SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Metodi e Modelli per le Scienze Molecolari

XXXV ciclo

Computational Methods and Models for Atomistic Simulations of Ion Hydration, Ion-Ligand Complexes, and Ion Transport in Channels

Settore Scientifico Disciplinare **CHIM/02**

Candidato
Luca Sagresti

Relatori

Prof. Giuseppe Brancato
Prof. Kenneth M. Merz Jr.

Anno accademico 2022–2023

Acknowledgments

I wish to thank my supervisor Professor Giuseppe Brancato of the Scuola Normale Superiore for guiding and assisting me, my co-supervisor Professor Kenneth Merz of Michigan State University for his valuable suggestions and the productive discussions I had with him and his group. I would like to thank Professor Giovanni Bottari of Universidad Autonoma de Madrid for imparting me significant knowledge about the experimental aspects of chemistry and for the fruitful collaborations that we have developed over the years. I am very grateful to Prof. Armagan Kocer University of Twente for introducing me to the fascinating world of ion channels and electrophysiology, and for the enriching idea exchanges with his group. I wish to extend my thanks to Dr. Walter Rocchia and Dr. Sergio Rampino, who, despite the short duration of my collaboration with them, have nonetheless given valuable insights to me.
I would also like to thank all the members, past and present, of my research group at Scuola Normale Superiore for their teachings and support. I deeply thank all my fellow PhD students that cheered me along the way.
An heartfelt thanks to all those who have been by my side and supported me throughout this lengthy journey, I would not have wanted anything different.

Abstract

Ions are omnipresent in Nature and play significant roles in a large number of biological and nanotechnological applications. For example, the investigation of their coordination with ligands is a recurrent theme in the scientific literature, with special reference to catalytic activity, nucleic acid cleavage, and anticancer drug studies. Besides, in the context of biological channels, ions may generate various kinds of ionic currents crucial for human physiological activities, like movement and heartbeats. Over the past seventy years, numerous experimental and theoretical techniques have been developed to address several properties of ions in aqueous solutions and in interaction with proteins, such as ion coordination, hydration free energy, ligand exchange times, and ionic currents in biochannels. In this regard, molecular dynamics (MD) simulations have proved very fruitful in providing a deep atomistic understanding of both complex and subtle phenomena involving ions. In this dissertation, novel computational approaches and applications are presented aiming at a better comprehension of different aspects of ion microsolvation, ion-ligand complex formation, and ion transport into protein channels, thus extending the range of available *in silico* techniques in this research area. The dissertation is structured into three parts. The first part of this thesis introduces a new computational methodology for analyzing the structural, thermodynamic, and kinetic properties of ion microsolvation, particularly effective in studying aqua-ion complex formation and solvent exchange in the first hydration shell, beyond the reach of standard MD simulations. The second part enhances the accuracy of force fields for ion-carboxylate interactions and subsequently presents a computational procedure for assessing stability constants and ligand exchange rates. This procedure, adaptable to different ions and ligands, shows promise in elucidating ion-ligand exchange mechanisms and predicting dissociation rates up to seconds, thus expanding applications of the method to more complex systems. The third part firstly discusses software developed for analyzing pore morphology and ion translocation pathways, and secondly the use of MD simulations and master-equations for the Kv4.3 potassium channel. In the latter, both techniques combined proved to be useful tools coupled with experiments to disclose the molecular causes of detrimental point mutations of the Kv4.3 potassium channel. Although the application areas of the above studies may appear diverse, each research work contributes consistently to a deeper understanding of the underlying molecular mechanisms characterizing ion solutions and strives to align computational models with experimental conditions, thus pushing the boundaries of the *in silico* research in this domain.

Publications

Presented in PhD thesis

1. L. Sagresti, L. Peri, G. Ceccarelli, G. Brancato. Stochastic Model of Solvent Exchange in the First Coordination Shell of Aqua Ions. *Journal of Chemical Theory and Computation*, 2022, 18(5), 3164-3173.

2. M. Jafari, Z. Li, L. Frank Song, L. Sagresti, G. Brancato, K. M. Merz Jr. Thermodynamics of Metal–Acetate Interactions. *The Journal of Physical Chemistry B*, 2024,128(3), 684-697.

3. L. Sagresti, L. Benedetti, K. M. Merz Jr., G. Brancato. Simulating the chemical equilibria of metal complexes: Insights into the ligand exchange mechanism. **In preparation**, 2024.

4. A. Raffo, L. Gagliardi, U. Fugacci, L. Sagresti, S. Grandinetti, G. Brancato, S. Biasotti, W. Rocchia. Chanalyzer: A Computational Geometry Approach for the Analysis of Protein Channel Shape and Dynamics. *Frontiers in Molecular Biosciences*, 2022, 9.

5. E. de Jong, A. Catte, L. Sagresti, N. Bhattacharjee, C. Tiecher, G. Brancato, A. Kocer. Unraveling the molecular origin of an inherited channelopathy in the voltage-gated potassium channel Kv4.3. *Proceedings of the National Academy of Sciences*, 2024 **Submitted**.

Further publications during PhD research training

1. L. M. Mateo, L. Sagresti, Y. Luo, D. M. Guldi, T. Torres, G. Brancato, G. Bottari. Expanding the Chemical Space of Tetracyanobuta-1,3-diene (TCBD) through a Cyano-Diels-Alder Reaction: Synthesis, Structure, and Physicochemical Properties of an Anthryl-fused-TCBD Derivative. *Chem. Eur. J.*, 2021, 27, 16049.

2. N. Barbosa, L. Sagresti, G. Brancato. Photoinduced azobenzene-modified DNA dehybridization: insights into local and cooperativity effects from a molecular dynamics study. *Phys. Chem. Chem. Phys.*, 2021, 23, 25170-25179.

3. L. Sagresti, S. Rampino. Charge-Flow Profiles along Curvilinear Paths: A Flexible Scheme for the Analysis of Charge Displacement upon Intermolecular Interactions. *Molecules*, 2021, 26, 6409.

4. A. Ferretti, S. Sinha, L. Sagresti, E. Araya-Hermosilla, M. Prato, V. Mattoli, A. Pucci, G. Brancato. One-step functionalization of mildly and strongly reduced graphene ox-

ide with maleimide: an experimental and theoretical investigation of the Diels–Alder [4+2] cycloaddition reaction. *Phys. Chem. Chem. Phys.*, 2022, 24, 2491-2503.

5. O. F. Vera, L. Sagresti, L. M. Mateo, T. Torres, G. Brancato, G. Bottari. Unprecedented "Off-pathway" [2+2] Cycloaddition-Retro-Electrocyclization Reaction between an Unsymmetric Alkyne and Tetracyanoquinodimethane. **In preparation**, 2024.

Contents

List of Figures

List of Tables

List of Acronyms

ABF	Adaptive Biasing Force
AP	Action Potential
CC	Coupled Cluster
CF	Crystal Field
CG	Coarse Grained
CK	Chapman-Kolmogorov
CN	Coordination Number
CPU	Central Processing Unit
CV	Collective Variable
DFT	Density Functional Theory
DFTB	Density Functional Based Tight Binding
DO	Drude Oscillator
ECD	Extra-Cellular Domain
FEP	Free Energy Perturbation
FES	Free Energy Surface
FFT	Fast Fourier Transform
FP	Fokker-Planck
FQ	Fluctuating Charge
GPU	Graphical Processing Unit
HFE	Hydration Free Energy
HOF	Heats of Formation
HPC	High Performance Computing
ICD	Intra-Cellular Domain
IDM	Induced Dipole Model
IOD	Ion-Oxygen Distance
IP	Ionization Potential
LD	Langevin Dynamics
LDA	Linear Discriminant Analysis
LGIC	Ligand-Gated Ion Channel
LJ	Lennard-Jones
MAD	Mean Absolute Deviation
MC	Monte Carlo
MD	Molecular Dynamics
MetaD	Metadynamics
MFPT	Mean First Passage Time
MO	Molecular Orbital
MP	Møller-Plesset
MPI	Messaging Passing Interface

μVT	Grand Canonical ensemble
MW	Multiple Walker
NMR	Nuclear Magnetic Resonance
NpT	Isobaric-Isothermal (Gibbs) ensemble
NVE	Microcanonical ensemble
NVT	Canonical ensemble
ODE	Ordinary Differential Equation
PB	Poisson Boltzmann
PCA	Principal Component Analysis
PCCA	Perron-Cluster Cluster Analysis
PDB	Protein Data Bank
PME	Particle Mesh Ewald
QM/MM	Quantum Mechanics/Molecular Mechanics
RDF	Radial Distribution Function
SDE	Stochastic Differential Equation
SF	Selectivity Filter
TI	Thermodynamic Integration
TICA	Time-lagged Independent Component Analysis
TM	Transition Metal
TMD	Trans-Membrane Domain
TPS	Transition Path Sampling
TPT	Transition Path Theory
TST	Transition State Theory
VB	Valence Bond
VC	Voltage Clamp
vdW	van der Waals
VES	Variationally Enhanced Sampling
VGIC	Voltage-Gated Ion Channel
VSD	Voltage-Sensing Domain
WT	Wild-Type
XRD	X-Ray Diffraction

Introduction

Charged entities have been known since at least 600 BC when Thales of Miletus observed the attraction of small objects to amber rubbed with fur. However, it was not until the 19th century that these entities were named ions, derived, coincidentally or as homage, from the Greek word *ienai* meaning "to go".
Ions are seen nowadays as particles, being them atoms or molecules, possessing a net electric charge where the name cation is given to ions with a net positive charge and anion to the ones with a net negative charge. Ions exist in many states of matter and are widely distributed in the natural world. They can exist as individual ions in the gaseous phase or as an electrically neutral combination of cations and anions in the condensed phase.
In this thesis, the focus will be exclusively on ions in aqueous solutions. Therefore, unless explicitly stated otherwise, the term "ions" should be understood as referring monoatomic ions in solutions where water is the solvent. Many branches of chemistry conduct studies on solutions, thus rendering ions, as particles with a net electric charge, significant actors. This significance is due to the strong interactions they can generate, owing to Coulomb's potential, which can crucially influence the outcomes of experiments.
It is evident that the complexity of interactions ions have with various system entities increases with the system's complexity. Modeling, describing, and quantitatively measuring an ion's interaction with a solvent molecule is simpler than understanding the effects caused by an excess or deficiency of ions in a human cell. However, simplicity in a system should not be equated with triviality. Instead, the study of simpler systems can aid scientists in uncovering mechanisms that might be lost in the chaotic complexity of larger systems. The mastery and application of such fundamental molecular insights can be extremely beneficial when transferred to larger and more realistic systems, yielding positive results in technology, the environment, biology, and other fields.
The subsequent sections introduce the topics of ions in aqueous solutions, ion-ligand complexes, and the transport of ions in ionic channels, including some considerations about the experimental measurements and the limitations of typical wet-lab techniques. The concluding part of this introductory chapter will examine the range of theoretical models and computational methods reported in the literature for studying such systems. Also, this chapter clarifies why only a subset of these methods, as elaborated in Chapter 1, were adopted during the PhD research project.

Ions in Aqueous Solutions

Apart from more subtle features related to the specific electronic configuration, a bare ion is characterized by three key attributes: its charge, mass, and size. The charge is typically expressed as $z_I e$, which represents an integer multiple (either positive or negative) of the elementary charge. The mass is conveyed in terms of the molar mass. However, determining the size of an ion poses challenges. Generally, monoatomic ions assume a spherical shape

unless subjected to strong electromagnetic fields, but estimating their size is difficult due to the external electron cloud [1]. In condensed phases, specific sizes can be assigned to monoatomic ions due to the strong repulsion between adjacent electronic shells, which will be the case of the ions studied here. The radius of an ion in water can be estimated using techniques similar to those used for crystals. Through X-ray and neutron diffraction experiments, it is possible to approximate the ion radius, r_I, with an uncertainty of only a few picometers, by measuring ion-water distances and assuming a fixed radius of 138 pm for a water molecule [1]. These three physical attributes form the foundation for modeling an ion at every level, and obtaining experimental values for them is crucial.

When studying different properties of ions in aqueous solutions, it is important to note that only quantities related to ion mobility (such as diffusion and conductivity) can be attributed to an ion as an individual characteristic. Other measurements must consider the fact that they involve the entire electrolyte. This is true for determining thermodynamic quantities like standard molar heat capacities (at constant pressure), entropies, standard molar enthalpy, and Gibbs energy of formation of ions in solution. These studies are typically conducted under specific conditions known as "infinite dilution" for the standard state (at $T^0 = 298.15$ K and $P^0 = 0.1$ MPa), under the reasonable assumption that an ion interacts solely with the surrounding solvent and not with other ions. The most common experimental procedure to reach such conditions involves dissolving a complete electrolyte in the solvent. This electrolyte consists of an equal number of cations and anions, forming a neutral compound. The concept of infinite dilution can be closely approximated through extrapolation to low concentrations. This approach is akin to dissolving a mole of electrolyte in an extremely large volume of solvent, or a minuscule amount of electrolyte in a finite volume of solvent. In such scenarios, the individual ionic contributions to the measured molar properties of the electrolyte at infinite dilution are cumulative. This is because each ion is exclusively surrounded by solvent molecules and is far enough to avoid ion-ion interaction. These individual contributions are then adjusted according to their stoichiometric coefficients, denoted as ν_+ and ν_-, within the electrolyte.

It is crucial to note that the thermodynamic properties associated with the solvation process differ from standard molar quantities. These characteristics relate to the concept of transferring an isolated ion from the gas phase into a solvent [2]. The Gibbs energy of solvation can be defined as the change in the chemical potential of a solute ion I at equilibrium, as shown in the following equation [2]

$$\Delta\mu_I^\star = \Delta\mu_I^{\star L} - \Delta\mu_I^{\star G} = k_{\mathrm{B}} T ln \left(\frac{\rho_I^G}{\rho_I^L}\right)_{eq} \tag{1}$$

Here, the superscript L and G denote the transfer of an ion I from an ideal gas phase (G) to a fixed position in a liquid (L). k_{B} is the Boltzmann constant, and ρ represents its number density. The corresponding molar entropy and enthalpy of solvation can be derived from simple thermodynamic relationships, starting from equation 1. Highlighting these properties, often referred to as hydration thermodynamic quantities in the literature, is vital because their values are typically those compared against theoretical models that consider solvent effects.

One can apply different models to ions in water. Perhaps, the easiest one treats ions as uniformly dispersed conductive particles with a charge, embedded in a continuous and homogeneous medium characterized by a specific dielectric constant. A more sophisticated model could take into account the molecular nature of water as it will be done in Chapters 2, 3 and 4. Consequently, if the focus is on the structure of solvated ions one can think of concentric solvation shell surrounding the ions, each one consisting of an average number

of solvent molecules. The solvation number of the first shell is usually approximated as an integer and is also termed as coordination number (CN) or hydration number in the case of aqueous solutions. CN is obtained by integration of the pair correlation function g_{IW} between ions and water

$$h = 4\pi\rho_W \int_0^{r'} g_{IW}(r)\, dr \tag{2}$$

where the value r' is usually taken at the minimum of the pair-correlation function after the first peak. This is usually evaluated from diffraction studies (X-ray or neutron diffraction) [1].

Switching from the structure of solvated ions toward their dynamics, the most important individual quantities are the conductivity and the self-diffusion (diffusion without external electric fields) as mentioned earlier in the section. The former can be measured with an alternating external electric field from the electrolyte conductivities and the transference numbers [1]. The latter can also be measured directly by using isotopically labeled ions in capillaries or diaphragm cells or by spin-echo NMR measurements of suitable nuclei [1]. The process of ion solvation involves a dynamic exchange of solvent molecules within the ions' solvation shells, where molecules continuously move in and out. This exchange rate of water molecules between the hydration shells of ions and the bulk water is indicative of the hydration's strength and, by extension, the influence of ions on the structure of water. The average duration for water molecules to diffuse away from neighboring molecules in bulk water can be deduced from the diameter of a water molecule, d_W, and the diffusion coefficient of pure water [3]. The average residence time of a water molecule within an ion's hydration shell, denoted as τ_{IW}, compared to that in the bulk, is derived from the activation Gibbs energy involved in this exchange ΔG_{exch} as [3]

$$\frac{\tau_{IW}}{\tau_W} = exp(\frac{\Delta G_{exch}}{RT}) \tag{3}$$

The rate constants, k_r, which represent the rate of water molecule release from cations' hydration shells, is expected to be inverse proportional to τ_{IW} values calculated from equation 3. These rates, determined through methods like ultrasound absorption [4] and NMR line width measurements [5], are obviously influenced by the competition between water molecules and anions for positions within the cations' coordination shells. At 25°C, these rate constants cover a range of nearly 17 orders of magnitude as reported in figure 1A taken from [6]. Significantly less data are available for exchange rate of cations in solvents different from water. Moreover, there is a notable scarcity of experimental data on the dehydration rates of anions [7]. These limitations points toward an area where computational models can come to in help of experimental studies.

When moving away from the scenario of infinite dilution, certain factors must be considered. As the concentration of electrolytes increases, the ions can no longer be viewed as interacting solely with the surrounding solvent molecules. In such cases, the influence of the ionic environment on the ideal behavior of dilute solutions is quantitatively captured by the activity coefficients $\gamma_\pm$. In solutions with ionizing substances, independently determining the activity coefficients for cations and anions is impractical due to the mutual influence of both ion types on the solution's properties. Therefore, the activity coefficients of individual ions should be linked to the activity coefficient of the electrolyte, assuming it remains undissociated. This leads to the introduction of the mean activity coefficient concept for ionic solutions, where $\gamma_\pm = 1$ denotes ideal behavior. Figure 1B, taken from [1], illustrates the deviation of the mean ionic activity coefficients from 1 with increasing salt concentration. As the concentration of electrolytes rises further, other nonlinear effects become significant. One such effect is ion-ion association. The formation of ion pairs (IPs) or

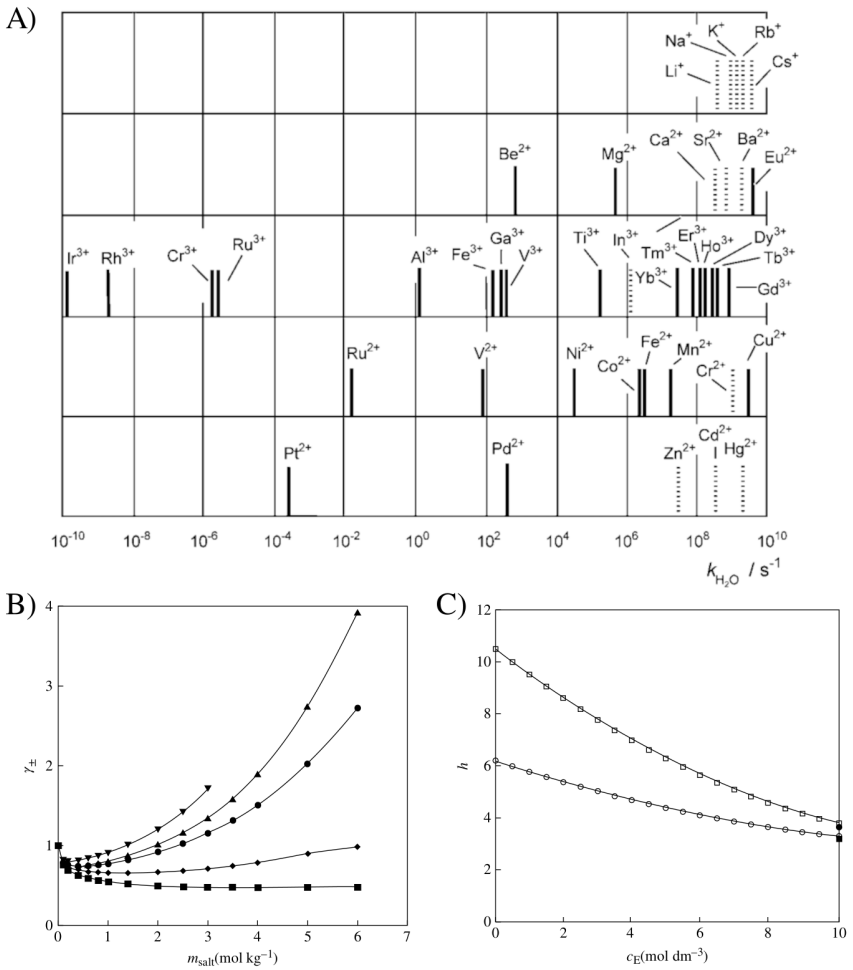

Figure 1: A) Water exchange rate constant of several metal ions, measured with NMR (continuous lines) or derived from complex formation reactions (dashed lines). B) Mean ionic activity coefficients of aqueous LiCl(circles), LiBr(upward triangles), LiI(downward triangles), NaCl (rhombi), and CsCl (squares) at 25°C. C) Hydration numbers h at different concentrations of aqueous salts at 25°C for LiCl (circles) and NaOH (squares).

solvent-shared ion pairs (SIPs) can alter the structure and dynamics of ionic solutions due to their stability and long-lived states. For example, one consequence of these formations is the reduction of the hydration number, as depicted in figure 1C taken from [8]. Although these significant effects will not be addressed in this thesis, it is important to be aware of their existence. Understanding these effects can be instrumental in selecting appropriate models and in interpreting experimental data, especially when compared with theoretical calculations.

Ion-Ligand Complexes in Aqueous Solutions

The discussion above can be extended to situations where small to medium-sized molecules (ranging from a few atoms to tens of atoms) are present in ionic solutions. This area of study falls under coordination chemistry, where the focus shifts from ions and water to ions and ligands that interact to varying degrees, forming complexes. The significance of

coordination occurring in a water solvent will be elaborated in the following paragraphs. In the context of metal complex formation in solution, two critical aspects must be considered: thermodynamic and kinetic stability. Thermodynamic stability is linked to the stability constant, also referred to as the formation or binding constant. This constant serves as an indicator of the interaction strength between reagents forming the metal complex. For instance, consider a metal M in an aqueous solution with a ligand L. Since metals are typically coordinated by water in their first solvation shell, the following substitution reaction can be represented as

$$[M(H_2O)_n] + L \rightleftharpoons [M(H_2O)_{n-1}L] + H_2O$$

The equilibrium constant of this reaction is defined by

$$K = \frac{[M(H_2O)_{n-1}L][H_2O]}{[M(H_2O)_n][L]}$$

where $[L]$ denotes the ligand concentration. In diluted solutions, this equation can be simplified by treating the water concentration and the number of water molecules attached to the metal as constants

$$K = \frac{[ML]}{[M][L]}$$

Metal complexes typically exhibit various coordination states, with different numbers of ligands coordinating the metal center. For example, considering the formation of the complex ML_2, the overall constant K_2 can be expressed as

$$M + 2L \rightleftharpoons ML_2$$
$$K_2 = \frac{[ML_2]}{[ML][L]}$$

In this scenario, it is common to distinguish between the overall (or cumulative) β and the step-wise constants K [9]. Instead of defining each complex formation step with the step-wise constants K_1 and K_2, the overall constant β_n can be formulated as the product of n step-wise constants

$$\beta_n = \prod_{i=1}^{n} K_i$$

From classical thermodynamics, the equilibrium constant can be related to the equilibrium Gibbs free energy [10]

$$\Delta G^0 = -RT \ln K_{eq}$$

where R is the ideal gas constant and T the temperature. The Gibbs free energy comprises two main components: enthalpy and entropy

$$\Delta G^0 = \Delta H^0 - T\Delta S^0$$

The enthalpic term is associated with bond strengths, while entropic effects are related to changes in the order/disorder of the entire system. A noteworthy observation arises when comparing complexes formed by monodentate and polydentate ligands, particularly those with analogous binding centers or donor atoms: complexes involving polydentate ligands exhibit greater stability than their monodentate counterparts. This phenomenon is known as the chelate effect, and polydentate ligands are often referred to as chelating agents. These topics will be explored in detail in Chapter 4 dedicated to the chelate effect.

The relationship between higher thermodynamic stability of a complex and a higher formation constant is straightforward. However, it is not necessarily true that a thermodynamically stable complex will also exhibit kinetic stability. For instance, the complex of nickel(II) with CN^- ions $[Ni(CN^-)_4]^{2-}$, despite its high stability constant ($log\beta_4 \simeq 30$ at T=25°C), is kinetically labile as demonstrated by its rapid exchange rate in studies using radioactive $^{14}CN^-$ [11]. Conversely, the hexaamminocobalt(III) complex $[Co(NH_3)_6]^{3+}$, while thermodynamically unstable in acidic solution, requires several days to replace all ammonia molecules with water [12].
To measure the stability constants of metal complex formation, various experimental techniques are employed. Since stability constants are deduced from the equilibrium concentrations of the involved species, most methods focus on measuring these concentrations and calculating the formation constant from their ratios. An exhaustive list of methods is beyond this thesis's scope, but a brief explanation of some common methods is provided.

- **Titration and potentiometry**. Commonly, metal complex formation involves an M^{n+} ion and anions of a weak acid H_lL. The concentrations of each species in solution can be linked to pH values, aiding in stability constant calculation. pH measurements can be conducted via titration or pH meters. Despite its simplicity and effectiveness, this method does not differentiate between isomer formations and requires careful consideration of ionic strength and activity coefficients [13].

- **Conductance method**. This method is suitable for measuring uncharged metal complex stability constants. It is implicitly assumed that such species do not contribute to solution conductivity, with measured conductivity depending solely on free ion concentration. Its precision is a key advantage, though it is best applied to diluted systems with very low ionic strength [14].

