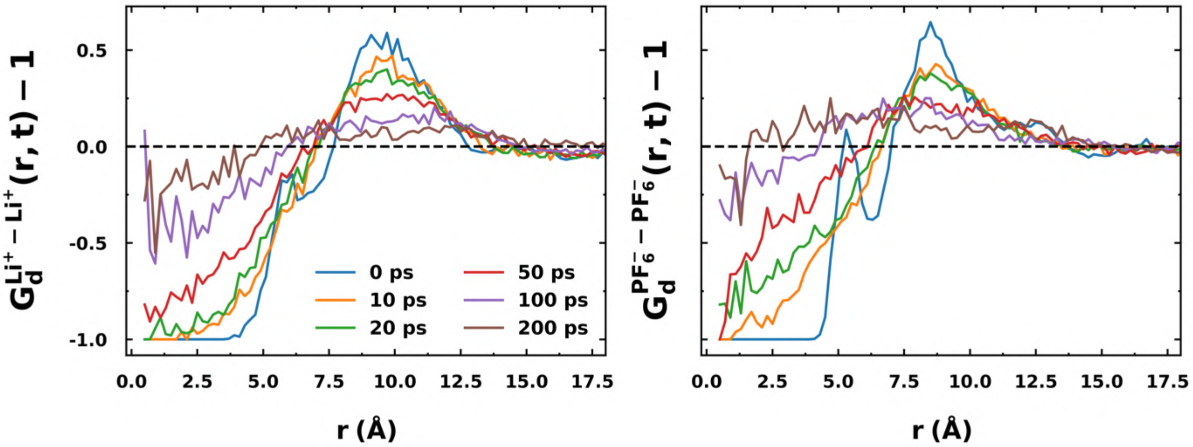


Figure 3.A.6. The distinct parts of Van Hove functions for cation-cation $\mathrm{G_d^{Li^+ - Li^+}}(\mathrm{r},\ \mathrm{t}) - 1$ and anion-anion $\mathrm{G_d^{PF_6^- - PF_6^-}}(\mathrm{r},\ \mathrm{t}) - 1$ at different times are shown for 0.36 M salt concentration.

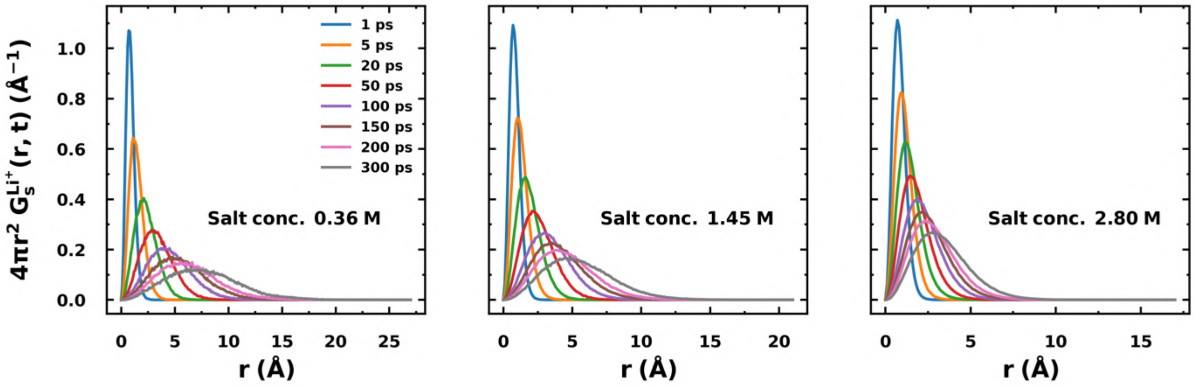


Figure 3.A.7. Self-parts of the Van Hove functions for Li$^+$ ions, $G_s^{Li^+}(r, t)$, are plotted at different times and at different salt concentrations.

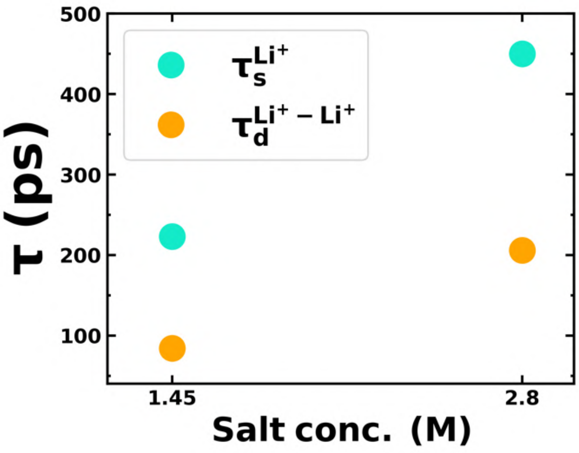


Figure 3.A.8. Li$^+$ − Li$^+$ correlation time scales $\left(\tau_d^{Li^+ - Li^+}\right)$ and Li$^+$ ion diffusive time scales $\left(\tau_s^{Li^+}\right)$ are shown at medium (1.45 M) and high (2.80 M) salt concentrations.

References:

1 J. B. Goodenough and K.-S. Park, J Am Chem Soc, 2013, 135, 1167–1176.

2 A. Opitz, P. Badami, L. Shen, K. Vignarooban and A. M. Kannan, Renewable and Sustainable Energy Reviews, 2017, 68, 685–692.

3 K. Xu, Chem Rev, 2004, 104, 4303–4418.

4 K. Xu, Chem Rev, 2014, 114, 11503–11618.

5 G. Gachot, S. Grugeon, M. Armand, S. Pilard, P. Guenot, J.-M. Tarascon and S. Laruelle, J Power Sources, 2008, 178, 409–421.

6 Y. Yamada, M. Yaegashi, T. Abe and A. Yamada, Chemical Communications, 2013, 49, 11194–11196.

7 Y. Yamada, K. Furukawa, K. Sodeyama, K. Kikuchi, M. Yaegashi, Y. Tateyama and A. Yamada, J Am Chem Soc, 2014, 136, 5039–5046.

8 J. Zheng, J. A. Lochala, A. Kwok, Z. D. Deng and J. Xiao, Advanced Science, 2017, 4, 1700032.

9 Y. Yamada and A. Yamada, J Electrochem Soc, 2015, 162, A2406.

10 K. Hayamizu, J Chem Eng Data, 2012, 57, 2012–2017.

11 A. Sirjoosingh, S. Alavi and T. K. Woo, J Phys Chem B, 2009, 113, 8103–8113.

12 M. H. Kowsari, S. Alavi, M. Ashrafizaadeh and B. Najafi, J Chem Phys, 2009, 130, 14703.

13 M. H. Kowsari, S. Alavi, B. Najafi, K. Gholizadeh, E. Dehghanpisheh and F. Ranjbar, Physical Chemistry Chemical Physics, 2011, 13, 8826–8837.

14 N. M. Vargas-Barbosa and B. Roling, ChemElectroChem, 2020, 7, 367–385.

15 K. Binder, J. Horbach, W. Kob, W. Paul and F. Varnik, Journal of Physics: Condensed Matter,2004, 16, S429.

16 K. D. Fong, H. K. Bergstrom, B. D. McCloskey and K. K. Mandadapu, AIChE Journal, 2020, 66, e17091.

17 H. K. Bergstrom, K. D. Fong, D. M. Halat, C. A. Karouta, H. C. Celik, J. A. Reimer and B. D. McCloskey, Chem Sci.

18 K. D. Fong, J. Self, B. D. McCloskey and K. A. Persson, Macromolecules, 2021, 54, 2575–2591.

19 K. D. Fong, J. Self, B. D. McCloskey and K. A. Persson, Macromolecules, 2020, 53, 9503–9512.

20 J. G. McDaniel and C. Y. Son, J Phys Chem B, 2018, 122, 7154–7169.

21 P. Kubisiak, P. Wróbel and A. Eilmes, J Phys Chem B, 2019, 124, 413–421.

22 S. Blazquez, J. L. F. Abascal, J. Lagerweij, P. Habibi, P. Dey, T. J. H. Vlugt, O. A. Moultos and C. Vega, J Chem Theory Comput, 2023, 19, 5380–5393.

23 H. K. Kashyap, H. V. R. Annapureddy, F. O. Raineri and C. J. Margulis, J Phys Chem B, 2011, 115.

24 H. S. Sachar, N. Marioni, E. S. Zofchak and V. Ganesan, Macromolecules, 2023, 56, 2194–2208.

25 N. Marioni, Z. Zhang, E. S. Zofchak, H. S. Sachar, S. Kadulkar, B. D. Freeman and V. Ganesan, ACS Macro Lett, 2022, 11, 1258–1264.

26 Ø. Gullbrekken, I. T. Røe, S. M. Selbach and S. K. Schnell, J Phys Chem B, 2023, 127, 2729–2738.

27 C. Fang, D. M. Halat, N. P. Balsara and R. Wang, J Phys Chem B, 2023, 127, 1803–1810.

28 A. K. Verma, A. S. Thorat and J. K. Shah, Journal of Ionic Liquids, 2024, 4, 100089.

29 P. Kubisiak and A. Eilmes, J Phys Chem B, 2020, 124, 9680–9689.

30 L. Van Hove, Physical Review, 1954, 95, 249.

31 Y. Shinohara, A. S. Ivanov, D. Maltsev, G. E. Granroth, D. L. Abernathy, S. Dai and T. Egami, Journal of Physical Chemistry Letters, 2022, 13, 5956–5962.

32 Y. Shinohara, R. Matsumoto, M. W. Thompson, C. W. Ryu, W. Dmowski, T. Iwashita, D. Ishikawa, A. Q. R. Baron, P. T. Cummings and T. Egami, Journal of Physical Chemistry Letters, 2019, 10, 7119–7125.

33 Y. Shinohara, W. Dmowski, T. Iwashita, D. Ishikawa, A. Q. R. Baron and T. Egami, Phys Rev Mater, 2019, 3, 065604.

34 X. Luo, H. Liu and S. J. Paddison, ACS Appl Polym Mater, 2021, 3, 141–152.

35 D. J. Xiong, M. Bauer, L. D. Ellis, T. Hynes, S. Hyatt, D. S. Hall and J. R. Dahn, J Electrochem Soc, 2018, 165, A126.

36 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, J Comput Chem, 2004, 25, 1157–1174.

37 K. G. Sprenger, V. W. Jaeger and J. Pfaendtner, J Phys Chem B, 2015, 119, 5882–5895.

38 N. Kumar and J. M. Seminario, The Journal of Physical Chemistry C, 2016, 120, 16322–16332.

39 J. N. Canongia Lopes and A. A. H. Pádua, J Phys Chem B, 2004, 108, 16893–16898.

40 C. Park, M. Kanduč, R. Chudoba, A. Ronneburg, S. Risse, M. Ballauff and J. Dzubiella, J Power Sources, 2018, 373, 70–78.

41 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, J Phys Chem, 1994, 98, 11623–11627.

42 A. Becke, Chem. Phys, 98, 5648.

43 C. Lee, W. Yang and R. G. Parr, Phys Rev B, 1988, 37, 785.

44 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji and others, Gaussian, Inc. Wallingford, CT, 2016, preprint.

45 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, J Phys Chem, 1993, 97, 10269–10280.

46 D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, J Comput Chem, 2005, 26, 1668–1688.

47 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, SoftwareX, 2015, 1, 19–25.

48 L. Martinez, R. Andrade, E. G. Birgin and J. M. Martinez, J Comput Chem, 2009, 30, 2157–2164.

49 G. Bussi, D. Donadio and M. Parrinello, J Chem Phys, 2007, 126, 14101.

50 H. J. C. Berendsen, J. P. M. van Postma, W. F. Van Gunsteren, A. DiNola and J. R. Haak, J Chem Phys, 1984, 81, 3684–3690.

51 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, J Chem Phys, 1995, 103, 8577–8593.

52 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, J Comput Chem, 1997, 18, 1463–1472.

53 N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, J Comput Chem, 2011, 32, 2319–2327.

54 R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domanski, D. L. Dotson, S. Buchoux, I. M. Kenney and others, in Proceedings of the 15th python in science conference, 2016, vol. 98, p. 105.

55 M. Maiti, A. N. Krishnamoorthy, Y. Mabrouk, N. Mozhzhukhina, A. Matic, D. Diddens and A. Heuer, Physical Chemistry Chemical Physics.

56 K. Kondo, M. Sano, A. Hiwara, T. Omi, M. Fujita, A. Kuwae, M. Iida, K. Mogi and H. Yokoyama, J Phys Chem B, 2000, 104, 5040–5044.

57 S. Hwang, D.-H. Kim, J. H. Shin, J. E. Jang, K. H. Ahn, C. Lee and H. Lee, The Journal of Physical Chemistry C, 2018, 122, 19438–19446.

58 E. R. Logan, E. M. Tonita, K. L. Gering, L. Ma, M. K. G. Bauer, J. Li, L. Y. Beaulieu and J. R. Dahn, J Electrochem Soc, 2018, 165, A705--A716.

59 F. Hanke, N. Modrow, R. L. C. Akkermans, I. Korotkin, F. C. Mocanu, V. A. Neufeld and M. Veit, J Electrochem Soc, 2019, 167, 13522.

60 D. Bedrov, J.-P. Piquemal, O. Borodin, A. D. MacKerell Jr, B. Roux and C. Schröder, Chem Rev, 2019, 119, 7940–7995.

61 E. J. Maginn, R. A. Messerly, D. J. Carlson, D. R. Roe and J. R. Elliot, Living J Comput Mol Sci, 2019, 1, 6324–6324.

62 B. Ravikumar, M. Mynam and B. Rai, The Journal of Physical Chemistry C, 2018, 122, 8173–8181.

63 M. Mynam, B. Ravikumar and B. Rai, J Mol Liq, 2019, 278, 97–104.

64 Y. Zhang and E. J. Maginn, Journal of Physical Chemistry Letters, 2015, 6, 700–705.

65 Y. Sasaki, Electrochemistry, 2008, 76, 2–15.

66 C.-C. Su, M. He, R. Amine, T. Rojas, L. Cheng, A. T. Ngo and K. Amine, Energy Environ Sci, 2019, 12, 1249–1254.

67 B. Wu, T. Iwashita and T. Egami, Phys Rev Lett, 2018, 120, 135502.

4

Correlations between ionic conductivity and co-solvent modulated water structure and dynamics in aqueous Zn-ion battery electrolytes

4.1 Introduction

Energy demands of the modern society have considerably increased because of massive technological and societal developments in the recent years[1,2]. Excessive consumption of fossil fuels and the consequential environmental issues have compelled researchers in exploring green sustainable energy resources[3]. Electrochemical energy storage devices, such as rechargeable batteries, play a vital role among the renewable energy sources[4–7]. Currently, non-aqueous lithium-ion batteries (LIBs) are dominating the electrochemical energy storage industry because of their high energy density and long life cycle[8]. However, the increasing concern about limited lithium resources, high cost and safety issues have seriously hindered the continuous large scale manufacturing of LIBs[9]. Sodium-ion batteries (SIBs) and potassium-ion batteries (KIBs) are plausible alternatives to LIBs because of the relative abundance of sodium (potassium) over lithium, but suffer from low energy density, high operating cost and critical security concerns[10–12]. These drawbacks of SIBs and KIBs have motivated the scientists to explore further alternatives that are associated with low cost, high energy density and long life cycles.[13]

Aqueous rechargeable batteries offer a promising alternative because of their low cost, inherent safety and environment friendliness[14,15]. Furthermore, the aqueous electrolyte systems offer at least one order of magnitude higher ionic conductivity than that of non-aqueous counterparts[16]. In a broad class of aqueous batteries, aqueous zinc ion batteries (AZIBs) have recently been studied extensively to support large scale production of energy storage systems[17–20]. The main components of a rechargeable AZIB cell are the Zn metal anode, Zn^{2+} host cathode, electrolyte and separator. Zn offers several advantages in aqueous electrochemical systems. Zn is relatively cheaper and shows relatively higher electrochemical stability in aqueous medium as compared to lithium or sodium[21]. Furthermore, Zn metal anode possesses high volumetric capacity and low redox potential (-0.76 V vs standard hydrogen electrode) suitable for battery operation in aqueous media[22].

In spite of the above advantages, poisoning of Zn metal anode due to corrosion and the consequent effects on battery lifespan have hampered the large scale commercialization of AZIBs[23–25]. The low redox potential of Zn anode triggers the hydrogen evolution reaction, leading to reduction of H_2O into H_2 during Zn deposition at the anode[26]. This leads to corrosion of Zn metal anode and formation of Zn dendrites. The uneven Zn platting with $[Zn(H_2O)_6]^{2+}$ solvation structure in AZIBs may be responsible for triggering such a poisoning. Several strategies have been developed over the years to mitigate Zn metal anode corrosion[27–30]. One of the strategies is to reduce the number of coordinated H_2O molecules inside the solvation shell of Zn^{2+} ions through either adequate electrolyte additives or adding an appropriate co-solvent with higher donor number than water[31,32]. A few experimental studies used propylene carbonate (PC) as a co-solvent in aqueous zinc triflate $(Zn(OTf)_2)$ electrolyte solution and showed the formation of hydrophobic solid electrolyte interface (SEI) along with decreased water activity on the Zn metal anode[33,34]. Another experimental study used methanol as a co-solvent in $ZnSO_4$ aqueous electrolyte and reported that the Zn^{2+} coordinated H_2O molecules interact with the co-solvent molecules, weakening Zn^{2+} solvation and water activity[35]. This significantly boosted the Zn reversibility, and dendrite free Zn platting enhanced the performance of the battery cell. Ethanol as a co-solvent was also used in aqueous $Zn(OTf)_2$ solution to boost the Zn reversibility and long-term durability of cell[36]. Another experimental study used dimethyl sulfoxide (DMSO) in aqueous $ZnCl_2$ solution and showed that DMSO replaces H_2O molecules inside Zn^{2+} solvation sheath and thus limits water decomposition and Zn dendrite growth[37]. Furthermore, several studies used tetramethylurea as a co-solvent to regulate the Zn^{2+} solvation structure[38,39]. This co-solvent electrolyte engineering for AZIBs has appeared as a very effective and straightforward technique to avoid Zn metal anode corrosion[40]. Apart from experimental studies, recently machine learning (ML) approaches have also been developed for AZIB electrolyte additive screening[41].

Despite the availability of several experimental studies on AZIB electrolyte engineering with different co-solvents, a detailed molecular level picture of how co-solvent mitigates anode poisoning via modulating structure and dynamics of water molecules that are around Zn^{2+} ions (solvation shell) and their correlations to ion transport and conductivity are yet to emerge. For this, we have performed all-atom classical molecular dynamics (MD) simulations of a representative experimental AZIB system[38], that is, 2M aqueous zinc triflate ($Zn(OTf)_2$) solutions with varying concentrations of tetramethylurea (TMU) as a co-solvent at room temperature. We have investigated the effects of co-solvent on water structure and dynamics, and explored correlations, if any, of the co-solvent altered water structure and dynamics to the calculated ionic conductivity. Figure 4.1 provides schematic diagrams of OTf^- ion and TMU molecule.

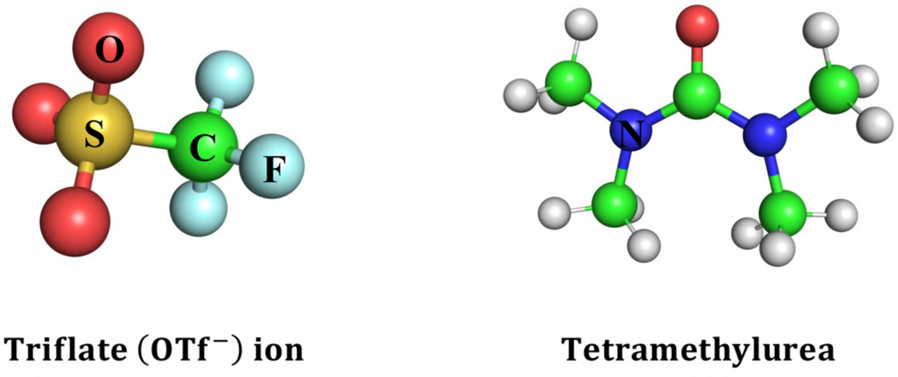


Figure 4.1. A ball-stick representation of OTf^- ion and tetramethylurea (TMU) molecule.

4.2 Computational details

All atom classical MD simulations were performed for five AZIB electrolyte solutions with varying TMU concentration (TMU vol% = 0, 10, 20, 30, 40). Electrolyte solution with no TMU (zero TMU vol%) is considered as the base electrolyte. The number of Zn^{2+} cations, OTf^- anions, water and TMU molecules used in this study are provided in Table 4.1. We have employed two different models for the electrolyte solutions: (i) OPC3 model for water[42] and the corresponding Zn^{2+} ion parameters[43] + an eight site potential model for TMU[44,45], and (ii) SPC/E model for water[46] and the corresponding Zn^{2+} ion parameters[47] + the same eight site potential model for TMU. The parameters for the OTf^- ions were taken from an earlier[48] wok and the same parameters were used for both the electrolyte models. Atomic partial charges of the ions were scaled by a factor 0.8 to include the effects of electronic polarization as done in several previous AZIB electrolyte simulation studies[49,50].

All the simulations reported in this study were performed using GROMACS v2022.5 software package[51]. Initial configurations of the molecules in the box of 50 Å × 50 Å × 50 Å dimensions were generated using the PACKMOL software[52]. These initial configurations were first energy minimized by the steepest descent method and then equilibrated in the NPT ensemble at a temperature of 300 K and pressure of 1 atm. until the densities converge sufficiently. Temperature and pressure were controlled through velocity rescaling method[53] with a time constant of 0.1 ps and Berendsen barostat[54] with a time constant of 1 ps, respectively. All simulations were performed under the periodic boundary conditions, and the long-range electrostatic interactions were handled by using the particle mesh Ewald (PME) summation method[55]. The cut-off distances for the electrostatic and van der Waals interactions were set to 12 Å. Newton's equations of motions were numerically integrated using the leap-frog algorithm with a time step of 1 fs. The equilibrated electrolyte salt concentrations in molarity (M) for all the five systems were ~ 2M (see Table 4.1) that matches with the experimental concentration[38]. The systems were further equilibrated for 1 ns in NVT ensemble starting from these 'NPT-equilibrated' solution structures before performing production simulations.

Initially, 400 ns of simulations were performed for all the five OPC3 water model electrolyte systems and only for the TMU vol% = 40 system modelled using the SPC/E water in NVT ensemble with a frame saving interval of 5 fs. It turned out that the OPC3 water model performed better than the SPC/E water model for estimating shear viscosities of the AZIB electrolyte solutions in the presence of TMU (see the discussion in the section 4.3.1). Therefore, production runs of 600 ns length for all the OPC3 water model electrolyte systems were performed in NVT ensemble with a frame saving interval of 1 ps. All the trajectories were analysed using various GROMACS utilities and the MDAnalysis python package[56].

Table 4.1. Number of cation (Zn^{2+}), anion (OTf^{-}), water and TMU molecules used are given. Salt concentrations in molarity (M) for each simulation box were calculated after NPT equilibrations.

System (TMU vol%)	Number of Zn^{2+}	Number of OTf^{-}	Number of waters	Number of TMU	Salt concentration (M)
0	95	190	2000	0	1.979
10	95	190	1795	31	1.977
20	95	190	1597	61	1.970
30	95	190	1393	92	1.979
40	95	190	1195	122	1.978

4.3 Results and discussions

4.3.1 Force field validation

The force field parameters were validated by comparing the ionic conductivities (σ) and shear viscosities (η) of both the base electrolyte system (zero TMU vol%) as well as in the presence of TMU with the available experimental data[38]. Ionic conductivities were calculated using the Onsager transport coefficients as discussed in section 2.1 of Chapter 2. Each 600 ns long production trajectory was divided into 10 equal segments. Segments were treated as independent trajectories. Subsequently, the transport coefficients were calculated from these 10 independent trajectories using Eqs. (2.2), (2.3) and (2.4). The net ionic conductivity (σ) was then calculated using Eq. (2.8) and the results are shown in Figure 4.2(a). All individual collective mean squared displacement plots for all the systems are shown in Figure 4.A.1-4.A.5.

Shear viscosities were calculated using the Green-Kubo (GK) integral formula given in Eq. (2.17). Every 400 ns long trajectories were divided into 40 equal parts and all these individual segments were treated as independent trajectories. To calculate shear viscosities, Eq. (2.17) was applied on each of these 40 independent trajectories and the average was taken. Dependence of the viscosity coefficients on the upper limit of the GK integration averaged over these 40 independent trajectories for various electrolyte systems are shown in Figure 4.A.6. Simulated viscosity values computed for the OPC3 water model systems are plotted in the inset of Figure 4.2(a) and for the SPC/E water model is provided in Table 4.2.

Table 4.2. Simulated shear viscosities (η) for the TMU vol% = 40 electrolyte system modelled with the OPC3 and SPC/E water are provided and compared with the experimental result. Errors represent the standard error.

Model	η (cP)
Experimental	7.00
OPC3 water electrolyte system	16.12 ± 1.54
SPC/E water electrolyte system	18.35 ± 1.91

As shown in Figure 4.2(a), in the absence of TMU, both the shear viscosity and ionic conductivity match remarkably well with the corresponding experimental values as the OPC3 water model was proven to accurately reproduce the dynamical properties of bulk water at room temperature[42,57]. However, in the presence of TMU, simulation data reproduced both the properties with reasonable accuracy at low TMU concentrations (up to 20 vol%), but largely overestimated the shear viscosity at the highest TMU concentration (see the inset of Figure 4.2(a)). As a result, simulated ionic conductivity was found to be ~ 33% less than the experimental value at this highest TMU concentration as shown in Figure 4.2(a). Using the SPC/E water model and the corresponding Zn^{2+} ion parameters for the highest TMU concentration system did not improve the estimation of the shear viscosity (see Table 4.2). Therefore, we decided to work with the OPC3 water model as this water model captured qualitatively correctly the experimental TMU concentration dependent viscosity and ionic conductivity of these AZIB electrolyte systems with remarkable quantitative accuracy at low TMU concentrations.

4.3.2 Ionic conductivity, ion-ion correlations and cation transference number

Different components of ionic conductivity are plotted in Figure 4.2(b) as a function of TMU concentration. It can be observed that, due to the combined effect of different ion-ion correlations, the net ionic conductivities are smaller than the ideal NE values (σ_{NE}). Moreover, in the absence of TMU, all the three components $\sigma^{++}_{\mathrm{distinct}}$, $\sigma^{--}_{\mathrm{distinct}}$ and $2\sigma^{+-}$ are negative. This indicates the presence of anti-correlated motion between cation-cation, anion-anion and cation-anion in AZIB electrolyte system in the absence of the co-solvent, TMU. This is expected as water, being a high dielectric constant (78.4)[58] solvent, facilitates ion disassociation in electrolyte solutions and thus produces a considerable number of free ions. As shown in Eq. (2.8), while cation-anion anti-correlation has positive contribution to the net ionic conductivity, anti-correlations between like-charged ions negatively impact the conductivity. Furthermore, Figure 4.2(b) suggests that the magnitude of these correlated ion motion contributions (to conductivity) decreases with increasing TMU concentration. Therefore, the negative impact of cation-cation and anion-anion anti-correlations to the net ionic conductivity decreases at higher TMU concentrations. Moreover, positive contribution coming from cation-anion anti-correlation also decreases with increasing TMU concentration. This decrease in anti-correlations in the presence of TMU is due to slower movement of cations and anions as captured by the respective mean squared displacements (see Figure 4.A.1-4.A.5). In addition, the contributions from each of these three different ion-ion correlations, although comparable in magnitude and remains negative across the TMU concentrations studied, gradually decreases to nearly vanish at the highest TMU concentration explored.

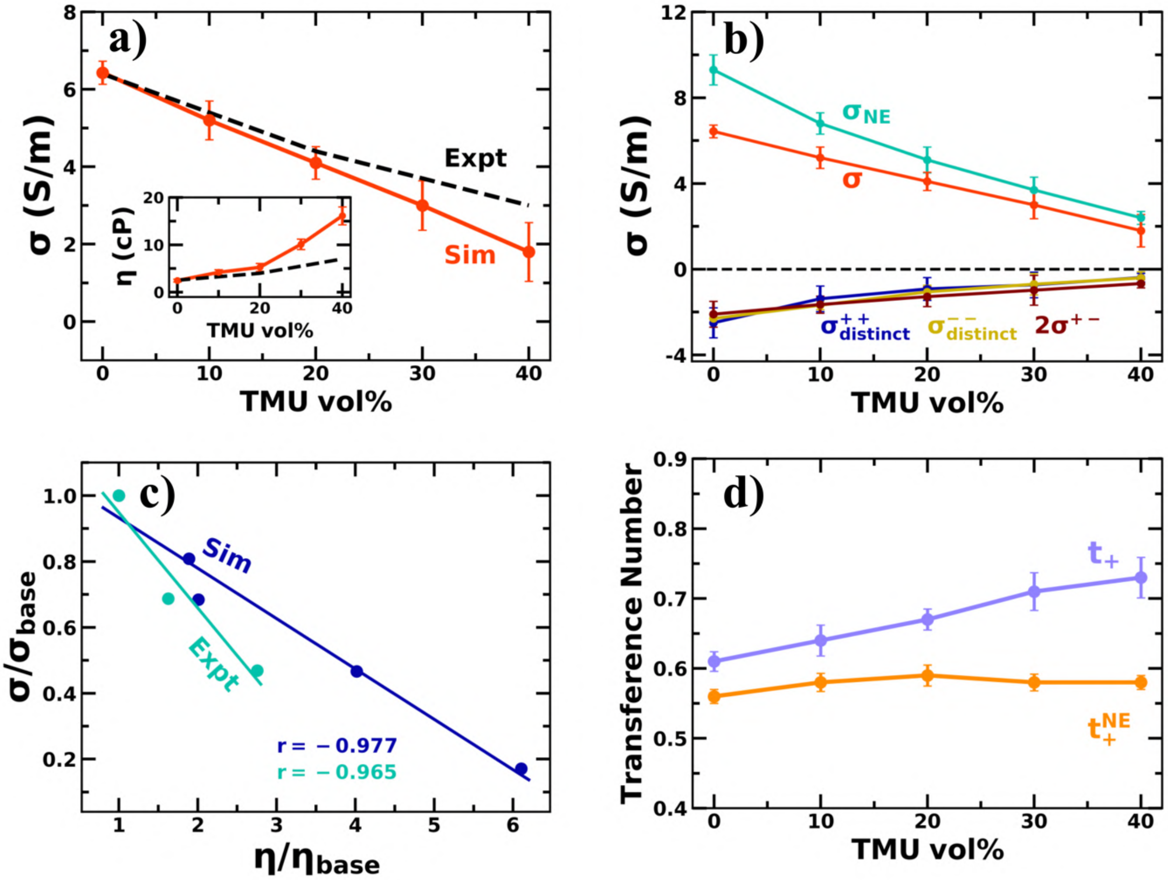


Figure 4.2. a) Simulated (Sim) ionic conductivities (σ) are plotted against TMU concentration (in vol%) and compared with the experimental[38] (Expt) values. Simulated and experimental viscosities (η) are shown in the inset. **b)** Different components of the total ionic conductivity arise due to different ion-ion correlations are plotted with respect to TMU concentration. **c)** Correlation between ionic conductivity and viscosity is shown using both of the simulated and experimental data. Quantities are normalized by the corresponding values of the base electrolyte solution (zero TMU vol%). Experimental[38] and simulated viscosities and ionic conductivities for the base electrolyte are: $\eta^{\mathrm{expt}}_{\mathrm{base}} = 2.54\ \mathrm{cP}, \eta^{\mathrm{sim}}_{\mathrm{base}} = 2.40 \pm 0.30\ \mathrm{cP}, \sigma^{\mathrm{expt}}_{\mathrm{base}} = 6.4\ \mathrm{S/m}$ and $\sigma^{\mathrm{sim}}_{\mathrm{base}} = 6.43 \pm 0.32\ \mathrm{S/m}$. Pearson correlation coefficient (r) values are also provided. **d)** Cation transference numbers are plotted as a function of TMU concentration. All error bars represent standard error of the mean.

The solution viscosity dependence of net ionic conductivity is shown in Figure 4.2(c), where both the quantities have been normalised by their respective values at the zero TMU vol%. As expected, conductivity decreases with solution viscosity. This inverse proportionality relation between ionic conductivity and viscosity is quantified by using the Pearson correlation coefficient (r) calculated using both experimental and simulation data. Similar r values suggest that our simulations correctly capture the experimental correlation between ionic conductivity and viscosity, although simulated viscosities were overestimated and conductivities were underestimated at higher TMU concentrations.

The true cation transference number t_+ that incorporates the effects of these ion-ion correlations for concentrated electrolyte solutions and its ideal solution value t_+^{NE} are calculated using Eqs. (2.9) and (2.10) and are plotted as a function of TMU concentration in Figure 4.2(d). In the absence of TMU, $t_+ \approx 0.61$ and $t_+^{NE} \approx 0.58$. These simulated transference numbers match well with the available experimental data[59]. Moreover, while t_+^{NE} remains almost unchanged irrespective of the TMU concentration, t_+ slightly increases as TMU concentration increases. This increase in Zn^{2+} transference number with increasing TMU concentration is due to the weakening of cation-anion interaction[60] with increasing TMU concentration as shown in Figure 4.2(b). Since the effects of cation-anion interaction is not taken into account in t_+^{NE}, the ideal Zn^{2+} transference number remains unchanged. Furthermore, we also observed that the values of t_+ are always larger than t_+^{NE} for all the systems and the differences become more prominent at higher TMU concentrations. Since the contributions of different ion-ion correlations are comparable in magnitude as shown in Figure 4.2(b), pairwise cancellation of their effects (Eq. (2.9)) leads to larger values for t_+ than t_+^{NE}. Therefore, while the combined effects of different ion-ion correlations decrease ionic conductivity from the ideal NE value, presence of these correlations causes the true cation transference number of AZIB electrolyte solutions to become larger than its ideal value approximation. Therefore, it can be concluded that correlated ion dynamics are playing an important role in dictating ion transport and performance of AZIB electrolytes in the presence of co-solvents.

From all these discussions, it may appear that increase of solution viscosity upon addition of the cosolvent is the reason for the observed reduction in ionic conductivity of AZIB electrolytes with increasing TMU concentration. However, this explanation lacks any microscopic insight because the presence of TMU and interaction with Zn^{2+} are likely to significantly impact both the water-water hydrogen bond (H-bond) network structure and dynamics of water molecules that are engaged in solvating the doubly charged cation. The residence time of water molecules in the solvation shell and the water-water H-bond lifetime are two particularly important factors as they may provide critical information regarding the vehicular mode[61] of Zn^{2+} transport in AZIB batteries. This molecular picture is completely missing in the mundane viscosity dependence of ionic conductivity and thus warrants a closer examination.

4.3.3 Arrangements of water molecules inside Zn^{2+} solvation shell

Here, we have investigated the effects of TMU on the microscopic arrangements of water molecules inside Zn^{2+} solvation shell. To estimate Zn^{2+} solvation shell size, we have calculated the radial distribution functions (RDFs) between $\mathrm{Zn}^{2+} - \mathrm{O(water)}$ and $\mathrm{Zn}^{2+} - \mathrm{O(TMU)}$ (see Figure 4.A.7). The minima of the RDFs appeared at 2.3 Å distance from Zn^{2+} ions. Thus, 2.3 Å was chosen as the radius of the solvation shell centred around Zn^{2+} ions for both water and TMU molecules. The coordination numbers of water and TMU molecules inside the Zn^{2+} ion solvation shell were calculated by integrating the corresponding RDFs up to $\mathrm{r} = 2.3$ Å (see Figure 4.A.7) and the results are shown in Figure 4.3(a). In the absence of TMU, Zn^{2+} ions are coordinated by ~ 6 water molecules. This observation matches with the results from several previous studies[47,62]. With the increase of TMU concentration in the solutions, TMU is gradually replacing water molecules inside the solvation shell of Zn^{2+}, but total coordination number is maintained at ~ 6. Interestingly, water is the major component inside Zn^{2+} solvation shell even at the highest TMU concentration studied (TMU vol%=40).

In order to probe the microscopic arrangements of water molecules inside Zn^{2+} solvation shell in the absence or presence of the co-solvent TMU, we have computed the tetrahedral angle distributions of water molecules that are located inside Zn^{2+} solvation shell for all the TMU concentrations and the results are shown in Figure 4.3(b). We have also added the tetrahedral angle distribution of pure bulk water in the same plot for comparison. As shown in Figure 4.3(b), the tetrahedral angle distributions of water molecules that are inside Zn^{2+} solvation shell differ significantly from that of bulk water. Two sharp peaks around 60^0 and 92^0 angles in the tetrahedral angle distributions and the complete absence of the characteristic peak[63] at $\sim 104^0$ demonstrate the significant modifications in the tetrahedral network structure of water molecules that are engaged in directly solvating Zn^{2+} cations and a representative snapshot of Zn^{2+} solvation shell is shown in Figure 4.3(c). This dramatic shift in the tetrahedral angle distribution of water molecules inside Zn^{2+} shell from that of neat water is expected due to the high charge density of Zn^{2+} ions which breaks the tetrahedral hydrogen bond network of water[64]. Moreover, these two sharp and well-separated peaks suggest that water molecules, while being engaged in solvation of Zn^{2+}, splits into two types of different orientational structures. The near-insensitivity of these two peak heights to TMU concentration may indicate that excluded volume effects due to steric repulsion along with electrostatic interactions regulate the maximum number of TMU molecules that can co-populate the Zn^{2+} solvation shell in these aqueous solutions. The interaction energy between Zn^{2+} and water (-49 kJ/mol)[65] is more than an order of magnitude higher than urea-water interaction energy (-4 kJ/mol)[66]. Thus, water interacts more strongly with Zn^{2+} ions than urea. Thus, TMU is not capable enough to modify the arrangements of water molecules inside Zn^{2+} solvation shell even if it enters into the solvation shell. TMU can only replace Zn^{2+} solvated water molecules because of its higher donor number than water[39].

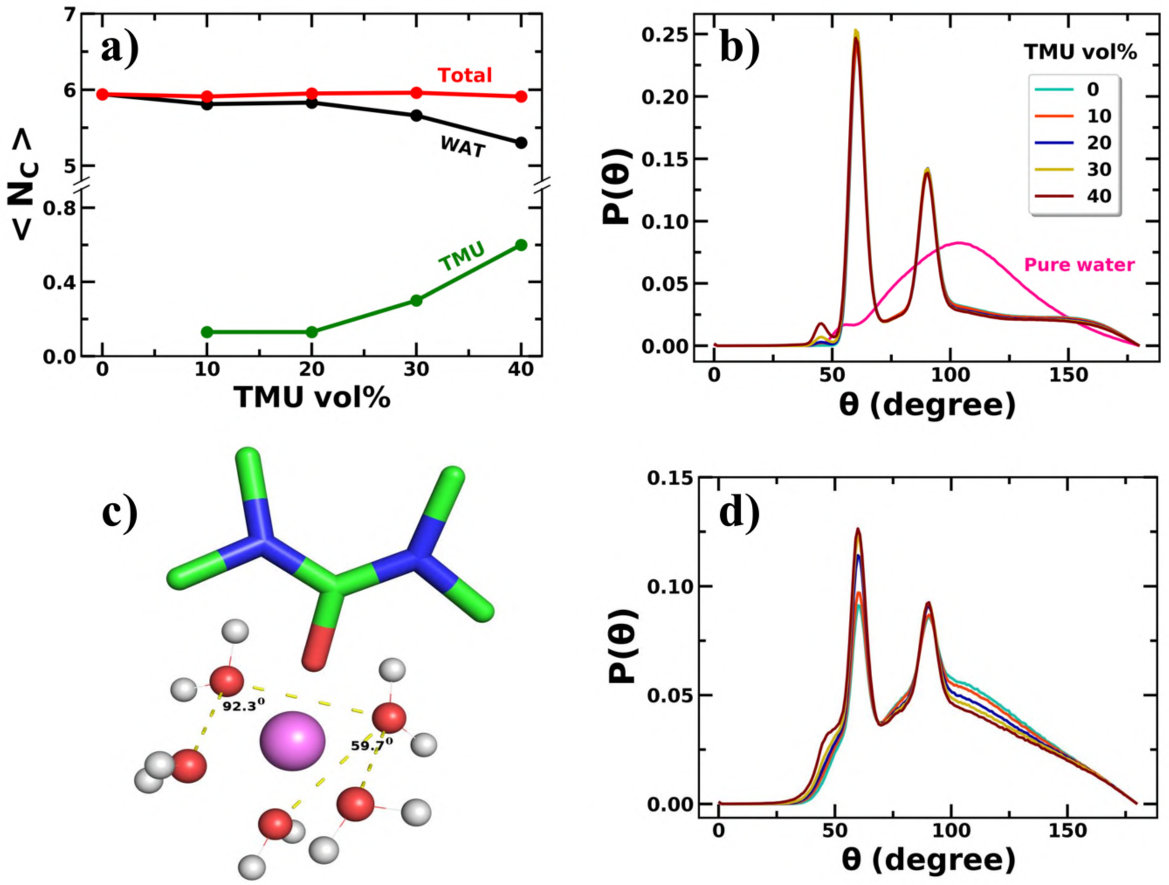


Figure 4.3. a) Coordination number of water (WAT) and tetramethylurea (TMU) molecules inside Zn^{2+} solvation shell as a function of TMU concentration (vol%). **b)** Tetrahedral angle distribution for water molecules residing inside Zn^{2+} solvation shell at different TMU concentrations. Tetrahedral angle distribution of pure bulk water is also added in the same plot for comparison. **c)** A representative snapshot of Zn^{2+} ion solvation shell is shown. The local arrangements of Zn^{2+} ion solvated water molecules (in ball-stick representation) that gives rise to the two angles where sharp peaks have been observed in the tetrahedral angle distribution are also highlighted. TMU molecules are shown in licorice representation. Zn^{2+} ion is coloured in violet, oxygen atoms are in red, carbon atoms are in green and nitrogen atoms are in blue. **d)** Tetrahedral angle distribution of water molecules residing inside the solvation shell of OTf^- ions at different TMU concentrations.

In Figure 4.3(d), we have also shown the tetrahedral angle distribution of water molecules inside the solvation shell of OTf^- ions. Again two sharp peaks in the same positions as that of Zn^{2+} have been observed, but this time the peak height around 60^0 angle slightly increases with the TMU concentration due to the much lower interaction energy between OTf^- ion and water (-8.34 kJ/mol)[67] as compared to the interaction energy between Zn^{2+} and water. However, the peak height around 92^0 remains almost unaffected.

4.3.4 Structure and dynamics of water molecules around Zn^{2+} ions and their correlations with ionic conductivity

Water molecules residing just above Zn^{2+} solvation shell are capable of forming H-bonds with water molecules that are inside Zn^{2+} solvation shell (within 2.3 Å distance from Zn^{2+}). Therefore, these outside waters can directly participate in Zn^{2+} solvation shell water exchange process by re-organising their H-bond network. The vicinity of Zn^{2+} ion is defined as a spherical region of radius 6 Å centred around Zn^{2+} ion. The average number of water-water and water-TMU H-bonds in this region are plotted in Figure 4.4(a). Figure 4.4(a) shows that the number of water-water H-bonds around Zn^{2+} ions decrease with increase in TMU concentration, whereas the average number of water-TMU H-bond increases.

To investigated the dynamical behaviour of water molecules around Zn^{2+} ions, we calculated the structural H-bond lifetime (τ_{HB}) of waters that are in the vicinity of Zn^{2+} ions using the structural H-bond autocorrelation function using Eq. (2.15). Each 600 ns long trajectories were divided into 3 equal segments and each segment was treated as independent trajectory. Structural H-bond autocorrelation functions were calculated for each independent trajectories for every system, fitted with a triple-exponential function given in Eq. (2.16) and averaged over the independent trajectories to calculate τ_{HB}. The autocorrelation functions are shown in Figure 4.A.8(a-c) and the fit parameters are provided in Table 4.A.1. The average τ_{HB} values are plotted against TMU concentration in Figure 4.4(b). We can see that the structural H-bond lifetime of water molecules in the vicinity of Zn^{2+} ion increases with TMU concentration.

We calculated the residence time (τ_{R}) of water molecules inside Zn^{2+} solvation shell using the residence time autocorrelation function given in Eq. (2.14) on the single 600 ns long trajectories and fitted with a stretched exponential function as discussed in section 2.3.1 of Chapter 2 to compute the time scales. The residence time autocorrelation functions are shown in Figure 4.A.8(d) and the fit parameters are provided in Table 4.A.2. The simulated τ_{R} values are plotted in the inset of Figure 4.4(b).

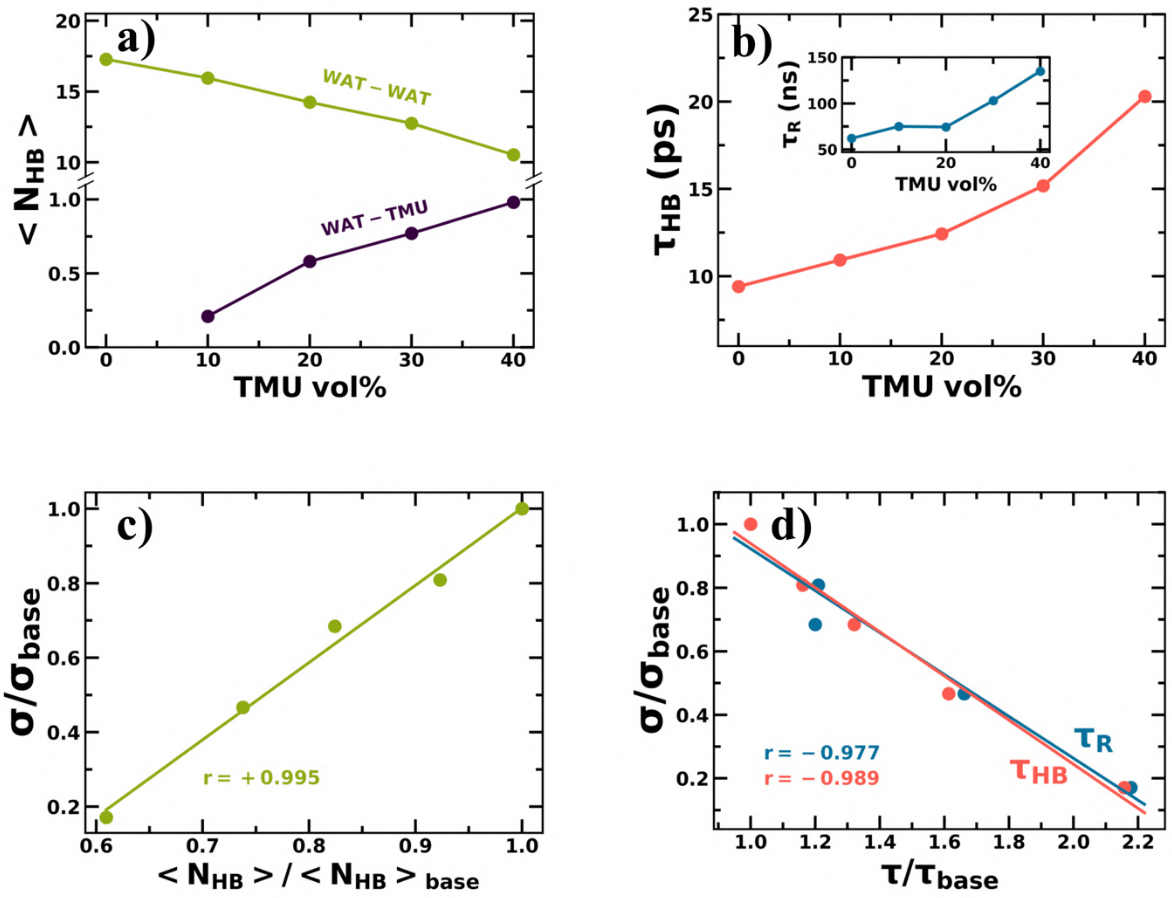


Figure 4.4. a) Average number of water-water (WAT-WAT) and water-tetramethylurea (WAT-TMU) hydrogen bonds in the vicinity of Zn^{2+} ions (within 6 Å distance from Zn^{2+}), **b)** Average structural H-bond lifetimes (τ_{HB}) of waters in the vicinity of Zn^{2+} ions are plotted as a function of TMU concentration (vol%). Symbol size represents standard error of the mean. Residence time of waters (τ_R) inside Zn^{2+} solvation shell (within 2.3 Å distance from Zn^{2+}) are shown in the inset. **c)** Correlation between ionic conductivity and average number of water-water hydrogen bonds in the vicinity of Zn^{2+} ions. **d)** Correlations between ionic conductivity and both the time scales are shown. All the quantities are normalized by the corresponding values of the base electrolyte solution, that is in the absence of TMU, $\langle N_{HB} \rangle_{base} = 17.28$, $\tau_{HB,base} = 9.41 \pm 0.02$ ps and $\tau_{R,base} = 62$ ns. Pearson correlation coefficients (r) are also provided.

It is interesting to note that water residence time inside Zn^{2+} solvation shell increases with TMU concentration. Thus, the presence of TMU decreases the number of water molecules inside Zn^{2+} solvation shell (see Figure 4.3(a)), although TMU compels water molecules to spend longer times inside Zn^{2+} solvation shell as compared to the same in the absence of TMU. Therefore, in the presence of

TMU, Zn^{2+} solvation shell water molecules remain H-boned with their neighbouring water molecules for longer durations, making water exchange inside Zn^{2+} solvation shell slower, in spite of the fact that the number of water-water hydrogen bond decreases. This suggests the vehicular mode of ion transport where Zn^{2+} moves with its solvation shell that contains both TMU and water[61].

In Figure 4.4(c), we have shown the correlation between ionic conductivity and average number of water-water H-bonds in the vicinity of Zn^{2+} ions with r ~ +0.99. This signifies strong positive correlation between these two quantities. This means that water-water H-bond in the vicinity of Zn^{2+} ions is critically important for Zn^{2+} mobility in AZIB in the presence of TMU. Similarly, correlations between ionic conductivity and τ_R and, τ_{HB} are shown in Figure 4.4(d). Figure 4.4(d) indicates that the TMU concentration dependency of ionic conductivity is strongly anti-correlated to these two microscopic timescales. This anti-correlation between co-solvent dependent behaviour of ionic conductivity and water H-bond lifetime in the vicinity of Zn^{2+} ions can be understood on a physical ground as follows: strong interaction between Zn^{2+} and its solvation shell water molecules promotes Zn^{2+} transport in a vehicular fashion, that is, Zn^{2+} moves along with its solvation shell water molecules[68,69]. With increase in TMU concentration, as the solvation shell water molecules remain H-bonded to their neighbouring water molecules for longer durations, this hinders the movement of Zn^{2+} ions. Macroscopically, this is reflected in the overall rise in solution viscosity with increasing TMU concentration (see Figure 4.2(a)). Subsequently, this causes slower ion transport and reduces ionic conductivity.

4.4 Conclusions

Addition of co-solvents with donor number higher than water in AZIB electrolytes is a promising approach to prevent Zn dendrite formation and Zn metal anode corrosion. In this work, we have provided a molecular level picture of how the presence of such co-solvents affect the structure and dynamics of water molecules around Zn^{2+} ions, highlighted the correlations between ionic conductivity and co-solvent induced modifications in water structure and dynamics. In addition, we have discussed the importance of various ion-ion correlations in regulating solution ionic conductivity and the true transference number of AZIB electrolyte solutions. We have performed all-atom classical molecular dynamics simulations of a representative AZIB electrolyte system: 2M $Zn(OTf)_2$ aqueous solution with varying concentration of tetramethylurea as a co-solvent at room temperature.

Structural analyses have shown that while tetramethylurea replaces water molecules inside Zn^{2+} solvation shell, its presence does not further alter the microscopic orientational arrangements of the water molecules surrounding Zn^{2+}ions. Furthermore, in the presence of tetramethylurea, the residence time of water molecules inside Zn^{2+} solvation shell increases due to the increase in H-bond lifetime between the solvation shell water molecules and their immediate neighbouring water molecules. Moreover, average number of water-water H-bonds around Zn^{2+} ions decreases with increase in tetramethylurea concentration.

Moreover, we have observed that ionic conductivity is positively correlated to the number of water-water H-bonds around Zn^{2+} ions, but anti-correlated to both the Zn^{2+} solvation shell water residence times and water-water H-bond lifetimes. The increase in water-water H-bond lifetimes around Zn^{2+} ions supports the notion of vehicular movement of Zn^{2+} in these solutions. The TMU concentration dependence of water H-bond lifetimes and residence times provides a microscopic explanation to the experimentally observed viscosity dependence of ionic conductivity for the studied system. Furthermore, we have also investigated the role of ion-ion correlations using the Onsager's transport coefficients in dictating ionic conductivity and the transference number of AZIB electrolytes. Our results, therefore, have not only provided microscopic insights into the co-solvent induced change in the solution conductivity in terms of modulated water structure and dynamics but also revealed an important relationship that macroscopic ionic conductivity constructs with the microscopic water H-bond network and dynamics. These findings probably suggest extensive and deeper investigations into co-solvent induced changes in bulk AZIB electrolytes as well as in the presence of electrochemical interfaces to develop more effective electrolyte engineering strategies.

Appendix 4.A

Table 4.A.1. Parameters obtained after fitting the structural H-bond auto correlation functions $C_{HB}(t)$. The structural H-bond lifetimes (τ_{HB}) were computed using Eq. (2.16) .

Set-1							
System (TMU vol%)	a_1	τ_1 (ps)	a_2	τ_2 (ps)	a_3	τ_3 (ps)	τ_{HB} (ps)
0	0.07	0.69	0.56	7.78	0.37	13.57	9.43
10	0.07	0.75	0.63	8.93	0.30	18.15	10.97
20	0.06	0.77	0.60	9.47	0.34	19.71	12.40
30	0.06	0.85	0.62	10.78	0.32	25.62	15.02
40	0.06	0.98	0.63	12.75	0.31	38.12	20.10
Set-2							
0	0.07	0.67	0.47	7.31	0.46	12.78	9.36
10	0.05	0.74	0.67	8.93	0.28	18.15	10.94
20	0.06	0.76	0.62	9.45	0.32	19.72	12.39
30	0.05	0.86	0.64	10.78	0.31	25.65	15.06
40	0.05	1.04	0.64	12.78	0.31	38.25	20.15
Set-3							
0	0.07	0.71	0.55	7.75	0.38	13.46	9.42
10	0.07	0.81	0.65	8.96	0.28	18.03	10.90
20	0.06	0.75	0.61	9.54	0.33	20.12	12.38
30	0.06	0.88	0.62	10.84	0.32	26.75	15.43
40	0.05	0.98	0.60	12.54	0.35	36.17	20.02

Table 4.A.2. Fit parameters in the equation: $C(t) = e^{-(t/\tau_R)^{\beta}}$ used to fit the simulated residence time autocorrelation functions. Here τ_R denotes the water residence time inside Zn^{2+} solvation shell.

System (TMU vol%)	β	τ_R (ns)
0	0.973	63.58
10	0.963	74.15
20	0.989	75.02
30	0.977	102.91
40	0.932	132.52

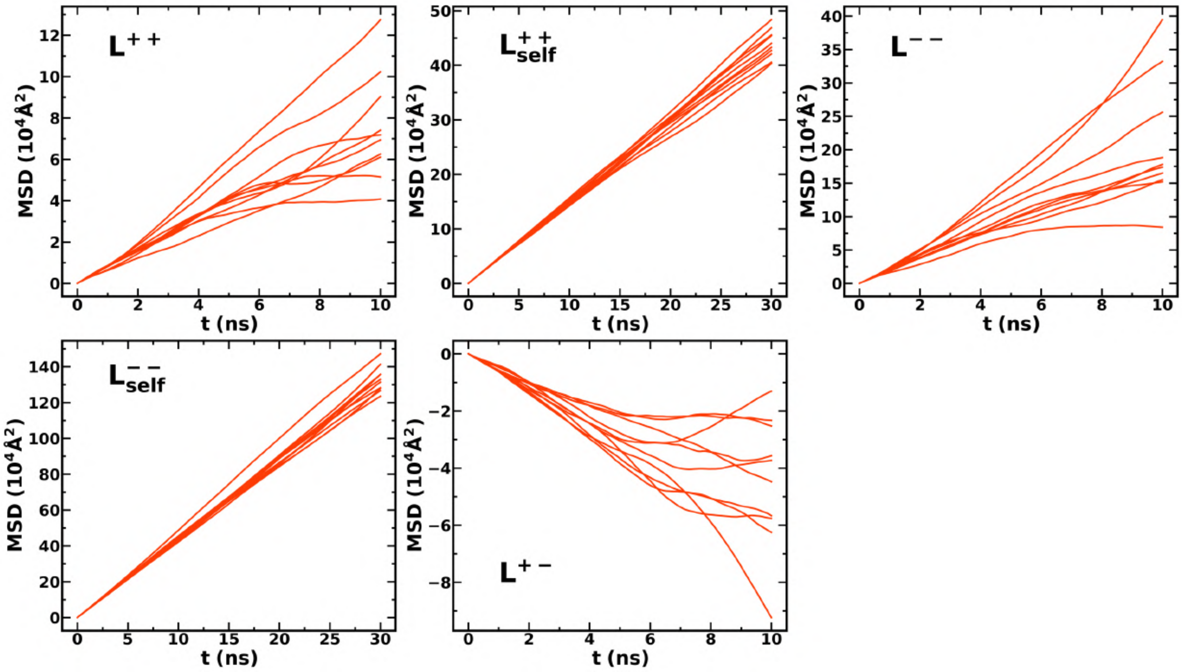


Figure 4.A.1. Collective mean squared displacement lines (MSD in general term) of 10 independent trajectories for the TMU vol% = 0 system. Slopes of these lines were used to calculate the Onsager coefficients.

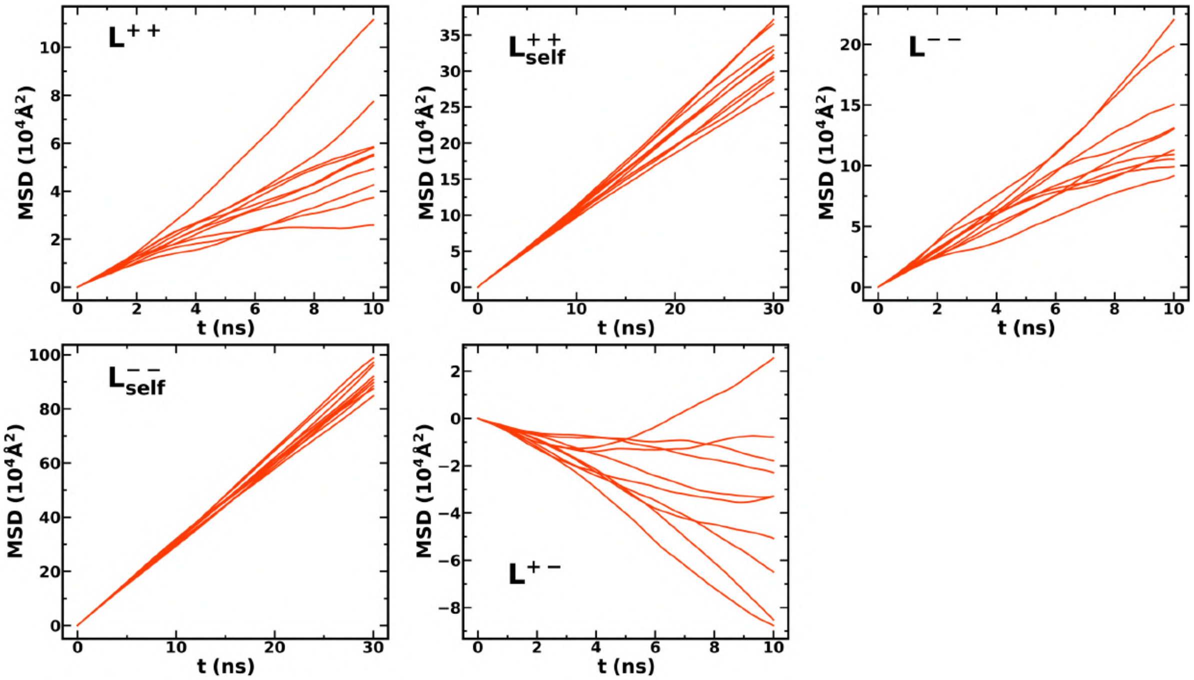


Figure 4.A.2. Collective mean squared displacement lines (MSD in general term) of 10 independent trajectories for the TMU vol% = 10 system. Slopes of these lines were used to calculate the Onsager coefficients.

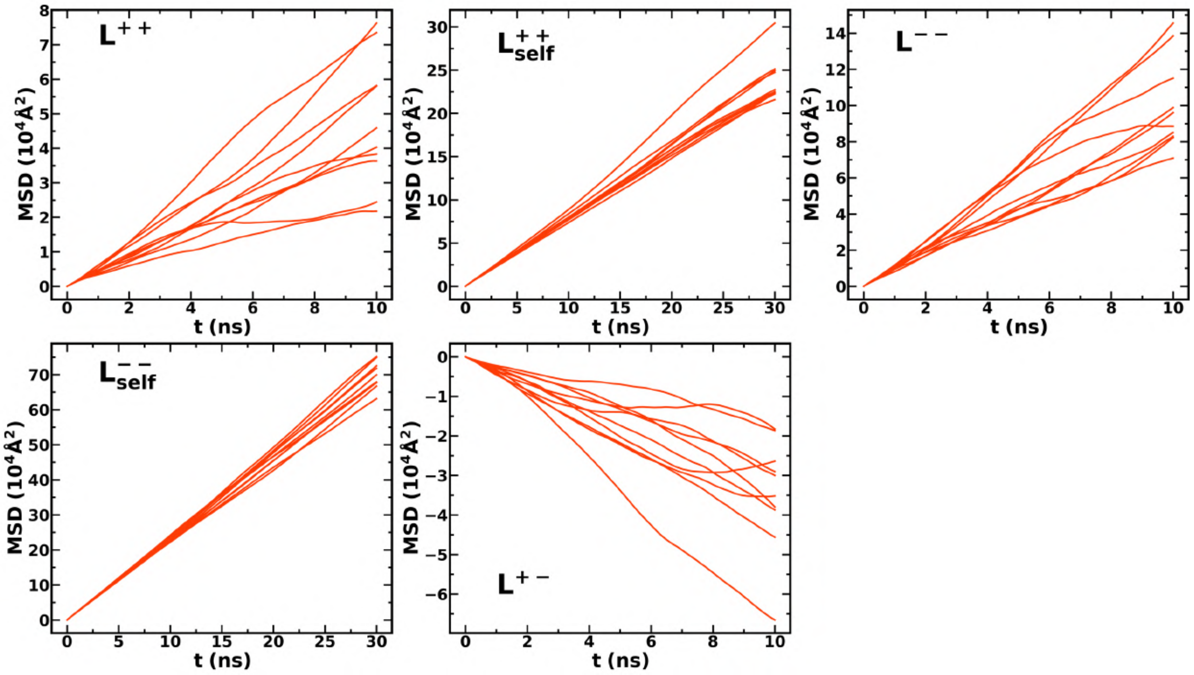


Figure 4.A.3 Collective mean squared displacement lines (MSD in general term) of 10 independent trajectories for the TMU vol% = 20 system. Slopes of these lines were used to calculate the Onsager coefficients.

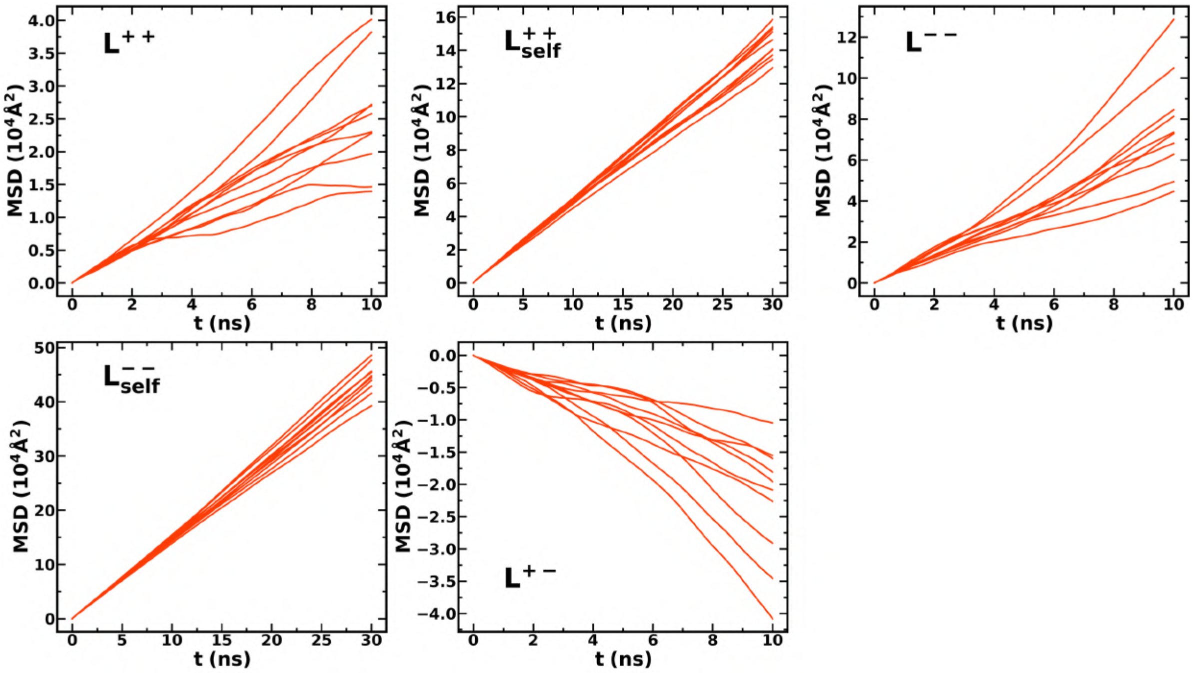


Figure 4.A.4 Collective mean squared displacement lines (MSD in general term) of 10 independent trajectories for the TMU vol% = 30 system. Slopes of these lines were used to calculate the Onsager coefficients.

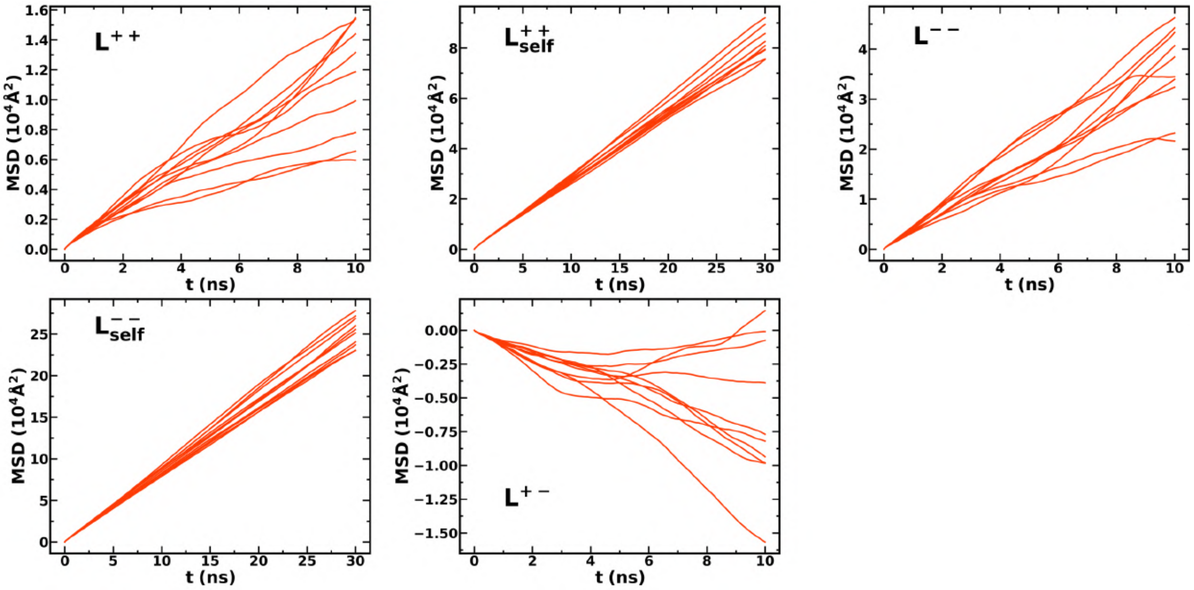


Figure 4.A.5 Collective mean squared displacement lines (MSD in general term) of 10 independent trajectories for the TMU vol% = 40 system. Slopes of these lines were used to calculate the Onsager coefficients.

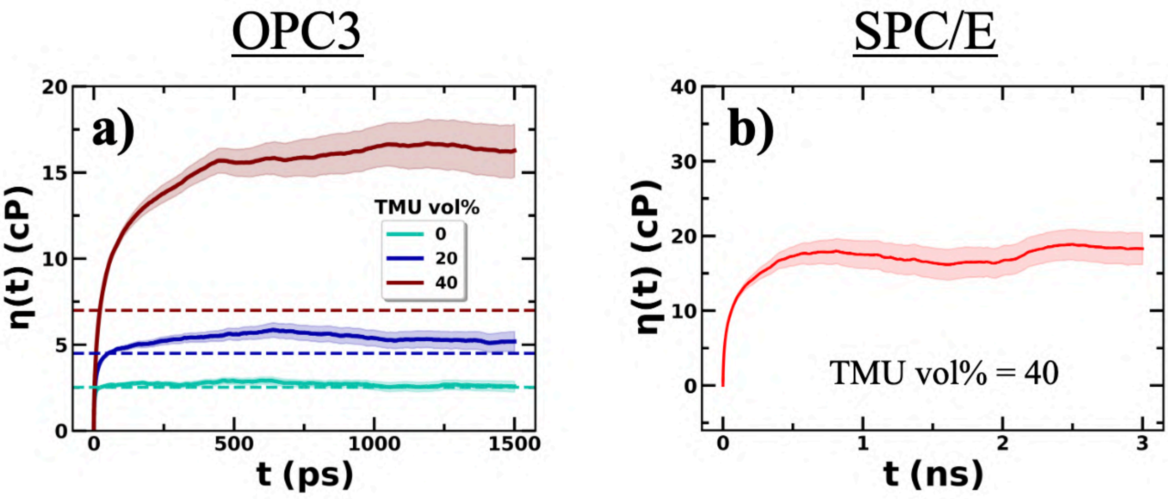


Figure 4.A.6. Dependence of the simulated shear viscosities on the upper limit of the Green-Kubo integration, denoted as η(t), averaged over 40 independent trajectories are shown for **a)** various electrolyte systems modelled with OPC3 water and **b)** the TMU vol% = 40 electrolyte system modelled with the SPC/E water. The shaded region represents the standard error of the mean. Dashed lines represent experimental values.

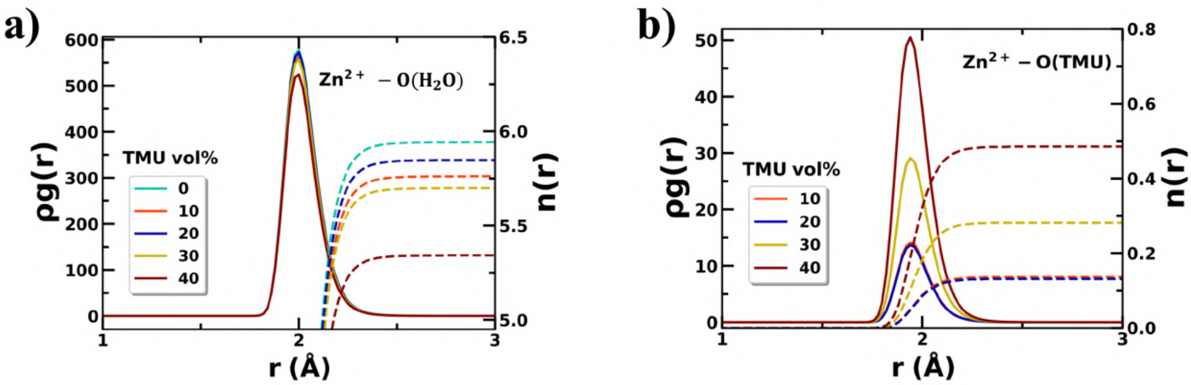


Figure 4.A.7. The radial distribution functions g(r) multiplied by the solvent molecule number densities (ρ) between **a)** Zn^{2+} and oxygen (O) atom of water (H_2O), denoted as $Zn^{2+} - O(H_2O)$ and **b)** Zn^{2+} and O atom of TMU, denoted as $Zn^{2+} - O(TMU)$ are plotted. The average number of solvent molecules $n(r) = \int_0^r dr\, \rho\, g(r)$ at a distance r from Zn^{2+} ions are plotted for different values of r. Coordination number of H_2O and TMU inside the Zn^{2+} solvation sheath is the value of the integral up to $r = 2.3$ Å, corresponds to the minimum of the corresponding g(r).