- **Spectrophotometric method**. This technique leverages the unique absorption properties of solutions when metal complexes form. Metal complex concentrations can be determined from UV-visible spectra using the Lambert-Beer law

$$log\left(\frac{I_0}{I}\right) = \epsilon C d$$

  where $\left(\frac{I_0}{I}\right)$ denotes the incident to transmitted light ratio, ϵ the absorptivity, C the molar concentration of the absorbing species, and d the light path in cm. Although rapid and straightforward, spectrophotometric measurements lack high precision but are useful for low solubility species unsuitable for direct titration [15].

- **Polarographic methods**. Here, stability constants are inferred by measuring changes in the half-wave potential $\Delta E_{1/2}$ in solutions with varying ligand concentrations, relative to the potential in the absence of ligands. It has been established that [16]

$$\Delta E_{1/2} = -\frac{RT}{nF} ln\left[\left(\frac{D_C}{D_M}\right)^{1/2} \sum_{i=1}^{n} \beta_i [L]^i\right] \tag{4}$$

  where R is the ideal gas constant, F the Faraday constant, T the temperature, and β_i the overall constants for the complex ML_i. D_C and D_M represent the diffusion coefficients of the complex and free metal, typically assumed equal [16]. In eq. 4, [L] is formally the ligand activity at the electrode surface, but can be approximated as the ligand bulk solution concentration in diluted solutions [13].

In aqueous solutions, every complex formation reaction can be regarded as a substitution reaction involving a ligand L and water molecules. Concerning the reaction mechanism of ligand substitution, four categories are identified based on the step that determines the reaction rate [17].
Consider the substitution reaction between two ligands X and Y

$$M - X + Y \rightarrow M - Y + X \tag{5}$$

Then, four types of substitution reactions can be described

- **Associative A**. The formation of the M-Y bond is completed before the M-X bond begins to break.
- **Interchange-associative $\mathbf{I}_a$**. The breaking of the M-X bond initiates before the complete formation of the M-Y bond, yet the bond formation remains the velocity determining step.
- **Dissociative D**. The M-X bond is entirely broken prior to the formation of the M-Y bond.
- **Interchange-dissociative $\mathbf{I}_d$**. The formation of the M-Y bond commences before the complete breaking of the M-X bond, with the bond rupture being the velocity determining step.

Distinguishing between these mechanisms is challenging as the kinetic law often does not permit clear differentiation [17]. A typical substitution reaction in neutral aqueous solutions can be represented as equation 5 with H_2O as X and ligand L as Y. Should the reaction follow a D mechanism, the initial step is the breaking of the metal-water bond, succeeded by the formation of the metal-ligand bond. In high ligand concentration conditions, the equation can be approximated by a pseudo-first-order law [10]

$$velocity = k_{obs}[M - OH_2] \tag{6}$$

where k_{obs} represents the experimentally observed rate constant, typically in units of s^{-1} (or M^{-1} s^{-1}, if dependent on a reagent concentration). Since eq. 6 exemplifies the typical pseudo-first order law and is also applicable to other mechanisms, distinguishing them becomes nearly impossible, necessitating *ad hoc* experiments to determine the step that dictates the reaction velocity [17].
Complex reaction rate constants and ligand exchange rates are usually determined experimentally through two methods: NMR and stopped-flow techniques.

- **NMR**. Changes in NMR spectrum parameters, namely the half-height peak width $W_{1/2}$ and the chemical shift $\Delta\nu$, occur during complex formation. These changes can be linked to the mean-life time of the complex through [18]:

$$\frac{1}{\tau_{complex}} = \pi(W_{1/2}^{complex} - W_{1/2}^{0}) \tag{7}$$

$$\frac{1}{\tau_{complex}} = \frac{\sqrt{2}\pi(\Delta\nu_0^2 - \Delta\nu_{complex}^2)}{2} \tag{8}$$

  Eq. 7 is suitable for extremely slow exchange dynamics, while eq. 8 applies to faster dynamics. However, $\tau_{complex}$ alone does not reveal the mechanism of the exchange reaction. To determine these quantities, equilibrium concentrations must be measured and the rate equations solved [19].

- **Stopped-flow techniques**. These techniques refer to experimental procedures where multiple reagents are rapidly combined and then halted in an observation cell. Here, a system characteristic, often light absorbance, is monitored for changes as the reaction progresses. These changes, reflecting variations in reagent concentration over time, are used to determine rate constants [20]. Specialized instruments are available, allowing the tracking of reactions occurring from 1 millisecond to hundreds of seconds.

In the study of ligand complexes, besides thermodynamic and kinetic aspects, complexes' geometry emerges as another crucial factor. These three aspects are interrelated, and an ideal comprehensive model would encompass all of them. While this aspect primarily stems from quantum mechanics (QM) considerations and will not be covered in this thesis, its significance warrants a brief overview on three of the most successful theoretical frameworks developed for understanding the coordination chemistry of metal-ligand complexes, addressing their geometrical, thermodynamic, kinetic, and electromagnetic characteristics.

- **Valence Bond Theory (VB)**. This theory conceptualizes the interactions between the metal center and ligands as Lewis acid-base dative bonds. It explains the magnetic properties and geometries of metal complexes through the hybridization of the metal's valence orbitals s, p, and d. While VB theory can elucidate the geometrical and magnetic properties of coordination complexes, it falls short in providing a quantitative interpretation of magnetic data and fails to account for the spectroscopic properties, as well as the thermodynamic or kinetic stability of metal complexes.

- **Crystal Field Theory (CF)**. Essentially a non-bond model, CF views the interactions between metal and ligands as purely electrostatic. Subsequent developments of this theory, known as Ligand Field (LF) theory, introduced a more covalent character to the metal-ligand bonds [21]. CF theory posits that the electric field of ligand electrons influences the electrons in the metal's d valence orbitals. These d orbitals, typically degenerate in an isolated metal atom, undergo splitting due to electronic repulsion between metal and ligand electrons. A notable achievement of CF theory is its ability to straightforwardly explain high-spin and low-spin octahedral complexes, based on the premise that different ligands generate crystal fields of varying strengths. However, CF and LF theories are limited in that they only consider electrostatic effects for the splitting of the d orbitals.

- **Molecular Orbital Theory (MO)**. MO theory accurately represents the sharing of electrons between the ion center and the ligands. It is based on constructing ligand group orbitals (LGO) through a linear combination of ligand orbitals, which can overlap with the metal's d orbitals, depending on the complex's geometry. MO theory is the current state-of-the-art for effectively treating metal-ligand interactions. Its efficacy is proven in explaining the electronic spectra of metal complexes [22] and certain unique effects, such as the Jahn-Teller effect [23].

Although not covered in detail, it is essential to recognize the existence of these theories and understand that any ion-ligand model will eventually involve approximations related to these theoretical descriptions.

Ion Transport in Ion Channels

The study of ion interactions with large biomolecules in aqueous solutions represents a crucial area of scientific inquiry, especially due to its profound implications for understanding the fundamental principles of living human cells and the mechanisms underlying health and disease. When ions are placed in water solutions alongside enzymes, proteins, transporters, and channels, their interactions become a focal point of research. This is because these interactions are central to numerous cellular processes, from the basic functioning of cells to complex physiological systems.

Intricate mechanisms within cells, such as enzymatic reactions, protein folding, signal transduction, and ion transport, are significantly influenced by the presence and behavior of ions in the cellular environment. For example, ions play a pivotal role in enzyme catalysis, often acting as cofactors that enable or enhance enzymatic activity [24]. Similarly, in proteins, ion interactions can affect structural stability and function [25]. Furthermore, ions are integral to the operation of transporters and channels, which regulate the movement of molecules and ions across cell membranes, a process essential for maintaining cellular homeostasis [26].

For example, potassium (K^+) is a very important ion for human cells. Maintaining the proper concentration of K^+ is critical for ensuring cellular functionality and overall physiological balance in the body. Key processes like neuromuscular excitability, which are fundamental for the contraction and relaxation of muscles, are heavily influenced by fluctuations in the potassium levels between intracellular fluid (ICF) and extracellular fluid (ECF). Such a delicate balance of potassium is paramount for normal physiological functioning [27, 28, 29]. Any imbalance, such as hyperkalemia (high K^+ concentration) or hypokalemia (low K^+ concentration), can lead to severe health issues like respiratory and metabolic acidosis, arrhythmias, and impaired muscular and neural function [30, 31].

An other example regards metal ions. They are present in around 1/3 of the entire protein data bank (PDB) [32] and they play a crucial role in the biological systems of living organisms, serving as essential components for various cellular functions while avoiding toxic overload. Organisms rely on a well-coordinated array of metal homeostasis factors, including acquisition and storage proteins, transporters, metallochaperones, and metal-sensing transcriptional regulators, to ensure a proper balance of these vital elements [33]. In addition, understanding how metalloproteins acquire the correct metal for proper function, following the Irving–William stability series, is vital [34].

Understanding these ion-biomolecule interactions in an aqueous environment is crucial not only for grasping the normal functioning of cells but also for deciphering how cellular processes may go awry, leading to disease. Thus, the study of ions in aqueous solutions in the context of large biomolecules is not just a pursuit of basic scientific knowledge but also a gateway to medical breakthroughs that could transform our approach to treating various diseases.

Having established the vast and significant role of ions in biology and biochemistry, it is clear that covering the entirety of this extensive field in literature is impractical. Therefore, the focus will now shift specifically to the interactions and transport of ions within the realm of ion channels. Given the vast nature of ion channels, it is crucial to first define what an ion channel is, identify its constituent parts, differentiate among the various families and their distinct roles, and finally, explore how they can be studied from a quantitative perspective.

Ion transport in human cells occurs through ion channels or ion pumps. Ion channels, also known as passive transporters, facilitate transport driven by concentration gradients or

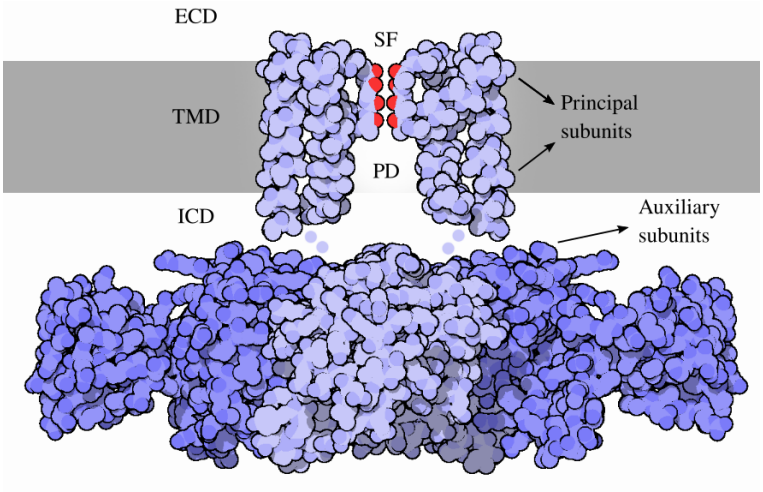

Figure 2: Structure of MthK. A prokaryotic potassium channel calcium-gated.

electrochemical forces. Conversely, ion pumps, termed active transporters, function against the concentration gradient [35]. More than 100 types of ion channels are recognized [36], and categorizing them is challenging due to their diverse roles in mediating ion transport in cells. Nonetheless, their general structure can be described.

As illustrated in figure 2, an ion channel typically comprises several principal subunits (monomeric or multimeric, which can be homomers or heteromers) that form the pore domain spanning the membrane. Additionally, ion channels may possess auxiliary subunits located either in the transmembrane domain (TMD) or in the cytoplasmic intracellular domain (ICD). These subunits can play significant roles, as they might mitigate the effects of certain detrimental mutations, a topic explored in Chapter 6.

However, the most intriguing aspect of an ion channel is arguably the selectivity filter. This concept, postulated by Hille [26], could explain for example the high efficiency of potassium channels in selectively allowing only K^+ cations to pass through, while excluding others like Na^+ or Ca^{2+}. The selectivity filter is situated on the peripheral part of the pore domain, enabling it to accurately select the appropriate cation to be taken up from the ECF or down into the ICF.

Ion channels can also be categorized based on their gating mechanisms, which refer to the processes that enable their opening or closing. A prominent group within this classification is the voltage-gated ion channels (VGICs). For instance, voltage-gated sodium channels are crucial for neuronal excitability, playing a pivotal role in initiating and propagating action potentials in neurons [37]. Voltage-gated calcium channels, on the other hand, are primarily involved in muscle contraction [38], while voltage-gated potassium channels are key in repolarizing action potentials, thereby influencing the duration and frequency of these potentials [26]. Another significant category is the ligand-gated ion channels (LGICs). In this case, the gating mechanism is controlled by the binding of a ligand (or ligands) to a receptor protein in the extracellular domain (ECD). This binding triggers a conformational change in the ion channel, allowing the flow of ions towards the intracellular domain (ICD) [36]. Families such as the nicotinic acetylcholine receptors and $GABA_A$ receptors are extensively studied examples of LGICs [35].

The significance of hydration in the mechanisms governing ion transport through ion channels merits emphasis. Over the past 30 years, two primary theories have been proposed regarding ion translocation in ion channels. The first, known as the direct knock-on (or

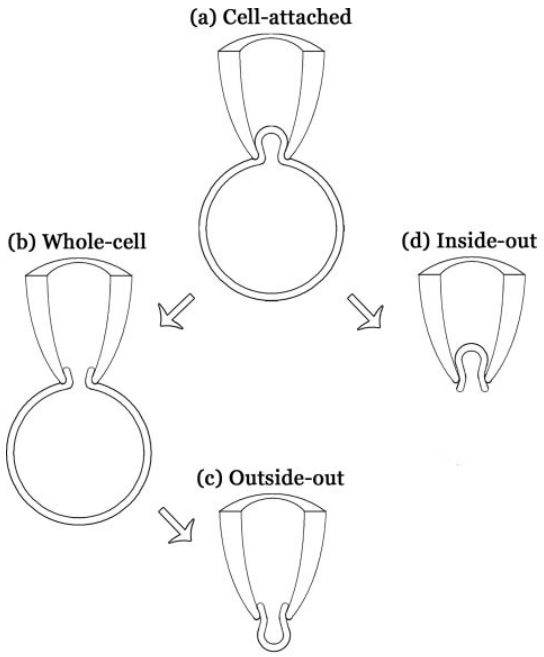

Figure 3: Four possible experimental setups for patch clamp experiments. A) Cell-attached. B) Whole-cell. C) Outside-out. D) Inside-out.

hard knock-on) mechanism [39], suggests that water is absent at the selectivity filter (SF) level and that an incoming ion displaces the previously entered ion, effectively pushing it through the channel. In contrast, the soft knock-on mechanism [39] hypothesizes the presence of water in the SF, which separates ions to prevent direct ion-ion contact. Although this issue remains unresolved, it is clear that water's role in ion channel studies is not peripheral but rather could be key in uncovering significant natural mechanisms. Computational research examining the translocation of Na^+ and K^+ through a sodium channel revealed that while the pore size should permit the passage of both ions along with their hydration shells, the geometry of the first hydration shell for K^+ requires significantly more energy to pass through due to a glutamate residue interacting more on its path, unlike the sodium hydration shell [40]. A recent work supporting the soft-knock mechanism has demonstrated the importance of water behavior involving the breaking and reforming of hydrogen bonds between water molecules and the backbone [41]. Thus, understanding water's behavior within the selectivity filter is crucial for a deeper comprehension of ion transport mechanisms.

A pertinent question arises regarding the methods used to study ion channels. In the literature, one can find studies employing fluorescence-based methods, flux-based assays, and sodium NMR spectroscopy [35], although these techniques do not directly measure ionic current. The following paragraphs will discuss a technique known as patch clamp, which is considered the gold standard for directly studying ion channel activity [42]. A significant advancement in this field was the development of techniques to create seals with high resistance (in the order of gigaohms) for precise measurements. In essence, patch clamp is a technique that relies on establishing a high-resistance seal between a pipette and the cell membrane to which it is attached. This setup ensures that the majority of the electrical current flows through the patch of the membrane covered by the pipette. Various configurations of this technique are illustrated in figure 3 taken from [43].

- **Cell-attached** This setup needs the pipette tip to be positioned onto the targeted cell, allowing the recording of individual ion channels present in the membrane area covered by the pipette's tip.
- **Inside-out** This setup is reached after the cell attached one if the pipette is retracted rapidly to expose the cell membrane's inner surface to the external bath solution. The inside-out patch is particularly valuable for examining the behavior of ion channels that are activated by intracellular ligands.
- **Whole-cell** This configuration gains access to the cell's interior and enables control over the entire cell's voltage through voltage clamping. This allows for the observation of cumulative currents from all ion channels in the cell membrane.
- **Outside-out** This setup is reached removing the pipette from a whole-cell setup. The external surface of the membrane faces the bath solution. This arrangement is optimal for the study of ion channels that are gated by extracellular ligands.

To select just specific ionic currents one can use "blockers" to avoid the presence of certain type of currents (i.e. removing sodium ions to avoid sodium currents) or particular range of voltages that do not activate particular ionic currents. A common procedure is to apply voltage protocols. Voltage protocols involve applying a series of controlled voltage steps or ramps to the cell, allowing the study of ion channel responses under different electrical conditions. These protocols can vary the membrane potential systematically, enabling the measurement of ion channel currents and their dynamics in response to specific voltage changes. This technique would be discussed and its results modeled in Chapter 6.

Modeling and Computational Approaches

Due to the significant size disparity between systems containing ions and water, potentially including ligands (comprising up to a few tens of atoms), and those involving ion transport in ion channels, this section will be divided into two parts. The first part will briefly overview the numerous theoretical models and simulation protocols developed over the past 60 years for studying ion-water and ion-ligand interactions. The second part will address the approaches proposed in the last decades for simulating ion channels.

Regarding the first part, methodologies can be further divided into two categories based on the level of theory: quantum mechanical (QM) models and classical models. Starting from the higher level, there are coupled cluster (CC) methods, often considered the gold standard. Although capable of achieving chemical accuracy for small molecules, CC methods have shown limitations in describing transition metal (TM) ions using single reference character. For improved results, a multi-reference character must be employed, which incurs very high computational costs [44]. Other post-Hartree-Fock methods involving perturbation theory, like MP2 and MP3, have not performed significantly better than density functional theory (DFT) calculations. In fact, for certain functionals, they even perform worse when used to compute dissociation enthalpy changes in a number of ion-containing complexes [45]. Composite schemes, which achieve higher accuracy through a combination of procedures at lower levels of theory, have garnered much attention for being computationally more efficient. The correlation consistent composite approach (ccCA) [46], for example, achieved a mean absolute deviation (MAD) of less than 4 kcal/mol when computing enthalpies of formation for 225 species containing TM ions [47]. DFT calculations, usually less demanding in terms of computational time than the previously mentioned methods, vary significantly in accuracy depending on the chosen functional and system under study [48]. Generally, it has been observed that hybrid functionals perform better in reproducing both higher-level theoretical results and experimental data for ionization potentials (IP) and heats of formation (HOF) [44, 49]. It is important to underline that most of these high level calculations have performed in gas phase, rarely these kind of calculations are performed taking into account the solvent and if it is done it is usually in DFT calculations modeled as implicit [50] or few water molecules are explicitly added to the model [51, 52] to mimic the first hydration shell.

Recent years have seen growing interest in methodologies broadly categorized as semi-empirical methods, which have shown much promise. Notably, density functional based tight binding (DFTB) models have successfully replicated the correct sequence of the three most stable isomers of a copper cluster with three glycine molecules, one histidine, and one pyridine within a solvation cluster of 84 water molecules [53]. This achievement was comparable to the results of DFT calculations performed at the BP86/TZVP level, but with significantly reduced computational expenses [53]. One last very appealing method involving the use of QM calculations are QM/MM methods, where a part is simulated at QM level and the other part is simulated with empirical force fields. Is a method still in its development stage however could suffer of the intrinsic level of theory used for the QM part. If a higher level of theory is used to reach better accuracy then the problem could be that the sampling of the QM/MM simulation is limited, on the other hand it has been shown that if semi-empirical methods are used for the QM part it could produce larger distortions on the system structure with respect to additive force field when comparing the experimental XRD structure [48].

In contrast, classical force fields for aqueous ions and ion-ligand solutions are typically represented as point charges and van der Waals spheres. They utilize a Coulomb potential

to model electrostatic interactions and a Lennard-Jones potential for modeling close contact repulsion and dispersion forces (further details in section 1.1.2). These functions, or modifications of them, are commonly employed in molecular dynamics simulations. Since they involve adjustable parameters, the choice lies in fitting these parameters based on either QM calculations (bottom-up approach) or experimental data (top-down approach). Fitting parameters to QM results is often done when experimental data are lacking, but it is important to remember that QM calculations consider many-body effects, which may not translate accurately to the simpler two-body interactions used in molecular mechanics. This can lead to overestimations of hydration free energies or coordination numbers for ions with water, as QM calculations typically only consider the first layer of water around an ion [48]. Alternatively, fitting classical force field parameters to experimental quantities (e.g., hydration free energies, coordination numbers, ion-oxygen distances, ion-ligand distances) can yield results more consistent with experiments. However, this approach often suffers from limited transferability due to the parameterization being highly system-specific, even among different water models [54]. An interesting approach for non-polarizable force fields is charge scaling, where charges are scaled by a factor of $1/\sqrt{\epsilon_d}$, with ϵ_d being the high-frequency (electronic) dielectric constant of the solvent under study [55].
Lastly, it is crucial not to overlook polarizable models, a significant subgroup of classical models. Their development became essential because polarization energies can significantly contribute to the system's total energy and cannot be disregarded. In the literature is possible to find three main types of polarizable models: the Fluctuating Charge (FQ) model, the Drude Oscillator (DO) model, and the Induced Dipole Model (IDM). The FQ model stands as one of the most basics among polarizable models. Despite its reliance on a point monopole approach, it accommodates charge variations in response to shifts in the chemical environment applying a second-order Taylor series expansion relative to the charges of the chemical potential [56]. DO model employs a Drude particle as the satellite particle for the atomic core's induced electronic cloud. When there is no electric field present, the Drude particle is positioned at the center of the atomic core. However, when the atom is subjected to an electric field, the Drude particle is displaced by a distance d from the atomic core [57]. Meanwhile IDM differences with respect to unpolarizable models lie in the calculation of electrostatic energy at each site, which considers the induced dipole and the electric field at that site. The induced dipole μ_i of a molecule is typically modeled, in the linear representation, as the product of its polarizability α_i and the total electric field E_i at that point [58]. Despite their undoubted greater accuracy compared to non-polarizable models, polarizable models face a couple of challenges. First, simulations using polarizable models may need smaller time steps than those with non-polarizable models to ensure energy conservation, which is calculated using extended Lagrangian algorithms [59] in many codes. Second, the increased accuracy often comes at the cost of more complex parametrization, which can also suffers from poor transferability, similar to non-polarizable models [48].
Proper modeling of large biomolecular systems like ion channels, along with their environments, requires a different scale of approximations compared to the methods previously discussed. For instance, in such systems, QM calculations might only be feasible for a small QM section of QM/MM simulations. To study ion interactions through various residues of the channel, the QM part's level of theory must be relatively low to sample sufficient configurations. This necessitates evaluating the trade-off between higher accuracy (though sometimes specially developed force fields could outperform semi-empirical methods in ion channel simulations [60]) and the feasible computational time. The most common approach is full-atomistic modeling with classical force fields, as was done in the first atomistic simulation of an ion channel in 1984 [61].

The limitations of classical force fields mentioned earlier are applicable here as well. However, the ability to parallelize classical MD codes on GPUs, enabling simulations of a few microseconds even for ion channels, remains a significant advantage. A challenge for fully-atomistic simulations of ion channels is sometimes the lack of high-resolution structural data from X-ray, NMR, or Cryo-EM. Previously, homology modeling was a key method, replicating the structure of a protein with a similar amino acid sequence [62]. More recently, AlphaFold has become a crucial tool in predicting the structure of large proteins from their sequence alone [63]. In theory, polarizable models could be used for ion channel simulations. However, the widely-used polarizable force field AMOEBA [64] has mainly been parameterized against a limited range of small molecules, peptides, and lipids, lacking parameters for many essential biomolecules, indicating substantial work still needs to be done in this area. A very promising approach in terms of accuracy and simulation time is coarse-graining (CG) models of ion channels. Marrink and colleagues successfully employed interaction beads representing multiple atoms. While some atomistic details were lost, they were able to observe the gating mechanism of the MscL channel, including its opening and closing, upon stretching of the membrane in which it was embedded [65].
As previously mentioned, observing the gating mechanism of ion channels and measuring realistic ion currents is currently unfeasible using the approaches described earlier. Even before a clear understanding of what an ion channel was, Hodgkin and Huxley developed a phenomenological model to explain the sodium and potassium conductance in the giant squid axon upon membrane excitation [66]. Their fundamental concept involved "gating particles," each of which needed to be open to conduct. For potassium currents, these particles were considered identical, whereas for sodium, three were identical and one was different. By modeling the particles as being in either an open or closed state using first-order kinetics, they were able to fit the experimental data to the gating particles' parameters. This simple yet groundbreaking idea is still utilized today. To provide greater flexibility to the model, a general Markov state model can be adopted. In this model, the entire channel is thought to exist in one of several possible states defined by the model. This approach will be detailed in section 1.3 and applied carefully in Chapter 6.
Finally continuum models cannot be left out of the discussion. To describe the electrostatic of a solvated ion channel the Poisson-Boltzmann (PB) equation is the most popular theoretical model, it assumes that the charges distribution ($\rho(\mathbf{r})$) is linked to the electrostatic potential ($\Psi(\mathbf{r})$) according to Boltzmann statistics [60]. However, the PB equation characterizes the equilibrium charge distribution. To model ion transport, which involves ion flux, the Poisson-Nernst-Planck equation is more appropriate [36]. The density current then can be described as the sum of two components: the density current opposing the concentration gradient (Fick's law)

$$J(\mathbf{r},t) = -D(\mathbf{r})\nabla c(\mathbf{r},t)$$

where $D(\mathbf{r})$ is a diffusion term and $c(\mathbf{r},t)$ is the concentration of the ion species under study. The second contribution comes from the electrical drift arising from a voltage potential

$$J(\mathbf{r},t) = -D(\mathbf{r})\frac{c(\mathbf{r},t)}{k_{\mathrm{B}}T}\nabla\Psi(\mathbf{r})$$

It is important to note that while modeling ion transport with these equations can be computationally efficient, many details may be overlooked due to the mean-field approximations.