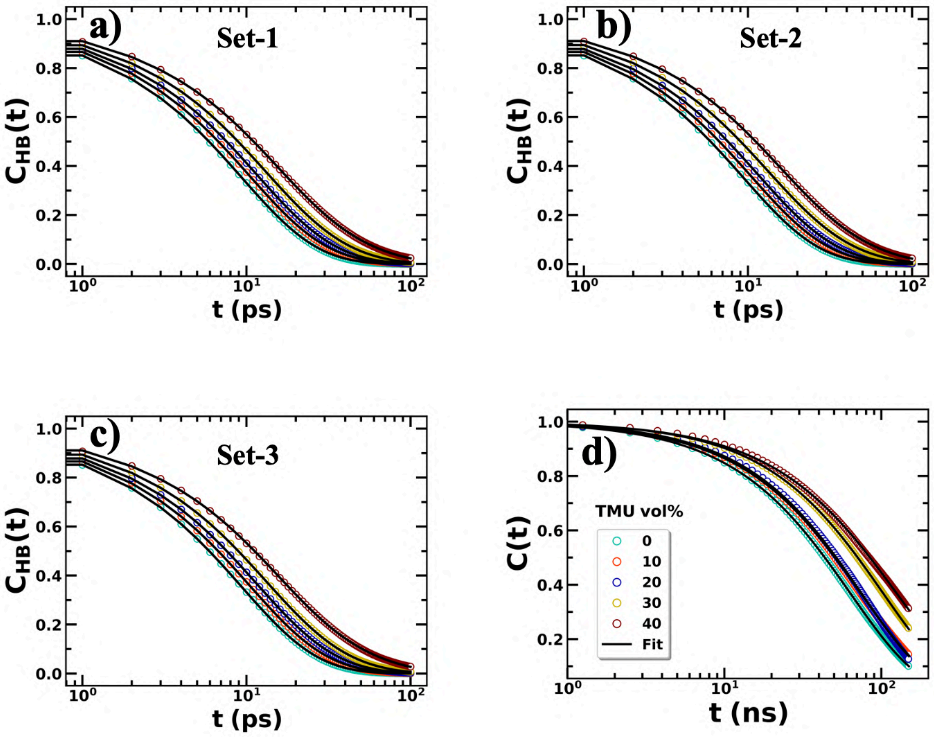


Figure 4.A.8. a)-c) Structural hydrogen bond autocorrelation functions, $C_{HB}(t)$ and **d)** residence time autocorrelation functions, C(t) are plotted. Fittings are also shown. X-axes are in log scale.

References:

1 O. Schmidt, A. Hawkes, A. Gambhir and I. Staffell, *Nat Energy*.

2 E. Telaretti and L. Dusonchet, *Renewable and Sustainable Energy Reviews*, 2017, **75**, 380–392.

3 D. Larcher and J. M. Tarascon, *Nature Chemistry 2014 7:1*, 2014, **7**, 19–29.

4 M. Armand and J. M. Tarascon, *Nature 2008 451:7179*, 2008, **451**, 652–657.

5 Y. Gogotsi and P. Simon, *Science (1979)*, 2011, **334**, 917–918.

6 B. Dunn, H. Kamath and J. M. Tarascon, *Science (1979)*, 2011, **334**, 928–935.

7 M. Winter and R. J. Brodd, *Chem Rev*, 2004, **104**, 4245–4269.

8 T. Kim, W. Song, D. Y. Son, L. K. Ono and Y. Qi, *J Mater Chem A Mater*, 2019, **7**, 2942–2964.

9 J. M. Tarascon and M. Armand, *Nature*, 2001, **414**, 359–367.

10 S. W. Kim, D. H. Seo, X. Ma, G. Ceder and K. Kang, *Adv Energy Mater*, 2012, **2**, 710–721.

11 S. Chen, C. Wu, L. Shen, C. Zhu, Y. Huang, K. Xi, J. Maier and Y. Yu, *Advanced Materials*, 2017, **29**, 1700431.

12 J. C. Pramudita, D. Sehrawat, D. Goonetilleke and N. Sharma, *Adv Energy Mater*.

13 J. W. Choi and D. Aurbach, *Nat Rev Mater*.

14 J. Liu, C. Xu, Z. Chen, S. Ni and Z. X. Shen, *Green Energy & Environment*, 2018, **3**, 20–41.

15 L. Li, Q. Zhang, B. He, R. Pan, Z. Wang, M. Chen, Z. Wang, K. Yin, Y. Yao, L. Wei, L. Sun, L. Li, Q. C. Zhang, R. Pan, K. B. Yin, L. T. Sun, B. He, Z. X. Wang, M. X. Chen, Z. Wang, L. Wei and Y. G. Yao, *Advanced Materials*, 2022, **34**, 2104327.

16 Y. Wang, J. Yi and Y. Xia, *Adv Energy Mater*, 2012, **2**, 830–840.

17 G. Fang, J. Zhou, A. Pan and S. Liang, *ACS Energy Lett*, 2018, **3**, 2480–2501.

18 C. Li, S. Jin, L. A. Archer and L. F. Nazar, *Joule*, 2022, **6**, 1733–1738.

19 N. Zhang, X. Chen, M. Yu, Z. Niu, F. Cheng and J. Chen, *Chem Soc Rev*, 2020, **49**, 4203–4219.

20 X. Jia, C. Liu, Z. G. Neale, J. Yang and G. Cao, *Chem Rev*, 2020, **120**, 7795–7866.

21 A. Konarov, N. Voronina, J. H. Jo, Z. Bakenov, Y. K. Sun and S. T. Myung, *ACS Energy Lett*, 2018, **3**, 2620–2640.

22 H. Jia, Z. Wang, B. Tawiah, Y. Wang, C. Y. Chan, B. Fei and F. Pan, *Nano Energy*, 2020, **70**, 104523.

23 B. Tang, L. Shan, S. Liang and J. Zhou, *Energy Environ Sci*, 2019, **12**, 3288–3304.

24 J. Shin, J. Lee, Y. Park and J. W. Choi, *Chem Sci*, 2020, **11**, 2028–2044.

25 W. Du, E. H. Ang, Y. Yang, Y. Zhang, M. Ye and C. C. Li, *Energy Environ Sci*, 2020, **13**, 3330–3360.

26 C. Li, X. Zhang, W. He, G. Xu and R. Sun, *J Power Sources*, 2020, **449**, 227596.

27 Y. Zhu, J. Yin, X. Zheng, A. H. Emwas, Y. Lei, O. F. Mohammed, Y. Cui and H. N. Alshareef, *Energy Environ Sci*, 2021, **14**, 4463–4473.

28 Y. Chai, X. Xie, Z. He, G. Guo, P. Wang, Z. Xing, B. Lu, S. Liang, Y. Tang and J. Zhou, *Chem Sci*, 2022, **13**, 11656–11665.

29 X. Xie, S. Liang, J. Gao, S. Guo, J. Guo, C. Wang, G. Xu, X. Wu, G. Chen and J. Zhou, *Energy Environ Sci*, 2020, **13**, 503–510.

30 J. Li, B. He, Y. Zhang, Z. Cheng, L. Yuan, Y. Huang, Z. Li, J. Li, B. He, Y. Zhang, Z. Cheng, L. Yuan, Y. Huang and Z. Li, *Small*, 2022, **18**, 2200567.

31 J. Hao, L. Yuan, C. Ye, D. Chao, K. Davey, Z. Guo and S. Z. Qiao, *Angewandte Chemie International Edition*, 2021, **60**, 7366–7375.

32 Y. Geng, L. Pan, Z. Peng, Z. Sun, H. Lin, C. Mao, L. Wang, L. Dai, H. Liu, K. Pan, X. Wu, Q. Zhang and Z. He, *Energy Storage Mater*, 2022, **51**, 733–755.

33 B. Kakoty, R. Vengarathody, S. Mukherji, V. Ahuja, A. Joseph, C. Narayana, S. Balasubramanian and P. Senguttuvan, *J Mater Chem A Mater*, 2022, **10**, 12597–12607.

34 F. Ming, Y. Zhu, G. Huang, A. H. Emwas, H. Liang, Y. Cui and H. N. Alshareef, *J Am Chem Soc*, 2022, **144**, 7160–7170.

35 Y. Zhu, J. Hao, Y. Huang and Y. Jiao, *Small Struct*, 2023, **4**, 2200270.

36 Z. Tian, H. Liu, M. Cheng, L. Cui, R. Zhang, X. Yang, D. Wu, D. Wang and J. Xia, *ACS Appl Mater Interfaces*, 2024, **16**, 21857–21867.

37 L. Cao, D. Li, E. Hu, J. Xu, T. Deng, L. Ma, Y. Wang, X. Q. Yang and C. Wang, *J Am Chem Soc*, 2020, **142**, 21404–21409.

38 Z. Li, Y. Liao, Y. Wang, J. Cong, H. Ji, Z. Huang and Y. Huang, *Energy Storage Mater*, 2023, **56**, 174–182.

39 J. Yang, Y. Zhang, Z. Li, X. Xu, X. Su, J. Lai, Y. Liu, K. Ding, L. Chen, Y.-P. Cai, Q. Zheng, J. Yang, Z. Li, X. Xu, X. Su, J. Lai, Y. Liu, K. Ding, L. Chen, Y.-P. Cai, Q. Zheng and Y. Zhang, *Adv Funct Mater*, 2022, **32**, 2209642.

40 P. Cui, T. Wang, Z. Wang, H. Geng, P. Song, F. Hu, J. You and K. Zhu, *Chemical Engineering Journal*, 2024, **500**, 156971.

41 M. Kim, M. Lee, I. Choi, J. Oh, S. Paik, A. Han, S. Lee, H. Hwang, J. Na and K. W. Nam, *Small*, 2025, 2411632.

42 S. Izadi and A. V. Onufriev, *Journal of Chemical Physics*, 2016, **145**, 74501.

43 Z. Li, L. F. Song, P. Li and K. M. Merz, *J Chem Theory Comput*, 2020, **16**, 4429–4442.

44 P. Belletato, L. C. G. Freitas, E. P. G. Arêas and P. S. Santos, *Physical Chemistry Chemical Physics*, 1999, **1**, 4769–4776.

45 A. Kuffel and J. Zielkiewicz, *Journal of Chemical Physics*, 2010, **133**, 49.

46 P. Mark and L. Nilsson, *Journal of Physical Chemistry A*, 2001, **105**, 9954–9960.

47 P. Li, B. P. Roberts, D. K. Chakravorty and K. M. Merz, *J Chem Theory Comput*, 2013, **9**, 2733–2748.

48 J. N. Canongia Lopes and A. A. H. Pádua, *J Phys Chem B*, 2004, **108**, 16893–16898.

49 Y. Zhang, E. J. Maginn, S. Tepavcevic, E. Carino, N. T. Hahn, N. Becknell, J. Mars, K. S. Han, K. T. Mueller and M. Toney, *Journal of Physical Chemistry Letters*, 2023, **14**, 11393–11399.

50 Y. Zhang, G. Wan, N. H. C. Lewis, J. Mars, S. E. Bone, H. G. Steinrück, M. R. Lukatskaya, N. J. Weadock, M. Bajdich, O. Borodin, A. Tokmakoff, M. F. Toney and E. J. Maginn, *ACS Energy Lett*, 2021, **6**, 3458–3463.

51 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19–25.

52 L. Martinez, R. Andrade, E. G. Birgin and J. M. Martinez, *J Comput Chem*, 2009, **30**, 2157–2164.

53 G. Bussi, D. Donadio and M. Parrinello, *J Chem Phys*, 2007, **126**, 14101.

54 H. J. C. Berendsen, J. P. M. van Postma, W. F. Van Gunsteren, A. DiNola and J. R. Haak, *J Chem Phys*, 1984, **81**, 3684–3690.

55 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J Chem Phys*, 1995, **103**, 8577–8593.

56 N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, *J Comput Chem*, 2011, **32**, 2319–2327.

57 T. Ando, *Journal of Chemical Physics*, 2023, **159**, 101102.

58 P. Jungwirth, J. K. Beattie, A. M. Djerdjev, G. G. Warr, K. Hänni-Ciunel, N. Schelero, R. Von Klitzing, D. Bratko, C. D. Daub, A. Luzar, O. Link, E. Vöhringer-Martinez, E. Lugovoj, Y. Liu, K. Siefermann, M. Faubel, H. Grubmüller, R. B. Gerber, Y. Miller, B. Abel, D. J. Tobias, N. Sengupta, M. Tarek, A. Frölich, F. Gabel, M. Jasnin, U. Lehnert, D. Oesterhelt, A. M. Stadler, M. Tehei, M. Weik, K. Wood, G. Zaccai, J. Qvist, E. Persson, C. Mattea, B. Halle, A. Ghosh, R. Kramer Campen, M. Sovago, M. Bonn, B. Born, S. J. Kim, S. Ebbinghaus, M. Gruebele, M. Havenith, A. P. Willard, D. Chandler, M. Salmeron, H. Bluhm, M. Tatarkhanov, G. Ketteler, T. K. Shimizu, A. Mugarza, X. Deng, T. Herranz, S. Yamamoto, A. Nilsson, M. Gallagher, A. Omer, G. R. Darling, A. Hodgson, C. L. Bishop, D. Pan, L. M. Liu, G. A. Tribello, A. Michaelides, E. G. Wang, B. Slater, M. Bertin, M. Meyer, J. Stähler, C. Gahl, M. Wolf, U. Bovensiepen, M. A. Ricci, F. Bruni, A. Giuliani, T. G. Lombardo, N. Giovambattista, P. G. Debenedetti, F.-X. Coudert, F. Cailliez, R. Vuilleumier, A. H. Fuchs, A. Boutin, S. Perkin, R. Goldberg, L. Chai, N. Kampf, J. Klein, M. Lee, B. Sung, N. Hashemi, W. Jhe, S. K. Reed, P. A. Madden, C. Vega, J. L. F. Abascal, M. M. Conde and J. L. Aragones, *Faraday Discuss*, 2008, **141**, 251–276.

59 X. Li, H. Yao, Y. Li, X. Liu, D. Yuan, Y. Chen, M. W. Wong, Y. Zhang and H. Zhang, *J Mater Chem A Mater*, 2023, **11**, 14720–14727.

60 J. Cong, Y. Wang, X. Lin, Z. Huang, H. Wang, J. Li, L. Hu, H. Hua, J. Huang, Y. C. Lin, H. Xu, Z. Li and Y. Huang, *J Am Chem Soc*, 2025, **147**, 8607–8617.

61 A. Mistry, Z. Yu, L. Cheng and V. Srinivasan, *J Electrochem Soc*, 2023, **170**, 110536.

62 Y. Marcus, *Chem Rev*, 1988, **88**, 1475–1498.

63 N. C. Maity, A. Baksi, K. Kumbhakar and R. Biswas, *J Photochem Photobiol A Chem*, 2023, **439**, 114600.

64 S. Yadav and A. Chandra, *Journal of Chemical Physics*.

65 K. Qiu, G. Ma, Y. Wang, M. Liu, M. Zhang, X. Li, X. Qu, W. Yuan, X. Nie and N. Zhang, *Adv Funct Mater*, 2024, **34**, 2313358.

66 P. Adhikary, K. D. Reddy and R. Biswas, *Chemical Physics Impact*, 2024, **8**, 100609.

67 I. Khan, K. A. Kurnia, F. Mutelet, S. P. Pinho and J. A. P. Coutinho, *Journal of Physical Chemistry B*, 2014, **118**, 1848–1860.

68 O. Borodin and G. D. Smith, *J Solution Chem*, 2007, **36**, 803–813.

69 E. Crabb, A. Aggarwal, R. Stephens, Y. Shao-Horn, G. Leverick and J. C. Grossman, *Journal of Physical Chemistry B*, 2024, **128**, 3427–3441.

5

Molecular thermodynamic origin of substrate promiscuity in the enzyme laccase: Toward a broad spectrum degrader of dye effluents

5.1 Introduction

Urbanization and industrialization are causing serious contamination of water bodies, resulting in harmful effects not only to the ecosystem but also to the health of all living creatures[1]. Among all the pollutants that are discharged into water bodies, textile industry effluents are complex and contain a wide variety of dyes[2]. Dyes are also used in cosmetics, pharmaceuticals, paper, leather etc[3–5]. These dye effluents are often carcinogenic. Furthermore, the absorption of light by textile dyes creates problems to photosynthetic aquatic plants and algae[6]. This emerging problem necessitates the need for new ways to treat and degrade these effluents[7,8].

Interestingly, white rot fungi are known to be some of the most efficient microorganisms capable of breaking down synthetic dyes[9]. They generate a class of enzymes called laccases (EC 1.10.3.2), which belong to the family of multicopper oxidases (MCOs)[10]. They can oxidize a variety of phenolic and non-phenolic compounds using oxygen as reactant and generate water as the only by-product[11]. That's why laccases are also known as "green catalysts". Due to the easy availability and capability of degrading a variety of synthetic dyes, implementation of laccase mediated processes for treating and improving water quality offers a promising economical and ecofriendly solution to a complex problem[12–16].

The ability of an enzyme to act upon a broad range of substrates is known as 'substrate promiscuity'[17,18]. Previous studies have suggested that hydrophobicity, flexibility and protonation states of the active site residues could play an important role in enzyme promiscuity[19]. A recent study[20] has proposed that substrate promiscuity is a continuum feature of enzymes and the presence of various conformational states of the active site that are capable of accommodating ligands with different shapes and sizes may be one of the mechanisms to achieve substrate promiscuity.

In this chapter, we discussed why laccases are substrate promiscuous by studying the binding of several dye molecules to the known active site of a laccase from the white-rot fungus Tramates Versicolor (TvL) (PDB: 1KYA)[21] as shown in Figure 5.1 using experimental, molecular docking and MD simulations. It is important to note that this crystallize structure of the laccase was in its active state, that is, bound to a ligand named 2,5-xylidine. The active site of this laccase from TvL made up of the following residues[22]: D150, A161, F162, P163, L164, D206, G392, H458, F332 and F337. Among these residues A161, F162, P163, L164 and G392 are non-polar, and D150 and D206 are negatively charged at pH 7. The catalytic site of the protein consists of four copper atoms. The one close to the active site is known as T1 copper and the other three copper atoms form a trinuclear copper cluster (TNC)[23]. T1 copper is connected with the TNC via a conserved cysteine-histidine bridge[24]. This T1 copper is involved in the oxidation of substrates[25,26]. Earlier studies have reported that H458 and D206 residues are important for the catalytic function of the enzyme[27]. These residues are involved in hydrogen bonding and π-stacking interactions with the substrates.

We have selected a set of five dye molecules with large variation in their net charge, size and shape: (i) Methyl Green, (ii) Crystal Violet, (iii) Thioflavin T, (iv) Coumarin 343, and (v) Brilliant Blue. Their chemical structures and net charges are shown in Figure 5.2.

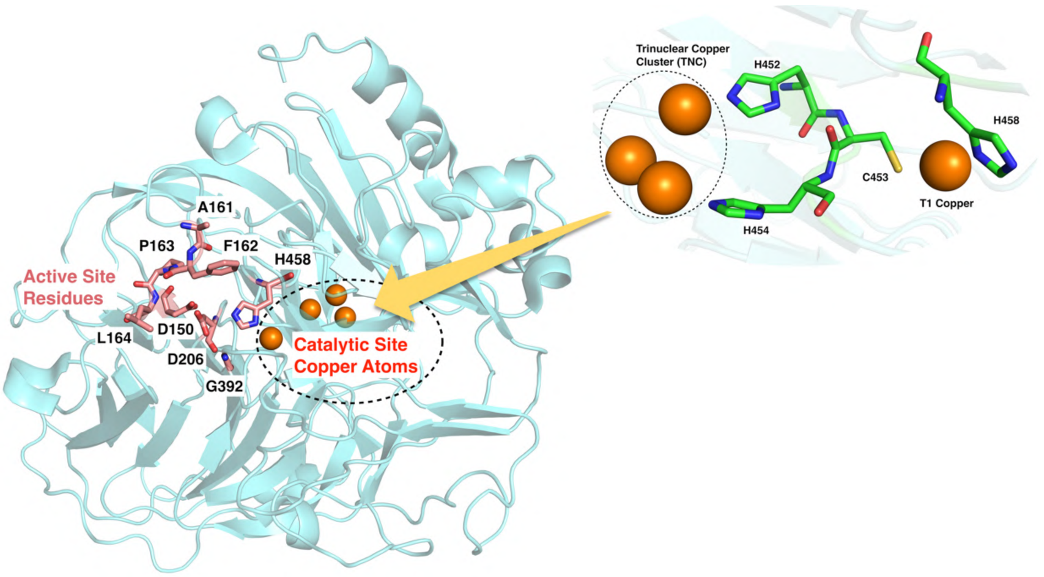


Figure 5.1: Crystallographic structure of the laccase from TvL (PDB ID: 1KYA). The active site residues, T1 copper and the trinuclear copper cluster (TNC) are shown separately (inset).

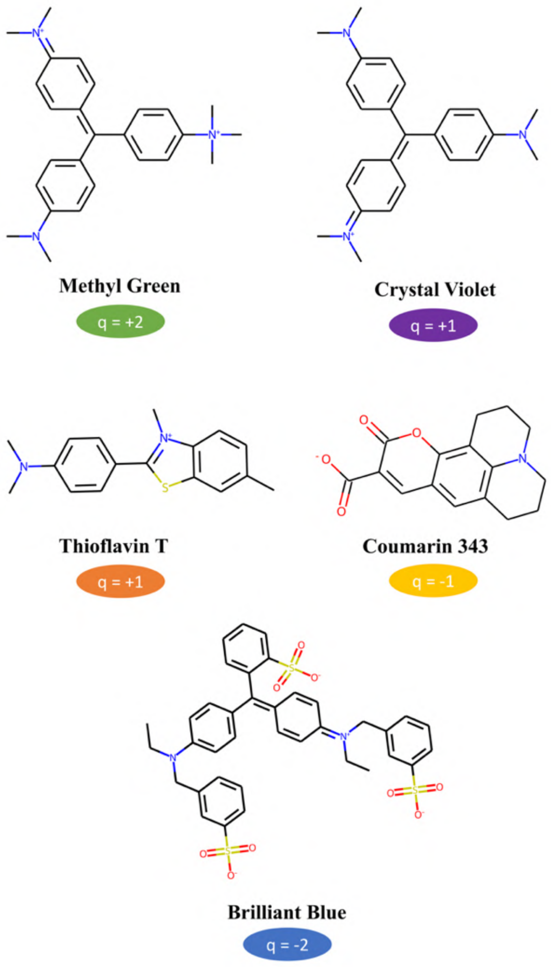


Figure 5.2: Chemical structures of the dye molecules used in this study.

5.2 Experimental and computational details

5.2.1 Experimental details

TvL Laccase and the dyes were purchased from Sigma-Aldrich. Stock solutions of laccase (concentration 8 mg/mL) were prepared in deionized water and stored at 4^0C[28,29]. The solutions were centrifuged to remove insoluble materials. An UV−visible spectrophotometer (UV-2600, Shimadzu) was used for recording steady state absorption. Solution concentrations for coumarin-343, brilliant blue, thioflavin T and methyl green were 10 μM, and solution concentration for crystal violet was 1 μM. These solution concentrations were used to maintain a uniform optical density (OD ~ 0.2) in all our experiments. 300 μL of laccase was added to the coumarin-343 solution, 100 μL of laccase was added to brilliant blue solution and 50 μL of laccase was added to crystal violet, methyl green and thioflavin T solution.

5.2.2 Molecular docking

Autodock Vina v1.2.2[30] was used to perform the molecular docking of the dye molecules into the active site of the laccase. The 3D structures of the dyes were downloaded from PubChem. The receptor grid used in molecular docking was defined based on the active site residues mentioned above. Several binding poses for the ligands were obtained. The binding pose with the lowest docking score was selected for initial MD simulation. The interactions (hydrogen bonds and hydrophobic interactions) of the protein with the ligands were analysed and depicted with LigPlot+[31]. Docking scores of the dye molecules in the active site of laccase crystallographic structure are given in Table 5.A.1.

5.2.3 MD simulation details

The T1 copper and tri-nuclear copper cluster (TNC) in laccase's catalytic site were parameterized using the Metal Centre Parameter Builder (MCPB.py)[32] tool included in the AmberTools20 package[33] to introduce bonded interactions between the copper ions and coordinating amino acid residues in laccase so that the copper coordination remains preserved during MD simulation runs. The atomic partial charges were derived from electrostatic calculations using Gaussian 16 software[34]. All coppers were modelled as Cu^{2+}. The LJ parameters for the copper ions were taken from a previous work[35]. Protonation state assignment to the titratable amino acid residues was performed in two steps. In the first step, we calculated the pK_a values of the titratable amino acid residues using PROPKA-3.4.0[36]. In the next step, we calculated deprotonated fractions of the deeply buried titratable amino acid residues in the protein based on the PROPKA prediction using explicit solvent constant pH molecular dynamics simulation (CpHMD)[37] at pH 7 by using Amber18 package[33]. The protonation states of the residues were then assigned based on all of these results. The protonation states of the copper coordinated histidine residues were determined based on the copper coordinating atom information as provided in the PDB file. For example, the δ-nitrogen in the side chain of H458 is one of the T1 copper coordinated atoms. So, H458 must be protonated at ε-nitrogen position. The final protonation states of the titratable residues are provided in Table 5.A.2. The structure and topology files were then prepared using the tleap interface of AmberTools[38]. The missing residue side chain atoms were added by the tleap program . The systems were inserted into a cubic TIP3P[39] water box with at least 1.2 nm between any atom of the protein and the edge of the box. The solvated systems were neutralized by adding counterions into the box. The number of counterions and water molecules in the simulation boxes are given in Table 5.A.3.

The ff14SB force field[40] was used to model all the amino acid residues. Dye molecules were represented using the General AMBER force field (GAFF2)[41] parameters generated using the Antechamber utility of AmberTools20. As shown in a previous study[42], the pK_a value of the carboxylic acid group present in coumarin 343 is 4.65 and the protonation and deprotonation of the coumarin 343 molecule occur at this carboxyl group. This result indicates that this dye is anionic at neutral pH (pH 7). Hence, the net charge of coumarin 343 is taken as -1. Charge of brilliant blue is -2, thioflavin T is +1, crystal violet is +1 and methyl green is +2. The partial atomic charges on the dyes were derived by first optimizing the structures using Becke's three-parameter exchange function combined with the Lee-Yang-Parr correlational functional (B3LYP)[43–45] and 6-311++G(2d,2p)[46,47] basis set using the Gaussian 16 package and then fitting the electrostatic potential surface using the RESP method[48].

All simulations were performed using GROMACS 2019.6[49]. All systems were first energy minimized using the steepest descent method and then two-step equilibration were performed in NVT and NPT ensembles. The temperature was maintained at 300K using velocity rescale method[50] and the pressure was kept constant at 1 atm using Parinello-Rahman barostat[51]. All simulations were performed using periodic boundary conditions and the long range electrostatic interactions were handled using the particle mesh Ewald (PME)[52] summation method. The cut-off distance for electrostatic and van der Waals interactions was set to 1.0 nm. Bonds containing hydrogen atoms were constrained with LINCS[53] and the integration time step was set to 2 fs. Trajectories were generated in protein apo state for 2 μs, in different dye bound states for 1 μs each and in protein apo state after deleting the five dye molecules from the active site for 1 μs each. Interaction statistics on the ligand bound trajectories were analysed using ProLIF program[54].

5.3 Results and discussions

5.3.1 Evidence of dye degradation by laccase from experiments

In order to confirm that these dye molecules can indeed degraded by laccases, we have carried out systematic UV-visible absorption spectroscopy of these dyes in the presence of laccase and the results are shown in Figure 5.A.1. It is quite evident that the absorption intensity of all the dye molecules studied dropped significantly after laccase treatment but the time elapsed to notice any significant drop in absorption intensity varies from a few hours to more than one day for these dye molecules. It is quite remarkable that laccase can bind and degrade the dye molecules with such large variation in charge (polar interactions) and size (non-polar interactions). UV-visible absorption spectroscopy results not only confirm the ability of laccase to degrade the chosen dye molecules in this study but it also shows that efficacy of dye degradation varies across the different molecules. Subsequent discussion will focus on exploring the molecular mechanism behind this experimentally observed substrate promiscuity using computational methodologies.

5.3.2 Molecular docking and initial MD results for laccase-coumarin 343 binding

At first, we have carried out rigid molecular docking of the five dye molecules near the active site of the crystal structure of laccase. The corresponding docking scores are reported in Table 5.A.1. Figure 5.A.2 shows the docked poses of all dye molecules corresponding to the lowest docking score. Distances between the dye molecules and D206 and H458 in the best docked poses near the active site of the crystal structure of laccase are listed in Table 5.A.4.

At first, we initiated MD simulation for the protein and coumarin 343 (C343) complex starting from the docked structure. After an initial 500ns run, we took three different structures from the trajectory where the dye C343 remains closely bound to the active site. Subsequently, three independent trajectories (500ns each) were generated from each of these structures. Interestingly, we found that the C343 binding poses and the conformation of the loop covering the active site varies significantly across these independent MD runs. Figure 5.A.3a shows the representative active site configurations for C343. Although a part of the C343 molecule remains close to the catalytic H458 residue, the overall binding pose may vary considerably. Moreover, the loop covering the active site (residues 159-164) shows significant variation in the conformation depending on the specific binding pose (Figure 5.A.3b). Moreover, the binding pose does not remain same as the docked configuration, which was based on the crystal structure and did not allow for conformational change in the active site. Moreover, hydrogen bond analysis between the protein and C343 (Figure 5.A.3d) showed that in the two binding poses for C343, two oxygen atoms of the carboxyl group made hydrogen bonds with the N^{ϵ} H -atom of H458 very frequently, but the third pose was not involved in hydrogen bonding with H458 due to a different orientation.

An important take home message from this analysis is that rigid docking based on the crystal structure of the receptor can often be misleading. Specifically for the laccase system the crystal structure corresponds to the active site being partially hidden in a closed pocket. Hence, all the dye molecules exhibit relatively lower docking score (affinity).

5.3.3 Results for all the five laccase-dye complexes

As observed earlier the receptor may exists in at least three distinct conformations (Figure 5.A.3b) depending on the binding pose of C343. Hence, we decided to dock again the other four dye molecules to all these conformations rather than the crystal structure. Distances between the dye molecules and D206 and H458 in the best docked poses near the active site of laccase conformations found in C343 study are listed in Table 5.A.5. Interestingly, this approach leads to significantly lower docking score (Figure 5.A.4) as these partially open loop conformations allow the larger dye molecules to be accommodated. For example, for Brilliant Blue the lowest docking score (-10.807 Kcal/mol in Figure 5.A.4a) was found for active site conformation corresponding to the binding pose-B of C343. This conformation of the active site offers the largest space for ligand binding. We find that both crystal violet and methyl green are giving lowest docking scores in the same protein conformation and their docked poses are also quite similar (Figure 5.A.4). Thioflavin T is giving the lowest docking score in the same protein conformation where C343 binding energy is the lowest and their binding poses are also quite similar. In all the binding poses the dye molecules are bound close to the catalytic residue H458 (Table 5.A.5) justifying the ability of this enzyme to degrade these molecules. While a part of the dye molecule interacts closely with the catalytic H458, the remaining part may adopt different orientations and stabilised by different residues around the flexible loop region, primarily through hydrophobic and non-polar interactions.

MD simulation trajectories starting from these new docked poses (selective to different loop conformations depending on lowest docking score) showed that all the ligands remain stable in the originally docked configuration at the simulation timescale. The binding poses of all the five dye molecules and the corresponding protein active site conformations are shown in Figure 5.3. In all these binding poses dye molecules stay close to the H458 residue in the binding pocket as shown by the distance plot between the centre-of-mass (COM) of the dye heavy atoms and the centre of mass of the H458 residue heavy atoms in Figure 5.3a. Binding poses of Crystal Violet and Methyl Green are identical. The loop (residues 159-164) can exists in significantly different conformations that can accommodate the dye molecules with different shapes and sizes.

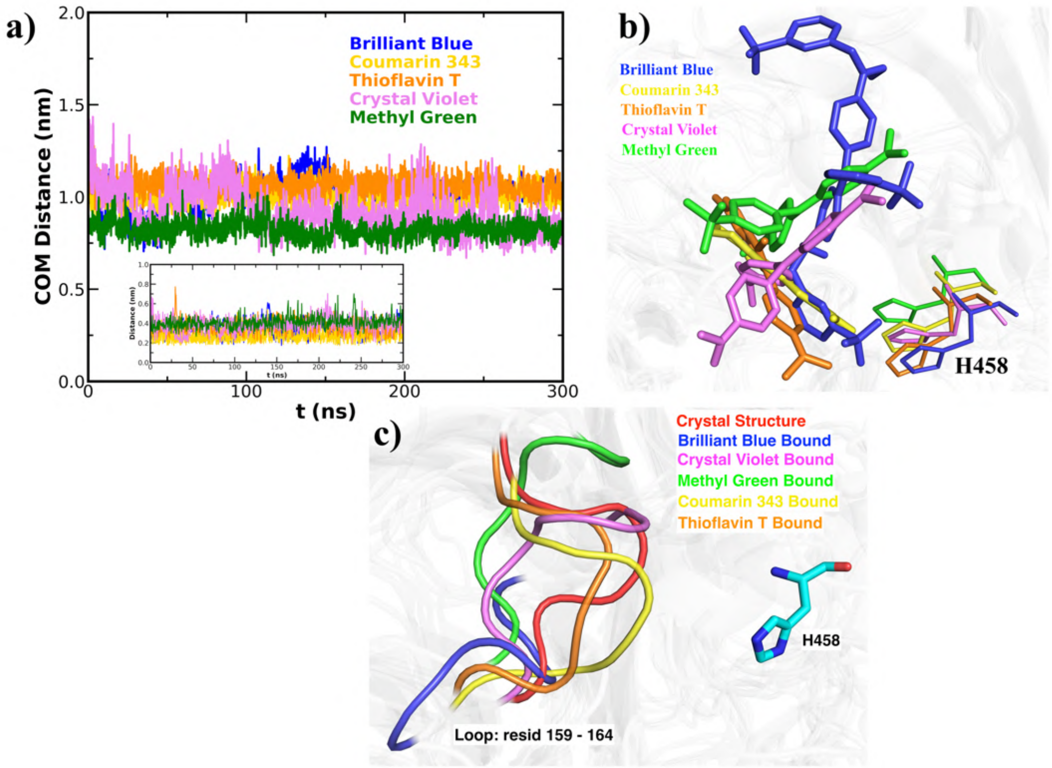


Figure 5.3: a) Distance between the centre-of-mass (COM) of dye molecule heavy atoms and the COM of H458 heavy atoms. Inset shows the minimum distance plots between dye molecules and H458 in stable bound poses. **b)** Superposition of the binding poses of all the five dye molecules. **c)** Various conformations of a loop in the active site of laccase in crystal structure and in dye bound poses.

In Figure 5.4, we show the 2d interaction diagrams[31,55] for the stable bound poses of the dye molecules in the active site of laccase. Brilliant Blue and C343 make hydrogen bonds with the side chain $N^{\epsilon}\ H$ - atom of H458 due to the presence of hydrogen bond acceptor atom in these dye molecules. As there are several hydrophobic amino acid residues present in the active site; L158, G159, P160, A161, P163 and L164, majority of the interactions between the protein and the dye molecules are hydrophobic in nature. In Figure 5.A.5, we show the interaction statistics between the protein and the ligands calculated using the trajectories where the dye molecules are stable in the active site of laccase. There also we can see that hydrophobic, π-stacking and hydrogen bonding interactions are predominant between the ligand and the protein, but the major interactions are hydrophobic in all the five cases.

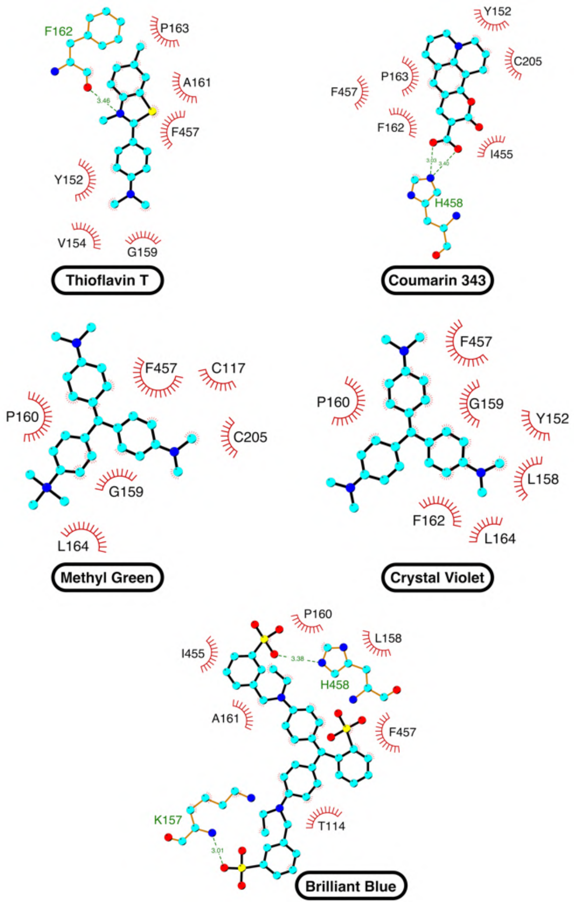


Figure 5.4*:* 2D interaction diagrams between the dye molecules and the protein in the stable binding poses of the dyes. O atoms are in red, N atoms in blue, C atoms in cyan and S atoms in yellow. Hydrogen bonds are represented by green dashed lines. Dye molecules are mostly interacting with hydrophobic residues.