Aim and Outline

Given the broad scope of topics addressed in this dissertation, defining a singular, overarching goal is challenging. Nonetheless, a general objective is outlined below and will be further elaborated upon in the case studies. This thesis is dedicated to introducing computational methods for extracting chemical and physical properties of systems containing ions in solution. The primary aim is not necessarily to attain the highest level of accuracy in deriving these properties, but to develop versatile methods applicable to a wide array of scenarios (such as different ions and diverse ligands). This approach seeks to enhance the current literature by advancing simulation algorithms towards time scales that align more closely with experimental conditions. The intention is to gather useful information for experimentalists and also facilitates a more direct correlation with experimental data. Such comparisons could lead to the more refined and self-consistent improvement of the models presented. The thesis is structured as follows

- Chapter 1: An in-depth discussion of the theories and methods employed in this thesis will be provided.
- Chapter 2: This chapter explores the development of a computational protocol for investigating the water exchange dynamics of aqua ions. It focuses on the combined use of enhanced sampling and stochastic methods to extract critical thermodynamic and kinetic data, with special attention to often overlooked kinetic properties.
- Chapter 3: The development of a 12-6-4 LJ interaction potential between metal ions and acetate is presented. The C_4 term of different metal ions against the carboxylate group has been carefully tuned against experimental binding free energies to achieve a discrepancy ≤ 0.3 kcal/mol.
- Chapter 4: The development of a computational methodology to study the ion-ligand interaction in water solution is discussed. The method presented in chapter 2 is refined and generalized to take into account the effect of different ligands binding the metal center and competing with water molecules. The methodology was applied to different amines with different denticity to gain insights onto the chelate effect.
- Chapter 5: A study of the pathway of ions translocating toward an engineered MscL channel is presented. This results are compared with a geometrical tool developed to monitor the size and the physico-chemical description of the surface facing the pore lumen.
- Chapter 6: A molecular dynamics study of the Kv4.3 voltage-gated potassium channel and two hereditary point mutations is presented. Together with a Markov state model to study the current response of the ion channel upon membrane excitation. Moreover, simulations' results are compared against experimentally available data.
- Chapter 7: The conclusions are presented, together with possible future improvements of the methodologies developed in the thesis to study more complex systems.

Chapter 1

Theoretical and Computational Background

1.1 Molecular Dynamics

Biological molecules and macromolecules exhibit a wide range of characteristic motions at various time and spatial scales. These motions encompass fast, localized movements involving small chemical groups and slower, long-range, collective motions responsible for processes such as molecule folding, domain swapping, and molecular docking.
Molecular dynamics (MD) techniques have the potential to atomistically track the structural and dynamic changes within molecular systems over time. They allow researchers to connect the outcomes of molecular simulations with physically and chemically relevant properties, facilitating comparisons with experimental results and, in some cases, predictions [67]. Molecular dynamics has found successful applications in the study of diverse biological phenomena, including ionic transport, lipid membranes, protein stability, entire proteins in solution with explicit solvent representations, membrane-embedded proteins, and large macromolecular complexes like nucleosomes [68, 69, 70, 71].
Over the last few decades, molecular dynamics simulations have evolved into a well-established and widely used technique for understanding complex biological processes. Today, routine simulations involve systems with approximately 100,000 atoms, while simulations with around 500,000 to 1 million atoms are common when sufficient computational resources are available. This significant progress is attributed to the utilization of high-performance computing (HPC) and the parallelizability of the fundamental molecular dynamics algorithm. Molecular dynamics codes like AMBER [72], CHARMM [73], GROMACS [74], and NAMD [75] have been adapted for use with the Message Passing Interface (MPI) and, more recently, with CUDA for GPU acceleration. The parallelization strategy typically involves spatial decomposition, where the system to be simulated is divided among processors or GPUs based on their positions in space. Each processor handles a specific region of space, regardless of which particles are present there. This approach minimizes inter-processor communication, significantly enhancing computational efficiency and enabling simulations over longer timescales.
In this section, it will be provided a brief overview of the theoretical foundations of molecular dynamics, with a particular focus on Langevin dynamics. However, given the vast scope of molecular dynamics simulations, a comprehensive exploration of all techniques and methods is beyond the scope of this section and the entire thesis. Instead, the focus will be on the tools implemented and utilized in Chapters 2, 4, and 5.

1.1.1 Statistical mechanics foundations

Microscopic information about the particles constituting a system, including their positions, velocities, and forces, can be leveraged to investigate the thermodynamics of the entire system. Statistical mechanics provides the link between macroscopic properties of a system and the characteristics of its individual components [76]. Under equilibrium conditions, the thermodynamic state of a molecular system is governed by an ensemble of microscopic states, also known as microstates. The choice of conserved thermodynamic parameters determines the appropriate statistical mechanics framework [77]. The following ensembles are commonly used:

- Microcanonical ensemble (NVE): This ensemble is characterized by a fixed number of particles (N), volume (V), and energy (E). However, it often does not align with the conditions of most experiments.
- Canonical ensemble (NVT): In this ensemble, the number of particles (N), volume (V), and temperature (T) are fixed. The temperature has a specified average value, while the total energy of the system (Hamiltonian $\mathcal{H}(\mathbf{r})$) can fluctuate.
- Isobaric-Isothermal (Gibbs) ensemble (NpT): This ensemble maintains a fixed number of particles (N), pressure (p), and temperature (T). Both pressure and temperature have specified average values, and the system's volume (V) can fluctuate.
- Grand canonical ensemble (μVT): In this ensemble, there are fixed values for chemical potential (μ), volume (V), and temperature (T). The volume and temperature are constant, similar to the canonical ensemble, but the system can exchange particles with a surrounding bath. The chemical potential of different species has a specified average value, while the number of particles (N) can fluctuate.

To evaluate a property of a molecular system within one of these ensembles, it depends on the coordinates and momenta of the individual particles within the system. The instantaneous value of a property A can be expressed as $A(\mathbf{q}^{3N}(t), \mathbf{p}^{3N}(t))$, where $\mathbf{q}^{3N}(t)$ and $\mathbf{p}^{3N}(t)$ represent the coordinates and momenta of N particles along the x, y, and z directions. Over time, the instantaneous value of A fluctuates due to interactions between the system's components.
Experimental measurements yield an average value of a property, which can be regarded as a time average. Theoretically, as the measurement time approaches infinity, the time average converges to the true average value, as

$$A_{ave} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t_0}^{\tau} A(\mathbf{q}^{3N}(t), \mathbf{p}^{3N}(t))\, dt \tag{1.1}$$

However, practically, it is infeasible to calculate these averages directly due to the large number of atoms or molecules in the system. Boltzmann and Gibbs introduced the concept of the *ergodic hypothesis* to address this issue. This hypothesis suggests replacing a single system evolving over time with a multitude of replicas considered simultaneously. The ensemble average, denoted as $\langle A \rangle$, can then be used in place of the time average.

$$\langle A \rangle = \int \int d\mathbf{q}^{3N}\, d\mathbf{p}^{3N} A(\mathbf{q}^{3N}, \mathbf{p}^{3N}) \rho(\mathbf{q}^{3N}, \mathbf{p}^{3N}) \tag{1.2}$$

where the angular brackets indicate an ensemble average. In this equation, $\rho(\mathbf{q}^{3N}, \mathbf{p}^{3N})$ is the probability distribution function of the ensemble, indicating the probability of finding

a configuration with positions between $\mathbf{q}^{3N}$ and $\mathbf{q}^{3N} + d\mathbf{q}^{3N}$ and momenta between $\mathbf{p}^{3N}$ and $\mathbf{p}^{3N} + d\mathbf{p}^{3N}$. The choice of the correct probability distribution function depends on the selected ensemble. In the canonical ensemble, the Boltzmann distribution function can be written as

$$\rho_{NVT}(\mathbf{q}^{3N}, \mathbf{p}^{3N}) = \frac{1}{Q} e^{-\frac{\left\langle \mathcal{H}(\mathbf{q}^{3N}, \mathbf{p}^{3N}) \right\rangle}{k_{\mathrm{B}} T}} \tag{1.3}$$

Here, $\mathcal{H}$ is the system's Hamiltonian, and $\left\langle \mathcal{H}(\mathbf{q}^{3N}, \mathbf{p}^{3N}) \right\rangle$ represents its expectation value. T is the temperature, and k_{B} is the Boltzmann constant. The partition function (Q) is one of the central quantities in equilibrium statistical mechanics and represents a measure of the number of microscopic states in the phase space accessible within a given ensemble. Again, in the case of the canonical ensemble for N identical particles it can be written as

$$Q_{NVT} = \frac{1}{N!} \frac{1}{h^{3N}} \int \int d\mathbf{q}^{3N}\, d\mathbf{p}^{3N} exp\left[-\frac{\mathcal{H}(\mathbf{q}^{3N}, \mathbf{p}^{3N})}{k_{\mathrm{B}} T} \right] \tag{1.4}$$

where h is the well-known Planck's constant and the factor N! arises from the indistinguishably of the particles. The factor $1/h^{3N}$ is required to ensure that the partition function is equal to the quantum mechanical result for a particle in a box.
In the 1950s [78], researchers developed algorithms to sample the Boltzmann distribution and obtain ensemble averages within the Monte Carlo framework. Over time, improvements like mimicking periodic boundary conditions [79] allowed simulations to integrate the equations of motion for a single microstate in phase space. While it may seem like a solution to the time average problem initially addressed by Boltzmann and Gibbs, it is important to note that simulation time is not infinite.
Due to the impracticality of analytically solving equations of motion for a large number of particles, molecular dynamics simulations involve breaking calculations into numerous short timesteps. As a result, the ensemble average is approximated by the ensemble average in equation 1.5, where M represents the number of time steps

$$A_{ave} \simeq \frac{1}{M} \sum_{i=1}^{M} A(\mathbf{q}_i^{3N}, \mathbf{p}_i^{3N}) \tag{1.5}$$

The choice of M and the size of time steps depend on the specific system under study, representing a crucial consideration in molecular dynamics simulations. In essence, if the sampling time is sufficiently long for the system to explore most of its phase space, equation 1.5 holds true, making MD an effective method for studying and predicting thermodynamic properties across a wide range of systems, from simple to complex.

1.1.2 Molecular mechanics

MD stands as one of the foremost methods for sampling the vast number of microscopic configurations within a molecular system. Through MD simulations, the time evolution of a particle system across phase space can be tracked as it responds to the forces governing its interactions. Given an initial phase space point defined by initial positions ($\mathbf{r}_0$) and momenta ($\mathbf{p}_0$), at a time $t > t_0$, the new positions and momenta are updated by integrating Newton's equations of motion

$$m_i \ddot{r}_i = -\nabla U_i(\mathbf{r}), \qquad \forall i = 1, \ldots, N \tag{1.6}$$

Here, m_i represents the mass of the i-th particle, $\mathbf{r}$ is a vector describing the configuration of the particles in the system, and $U(\mathbf{r}) : \mathbb{R}^{3N} \to \mathbb{R}$ denotes the potential energy function, which will be discussed in detail in the following section. This process is repeated multiple times until the desired number of time steps is reached, as depicted in Figure 1.1.

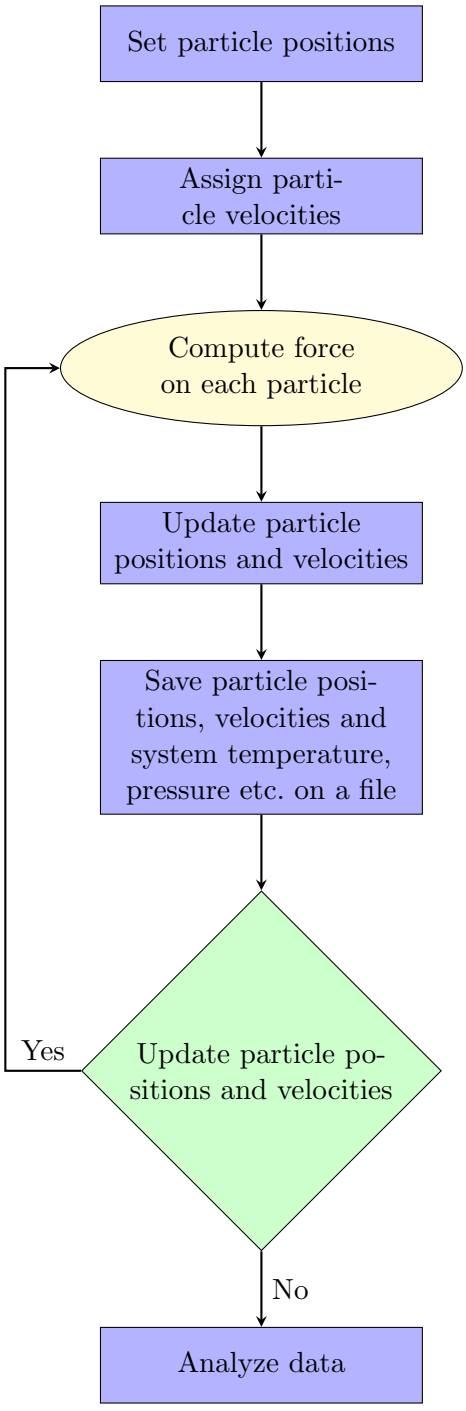

Figure 1.1: Flowchart of a molecular dynamics simulation iterative process.

The force field

Depending on how the potential energy surface is calculated, MD methods can be categorized as either "ab initio", where electronic structure is explicitly considered using quantum mechanics, or "classical" MD, which employs classical Newtonian mechanics to describe the system interactions.
Many problems in molecular modeling are too large to be tackled using quantum mechanics. Quantum mechanical methods primarily deal with electrons, and even semi-empirical approaches that simplify electron calculations still involve a substantial number of particles and are computationally intensive. In contrast, classical MD uses force field methods that neglect electronic motion and compute the energy solely based on nuclear positions. This approach allows molecular mechanics to handle systems with a significant number of atoms [80]. Although molecular mechanics cannot provide properties dependent on the electronic distribution within a molecule, it can yield accurate results for a wide range of properties in a fraction of the time required by quantum mechanics.
Molecular mechanics relies on a simplified model of interactions, encompassing bond stretching, angle opening, and single bond rotations. Despite this simplicity, the force field's empirical analytical functions, such as Hooke's law, Coulomb's potential, Lennard-Jones potential, etc., contain parameters that enable broad applicability across various problems after being tested on a few model systems. Force fields comprise components that describe specific interactions, either bonded or non-bonded. Over the years, parameter sets have been developed to make these interactions as specific as possible for different atomic interactions, accounting for various hybridizations and chemical contexts. Among the numerous parametrized and optimized force fields documented in the literature, AMBER [72], CHARMM [73], and GROMOS [74] are widely recognized and accepted by the scientific community. The differences between these force fields often stem from their parameterization approaches. For instance, AMBER derives its parameters from quantum mechanical calculations, CHARMM predominantly relies on experimental data, and GROMOS starts with experimental data and subsequently refines van der Waals interactions of specific chemical groups (e.g., aliphatic and aromatic) based on molecular dynamics simulations of model liquid alkanes.
A general form of a force field, as mentioned earlier, consists of both bonded and non-bonded interactions and can be expressed as

$$U(\mathbf{r}) = U(\mathbf{r})_{bonded} + U(\mathbf{r})_{non-bonded} \tag{1.7}$$

where the first term could be further expanded into three contributions as

$$U(\mathbf{r})_{bonded} = U(\mathbf{r})_{bonds} + U(\mathbf{r})_{angle} + U(\mathbf{r})_{dihedrals} \tag{1.8}$$

The first term in 1.8 models the interaction between pairs of bonded atoms, usually modelled by an harmonic potential as

$$U(\mathbf{r})_{bonds} = \sum_{bonds} \frac{k_i^b}{2}(r_i - r_{0,i})^2 \tag{1.9}$$

that gives an increase in energy as the i-th bond length r_i deviates from the reference value $r_{0,i}$. The second term in 1.8 is a summation over all the valence angles (the angles formed between three subsequent bonded atoms) in the system, using an harmonic potential of the form

$$U(\mathbf{r})_{angles} = \sum_{angles} \frac{k_j^a}{2}(\theta_j - \theta_{0,j})^2 \tag{1.10}$$

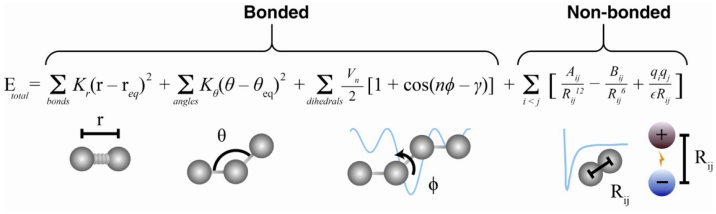

Figure 1.2: Sketch of the single components of a generic force field.

that gives an increase in energy as the j-th angle θ_j differs from the reference value $\theta_{0,j}$ both opening or closing the valence angle. The third term in 1.8 takes into account any torsional contribution as one bond rotates respect to another and could be expressed as an angular spring equation between two planes formed by first three and last three atoms of four consecutively bonded part of a molecule as sketched in figure 1.2. It contributes to the potential energy as

$$U(\mathbf{r})_{dihedrals} = \sum_{dihedrals} k_l^d[1 + cos(n_l\phi_l - n_l\phi_{0,l})] \tag{1.11}$$

and it has a sinusoidal behaviour as depicted in figure 1.2 reaching its minimum when the angle between the two planes differs of 0° from the reference value and its maximum when the angle between the two planes differs of 180° from the reference value. As for what concern the non-bonded part of the force field, historically two potentials have been taken into account.

$$U(\mathbf{r})_{non-bonded} = U(\mathbf{r})_{coulomb} + U(\mathbf{r})_{vdW} \tag{1.12}$$

The first term is the Coulombian potential, used to compute the electrostatic interactions and it can be written as

$$U(\mathbf{r})_{coulomb} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.13}$$

where the notation below the summation should avoid any double counting between the i-th and j-th charges. Finally, the second term of the non-bonded potential is counting for the non-electrostatic intra-molecular and inter-molecular interactions, also known as van der Waals interactions (vdW). This is usually modelled as a Lennard-Jones (LJ) potential

$$U(\mathbf{r})_{vdW} = \sum_{i<j} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.14}$$

where the attractive part $1/r^6$ models the instantaneous dipole-induced dipole interaction and the repulsive part $1/r^{12}$ mimic the large repelling force to avoid compenetration of two atoms within their respective electronic clouds. The term ϵ_{ij} represents the intensity of the London's dispersion interaction and σ_{ij} is the equilibrium distance between the i-th and j-th particles. These two terms could be extrapolated from experimental data or quantum mechanical calculations. In the last 20 years, the vdW interaction term has seen substantial refinements and modifications due to limitations of conventional LJ potentials in describing complex biological systems like protein secondary structures and stability, as noted in references [74], [80], and [81].

12-6-4 Lennard-Jones potential

The primary focus of this thesis is on ions, and while the goal is not to calculate their interactions with water and residues within chemical accuracy, it is essential to utilize the

best available framework for modeling them at the MD level. The limitations of the 12-6 LJ non-bonded potential, as discussed previously, can become more apparent when dealing with metal ions. This is because Li and Merz [54] highlighted that charge-induced dipole interactions, which vary as $1/r^4$, are often overlooked in the standard LJ potential but play a significant role in highly charged systems. They introduced a C_4 term into the LJ potential in 1.14, leading to the following expression

$$U(r_{ij})_{LJ} = \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6} - \frac{C_4^{ij}}{r_{ij}^4} \tag{1.15}$$

Here, $C_{12}^{ij} = 4\epsilon_{ij}\sigma_{ij}^{12}$ and $C_6^{ij} = 4\epsilon_{ij}\sigma_{ij}^6$, and they introduced $C_4^{ij} = \kappa C_6^{ij}$. This clever modification of the LJ potential, referred to as 12-6-4, allows the use of standard LJ parameters, σ and ϵ, from existing force fields. The only parameter that needs tuning is κ, which accounts for ion-induced dipole interactions. This approach enables simultaneous fitting to experimental measurements of metal ion-water coordination number (CN), ion-oxygen distance (IOD), and hydration free energy (HFE), resulting in a better parameterization of metal ions compared to the standard LJ potential. In their seminal work, this was done for various divalent cations against different water models. Due to the additive nature of LJ interactions, 12-6-4 parameters for metal ions have also been developed for different atom types [82] or biological residues [83].
To calculate the C_4 interaction term between a metal ion and a generic atom type, one can follow this rule [54]

$$C_4(M-A) = C_4(M-H_2O)\frac{\alpha_0(A)}{\alpha_0(H_2O)} \tag{1.16}$$

In this equation, the computed C_4 term for the metal ion with a specific water model is scaled by the ratio of the polarizability of the atom type of interest, $\alpha_0(A)$, and the polarizability of a water molecule in that water model, $\alpha_0(H_2O)$.

The choice of the time step

As previously mentioned, the integration of Newton's equations of motion cannot be carried out analytically and must be performed numerically, typically using finite differences. This numerical integration lies at the heart of any molecular dynamics simulation program. Various algorithms are available for this task, each with its own advantages and limitations. The fundamental concept involves breaking down the integration of differential equations into numerous small fixed stages, each separated in time by a timestep, denoted as Δt. Initially, a configuration is provided to the simulation package, either from experiments or quantum calculations, which assigns the initial positions for each component of the system. The starting velocities are usually generated by sampling from a Maxwell-Boltzmann distribution at a specific temperature, T

$$p(v_{r_i}) = \left(\frac{m_i}{2\pi k_B T}\right)^{1/2} exp\left[-\frac{m_i v_{r_i}^2}{2k_B T}\right] \tag{1.17}$$

This process is repeated for assigning velocities in the x, y and z directions for each particle. With the initial positions, it becomes possible to compute the total potential, denoted as $U(\mathbf{r})$ and derive the forces acting on each particle thereby determining their accelerations. In principle, now that all the necessary components are known, it is possible to calculate the new positions and velocities at time $t + \Delta t$. The key assumption in this process is that the force remains constant during the time step. In this context, the algorithms make

use of Taylor series expansions to approximate positions, velocities, and accelerations (and potentially higher-order derivatives) at each time step.

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)(\Delta t)^2 + \dots \tag{1.18}$$

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \mathbf{a}(t)\Delta t + \frac{1}{2}\mathbf{b}(t)(\Delta t)^2 + \dots \tag{1.19}$$

$$\mathbf{a}(t+\Delta t) = \mathbf{a}(t) + \mathbf{b}(t)\Delta t + \frac{1}{2}\mathbf{c}(t)(\Delta t)^2 + \dots \tag{1.20}$$

where $\mathbf{b}$ and $\mathbf{c}$ are the third and fourth time derivative of the positions. One well-known algorithm, the Verlet algorithm, calculates new positions at $(t + \Delta t)$ using positions and accelerations at time t, as well as positions from the previous step $(t-\Delta t)$. Other algorithms, such as the leap-frog and velocity Verlet, share the same underlying principles based on Taylor series expansions. The differences between them are largely technical, affecting how dynamic properties are computed, whether in a single step or in multiple steps. The Verlet algorithm establishes relationships between these positions at time $t - \Delta t$ and $t + \Delta t$ and the velocities at time t [77]. Writing them down as

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)(\Delta t)^2 + \dots \tag{1.21}$$

$$\mathbf{r}(t-\Delta t) = \mathbf{r}(t) - \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)(\Delta t)^2 + \dots \tag{1.22}$$

adding these two quantities leads to

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \mathbf{a}(t)(\Delta t)^2 + \mathcal{O}((\Delta t)^3) \tag{1.23}$$

where everything is known in order to compute the update of the positions.
In order to update the velocities the difference of the two quantities above can be taken and divided by $2\Delta t$

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} \tag{1.24}$$

It can be seen here, that at a computational memory level it is necessary to store $\mathbf{r}(t)$, $\mathbf{r}(t-\Delta t)$ and $\mathbf{a}(t)$ at each step. Lastly, it can be seen that at $t = 0$ there is obviously only one set of positions and so it is necessary to employ some other means to obtain positions at $t - \Delta t$. One way to obtain $\mathbf{r}(t - \Delta t)$ is to use the Taylor series, truncated after the first term[80]. Thus, $\mathbf{r}(-\Delta t) = \mathbf{r}(0) - \mathbf{v}(0)\Delta t$.
Regardless of the integration algorithm used to simulate molecular dynamics, a significant portion of computational resources is devoted to calculating the interaction potential for each particle in the system. This aspect heavily influences the overall simulation time and limits the size of the system that can be effectively studied. Therefore, the choice of the time step is a critical factor, as a larger time step can expedite the phase space sampling but also carries the risk of destabilizing the simulation. In situations where atoms approach each other closely, a substantial repulsive force is encountered, leading to a sudden increase in energy, and the atoms subsequently move apart with higher velocities. When different particles have different time steps, the ones with the shortest time steps tend to move faster precisely in regions where smaller steps would be more appropriate. This disparity in time steps can result in significant errors, potentially causing the simulation to fail. While using smaller time steps can improve accuracy, it comes at the cost of increased computational time for a given simulation duration. Striking the right balance between accurately capturing the trajectory and efficiently sampling the phase space is a key challenge. While

there is no universal rule for selecting a time step, a common approach is to base it on the fastest motion in the system. In molecular dynamics systems, this often corresponds to the stretching vibrations of hydrogen bonds, which occur on the order of $\simeq 10$ fs. Using a time step of 1 femtosecond can provide a good compromise for accurately tracking the trajectory while still efficiently sampling the phase space.
In the last decades, many algorithms [84] (SETTLE, RATTLE, SHAKE, LINCS, etc.) were introduced to overcome this problem. The basic idea was to apply constraints to the system under study, a constraint algorithm is a method for satisfying the Newtonian motion of a rigid body which consists of mass points. In this case usually a restraint is used to ensure that the distance between mass points is maintained. The general steps involved are [84]

- Choose novel unconstrained coordinates (internal coordinates).
- Introduce explicit constraint forces.
- Minimize constraint forces implicitly by the technique of Lagrange multipliers or projection methods.

Constraint algorithms achieve computational efficiency by neglecting motion along some degrees of freedom. These algorithms introduce constraints to the system, effectively limiting certain degrees of freedom. For instance, the SHAKE algorithm freezes the lengths of covalent bonds involving hydrogen atoms [84]. This approach is acceptable when the constrained degrees of freedom have a negligible impact on the phenomena under investigation. In biomolecular systems like proteins or lipid membranes, the vibrations of hydrogen atoms are often of relatively little significance in terms of overall behavior, making constraint algorithms a practical choice.
Additionally, some molecular dynamics packages employ the multiple timestep method. With this approach, different sets of non-bonded interactions are calculated at different frequencies. For example, a smaller timestep may be used for a group of atoms within a certain radius, as they exhibit faster motion, while a longer timestep is applied to other particles outside this range. This strategy can significantly reduce computational costs, especially when certain interactions evolve more rapidly over time than others [80].

Periodic boundary conditions

The proper handling of boundaries and boundary effects is of paramount importance in simulation methods, as it allows the computation of macroscopic properties using a relatively small number of particles. In experimental scenarios, the number of particles can be up to 20 orders of magnitude greater than the number of particles that can be realistically simulated. This means that particles in close proximity to or in contact with boundaries can significantly impact the simulated system. Periodic boundary conditions offer a solution by allowing simulations to be conducted with a limited number of particles while ensuring that particles experience forces as if they were in a bulk fluid. This is achieved by replicating the simulation box in all directions to create a periodic array.
In Figure 1.3, a two-dimensional example is shown. In a typical three-dimensional periodic array, the simulation box is surrounded by 26 nearest neighbors. It's worth noting that while a cubic cell is the simplest periodic system to conceptualize and implement in code, different cell shapes may be more suitable for specific simulations (e.g., triclinic, truncated octahedron, rhombic dodecahedron, etc.). Choosing a periodic cell that aligns with the underlying geometry of the system is often beneficial, especially for systems with spherical configurations. Most molecular dynamics packages now offer various cell geometries. One

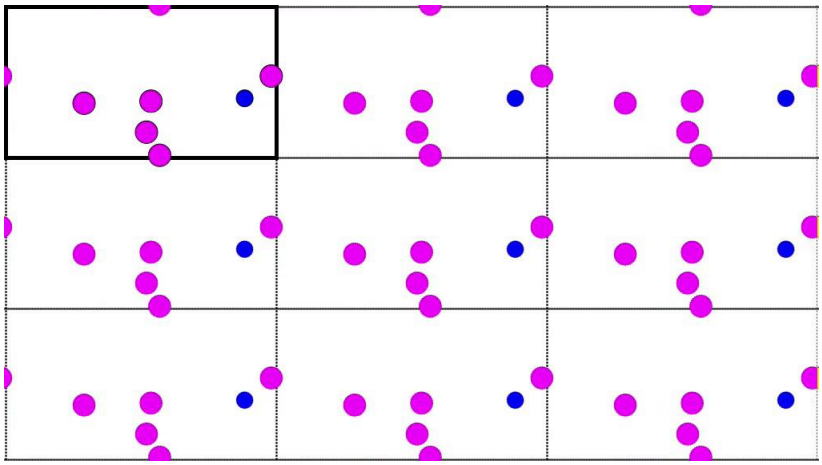

Figure 1.3: Representation of two-dimensional periodic boundary conditions. The central simulation cell is replicated in both the x and y dimensions.

advantage of using periodic boundary conditions is that atoms near the box boundaries can interact with atoms in neighboring cells within a specified cutoff radius, denoted as r_c. It's crucial to select this radius carefully to avoid artifacts resulting from self-interactions of particles with themselves. A common convention is to set the cutoff radius to be smaller than half of the smallest dimension of the box [85]. This ensures that interactions are calculated only with the closest neighboring particles, following the minimum image convention. If a particle were to exit the central box during the simulation, one of its periodic images enters through the opposite face to maintain a constant number of particles in the simulation box.
While periodic boundaries are a well-established tool in computer simulations, they do have limitations. Notably, fluctuations of a physical quantity cannot have a wavelength greater than the length of the cell. The range of interactions present in the system is also a critical consideration. If the cell size is significantly larger than the range over which interactions occur, there should be no issues. For short-range interactions like the LJ potential (which decays as $\simeq r^{-6}$), a simulation box width of a few tens of Angstroms is usually sufficient [85]. However, dealing with long-range electrostatic interactions can be more challenging, and it may be necessary to accept some degree of "long-range order" imposed on the system [77]. Empirical evaluation of the effects of imposing periodic boundaries can involve comparing simulation results obtained with different cell shapes and sizes or matching simulation results with experimental data.
Lastly, non-periodic boundary methods exist and have been implemented in simulation programs in recent decades, particularly in studies of protein-ligand complexes or systems far from equilibrium. However, these approaches require careful consideration of the algorithms used, and additional precautions are often necessary, such as applying constraints to atoms outside the reaction zone in the case of protein-ligand interactions [80].

Long-range interactions

Long-range interactions pose a significant computational challenge in molecular dynamics simulations. In principle, the computational complexity of evaluating pair interactions scales as $\mathcal{O}(N^2)$, where N represents the number of atoms in the system. To mitigate the computational cost of energy calculations, a cutoff radius is often imposed, as previously mentioned. This approach, suitable for LJ interactions, can be further enhanced by mul-

tiplying the cutoff radius by an appropriate function that smoothly approaches zero at r_c. However, for electrostatic interactions governed by Coulomb's law, which exhibit a less pronounced dependence on the distance between atoms ($\simeq r^{-1}$), the effectiveness of a cut-off radius remains an open question. One of the most widely used techniques in molecular dynamics to address this challenge is the particle mesh Ewald (PME) method. PME avoids severe approximations, especially in highly charged systems, where incorrect data interpretation could occur. In the following paragraphs, it will be briefly explained how the Ewald summation or particle mesh algorithm works, starting with Ewald's original idea.
Historically, Ewald introduced an ingenious method to compute Madelung's constant for perfect ionic crystals by transforming a slow, conditionally convergent lattice series, into the sum of two fast absolutely convergent series. One in real space and one in reciprocal space. In the presence of spatial periodic symmetry, the contribution of charge-charge interactions between charges in the central box and all images of particles in neighboring boxes is given by [86]

$$U(\mathbf{r})_{coulomb} = \frac{1}{4\pi\epsilon}\frac{1}{2}\sum_{\mathbf{n}}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{|r_{ij}+\mathbf{n}|} \tag{1.25}$$

where the slowly and conditioned convergent series on the right-hand term depends purely on the geometric structure of the periodic array. To solve this problem the first assumption was to assume a Gaussian charge distribution, instead of a point charge, of the form [86]

$$\rho_i(r) = \frac{q_i\alpha^3}{\pi^{3/2}}exp(-\alpha^2 r^2) \tag{1.26}$$

where it is easy to verify that the electrostatic potential $\phi(r)$ between the i-th and j-th charge distributions is equal to

$$\phi_{ij}(r_{ij}) = -\frac{q_i q_j}{r_{ij}}erf(\alpha r_{ij}) \tag{1.27}$$

where α is called the Ewald splitting parameter and erf(x) is the error function defined as

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt \tag{1.28}$$

that increases monotonically from 0 to 1 as x varies from zero to infinity. It is also helpful to recall that the error function satisfy the relationship $erf(x)+erfc(x)=1$ where erfc(x) is the complementary error function. The last passage introduced by Ewald was to rewrite the potential created by point charges distribution with the one obtained with Gaussian charges distribution as[86]

$$-\frac{q_i q_j}{r} \equiv -\frac{q_i q_j}{r}erf(\alpha r) - \frac{q_i q_j}{r}erfc(\alpha r) \tag{1.29}$$

In Equation 1.29, the well-known Coulomb potential, singular at the origin and long-range in nature, was decomposed into a first term, regular at the origin and long-range, and a second term, singular at the origin and short-range. This decomposition allowed the first term on the right-hand side of Equation 1.29 to be computed in reciprocal space using Fourier transform, while the second term could be easily computed in real space due to its fast convergence. Balancing the real-space and reciprocal-space summations is crucial. The former converges more rapidly for large values of α, whereas the latter converges more rapidly for small α. A common value for α is $5/L$, where L represents the length scale of the

simulation box, and it is suggested to use 100-200 reciprocal vectors k to achieve acceptable results [80].
Additionally, the sum of Gaussian functions in real space includes the interaction of each Gaussian with itself. To obtain an accurate Ewald summation, a third self-term U^{self} must be subtracted, and a fourth contribution from the surrounding medium U^{medium} must be added to Equation 1.29. The Ewald summation method has been extensively implemented in computational simulations and has shown good agreement with expected values for various systems, including ionic melts and biomolecules such as lipid bilayers, proteins, and DNA [77].
However, it is important to note that the Ewald summation is computationally expensive to implement. It was demonstrated [87] that this method can reduce the computational workload of electrostatic pair interactions from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^{3/2})$ with an appropriate choice of the parameter α discussed earlier. Several methods have been proposed to accelerate the computationally intensive reciprocal space part of the calculation, such as using polynomial approximations. Nevertheless, these approaches do not completely solve the unfavorable scaling with N. The most efficient way to address this issue was to modify the problem slightly so that the fast Fourier transform (FFT) algorithm could be used for the reciprocal space summation. Since the FFT algorithm scales as $\mathcal{O}(N \log N)$, it provides significant advantages over the standard Ewald summation. However, implementing the FFT method requires replacing continuous atomic point charges with a grid-based charge distribution, where each atomic charge is distributed among surrounding fictitious grid points to reproduce the original charge's potential. The introduction of the particle mesh Ewald method brought a revolution to molecular dynamics in the 1990s. However, it's important to note that it is not the only method available. For systems with significant charge fluctuations, other techniques like reaction field [88] or fast multipole methods [89] may be more suitable.

1.1.3 Langevin dynamics

In this section, we will explore an alternative approach that deviates from the conventional Newtonian dynamics typically used in molecular dynamics simulations. While Newton's equations of motion are numerically solved to generate trajectories of a system in the microcanonical ensemble, the canonical ensemble, where the system's temperature remains constant rather than its energy, is more commonly used in statistical mechanics. Over the years, various methods have been proposed for conducting molecular dynamics simulations in the canonical ensemble.
One particularly appealing class of methods involves the application of Langevin dynamics simulations, which have been employed in all the MD studies presented in this work (Chapters 2, 4, 5, 6). Moreover, Langevin dynamics is inherently stochastic, making it suitable for several insightful considerations in section 1.3.
Originally designed to model a bath of lighter particles, Langevin equations of motion were later adapted to sample the canonical ensemble when Schneider et al. [90] recognized the potentialities of using Langevin equations as a thermostat. This approach assumes that the particles undergo collisions with much lighter ones, representing the heat bath. To describe these collisions, a frictional term and a random force were introduced, resulting in two

coupled equations governing Langevin dynamics. These equations can be expressed as

$$d\mathbf{q}(t) = \mathbf{M}^{-1}\mathbf{p}(t)dt \tag{1.30}$$

$$d\mathbf{p}(t) = -\nabla U(\mathbf{q}(t))dt - \gamma\mathbf{M}^{-1}\mathbf{p}(t)dt + \sqrt{\frac{2\gamma}{\beta}}dW(t) \tag{1.31}$$

where $\mathbf{M}$ is the mass matrix, γ is the frictional coefficient, $\beta = 1/k_{\mathrm{B}}T$ and W(t) is a Wiener process that underlies the random force (white-noise). An important point to underline is that, if the damping force is very soft and the stochastic forces are small, equation 1.31 could be considered as a perturbation of Newtonian dynamics.

Langevin thermostat

In this section, Langevin thermostat will be described carefully. As mentioned repeatedly, when sampling a canonical ensemble, it is essential to maintain a constant temperature, and this is achieved through the use of thermostats. There are numerous algorithms available for implementing thermostats in molecular dynamics simulations, and while it is beyond the scope of this work to cover all of them in detail, the focus will be on the application of Langevin dynamics in conjunction with an appropriate thermostat.
In 1978, Schneider and Toll [90] were the first to demonstrate, under certain significant assumptions, that Langevin dynamics could effectively conserve temperature during the integration of equations 1.30 and 1.31. To satisfy the dissipation-fluctuation theorem, it is often assumed that the stochastic force follows a Gaussian distribution and possesses specific statistical properties [90]

$$\langle W(t)\rangle = 0 \tag{1.32}$$

$$\langle W(t)W(t')\rangle = 2\gamma k_{\mathrm{B}}T\delta(t-t') \tag{1.33}$$

Developing accurate integrators for Langevin dynamics poses challenges due to the non-analytical nature of the stochastic force, rendering traditional methods like the Verlet scheme (discussed in section 1.1.2) invalid due to the inability to perform Taylor expansions [91].
The initial step was to demonstrate that if a system evolves according to Langevin equations of motion, the energy can be nearly conserved with an appropriate choice of the friction coefficient. From the equations above, a collisional time τ_c for the particles in the system with the heat bath can be derived, and it is proportional to $1/\gamma$. In the original 1978 derivation [90], it was established that the energy of the system can be considered nearly conserved if the collisional time is much larger than the characteristic timescales of single-particle dynamics. This condition helps prevent the dynamics from becoming overdamped. Additionally, another condition was imposed on the collisional time τ_c, to be much smaller than the characteristic timescales associated with any system transitions. To quantify this explanation more precisely, it will be explored the revised and more detailed derivation presented by Bussi and Parrinello [92]. They introduced a quantity $\Delta\hat{E}$ known as the effective energy increment, which can be expressed for the i-th time step as [92]

$$\begin{aligned}\Delta\hat{E} = \sum_{n=1}^{3N}\Bigg[&(q_n(t_{i+1}) - q_n(t_i))\left(\frac{f(q_n(t_{i+1})) + f(q_n(t_i))}{2}\right) + \\ &+ (U(q_n(t_{i+1}) - U(q_n(t_i))) + \frac{\Delta t^2}{8m_n}\left(f^2(q_n(t_{i+1}) - f^2(q_n(t_i))\right)\Bigg]\end{aligned} \tag{1.34}$$

where $f(q_n(t_i))$ is the force at i-th time step acting on the n-th coordinate of the system, $U(q_n(t_i))$ is the potential at i-th time step acting on the n-th coordinate of the system and Δt is the time step. Once explicitly presented, it could be rewritten in a more fashion formulation but in a more compact notation

$$\Delta\hat{E} = \Delta q\left(\frac{f(q_{i+1}) + f(q_i)}{2}\right) + \Delta U + \frac{\Delta t^2}{8m}\Delta(f^2) \tag{1.35}$$

where it can be seen that when Δt is small enough the effective energy is approximately constant, since the first and second terms tend to compensate each other and the third term vanishes on the order of Δt^2.
Finally, an implementation of the Langevin dynamics in the Verlet scheme will be presented, with one of the examples that proved its validity to the scientific community, deciding to implement the algorithm in the major simulation packages. Starting with the continuous-time Langevin equations and integrating equation 1.31 over a (small) time interval $dt = \Delta t$ between t_i and t_{i+1} and, for sake of simplicity, on just the n-th particle of the system

$$\int_{p_i}^{p_{i+1}} dp'_n = \int_{t_i}^{t_{i+1}} -\nabla U_n\, dt' - \int_{t_i}^{t_{i+1}} \gamma\frac{dq_n}{dt'}\, dt' + \sqrt{\frac{2\gamma}{\beta}}\int_{t_i}^{t_{i+1}} W(t')\, dt' \tag{1.36}$$

that, without approximations and dropping the index n indicating the single particle on which it is computed, can be rewritten as

$$m(v^{i+1} - v^i) = \int_{t_i}^{t_{i+1}} f\, dt' - \gamma(q^{i+1} - q^i) + \eta^{i+1} \tag{1.37}$$

where it has been defined $f_n \equiv -\nabla U_n$ and

$$\eta^{i+1} \equiv \sqrt{\frac{2\gamma}{\beta}}\int_{t_i}^{t_{i+1}} W(t')\, dt' \tag{1.38}$$

is a Gaussian random number with $\langle\eta^i\rangle = 0$ and $\langle\eta^i\eta^j\rangle = 2\gamma k_{\mathrm{B}}T\Delta t\delta_{i,j}$. In a similar way equation 1.30 could be integrated to give

$$q^{i+1} - q^i = \int_{t_i}^{t_{i+1}} v\, dt' \approx \frac{\Delta t}{2}(v^{i+1} + v^i) \tag{1.39}$$

where equations 1.37 and 1.39 should remind of a starting point to introduce a Verlet scheme. In fact, inserting v^{n+1} from 1.37 into 1.39 and approximating the integral of the deterministic force f such that both equations are correct to second order in Δt[91]:

$$q^{i+1} = q^i + b\Delta t v^i + \frac{b\Delta t^2}{2m}(f^{i+1} + f^i) + \frac{b\Delta t}{2m}\eta^{i+1} \tag{1.40}$$

$$v^{i+1} = v^i + \frac{\Delta t}{2m}(f^{i+1} + f^i) - \frac{\gamma}{m}(q^{i+1} - q^i) + \frac{1}{m}\eta^{i+1} \tag{1.41}$$

where $b \equiv 1/(1+\frac{\gamma\Delta t}{2m})$. In this regard, equations 1.40 and 1.41 constitute a simple functional Verlet-type scheme for solving stochastic Langevin equations.
An easy example on how reliable could be the method just introduced, is the ability of reproducing the correct results for the case of thermal diffusion with f=0. Equation 1.41 repeated n times, introducing $\alpha \equiv \frac{1-\frac{\gamma\Delta t}{2m}}{1+\frac{\gamma\Delta t}{2m}}$ results [91]

$$v^i = a^i v^0 + \frac{b}{m}\sum_{k=0}^{i-1} a^k\eta^{i-k} = a^i v^0 + \frac{b\sqrt{2\gamma}}{m\sqrt{\beta}}\sum_{k=0}^{i-1} a^k\sigma^{i-k} \tag{1.42}$$

where σ is a standard Gaussian random number with mean equal to 0 and variance equal to 1. Summing over the random numbers in the above equation yields another Gaussian random number, and substituting b it is possible to have

$$v^i = a^i v^0 + \sqrt{1 - a^{2i}} \sqrt{\frac{1}{\beta m}} \sigma \tag{1.43}$$

For large n, $a^n \ll 1$, and it is immediate to find that the velocity is characterized by a Maxwell-Boltzmann distribution with zero mean and

$$\langle (v^i)^2 \rangle = \frac{k_\mathrm{B} T}{m} \tag{1.44}$$

which results in reproducing the exact expectation for the average kinetic energy (thermal energy)

$$E_k = \frac{1}{2} m \langle (v^i)^2 \rangle = \frac{1}{2} k_\mathrm{B} T \tag{1.45}$$

Is it also possible find a value for the simulated diffusion coefficient taking the equation 1.40 where, with similar steps done above for the velocities, one recovers for the positions

$$q^i = q^0 + \frac{m}{\gamma} v^i + \sqrt{i \Delta t \frac{2 k_\mathrm{B} T}{\gamma}} \sigma \tag{1.46}$$

Finally, the diffusion coefficient can be computed as

$$D = \lim_{i \Delta t \to \infty} \frac{\langle q^i - q^0 \rangle^2}{2 i \Delta t} = \frac{k_\mathrm{B} T}{\gamma} \tag{1.47}$$

which agrees with Einstein's fluctuation-dissipation relationship for dynamics in presence of a friction coefficient. In the study conducted by Gronbech et al. [91], they provide additional examples where their Verlet scheme for Langevin dynamics performs effectively, accurately predicting known results for the simple models under investigation. This Verlet-based algorithm represents a starting point for algorithms implementing Langevin dynamics in major simulation programs.

1.2 Accelerate Rare Events

Molecular dynamics simulations, as described in section 1.1, have become an invaluable tool to explore complex systems as crystal nucleation, fullerene formation, protein folding, or prebiotic reactions forming complex molecules. These transformations, characterized by metastable structures separated by energy barriers, involve activated processes where the system transitions over time between states. The exploration of these phenomena, often termed the "science of change", heavily relies on computational methods due to experimental constraints. The limitations of MD simulations become particularly evident in the study of biological molecules, which frequently exhibit rugged energy landscapes with multiple local minima and formidable energy barriers. Traditional simulations may lead to prolonged entrapment in non-functional states, impeding the characterization of dynamic behavior.
Assuming a specific MD simulation algorithm that samples a designated ensemble as described in section 1.1.1, the chosen algorithm must satisfy two crucial criteria: effective sampling of the chosen distribution and ergodicity (eq. 1.2), ensuring eventual coverage of the entire configuration space. However, the inherent correlation between samples and the time required to approach ergodic behavior may pose challenges, potentially extending beyond the practical time scale of computer simulations. This challenge underscores the necessity for enhanced sampling methods to efficiently surmount energy barriers, ensuring a more thorough exploration of conformational space.
MD simulations are known generally to suffer of three main limitations [93]:

- Accuracy with respect to the underlying model (force field as described in sec. 1.1) (higher level of theory as QM/MM, semi-empirical, *ab initio* can lead to more accurate results).
- Huge dimensionality of the output of the system trajectories, methods and algorithms need to be developed to understand the complexity of the system under study by lowering the number of variables to better interpret the results.
- Due to the constraints imposed by the necessity for a small timestep to maintain stable and accurate integration, the temporal scales amenable to sampling in molecular dynamics simulations frequently fall short of the researcher's target processes.

The second challenge, involving the identification of a low-dimensional projection for interpreting MD simulations, is intricately connected to several algorithms discussed herein, falling under the overarching category of "collective variable (CV)-based methods" will be discussed briefly in the next paragraph. The configurational ensemble of the systems of interest typically exhibits pronounced peaks, featuring multiple metastable and well-defined states with high probabilities. Conversely, the regions between these states often have probabilities approaching zero in the complete high-dimensional Cartesian space. This clarifies why strategies reducing the dimensionality of the sampled space by projecting it onto a lower-dimensional surface can prove successful, especially when they expedite sampling along a pertinent CV linking metastable states.
This section is mainly dedicated to describe the importance of approaches proposed to tackle the third challenge, commonly encompassed under the umbrella term "enhanced sampling MD simulation" techniques and in particular with a focus on the method named metadynamics. It is worth noting that the realm of the methods developed to ensure a fast and correct sampling of the configurational space is huge and goes far beyond the scope of this section, while other comprehensive reviews on this subject have been written, such as those referenced in [94, 95, 96].

1.2.1 Collective variables

In a system of N atoms with coordinates $\mathbf{r} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ it is possible to define collection of d collectively intuitive variables $\mathbf{q}(\mathbf{r}) = \{q_1(\mathbf{r}), \ldots, q_d(\mathbf{r})\} \in \mathbb{R}^d$ representing a coarse-grained depiction of the system, where $q_i(\mathbf{r})$ represents the i-th collective variable. It can be defined a function ξ from $\mathbb{R}^{3N} \to \mathbb{R}^d$ mapping configurations $\mathbf{r}$ to a lower-dimensional representation, this process is often called in literature dimensionality reduction. When setting $\mathbf{q}(\mathbf{r}) = \mathbf{s}$, a partition function at temperature T analogous to eq. 1.4, can be formulated as

$$Q(\mathbf{s}) = C \int d\mathbf{r} e^{-\beta U(\mathbf{r})} \delta(\mathbf{q}(\mathbf{r}) - \mathbf{s}) \tag{1.48}$$

where C is a constant independent of $\mathbf{s}$, $\beta = 1/k_{\mathrm{B}}T$, and $U(\mathbf{r})$ is the potential energy function describing all the system's interactions. Subsequently a free energy surface (FES) $F(s)$ can be defined as

$$F(\mathbf{s}) = -k_{\mathrm{B}} T ln(Q(\mathbf{s})) + \tilde{C} \tag{1.49}$$

FES, as a simplified model of complex systems, should effectively capture key macroscopic states in molecules and materials. This encompasses stable molecular configurations and phase changes of materials. CVs function as tools for encapsulating long-lasting collective behaviors, effectively differentiating macroscopic variations in structure or aggregation within a system. Their essential attribute is the capacity to distinguish between metastable states, thereby clarifying transitions in rare events and the slower dynamics or kinetics associated with these transitions. The term "slow CVs" refers to those CVs adept at tracking these prolonged processes. These slow CVs are crucial not just for understanding the kinetics of infrequent events, but also for deriving vital insights for analysis, particularly when aligning with experimental data. Reducing the description of a high-dimensional system to a handful of coordinates implies potential renormalization across the other coordinates. This indicates that these other coordinates either average out over quicker timescales relative to the resolved CVs, or they do not significantly influence the process in focus. Slow CVs are crucial in improving sampling efficiency. By introducing a bias to the system's energy function, they increase the likelihood of accessing less frequently visited states.
Historically researchers have been employed different kind of physically intuitive CVs that can be listed as

- Euclidean distances: These measure the distances between different molecular conformations and can be used to quantify how slowly these conformations interconvert.
- Angles: Angles between atoms or groups of atoms can be used as collective variables to capture specific structural changes or conformational transitions.
- Interatomic distances: The distances between specific pairs of atoms or groups of atoms can provide insights into the interactions and dynamics of the system.
- Formation of specific contacts: Collective variables can be defined based on the formation or breaking of specific contacts between atoms or groups of atoms, allowing for the study of binding/unbinding processes.
- Ramachandran angles: These angles describe the backbone dihedral angles in proteins and can be used to study protein folding and conformational changes.
- Solvent accessible surface area: This collective variable measures the surface area of a molecule that is accessible to solvent molecules and can be used to study protein-ligand interactions and binding events.