So far we have established that the active site loop of laccase can exists in multiple distinct conformations to accommodate diverse substrate molecules. It would be interesting to explore whether these conformational states may exist in the "apo" state of the enzyme as well (conformational selection)[55], or these conformations can exist only in the presence of the substrate (induced fit)[56]. In order to explore the mechanism of dye binding to the active site of laccase, we exhaustively sampled the apo form of protein (for 2 μs). Moreover, we ran another set of apo trajectories starting at the respective substrate bound conformations, but after deleting the substrate molecules to generate the apo state. We performed Principal Component Analysis (PCA)[57,58] on these trajectories using MDAnalysis python package[59] to characterise the conformational space of the protein. The theory of PCA is discussed in section 2.4.1 of Chapter 2. Results are shown in Figure 5.A.6. We have found that there exist several distinct stable protein conformations where thioflavin T, methyl green and crystal violet can be bound. These conformations may exist as metastable states in the apo form (without ligand) as well. These results suggest that 'conformational selection' might be the dominant mechanism of ligand binding for some of the dye molecules, but more detailed kinetic analysis along with enhanced sampling simulations need to be taken up for a more exhaustive and quantitative understanding of the ligand binding mechanism in such systems with flexible receptor site.

5.3.4 Binding energy calculation

The binding energy calculations using Molecular Mechanics/Generalized Born Surface Area (MM/GBSA)[60] method are shown in Table 5.1. The theory behind MM/GBSA method is discussed in section 2.6.1 of Chapter 2. Interestingly, the overall binding energy for each dye molecule is quite comparable, although the relative contributions of the individual terms vary significantly. The magnitudes of non-polar solvation energy and van der Waals energy are the highest for Brilliant Blue as it is the largest dye among all. Change in electrostatic interaction energy is positive for the negatively charged dye molecules Coumarin 343 and Brilliant Blue, whereas it is negative for the positively charged dye molecules Thioflavin T, Crystal Violet and Methyl Green. This happens due to the presence of negatively charged D150 and D206 residues in the active site creating a negative electrostatic potential surface as shown in Figure 5.5. The electrostatic interactions are unfavorable for binding for negatively charged dye molecules. The change in polar solvation energy opposes that of change in the electrostatic energy component due to the dielectric screening by the solvent.

Table 5.1: Binding energy for all the dye molecules computed using MM/GBSA method. The error bars represent the standard error of mean of the corresponding energy terms.

Energy terms	**Brilliant Blue (Charge -2) (kcal/mol**	**Coumarin 343 (Charge -1) (kcal/mol)**	**Thioflavin T (Charge +1) (kcal/mol)**	**Crystal Violet (Charge +1) (kcal/mol)**	**Methyl Green (Charge +2) (kcal/mol)**
van der Waal Energy	-55.18 ± 0.06	-41.92 ± 0.10	-42.45 ± 0.03	-40.68 ± 0.05	-43.46 ± 0.04
Electrostatic Energy	95.01 ± 0.23	53.96 ± 0.60	-148.87 ± 0.11	-117.33 ± 0.11	-243.57 ± 0.17
Polar Solvation Energy	-65.94 ± 0.21	-36.62 ± 0.61	162.63 ± 0.11	130.55 ± 0.11	257.94 ± 0.17
Electrostatic Energy + Polar Solvation Energy	29.06 ± 0.32	17.34 ± 0.86	13.76 ± 0.16	13.22 ± 0.15	14.37 ± 0.24
Non-Polar Solvation Energy	-7.33 ± 0.01	-4.33 ± 0.01	-4.35 ± 0.01	-4.68 ± 0.01	-4.95 ± 0.01
Total Binding Energy	**-33.48 ± 0.06**	**-28.97 ± 0.11**	**-33.06 ± 0.03**	**-32.12 ± 0.05**	**-34.03 ± 0.04**

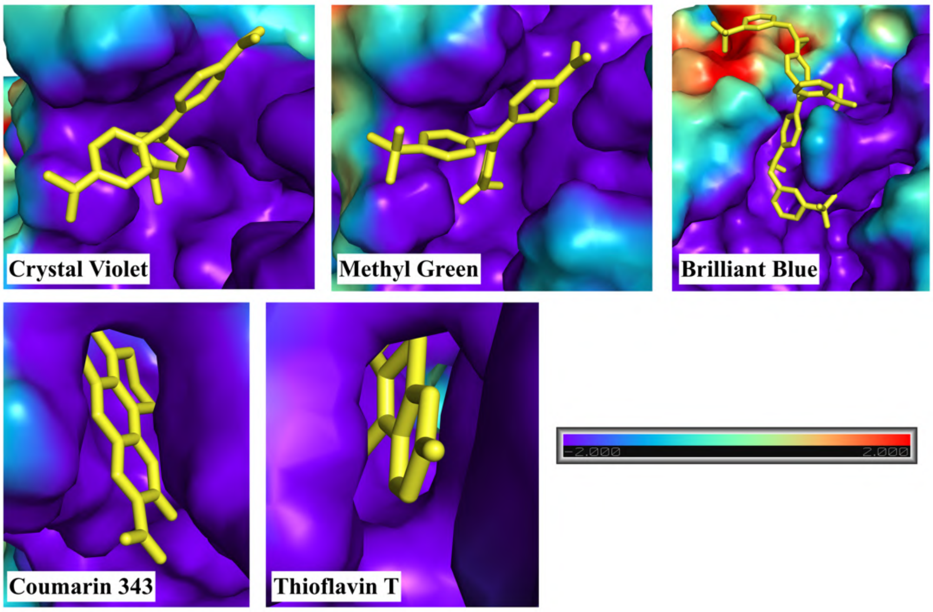


Figure 5.5: Electrostatic potential surface of the protein active site: Due to the presence of D150 and D206 residues in the binding site, binding pocket has a dominant negative electrostatic potential. The color bar representing the electrostatic potential surface is in the units of $\frac{k_B T}{e}$.

Interestingly, although the magnitudes of each type of energy terms differ for each dye molecules, a large cancellation between them leads to almost the same binding energy for all the five dye molecules. The change in the total polar terms including electrostatic and polar solvation energy is positive in all five cases, which disfavors binding across all ligands. However, the non-polar terms, i.e. van der Waals and non-polar solvation energies are favoring the binding. Residue-wise contributions towards the overall binding energies have been shown in the Figure 5.A.9.

While MM/GBSA can provide a reasonable estimate of the relative binding affinity trends between the dye molecules, it misses several factors like explicit solvation, entropic contributions etc. Hence, we have also calculated the absolute binding free energies of two representative dye molecules Coumarin 343 (charge -1) and Crystal Violet (charge +1) using the more robust Thermodynamic integration (TI) approach[61–63]. The theory behind TI is discussed in section 2.6.2 of Chapter 2 and the λ-values used are provided in Table 5.A.6. The results are shown in Table 5.A.7, Figure 5.A.10 and Figure 5.A.11. These binding free energy values also confirm that despite having large difference in charge and shape, Coumarin 343 and Crystal Violet have almost similar binding free energy (about -9 kcal/mol). This is in agreement with our proposal about ligand promiscuity and the qualitative trends observed from the MM/GBSA calculations.

5.4 Conclusions

In conclusion, we have demonstrated that due to the inherent conformational adaptability (plasticity) in the laccase active site rendered by the multiple possible conformational states of the surrounding loop regions, laccase can bind a plethora of dye molecules with largely varying charges and size. There is a clear relationship between the shapes of the dye molecules and the corresponding shape or conformation of the protein active site. The mechanism of binding may vary from a conformation selection model to induced fit model.

The MM/GBSA and TI binding energy calculations have shown that although the contribution of different energy terms to the overall binding energy may differ largely in magnitude, a large cancellation between them results in almost similar binding energy values for the different dye molecules. The interaction between the protein and the dyes is primarily hydrophobic in nature even for the charged dyes due to cancellation between the direct Coulomb interaction and the polar solvation energy component. Thus, our results provide a molecular thermodynamic basis of the substrate promiscuity in laccase with potential implications in designing a broad-spectrum enzymatic approach towards treatment of industrial dye effluents. Moreover, these insights should be relevant to other classes of enzymes demonstrating substrate promiscuity as well.

Appendix 5.A

Table 5.A.1: Docking scores for each dye molecule docked in the active site of laccase in crystallize conformation. All the docking scores are in Kcal/mol.

Dye Molecules	Pose-1	Pose-2	Pose-3	Pose-4
Coumarin 343	-6.846	-6.590	-6.503	-6.370
Thioflavin T	-6.334	-5.729	-5.433	-4.972
Crystal Violet	-5.205	-5.195	-5.054	-5.051
Methyl Green	-5.180	-5.156	-5.105	-5.100
Brilliant Blue	-8.240	-8.199	-8.134	-7.879

Table 5.A.2: Protonation states of the titratable residues used in this study.

Residue Names & ID (Three letter codes)	Protonation states
All TYR	Protonated (charge 0)
All LYS	Protonated (charge +1)
All ARG	Protonated (charge +1)
All ASP	Deprotonated (charge -1)
All GLU	Deprotonated (charge -1)
HIS (ID:: 153, 216, 402)	Protonated at both **δ**-N & **ε**-N positions (HIP). (charge +1)
HIS (ID:: 64, 109, 111, 398, 400, 452, 454)	Protonated at **δ**-N positions only (HID). (charge 0)
HIS (ID::55, 66, 91, 306, 395, 458)	Protonated at **ε**-N positions only (HIE). (charge 0)

Table 5.A.3: The number of counterions and water molecules in the simulation boxes.

Protein State	Number of water molecules	Number of Na^+ ions
Protein apo state	17230	14
Brilliant Blue (charge -2) bound state	17687	16
Crystal violet (charge +1) bound state	18226	13
Methyl Green (charge +2) bound state	18648	12
Thioflavin T (charge +1) bound state	18493	13
Coumarin 343 (charge -1) bound state	17228	15
Protein apo state after deleting brilliant blue dye from active site	20141	14
Protein apo state after deleting crystal violet dye from active site	17907	14
Protein apo state after deleting methyl green dye from active site	18124	14
Protein apo state after deleting thioflavin T dye from active site	18131	14
Protein apo state after deleting coumarin 343 dye from active site	17944	14

Table 5.A.4: Distance between the centre of mass (COM) of the dye molecules in the docked poses and the centre of mass of D206 & H458 individually and minimum distance between dye heavy atoms and D206 & H458 heavy atoms in the crystallographic protein conformation.

Dye Molecules	COM Distance (Å)		Minimum Distance (Å)	
	From COM of D206	From COM of H458	Min distance From D206	Min distance From H458
Brilliant Blue	9.73	10.17	3.42	2.94
Crystal Violet	10.55	9.26	3.24	3.81
Methyl Green	10.59	9.31	3.19	3.85
Thioflavin T	8.06	9.17	4.08	3.50
Coumarin-343	9.82	8.36	3.05	2.83

Table 5.A.5: Distance between the centre of mass (COM) of the dye molecules in the docked poses and the centre of mass of D206 & H458 individually and minimum distance between dye heavy atoms and D206 & H458 heavy atoms in the protein conformations obtained from Coumarin-343 study.

Dye Molecules	**COM Distance (Å)**		**Minimum Distance (Å)**	
	From COM of D206	**From COM of H458**	**Min distance From D206**	**Min distance From H458**
Brilliant Blue	8.45	7.09	3.30	2.83
Crystal Violet	5.29	8.07	3.46	3.86
Methyl Green	5.25	7.86	3.41	3.57
Thioflavin T	6.08	9.86	4.25	3.56

Table 5.A.6: λ values used to decouple coulomb and van der Waals interactions between the dye molecules and the surrounding environments, apply restraints to the dye molecules.

Coumarin 343		**Crystal Violet**	
Transformation	**λ Values**	**Transformation**	**λ Values**

Dye decoupling in water without the protein	Coul= 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.
	VdW=0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.83, 0.85, 0.90, 0.95, 1.00
		Dye decoupling in water without the protein	Coul= 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.
			VdW=0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00

Dye decoupling while bound to the protein	Coul= 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.
	VdW=0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.83, 0.85, 0.90, 0.95, 1.00
		Dye decoupling while bound to the protein	Coul= 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.
			VdW=0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.83, 0.85, 0.90, 0.95, 1.00

Apply restraints to the dye while bound to the protein	0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, 1.0	Apply restraints to the dye while bound to the protein	0.0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.5, 0.75, 1.0

Table 5.A.7: Binding free energies of coumarin 343 and crystal violet using TI.

Coumarin 343 (Charge -1)		**Crystal Violet (Charge +1)**	
Components	**Value (Kcal/mol)**	**Components**	**Value (Kcal/mol)**

$\Delta G^{\mathbf{prot}}_{\mathbf{elec+vdw+rest}}$
(Decoupling of coumarin 343 from protein in the presence of restraints)
	102.837 ± 0.197
	(elec= 84.323 ± 0.059,
	vdw = 17.561 ± 0.206,
	rest = 0.953 ± 0.013)
		$\Delta G^{\mathbf{prot}}_{\mathbf{elec+vdw+rest}}$
		(Decoupling of crystal violet from protein in the presence of restraints)
			31.462 ± 0.267
			(elec= 21.091 ± 0.062,
			vdw = 8.705 ± 0.277,
			rest = 1.666 ± 0.037)

$\Delta G^{\mathbf{solv}}_{\mathbf{elec+vdw}}$ (Decoupling of coumarin 343 from water without protein)	87.179 ± 0.073 (elec= 85.470 ± 0.037, vdw = 1.720 ± 0.063)	$\Delta G^{\mathbf{solv}}_{\mathbf{elec+vdw}}$ (Decoupling of crystal violet from water without protein)	16.882 ± 0.084 (elec= 17.972 ± 0.017, vdw = -1.091 ± 0.082)
$\Delta G^{\mathbf{solv}}_{\mathbf{rest\_on}}$ (Add restraints on decoupled coumarin 343 in water)	6.367	$\Delta G^{\mathbf{solv}}_{\mathbf{rest\_on}}$ (Add restraints on decoupled crystal violet in water)	6.081
$\Delta G_{\mathbf{bind}}$	**-9.291 ± 0.210 (elec= 1.147 ± 0.069, vdw = -15.841 ± 0.215)**	$\Delta G_{\mathbf{bind}}$	**-8.499 ± 0.280 (elec= -3.119 ± 0.064, vdw = -9.796 ± 0.289)**

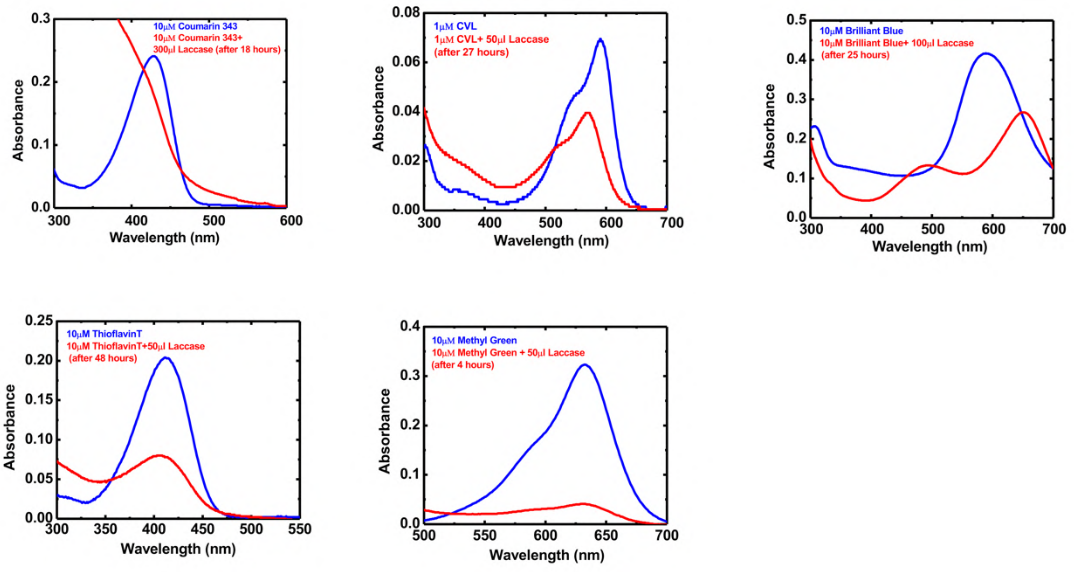


Figure 5.A.1: UV-Visible absorption spectra of the solutions of pure dye molecules and with the addition of laccase.

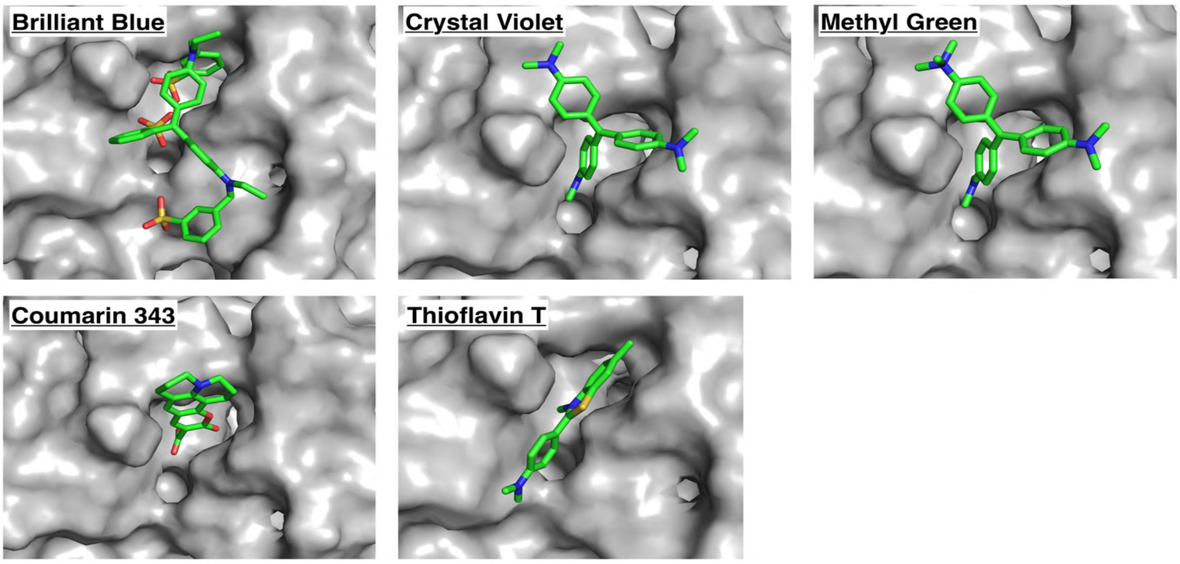


Figure 5.A.2: Docked poses of the dye molecules in the active site of the crystallize protein conformation corresponding to the lowest docking scores (pose-1 from Table 5.A.1).

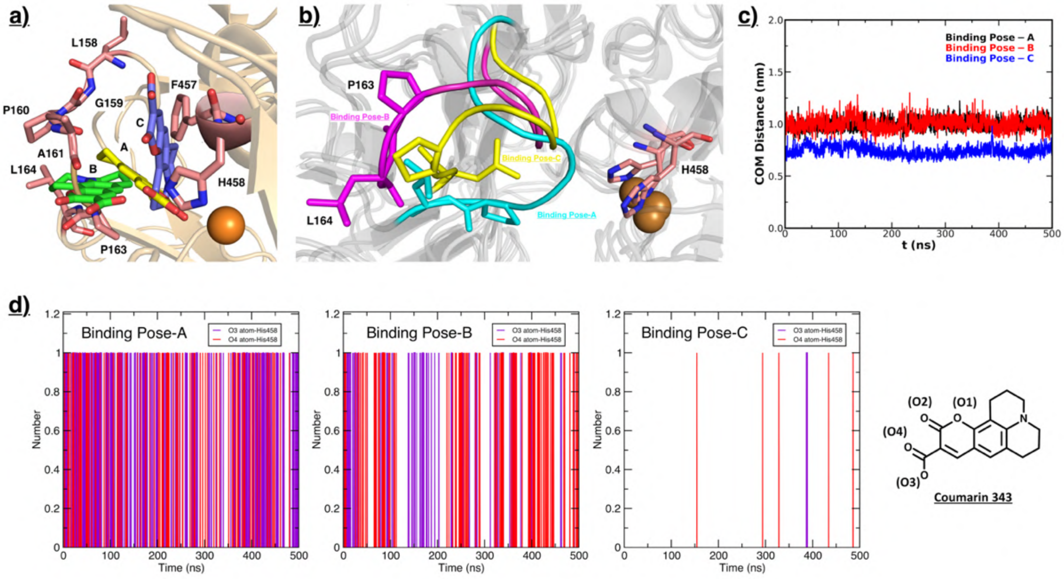


Figure 5.A.3: a) Three different stable binding poses (A, B, C) of coumarin-343 in the substrate active site of laccase. **b)** Three different conformations of a loop in the active site corresponding to the three different binding poses of coumarin-343. **c)** Distance plot between centre-of-mass (COM) of coumarin 343 heavy atoms and COM of H458 heavy atoms in the three stable ligand bound trajectories. **d)** Hydrogen bond analysis in these three stable bound poses of coumarin 343. Hydrogen bonds are calculated between the carboxyl oxygen atoms of coumarin 343 (hydrogen bond acceptor atoms) and amide nitrogen atom (hydrogen bond donor) in the side chain of H458. Cut-off distance is taken as 0.35 nm.

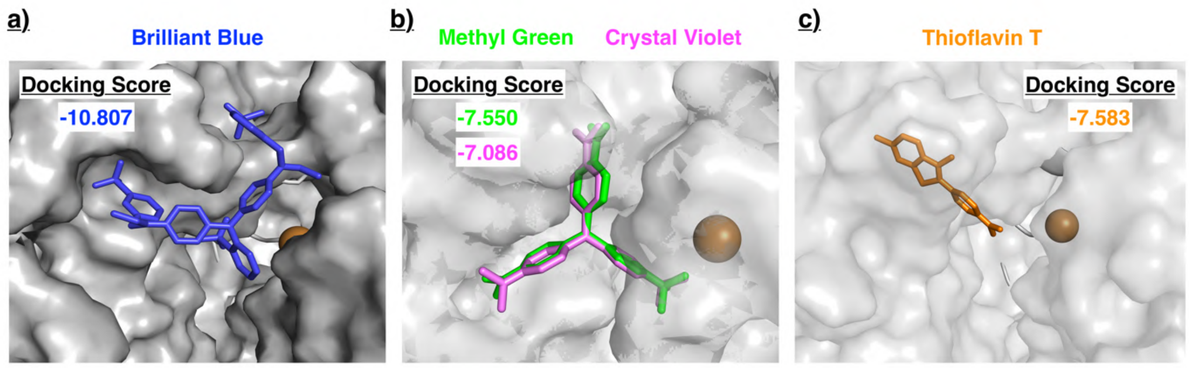


Figure 5.A.4: a) For brilliant blue, lowest docking score (-10.807 Kcal/mol) pose was found in the protein active site conformation corresponding the binding pose-B of coumarin 343. This conformation of the active site offers largest space for ligand binding and since brilliant blue is also the largest dye, this is acceptable. **b)** Methyl green and crystal violet docked poses are identical. For both of them lowest docking score poses were found in the protein active site conformation corresponding to the binding pose-C of coumarin 343. As their shapes are identical, this is also acceptable. **c)** For thioflavin T, lowest docking score (-7.583 Kcal/mol) pose was found in the protein active site conformation corresponding to the binding pose-A of coumarin 343. All the docking scores in the picture are in Kcal/mol.

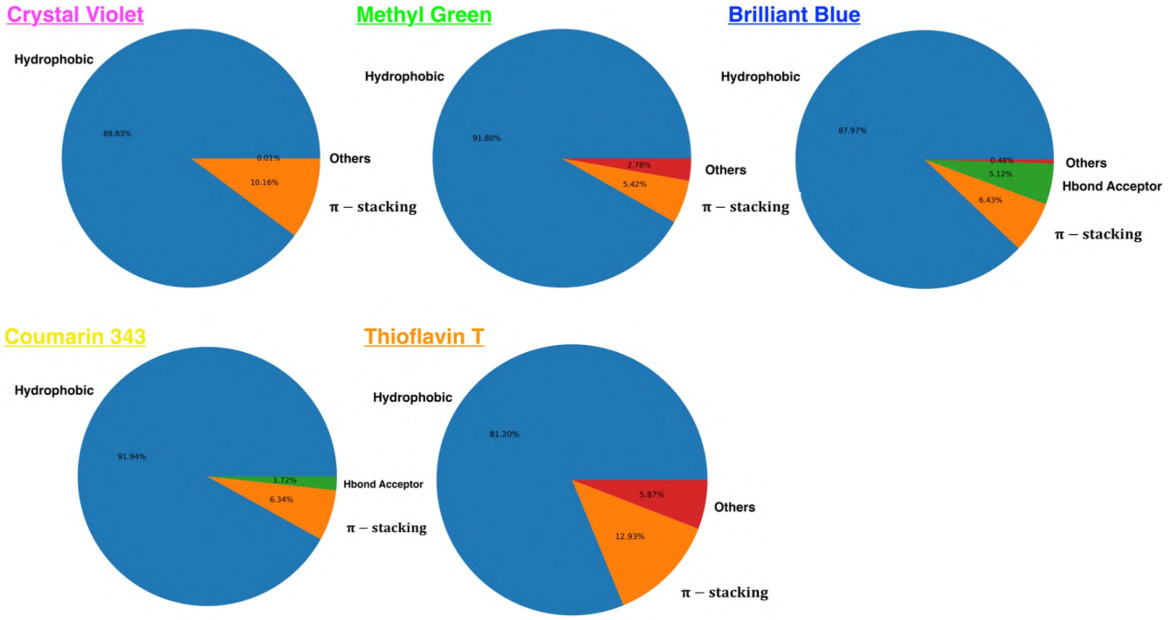


Figure 5.A.5: Protein-ligand interaction statistics calculated using the stable ligand bound trajectories of the protein-ligand complex. Major interactions are hydrophobic, π-stacking in all the five cases.

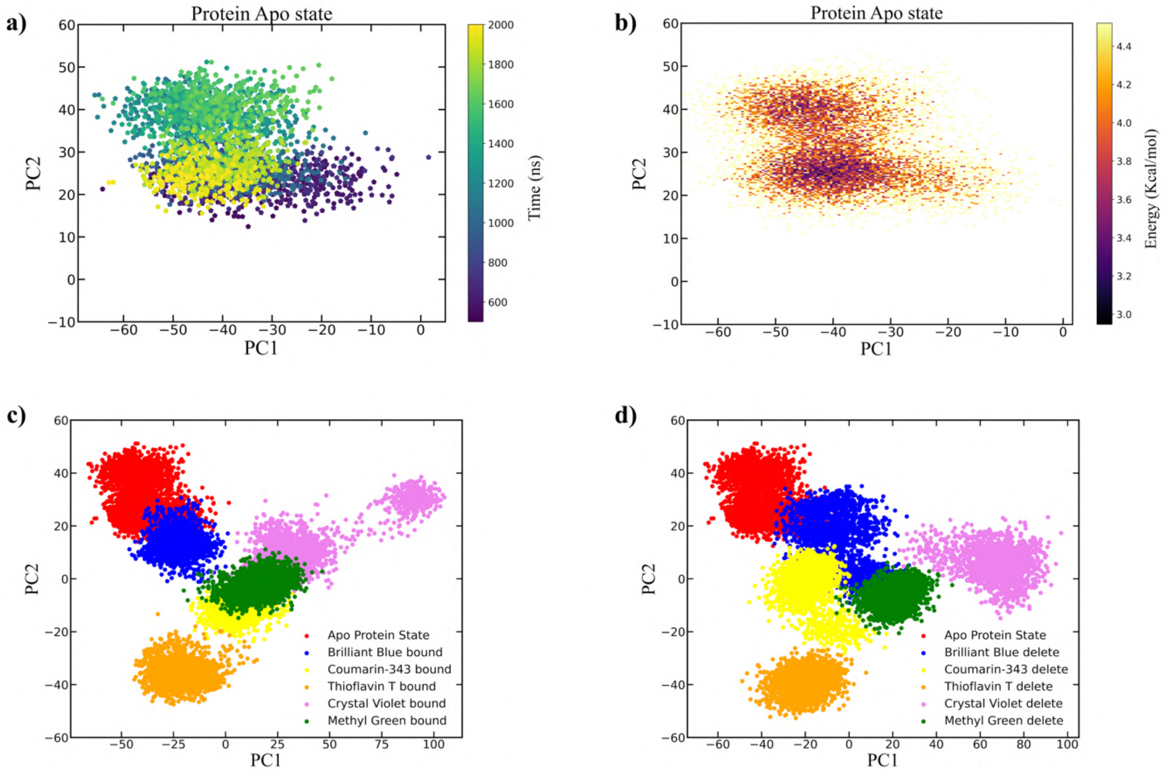


Figure 5.A.6: **a)** Protein apo state conformational space sampled for 2 µs during the simulation projected on PC1-PC2 plane. **b)** Free energy surface for protein apo state (for 2 µs). **c)** Protein apo state and dye bound state (1 µs each) conformational spaces are projected on PC1-PC2 plane. **d)** Protein apo state and protein apo state (after deleting dye molecules from the active site in their bound poses 1 µs each) conformational spaces are projected on PC1-PC2 plane.

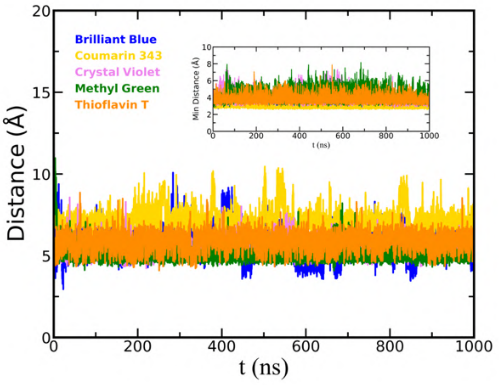


Figure 5.A.7: Distance between the COM of dye molecules and the COM of the active site residue heavy atoms for 1 µs long ligand bound trajectories. Inset shows the minimum distance plot between the dye heavy atoms and a group of D206 & H458 heavy atoms.

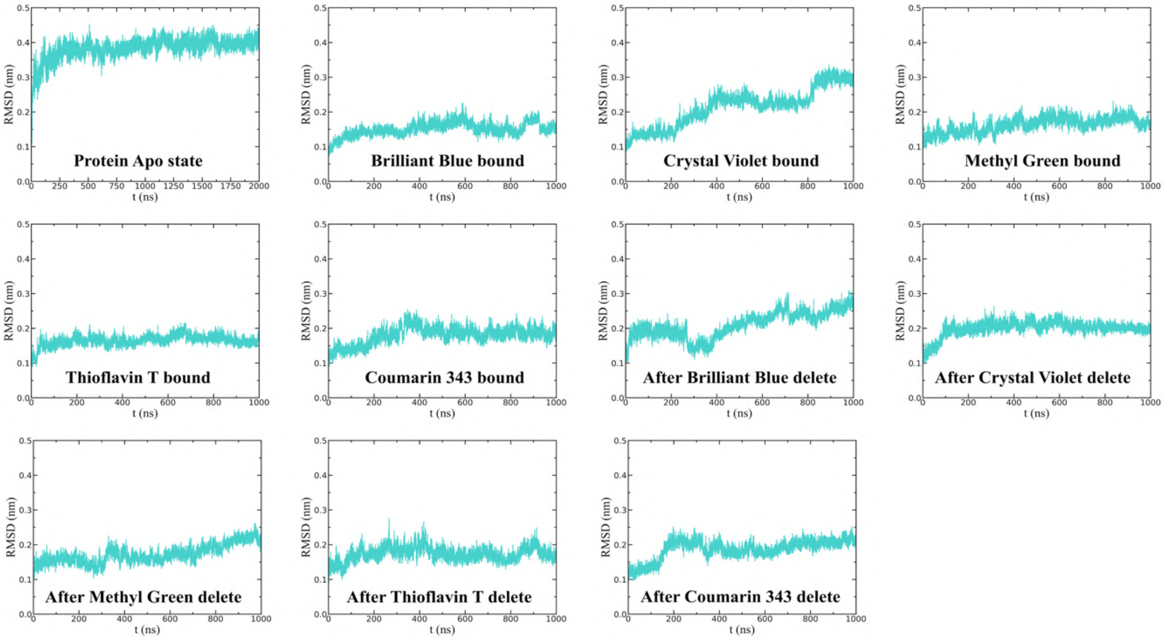


Figure 5.A.8: Root mean squared displacement (RMSD) of protein backbone atoms from the various MD simulations.

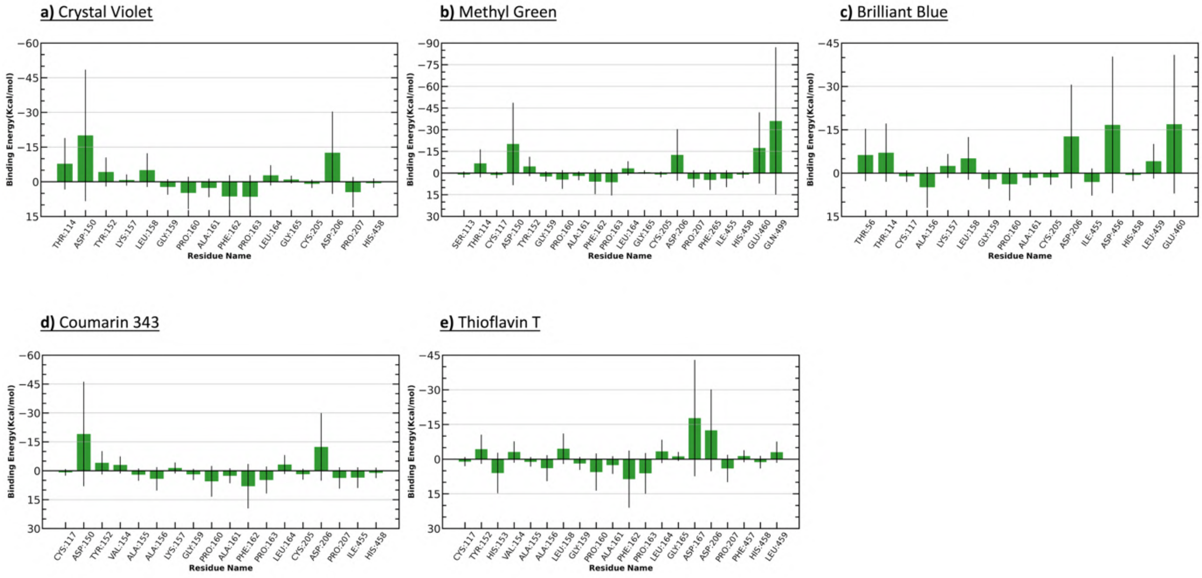


Figure 5.A.9: Contribution of the residues in the binding pocket of the laccase to the MM-GB/SA binding energy of each dye molecule in their binding poses. Negatively charged Asp-150, Asp-206, Asp-456 have significant contributions to the binding energy. This shows that strength of the electrostatic interactions between the charged dyes and these charged amino acid residues overwhelmed the hydrophobic interactions, although statistics have shown that most of the interactions between the ligands and the protein are hydrophobic.

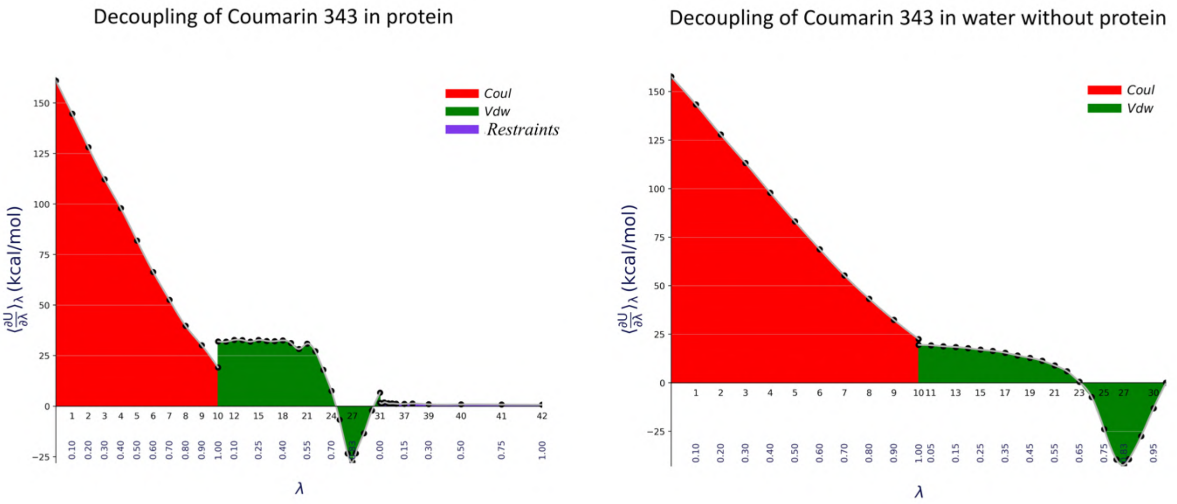


Figure 5.A.10: $\langle\frac{\mathrm{dU}}{\mathrm{d}\lambda}\rangle_\lambda$ vs λ plots for coumarin 343 decoupling from protein and coumarin 343 decoupling in water without protein. Free energy difference is the area under the curve. Cubic interpolation line is represented by silver colour. Different energy components are shown in colour: red for electrostatic, green for van der Waals and violet for attaching restraints to the ligand.

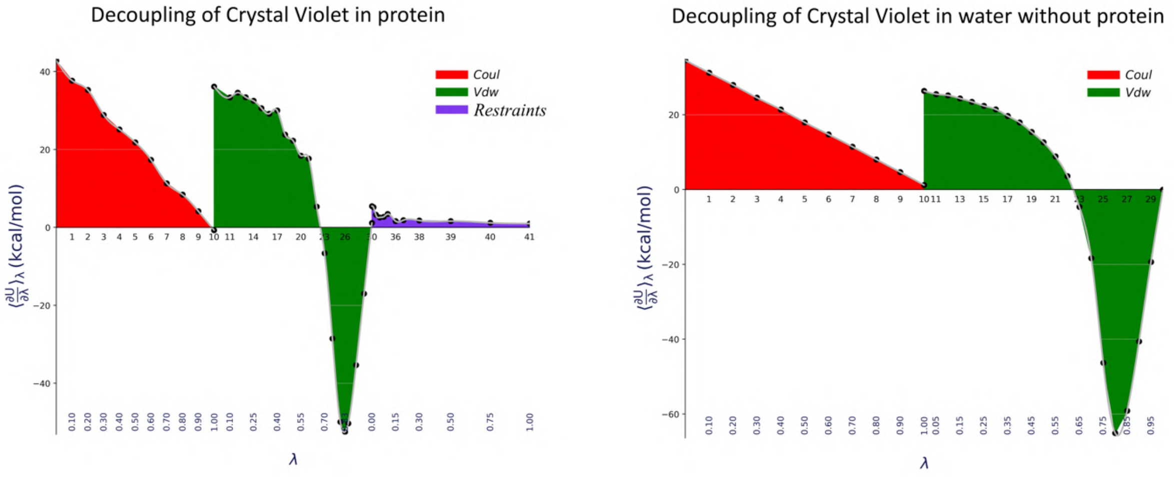


Figure 5.A.11: $\langle\frac{\mathrm{dU}}{\mathrm{d}\lambda}\rangle_\lambda$ vs λ plots for crystal violet decoupling from protein and crystal violet decoupling in water without protein. Free energy difference is the area under the curve. Cubic interpolation line is represented by silver colour. Different energy components are shown in colour: red for electrostatic, green for van der Waals and violet for attaching restraints to the ligand.

References:

1 L. Lin, H. Yang and X. Xu, Front Environ Sci, 2022, 10, 975.

2 H. Ben Slama, A. Chenari Bouket, Z. Pourhassan, F. N. Alenezi, A. Silini, H. Cherif-Silini, T. Oszako, L. Luptakova, P. Golińska and L. Belbahri, Applied Sciences, 2021, 11, 6255.

3 M. Wainwright, Dyes and Pigments, 2008, 76, 582–589.

4 S. Adeel, S. Abrar, S. Kiran, T. Farooq, T. Gulzar and M. Jamal, Handbook of Renewable Materials for Coloration and Finishing, 2018, 189–211.

5 M. C. Kannaujiya, R. Kumar, T. Mandal and M. K. Mondal, Process Safety and Environmental Protection, 2021, 155, 444–454.

6 R. Kant and R. Kant, Nat Sci (Irvine), 2011, 4, 22–26.

7 T. Shindhal, P. Rakholiya, S. Varjani, A. Pandey, H. H. Ngo, W. Guo, H. Y. Ng and M. J. Taherzadeh, Bioengineered, 2021, 12, 70–87.

8 C. R. Holkar, A. J. Jadhav, D. V. Pinjari, N. M. Mahamuni and A. B. Pandit, J Environ Manage, 2016, 182, 351–366.

9 H. Claus, G. Faber and H. König, Appl Microbiol Biotechnol, 2002, 59, 672–678.

10 E. I. Solomon, U. M. Sundaram and T. E. Machonkin, Chem Rev, 1996, 96, 2563–2605.

11 J. Su, J. Fu, Q. Wang, C. Silva and A. Cavaco-Paulo, Taylor & Francis, 2018, preprint.

12 L. Arregui, M. Ayala, X. Gómez-Gil, G. Gutiérrez-Soto, C. E. Hernández-Luna, M. de Los Santos, L. Levin, A. Rojo-Dominguez, D. Romero-Martinez, M. C. N. N. Saparrat, others, M. Herrera De Los Santos, L. Levin, A. Rojo-Domínguez, D. Romero-Martínez, M. C. N. N. Saparrat, M. A. Trujillo-Roldán, N. A. Valdez-Cruz, M. de Los Santos, L. Levin, A. Rojo-Dominguez, D. Romero-Martinez, M. C. N. N. Saparrat and others, Laccases: structure, function, and potential application in water bioremediation, 2019, vol. 18.

13 S. Witayakran and A. J. Ragauskas, John Wiley & Sons, Ltd, 2009, preprint.

14 S. F. F. Zofair, S. Ahmad, M. A. Hashmi, S. H. Khan, M. A. Khan and H. Younus, J Environ Manage, 2022, 309, 114676.

15 V. Pande, T. Joshi, S. C. Pandey, D. Sati, S. Mathpal, V. Pande, S. Chandra and M. Samant, J Biomol Struct Dyn.

16 J. Yang, W. Li, T. Bun Ng, X. Deng, J. Lin and X. Ye, Front Microbiol, 2017, 8, 832.

17 K. Hult and P. Berglund, Trends Biotechnol, 2007, 25, 231–238.

18 P. Singla and R. D. Bhardwaj, Enzyme promiscuity–A light on the “darker” side of enzyme specificity, 2020, vol. 38.

19 A. Babtie, N. Tokuriki and F. Hollfelder, Curr Opin Chem Biol, 2010, 14, 200–207.

20 D. Thakur and S. B. Pandit, J Struct Biol, 2022, 214, 107835.

21 T. Bertrand, C. Jolivalt, P. Briozzo, E. Caminade, N. Joly, C. Madzak and C. Mougin, Biochemistry, 2002, 41, 7325–7333.

22 R. Mehra, A. S. Meyer and K. P. Kepp, RSC Adv, 2018, 8, 36915–36926.

23 J. L. Cole, P. A. Clark and E. I. Solomon, J Am Chem Soc, 1990, 112, 9534–9548.

24 S. M. Jones and E. I. Solomon, Electron transfer and reaction mechanism of laccases, 2015, vol. 72.

25 I. Bento, C. S. Silva, Z. Chen, L. O. L. O. Martins, P. F. Lindley and C. M. Soares, Mechanisms underlying dioxygen reduction in laccases. Structural and modelling studies focusing on proton transfer, BioMed Central, 2010, vol. 10.

26 S. Riva, Trends Biotechnol, 2006, 24, 219–226.

27 R. Mehra, J. Muschiol, A. S. Meyer and K. P. Kepp, Sci Rep, 2018, 8, 1–16.

28 K. Junker, R. Kissner, B. Rakvin, Z. Guo, M. Willeke, S. Busato, T. Weber and P. Walde, Enzyme Microb Technol, 2014, 55, 72–84.

29 S. Kurniawati and J. A. Nicell, Bioresour Technol, 2008, 99, 7825–7834.

30 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, J Comput Chem, 2009, 30, 2785–2791.

31 R. A. Laskowski and M. B. Swindells, J Chem Inf Model, 2011, 51, 2778–2786.

32 P. Li, K. M. Merz Jr and K. M. Merz, MCPB. py: a python based metal center parameter builder, ACS Publications, 2016, vol. 56.

33 D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, J Comput Chem, 2005, 26, 1668–1688.

34 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji and others, Gaussian, Inc. Wallingford, CT, 2016, preprint.

35 P. Li, B. P. Roberts, D. K. Chakravorty and K. M. Merz Jr, J Chem Theory Comput, 2013, 9, 2733–2748.

36 M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, J Chem Theory Comput, 2011, 7, 525–537.

37 J. M. Swails, D. M. York and A. E. Roitberg, J Chem Theory Comput, 2014, 10, 1341–1352.

38 D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, Comput Phys Commun, 1995, 91, 1–41.

39 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, J Chem Phys, 1983, 79, 926–935.

40 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, J Chem Theory Comput, 2015, 11, 3696–3713.

41 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, J Comput Chem, 2004, 25, 1157–1174.

42 R. E. Riter, E. P. Undiks and N. E. Levinger, J Am Chem Soc, 1998, 120, 6062–6067.

43 A. D. Becke, J Chem Phys, 1993, 98, 5648–5652.

44 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, J Phys Chem, 1994, 98, 11623–11627.

45 C. Lee, W. Yang and R. G. Parr, Phys Rev B, 1988, 37, 785.

46 A. D. McLean and G. S. Chandler, J Chem Phys, 1980, 72, 5639–5648.

47 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, J Chem Phys, 1980, 72, 650–654.

48 C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, J Phys Chem, 1993, 97, 10269–10280.

49 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindah, SoftwareX, 2015, 1–2, 19–25.

50 G. Bussi, D. Donadio and M. Parrinello, J Chem Phys, 2007, 126, 14101.

51 M. Parrinello and A. Rahman, J Appl Phys, 1981, 52, 7182–7190.

52 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, J Chem Phys, 1995, 103, 8577–8593.

53 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, J Comput Chem, 1997, 18, 1463–1472.

54 C. Bouysset and S. Fiorucci, J Cheminform, 2021, 13, 1–9.

55 T. R. Weikl and F. Paul, John Wiley & Sons, Ltd, 2014, preprint.

56 H. X. Zhou, Biophys J, 2010, 98, L15–L17.

57 C. C. David and D. J. Jacobs, Methods in Molecular Biology, 2014, 1084, 193–226.

58 M. A. Balsera, W. Wriggers, Y. Oono and K. Schulten, Journal of Physical Chemistry, 1996, 100, 2567–2572.

59 N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, J Comput Chem, 2011, 32, 2319–2327.

60 T. Hou, J. Wang, Y. Li and W. Wang, J Chem Inf Model, 2011, 51, 69–82.

61 S. Bruckner and S. Boresch, J Comput Chem, 2011, 32, 1320–1333.

62 M. Lawrenz, R. Baron, Y. Wang and J. A. McCammon, Methods in Molecular Biology, 2012, 819, 469–486.

63 P. V. Klimovich, M. R. Shirts and D. L. Mobley, J Comput Aided Mol Des, 2015, 29, 397–411.

6

Unveiling dye effluent binding mechanism and allostery in laccase using molecular dynamics combined with machine learning approaches

6.1 Introduction

In Chapter 5, We have discussed the molecular origin of substrate promiscuity in laccase using UV-visible absorption spectroscopy and classical MD simulations. We discovered that a loop (loop-1; resids:G159-P160-A161-F162-P163-L164) in the laccase active site can have various distinct conformations while remain bound to the six ligands; coumarin 343, thioflavin T, methyl green, crystal violet, brilliant blue and 2,5-xylidine. Furthermore, binding energy calculations revealed surprisingly similar binding affinity of the ligands irrespective of their different shape and charge.

The mechanism of ligand binding to biomolecules generally proceeds via two competing pathways[1,2]: (i) "conformational selection mechanism", where ligand binds to a pre-existing binding competent conformation of the protein (could be a less populated or high energy one) and stabilizes this preferred conformation; (ii) "induced fit mechanism", where ligand binds to the less binding competent but predominant protein conformation followed by a protein conformational change to stabilize the protein-ligand complex. For an ideal induced fit mechanism, the ligand induced conformation should not be present in apo protein. As mentioned earlier, we have found various distinct ligand bound conformations of laccase. Therefore, a question naturally arises whether the dye molecule binding to the laccase proceeded via conformational selection or induced fit mechanism. Precise identification of the mechanism requires extensive calculation of equilibrium rate constants or flux through each pathway[3,4]. However, one possible way to qualitatively answer this question from MD simulation data is to find as many as possible metastable apo conformations of the laccase and compare the bound conformations with them. A recent study reveals the binding pathway of a pollutant to the laccase[5].

In the last two decades, Markov State Model (MSM)[6,7], build from MD simulations, becomes the most popular kinetic model to describe protein conformational changes by capturing the dynamics through transitions among discrete metastable conformational states. In spite of great advances in making algorithms for building reliable MSMs, finding good feature dataset and a low dimensional projection space that best describes the conformational landscape of a biomolecule is still not a trivial task[8].

For the last couple of years, machine learning and artificial intelligence (AI/ML) revolutionize the field of computational chemistry[9]. Particularly, ML-based methods for analysing and enhancing MD simulations unlock new ways of seeing the unseen[10,11]. With this motivation, we have also used numerous ML techniques to complement MD simulations. Useful features can be identified using the Random Forests (RF) classification algorithm[12]. RF classifier, a supervised ML-based method, is one of the most popular feature selection methods that can identify a small subset of important features from a large set of input variables. RF classifier ranks the input variables using some feature importance metric. This algorithm assigns lower importance score to the large amount of input variables that are unnecessary to describe the processes of interest. This way, this method filters out the insignificant input variables. A recent study showcased the ability of the RF classifier to identify key amino acid residue pairs that can distinguish between the apo and ligand bound protein conformations[13].

Time-lagged independent component analysis (TICA)[14,15] is a widely used linear dimensionality reduction method, which can efficiently learn the slow processes and generates projections of the high dimensional data along some slow degrees of freedom. Over the years, TICA has played an important role in the quantitative description of various biomolecular processes[16–18]. However, often, when dealing with high dimensional data, it is not possible to differentiate well the states of a system using a linear dimensionality reduction method[19]. In the recent years, non-linear dimensionality reduction methods, such as UMAP[20], t-SNE[21], variational autoencoder[22] and time-lagged autoencoder[23] have gain popularities in building low dimensional spaces that discriminates well the states of a system. Among these, variational autoencoder (VAE), a neural network based deep learning model, has shown its potential to generate low dimensional latent space where different states of complex chemical and biomolecular systems are well discriminable[24–27]. The performance of linear and non-linear dimensionality reduction methods depend on the choice of input feature dataset and complexity of the system.

In this present chapter, we have discussed the possible dye binding mechanism to laccase using classical MD simulations coupled with several aforementioned ML approaches. We first identified useful residue pairs of the protein as features using the RF classifier. Subsequently, we have compared the two different classes of dimensionality reduction methods: (i) linear method TICA and (ii) non-linear method VAE to obtain the best low dimensional representation for the laccase apo conformational landscape. We then performed kinetic clustering using VAMPnets[28] to identify metastable apo conformations of the laccase. We have also calculated the transition rates between the metastable conformational states of laccase using MSM. Finally, we have provided an allosteric connection between loop-1 and another loop that is far away from loop-1.

6.2 Computational details

6.2.1 Equilibrium MD simulation protocols

Force field information and other modelling details of the protein are provided in the previous Chapter 5. For this present study, we performed 100 classical MD simulations of 500 ns each starting from 100 different conformations of laccase obtained from metadynamics simulations. Metadynamics parameter details for this present work is discussed in section 6.2.4. Thus, a total of 50 µs trajectories were generated. All the systems were solvated with ~ 19000 TIP3P[29] waters and neutralized by adding 14 Na^+ ions into the simulation box. All the simulations were performed using GROMACS 2024.2 software[30]. All systems were first energy minimized using the steepest descent method followed by equilibrations in NVT and NPT ensembles. The temperature was maintained at 300K using velocity rescale method[31] and the pressure was kept constant at 1 atm using Parinello-Rahman barostat[32]. All simulations were performed under periodic boundary conditions and the long range electrostatic interactions were handled using the particle mesh Ewald (PME) summation method[33]. Both the cut-off distances for electrostatic and van der Waals interactions were set to 1.0 nm. Bonds containing hydrogen atoms were constrained with LINCS[34] and the integration time step was set to 2 fs. Frames were saved every 100 ps and the first 100 ns from each trajectory was discarded from further analyses.

6.2.2 Random forests classifier details

We took the five 1 µs long dye molecule bound trajectories and last 500 ns of a 1 µs long apo trajectory generated after deleting the 2,5-xylidine ligand from the crystallized active structure of laccase for RF analysis. Frames of each trajectory were labelled to distinguish between the six different conformational states. At first a frame was extracted from the apo trajectory and considered as a reference structure. Then, from this reference structure, we selected all the pairs of heavy atoms which were more than three residues apart, calculated the distances between these pairs and identified ~ 1500 residue pairs such that the heavy atom minimum distance for each pair was less than or equal to 4.5 Å using MDTraj python package[35]. $C_\alpha - C_\alpha$ atom distances of the residue pairs for all the six trajectories were used as input dataset for RF classifier. 70% of the input data was used for training and 30% for validation. In this work, we employed the default RF method implemented in the scikit-learn python package[36] to identify a small number of features from this large set of distance pairs. The default method uses the mean decrease in impurity of gini impurity[12] as the feature importance score. The RF classifier consisted of 10 estimators. Accuracy score for the validation dataset was 0.998. Cross-validation was performed for 200 different input datasets generated randomly.

6.2.3 HDBSCAN clustering

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identifies clusters of varying densities without specifying the number of clusters beforehand. We converted the six 500 ns long apo trajectories (last 500 ns of 1 μs long trajectories generated after deleting the six ligands from the laccase active site, referred as "ligand delete apo trajectories") to feature dataset using the selected $C_\alpha - C_\alpha$ atom distance pairs obtained from the RF analysis using PyEMMA 2.5.7[37] and subsequently performed HDBSCAN clustering on this high dimensional feature space using the scikit-learn python package[36]. One of the main parameters of HDBSCAN clustering method is, *min_samples*, defines the minimum number of points in a cluster. We used *min_samples* = 1000 and 2000 to test convergence of the results.

6.2.4 TICA-Metadynamics simulation

We performed time-lagged independent component analysis (TICA)[14,15] on the six 500 ns long "ligand delete apo trajectories" used in HDBSCAN clustering with the aforementioned set of selected distance pairs as features using PyEMMA 2.5.7[37] with 20 ns lag-time. Theory of TICA is presented in section 2.4.2 of Chapter 2. The first four TICA components were used as collective variables (CVs) to perform four TICA-Metadynamics (TICA-MetaD)[38,39] enhanced sampling simulations of 100 ns each starting from different structures. Theory of metadynamics is discussed in section 2.7.1 of Chapter 2. These TICA-MetaD simulations were performed using GROMACS 2021.5[30] patched with PLUMED 2.7.4[40]. In our TICA-MetaD simulations, the Gaussian bias was deposited every 500 steps with a height of 1.2 kJ/mol and width of 0.5 nm. The height of the Gaussian hill was tempered with a bias factor of 8. All the metadynamics simulation trajectories were projected on the same four TICA eigenvectors (see Figure 6.A.1a-c) and 100 different structures spanning the whole 2D TICA space (see Figure 6.A.1d-f) were selected to generate 100 short equilibrium trajectories as discussed in section 6.2.1. We finally performed another TICA on the combined 43 μs discrete equilibrium apo trajectories (100 × 400 ns generated in the present study and 6 × 500 ns used in HDBSCAN clustering) using 20 ns of lag-time.

6.2.5 VAE latent space construction

Theory of VAE is discussed in section 2.4.3 of Chapter 2. We trained the VAE on the total of 43 μs discrete equilibrium apo trajectories used for TICA using PyTorch for constructing a 2-dimensional latent space to visualize the apo conformational landscape of laccase. We employed the same set of distance pairs used in HDBSCAN clustering and TICA to construct a large dataset from the trajectories for VAE training and validation. Subsequently, this dataset was normalized using the StandardScaler function of the scikit-learn python package[36]. The encoder block of VAE comprised of four fully connected hidden layers with nodes 300-200-100-20, leading to a latent space layer with 2 nodes and the decoder part mirrors the encoder. We used *tanh* activation function in the first hidden layer followed by the ReLU activation function for all the other hidden layers. 70% of the total input data was used to train VAE model with 20 epochs and 30% data was used for validation. The model was trained using the Adam optimizer[41] with a learning rate of 0.00001. Mini-batch learning was employed with a batch size of 500. See Figure 6.A.2 for the VAE loss of training and validation.

6.2.6 VAMPnets model details

We performed VAMPnets analysis on the same data used for the VAE latent space construction using Deeptime[42]. Theory of VAMPnets is discussed in section 2.5.1 in Chapter 2. We used 350 ns of lag-time to generate instantaneous and time-lagged datasets for VAMPnets training and validation. The input data was normalized using the StandardScaler function of the scikit-learn python library[36] before feeding to the neural networks. VAMPnets neural network lobes for both the instantaneous and time-lagged data shared the same architecture with three fully connected hidden layers with nodes 36-24-16 and a output layer with varying dimensions: 4, 5 and 6. Therefore, 4, 5 and 6-clusters VAMPnets models were investigated for our system. ReLU activation function was used for all the hidden neurons, while the output layer used a softmax activation function in order to obtain fuzzy discretization of the state space to the target number of clusters. 80% of the total input data was used to train the models with 100 epochs and 20% data was used for validation for each output dimensions. The models were trained using the Adam optimizer[41] with a learning rate of 0.0001. Mini-batch learning was employed with a batch size of 500. See Figure 6.A.3 for the VAMP2[43] scores of training and validation.

6.2.7 Markov state modelling

Theory of MSM is discussed in section 2.5 of Chapter 2. In this present study, we built a MSM directly from the VAMPnets 5-clusters discretization for a lag-time of 70 ns from the cumulative 43 μs discrete trajectories using Deeptime[42]. Inter-state transition rates and stationary probabilities of the states were calculated from the MSM.

6.3 Results and discussions

6.3.1 Feature selection using RF classifier and identifying conformational clusters in the high dimensional feature space

We performed the RF analysis to identify the key residue pairs of the laccase which can be used as features. Feature importance was derived for each input features and used as a measure of relative importance of a feature over others. The top 38 features capture 80% of the total feature importance as shown in Figure 6.1a. These 38 residue pairs are shown in Figure 6.1b by lines connecting the C_α atoms of the respective residues and listed in Table 6.1. Interestingly, we found that many residues out of these 38 pairs present in the active site loop (G159-P160-A161-F162-P163-L164) as well as in the catalytic site of laccase (see Figure 6.1b). To assess the uniqueness of these identified residue pairs, we repeated the whole RF protocol for 200 replicates. The probability of observing at least *n* number of different residue pairs out of the 38 pairs is plotted in Figure 6.A.4. $\mathrm{C}_\alpha - \mathrm{C}_\alpha$ atom distances of these 38 residue pairs were used as features for dimensionality reduction, deriving CVs for enhanced sampling and subsequent kinetic modelling of our system.

Before constructing the low dimensional projection space for the laccase apo conformational landscape, we opted to identify conformational clusters in this 38 dimensional feature space derived from the RF classifier. We performed HDBSCAN clustering with different minimum cluster sizes on this high dimensional space. We found that the six apo trajectories belong to six different clusters in this high dimensional feature space (see Figure 6.A.5). Moreover, identical clusters have been obtained for different minimum cluster sizes, showing the convergence of the results (Figure 6.A.5). In the next section 6.3.2, this identification of different conformational clusters in the high dimensional feature space will be used to find the best low dimensional projection space for the laccase apo conformational landscape.

Table 6.1. Selected residue pairs from the RF analysis.

F337 – A461	T105 – S225	K157 – E460
F44 – V126	P163 – D206	V425 – H454
F162 – F457	H66 – R243	F162 – F337
P4 – G38	A390 – H395	F337 – F344
G401 – N445	H402 – W449	A155 – G159
F397 – F404	C117 – A156	P163 – F457
Q70 – F97	W65 – Q70	F337 – H395
F162 – A461	N336 – A461	P446 – N478
H452 – F463	F330 – T428	I82 – F344
A329 – I339	F450 – L459	M328 – N340
T114 – F457	T114 – A156	Q70 – W449
V154 – G159	F44 – V99	I356 – A475
A156 – F457	F162 – G334	

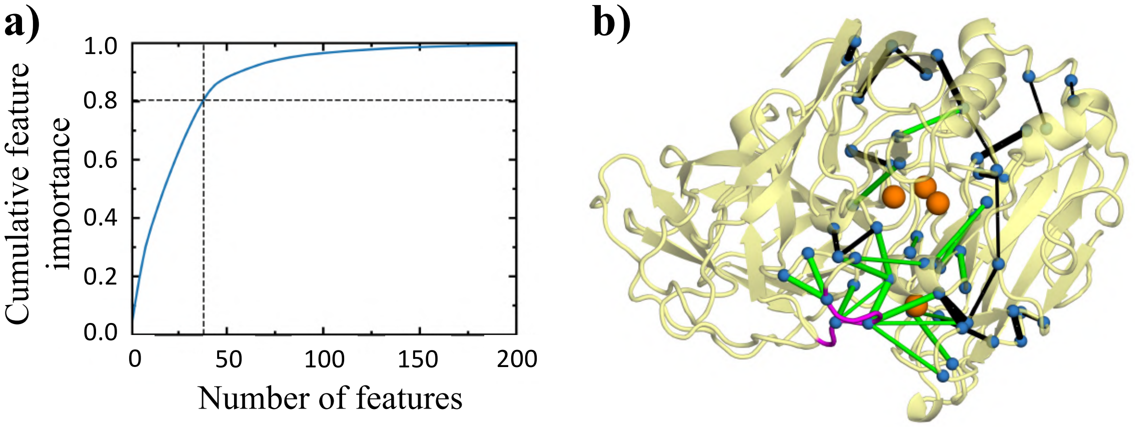


Figure 6.1. a) Cumulative feature importance for the top 200 features. The top 38 features capture 80% of the total feature importance as shown by the black dashed lines. **b)** 38 selected residue pairs identified using the RF classifier are shown on the laccase 3D structure. The lines connect the C_α atoms (blue spheres) of the residue pairs. The thickness of a connection represents the importance of the respective residue pair. The green lines involve the residues present in the active site loop (G159-P160-A161-F162-P163-L164) and in the catalytic site of laccase. The active site loop is shown in magenta colour and the copper atoms are shown as orange spheres.

6.3.2 Identifying the best low dimensional projection space for the laccase apo conformational landscape: TICA vs VAE

In this section, we discuss and compare different low dimensional projection spaces for our system. We have performed TICA on this cumulative 43 μs discrete trajectories. In addition, a 2-dimensional VAE latent space was also constructed using the same input feature dataset for the new TICA. The projections of the six "ligand delete apo trajectories", on the TICA eigenvectors and on the VAE latent space are shown Figure 6.2a-b, respectively. Both the projection plots are coloured according to the clusters obtained from the HDBSCAN method performed on the high dimensional feature space. Note that each "ligand delete apo trajectories" belongs to different clusters on the high dimensional feature space as observed in the previous section. Here, we can also see that on both of these low dimensional projection spaces, each trajectory also belongs to separate clusters. However, substantial overlaps between these different clusters have been observed on the 2-dimesional TICA space, whereas all the clusters are well separated on the 2-dimesional VAE latent space. Thus, the VAE latent space correctly mapped the high dimensional feature space for our system.

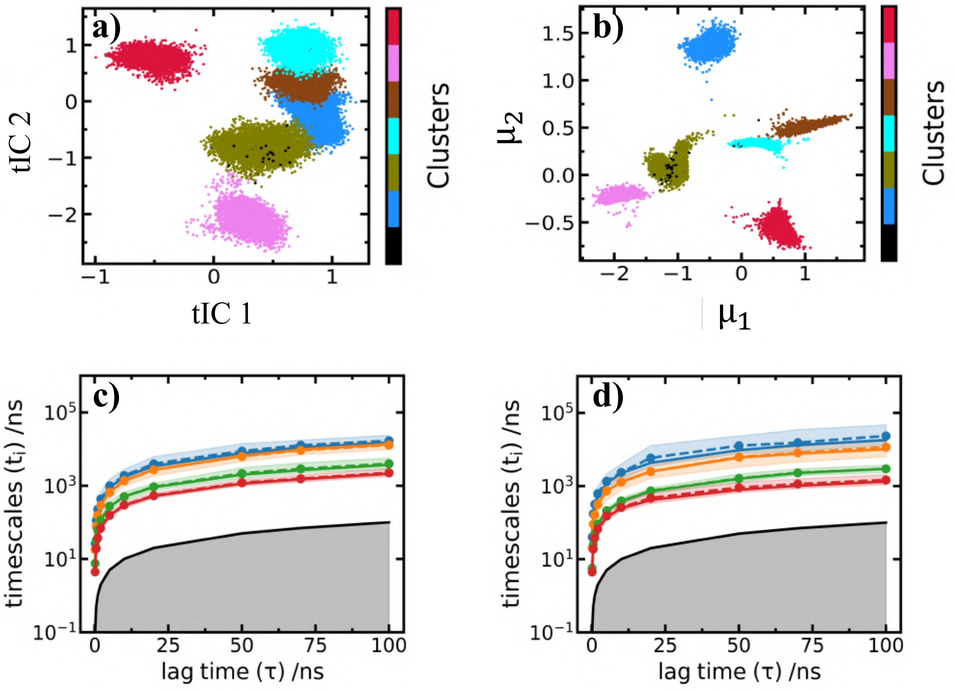


Figure 6.2. a-b) Projections of the six "ligand delete apo trajectories" on the first two TICA eigenvectors (tIC 1 and tIC 2) and on the 2-dimensional VAE latent space (μ_1 and μ_2), respectively. Both the plots are coloured according to the clusters obtained from HDBSCAN analysis performed on the high dimensional feature space. **c-d)** Implied timescale vs lag-time plots from K-means microstate clustering on the high dimensional TICA space and 2-dimensional VAE latent space, respectively. Standard errors are shown. Grey area covers the area where implied timescale is less than or equal to lag-time.

Furthermore, we calculated the implied timescales (see section 2.5 of Chapter 2) to identify the time scales of the slowest processes to assess the quality of the projections. We performed K-means clustering on the 10-dimensional TICA space as well as on the 2-dimesional VAE latent space and subsequently estimated the implied timescales for various lag-times. Implied timescale vs lag-time plots derived from TICA and VAE projection spaces are shown in Figure 6.2c-d, respectively. Similar time scales were predicted in both of the scenarios. However, the errors associated with the largest time scale are marginally higher for the VAE latent space compared to that of the TICA projection space. These results indicate that the 2-dimensional VAE latent space not only correctly mapped the high dimensional feature space, it also correctly encodes the same slow processes learned by TICA even though there was no time information in VAE training. Thus, VAE outperformed TICA as a dimensionality reduction method in our case both in terms of correctly mapping the high dimensional space as well as encoding the slow processes for our system. So, we choose the VAE latent space as the best low dimensional space for describing the laccase apo conformational landscape.

6.3.3 Comparing the performance of different kinetic clustering methods to identify laccase metastable apo conformational states: PCCA+ vs VAMPnets

As the superiority of the VAE latent space as a dimensionality reduction method over TICA for our system has been established, we built a MSM using 70 ns of lag-time using the K-means clustering performed on the 2-dimensional VAE latent space and subsequently coarse-grained the model using PCCA+[44] kinetic clustering method to identify metastable apo conformational states of laccase. Initially, PCCA+ kinetic clustering was performed with 3 output states. These 3 metastable states are shown in Figure 6.A.6 on the 2-dimensional VAE latent space. We observed that the PCCA+ generated metastable states were not well separated on the VAE projection space. Moreover, PCCA+ clustering to more than 3 metastable states could not be performed because of the appearance of negative elements while coarse-graining the corresponding transition matrices.

To examine whether we can obtain more than 3 metastable states, we performed VAMPnets with 4, 5 and 6 output states. Subsequently, implied time scales for various lag-times were calculated for the 4, 5 and 6-clusters VAMPnets models. In Figure 6.3a-b, we have shown the implied time scales at various lag-times for the 4 and 5-clusters VAMPnets models, respectively and the same plot for the 6-clusters VAMPnets model is shown in Figure 6.A.7a. 4 and 5-clusters VAMPnets models produced least errors for the largest implied time scale, whereas the errors associated with that of the 6-clusters VAMPnets model are pretty high (see Figure 6.A.7a). Moreover, as shown in Figure 6.3a-b, the time scales were predicted similarly for the 4 and 5-clusters VAMPnets models, showing convergence of the results. Furthermore, note that the higher implied timescales were predicted from the VAMPnets than TICA, consistent with observations from a recent study[45]. In Figure 6.3c-d, we have shown the VAMPnets generated 4 and 5-state clustering on the 2-dimensional VAE latent space, respectively. 6 VAMPnets clusters are shown in Figure 6.A.7b. It can be seen that all the VAMPnet generated clusters are well separated on the VAE latent space unlike the PCCA+ generated clusters. In addition, VAMPnets has been able to identify more metastable apo conformational states of laccase than PCCA+. Finally, we decided to select the metastable conformations obtained from the 5-clusters VAMPnets model for further analyses.

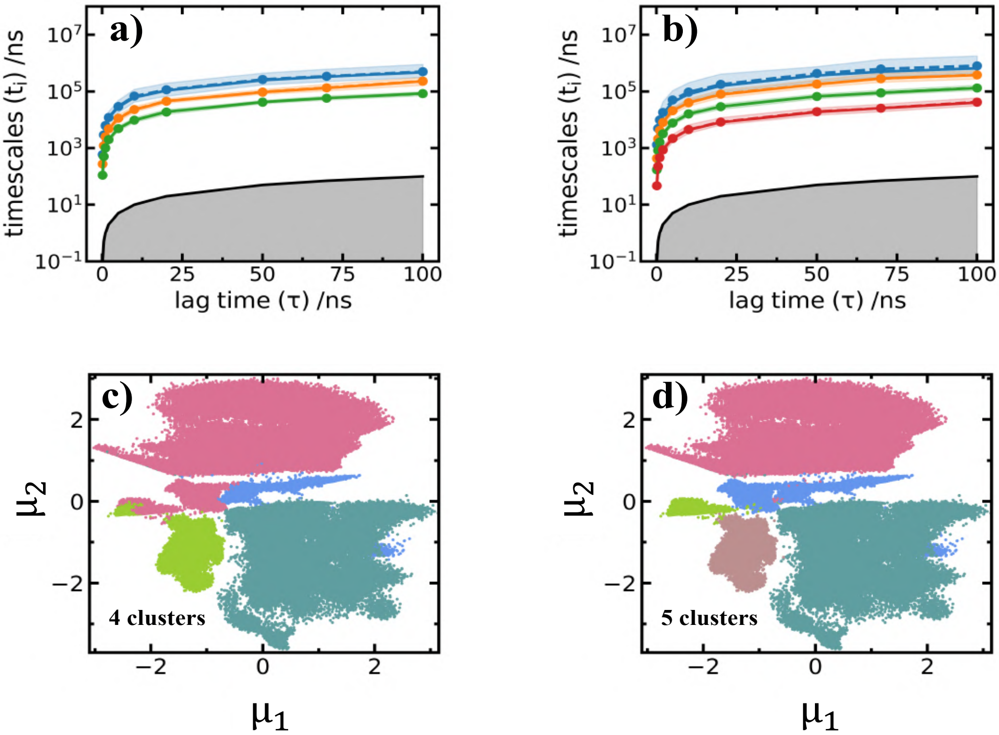


Figure 6.3. a-b) Implied timescale vs lag-time plots for the 4 and 5-clusters VAMPnets models, respectively. Standard errors are shown. Grey area covers the area where implied timescale is less than or equal to lag-time. **c-d)** 4 and 5 VAMPnets clusters are shown on the VAE latent space. Different clusters are shown using different colours.

6.3.4 Laccase apo free energy surface, metastable conformational states and Markov state model

The final five apo metastable conformational states of the laccase, denoted as S1-S5, obtained from the 5-clusters VAMPnets model are shown on the 2-dimensional VAE latent space in Figure 6.4a. A MSM was constructed from this five state assigned discreate trajectories using a lag-time of 70 ns. The MSM passed the Chapman-Kolmogorov (CK) test (Figure 6.A.8). The reweighted laccase apo free energy surface (FES) is shown in Figure 6.4b. Representative conformations of the laccase active site in these five metastable states are shown in Figure 6.4c. Note that five distinct conformations of the laccase active site have been observed in these five metastable states.

Inter-state transition rates and stationary probabilities of these five metastable states were calculated from the MSM. Highest stationary probability of the state S2 makes it the predominant laccase apo state followed by the state S4 and others. We have provided a kinetic network in Figure 6.4d by omitting transitions with rates smaller than 1 ms^{-1}, i.e mean first passage time (MFPT) greater than 1000 μs for visual clarity. It can be seen that the transition rates between different laccase apo metastable conformations are very low (all in the order of ms^{-1}). Moreover, S1 → S4 transition is the fastest among all, while the rates of all the other transitions are at least one order of magnitude smaller.

Several studies have proposed hydrogen bond as a good reaction coordinate to understand protein conformational changes[46–48]. Therefore, to understand why the inter-state transitions have such sluggish kinetics and provide a physical picture associated with these transitions in terms of inter-residue hydrogen bond breaking and formation, we performed hydrogen bond occupancy analysis[49] using 50 conformations taken from each of the five metastable state free energy basins. We generated a collection of residue pairs that were found to be hydrogen bonded in at least one-third conformations out of 50 and the corresponding occupancy values were calculated for all the five states. Then the lists of occupancy values were compared for each pair of states to calculate the percentage dissimilarities in hydrogen bond networks. In Figure 6.4e, we have shown the percentage dissimilarities in hydrogen bond networks for each pair of metastable states and found that the dissimilarities in hydrogen bonding for each pair of metastable states are more than 40%. Therefore, all the metastable conformational states are significantly different from each other in terms of presence or absence of inter-residue hydrogen bonds. In Figure 6.4f, we have shown those inter-residue hydrogen bonds that need to form and break to make a transition from the second most populated apo state S4 to the most populated state S2 of laccase. Similar pictures for some of the other inter-state transitions are shown in Figure 6.A.9. It can be observed that these hydrogen bonds are scattered throughout the entire protein, not localised only around the active site. All these results indicate that to make a transition between any two metastable states, the protein has to perform a large amount of hydrogen bond reformation (breaking and making) at a global level, not only limited around the active site. This could be a possible reason of observing such slow transitions between the metastable apo conformations of laccase.