While physical intuition remains a cornerstone for scientists [97, 98, 99], the vast amount and dimensionality of generated data necessitate automatic methods for extracting meaningful CVs from simulation data. The goal is to find CVs in a principled and systematic manner, minimizing ambiguity in the interpretation of simulation results and facilitating a direct connection to experimental observations. A rigorous theory has been formulated, asserting that optimal slow CVs are represented by the eigenfunctions of the dynamical operator underlying MD simulations [100],

$$\mathcal{T}\phi_i = \lambda_i \phi_i \tag{1.50}$$

where $\mathcal{T}$ is the transfer operator that propagates the probability densities of molecular configurations, ϕ_i are its eigenfunctions, and λ_i are the associated eigenvalues. Though exact computation is generally unfeasible, recent advancements involve estimating these eigenfunctions from MD simulation data. One possibility is to use a variational principle applied to the ϕ_i eigenfunctions that maximally uncorrelate the ones with the biggest eigenvalues associated [101] or by using a time-lagged independent component analysis (TICA) applied to system coordinates in order to maximize the covariance matrix in particular states [102]. It is worth noting that in the last few years the development of machine learning (ML) and especially deep learning (DL) algorithms paved the way to the possibility of describe the complexity of molecular systems beyond the number of CVs (usually ≤ 3) that researchers can design. The literature on this field is growing exponentially [103] and treat all the methods goes far beyond the scope of this section, however a brief introduction and a list of the most promising methods will be given below.
ML and DL methods involved in finding optimal CVs usually needs two ingredients. The first one is a model, this could be linear, not linear and even a neural network (NN) that takes as input the cartesian coordinates of the system $\mathbf{r}$, or eventually a subset of them, and has a number of CVs $\mathbf{q}$ as output. The second ingredient would be a loss function in which the related model tries to optimize (usually minimizing the loss function) the goal of the model that in general would be to describe the greater possible variability of the system with the lower amount of CVs. Among them there are

- Principle Component Analysis (PCA): PCA is a dimensionality reduction algorithm that decomposes the sample covariance matrix to find directions with large variance. It can be used to train linear combinations of coordinates as CVs.

- Sketch Map: This method trains CVs based on preserving structural similarity. It uses non-linear distance matching to preserve clustering information in the Cartesian coordinates of the system [104].

- Autoencoder: Within the architecture of an autoencoder model, it comprises both an encoder and a decoder. The encoder, represented by the function f_e, is responsible for transforming high-dimensional inputs $\mathbf{r}$ into lower-dimensional latent variables $\mathbf{q}$. Conversely, the decoder, denoted as f_d, reconstructs $\mathbf{q}$ back into high-dimensional outputs $\mathbf{r}'$ [105].

- Deep Linear Discriminant Analysis (deep-LDA): LDA was proposed initially as a classification model and here a NN finds the best $\mathbf{q}_i$ that maximize the discrimination between systems' metastable states (two or more) given in input molecular descriptors recorded during MD runs in different basins [106].

1.2.2 Enhanced sampling and free energy methods

This section is intended to briefly introduce the field of enhanced sampling in MD simulations in order to have all the theoretical tools to dig deeper into metadynamics that would be the focus of the next section. This short review on enhanced sampling methods does not want to be exhaustive, for which the literature is abundant and evolving continuously [93, 107, 108].
In this section the focus lies on presenting techniques for expediting the sampling of a specified equilibrium probability distribution. The present dissertation will refrain from delving into purely exploratory methods that lack applicability in recovering equilibrium statistics and will not cover methodologies primarily designed for characterizing kinetic rates, for which the reader is referred to section 1.3. Finally, pathway-based methods facilitating the retrieval of statistical ensembles will not be treated but it is worth mentioning that, among them, Transition Path Sampling (TPS) was one of the most studied and successfully applied and worth a couple of sentences.
TPS directs its attention to the pivotal segment of a simulation, namely, the transition. For instance, in the context of a protein undergoing folding, TPS aims to precisely replicate those critical folding moments. When dealing with a system characterized by two stable states, A and B, the system typically spends extended periods in these states and occasionally transitions between them. TPS assigns probabilities to various transition pathways, enabling the construction of a Monte Carlo random walk in the path space of these trajectories. This process generates an ensemble encompassing all transition paths, from which crucial information like reaction mechanisms, transition states, and rate constants can be extracted [109]. The shooting move, an influential and efficient algorithm, is a cornerstone of TPS. Consider a classical many-body system described by coordinates $\mathbf{r}$ and momenta $\mathbf{p}$. Molecular dynamics generates a path as a set of (r_t, p_t) at discrete times t in [0,T], where T is the length of the path. For a transition from A to B, $(\mathbf{r}_0, \mathbf{p}_0)$ lies in A, and $(\mathbf{r}_T, \mathbf{p}_T)$ lies in B. Randomly select one of the path times, and slightly modify the momenta $\mathbf{p}$ to $\mathbf{p} + \delta\mathbf{p}$, where $\delta\mathbf{p}$ is a random perturbation consistent with system constraints, such as the conservation of energy and linear and angular momentum. A new trajectory is then simulated from this point, both backward and forward in time until one of the states is reached. Given the transition region's nature, this process is relatively quick. If the new path still connects A to B, it is accepted; otherwise, it is rejected, and the procedure repeats. This iterative process gradually samples the ensemble through accepted paths [110].

Thermodynamic Integration

Among the methods going directly for the free energy estimation it is impossible not to cite thermodynamic integration (TI). TI encompasses a range of methods for estimating free energy, expressing the derivative of free energy concerning a continuous parameter as an ensemble average of the energy's derivative with respect to the same parameter. In cases where the reduced potential energy is a function u_λ of a smooth coupling parameter λ, linking all relevant states, a continuous free energy $f(\lambda)$ is defined. The derivative of this free energy, often termed the "mean force" at a fixed λ, is expressed as:

$$\frac{df}{d\lambda} = \langle \frac{\partial u_\lambda}{\partial \lambda} \rangle_\lambda \tag{1.51}$$

Here, the averaging is performed over samples acquired with the specified value of λ. For the i-th state and j-th state, corresponding to values λ_i and λ_j of the coupling parameter,

the free energy difference Δf_{ij} between these states is given by [93]

$$\Delta f_{ij} = \int_{\lambda_i}^{\lambda_j} d\lambda \langle \frac{\partial u_\lambda}{\partial \lambda} \rangle_\lambda \tag{1.52}$$

When the i-th state and j-th state are nearby along λ with a sufficiently small difference $\Delta\lambda_{ij}$, the integral can be resolved using well-known methods for numerical integration.

Out of Equilibrium Methods

Out-of-equilibrium driven approaches involve imposing a predefined trajectory for a collective variable or an alchemical parameter λ to navigate configuration space. This category of methods results in simulations where the original distribution is altered, and equilibrium ensemble convergence is not achieved. Under specific conditions the equilibrium distribution can be restored. The trajectory may unfold rapidly, causing the system, even if initially at equilibrium, to deviate from equilibrium.
Out-of-equilibrium pulling exhibits different behavior depending on the transformation rate. In the limit of infinitely slow (quasistatic) transformation, all orthogonal degrees of freedom are continuously relaxed, recovering equilibrium properties. This is exemplified by the "slow growth" approach [111], employing very gradual switching from energy U_A to U_B. The work W performed along this path approximates the reversible work, representing the free energy difference from A to B. The other possible limit has infinitely fast switching, comparing the energy of a configuration for two different Hamiltonians without any relaxation, utilizing a Free Energy Perturbation (FEP) approach [112]. In intermediate cases, the free energy difference can be estimated by weighting non-equilibrium trajectories [113]. Denoting W_λ as the total non-equilibrium work exerted by the bias over a trajectory up to a value λ, the Jarzynski identity states

$$e^{-\beta(F_\lambda - F_0)} = \langle e^{-\beta W_\lambda} \rangle_E \tag{1.53}$$

Here, the average is taken over the equilibrium ensemble. Out-of-equilibrium methods are not commonly employed for estimating free energy differences due to the high variance of the exponential averaging free energy estimator. The notable variance arises from the fact that the work values contributing the most to the average have small probabilities [93].

Adaptive Bias

Numerous methods have been developed falling into this category, however here just adaptive biasing force (ABF) and variationally enhanced sampling (VES) will be presented.
In ABF simulations [114], the free energy gradient with respect to selected collective variables $\mathbf{q} = \xi(\mathbf{r})$ is estimated. This estimate is then utilized to apply a time-dependent external force $F_t^{\mathrm{ABF}}(\mathbf{q})$ that opposes the estimated free energy gradient, facilitating enhanced sampling along those coordinates. Usually it can be adopted the following scheme:

1. Choose a small number of collective variables, $\mathbf{q} = \xi(\mathbf{r})$.

2. Estimate the gradient of the free energy surface as: $\nabla_{\mathbf{q}} A(\mathbf{q}) = -\langle F_\xi(\mathbf{r}) \rangle_{\xi(\mathbf{r})=\mathbf{q}}$.

3. Once enough sampling is accumulated to provide a reliable estimate of the average force at the current position in the CV $\mathbf{q}$, apply a biasing force $F_t^{\mathrm{ABF}}(\mathbf{q})$ equal to the opposite of this average force.

4. This force, on average, nullifies the forces acting along $\mathbf{q}$, effectively smoothing the free energy barriers and fastening the diffusion in the CV space.

5. Upon convergence, the biasing force becomes the negative of the gradient of the free energy, rendering $\mathbf{q}$ subject to a flat effective free energy landscape. Numerical integration methods can be employed to estimate the FES directly from the gradient.

The heart of ABF lies in the fact that a time-dependent biasing force $F_t^{\mathrm{ABF}}(\mathbf{q})$ converges towards the free energy gradient $\nabla_{\mathbf{q}} A$ for extended sampling times.
VES is a more recently introduced method and it is based on a variational principle [115]. In VES, a bias potential $U^{bias}(\mathbf{q})$ is formulated by minimizing a convex functional [93]:

$$\Omega[U^{bias}] = \frac{1}{\beta} \ln \frac{\int d\mathbf{q} e^{-\beta A(\mathbf{q}) + U^{bias}(\mathbf{q})}}{\int d\mathbf{q} e^{-\beta A(\mathbf{q})}} + \int d\mathbf{q} p_{tg}(\mathbf{q}) U^{bias}(\mathbf{q}) \tag{1.54}$$

where $p_{tg}(\mathbf{q})$ represents a user-chosen target distribution. The functional $\Omega[U^{bias}]$ has a global minimum given by

$$U^{bias}(\mathbf{q}) = -A(\mathbf{q}) - \frac{1}{\beta} \ln p_{tg}(\mathbf{q}) + C \tag{1.55}$$

where C is a meaningless constant. This bias potential yields a biased CV distribution that matches the target distribution $\tilde{\rho}(\mathbf{q}) = p_{tg}(\mathbf{q})$. Therefore, the selected target distribution directly influences the sampling of CVs during the minimization of $\Omega[U^{bias}]$ [93]. Thus, by opting for a target distribution that simplifies sampling compared to the equilibrium distribution, the efficiency of CV sampling is improved. Additionally, the FES can be readily obtained from the bias potential, as detailed in equation 1.55.

Generalized ensemble and replica exchange methods

A wide range of simulation methodologies falling under the umbrella of generalized ensemble [116], or sometimes referred to as extended ensemble [117], algorithms has gained popularity in the last two decades. These approaches adopt a strategy that is distinct from the methods discussed thus far. In this category of methods, the original configurational distribution is maintained, and sampling is improved by leveraging transitions to other ensembles. The primary algorithmic classes within this category include replica exchange [118], which encompasses parallel tempering [119] and Hamiltonian exchange [118]. Here replica exchange will be briefly introduced.
In a replica exchange simulation, K separate simulations are conducted, each in one of the K thermodynamic states. The current state of the replica exchange simulation is represented by (X, S), where X is a vector of configurations for all replicas, denoted as $X \equiv \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$, and $S \equiv \{s_1, \ldots, s_K\} \in \mathcal{S}_K$ is the set of state labels $\{1, \ldots, K\}$ associated with each replica configuration $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. The joint probability density of the entire set of simulations Q is given by [93]:

$$Q(X, S) \propto \prod_{i=1}^{K} \nu_{s_i}(\mathbf{x}_i) \propto \exp\left(- \sum_{i=1}^{K} u_{s_i}(\mathbf{x}_i) \right) \tag{1.56}$$

where μ_{s_i} and u_{s_i} are the normalized probability distributions and reduced energies of state s_i, respectively. The conditional densities, given a specific order of the replicas S, are:

$$Q(X|S) = \prod_{i=1}^{K} \frac{e^{-u_{s_i}(\mathbf{x}_i)}}{\int e^{-u_{s_i}(\mathbf{x}_i)} \, dx_i} \tag{1.57}$$

and

$$Q(S|X) = \exp\left(-\sum_{i=1}^{K} u_{s_i}(\mathbf{x}_i)\right) \Big/ \sum_{S' \in \mathcal{S}_K} \exp\left(-\sum_{i=1}^{K} u_{s'_i}(\mathbf{x}_i)\right) \tag{1.58}$$

These equations are more complex than those for a single simulation as they describe the state of all replicas, which are coupled together. With the two above equations, one can describe jumps in coordinate space and jumps replicas.
In conventional replica exchange simulation techniques, when considering a new permutation S for the state of the entire system set (X, S), exchanges are typically limited to states that are immediate neighbors [119]. A usual approach is to try exchanging either pairs of states like $\{(1,2),(3,4),\ldots\}$ or $\{(2,3),(4,5),\ldots\}$, with each pair being selected with the same likelihood. For each pair (i,j), the exchange of states i and j, corresponding to configurations x_i and x_j, is processed independently. The approval of this exchange follows the standard Metropolis probability criteria. [93]:

$$P_{\text{accept}}(\mathbf{x}_i, i, \mathbf{x}_j, j) = \min\left(1, \frac{e^{-(u_i(\mathbf{x}_j)+u_j(\mathbf{x}_i))}}{e^{-(u_i(\mathbf{x}_i)+u_j(\mathbf{x}_j))}}\right) \tag{1.59}$$

1.2.3 Metadynamics

In this section it will be given a focus on metadynamics [120], a successfully and widely adopted methodology in the field of enhanced sampling schemes that fell in the adaptive bias sampling algorithms briefly mentioned in the previous section.
In the metadynamics approach, an external bias potential that depends on its history and is a function of CVs, as detailed in sec. 1.2.1, is introduced to the system's Hamiltonian. This potential is constructed as a combination of Gaussians distributed along the trajectory in the CV space $\mathbf{s} = \mathbf{q}(\mathbf{r})$. The purpose is to discourage the system from revisiting configurations that have been sampled before.
At time t, the metadynamics bias potential can be expressed as:

$$V_G(\mathbf{s}, t) = \int_0^t dt' \omega \exp\left(-\sum_{i=1}^{d} \frac{(\mathbf{q}_i - \mathbf{q}_i(t'))^2}{2\sigma_i^2}\right) \tag{1.60}$$

where ω is an energy rate, and σ_i is the width of the Gaussian for the i-th CV. The energy rate ω is typically constant and often expressed in terms of a Gaussian height W and a deposition stride τ_G:

$$\omega = \frac{W}{\tau_G} \tag{1.61}$$

Apply metadynamics algorithm offers several advantages:

- Acceleration of sampling: Metadynamics accelerates the sampling of rare events by steering the system away from local free-energy minima.
- Exploration of new pathways: It enables the exploration of new reaction pathways as the system tends to escape minima, passing through the lowest free-energy saddle point.
- No a priori knowledge required: Unlike umbrella sampling, metadynamics inherently explores low free-energy regions first without the need for prior knowledge of the landscape.

- Unbiased free energy estimate: After a transient period, the bias potential V_G provides an unbiased estimate of the underlying free energy:

$$V_G(\mathbf{s}, t \to \infty) = -F(\mathbf{s}) + C \tag{1.62}$$

where C is an irrelevant additive constant, and the free energy $F(\mathbf{s})$ is defined as:

$$F(\mathbf{s}) = -\frac{1}{\beta} \ln \left(\int d\mathbf{r} \delta(\mathbf{q}(\mathbf{r}) - \mathbf{s}) e^{-\beta U(\mathbf{r})} \right) \tag{1.63}$$

where $\beta = (k_\mathrm{B}T)^{-1}$, k_B is the Boltzmann constant, T is the temperature, and $U(\mathbf{r})$ is the potential energy function.
The correctness of the relation in eq. 1.62 has been empirically demonstrated through extensive testing on simplified models and comparison with other free-energy methods on complex systems. Moreover, a formal proof of eq. 1.62 has been provided under the assumption that, in the absence of bias, the stochastic dynamics in the CVs space is memoryless. Under the same assumption, the error in the FES reconstruction has been proven, both empirically and theoretically, to be [121]

$$\varepsilon \propto \sqrt{\frac{\omega}{D(\mathbf{s})\beta}} \tag{1.64}$$

where $D(\mathbf{s})$ is the intrinsic system diffusion coefficient in the CVs space. The practical application of this equation may be challenging, and in most studies, the error in the free-energy profile is estimated through the comparison of independent runs [122] or block averaging [123].
Within this approach a couple of major drawbacks arise:

1. The bias is pushing the system at exploring high energies configurations, thus convergence in the desired region of CV space is not guaranteed and not automatically granted.

2. Being a method falling in the CV based algorithm, identify a set of appropriate CVs describing complex processes could be far from being easy.

However, regarding the first problem a solution was proposed through well-tempered metadynamics (WT-MetaD) [124]. In WT-MetaD the bias deposition rate diminishes over simulation time, achieved through a different expression for the bias potential with respect to 1.62,

$$V(\mathbf{s}, t) = k_\mathrm{B}\Delta T \ln \left(1 + \frac{\omega N(\mathbf{s}, t)}{k_\mathrm{B}\Delta T} \right) \tag{1.65}$$

where $N(\mathbf{s}, t)$ is the histogram of the $\mathbf{s}$ CVs collected during the WT-MetaD simulation and ω is an input parameter with the dimension of temperature. This formulation can be easily applied by rescaling the Gaussian height W as

$$W = \omega \tau_G e^{-\frac{V_G(\mathbf{s},t)}{k_\mathrm{B}\Delta T}} \tag{1.66}$$

It is possible to note a couple of main differences with standard metadynamics. Firstly, the deposition rate of the bias decreases as $1/t$. Secondly, the bias potential does not fully fill the free energy surface, but it converges to [125]

$$V_G(\mathbf{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\mathbf{s}) + C \tag{1.67}$$

where C is a constant that does not intervene in any physical property of the system. With respect to standard metadynamics, here the algorithm assures that the bias converges to its limiting value in a single run. For time approaching to infinity, the probability distribution of $\mathbf{s}$ becomes [125]

$$P(\mathbf{s}) \propto e^{-\frac{F(\mathbf{s})}{k_B(T+\Delta T)}} \tag{1.68}$$

Where it is easy to see that for $\Delta T \to 0$, ordinary MD is recovered, while the limit $\Delta T \to \infty$ corresponds to standard metadynamics. Thus, tuning ΔT allows the regulation of the extent of FES exploration, avoiding excessive accumulation of the bias and potentially conserving computational resources, particularly when employing a large number of CVs.
Following the introduction of the primary MetaD algorithm and its widely utilized variant, WT-MetaD, the subsequent paragraphs will delve into various common practices and widely adopted configurations. These discussions are valuable not only for enhancing the efficiency of MetaD but also because they have been incorporated into many of the studies presented in this thesis.
A common algorithmic optimization employed in MetaD simulations involves accumulating the bias on a grid [126]. Although initially developed primarily for ABF, this procedure can be readily adapted for MetaD. In lengthy simulations lasting tens or hundreds of nanoseconds, the computation of the deposited bias (equation 1.60) at every integration timestep becomes computationally intensive, leading to potential slowdowns. To address this issue, the bias is accumulated on a multidimensional grid $\boldsymbol{\gamma}$ with dimensions $(N_1 \times \ldots \times N_d)$, where N_i represents the user-defined discretization along each of the d CVs introduced. Let Ω denote the CV space; this approach ensures that the system experiences bias based on its position relative to the CV space at a given time during the MetaD simulation, denoted as $V_G^{\Delta\Omega}((\Delta q_1, \ldots, \Delta q_d), t')$, where $\Delta\Omega$ represents the corresponding discretized hypercube of the CV space in which $\mathbf{s}$ is located at time t'. The grid $\boldsymbol{\gamma}$ is typically updated every τ_γ timesteps, during which only the newly added Gaussian kernels are incorporated into the grid. This optimization accelerates both the reading and application of the bias to the simulation.
Another crucial aspect that needs to be analyzed concerns the width of the Gaussian kernels in Eq. 1.60. In this equation, the width of the Gaussian in each CV direction can vary, but it is fixed over time. This is represented by σ_i in the Gaussian kernel. However, a more detailed description should consider that the Gaussian kernels are generally multivariate Gaussian functions, and the width is derived from a variance matrix σ_{ij} that was assumed to be diagonal in the previous explanation. There are two choices regarding the covariance σ_{ij}. The first choice is known as dynamically-adapted Gaussians [127] (also called diffusion-adaptive), and the covariance can be expressed as:

$$\sigma_{ij}^2(t) = \frac{1}{\tau_D} \int_0^t dt' \, [\mathbf{s}_i(t') - \bar{\mathbf{s}}_i(t')][\mathbf{s}_j(t') - \bar{\mathbf{s}}_j(t')] e^{-(t-t')/\tau_D} \tag{1.69}$$

where τ_D is a free parameter that determines how long to record variations in the i-th and j-th directions to adapt the covariance σ_{ij} at the moment of kernel deposition. This procedure is applied every τ_D. The second approach is referred to as geometry-adapted Gaussians [127], and the covariance can be expressed as:

$$\sigma_{ij}^2(\mathbf{s}) = \sigma_G^2 \sum_\alpha \frac{\partial q_i}{\partial r_\alpha} \frac{\partial q_j}{\partial r_\alpha} \tag{1.70}$$

where σ_G is a free parameter carefully chosen to encompass the oscillations of the CVs, and in this case, the covariance σ_{ij} is adjusted based on the atomic displacement of the system

$\partial \mathbf{r}$ involved in the CVs $\mathbf{s}$.
It was stated previously that WT-MetaD has been developed to solve a critical problem as the convergence of the deposited bias in a specific region of the CVs space. Another possibility is to employ *ad hoc* boundary treatments. One is a classic harmonic restraint of the form

$$U^{rest}(\mathbf{s}) = \frac{1}{2}\sum_i k_i(q_i(t) - \tilde{q}_i) \tag{1.71}$$

where k_i is the force constant restraining the i-th CV around the $\tilde{q}_i$ position. Another possibility is to employ the solution proposed in [128] where the metadynamics algorithm is setting the bias force to zero beyond a defined boundary. For instance, when conducting metadynamics on a CV and focusing solely on the free energy for $\mathbf{s} > \mathbf{s}_w$, the history-dependent potential is still updated according to Eq. 1.60. However, the bias force is then set to $\frac{dV_G}{\partial q} = 0$ for $\mathbf{s} < \mathbf{s}_w$. It is essential to note that the update to V_G implies the addition of Gaussians even when $\mathbf{s} < \mathbf{s}_w$. The tails of these Gaussians impact V_G in the relevant region (where $\mathbf{s} > \mathbf{s}_w$). Consequently, in the region $\mathbf{s} < \mathbf{s}_w$, the force on the system originates solely from the system's potential, while in the region $\mathbf{s} > \mathbf{s}_w$, it arises from both metadynamics and the system's potential.
It is crucial to highlight a highly potent variant of MetaD known as Multiple-Walkers Metadynamics (MW-MetaD) [129]. This approach underscores the capability of exploring the FES through MPI parallelization. While the fundamental theory underlying MetaD and WT-MetaD remains unchanged, a significant enhancement is introduced in the bias potential (Eq. 1.60), involving a summation over N_w walkers:

$$V_G(\mathbf{s}, t) = \sum_{k=1}^{N_w} \int_0^t dt' \omega \exp\left(-\sum_{i=1}^{d} \frac{(\mathbf{q}_{ik} - \mathbf{q}_{ik}(t'))^2}{2\sigma_i^2} \right) \tag{1.72}$$

Here, the width of Gaussian kernels σ is assumed to be constant over both time and space, maintaining consistency across different walkers for simplicity. It is noteworthy that associating each of the N_w walkers with a distinct replica of the biased simulation facilitates efficient sharing of information related to the deposited bias. Consequently, each walker experiences a bias that discourages exploration of regions already visited in the CV space by others. It is essential to acknowledge that while this implementation significantly accelerates the exploration of CVs, the speedup may not be linear [129]. However, it is crucial to recognize that, like other MetaD algorithms, MW-MetaD can still encounter common problems like the one involving convergence of the bias.
One important variation of MetaD is known as parallel-bias metadynamics (PB-MetaD) [130]. It addresses the challenge of maintaining a low number of CVs while efficiently sampling all slow processes in a system. Typically, in standard MetaD, due to computational and memory constraints, the number of CVs biased is usually less than 4. Although studies employing concurrent MetaD [131] have been conducted, applying more than one MetaD simultaneously to different CVs, the non-orthogonality of CVs can complicate retrieving the correct probability distribution. PB-MetaD aims to simplify this multidimensional problem into multiple monodimensional problems. Unlike concurrent MetaD, which applies all biases simultaneously, PB-MetaD applies only one bias at a time. If a new discrete variable $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ is introduced, the system samples the probability distribution:

$$P(\mathbf{s}, \boldsymbol{\eta}, t) \propto \exp\left(-\beta U(\mathbf{s}) + \sum_{i=1}^{d} \eta_i V_G(q_i, t) \right) \tag{1.73}$$

where $\boldsymbol{\eta}$ is a vector where, at each step, only one element can take the value of 1, and the remaining elements are 0. The probability distribution can be sampled using MD for the coordinate part and Monte Carlo for the discrete variable part $\boldsymbol{\eta}$. During PB-MetaD simulations, Gaussian kernels are added to the corresponding $V_G(q_i, t)$. Since no thermodynamic property of $\boldsymbol{\eta}$ is of interest, it is possible to marginalize along this variable [130]:

$$P(\mathbf{s}, t) = \sum_{i=1}^{d} P(\mathbf{s}, \eta_i, t) \propto \exp(-\beta U(\mathbf{s}) + V_{PB}(q_1, \ldots, q_d, t)) \tag{1.74}$$

where

$$V_{PB}(q_1, \ldots, q_d, t) = -\frac{1}{\beta} \ln \left(\exp \left(-\sum_{i=1}^{d} \beta V_G(q_i, t) \right) \right) \tag{1.75}$$

In many MD studies, identical atoms or molecules with the same physicochemical properties are common. To handle an ever-increasing number of CVs, these CVs can be grouped into families, and the PB-MetaD formalism can be applied [132]. This approach not only reduces the dimensional problem but also aids in exploring the bias profile. If the d CVs are partitioned into f families, each with a variable k members such that $\sum_{j=1}^{f} k_j = d$, then each k_j member of the corresponding family deposits the bias along the same partitioned CV. Moreover, the bias acts on all members of the same family, and just one free energy profile will be constructed for each partition family (PF) instead of one for each CV.
Finally, it is worth concluding this section with a very promising variant of MetaD named PTWT-MetaD [133]. This hybrid algorithm combines adaptive bias (WT-MetaD) and replica exchange (parallel tempering). The Metropolis probability in Eq. 1.59 would take into account also the exponential of the bias for the configurations under exchange. If accepted, the coordinates of replica i and j would be swapped, and the j-th replica would experience the temperature T_i and the bias $V_{G,i}$, and vice versa for the i-th replica. PTWT-MetaD benefits from the advantages of WT-MetaD regarding the sampling efficiency and on the other hand PT mitigates the impact of excluding a slow degree of freedom in selecting the CVs with respect to standard MetaD.

Reweighting and Statistics recovering

In this short section it would be described how to recover statistically correct physical quantities for a simulation where MetaD has been applied. It is worth noting that this discussion can be extended to other MetaD variants and also other enhanced sampling methods but this would be beyond the scope of this section.
When a bias is applied for a set of CVs $\mathbf{s}$, the equilibrium probability distribution attained equilibrium under the influence of the underlying potential and the bias applied as

$$P(\mathbf{s}, t) = e^{-\beta(V_G(\mathbf{s},t) - c(t))} \cdot P_0(\mathbf{s}) \tag{1.76}$$

where $P_0(\mathbf{s}$ is the unbiased Boltzmann probability density and c(t) is defined as [134]

$$c(t) = \frac{1}{\beta} \log \frac{\int d\mathbf{s} e^{-\beta F(\mathbf{s})}}{\int d\mathbf{s} e^{-\beta(F(\mathbf{s}) + V(\mathbf{s},t))}} \tag{1.77}$$

From equation 1.76 it is possible to see that for every operator $\mathcal{O}(\mathbf{r})$, it is possible to write its equilibrium average as

$$\mathcal{O}_0(\mathbf{r}) = \langle \mathcal{O}(\mathbf{r}) e^{\beta(V_G(\mathbf{s},t) - c(t))} \rangle \tag{1.78}$$

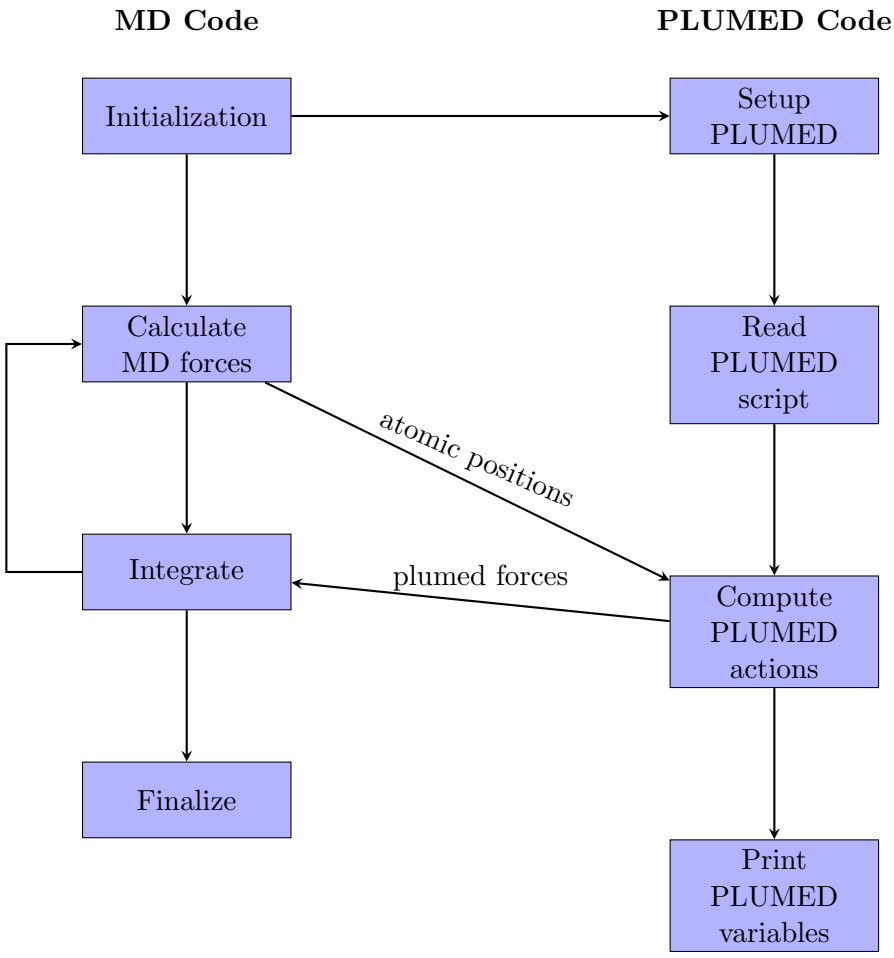

Figure 1.4: Schematic workflow of the interface between PLUMED software and a generic MD code.

It is noteworthy that Eq. 1.78 is intended for a generic operator that could depend on coordinates other than the ones biased. The focus then shifts to computing the function c(t). This can be computed explicitly between two deposition of Gaussian kernels τ_G through finite differences (indicated by the approximation symbol in the following equation) as [134]

$$e^{\beta c(t)} \approx \int d\mathbf{s} \left[e^{\gamma \frac{V(\mathbf{s},t+\Delta t)}{k_\mathrm{B}\Delta T}} - e^{\gamma \frac{V(\mathbf{s},t)}{k_\mathrm{B}\Delta T}} \right] \tag{1.79}$$

with $\gamma = \frac{(T+\Delta T)}{T}$ being the factor introduced in Eq. 1.67.

The PLUMED Software

The theoretical framework concerning CVs and biases has been translated into algorithms within the community-developed software PLUMED [135]. This code is implemented in C++ and has been parallelized using MPI. In Figure 1.4, the interaction between the PLUMED code and the MD engine is illustrated. Notably, PLUMED has also been interfaced with codes that perform *ab initio* MD, although the interface operates at the same level as pure MD codes.

The PLUMED interface must be called through a PLUMED script usually structured as

1. Definition of the units of measurements
2. Definition of group of atoms or virtual atoms (i.e. center of mass of a group of atoms)
3. Definition of one or more CVs
4. Eventual definition of differentiable operations on the defined CVs
5. Definition of the bias algorithm (MetaD, WT-MetaD, ABF, etc.)
6. Eventual definition of restraints on the CVs space to sample
7. Definition of the parameters to print on files

The code provides various routines for the analysis and post-processing of data generated through enhanced simulations. For instance, it offers the capability to reconstruct $F(\mathbf{s})$ from the deposited Gaussian kernels in MetaD simulations, as outlined in Eq. 1.63. Additionally, it allows the extraction of the unbiased Boltzmann probability density in Eq. 1.78, enabling the correct estimation of the equilibrium average for any arbitrary operator.

1.3 Stochastic Processes and Markov State Models

In this section, the concept of a stochastic process and its role in modeling physical processes will be initially introduced. Subsequently, Markovian processes, a subclass of stochastic processes, will be discussed. The formalism of Markov processes will be utilized to employ a wide variety of tools for extracting properties of the system that might otherwise appear to vary unpredictably over time. Finally, the contemporary use of Markov State Models (MSM) in the analysis of molecular dynamics simulation trajectories will be delved into since it will be one of the main tool applied in Chapters 2 and 4. The section will conclude with a brief mention of the application of MSM to model the response of biological ion channels to changes in membrane potential, where a practical and computational implementation will be treated in Chapter 6.

For a given random or stochastic time-dependent variable $\mathbf{X}(t)$, any function mapping it can be considered a stochastic process. The realizations of $\mathbf{X}(t)$, denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_n$, at different times $t_1, \ldots, t_n$ (with times increasing from left to right, i.e., $t_1 \leq t_2, \ldots, \leq t_n \leq t_{n+1}$) can be measured. If a set of joint probability densities exists

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots; \mathbf{x}_{n+1}, t_{n+1}) \tag{1.80}$$

then the system can be completely described. In terms of these joint probability density functions, conditional probability densities can be defined as follows

$$p(\mathbf{x}_{n+1}, t_{n+1}; \mathbf{x}_{n+2}, t_{n+2}; \ldots | \mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots) = \frac{p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots; \mathbf{x}_{n+1}, t_{n+1}; \mathbf{x}_{n+2}, t_{n+2}; \ldots)}{p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots)} \tag{1.81}$$

The interpretation of conditional probabilities is intrinsic to an evolution equation, suggesting that these probabilities predict future values of $\mathbf{X}(t)$ (i.e., $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots$ at times $t_{n+1}, t_{n+2}, \ldots$), given the knowledge of the past (values $\mathbf{x}_1, \mathbf{x}_2, \ldots$ at times $t_1, t_2, \ldots$).

To define a stochastic process, one needs to know at least all possible joint probabilities of the form given in Eq. 1.80. If such knowledge defines the process, it is referred to as a separable stochastic process. All processes considered hereafter are assumed to be separable. The simplest kind of stochastic process is complete independence:

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots) = \prod_i p(\mathbf{x}_i, t_i) \tag{1.82}$$

This implies that the value of $\mathbf{X}$ at time t is completely independent of its values in the past. An even more special case occurs when the $p(\mathbf{x}_i, t_i)$ are independent of t_i, meaning that the same probability law governs the process at all times.

An interesting subset of stochastic processes are the ones that satisfy the Markov assumption that the conditional probability of a stochastic process is determined by the realization of just the most recent condition

$$p(\mathbf{x}_{n+1}, t_{n+1}; \mathbf{x}_{n+2}, t_{n+2}; \ldots | \mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots) = p(\mathbf{x}_{n+1}, t_{n+1}; \mathbf{x}_{n+2}, t_{n+2}; \ldots | \mathbf{x}_n, t_n) \tag{1.83}$$

It seems a weak assumption, but it is instead very powerful because it allows to build a chain of simple conditional probability densities just following Eq. 1.83 as

$$p(\mathbf{x}_n, t_n; \mathbf{x}_{n+1}, t_{n+1}; \mathbf{x}_{n+2}, t_{n+2}; \ldots) = p(\mathbf{x}_n, t_n) p(\mathbf{x}_{n+1}, t_{n+1} | \mathbf{x}_n, t_n) p(\mathbf{x}_{n+2}, t_{n+2} | \mathbf{x}_{n+1}, t_{n+1}) \ldots \tag{1.84}$$

where an arbitrary joint probability density can be simply rewritten as a product of simpler conditional probability densities where the condition is the knowledge of the previous realizations of the stochastic variable at time t_{i-1}.
If a simpler case with three consecutive realizations of the Markov process is taken from equation 1.84 where $t_n \leq t_{n+1} \leq t_{n+2}$ then it is possible to find a useful identity, in literature called Chapman-Kolmogorov equation, by marginalization over $\mathbf{x}_{n+1}$

$$p(\mathbf{x}_n, t_n; \mathbf{x}_{n+2}, t_{n+2}) = p(\mathbf{x}_n, t_n) \int d\mathbf{x}_{n+1} p(\mathbf{x}_{n+1}, t_{n+1}|\mathbf{x}_n, t_n) p(\mathbf{x}_{n+2}, t_{n+2}|\mathbf{x}_{n+1}, t_{n+1}) \quad (1.85)$$

where using the definition of conditional probability is possible to write the left hand term of equation 1.85 as

$$p(\mathbf{x}_{n+2}, t_{n+2}|\mathbf{x}_n, t_n) = \frac{p(\mathbf{x}_n, t_n; \mathbf{x}_{n+2}, t_{n+2})}{p(\mathbf{x}_n, t_n)} \quad (1.86)$$

and by dividing both sides of Eq. 1.85 by $p(\mathbf{x}_n, t_n)$ one obtains the standard definition of the Chapman-Kolmogorov equation as

$$p(\mathbf{x}_{n+2}, t_{n+2}|\mathbf{x}_n, t_n) = \int d\mathbf{x}_{n+1} p(\mathbf{x}_{n+1}, t_{n+1}|\mathbf{x}_n, t_n) p(\mathbf{x}_{n+2}, t_{n+2}|\mathbf{x}_{n+1}, t_{n+1}) \quad (1.87)$$

It is crucial to note a few key points related to Equation 1.87. Firstly, as only the Markov assumption and the definition of conditional probability have been utilized to derive the Chapman-Kolmogorov equation, it is inferred that any Markov process must adhere to Equation 1.87. Secondly, since Equation 1.87 represents a highly intricate nonlinear functional equation establishing interconnections among all conditional probabilities $p(\mathbf{x}_i, t_i|\mathbf{x}_j, t_j)$ it allows for numerous solutions. The exploration of all possible solutions and discussions of Equation 1.87, including those for Wiener and Ornstein-Uhlenbeck processes, can be found in detail in books [136] and [137]. Instead, here it will be presented a couple of useful cases of Equation 1.87 namely, jump processes, leading to the interpretation of the Chapman-Kolmogorov equation as a master equation, and diffusion processes, leading to the Fokker-Planck equation.
Since Equation 1.87 will be the starting point for many derivations in the following sections, it is convenient to express it in another known form without loss of generality:

$$p(\mathbf{x}, t|\mathbf{x}', t') = \int d\mathbf{x}' W(\mathbf{x}, t|\mathbf{x}', t') p(\mathbf{x}', t') \quad (1.88)$$

where the chain of conditional probabilities in equation 1.87 has been substituted by the more compact term $W(\mathbf{x}, t|\mathbf{x}', t')$, which can be interpreted as the transition probability of the stochastic, or Markov, process to be $\mathbf{x}$ at a certain time $t \geq t'$ given that it was at $\mathbf{x}'$ at time t'. Moreover, it is also useful to write the differential Chapman-Kolmogorov equation starting from eq. 1.88 and differentiating over time the probability distribution $p(\mathbf{x}, t)$ using finite differences

$$p(\mathbf{x}, t + \Delta t|\mathbf{x}', t) - p(\mathbf{x}, t|\mathbf{x}', t) = \frac{\partial}{\partial t} p(\mathbf{x}, t|\mathbf{x}', t) + \mathcal{O}(\Delta t^2) \quad (1.89)$$

it can be shown that the differential Chapman-Kolmogorov equation denoted as the right term in the above equation can be written as [138]

$$\frac{\partial}{\partial t} p(\mathbf{x}, t) = \mathcal{L}_{KM} p(\mathbf{x}, t) \quad (1.90)$$

where it is important to note that the probability distribution in eq. 1.90 are still conditional probabilities which are not explicitly shown for the sake of simplicity in notation. Finally, $\mathcal{L}_{KM}$ is the Kramers-Moyal operator [138] and can be expressed as

$$\mathcal{L}_{KM} = \sum_{n=0}^{\infty} \left(-\frac{\partial}{\partial x} \right)^n L^{(n)}(\mathbf{x}, t) \tag{1.91}$$

with

$$L^{(n)}(\mathbf{x}, t) = \lim_{\Delta t \to 0} \frac{M_n(\mathbf{x}, t, \Delta t)}{\Delta t \cdot n!} \tag{1.92}$$

with M_n being the moments of the transition probability W in equation 1.88 defined as

$$\begin{aligned} M_n(\mathbf{x}', t, \Delta t) = \langle (\mathbf{x}(t + \Delta t) - \mathbf{x}(t)) \rangle|_{\mathbf{x}'=\mathbf{x}(t)} \\ = \int (\mathbf{x} - \mathbf{x}')^n W(\mathbf{x}, t + \Delta t | \mathbf{x}', t) d\mathbf{x}' \end{aligned} \tag{1.93}$$

As mentioned earlier, equation 1.88 possess different solutions if one assumes different functions for the transition probability $W(\mathbf{x}, t|\mathbf{x}', t')$ and for the probability distribution $p(\mathbf{x}', t')$. In the following, it will be assumed the probability distribution to be twice differentiable in $\mathbf{x}$ and so all the Kramers-Moyal terms of eq. 1.91 of order greater than 2 will be considered zero.

Jump Processes - Master Equation

One of the simplest cases related to equation 1.90 is when only the 0-th term of the Kramers-Moyal series survives, leading to the so-called jump processes. In this case, L^0 is written as

$$L^0 = \int d\mathbf{x}' W(\mathbf{x}, t|\mathbf{x}', t) - W(\mathbf{x}', t|\mathbf{x}, t) \tag{1.94}$$

it is easy to see that this approach results in a Chapman-Kolmogorov equation of the form of a master equation [137]

$$\frac{\partial}{\partial t} p(\mathbf{x}, t) = \int d\mathbf{x}' W(\mathbf{x}|\mathbf{x}') p(\mathbf{x}', t) - W(\mathbf{x}'|\mathbf{x}) p(\mathbf{x}, t) \tag{1.95}$$

furthermore equation 1.95 gain significant relevance in physical and chemical applications when the range of state space is discrete leading to

$$\frac{\partial}{\partial t} p_n(t) = \sum_n \left(W_{nn'} p_{n'}(t) - W_{n'n} p_n(t) \right) \tag{1.96}$$

the resulting master equation represents a balance between the increase and decrease in the probabilities of individual states n. The initial component corresponds to the augmentation of state n resulting from transitions originating from other states n', while the subsequent component reflects the reduction due to transitions from state n to other states. Equation 1.96 is used to model chemical reactivity, birth-death processes and as will be shown in Chapter 6 also to model possible conformational states of conductive biological channels under a potential stimuli.

Diffusion Processes - Fokker-Planck equation

Finally, another important solution of Equation 1.88 involves diffusion processes, where the surviving Kramers-Moyal terms in Equation 1.90 are only the first and second terms. This results in a time evolution of the probability density described by the Fokker-Planck equation

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[L^{(1)}(\mathbf{x},t)p(\mathbf{x},t)\right] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[L^{(2)}(\mathbf{x},t)p(\mathbf{x},t)\right] \tag{1.97}$$

Here, $L^{(1)}$, sometimes denoted as A_i, is referred to as the "convection term" or "drift term", while $L^{(2)}$, found in literature as B_{ij}, is usually termed the "diffusion term" or "fluctuation term". Before introducing specific cases of equation 1.97, it is crucial to establish the strong relationship that the Fokker-Planck equation has with a significant stochastic differential equation (SDE) introduced in Section 1.1.3 as the Langevin equation and explicitly formulated in Equation 1.31. Suppose to introduce an Ito's drift-diffusion process or, from a more physical perspective, a particle or a system of particles through the SDE

$$\dot{\mathbf{x}} = \mu_i(\mathbf{x},t) + \sigma_{ij}(\mathbf{x},t)\eta(t) \tag{1.98}$$

Here the most general formulation has been used, where both the functions μ_i and σ_{ij} depends on the stochastic variable $\mathbf{x}$ and on time and where $\eta(t)$ is a noise term (or Wiener process) that is normally distributed around 0 with variance equal to t. The situation mirrors the derivation of the differential Chapman-Kolmogorov equation for the probability distribution $p(\mathbf{x},t)$. By applying the Kramers-Moyal coefficients to the stochastic variable with the correct reformulation, the Kramers-Moyal coefficients become for the first term

$$L^{(1)}(\mathbf{x},t) = \lim_{\Delta t\to 0}\frac{\langle(\mathbf{x}(t+\Delta t)-\mathbf{x}(t))\rangle|_{\mathbf{x}'=\mathbf{x}(t)}}{\Delta t} = \mu_i(\mathbf{x},t) + \sigma_{kj}(\mathbf{x},t)\frac{\partial}{\partial x_k}\sigma_{ij}(\mathbf{x},t) \tag{1.99}$$

and for the second term

$$L^{(2)}(\mathbf{x},t) = \lim_{\Delta t\to 0}\frac{\langle(\mathbf{x}(t+\Delta t)-\mathbf{x}(t))\rangle^2|_{\mathbf{x}'=\mathbf{x}(t)}}{2\Delta t} = \sigma_{ik}(\mathbf{x},t)\sigma_{jk}(\mathbf{x},t) \tag{1.100}$$

This leads to a non-linear equation for the probability distribution of a multidimensional stochastic variable in the form of the form

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\frac{\partial}{\partial x_i}\left[\left(\mu_i(\mathbf{x},t) + \sigma_{kj}(\mathbf{x},t)\frac{\partial}{\partial x_k}\sigma_{ij}(\mathbf{x},t)\right)p(\mathbf{x},t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\left[\sigma_{ik}(\mathbf{x},t)\sigma_{jk}(\mathbf{x},t)p(\mathbf{x},t)\right] \tag{1.101}$$

that is of the same form of equation 1.97 explicitly stating the relationship between an Ito's diffusion process described by a SDE and the Fokker-Planck equation.
When Langevin dynamics was introduced in Section 1.1.3, it was formalized through an SDE of the same form as the diffusion process in equation 1.98. Therefore, if the interest is on the physical applications of the Fokker-Planck equation, it can be observed that equation

1.101 can be used to describe the evolution of a probability density related to a physical system in a non-homogeneous medium governed by some potential and a diffusion term. Plugging in the physical functions present in the over-damped Langevin equation 1.1.3 into equation 1.98, the Langevin SDE can be written once again as

$$\gamma(\mathbf{x})M\dot{\mathbf{x}} = -U'(\mathbf{x}) + \sqrt{2\gamma(\mathbf{x})Mk_{\mathrm{B}}T}\eta(t) \tag{1.102}$$

where the friction coefficient γ has been substituted by a more general friction function $\gamma(\mathbf{x})$ positive definite in the domain of the stochastic variable and the prime notation has been introduced to denote the derivative with respect to the stochastic variable $\mathbf{x}$ to have a more compact notation. If a diffusion coefficient is defined as $D(\mathbf{x}) \equiv \frac{k_{\mathrm{B}}T}{\gamma(\mathbf{x})M}$, then comparing equation 1.98 and equation 1.102 makes it straightforward to note that the drift term and the diffusion term in the Ito's diffusion process take the form of the following time-independent functions

$$\mu(\mathbf{x}) = \frac{-U'(\mathbf{x})D(\mathbf{x})}{k_{\mathrm{B}}T} \tag{1.103}$$

$$\sigma(\mathbf{x}) = \sqrt{2D(\mathbf{x})} \tag{1.104}$$

Thus, using equation 1.101 with the newly derived drift and diffusion term and with $\sigma'(\mathbf{x}) = \frac{D'(\mathbf{x})}{\sqrt{2D(\mathbf{x})}}$, the associated Fokker-Planck equation is written as

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -\frac{\partial}{\partial x_i}\left[\left(\frac{-U'(\mathbf{x})D(\mathbf{x})}{k_{\mathrm{B}}T} + D'(\mathbf{x})\right)p(\mathbf{x},t)\right] + \frac{\partial^2}{\partial x_i \partial x_j}\left(D(\mathbf{x})p(\mathbf{x},t)\right) \tag{1.105}$$

Now, let's briefly discuss two limiting cases of equation 1.105. The first, although trivial, is interesting and concerns Brownian motion, where the particle's motion is purely diffusive without any potential acting on the system. In this case, there is no drift $U(\mathbf{x}) = 0$ and with constant diffusion $D(\mathbf{x}) = D$ equation 1.105 then reduces to

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = D\frac{\partial^2}{\partial x_i \partial x_j}p(\mathbf{x},t) \tag{1.106}$$

where the stationary solution leads to the probability distribution

$$p(\mathbf{x},t) = \frac{1}{\sqrt{4\pi Dt}}e^{-\frac{\mathbf{x}^2}{4Dt}} \tag{1.107}$$

This result indicates that the particle undergoes random motion around the mean value (set to 0), with a variance equal to $2Dt$. Notably, this variance is consistent with the Einstein relation involving the mean square displacement, diffusion, and elapsed time.
The second interesting case involves the presence of a drift term $U(\mathbf{x})$ alongside a constant diffusion term $D(\mathbf{x}) = D$. Thus, equation 1.105 reduces to

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = -D\frac{\partial}{\partial x_i}\left(\frac{-U'(\mathbf{x})}{k_{\mathrm{B}}T}p(\mathbf{x},t)\right) + D\frac{\partial^2}{\partial x_i \partial x_j}p(\mathbf{x},t) \tag{1.108}$$

where the stationary solution leads to a probability distribution

$$p(\mathbf{x},t) = Ce^{-\frac{U(\mathbf{x})}{k_{\mathrm{B}}T}} \tag{1.109}$$

Here C is a normalization constant and interestingly the probability distribution represents the Maxwell-Boltzmann equilibrium distribution.
However, equation 1.105 in its more general form involves both drift and diffusion terms that depend on the stochastic variable. Different approximations must be made depending on the system under study, and more in-depth studies are needed. These aspects will be further discussed in Chapter 2, along with an application to solvent exchange around different cations.