After finding these metastable apo conformational states of the laccase, we are now in a position to qualitatively investigate whether different dye molecule binding to the laccase occur via conformation selection or induced fit mechanism.

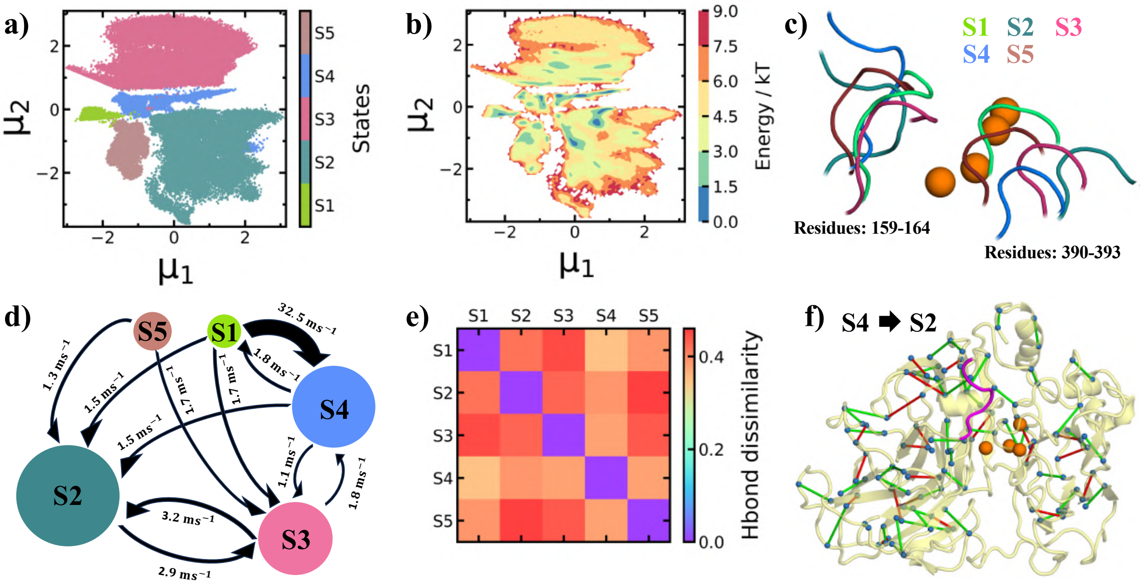


Figure 6.4. a) Apo metastable states (S1-S5) of the laccase obtained from the 5-cluster VAMPnets model are shown on the 2-dimensional VAE latent space. **b)** Reweighted laccase apo free energy surface. **c)** Representative conformations of the active site of the laccase in these five states. Copper atoms are shown as orange spheres. **d)** A kinetic network shows transition rates between the metastable states. Transitions with rates smaller than 1 ms^{-1} are not shown for visual clarity. Width of a arrow is proportional to the corresponding transition rate, whereas circle size represents the stationary probability of the respective metastable state. **e)** Percentage dissimilarities for a collection of inter-residue hydrogen bonds (Hbond) between each pair of states. **f)** Inter-residue hydrogen bonds that need to form (green lines) and break (red lines) for the S4 → S2 transition. The lines connect the C_α atoms (blue spheres) of the respective residue pairs. The active site loop (G159-P160-A161-F162-P163-L164) is shown in magenta colour and the copper atoms are shown as orange spheres.

6.3.5 Dye molecule binding to laccase: Conformational selection or Induced fit

In this section, we will try to qualitatively answer whether the five dye molecules, e.g brilliant blue, coumarin 343, crystal violet, methyl green and thioflavin T, bind to the laccase via conformational selection or induced fit by projecting the dye bound laccase conformations on the apo FES. The results are shown in Figure 6.5a. This figure shows that projections of all the five dye bound laccase conformations reside on different regions of the apo FES. A close inspection further reveals that different dye bound laccase conformations are actually correspond to different laccase apo metastable states: (i) brilliant blue bound laccase conformations are residing in the S3 state, (ii) methyl green, coumarin 343 and thioflavin T bound conformations are residing in the S4 state and (iii) crystal violet bound conformations are residing in the S1 state. Therefore, these laccase apo metastable conformations are dye molecule binding competent conformations. Also note that as discussed in the previous section 6.3.4, among these five metastable states, S2 is the predominant one. Thus, all the five dye molecules bind to higher energy laccase apo conformations. From all these observations, we hypothesize that these five dye molecule binding to the laccase proceeded possibly via conformational selection mechanism.

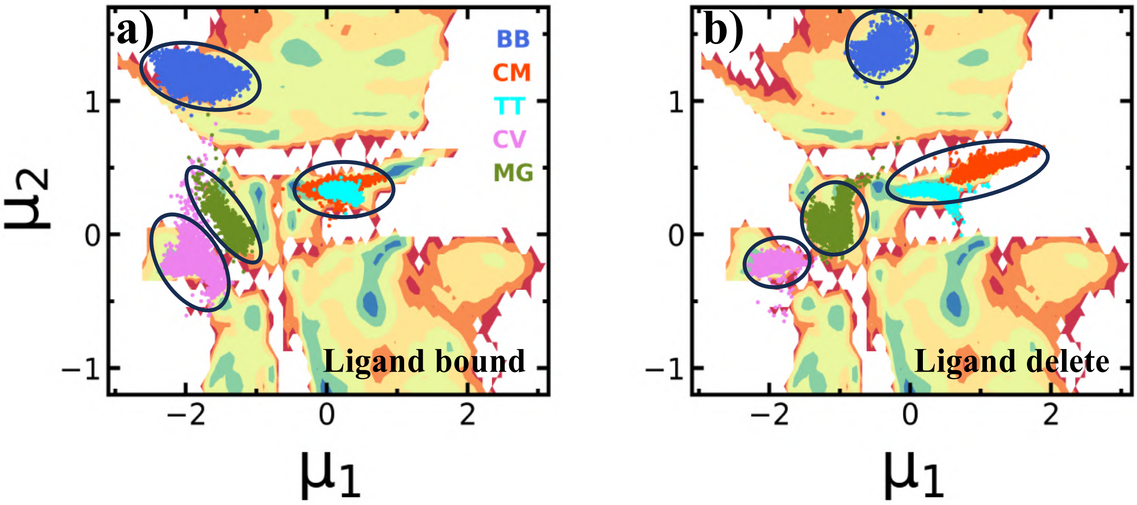


Figure 6.5. a) Projections of the brilliant blue (BB), coumarin 343 (CM), crystal violet (CV), methyl green (MG) and thioflavin T (TT) bound protein conformations (labelled as Ligand bound) on the laccase apo FES. **b)** Projections of the last 500 ns of the respective apo protein trajectories generated after deleting the dyes from the laccase active site (labelled as Ligand delete).

In Figure 6.5b, we have shown the projections of the last 500 ns of the respective apo protein trajectories generated after deleting the dyes from the laccase active site. It can be seen that while the protein moves to the local minimum after removing brilliant blue dye, it resides roughly in the same locations of the respective regions on the FES after deleting the other four dyes from the laccase active site. Laccase apo and holo FES should be compared to calculate how much the population of these binding competent metastable states change in holo as compared to apo in order to develop a deeper understanding of the dye molecule binding phenomenon.

6.3.6 Allostery in laccase

Interestingly, we have observed that another loop (loop-2; resids:331-337) shows significant conformational difference in the 5 laccase apo metastable states as shown in Figure 6.6a. In the state S2, this loop is disordered, whereas in the states S4 and S5, this loop forms anti-parallel β-sheet and helix respectively. Loop-2 is around 2 nm away from the loop-1 (resids:G159-P160-A161-F162-P163-L164) and forms the other end of the laccase active site. Moreover, we have observed that residues F332 and F337 residing in loop-2 are in different orientations in different metastable states. Previous experimental mutation study has shown that F332A mutation dramatically affects the catalytic efficiency of laccase[50]. Therefore, from all these observation, we propose that there is an allosteric connection between the loop-1 and loop-2 in laccase.

We have investigated this allosteric connection using mutual information theory and *correlationplus* python package[51]. In Figure 6.6b, we have shown the correlations between pairs of residues in these two loops. We can observed that the motions of the two loops are highly correlated. In Figure 6.6c, we have provided an possible allosteric pathway starts from the residue F332 (source) in loop-2 and ends at the residue F162 (sink) in loop-1 derived from the correlation values. We have also observed significant change in some inter-residue hydrogen bond occupancies around these two loops. These residue pairs are also shown in Figure 6.6c.

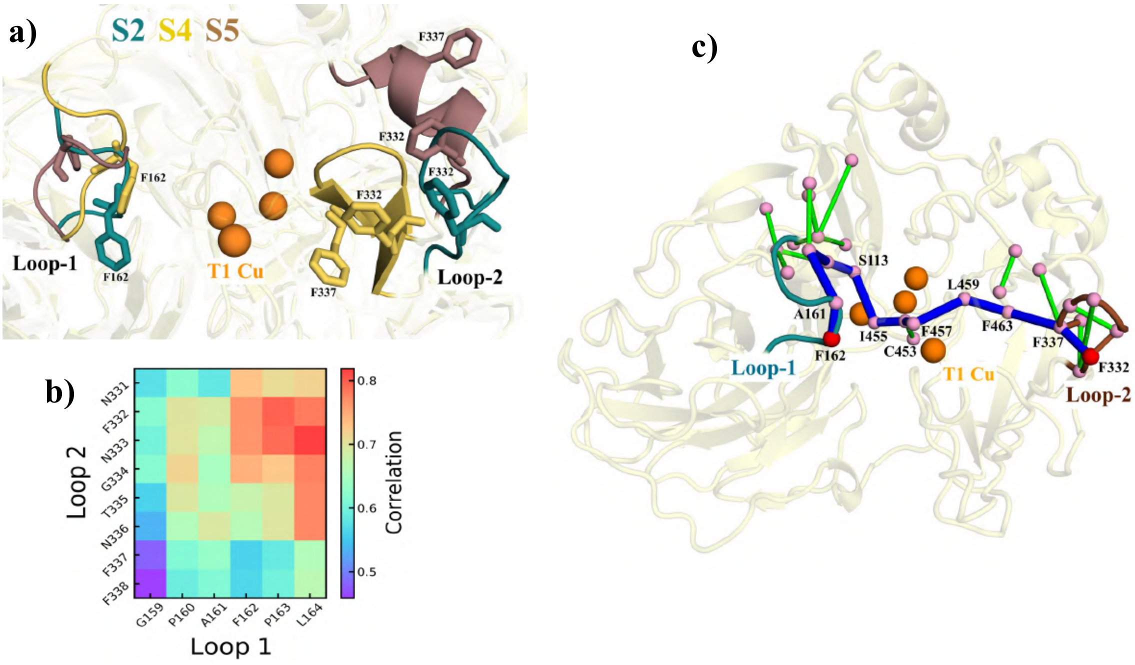


Figure 6.6. a) Different conformations of the loop-2 (resids: 331-337) in different metastable states are shown. **b)** Correlations between the pairs of residues residing in these two loops are derived from linear mutual information theory. **c)** Possible allosteric pathway starts from F332 (source, C_α atom as red sphere) in loop-2 and ends at F162 (sink, C_α atom as red sphere) in loop-1. The residues residing on this pathway are also shown. The green lines connect the C_α atoms (pink spheres) of the residue pairs that show significant difference in inter-residue hydrogen bond occupancies. Cu atoms are shown as orange spheres.

6.4 Conclusions

We have discussed the dye binding mechanism to laccase using classical MD simulations coupled with various machine learning approaches. We started our analysis of simulation data using the Random Forests classifier to identify important set of residue pairs as features for our system. Using the selected residue pairs and HDBSCAN clustering method, we discovered the existence of distinct apo conformational clusters in the high dimensional feature space. Subsequently, we compared the results of two different classes of dimensionality reduction methods: (i) a linear method, time-lagged independent component analysis (TICA) and (ii) a non-linear method, variational autoencoder (VAE) to obtain the best low dimensional representation for our system. The results showed that 2-dimensional

VAE latent space not only correctly mapped the high dimensional feature space, it also encoded the same slow processes learned by TICA even though there was no time information in VAE training. Thus, We conclude that VAE outperformed TICA as a dimensionality reduction method for our system.

We next performed kinetic clustering using VAMPnets deep learning method. A Bayesian MSM built from the VAMPnets clustering to 5 metastable states showed that the transition between these states were very slow (transition rates were in the order of ms^{-1}). Hydrogen bond occupancy analysis showed that all these states are significantly different from each other in terms of inter-residue hydrogen bond network. In addition, we also found out that these hydrogen bonds are scattered throughout the entire protein, not localised only around the active site. Therefore, we conclude that to make a transition between any two metastable states, the protein has to perform a lots of hydrogen bond reformation at a global level. This physical picture provides a way to understand the observed slow transitions between the metastable states.

Projections of the five dye molecule bound protein conformations on the apo free energy surface revealed that all the laccase bound conformations corresponded to different aforementioned apo metastable conformational states of the protein. Therefore, the dye molecules bind to the pre-existing conformations of the laccase. From all these observations, we hypothesize that these five dye molecule binding to the laccase proceeded possibly via conformational selection mechanism. However, laccase apo and holo free energy surface should be compared to calculate how much the population of these binding competent metastable states change in holo as compared to apo in order to develop a deeper understanding of the dye binding phenomenon.

We have also observed significant conformational change occurred in another loop in these metastable states, far away from the active site loop. Together with experimental mutation data and allosteric analyses, we have proposed an allosteric connection between this loop and our previously identified active site loop.

Appendix 6.A

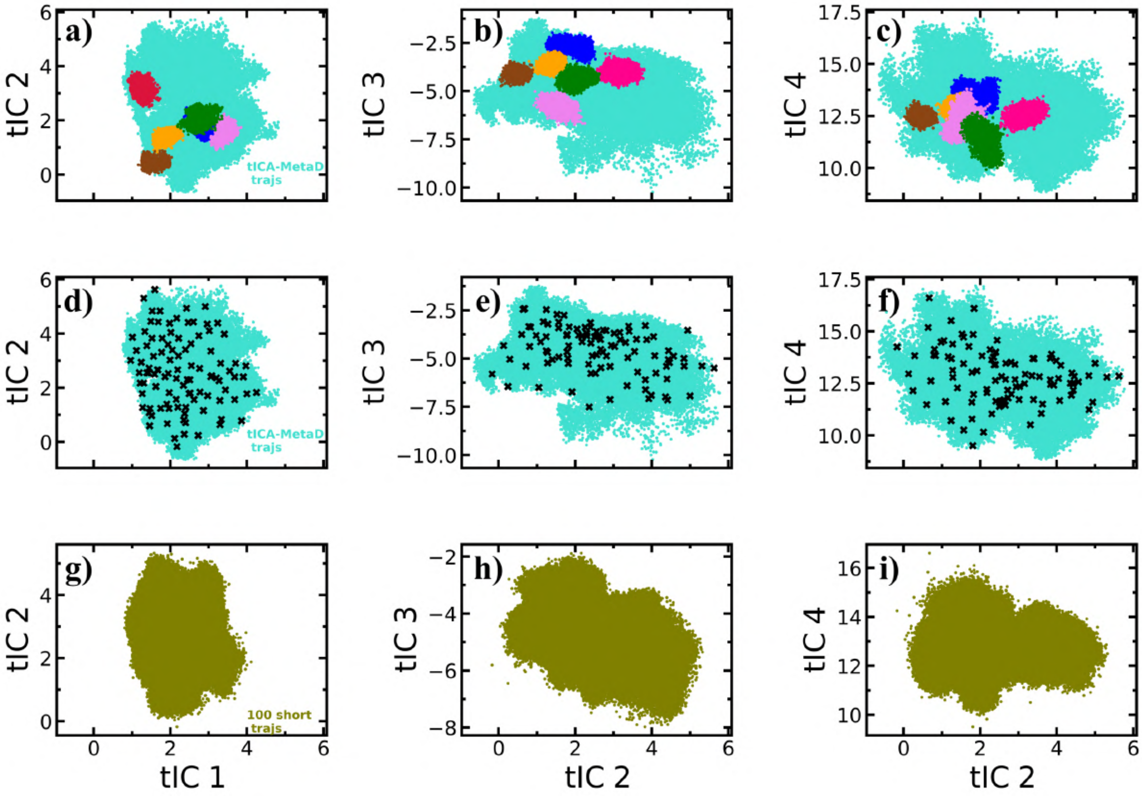


Figure 6.A.1. a) – c) Projections of the 6 apo "ligand delete apo trajectories" and 5 TICA-Metadynamics trajectories are shown on different TICA eigenvectors. **d) – f)** The crosses denote 100 different conformations used to generate 100 short equilibrium trajectories. **g) – i)** 100 short equilibrium trajectories are projected on different TICA eigenvectors.

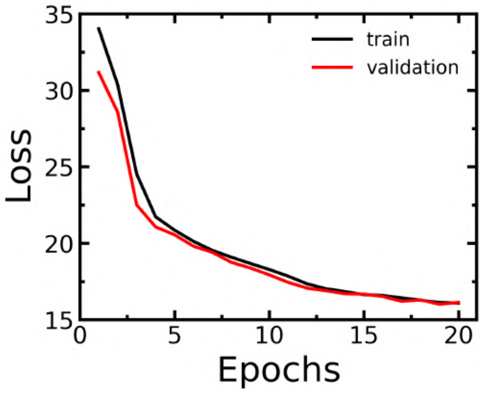


Figure 6.A.2. Variational autoencoder loss of training and validation.

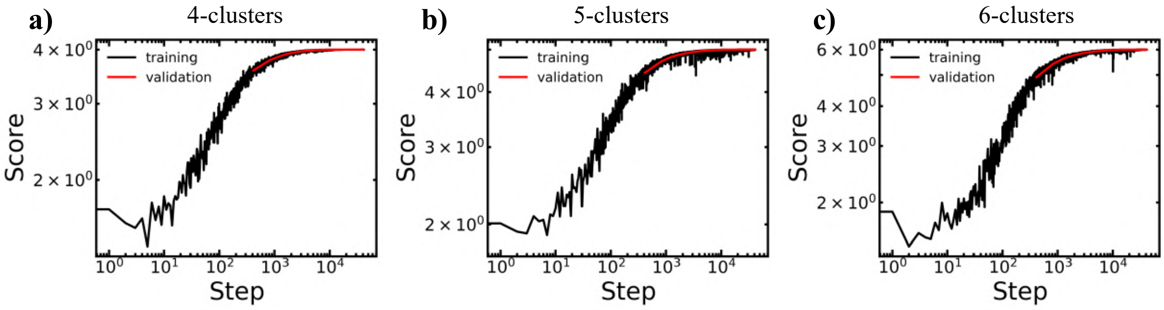


Figure 6.A.3. a) – c) VAMP2 scores for 4, 5, and 6-cluster VAMPnets model training and validation.

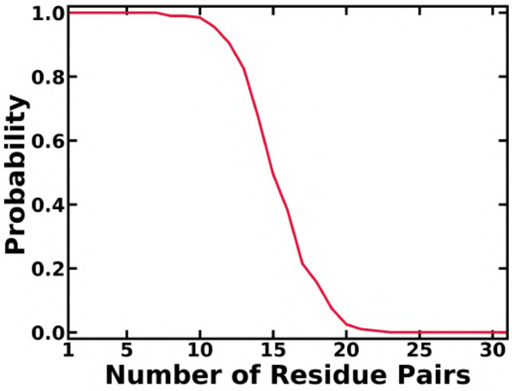


Figure 6.A.4. Probability of observing at least n number of different residue pairs out of the 38 selected residue pairs over replicates of Random Forest analysis.

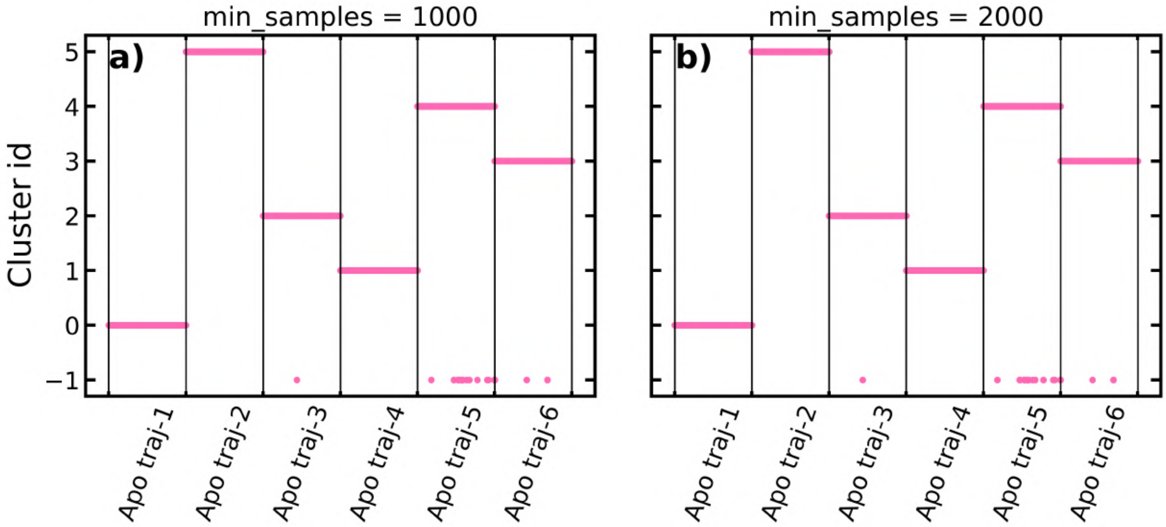


Figure 6.A.5. HDBSCAN assigned clusters for the six "ligand delete apo trajectories" for two *min_samples* parameter values.

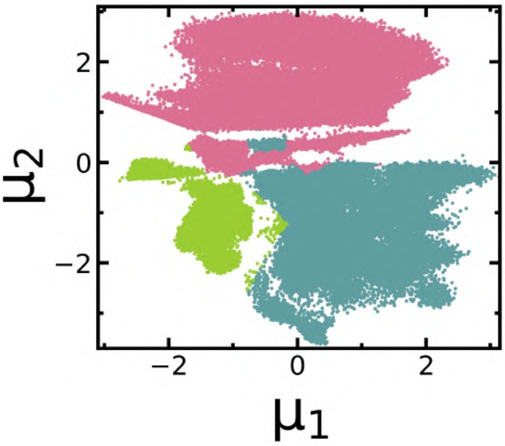


Figure 6.A.6. 3 clusters obtained from the PCCA+ kinetic clustering method are shown on the VAE projection space. Different clusters are shown using different colours.

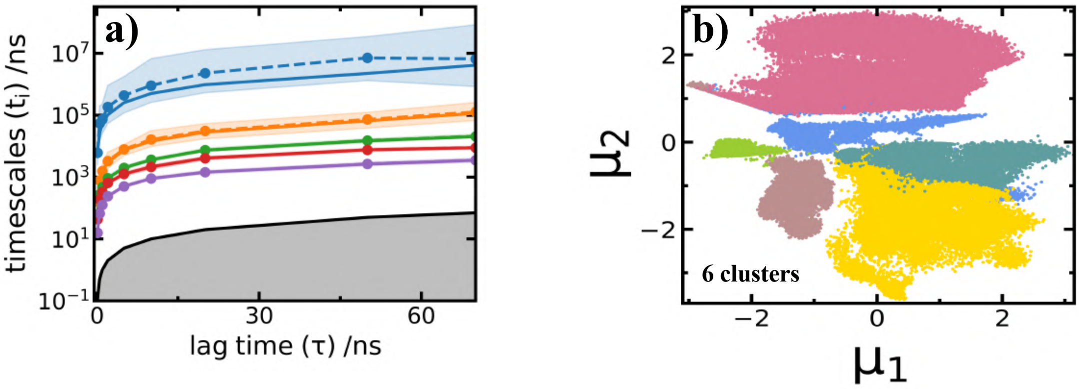


Figure 6.A.7. a) Implied timescale vs lag-time plots for the 6-cluster VAMPnets model. Standard errors are shown. Grey area covers the area where implied timescale is less than or equal to lag-time. **b)** The 6 clusters are shown on the VAE projection space. Different clusters are shown by different colours.

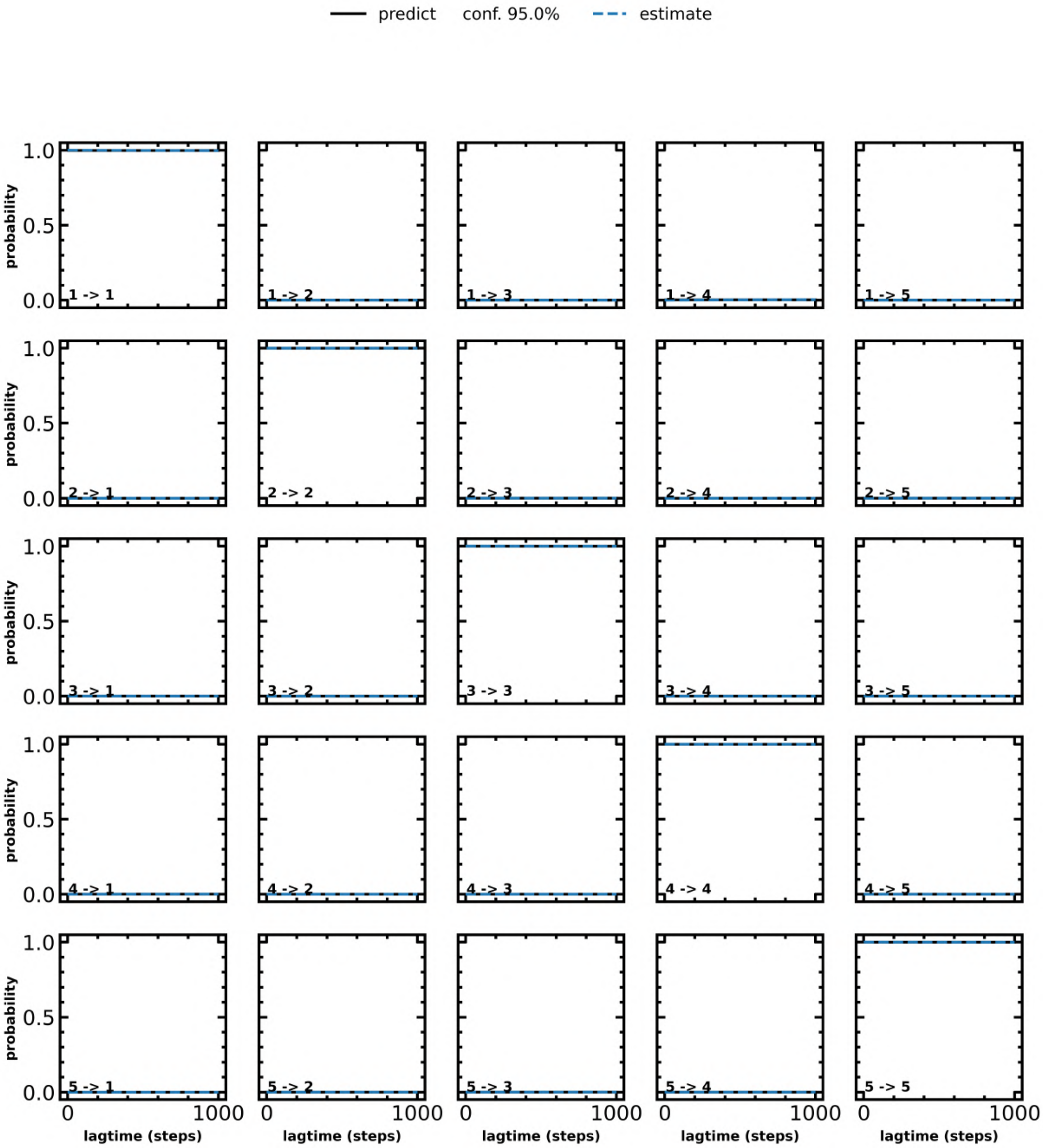
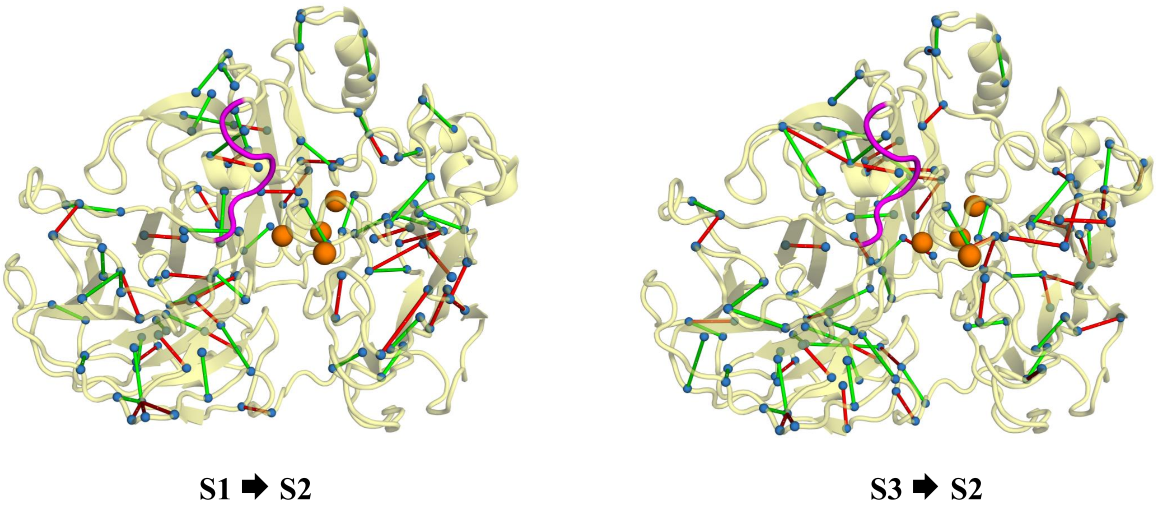


Figure 6.A.8. Chapman-Kolmogorov (CK) test 5 state MSM.



S1 ➡ S2

S3 ➡ S2

Figure 6.A.9. Inter-residue hydrogen bonds that need to form (green lines) and break (red lines) for the corresponding transition. The lines connect the C_{α} atoms (blue spheres) of the respective residue pairs. The active site loop (G159-P160-A161-F162-P163-L164) is shown in magenta colour and the copper atoms are shown as orange spheres.

References

1 H. X. Zhou, *Biophys J*, 2010, **98**, L15–L17.

2 D. E. Koshland Jr, *Angewandte Chemie International Edition in English*, 1995, **33**, 2375–2378.

3 A. D. Vogt and E. Di Cera, *Biochemistry*, 2012, **51**, 5894–5902.

4 G. G. Hammes, Y. C. Chang and T. G. Oas, *Proc Natl Acad Sci U S A*, 2009, **106**, 13737–13741.

5 A. Biswas and M. Radhakrishna, *Journal of Physical Chemistry B*, 2025, **129**, 3761–3775.

6 B. E. Husic and V. S. Pande, *J Am Chem Soc*, 2018, **140**, 2386–2396.

7 V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, **52**, 99–105.

8 S. Doerr, I. Ariz-Extreme, M. J. Harvey and G. De Fabritiis, .

9 R. K. Cersonsky, B. Cheng, M. De Vivo and P. Tiwary, *J Chem Theory Comput*.

10 Y. Wang, J. M. Lamim Ribeiro and P. Tiwary, *Curr Opin Struct Biol*, 2020, **61**, 139–145.

11 S. Mehdi, Z. Smith, L. Herron, Z. Zou and P. Tiwary, *Annu Rev Phys Chem*, 2024, **75**, 347–370.

12 L. Breiman, *Mach Learn*, 2001, **45**, 5–32.

13 N. Ahalawat, M. Sahil and J. Mondal, *J Chem Theory Comput*, 2023, **19**, 2644–2657.

14 L. Molgedey and H. G. Schuster, *Phys Rev Lett*, 1994, **72**, 3634.

15 G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, *Journal of Chemical Physics*.

16 C. R. Schwantes and V. S. Pande, *J Chem Theory Comput*, 2013, **9**, 2000–2009.

17 S. Schultze and H. Grubmüller, *J Chem Theory Comput*, 2021, **17**, 5766–5776.

18 N. Ahalawat and J. Mondal, *J Am Chem Soc*, 2018, **140**, 17743–17752.

19 L. Bonati, V. Rizzi and M. Parrinello, *Journal of Physical Chemistry Letters*, 2020, **11**, 2998–3004.

20 L. McInnes, J. Healy and J. Melville, .

21 L. Van Der Maaten and G. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

22 D. P. Kingma and M. Welling, *Foundations and Trends in Machine Learning*, 2019, **12**, 307–392.

23 C. Wehmeyer and F. Noé, *Journal of Chemical Physics*.

24 D. Maity and S. Chakrabarty, *J Chem Theory Comput*, 2025, **21**, 1916–1928.

25 Z. Belkacemi, M. Bianciotto, H. Minoux, T. Lelièvre, G. Stoltz and P. Gkeka, *Journal of Chemical Physics*, 2023, **159**, 24122.

26 S. Bhattacharya and S. Chakrabarty, *Biophys Chem*, 2025, **319**, 107389.

27 D. Wang, Y. Wang, L. Evans and P. Tiwary, *J Chem Theory Comput*, 2024, **20**, 3503–3513.

28 A. Mardt, L. Pasquali, H. Wu and F. Noé, *Nature Communications 2018 9:1*, 2018, **9**, 1–11.

29 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J Chem Phys*, 1983, **79**, 926–935.

30 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindah, *SoftwareX*, 2015, **1–2**, 19–25.

31 G. Bussi, D. Donadio and M. Parrinello, *J Chem Phys*, 2007, **126**, 14101.

32 M. Parrinello and A. Rahman, *J Appl Phys*, 1981, **52**, 7182–7190.

33 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J Chem Phys*, 1995, **103**, 8577–8593.

34 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J Comput Chem*, 1997, **18**, 1463–1472.

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

36 F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, andÉdouard Duchesnay and Fré. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

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

38 M. Oh, G. C. A. da Hora and J. M. J. Swanson, *J Chem Theory Comput*, 2023, **19**, 8886–8900.

39 M. M. Sultan and V. S. Pande, *J Chem Theory Comput*, 2017, **13**, 2440–2447.

40 G. A. Tribello, M. Bonomi, G. Bussi, C. Camilloni, B. I. Armstrong, A. Arsiccio, S. Aureli, F. Ballabio, M. Bernetti, L. Bonati, S. G. H. Brookes, Z. F. Brotzakis, R. Capelli, M. Ceriotti, K.-T. Chan, P. Cossio, S. Dasetty, D. Donadio, B. Ensing, A. L. Ferguson, G. Fraux, J. D. Gale, F. L. Gervasio, T. Giorgino, N. S. M. Herringer, G. M. Hocky, S. E. Hoff, M. Invernizzi, O. Languin-Cattöen, V. Leone, V. Limongelli, O. Lopez-Acevedo, F. Marinelli, P. F. Martinez, M. Masetti, S. Mehdi, A. Michaelides, M. H. Murtada, M. Parrinello, P. M. Piaggi, A. Pietropaolo, F. Pietrucci, S. Pipolo, C. Pritchard, P. Raiteri, S. Raniolo, D. Rapetti, V. Rizzi, J. Rydzewski, M. Salvalaglio, C. Schran, A. Seal, A. S. Zadeh, T. F. D. Silva, V. Spiwok, G. Stirnemann, D. Sucerquia, P. Tiwary, O. Valsson, M. Vendruscolo, G. A. Voth, A. D. White and J. Wu, *J Chem Phys*.

41 D. P. Kingma and J. L. Ba, *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

42 M. Hoffmann, M. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, N. Kutz, S. L. Brunton and F. Noé, *Mach Learn Sci Technol*, 2021, **3**, 015009.

43 H. Wu and F. Noé, *J Nonlinear Sci*, 2020, **30**, 23–66.

44 S. Röblitz and M. Weber, *Adv Data Anal Classif*, 2013, **7**, 147–179.

45 S. J. McElhenney and J. Yu, *J Chem Theory Comput*, 2025, **21**, 4829.

46 C. Scaletti, P. P. S. Russell, K. J. Hebel, M. M. Rickard, M. Boob, F. Danksagmüller, S. A. Taylor, T. V. Pogorelov and M. Gruebele, *Proceedings of the National Academy of Sciences*, 2024, **121**, e2319094121.

47 Y. Kasprzak, J. Rückert, N. Ludolph, C. G. Hübner and H. Paulsen, *Journal of Chemical Physics*, 2025, **162**, 74107.