1.3.1 Markov models in molecular dynamics

In the preceding section, the introduction of Markov processes was undertaken, emphasizing their utility in simplifying the modeling aspects of physical systems when applicable. Over the past decade, a vibrant area of research has witnessed the development of methods that bridge the gap between the theoretical tools associated with Markov processes and dynamical molecular systems, particularly trajectories obtained from MD simulations [139]. In the literature, the application of mathematical tools related to Markov processes to MD simulations is commonly referred to as Markov State Models (MSM). This nomenclature can be justified for several reasons. Firstly, trajectories in MD simulations are often inherently discrete (refer to section 1.1), and the number of molecular macrostates of interest, described through CVs (see section 1.2.2), is typically limited. Secondly, historically, early MSMs were employed to study conformational changes in proteins, focusing on a small number of states of interest [140]. Consequently, the term "Markov state models" is used, as it is a common practice in MD simulations to identify a set of states that sufficiently capture the dynamics of the molecular system. If the transitions between these states adhere to the Markov assumption (see equation 1.83), methods associated with solving the master equation (see section 1.3 and equation 1.96) can be applied. This enables the extraction of equilibrium probabilities for specific Markov states and, subsequently, the determination of transition times between different Markov states [141]. The workflow is summarized in Figure 1.5, and each step will be described in the subsequent paragraphs. The first step in building a MSM from MD trajectories involve the process of clustering the data coming from the trajectories (it can be one long MD trajectory or many short MD trajectories). This can be seen as a competitive process with the one of defining good CVs, instead they are two different processes and they need to be carefully assessed together in order to decrease the possibility to build an incorrect or non-representative MSM. In fact, the clustering part is usually performed over a pre-screened trajectory where the useless molecular descriptors have been discarded or even better over a good CVs space.
The primary objective at this stage is to establish a clustering that is kinetically relevant, employing geometric criteria. This clustering should effectively group together conformations that the system can swiftly transition between. Given this criterion, multiple clusterings may be deemed suitable, and there may not be a singular correct solution. The resulting microstate model derived from this clustering can then serve the purpose of establishing a quantitative link with experimental data or act as an initial framework for kinetic clustering. Several critical decisions need to be made during this phase, including the selection of a distance metric, the choice of a clustering algorithm and determining the number of clusters to generate.
One of the possible clustering algorithm could be selected among the k-centers clustering family. Thus, keeping the notation used within the discussion on CVs (section 1.2.1) and denoting $\tilde{\mathbf{s}}$ a sampled point in the CV space, the objective is to form a collection of clusters

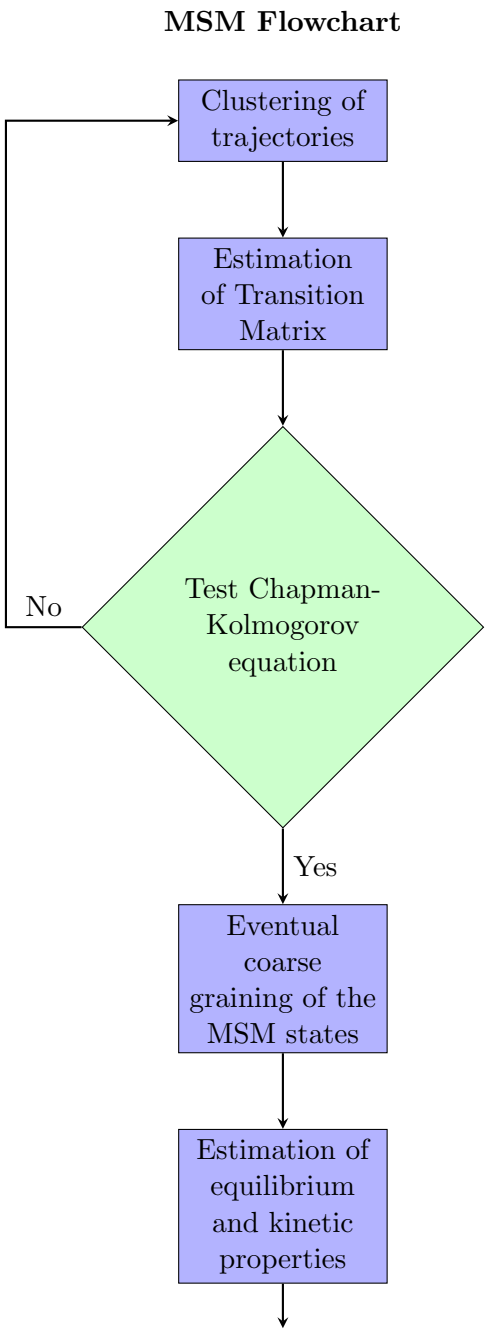

Figure 1.5: Schematic workflow to build, test and analyze a MSM constructed from MD trajectories.

with roughly equal radii, optimizing the objective function [142]

$$\min_{f} \max_{i} d(\tilde{\mathbf{s}}_i, f_k(\tilde{\mathbf{s}}_i)) \tag{1.110}$$

where $f_k(\tilde{\mathbf{s}})$ maps a conformation $\tilde{\mathbf{s}}$ to the nearest cluster center, and $d(\tilde{\mathbf{s}}_i, \tilde{\mathbf{s}}_j)$ represents the distance between two conformations. The minimization is performed over all possible clusterings f with k states, and the maximum is taken over all conformations in the dataset. The cluster's radius is defined as the maximum distance between any data point in that cluster and the cluster's center.
A notable advantage of the k-centers approach is its ability to distribute conformational space more evenly compared to other algorithms, achieving a balance in state volumes. From an intuitive perspective, this algorithm can be conceptualized as forming clusters with approximately equal volumes [143].
Another important family of algorithms belongs to the k-medoids clusterings, these algorithms are designed to minimize the average distance between data points and their assigned center, optimizing the objective function

$$\frac{1}{N} \sum_{i=1}^{N} d(\tilde{\mathbf{s}}_i, f_k(\tilde{\mathbf{s}}_i))^2 \tag{1.111}$$

where N is the number of data points. A notable advantage of k-medoids, compared to k-centers, is its tendency to create clusters with a more balanced distribution of samples [144].
Finally, one of the most used family of clustering algorithms is indeed k-means [145], where it arranges the data to minimize the cost function

$$\sum_{i=1}^{k} \sum_{\tilde{\mathbf{s}}_j \in S_i} d(\tilde{\mathbf{s}}_j, \boldsymbol{\mu}_i)^2 \tag{1.112}$$

where S_i represents clusters with centers of mass $\boldsymbol{\mu}_i$ and $\tilde{\mathbf{s}}_j$ denotes data points associated with their respective clusters. The outcome of clustering algorithms of this kind heavily depends on the initialization, and various initialization setups have been proposed to address this issue. In Chapter 4, two specific types, namely *uniform* [145] and *kmeans++* [146], were utilized. The *uniform* initialization aims for a setup that uniformly covers the spatial configuration of the dataset to a certain extent. On the other hand, *kmeans++* initializes centers by selecting random points uniformly from the provided dataset, using probability weights to determine new centers in the next iteration based on the shortness of the distance to the already placed cluster centers. It is worth noting that, despite k-medoids and k-means having similar formulations, a notable difference lies in the fact that the former finds cluster centers as actual data points $\tilde{\mathbf{s}}_i$, while the latter may find cluster centers that are not guaranteed to be data points, potentially lying between different data points.
After addressing the clustering issue, the subsequent problem involves the estimation of the transition matrix. It is crucial to emphasize that in the practical implementation of MSM on MD trajectories, it is challenging, if not impossible, to know *a priori* whether a specific type of discretization and the subsequent transition probability between states will satisfy the Markov assumption. Therefore, a trial-and-error approach is necessary, involving different clustering algorithms (and varying numbers of centers) to compute, evaluate, and test different MSMs until an appropriate state model satisfying the Markov assumption and the Chapman-Kolmogorov equation 1.85 is obtained.

Once a set of k centers (or microstates) is determined, with every data point falling into one of them, the next step is to construct a transition matrix W_{ij}. This matrix is defined as a row stochastic matrix

$$W_{ij} = \frac{C_{ij}}{\sum_k C_{ik}} \tag{1.113}$$

where C_{ij} is simply the value of the transition count matrix counting how many transitions have been recorded starting in i and finishing in j. The problem is now shifted toward the computation of the transition count matrix C_{ij}. This inevitably introduce another crucial parameter of a MSM that is the lag-time τ that defines the window of time in which recording the transition between two states. In literature, one can find the independent count approach where $C_{ij}(0) \rightarrow C_{ij}(\tau)$, $C_{ij}(\tau) \rightarrow C_{ij}(2\tau)$ and so on. Or the sliding window approach where $C_{ij}(0) \rightarrow C_{ij}(\tau)$, $C_{ij}(\Delta t) \rightarrow C_{ij}(\Delta t+\tau)$ and so on where Δt is the difference in time between two subsequent values of the MD trajectory $\tilde{\mathbf{s}}$ [143].
As mentioned earlier, the next step is to determine if the constructed MSM model and the associated transition matrix satisfy Markovianity. The simplest way to verify this is to utilize the Chapman-Kolmogorov equation 1.87. It is important to highlight that an MSM satisfying equation 1.87 is a necessary condition to ensure Markovianity, although sometimes it may not be sufficient.
Using equation 1.88 translated to matrix notation with successive times being $n\tau$ where n is a positive integer, one can find the useful relationship descending from directly from the Chapman-Kolmogorov equation

$$W(i, n\tau|j, t) = W^n(i, \tau|j, t) \tag{1.114}$$

that in compact notation is usually written as $W_{ij}(n\tau) = W^n_{ij}(\tau)$. One of the possible validation test is named implied timescale analysis and regards the analysis of the k eigenvalues λ_i of the transition matrix W_{ij} using equation 1.114. In fact, if equation 1.114 holds true then $\lambda_i(n\tau) \approx (\lambda_i(\tau))^n$ where in practical applications the equivalence symbol can be substituted with the approximate $\approx$ symbol due to the finiteness of the data available. Thus, the implied timescales defined as [147]

$$\theta_i(\tau) = -\frac{\tau}{ln\lambda_i(\tau)} \tag{1.115}$$

should be equivalent for $\theta_i(n\tau) \approx \theta_i(\tau)$. This has been used extensively as a validation test to check if the underlying model represented by the probability transition matrix can be considered Markovian [147] [143]. This test is usually represented graphically as can be seen in figure 1.6. Every $n = 1, 2, \ldots$ multiple of the lag-time τ, the eigenvalues of equation 1.115 are arranged from the greatest to the lowest, which in terms of timescales is ordering them from the slowest to the fastest and plot against the lag-time $n\tau$ associated. After a certain amount of time the implied timescales of a MSM will stabilize around a constant value that can be thought as the time needed to relax that particular mode of the model. The correct lag-time to use for further evaluation of the MSM would be a trade-off between the one able to resolve the slowest eigenvalues reaching the equilibrium (so the longer the better) and the one able to resolve the maximum number of eigenvalues even the ones associated to the fastest relaxation times (the shorter the better). In figure 1.6 a good lag-time of choice could be around 400 timesteps or 40 ps resolving the four slowest implied timescales that at the same time have reached plateau at that particular lag-time.
Another useful validation to perform on a MSM is the one proposed in [148]. Let $\boldsymbol{\pi}$ be the stationary probability of the Markov model associated to the transition matrix $W_{ij}(\tau)$. The

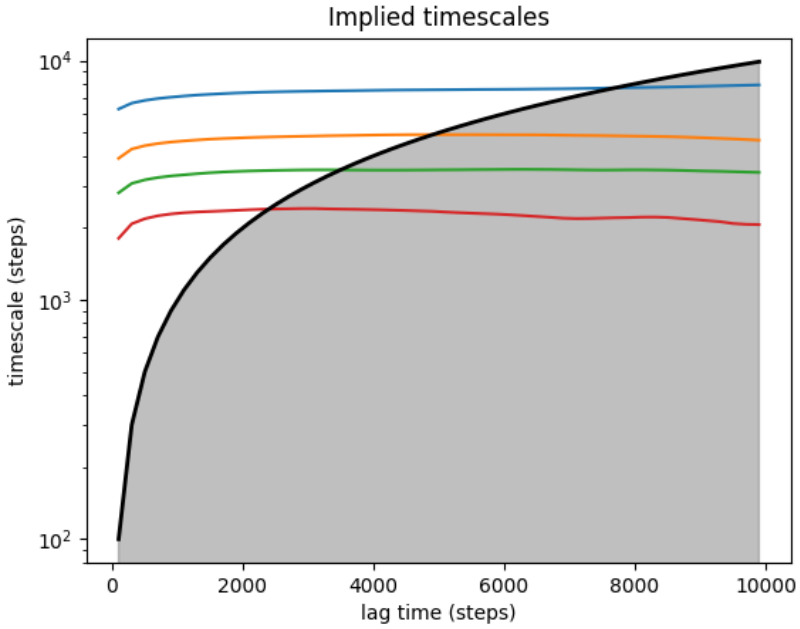

Figure 1.6: Implied timescales validation test for MSM. The four slowest implied timescales of the model have been computed through equation 1.115 and plotted on a log-scale against different $n\tau$ lag-times, where each step is 0.1 ps. The grey shaded area is the one under which the MSM cannot resolve any dynamic being the lag-time greater than the implied timescale of the MSM.

new stationary distribution corresponding to a restricted set A of microstates is given by

$$w_i = \begin{cases} \frac{\pi_i}{\sum_{j\in A}\pi_j}, & \text{if } i \in A \\ 0, & \text{if } i \notin A \end{cases} \tag{1.116}$$

where π_i is the probability of state i. In order to test the MSM it is possible to conduct a "relaxation experiment" on each set. Starting with $\mathbf{w}^A$ as an initial probability vector for each of the sets under consideration, the probability of finding the system in that set at times $n\tau$ can be computed either through the observed trajectory data or the Markov model and then subsequently compared. The probability to be at set A after time $n\tau$ extracted from the trajectories of MD simulations is given by

$$p_{\mathrm{MD}}(A, A; n\tau) = \sum_{i\in A} w_i^A p_{\mathrm{MD}}(i, A; n\tau) \tag{1.117}$$

Here, $p_{\mathrm{MD}}(i, A; n\tau)$ is the trajectory-based estimate of the probability to be in set A at time $n\tau$ when starting from state i at time 0

$$p_{\mathrm{MD}}(i, A; n\tau) = \frac{\sum_{j\in A} c_{ij}^{\mathrm{obs}}(n\tau)}{\sum_{j=1}^{n} c_{ij}^{\mathrm{obs}}(n\tau)} \tag{1.118}$$

In the other hand is possible to compute the probability to be at A using the Markov model as

$$p_{\mathrm{MSM}}(A, A; n\tau) = \sum_{i\in A} \left[(\mathbf{w}_A)^T W_{ij}^n(\tau)\right]_i \tag{1.119}$$

Evaluating the validity of the Markov model involves examining the extent to which the assumed equality holds

$$p_{\mathrm{MD}}(A, A; n\tau) \approx p_{\mathrm{MSM}}(A, A; n\tau) \tag{1.120}$$

that, essentially, is a test of the Chapman-Kolmogorov equation.
Assuming that a MSM has been validated with its associated probability transition matrix $W_{ij}(\tau)$. Since, typically, the number of centers of the clustering algorithms (or microstates of the MSM) has been increased to a large number to ensure Markovianity, there is a risk of losing the physical significance of certain states that, in reality, could be similar from both thermodynamic and kinetic perspectives. In chapter 4, this important problem will be addressed in connection with the microstates of the MSM and the experimentally measurable metal-ligand coordination states.
One of the main tool used to assess this issue is to perform Perron Cluster Cluster Analysis (PCCA) on the MSM. PCCA utilizes the eigenspectrum of $W_{ij}(\tau)$ to create coarse-grained models [149]. The part of the name "Perron Cluster" refers to a group of eigenvalues clustered near the largest eigenvalue and separated from the rest of the eigenspectrum by a significant gap [143]. In the PCCA approach, all microstates initially form a single macrostate, and then, through iterative steps, the most kinetically diverse macrostate is progressively divided into two smaller states based on the next slowest right eigenvector. PCCA+ [150] represents a more robust version of PCCA that avoids the drawback of error propagation. This feature is achieved by considering the relevant eigenvectors simultaneously rather than sequentially.
The last step that was mentioned in Figure 1.5 regards the evaluation of dynamical properties of the system that has been modeled with a MSM. Firstly, it is important to note that from a valid MSM it is possible to easily extract some equilibrium properties from the eigenvector associated to the biggest eigenvalue. Since the probability transition matrix that obeys Markov assumption is a row stochastic matrix, it possesses an eigenspectrum $0 \leq \lambda_i \leq 1$ where the eigenvector associated to the eigenvalue $\lambda_1 = 1$ is the vector of the equilibrium probabilities of each microstate of the MSM π_i [143]. Assuming that a cluster of the microstates has been created and there are C_i clusters representing realistic realization of the MSM, then the new equilibrium probabilities π_i can be summed over the same cluster as $\pi_i = \sum_{k \in C_i} \pi_k$. In this way it is possible to link those probabilities to actual physical quantities as

$$\pi_i \approx \frac{e^{-\beta U(\mathbf{s}_i)}}{Z(\beta)} \tag{1.121}$$

where $Z(\beta) = \int d\mathbf{s} e^{-\beta U(\mathbf{s})}$ is the partition function and $U(\mathbf{s})$ is the FES of the system felt by the MSM. It is straightforward how is it possible to extract relative thermodynamic quantities as ratio of two equilibrium probabilities as

$$\Delta G_{ij} = -k_{\mathrm{B}} T ln \frac{\pi_i}{\pi_j} \tag{1.122}$$

where the assumption of the underlying FES being an actual Gibbs free energy has been made.
Finally, it is possible to extract kinetic information from a valid MSM using for example Transition Path Theory (TPT) for Markov chains, as described in [151] and here briefly described. The computation of the statistics of transition pathways relies on computing the committor probability q_i^+. This quantity represents the probability of a system starting at state i, that it will reach one of the states of the set B first rather than the states belonging to the set A [110]. It is trivial to note that states in A have $q_i^+ = 0$ and states in B have

$q_i^+ = 1$. Regarding intermediate states between the two sets the committor increases from A to B and can be computed by solving the system of equations

$$-q_i^+ + \sum_{k\in I} W_{ik} q_k^+ = -\sum_{k\in B} W_{ik} \quad \text{for } i \in I \tag{1.123}$$

It is important to define also the backward-committor probability q_i^-. this quantity represents the probability, when starting at state i, that the system visited one of the states in set A before in time rather than in one of the states of set B. If detailed balance is satisfied it is simple to obtain $q_i^- = 1 - q_i^+$. Examining the probability flux between two arbitrary states i and j, indicated by $\pi_i W_{ij}$ (the absolute probability of locating the system at this transition), the attention is directed towards trajectories transitioning from A to B without going back to A. The flux for these reactive trajectories is determined by multiplication of the flux by the probability of originating from A and progressing to B

$$f_{ij} = \pi_i q_i^- W_{ij} q_j^+ \tag{1.124}$$

The net flux, f_{ij^+}, is defined as

$$f_{ij^+} = \max\{0, f_{ij} - f_{ji}\} \tag{1.125}$$

and establishes a network of fluxes emanating from states A and entering states B. The network adheres to flux conservation principles, ensuring that for every intermediate state i, the input flux matches the output flux. The total flux F for the transition $A \to B$ is expressed as follows

$$F = \sum_{i\in A}\sum_{j\notin A} \pi_i W_{ij} q_j^+ = \sum_{i\in B}\sum_{j\notin B} \pi_i W_{ij}(1 - q_i^+) \tag{1.126}$$

The total flux value provides the expected number of observed $A \to B$ transitions per time unit τ that an infinitely long trajectory would generate. However, the most significant kinetic quantity that can be extracted is the reaction rate constant k_{AB} [151], defined as

$$k_{AB} = \frac{F}{\tau \sum_{i=1}^{m} \pi_i q_i^-} \tag{1.127}$$

This quantity is of extreme importance when MSMs need to be compared with experimental realizations and this is one of the critical properties that will be addressed in Chapter 4 and for which the implementation of MSM in MD studies has reached a peak of interest in the last ten years [152, 153].

1.3.2 Markov models of active ion channels

In the introduction, the biological significance of ion channels has been emphasized, highlighting their intrinsic non-linear current response when subjected to external stimuli such as voltage or possible ligands attaching the ion channel. It is widely accepted that the opening and closing of a single ion channel follow a stochastic nature [154]. This observation has spurred a burgeoning field in computational biophysics, where concepts discussed in Section 1.3 are applied to model ion channel dynamics. This section will delve into key concepts, progressively generalizing towards the introduction of crucial tools that pave the way for discussions in Chapter 6.
In the simplest model of ion channel dynamics, the channel can exist in one of two states:

$$C \underset{k_{OC}}{\overset{k_{CO}}{\rightleftharpoons}} O$$

Figure 1.7: A simple two-state (closed-open) model for ion channels.

open or closed, as illustrated in the scheme in figure 1.7. Here, k_{CO} and k_{OC} are two reaction rates governing the model's dynamics, and they can be either constant or dependent on a predefined variable. In this case, the underlying model represents a jump process, and the temporal evolution of this kinetic scheme can be computed by solving the associated system of equations

$$\frac{d}{dt}p_o(t) = k_{CO}p_c(t) - k_{OC}p_o(t) \tag{1.128}$$

$$\frac{d}{dt}p_c(t) = k_{OC}p_o(t) - k_{CO}p_c(t) \tag{1.129}$$

This is analogous to equation 1.96. The solution for the probability density functions $p_i(t)$ is straightforward once the constraint $p_c(t) + p_o(t) = 1$ for every t is applied. The probability of finding the channel in the open state, $p_o(t)$, reaches a stable equilibrium for $t \to \infty$:

$$p_o(t \to \infty) = \frac{k_{CO}}{k_{CO} + k_{OC}} \tag{1.130}$$

However, advancing beyond a simple two-state model is essential to develop a more comprehensive kinetic model for a particular ion channel under study. This leads to a discrete master equation similar to equation 1.96

$$\frac{d}{dt}p_n(t) = \sum_n \left(k_{nn'}p_{n'}(t) - k_{n'n}p_n(t)\right) \tag{1.131}$$

This equation represents a complex system to solve, particularly as the number of states N in the model increases. When applied to ion channels, MSMs often have to address the resolution of Equation 1.131, where rates may depend on factors such as the cell membrane potential V, making the solution even more challenging. Four methodologies for solving these master equations will be discussed here.

One possible solution involves the discretization of time, replacing t with Δt, akin to the Euler scheme for solving ordinary differential equations (ODEs). In this approach, a sequence of probabilities associated with equation 1.131 is computed for small time steps Δt [155]. Despite not being computationally efficient, this method was implemented in the study of the dynamics of the voltage-gated potassium channel Kv4.3 and two of its possible mutations, as described in Chapter 6.

A second possibility is to solve the matrix equivalent of Equation 1.131, taking advantage of linear algebra tools and computationally faster routines involving matrix problem solutions. This method can be expressed as:

$$\frac{d\mathbf{p}(t)}{dt} = W^T\mathbf{p}(t) \tag{1.132}$$

where $\mathbf{p}(t)$ is the probability vector and W is the well-known transition matrix [36].

A third approach involves solving Equation 1.131 using a Monte Carlo scheme. In this method, the state at a successive time $t_{n+1} = t_n + \Delta t$ is determined by dividing the time interval into N non-overlapping time intervals, weighted by the respective rates. At $t_{n+1} = t_n + \Delta t$, the channel's state is updated based on a random number drawn from

$$C \underset{k_{OC}}{\overset{\mu k_{CO}}{\rightleftharpoons}} O$$

Figure 1.8: The simple two state (close-open) model for ion channels with the mutation parameter μ modifying the closed to open rate.

$$C \underset{k_{OC}}{\overset{k_{CO}}{\rightleftharpoons}} O \underset{k_{BO}}{\overset{k_{OB}}{\rightleftharpoons}} B$$

Figure 1.9: The simple two state (close-open) model for ion channels with an open blocker B that can act as a drain state for the open state.

a uniform distribution. This stochastic approach has been shown to converge to the same $p_i(t)$ as the previous methods with a sufficient number of Monte Carlo simulations [156]. Finally, the Gillespie algorithm, or Gillespie's direct method [157], is a well-known approach to solve Equation 1.131, but it will not be discussed here.
Regardless of the methodology used to solve Equation 1.131, the goal is to find the time-dependent probabilities $p_i(t)$ and, if it exists, the relative stable equilibrium as time approaches infinity $p_i(t \to \infty)$. In principle, there is no limit to the number of states and rates that can be included or designed for MSMs applied to ion channels.
However, it is essential to note that experimentally, it is possible to qualitatively measure an ion channel in just three states: closed, inactivated, and open. Additionally, quantitative measurements can only be made for the variation of the ion channel current over time, proportional to the time-dependent probability of finding the ion channel in the open state. In the literature, models often present different closed and inactive states [158, 159]. These states not only have a mathematical essence but also provide better mechanistic insights into the intricate dynamics of various ion channels.
Moreover, while everything discussed about MSMs for ion channels is typically relative to the wild type, it is crucial for MSMs to go beyond mere description and explain critical mutations of ion channels. A MSM should be able to incorporate parameters into an existing wild-type model, offering a framework to understand the effects of specific mutations. For example, in a simple two-state model (Figure 1.7), a closed-to-open mutation can be introduced by modifying the closed-to-open rate with a mutation parameter μ, as depicted in Figure 1.8. Finally, an interesting application involves incorporating theoretical drugs into MSMs [156]. Although these drugs may not have a pharmaceutical counterpart, they can be introduced into a model as blockers to hinder specific states. Figure 1.9 illustrates how a theoretical open blocker B can be implemented in a MSM to modify the dynamics of an ion channel, mimicking the potential behavior of a drug.