48 P. Pal, S. Chakraborty and B. Jana, *Journal of Physical Chemistry B*, 2022, **126**, 10822–10833.

49 K. Sinha, A. Kumawat, H. Jang, R. Nussinov and S. Chakrabarty, *Biophys J*, 2024, **123**, 57–67.

50 C. Galli, P. Gentili, C. Jolivalt, C. Madzak and R. Vadalà, *Appl Microbiol Biotechnol*, 2011, **91**, 123–131.

51 M. Tekpinar, B. Neron and M. Delarue, *J Chem Inf Model*, 2021, **61**, 4832–4838.

7

WeTICA: A directed search weighted ensemble based enhanced sampling method to estimate rare event kinetics in a reduced dimensional space

7.1 Introduction

MD simulation is an indispensable tool to understand different biomolecular processes and calculate associated thermodynamic and kinetic properties, for example, binding energy of a ligand to a target receptor[1,2], free energy landscape of proteins[3] and kinetic rate constants associated with ligand dissociation (and association) from (to) a target receptor[4–6]. Despite the remarkable advances in MD software and hardware that enable us to access millisecond time scales at atomistic resolution, normal MD simulations usually struggle to overcome the barriers associated with different processes making such events rare and difficult to capture.

A plethora of methods, known as enhanced sampling techniques, have been developed as a solution for sampling rare events. Some of the popular methods are Metadynamics (MetaD)[7–12] and its different variants[13], adaptive biasing force (ABF)[14–16], Gaussian-accelerated molecular dynamics (GaMD)[17,18], Replica exchange methods[19–23], $\tau -$Random acceleration molecular dynamics (τRAMD)[24,25] and various methods employing machine learning based techniques[26–29]. These enhanced sampling methods not only facilitate quick exploration of the rugged biomolecular free energy landscape, but also enable the calculation of rare event kinetics[30–35] and building kinetic models like the Markov state model (MSM)[36–39].

As discussed in section 2.7.2 of Chapter 2, weighted ensemble (WE)[40–44] simulation is a special class of enhanced sampling method capable of calculating kinetics of rare events. WE has proven to be useful in studying many biologically relevant systems[45–51]. WESTPA[52,53] is a popular toolkit to perform and analyse WE simulations. Generally, in conventional WE protocol, the configurational space is mapped onto a low dimensional collective variable (CV) space that describes the transition of interest, followed by dividing it into bins. Several trajectories (called "walkers" in WE framework) pre-assigned with probabilities or weights are simultaneously initiated from the initial structure. Walkers that reach new bins are cloned, that is, two new simulations will be started from that configuration, and walkers residing in the same bin will be merged[54]. In this way, the simulations evolve under the natural dynamics of the system with trajectory resampling (cloning + merging) for many iterations to enhance the sampling of rare events. WE simulation protocol has been successfully combined with other methods like milestoning (WEM)[55,56], gaussian accelerated molecular dynamics (GaMD-WE)[57], neural networks (DeepWEST)[58] to improve sampling and computing thermodynamic and kinetic properties. Since the resampling decisions are made based on the exploration of new bins in the CV space, the choice of appropriate CV space and binning schemes affect the algorithm's performance. Different binning techniques like voronoi polyhedral[59,60], finite temperature string method[61], minimal adaptive binning (MAB)[62], mean first passage time (MFPT) binning[63] have been developed to partition the CV space. Despite these mathematical developments, determination of good CV space and optimizing the binning scheme are still non-trivial tasks.

Resampling Ensembles by Variation Optimization or REVO[64,65] is a popular "binless" WE simulation algorithm based on the optimization of a quantity called "trajectory variation", which can be considered as a metric of how different the walkers are from each other. The use of binless WE method can bypass the hurdles of optimizing binning schemes to obtain quantitatively accurate and converged results. In a recent study[66], REVO was used to calculate residence times (in minutes time scales) of various inhibitors of soluble epoxide hydrolase (sEH). Despite the success of this algorithm, the current implementation of REVO in WE software Wepy[67] cannot be directly used to study other biologically relevant diverse class of problems like protein folding-unfolding or transition between different metastable conformational states of biomolecules.

To broaden the applicability of binless WE methods, in this chapter we have discussed about the development of a new binless WE algorithm utilizing the fundamental principles of the REVO algorithm. Our algorithm , named WeTICA, has been implemented based on the Wepy[67] codebase. Our proposed protocol uses a fixed pre-defined low dimensional linear CV space to drive the WE simulations toward the specified target state. In this work, we have demonstrated the performance of this new algorithm using projections along time-lagged Independent Component Analysis (TICA)[68] eigenvectors (see section 2.4.2) as CVs to recover the unfolding kinetics of three benchmark proteins: 1) TC5b Trp-cage mutant, 2) TC10b Trp-cage mutant and 3) Protein G with known unfolding times spanning the range between 3 μs − 40 μs.

7.2 Method

We have defined a new "trajectory variation" (V) function as follows:

$$V = \sum_i V_i = \sum_i \frac{1}{d^i_{\text{from target}}} \qquad (7.1)$$

where $d^i_{\text{from target}}$ is the distance of walker i measured from the target state conformation using some distance metric and the summation is over all the walkers at each cycle. Although the original algorithm of REVO used a different form for the variation function[64], our algorithm works similarly by maximizing this parameter V (Eq. (7.1)) using trajectory resampling.

Root mean squared distance (RMSD) between protein backbone atoms is a very high dimensional distance metric routinely used to quantify the difference between two conformational states. However, the RMSD CV suffers from a significant degeneracy problem at larger values. Thus, having a large and same RMSD value of two walkers with respect to some distant reference state does not necessarily mean that they have the same conformation/state. This degeneracy problem becomes increasingly pronounced with higher dimensionality of the problem.

The choice of a lower dimensional CV space not only tackles the degeneracy issue associated with high dimensions, but choosing appropriate CVs can separate the metastable states of interest and capture the transition processes between the states. However, we should always keep in mind that any projection from a higher dimensional space to lower dimensional space has other demerits as well. Time-lagged independent component analysis (TICA)[69,70] is capable of capturing slow modes in a molecular process. TICA is a linear dimensionality reduction method where the actual high dimensional descriptor data sets are projected on some leading TICA eigenvectors to visualize important metastable states in a low dimensional representation (see section 2.4.2 of Chapter 2). Furthermore, TICA components are not only widely used to build MSMs for studying kinetics from MD simulation data[71,72] , their use as CVs for enhanced sampling is also established[73]. Thus, we decided to project all the walkers after every cycle on some pre-defined TICA eigenvectors using some descriptor datasets and then calculate the Euclidean distances between the projections of the walkers and the projection of the target state conformation. So, the TICA eigenvectors serve as CVs in our algorithm. Note that the eigenvectors will be fixed throughout the entire simulation and due to the flexibility of this algorithm, any larger number of CVs (TICA eigenvectors) can be used without significant computational cost, unlike other popular enhanced sampling methods like umbrella sampling or metadynamics. Moreover, this algorithm can be used with any linear dimensionality reduction methods and is not just limited to TICA.

The $\mathrm{d}^{\mathrm{i}}_{\mathrm{from\ target}}$ parameter in Eq. (7.1) is the Euclidean distance between the projection of the *i*-th walker and the projection of the target state conformation on this fixed TICA projection plane. We have implemented two different sets of input featurization schemes: 1) pair-wise distances between the C_{α} atoms, and 2) distances between any set of selected atom pairs. Apart from calculating $\mathrm{d}^{\mathrm{i}}_{\mathrm{from\ target}}$, we also calculate distance between every pair of walkers d_{ij} to make the merging decision based on some distance cutoff on the same TICA projection plane.

Since variation V_{i} of the walker i is defined as the inverse of its distance from target (Eq. (7.1)), the walker that is closest to the target has the largest variation value. As a result, this algorithm selects the walker *i* which is closest to the target for cloning and the walker *j* which is farthest from the target as the first candidate for merging to fulfill the objective of maximizing total variation *V*. Then it will search for another walker *k* (excluding *i* and *j*) for the second candidate to form the merging pair with walker *j* if the distance between walker *j* and walker *k* is less than some predefined cut-off distance $\mathrm{d}_{\mathrm{merge}}$, such that $\mathrm{d}_{\mathrm{jk}} \leq \mathrm{d}_{\mathrm{merge}}$ and $\mathrm{w}_{\mathrm{j}} + \mathrm{w}_{\mathrm{k}} < \mathrm{p}_{\mathrm{max}}$. Here w_{j} and w_{k} are the weights of the walker *j* and walker *k*, respectively, and $\mathrm{p}_{\mathrm{max}}$ is the maximum weight that a walker can hold. We set $\mathrm{p}_{\mathrm{max}} = 0.20$ and any walker whose weight after cloning becomes less than $\mathrm{p}_{\mathrm{min}} = 10^{-12}$ is prohibited from cloning.

After selecting a suitable merging pair, a walker from that merging pair (*j*, *k*) is randomly selected for continuation with the total weight $(\mathrm{w}_{\mathrm{j}} + \mathrm{w}_{\mathrm{k}})$ and the other walker is discontinued. Like the original algorithm of REVO, this selection scheme goes on until V reaches a local maximum or no walkers are left for cloning and merging, at which point the ensemble is again propagated forward in time by the MD integrator.

To correctly calculate rate, it is necessary to recognise when the system has moved from the initial state basin and arrived in the desired one even if the precise knowledge of the corresponding transition state is not known[12]. We have used the target state conformation to steer the process to the correct direction. When a walker reaches inside a hypersphere of radius $\mathrm{d_{warp}}$ centred around the projection of the target state conformation on the same projection plane, it is considered that the walker crosses into the final desired state basin. Then the corresponding weight of the walker will be saved and that walker will be re-initiated with the starting conformation. This process is called *warping* and the corresponding walker is called *warped* walker. Mean first passage time (MFPT) is calculated directly from the weights of the warped walkers as follows:

$$\mathrm{MFPT} = \frac{\mathrm{T}}{\sum_{(\mathrm{i,t}) \in \mathcal{U}} \mathrm{w_{i,t}}} \tag{7.2}$$

where *T* is the total simulation time, $\mathcal{U}$ is a set of tuples denoting the walker indices *i* and the time point *t* when the walker is warped, $\mathrm{w_{i,t}}$ is the weight of the warped walker *i* at that time point *t*. A full workflow diagram illustrating our algorithm is presented in Figure 7.1. Table 7.A.1 summarizes the key differences between the original REVO algorithm and our proposed algorithm.

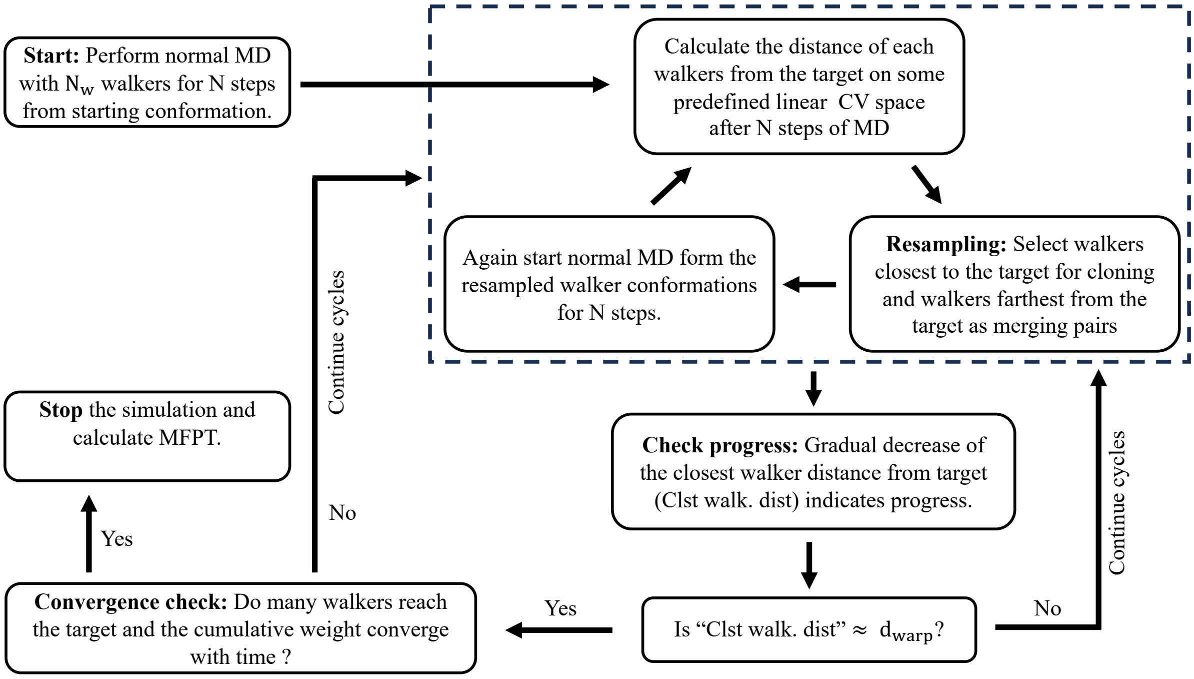

Start: Perform normal MD with $\mathrm{N_w}$ walkers for N steps from starting conformation.
Calculate the distance of each walkers from the target on some predefined linear CV space after N steps of MD
Resampling: Select walkers closest to the target for cloning and walkers farthest from the target as merging pairs
Again start normal MD form the resampled walker conformations for N steps.
Continue cycles
Stop the simulation and calculate MFPT.
Check progress: Gradual decrease of the closest walker distance from target (Clst walk. dist) indicates progress.
Continue cycles
No
Yes
Convergence check: Do many walkers reach the target and the cumulative weight converge with time ?
Yes
Is "Clst walk. dist" ≈ $\mathrm{d_{warp}}$?
No

Figure 7.1: Workflow diagram illustrating the WeTICA algorithm.

7.3 Guidelines for choosing WeTICA simulation parameters

Setting up a WeTICA simulation requires careful optimisation of certain input parameters to ensure efficient sampling of the productive trajectories. Here we provide some specific guidelines for choosing these parameters such as the number of walkers, generation and dimensionality of the CV space, walker merging distance $\mathrm{d}_{\mathrm{merge}}$ and the walker warping distance $\mathrm{d}_{\mathrm{warp}}$ that aim to an efficient and correct computation of MFPTs.

7.3.1 Number of walkers: Increasing the number of walkers will speed up the search process, but running greater number of simulations simultaneously will increase the computational wall time due to the limited number of computational resources. Thus, we advise to choose the number of walkers keeping a balance between these two factors.

7.3.2 Generation and dimensionality of CV space: The TICA eigenvectors should be computed before starting WeTICA simulations from unbiased MD trajectories (mostly in the two end states). As test cases, we have performed WeTICA simulations using TICA eigenvectors generated from both long unbiased trajectory and two short end-state simulation trajectories as discussed in the later sections of this chapter. It turns out that this method is working in both cases.

Since the major goal of this kind of enhanced sampling methods is to compute the kinetics of transition between the states, choose the minimum number of TICA eigenvectors as CVs that can ensure a proper separation of the two end states to avoid degeneracy arising due to large number of dimensions.

7.3.3 The choice of $\mathbf{d}_{\mathbf{merge}}$: Create a distribution using the distances between all the pair-wise projection points of the initial state ensemble conformations and then use the position of the minimum of this distance distribution as $\mathrm{d}_{\mathrm{merge}}$. This choice of $\mathrm{d}_{\mathrm{merge}}$ will resemble the approximate size of the initial state basin on that CV space and thus ensure that the merging walker pairs belong to the same basin.

7.3.4 The choice of $\mathbf{d}_{\mathbf{warp}}$: Create a distribution using the distances between all the pair-wise projection points of the target state ensemble conformations and then use the position of the minimum of this distance distribution as $\mathrm{d}_{\mathrm{warp}}$. This choice of $\mathrm{d}_{\mathrm{warp}}$ will resemble the approximate size of the target state basin on that CV space and thus ensure that the walkers have arrived in the target state basin when they are warped.

However, if the target state has high entropy (large number of possible configurations) with a flat basin in the free energy landscape, for example, targets in the protein unfolding or ligand unbinding processes, it might be difficult to sample enough data using short trajectories to obtain approximate size of the target state basin. In such cases, if the value for $\mathrm{d_{warp}}$ is not carefully chosen, the walkers may diffuse randomly without entering the hypersphere of radius $\mathrm{d_{warp}}$ centred around the representative target state conformation even though they have already reached the target state basin. This will slow down the convergence of the WeTICA simulation. In such cases, we advise to test various values of $\mathrm{d_{warp}}$ while ensuring sufficient escape form the initial state basin and choose the largest value as $\mathrm{d_{warp}}$ to speed up convergence and accurate calculation of rate constants.

7.4 Computational details

7.4.1 TC5b mutant of Trp-cage

The NMR structure (PDB ID: 1L2Y[74]) of the 20-residue TC5b Trp-cage mutant was modelled using CHARMM36 force filed[75]. The protein was solvated in a cubic box of ~44 Å side length containing 2679 TIP3P[76] water molecules and one $\mathrm{Cl^-}$ ion to neutralize the system. The system was then energy minimized using the steepest descent algorithm and equilibrated for 200 ps in NPT ensemble at 300K temperature and 1 atm pressure using GROMACS v2019.6[77]. Production run was conducted in NPT ensemble for 1.85 μs at the same temperature and pressure. Temperature and pressure were maintained using velocity rescale method[78] with time constant of 0.1 ps and Parrinelo-Rahman barostat[79] with a time constant of 2 ps respectively. All simulations were performed under periodic boundary conditions and the long range electrostatic interactions were handled using the Particle Mesh Ewald (PME)[80] summation method. The cut-off distances for electrostatic and van der Waals interactions were set to 10 Å. Bonds containing hydrogen atoms were constrained with LINCS[81]. Leap-frog integrator was used with an integration time step of 2 fs. Frames were saved at every 20 ps interval. After the completion of the production simulation, fraction of native contacts (Q) were calculated using MDTraj python package[82] for all the frames of the entire 1.85 μs trajectory with respect to the first frame considered as the folded structure. A snapshot with Q ~ 0.2 was taken as a representative of the unfolded state ensemble. This representative unfolded conformation was used as the target state conformation for the subsequent WeTICA simulations. Folded and the representative unfolded structures for this Trp-cage mutant are shown in Figure 7.2a.

We choose 153 $\mathrm{C_\alpha - C_\alpha}$ atom pair wise distances ($\mathrm{C_\alpha}$ atoms are at least 2 residues apart i.e i and i+3) as the feature to calculate TICA eigenvectors trained on the entire 1.85 μs long folding-unfolding trajectory with a lag-time of 10 ns using PyEMMA v2.5.12 Python package[83]. The first two TICA eigenvectors (TIC1 and TIC2) were used as CVs to perform several independent WeTICA simulations in NPT ensemble starting from the solvated folded structure. Details of the WeTICA simulation parameters are provided in Table 7.1.

7.4.2 TC10b mutant of Trp-cage

We took the 208 μs long explicit solvent K8A mutant of the 20-residue Trp-cage mini protein (TC10b) simulation trajectory generated by D.E Shaw Research[84]. The closest experimental structure is PDB ID: 2JOF[85]. The protein was modelled using CHARMM22* force field[86] and solvated with TIP3P[76] water molecules. We calculated the Q values of every frames of this trajectory with respect to the first frame considered as the folded structure. A snapshot with Q ~ 0.2 was taken as a representative of the unfolded state. This representative unfolded conformation was used as the target state conformation for the subsequent WeTICA simulations. Folded and the representative unfolded structures for this Trp-cage mutant are shown in Figure 7.2b.

This folded structure was solvated in a cubic box of ~ 43 Å side length with 2505 TIP3P[76] water molecules, 0.65 mM NaCl concentration and modelled using CHARMM22* force field[86] with Asp & Arg side chains in their charged states. Again a set of 153 $\mathrm{C}_\alpha - \mathrm{C}_\alpha$ atom pair wise distances was used as the feature to calculate TICA eigenvectors trained on the entire 208 μs long Anton folding-unfolding trajectory of this mutant using a lag-time of 10 ns. First two TICA eigenvectors were used as CVs to perform several independent WeTICA simulations in NVT ensemble starting from the solvated folded structure. Details of the WeTICA simulation parameters for this Trp-cage mutant are provided in Table 7.1.

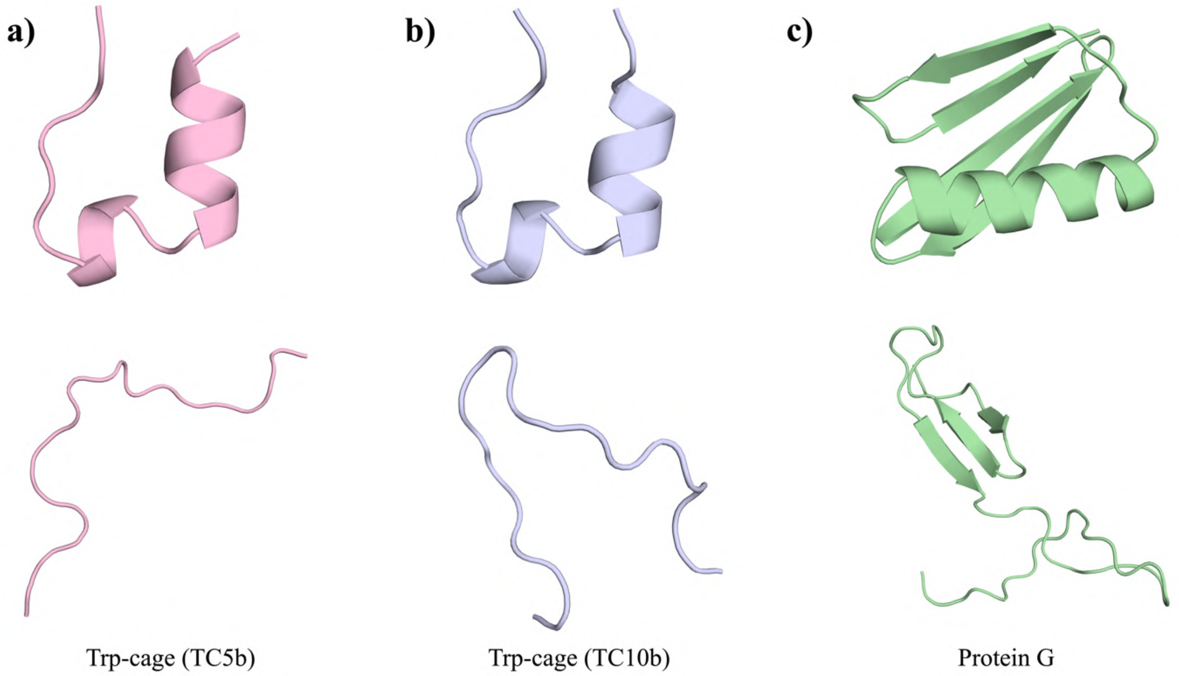


Figure 7.2: Upper row shows the folded structures and the lower row shows the representative unfolded structures in cartoon representation for **a)** TC5b Trp-cage mutant **b)** TC10b K8A Trp-cage mutant and **c)** Protein G, respectively. These unfolded structures were used as the target state conformations in the corresponding WeTICA simulations.

7.4.3 Protein G

We took two trajectory segments of 2 μs long sampled the folded and unfolded states from the 168 μs long explicit solvent N37A/A46D/D77A triple mutant of the redesigned protein G variant NuG2 (closest experimental structure PDB ID:1MI0[87]) simulation trajectory generated by D.E Shaw Research[84]. The protein was modelled using CHARMM22* force field[86] and solvated with TIP3P[76] water molecules. We calculated the Q values of every frames of these two trajectories. A snapshot with Q ~ 0.2 was taken as a representative of the unfolded state. This representative unfolded conformation was used as the target state conformation for the subsequent WeTICA simulations. Folded and the representative unfolded structures are shown in Figure 7.2c.

TICA eigenvectors were calculated using $C_\alpha - C_\alpha$ atom pair wise distances as the feature for a lag-time of 20 ns trained on these two end-state trajectories. The folded structure was modelled using CHARMM22* force field[86] with Asp, Glu and Lys side chains in their charged states and solvated with 100 mM NaCl salt concentration in a cubic box of ~ 57 Å side length with 5438 TIP3P[76] water molecules. The solvated structure was energy minimized and equilibrated at 350K temperature. The first two eigenvectors were used to perform several independent WeTICA simulations in NVT ensemble from the equilibrated solvated folded structure. Details of the WeTICA simulation parameters are provided in Table 7.1.

All WE simulations were performed using a modified version of the Wepy[67] v1.1.0 software. Dynamics was performed using OpenMM v7.5.1[88]. Validation of the chosen WeTICA simulation parameters will be addressed in the next section while discussing results.

Table 7.1: Details of the WeTICA simulation parameters for the three systems.

Parameters	**Trp-cage (TC5b)**	**Trp-cage (TC10b)**	**Protein G**
No. of walkers N_w	24	24	24
Temperature	300K	290K	350K
Feature	$C_\alpha - C_\alpha$ atom pair-wise distances	$C_\alpha - C_\alpha$ atom pair-wise distances	$C_\alpha - C_\alpha$ atom pair-wise distances
Eigenvectors	First two TICA eigenvectors.	First two TICA eigenvectors.	First two TICA eigenvectors.
Non-bonded cut-off	10 Å	9.0 Å	9.5 Å
Integration time step	2 fs	2 fs	2 fs
Resampling interval	20 ps	20 ps	20 ps
Merge dist. d_{merge}	0.50	0.50	0.25
Warp dist. d_{warp}	0.75	0.75	1.40

7.5 Results and discussions

7.5.1 Unfolding kinetics of TC10b Trp-cage mutant

We first calculated the unfolding time (τ_u) of the TC10b Trp-cage mutant using our protocol. Simulated unfolding time of 3 ± 1 μs at 290K temperature was reported from the direct analysis of the 208 μs long unbiased Anton trajectory[84] of this mutant as well as from a high resolution MSM built from the same trajectory[89]. Projections of all the frames of the entire Anton trajectory on the first two TICA eigenvectors (TIC1 and TIC2) used in WeTICA simulations were labelled with the Q values of the corresponding conformations (see Figure 7.A.1). A clear separation between the folded and unfolded state ensembles on this projection plane validates the choice of this projection space to run the WeTICA simulations.

We set $\mathrm{d_{merge}} = 0.50$ and $\mathrm{d_{warp}} = 0.75$ for walker merging and warping, respectively. Notice that $\mathrm{d_{merge}} = 0.50$ and $\mathrm{d_{warp}} = 0.75$ correspond to the locations close to the first minima of the distance distributions between pairs of projection points for the folded and unfolded state basins, respectively (Figure 7.A.2). These cut-off distances were chosen to resemble the approximate size of the folded and unfolded state regions on this TICA space. Thus, this choice of the $\mathrm{d_{warp}}$ ensures the arrival of the walkers in the unfolded state basin. In Figure 7.3a, we have shown the projections of a normal MD trajectory segment and WeTICA unfolding trajectory on the free energy surface (FES) derived using the projection of the full Anton trajectory. We can see that in normal MD the system got stuck in the folded state basin for 500 ns whereas in WeTICA simulation it has reached the unfolded region within ~ 6 ns. This highlights the sampling efficiency of our algorithm. In Figure 7.3b, we have plotted the Q values of this WeTICA unfolding trajectory with time. In the inset of the Figure 7.3b, we have shown the distribution of the Q values of the warped walker conformations. This Q value distribution has a broad range (0.7 – 0.2) and also shows that all the walkers indeed left the folded state basin and crossed into the unfolded state region when they were warped by the algorithm to calculate unfolding kinetics. Note that due to the broad spread of the unfolded state basin as compared to the folded one (Figure 7.3a), larger values for $\mathrm{d_{warp}}$ can be tested while ensuring proper escape of the walkers form the folded state basin $(\mathrm{Q} \gtrsim 0.80)$ to speed up convergence as discussed in Section 7.3.4. The same folded conformation was used in the fraction of native contact analysis for both normal MD and WeTICA unfolding trajectories.

The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted in Figure 7.3c. Average aggregated unfolding probability was used to subsequently calculate the unfolding time (τ_{u}) of this Trp-cage mutant using Eq. (7.2). The final result is shown in Figure 7.3d. Unfolding time calculation converged to the reported value[84,89] $(3 \pm 1\ \mu\mathrm{s})$ within few "ns" of total simulation time. Computed unfolding time of 3.77 ± 0.15 μs (average and standard error of the data shown in the inset of Figure 7.3d) matches reasonably well with the previously reported simulated unfolding time. Therefore, our methodology not only enhances the sampling rate compared to unbiased MD, it also successfully reproduces the unfolding time with significantly less computational cost.

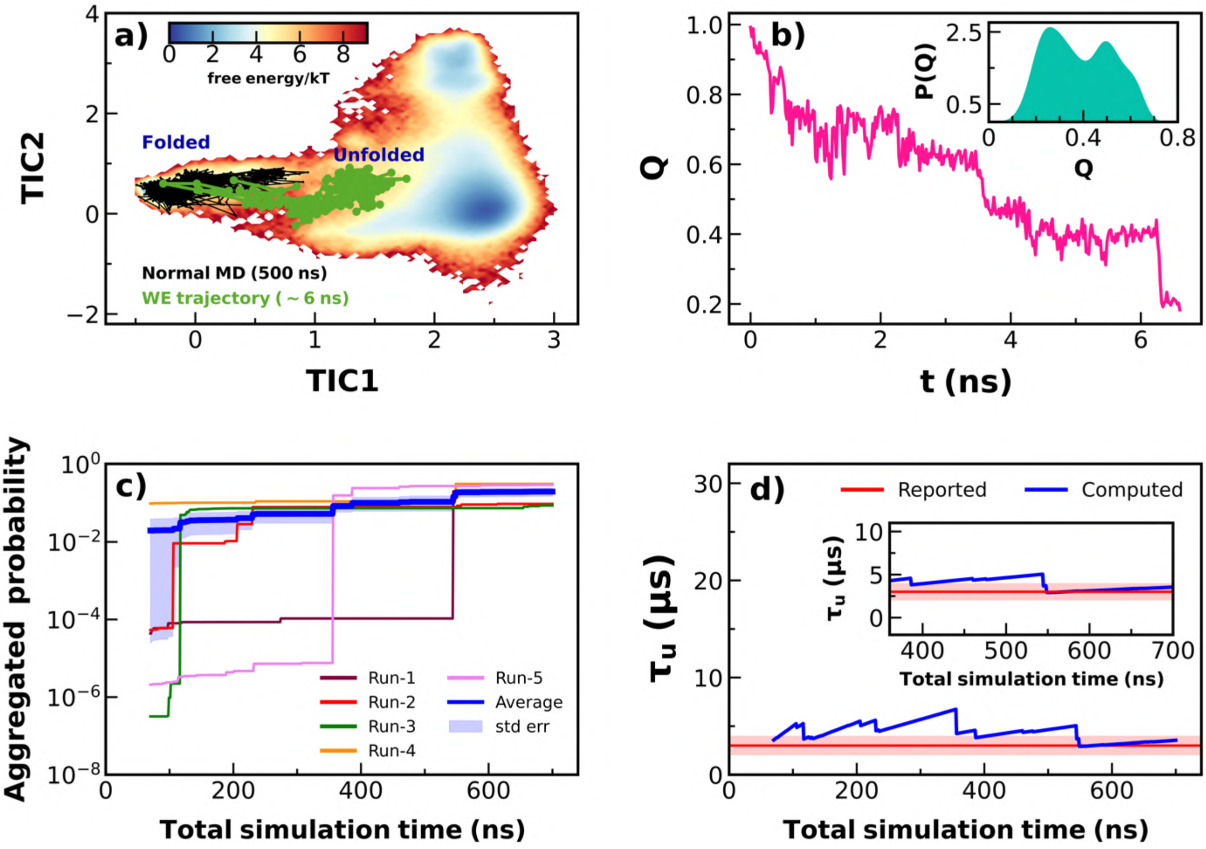


Figure 7.3: **a)** Projections of the normal MD trajectory segment and WeTICA unfolding trajectory on the first two TICA eigenvectors (TIC1 and TIC2) are shown on the free energy surface derived from the Anton trajectory. **b)** Fraction of native contacts (*Q*) with time are plotted for this WeTICA unfolding trajectory. Inset shows the distribution of the *Q* values of the warped walker conformations. **c)** The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted. Blue shaded region represents the standard error of the mean (std err). **d)** Computed unfolding time (τ_u) is plotted as a function of the total simulation time. Reported value (red line) represents the previously simulated unfolding time (3 ± 1 μs) for the TC10b mutant of Trp-cage protein. Inset shows a close view of the converged region. Red shaded regions represent the standard error of the reported value.

7.5.2 Unfolding kinetics of TC5b Trp-cage mutant

We next calculated the unfolding time (τ_u) of the Trp-cage TC5b variant using our methodology. The experimental unfolding time of this mutant is 12.7 μs at 296K temperature[90]. Projections of all the frames of the 1.85 μs long normal MD trajectory on the first two TICA eigenvectors (TIC1 and TIC2) used in WeTICA simulations were labelled with the Q values of the corresponding conformations (Figure 7.A.3). A clear separation between the folded and unfolded state basins for this system validates the use of this projection space for the WeTICA simulations.

We choose $d_{merge} = 0.50$ and $d_{warp} = 0.75$ for walker merging and warping, respectively. Note that $d_{merge} = 0.50$ and $d_{warp} = 0.75$ again correspond to the locations close to the positions of the first minima of the distance distributions between pairs of projection points for the folded and unfolded state basins, respectively (Figure 7.A.4). In Figure 7.4a, we have shown the comparison between a segment of normal MD trajectory and WeTICA unfolding trajectory on the free energy surface (FES) derived using the full normal MD trajectory. We can see that in normal MD the system got stuck in the folded basin for 500 ns whereas in WeTICA simulation the protein unfolds within ~ 20 ns. In Figure 7.4b, we have plotted the Q values of this WeTICA unfolding trajectory with time. In the inset of the Figure 7.4b, we have shown the distribution of the Q values of the warped walker conformations. This distribution of the Q values has a broad range (0.7 – 0.2) and also shows that all the walkers indeed left the folded state basin ($Q \gtrsim 0.80$) and arrived at the unfolded state basin when they were warped by the algorithm. As discussed in section 7.3.4, larger values for d_{warp} can be chosen in this case also.

The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted in Figure 7.4c. Average aggregated unfolding probability was used to subsequently calculate the unfolding time (τ_u) of this Trp-cage mutant. The final result is shown in Figure 7.4d. The calculated unfolding time converged to the experimental value[90] (12.7 μs) within at least one order of magnitude less cumulative WE simulation time than the unfolding time scale. Computed unfolding time of 12.87 $\pm$ 0.09 μs (average and standard error of the data shown in the inset of Figure 7.4d) matches reasonably well with the experimental unfolding time.

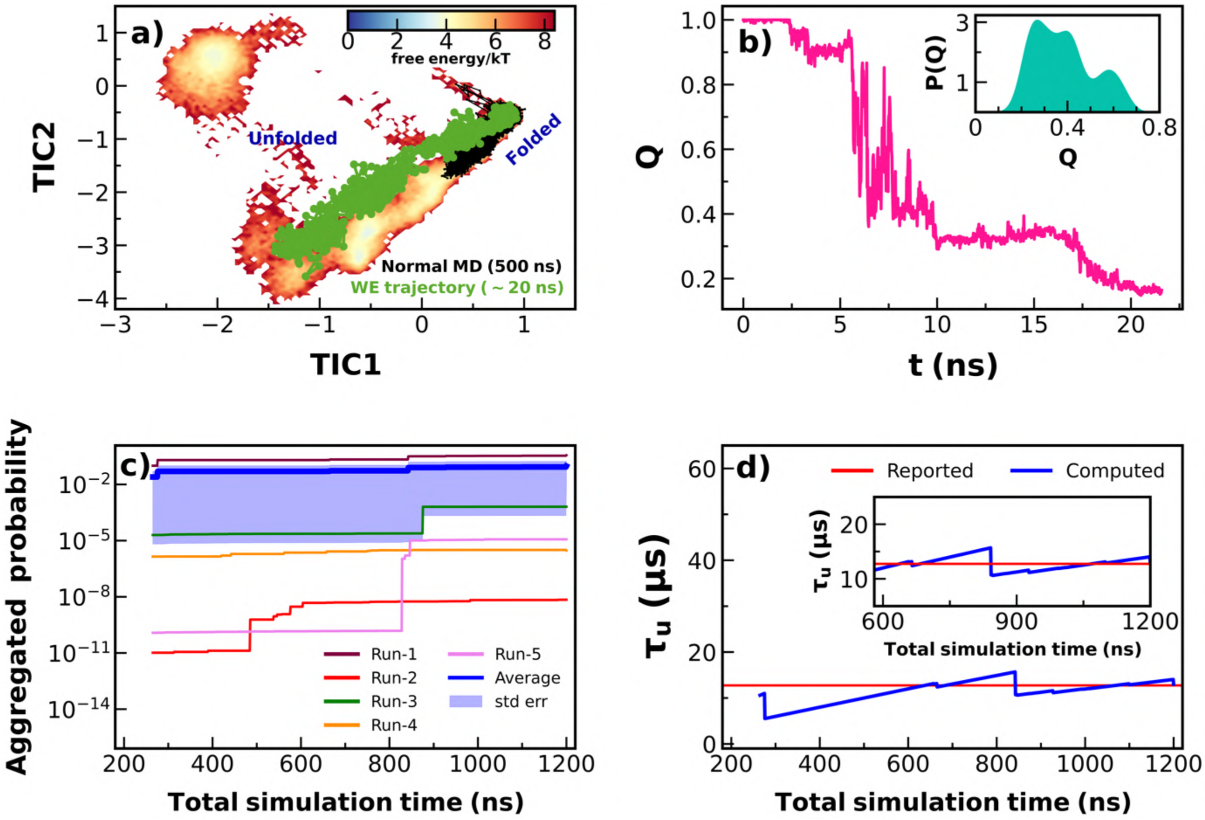


Figure 7.4: a) Projections of the normal MD trajectory segment and WeTICA unfolding trajectory on the first two TICA eigenvectors (TIC1 and TIC2) are shown on the free energy surface derived from the 1.85 μs long normal MD trajectory. **b)** Fraction of native contacts (*Q*) with time are plotted for this WeTICA unfolding trajectory. Inset shows the distribution of the *Q* values of the warped walker conformations. **c)** The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted. Blue shaded region represents standard error of the mean (std err). **d)** Computed unfolding time (τ_u) is plotted as a function of the total simulation time. Reported value (red line) represents the experimental unfolding time (12.7 μs) for the TC5b mutant of Trp-cage protein. Inset shows a close view of the converged region.

For both the mutants of the Trp-cage mini protein, WeTICA simulations produced reasonably accurate unfolding times with only few “ns” of simulations. However, we calculated the TICA eigenvectors (CVs) for these two systems using long enough unbiased trajectories where several folding-unfolding events have been observed. Thus, by construction, the TICA eigenvectors are embedded with all the necessary information related to the slow folding-unfolding processes. But in most of the practical scenarios, the appropriate CVs that can capture the transition of interest are not known *a priori* due to the time scale limitation of unbiased MD simulation. On the other hand we can always generate TICA eigenvectors trained on two short end-state trajectories with perhaps no hopping between the two states. Now the question naturally arises: will our method be able to calculate rare event kinetics using CVs derived from such two end-state simulation trajectories? We will discuss this scenario using the example of the unfolding of Protein G in the next section 7.5.3.

7.5.3 Unfolding kinetics of Protein G

Simulated unfolding time of 37 ± 10 μs at 350K temperature was reported from the direct analysis of several μs long Anton trajectories of redesigned Protein G[84] variant NuG2. The projections of the two Anton simulation trajectory segments belong to the folded and unfolded states on the first two TICA eigenvectors as used in WeTICA simulations are shown in Figure 7.5a. Each projection point is labelled according to the Q value of the corresponding conformation. Note that this time the folded and unfolded state regions are disconnected on this TICA projection plane as compared to the previous two examples. We set $\mathrm{d_{merge}} = 0.25$, which corresponds to the position of the minimum of the distance distribution between pairs of projection points for the folded state basin (Figure 7.A.6). As we have discussed in section 7.3.4 and assuming that only 2 μs unbiased trajectory may not be sufficient to sample the whole unfolded state of large proteins like Protein G, we cannot confidently calculate the approximate size of the unfolded state region for large systems using short trajectories. Thus, we decided to check different values for the $\mathrm{d_{warp}}$ parameter while ensuring proper escape of the walkers form the folded state basin $(\mathrm{Q} \gtrsim 0.80)$. This is accomplished by monitoring the Q values of three productive WeTICA unfolding trajectories generated using three different values for $\mathrm{d_{warp}} = 1.0, 1.2$ and 1.4 with time. The results are shown in Figure 7.5b. We can see that in all three cases, the Q values of the walkers are within 0.7-0.6 when they were warped by the algorithm. This ensures that all the walkers have indeed left the folded state basin sufficiently before getting warped by the algorithm. Thus, we decided to take the maximum value, that is, $\mathrm{d_{warp}} = 1.4$ for all the independent WeTICA simulations of Protein G to speed up convergence and accurate calculation of unfolding MFPT.

The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted in Figure 7.5c. This average aggregated unfolding probability was used to subsequently calculate the unfolding time (τ_u) of Protein G. The final result is shown in Figure 7.5d. The calculated unfolding time converged to the previously reported value[84] $(37 \pm 10$ μs) within more than one order of magnitude less cumulative WE simulation time than the unfolding time scale. Computed unfolding time of 47.35 ± 0.39 μs (average and standard error of the data shown in the inset of Figure 7.5d) also matches reasonably well with the reported value.

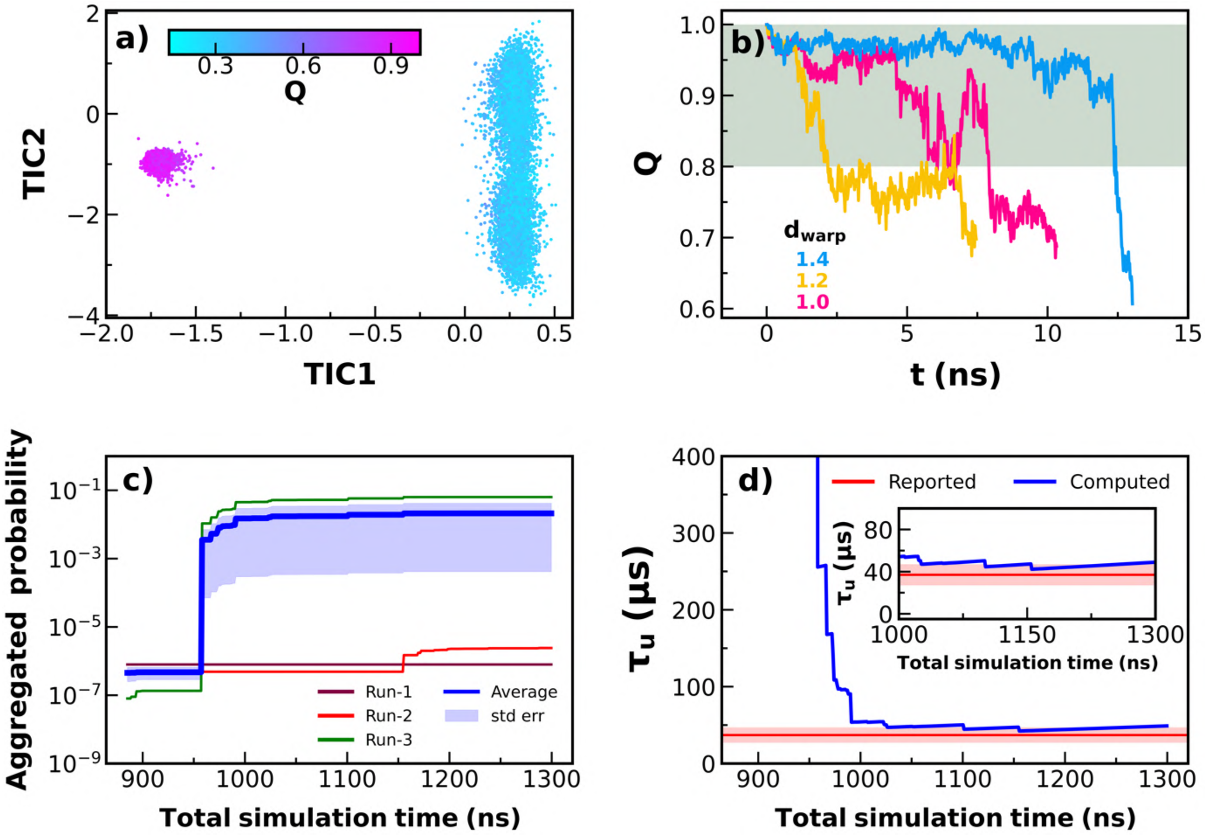


Figure 7.5: a) Frames from the two unbiased Anton trajectory segments (2 μs each) belong to the folded and unfolded states after projecting on the first two TICA eigenvectors (TIC1 and TIC2) are shown on a scatter plot. Each point is coloured according to the fraction of native contact (Q) value of the corresponding conformation. **b)** Fraction of native contacts (Q) of three productive WeTICA unfolding trajectories generated using three different values for d_{warp} are plotted with time. The shaded region represents the folded basin ($Q \gtrsim 0.80$). **c)** The cumulative sum of the unfolding probabilities from each independent WeTICA simulations as well as the average result as a function of the total simulation time are plotted. Blue shaded region represents standard error of the mean (std err). **d)** Computed unfolding time (τ_u) is plotted against the total simulation time. Reported value (red line) represents the previously simulated unfolding time (37 ± 10 μs). Inset shows a close view of the converged region. Red shaded regions represent the standard error (std err) of the reported value.

From all the above mentioned results for the three systems, we can conclude that our methodology significantly enhances the sampling and capable of computing kinetic rate constants with reasonable accuracy with or without a priori knowledge of correct CVs that can capture unfolding. The computed unfolding times for the three proteins are provided in Table 7.2 and several information, for example, total number of unfolding events and total number of WE cycles related to the independent WeTICA simulations for the three systems are provided in Table 7.A.2.

Table 7.2: Computed unfolding times (τ_u) with standard errors for the three proteins are provided and compared with the previously reported simulated (Sim) and experimental (Expt) values.

Systems	**τ_u (μs) produced in this work**	**τ_u (μs) reported in literature**[84,90]
TC10b Trp-cage mutant	3.77 ± 0.15	3 ± 1 (Sim)
TC5b Trp-cage mutant	12.87 ± 0.09	12.7 (Expt)
Protein G	47.35 ± 0.39	37 ± 10 (Sim)

7.6 Conclusions

In this work, we have developed a new “binless” WE simulation method named WeTICA utilizing the fundamental ideas from the REVO WE algorithm. Our proposed protocol uses fixed predefined linear CV space to drive the WE simulations towards the specified target state. We have demonstrated the performance of this new algorithm by recovering the unfolding kinetics of three proteins: 1) TC5b Trp-cage mutant, 2) TC10b Trp-cage mutant and 3) Protein G using Time-lagged Independent Component Analysis (TICA) eigenvectors as our predefined CVs.

For the two Trp-cage mutants, first we calculated TICA eigenvectors using long trajectories where several folding-unfolding events occurred. The first two TICA eigenvectors were used as CVs to guide the WeTICA simulations toward the target unfolded state. The calculated unfolding times converged to the reported unfolding times (in µs order) within feasible cumulative simulation time (within few “ns”) for both the mutants.

In the third case of Protein G, we generated TICA eigenvectors trained on two end-state trajectories where the system was trapped inside the folded and unfolded state basins with no hopping between the two states. Our protocol again successfully reproduced the kinetics of Protein G unfolding using TICA eigenvectors computed from these two end state trajectories within practical simulation time.

Thus, our proposed protocol is working with or without *a priori* knowledge of the CVs that can capture the transitions of interest (in this case, unfolding). Although we used TICA to construct our CV space, generality of our algorithm allow the use of any linear projection methods e.g recently developed harmonic linear discriminant analysis (HLDA), which was proven as good CV for studying transitions between two metastable states[91,92]. The performance of our method can be tested with this kind of newly developed linear CVs in future. Moreover, this new way of walker selection for resampling can also be used on more sophisticated nonlinear CV space e.g variational autoencoder (VAE) latent space[93,94] for further improvements of binless WE methods. We believe that the generality and efficiency of the WE algorithm presented here will be helpful to study kinetics of a diverse class of biologically relevant problems. All the data required to reproduce our results and the python codes to run WeTICA are available at https://github.com/TeamSuman/WeTICA .

Appendix 7.A

Table 7.A.1: Key comparisons between the original REVO and our algorithm

Original REVO algorithm	**Our Algorithm**
1) The current implementation is only limited to the study of ligand binding/unbinding and cannot be directly used for other diverse classes of problems like protein folding-unfolding or transition between different metastable states of biomolecules.	**1)** The goal is to generalize the scope of REVO-based binless methods to a diverse class of problems, not limited to ligand binding/unbinding.
2) Cloning and merging of walkers are decided based on ligand RMSD between a pair of walkers d_{ij}.	**2)** Cloning and merging are decided based on projections of the walkers on a lower dimensional CV space, where we calculate the Euclidean distance d_{ij} between those projections of walkers and also their distance from the target conformation $d^{i}_{\text{from target}}$.
3) Variation *V* is defined as $$V = \sum_i V_i = \sum_i \sum_j \left(\frac{d_{ij}}{d_0}\right)^{\alpha} \phi_i \phi_j$$	**3)** Variation *V* is defined as $$V = \sum_i V_i = \sum_i \frac{1}{d^{i}_{\text{from target}}}$$
4) Clone that walker which is farthest from all other walkers.	**4)** Clone that walker which is closest to the target state conformation.

Table 7.A.2: Information related to all the independent WeTICA simulations.

System	Run	Number of warping events	Total number of cycles	Total simulation time (ns)
Trp-cage (TC10b)	1	311	1461	701.28
	2	264	1460	700.80
	3	113	1465	703.20
	4	159	1466	703.40
	5	127	1470	705.60
Trp-cage (TC5b)	1	36	2502	1200.96
	2	48	2505	1202.40
	3	87	2515	1207.20
	4	48	2501	1200.48
	5	53	2498	1199.04
Protein G	1	113	2800	1344.00
	2	90	2805	1346.40
	3	126	2810	1348.80

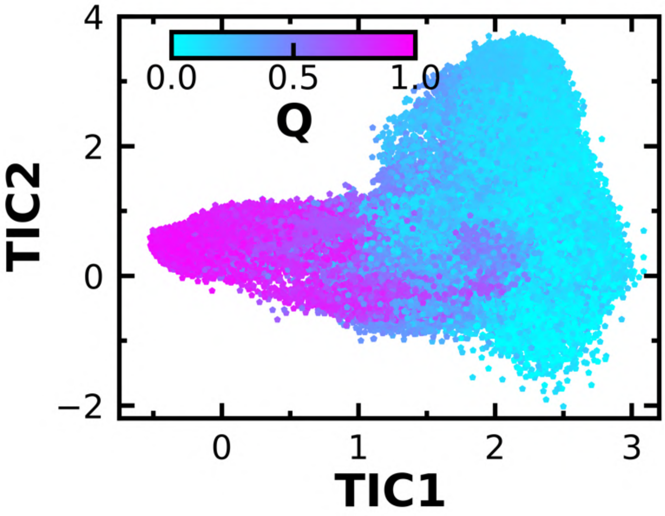


Figure 7.A.1: Frames of the Anton trajectory of the TC10b Trp-cage mutant after projecting on the first two TICA eigenvectors (TIC1 and TIC2) are shown as a scatter plot. Each point is coloured according to the fraction of native contact (Q) value of the corresponding conformation.

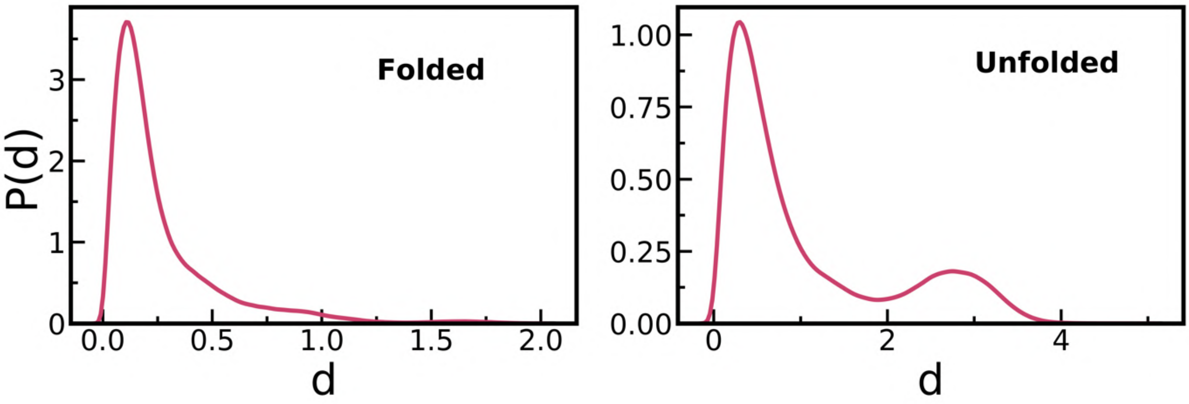


Figure 7.A.2: Distribution of pairwise distances between the projection points of the Anton trajectory frames on the TIC1 vs TIC2 plane in the folded and unfolded state regions for TC10b Trp-cage mutant.

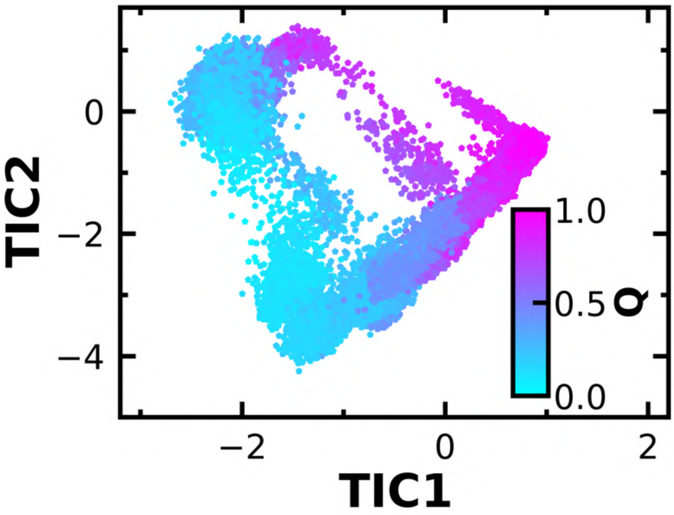


Figure 7.A.3: Frames of the normal MD trajectory of the TC5b Trp-cage mutant after projecting on the first two TICA eigenvectors (TIC1 and TIC2) are shown as a scatter plot. Each point is coloured according to the fraction of native contact (Q) value of the corresponding conformation.

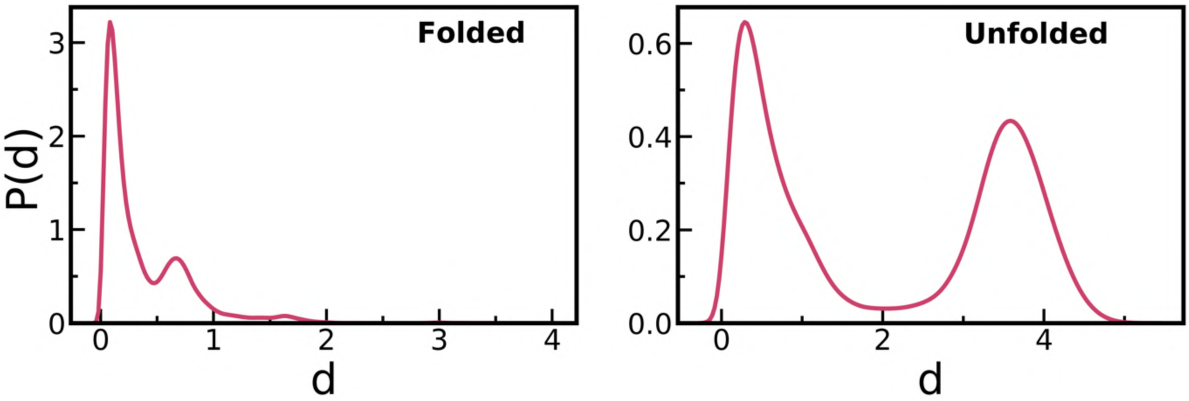


Figure 7.A.4: Distribution of pairwise distances between the projection points of the 1.85 µs long normal MD simulation trajectory frames on the TIC1 vs TIC2 plane in the folded and unfolded state basins for TC5b Trp-cage mutant.

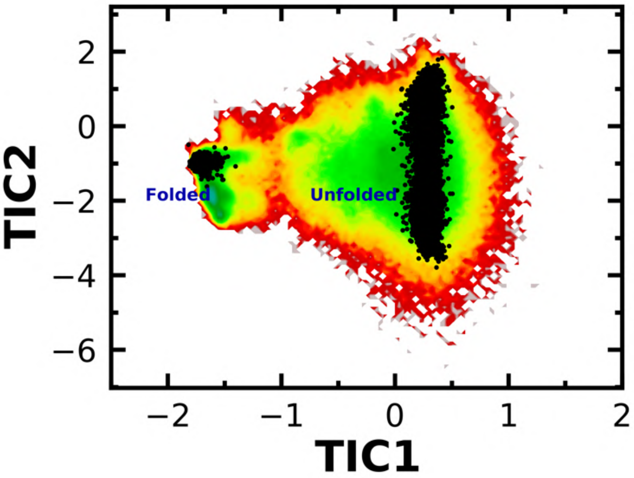


Figure 7.A.5: The projections (in black colour points) of the folded and unfolded state trajectories on the TICA eigenvectors trained using only the end-state simulation trajectories on the free energy surface derived using TICA eigenvectors trained on the full 168 μs long Anton trajectory of Protein G. This shows that the TICA eigenvectors calculated from the end-state simulations capture the same slow process as the full length trajectory TICA model.

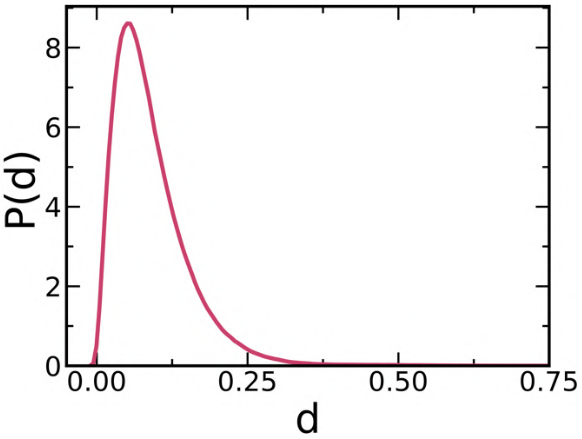


Figure 7.A.6: Distribution of pair-wise distances between the projection points of the 2μs long Anton trajectory frames on the TIC1 vs TIC2 plane in the folded state basin for Protein G.

References:

1 H.-J. Woo and B. Roux, Proceedings of the National Academy of Sciences, 2005, 102, 6825–6830.

2 L. Duan, X. Liu and J. Z. H. Zhang, J Am Chem Soc, 2016, 138, 5722–5728.

3 S. Iida, H. Nakamura and J. Higo, Biochemical Journal, 2016, 473, 1651–1662.

4 F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schoneberg and others, .

5 S. Iida and T. Kameda, J Chem Inf Model, 2023, 63, 3369–3376.

6 P. Tiwary, V. Limongelli, M. Salvalaglio and M. Parrinello, Proc Natl Acad Sci U S A, 2015, 112, E386–E391.

7 A. Barducci, M. Bonomi and M. Parrinello, Wiley Interdiscip Rev Comput Mol Sci, 2011, 1, 826–843.

8 A. Laio, A. Rodriguez-Fortea, F. L. Gervasio, M. Ceccarelli and M. Parrinello, J Phys Chem B, 2005, 109, 6714–6721.

9 G. Bussi, A. Laio and M. Parrinello, Phys Rev Lett, 2006, 96, 90601.

10 A. Barducci, G. Bussi and M. Parrinello, Phys Rev Lett, 2008, 100, 20603.

11 A. Laio and F. L. Gervasio, Reports on Progress in Physics, 2008, 71, 126601.

12 P. Tiwary and M. Parrinello, Phys Rev Lett, 2013, 111, 230602.

13 M. Invernizzi and M. Parrinello, Journal of Physical Chemistry Letters, 2020, 11, 2731–2736.

14 E. Darve, D. Rodriguez-Gómez and A. Pohorille, J Chem Phys.

15 J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille and C. Chipot, J Phys Chem B, 2015, 119, 1129–1151.

16 E. Darve and A. Pohorille, J Chem Phys, 2001, 115, 9169–9183.

17 J. Wang, P. R. Arantes, A. Bhattarai, R. V Hsu, S. Pawnikar, Y. M. Huang, G. Palermo and Y. Miao, Wiley Interdiscip Rev Comput Mol Sci, 2021, 11, e1521.

18 Y. Miao, V. A. Feher and J. A. McCammon, J Chem Theory Comput, 2015, 11, 3584–3595.

19 Y. Sugita and Y. Okamoto, Chem Phys Lett, 1999, 314, 141–151.

20 W. Zhang, C. Wu and Y. Duan, J Chem Phys.

21 D. Sindhikara, Y. Meng and A. E. Roitberg, J Chem Phys.

22 E. Rosta and G. Hummer, J Chem Phys.

23 L. Wang, R. A. Friesner and B. J. Berne, Journal of Physical Chemistry B, 2011, 115, 9431–9438.

24 D. B. Kokh, M. Amaral, J. Bomke, U. Grädler, D. Musil, H.-P. Buchstaller, M. K. Dreyer, M. Frech, M. Lowinski, F. Vallee and others, J Chem Theory Comput, 2018, 14, 3859–3869.

25 D. B. Kokh, B. Doser, S. Richter, F. Ormersbach, X. Cheng and R. C. Wade, J Chem Phys.

26 M. M. Sultan, H. K. Wayment-Steele and V. S. Pande, J. Chem. Theory Comput., 2018, 14, 1887–1894.

27 J. M. L. Ribeiro, P. Bravo, Y. Wang and P. Tiwary, Journal of Chemical Physics, 2018, 149, 72301.

28 W. Chen and A. L. Ferguson, J Comput Chem, 2018, 39, 2079–2102.

29 S. Mehdi, Z. Smith, L. Herron, Z. Zou and P. Tiwary, Annu Rev Phys Chem, 2024, 75, 347–370.

30 R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni and M. Parrinello, J Am Chem Soc, 2017, 139, 4780–4788.

31 Y. Wang, J. M. L. Ribeiro and P. Tiwary, Nature Communications 2019 10:1, 2019, 10, 1–8.

32 R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. USA, 2005, 102, 6732–6737.

33 Y. Wang, O. Valsson, P. Tiwary, M. Parrinello and K. Lindorff-Larsen, Journal of Chemical Physics, 2018, 149, 72309.

34 D. Ray, N. Ansari, V. Rizzi, M. Invernizzi and M. Parrinello, J Chem Theory Comput, 2022, 18, 6500–6509.

35 D. Ray and M. Parrinello, J Chem Theory Comput, 2023, 19, 5649–5670.

36 J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte and F. Noé, J Chem Phys.

37 B. E. Husic and V. S. Pande, J Am Chem Soc, 2018, 140, 2386–2396.

38 V. S. Pande, K. Beauchamp and G. R. Bowman, Methods, 2010, 52, 99–105.

39 J.-H. Prinz, B. Keller and F. Noé, Physical Chemistry Chemical Physics, 2011, 13, 16912–16927.

40 G. A. Huber and S. Kim, Biophys J, 1996, 70, 97–110.

41 D. Bhatt, B. W. Zhang and D. M. Zuckerman, J Chem Phys.

42 B. Abdul-Wahid, H. Feng, D. Rajan, R. Costaouec, E. Darve, D. Thain and J. A. Izaguirre, J Chem Inf Model, 2014, 54, 3033–3043.

43 J. Copperman and D. M. Zuckerman, J Chem Theory Comput, 2020, 16, 6763–6775.

44 E. Hellemann and J. D. Durrant, J Chem Theory Comput, 2023, 19, 5677–5689.

45 J. L. Adelman and M. Grabe, J Chem Theory Comput, 2015, 11, 1907–1918.

46 U. Adhikari, B. Mostofian, J. Copperman, S. R. Subramanian, A. A. Petersen and D. M. Zuckerman, J Am Chem Soc, 2019, 141, 6519–6526.

47 A. S. Saglam and L. T. Chong, Chem Sci, 2019, 10, 2360–2372.

48 S. D. Lotz and A. Dickson, J Am Chem Soc, 2018, 140, 618–628.

49 T. Dixon, S. D. Lotz and A. Dickson, J Comput Aided Mol Des, 2018, 32, 1001–1012.

50 T. Sztain, S.-H. Ahn, A. T. Bogetti, L. Casalino, J. A. Goldsmith, E. Seitz, R. S. McCool, F. L. Kearns, F. Acosta-Reyes, S. Maji and others, Nat Chem, 2021, 13, 963–968.

51 S.-H. Ahn, B. R. Jagger and R. E. Amaro, J Chem Inf Model, 2020, 60, 5340–5352.

52 M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe and others, J Chem Theory Comput, 2015, 11, 800–809.

53 J. D. Russo, S. Zhang, J. M. G. Leung, A. T. Bogetti, J. P. Thompson, A. J. Degrave, P. A. Torrillo, A. J. Pratt, K. F. Wong, J. Xia, J. Copperman, J. L. Adelman, M. C. Zwier, D. N. Lebard, D. M. Zuckerman and L. T. Chong, J Chem Theory Comput, 2022, 18, 638–649.

54 D. M. Zuckerman and L. T. Chong, Annu Rev Biophys, 2017, 46, 43–57.

55 D. Ray and I. Andricioaei, Journal of Chemical Physics.

56 D. Ray, S. E. Stone and I. Andricioaei, J Chem Theory Comput, 2022, 18, 79–95.

57 S.-H. Ahn, A. A. Ojha, R. E. Amaro and J. A. McCammon, J Chem Theory Comput, 2021, 17, 7938–7951.

58 A. A. Ojha, S. Thakur, S. H. Ahn and R. E. Amaro, J Chem Theory Comput, 2023, 19, 1342–1359.

59 B. W. Zhang, D. Jasnow and D. M. Zuckerman, J Chem Phys.

60 A. Dickson and C. L. Brooks III, J Phys Chem B, 2014, 118, 3532–3542.

61 J. L. Adelman and M. Grabe, Journal of Chemical Physics, 2013, 138, 44105.

62 P. A. Torrillo, A. T. Bogetti and L. T. Chong, Journal of Physical Chemistry A, 2021, 125, 1642–1649.

63 D. Aristoff, J. Copperman, G. Simpson, R. J. Webber and D. M. Zuckerman, Journal of Chemical Physics.

64 N. Donyapour, N. M. Roussey and A. Dickson, J Chem Phys.

65 N. M. Roussey and A. Dickson, J Comput Chem, 2023, 44, 935–947.

66 S. Bose, S. D. Lotz, I. Deb, M. Shuck, K. S. S. Lee and A. Dickson, J Am Chem Soc, 2023, 145, 25318–25331.

67 S. D. Lotz and A. Dickson, ACS Omega, 2020, 5, 31608–31623.

68 S. Schultze and H. Grubmüller, J Chem Theory Comput, 2021, 17, 5766–5776.

69 Y. Naritomi and S. Fuchigami, J Chem Phys.

70 G. Pérez-Hernández and F. Noé, J Chem Theory Comput, 2016, 12, 6118–6129.

71 G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis and F. Noé, J Chem Phys.

72 C. R. Schwantes, D. Shukla and V. S. Pande, Biophys J, 2016, 110, 1716–1719.

73 M. M. Sultan and V. S. Pande, J Chem Theory Comput, 2017, 13, 2440–2447.

74 J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, Nature Structural Biology 2002 9:6, 2002, 9, 425–430.

75 J. Huang and A. D. Mackerell, J Comput Chem, 2013, 34, 2135–2145.

76 P. Mark and L. Nilsson, Journal of Physical Chemistry A, 2001, 105, 9954–9960.

77 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, SoftwareX, 2015, 1, 19–25.

78 G. Bussi, D. Donadio and M. Parrinello, J Chem Phys, 2007, 126, 14101.

79 M. Parrinello and A. Rahman, J Appl Phys, 1981, 52, 7182–7190.

80 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, J Chem Phys, 1995, 103, 8577–8593.

81 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, J Comput Chem, 1997, 18, 1463–1472.

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

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

84 K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, Science (1979), 2011, 334, 517–520.

85 B. Barua, J. C. Lin, V. D. Williams, P. Kummler, J. W. Neidigh and N. H. Andersen, Protein Engineering, Design \& Selection, 2008, 21, 171–185.

86 S. Piana, K. Lindorff-Larsen and D. E. Shaw, Biophys J.

87 S. Nauli, B. Kuhlman, I. Le Trong, R. E. Stenkamp, D. Teller and D. Baker, Protein Science, 2002, 11, 2924–2931.

88 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, PLoS Comput Biol, 2017, 13, e1005659.

89 H. Sidky, W. Chen and A. L. Ferguson, J Phys Chem B, 2019, 123, 7999–8009.

90 L. Qiu, S. A. Pabit, A. E. Roitberg and S. J. Hagen, J Am Chem Soc, 2002, 124, 12952–12953.

91 D. Mendels, G. Piccini and M. Parrinello, Journal of Physical Chemistry Letters, 2018, 9, 2776–2781.

92 D. Mendels, G. Piccini, Z. F. Brotzakis, Y. I. Yang and M. Parrinello, Journal of Chemical Physics, 2018, 149, 194113.

93 S. Mansoor, M. Baek, H. Park, G. R. Lee and D. Baker, J Chem Theory Comput, 2024, 20, 2689–2695.

94 Z. Belkacemi, M. Bianciotto, H. Minoux, T. Lelièvre, G. Stoltz and P. Gkeka, Journal of Chemical Physics, 2023, 159, 24122.

8

Concluding remarks and future problems

8.1 Concluding remarks

In this thesis, we have investigated how the intermolecular interactions govern the dynamics and functionality of chemical and biological systems using molecular dynamics (MD) simulations complemented by the theories of physical chemistry, statistical physics and machine learning approaches. The chapter-wise conclusions are discussed at the end of the respective chapters. Therefore, we opt not to re-iterate them here. Instead, in this chapter we have discussed the broad conclusions about our works and highlighted some intriguing problems which can be explored in near future.

Our studies have shown how ion-ion, ion-solvent and solvent-solvent interactions affect ion transport in battery electrolyte solutions. We have used both the Onsager' transport theory and Van Hove function in understanding various ion-ion dynamical correlations and investigated how these correlations shape the complex transport phenomena in concentrated battery electrolyte systems. Furthermore, we have also investigated how the interactions between a co-solvent and water molecules in aqueous zinc-ion battery electrolyte solutions dictate the transport of zinc cations. We have shown that how the interactions between co-solvent and water molecules alter the structure and dynamics of waters around zinc cations. We have found that this altered structure and dynamics of water molecules around zinc ions are strongly correlated with the ionic conductivity of the solutions. Our results and analysis procedures will help in developing molecular level investigations of ion transport for designing next generation battery electrolytes.

Next, we have also investigated the interactions between several dye molecules, which are used in textile industry, with an enzyme known as Laccase and shown the ability of this enzyme to degrade a variety of dye molecules irrespective of their different shape and charge. Our studies have provided microscopic explanations behind laccase's substrate promiscuous nature and shown that the dye molecule binding to the enzyme possibly happened via conformational selection mechanism. Moreover, we have also shown in our studies that how different machine learning (ML) methods can be used to extract useful information from the simulation data of complex biomolecules. Furthermore, we have also developed an enhanced sampling algorithm to study kinetics of different biological processes. Our method has proven to be successful in recovering the unfolding kinetics of various small proteins and has the potential to be used in studying a diverse class of biological processes.

Now, we will discuss some of the possible future areas of research based on this thesis work.

8.2 Future problems

8.2.1 Development of machine learning potentials (MLPs)

In Chapter 3 & 4, we have seen that the conventional classical force fields have failed to correctly model the intermolecular interactions at high salt concentration used in battery electrolytes. As a result, we have observed that ion transport properties, such as ionic conductivity lacks quantitative accuracy at high salt concentration. Machine learning potentials (MLPs) have revolutionised the field of atomistic simulations in recent years[1–3]. MLPs have enable the simulation of variety of systems with long timescales processes at *ab initio* level accuracy[3]. Use of MLPs are greatly enhancing the accuracy and efficiency of molecular modelling. Therefore, development of MLPs to study ion transport in concentrated battery electrolyte solutions is a promising area of future research.

8.2.2 Understanding ion transport near electrolyte-electrode interfaces

In this thesis, we have investigated complex ion transport phenomena in bulk battery electrolyte solutions. However, several studies have shown that ion-ion interactions and ion transport properties are significantly different in confined electrolytes and/or in the presence of electrochemical surfaces[4–6]. Moreover, several classical MD simulation studies have also shown the prevalence of ion-ion interactions under confinement. Therefore, investigating complex ion transport phenomena under confinement using classical MD simulations using MLPs can be explored in future works for making better electrolyte engineering strategies.

8.2.3 Development of binless weighted ensemble path sampling methods

In Chapter 7, we have discussed not only the importance of developing binless weighted ensemble (WE) based path sampling methods, we have also developed a binless WE method called 'WeTICA' and showed the success of this method in studying the unfolding kinetics of several small test protein systems. Our method can be tested to study more complex biological processes in future. Moreover, one can also work in developing more advanced and sophisticated binless WE based path sampling methods in future.

References:

1 F. L. Thiemann, N. O'Neill, V. Kapil, A. Michaelides and C. Schran, Journal of Physics: Condensed Matter, 2024, 37, 073002.

2 J. Behler, Journal of Chemical Physics, 2016, 145, 170901.

3 N. Yao, X. Chen, Z. H. Fu and Q. Zhang, Chem Rev, 2022, 122, 10970–11021.

4 M. C. F. Wander and K. L. Shuford, Journal of Physical Chemistry C, 2010, 114, 20539–20546.

5 D. Nicholson and N. Quirke, DOI:10.1080/0892702031000078427.

6 X. Chen and X. Kong, Nano Lett, 2023, 23, 5194–5200.