Chapter 2

Stochastic Model of Solvent Exchange in the First Coordination Shell of Aqua Ions

In this Chapter is reported an effective computational strategy aiming at providing a detailed picture of solvent coordination and exchange around aqua ions, thus including the main structural, thermodynamic, and dynamic properties of ion microsolvation, such as the most probable first-shell complex structures, the corresponding free energies, the interchanging energy barriers, and the solvent exchange rates. Assuming the solvent coordination number as an effective reaction coordinate and combining MD simulations with enhanced sampling and master-equation approaches, here is described a stochastic model suitable for properly calculating, at the same time, thermodynamics and kinetics of ion-water coordination. The model is successfully tested towards various divalent ions (Ca^{2+}, Zn^{2+}, Hg^{2+}, and Cd^{2+}) in aqueous solution, considering also the case of a high ionic concentration. Results show a very good agreement with those issuing from brute-force MD simulations, when available, and support the reliable prediction of rare ion-water complexes and slow water exchange rates not easily accessible to usual computational methods. This chapter is based on an article already published by Sagresti et. al [160].

2.1 Introduction

As discussed in the first part of the Introduction solvation of aqua ions, namely ion microsolvation, plays a key role in the study of aqueous solution structures [161], catalytic activity [162], ion transport [163], materials design [164], and so on. Gaining a detailed molecular understanding of such dynamic environments can deepen our comprehension of more complex mechanisms such as the water exchange mechanism in the first hydration shell, the solvent reorganization energy between ion redox couples [165], or the electrostriction effect in ionic solutions [166]. Yet, in most cases ion coordination is described through simple structural parameters, such as the average ion-water distance or the average coordination number, which are often insufficient to understand the variable behavior ions show in many circumstances, such as the debated "Gadolinium break" [167] or the nonlinear solvent response induced by redox reactions [165].
The free energy profile of ion coordination offers a more comprehensive understanding of ion solvation in aqueous solutions [168, 169], as suggested by the quasi-chemical theory of Pratt and colleagues [170, 171]. This profile illustrates the most accessible ion-water con-

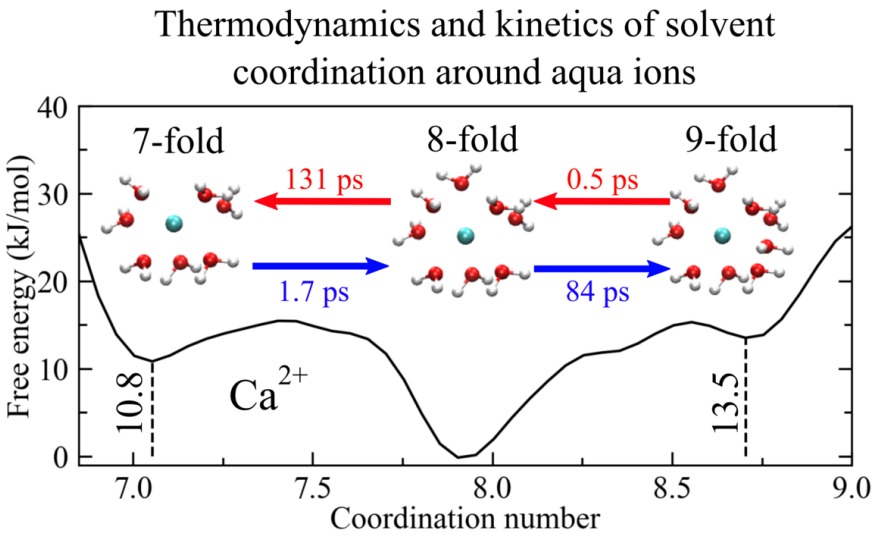

Figure 2.1: Idea behind the work presented in this chapter. Through a local variable it is possible to explore thermodynamics and kinetics of ions microsolvation.

figurations, provides energy differences among complexes, and estimates the energy barrier for water exchange events, clarifying whether the water exchange mechanism is dissociative or associative [172]. Additionally, the profile enriches our physico-chemical knowledge of ion solvation by adding a dimension to our understanding of these processes.
In this context, a metadynamics (MetaD) [120] based method has been proposed to obtain accurate free energy profiles for ion coordination in aqueous solutions, addressing the need for extended molecular sampling in evaluating free energy contributions to ion-water coordination [169]. This method provides a complete structural and thermodynamic picture of ion coordination through relatively short MetaD simulations, applicable to a variety of mono-, di-, and tri-valent ions, without bias towards specific water exchange pathways or mechanisms [169, 173].
While understanding the structural and thermodynamic aspects of ion microsolvation is crucial, assessing water exchange dynamics in the first coordination shell is equally important, as it relates to the kinetics of processes like ion transport in protein channels or ion-catalyzed reactions. Standard MD simulations generally cannot cover for the wide range of water exchange rates observed in NMR experiments (see Fig. 1A) [174, 162, 175], leading to the development of methodologies [176, 177, 178] especially within the framework of the transition state theory (TST). Two methodologies emerged as the most frequently adopted: the reactive flux [179] and transition path sampling (TPS) techniques [110]. As an alternative to TST-based methods, reaction rates can also be determined from a master-equation approach using the concept of mean first-passage time (MFPT) [180]. In this work, starting from the notion of free energy landscape of ion-water coordination [169] as seen above, we propose an effective computational strategy to estimate ion coordination and water exchange rates in the first solvation shell around aqua ions. In particular, the exchange rates are determined in terms of MFPTs between different ion-water configurations, as obtained by a purposely developed stochastic model. The model, which is based on the one-dimensional Fokker-Planck (FP) equation, assumes that the exchange process is Markovian given a suitable discretized reaction coordinate (i.e., water coordination num-

ber, s). In addition to the free energy function ($\Delta F(s)$), the key ingredient of the stochastic model is represented by the position-dependent diffusion coefficient, $D(s)$. Here, $D(s)$ was evaluated following the method proposed by Hummer [181], which is based on the calculation of the transition rate matrix assuming detailed balance at equilibrium. The present kinetic model was successfully tested against results issuing from direct MD simulations by considering Ca^{2+}, Zn^{2+}, Hg^{2+} and Cd^{2+} in aqueous solution as test cases. While most tests were performed on dilute solutions, in one case we also showed the application to a high molar concentration. Besides, we devised an effective methodology to address the case of rare exchange events not accessible to standard MD, thus allowing the reliable prediction of slow rates at an affordable computational cost. As a further important result obtained in this study, we showed, through the application of a committor analysis, that the water coordination number is not only a convenient and intuitive collective variable for describing ion-water coordination but also a physically sound "reaction coordinate" for the exchange process [109, 182].

2.2 Theory and Methods

2.2.1 Free energy of ion coordination

In this work, similarly to previous studies [168, 170], we made the assumption of describing the first hydration shell around a given ion in terms of the water coordination number, hereafter denoted as s, as an effective collective variable for the solvation process. The free energy of ion coordination in aqueous solution, $\Delta F(s)$, was conveniently expressed as a function of the solvent coordination number (see, e.g., Fig. 2.2), which was defined as a continuous parameter according to the method described in ref. [169]. For a given ion-water molecular configuration, the coordination number was, then, expressed as:

$$s = \sum_{i}^{N} \left(1 - \frac{1}{1 + e^{-a(r_i - r_0)}} \right) \tag{2.1}$$

where the sum is extended over the total number, N, of solvent molecules, r_i is the ion-oxygen distance of the i-th water molecule and r_0 and a are, respectively, the ion-oxygen cutoff distance and the parameter of the switching (exponential) function that smoothly goes from 1 to 0 across r_0 (see ref. [169] for more details). In particular, for each ion considered, the parameter r_0 was set to the distance of the well minimum following the first peak of the corresponding ion-oxygen radial distribution function (RDF) to see definition and how it was computed) (i.e., r_0 was within the range of 3.0-3.4 Å, Fig. 2.3). This choice was based on the idea of including all solvent molecules in the first coordination shell. The smoothing parameter a was set in all simulations to 4.0 $Å^{-1}$, according to some tests performed in our previous work. [169]. From extended samplings of the configurational space, as obtained by either standard MD or MetaD simulations, the free energy landscape of ion coordination, $\Delta F(s)$, was evaluated for all cations under scrutiny in this study (Figure 2.4). Note that the statistical error affecting $\Delta F(s)$ can be made systematically small by extending the configurational sampling. Then, according to the present method, accurate estimates of $\Delta F(s)$ can be obtained at an affordable computational cost, providing that a reliable ion-water interaction potential is employed.

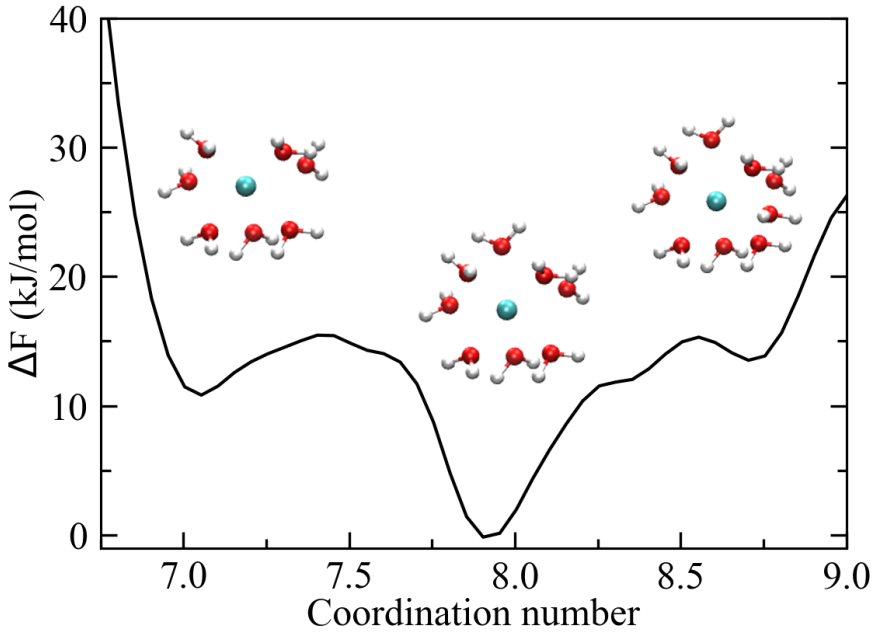

Figure 2.2: Free energy change (ΔF) as a function of the (continuous) solvent coordination number of Ca^{2+} in aqueous solution, obtained according to the equation 2.1 Representative ion-water complexes with seven-, eight- and nine-fold coordination are depicted as insets.

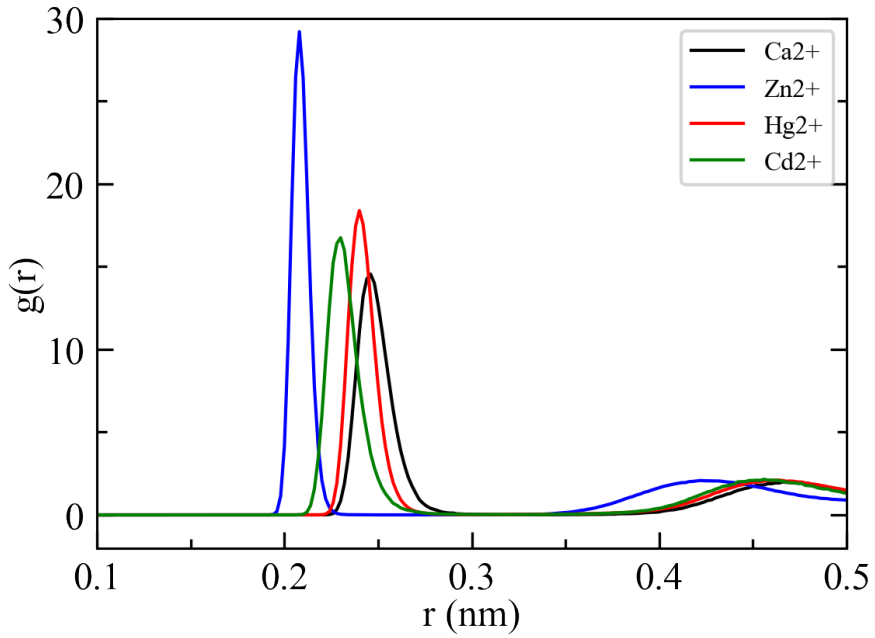

Figure 2.3: Computed radial distribution function (RDF) of Ca^{2+} (black), Zn^{2+} (blue), Hg^{2+} (red) and Cd^{2+} (green) in water (i.e., Ion-O RDF).

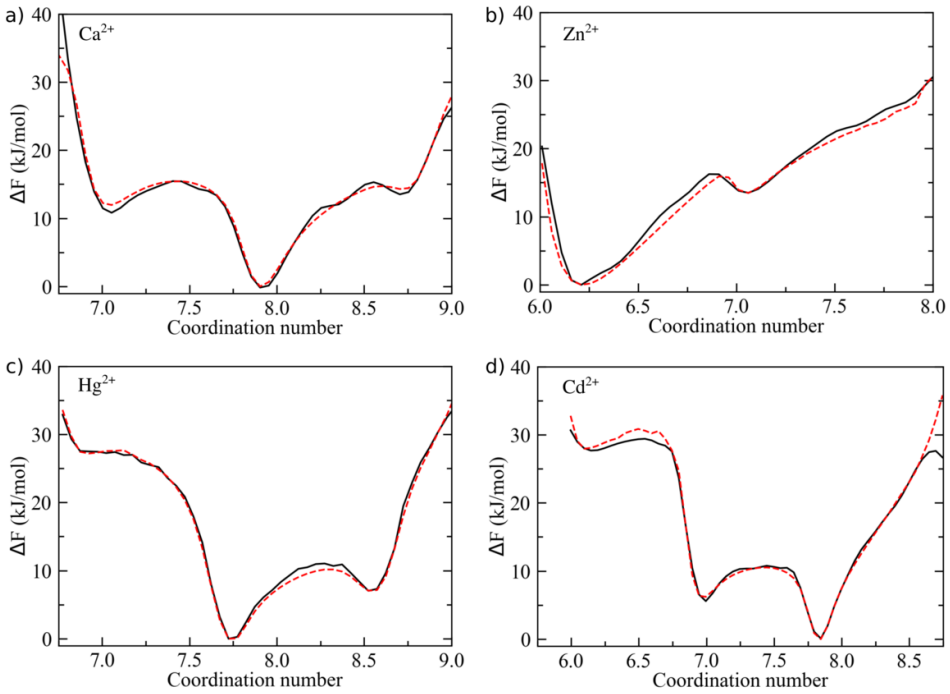

Figure 2.4: Free energy landscape, $\Delta F(s)$, of ion coordination as issuing from standard MD (red dashed line) and MetaD (black solid line) simulations for (a) Ca^{2+}, (b) Zn^{2+}, (c) Hg^{2+} and (d) Cd^{2+}.

2.2.2 Water exchange dynamics

A stochastic kinetic model was developed to describe the water exchange dynamics in the first solvation shell, that is to estimate the water exchange rates between different ion-water configurations. Assuming the dynamical process is Markovian for a proper coarse-grained discretization of the reaction coordinate (i.e., ion-water coordination number), a kinetic model based on the one-dimensional FP equation, also known as the Smoluchowski equation, was developed [183]:

$$\frac{\partial p(s,t)}{\partial t} = \nabla \cdot D(s)(\nabla - \beta F(s))p(s,t) \tag{2.2}$$

where $p(s,t)$ is the time-dependent probability distribution density, $\beta = (k_{\mathrm{B}}T)^{-1}$ is the Boltzmann factor (i.e., the inverse of the Boltzmann constant, k_{B}, times the temperature, T) and $D(s)$ is the position-dependent diffusion coefficient of s. As an alternative approach, the water exchange dynamics can be equivalently described by the (overdamped) Langevin equation (LE) which expresses the (stochastic) equation of motion of the coordinate s:

$$ds_t = (\nabla D(s_t) - D(s_t)\beta \nabla F(s_t))\, dt + \sqrt{2D(s_t)}dW_t \tag{2.3}$$

where $s_t \equiv s(t)$ and dW_t is the Wiener process. In the latter case, transition rates are obtained by averaging the arrival times over multiple Langevin dynamics (LD) simulations. Along with $F(s)$, $D(s)$ is the second most important ingredient needed to fully define the kinetic model and it was evaluated as described in the following.

2.2.3 Mean first passage time

Water exchange rates were evaluated in terms of MFPTs between different coordination number configurations. Note that MFPTs can be evaluated in different ways. In particular, if the coordination space is accessible to standard MD simulations, MFPTs can be directly obtained from the analysis of the corresponding trajectories. Otherwise, in case of rare transitions, MFPTs can be evaluated from the present kinetic model by solving numerically either the FP equation or the corresponding backward Kolmogorov-Chapman equation [184]. Exploiting the same kinetic model, MFPTs can also be obtained from the equivalent Langevin dynamics simulations.

First, on the basis of the computed free energy profile ($F(s)$) for a given ion-water system, the coordination number space was partitioned into a discrete number of consecutive coordination states, s_i (i.e., different regions of s), in correspondence to the free energy local minima, each one limited by adjacent energy barriers (Fig. 2.5a). Accordingly, the MFPT was defined as the average time spent by the system in each coordination state before jumping to a different one. Since ions generally showed three or more main coordination states, the MFPT, τ_{ij}, to jump from a given state s_i to an adjacent state s_j (with $j = i \pm 1$) was defined as the ratio between the overall residence time in the i-th state (τ_i) and the number of $i \to j$ state transitions (n_{ij}), that is $\tau_{ij} = \tau_i / n_{ij}$.

From standard MD (or Langevin dynamics) simulations, the τ_{ij}'s were obtained by initially assigning each configuration of the trajectory to a unique coordination state according to a history-based algorithm: each configuration sampled at a given time t was assigned to state s_i if, at a previous time t' with $t' < t$, the coordinate $s(t)$ crossed the local minimum configuration of s_i (see Fig. 2.5 b and c). This choice prevented the counting of spurious jumps between states (i.e., fast barrier recrossings) while, at the same time, it ensured that transitions occurred only between adjacent coordination states. To validate this procedure, we compared the population of states (s_i) issuing from such a history-based method with the one obtained by mapping directly each configuration of the MD trajectory onto the state s_i corresponding to the partition visited. As a result, no significant differences appeared (Fig. A.1). While other time-based criteria were also tested for assigning MD configurations to a given state s_i, as for example the use of a "minimum residence time", the procedure described above appeared as the most satisfactory for treating the ion-water coordination dynamics. Also, note that the present methodology is conceptually similar to the one used by Milestoning [185].

Within the framework of our stochastic model (i.e., a birth-death process where transitions are allowed only between adjacent coordination states), the conditional MFPT between state s_i and s_j (starting from s_i at $t = 0$) can be expressed as [137]:

$$\tau_{i,j} = \frac{\int_0^\infty t p_{i,j}(t)\, dt}{\int_0^\infty p_{i,j}(t)\, dt} \tag{2.4}$$

where $p_{i,j}(t)$ is the conditional probability density that the system reaches s_j at time t upon starting from s_i at time zero ($p_{i,j}(t) = p(s_j, t|s_i, 0)$). The integral $\int_0^\infty p_{i,j}(t)\, dt$ corresponds to the "splitting" probability to end up in s_j (see, e.g., ref. [137]). Eq. 2.4 can be solved, once the probability density $p(s,t)$ is known, by assuming an adsorbing (at the ending state) and a reflecting (preceding the starting state) boundary condition. Among other methods, the latter can be integrated numerically with the Crank-Nicolson scheme [186]. As a more convenient alternative, the MFPT can be also obtained directly from the adjoint equation of the FP (i.e., also known as backward Chapman-Kolmogorov equation) by solving the

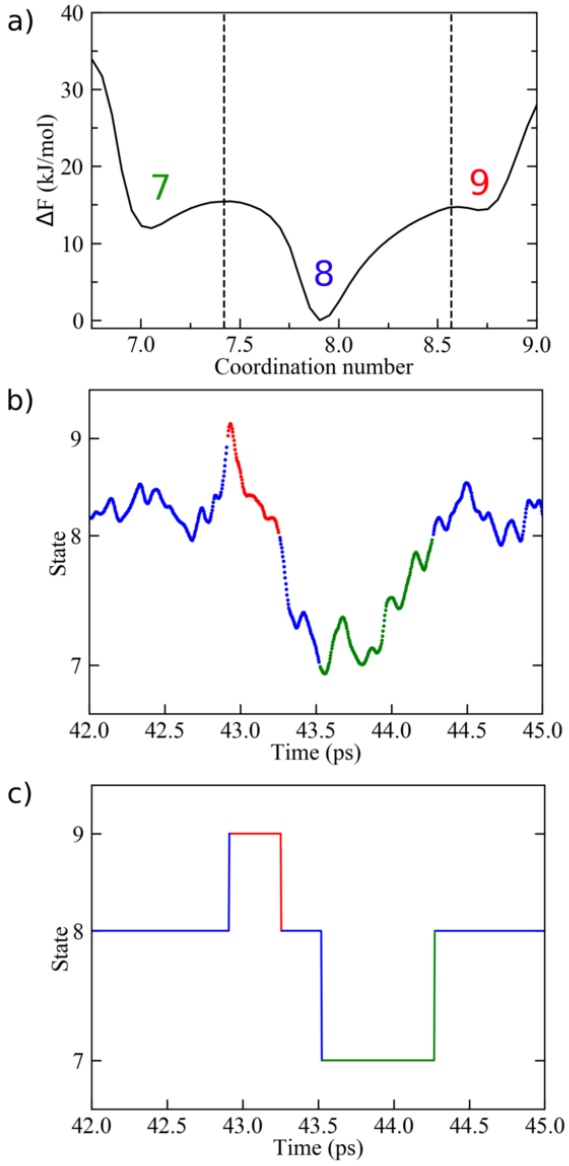

Figure 2.5: a) Partitioning (dashed lines) of the coordination number space in contiguous regions representing different metastable states (i.e., 7, 8 and 9) on the basis of the free energy profile of Ca^{2+} in water. b) Trajectory of the s coordinate (solid line) during a given time interval of the Ca^{2+} MD simulation. At each time step, the system is assigned to one of the possible coordination states according to the history-based method described in the text. Green, blue and red color correspond to state 7, 8 and 9, respectively. c) The same trajectory, after the assignment, is converted into a discrete number representation (i.e., coordination state number). Note that the overall residence time, τ_i, of the system in the state s_i is given by the sum of all time intervals assigned to s_i.

integral (see, e.g., ref. [184]):

$$\tau_{i,j} = \int_i^j \frac{e^{\beta F(z)}}{D(z)}\, dz \int_{-\infty}^{z} e^{-\beta F(y)}\, dy \tag{2.5}$$

Furthermore, an approximate well-known result for $\tau_{i,j}$ is provided by the Kramers theory [187], which, in the limit of an overdamped dynamics, gives the compact formula [180]:

$$\tau_{i,j} = \frac{2\pi\gamma}{\omega_i \omega_j} e^{\frac{\Delta F^\dagger}{k_\mathrm{B}T}} \tag{2.6}$$

where γ is the friction coefficient ($\gamma = k_\mathrm{B}T/D$), ω_i and ω_j are the angular frequencies at the well bottom of s_i and s_j, respectively, and $\Delta F^\dagger$ is the energy barrier for the $i \to j$ transition. The angular frequency can be approximated as $\omega_{i/j} = \sqrt{|\Delta F''(s_{i/j})|}$. In this work, the Kramers' MFPT was also evaluated, for the sake of comparison, assuming that the constant coefficient D was given by $D = \frac{1}{D(i)-D(b)} \int_i^b D(s)\, ds$, with $D(i)$ and $D(b)$ the diffusion at the bottom of s_i and at the peak of the barrier, respectively. Statistical errors of the τ's were estimated from the exponential fit of the distribution of the arrival times, as obtained from the MD simulations. In case of the stochastic approach, the same errors were estimated from the corresponding uncertainty of the diffusion coefficient ($D \pm \delta D$) as described in the next section.

2.2.4 Position-dependent diffusion coefficient

Following the method proposed by Hummer [181], a position-dependent diffusion coefficient along the coordination number, $D(s)$, was obtained from a master equation approach, upon partitioning evenly the configurational space into N non-overlapping regions of width Δs:

$$\dot{p}_i(t) = \sum_j \mathbf{R}_{ij} p_j(t) \tag{2.7}$$

where $p_i(t)$ is the probability of being in the region i at time t and $\mathbf{R}_{ij}$ is the transition rate matrix with constant coefficients. The solution of this equation can be expressed in terms of the propagator [181]:

$$p(j, t|i, 0) = \left(e^{t\mathbf{R}}\right)_{ji} \tag{2.8}$$

which expresses the probability of finding the coordinate s within the region j at a time t, after starting at i at time $t = 0$. The rate matrix $\mathbf{R}_{ji}$ is related to the position-dependent diffusion coefficient through the equation:

$$D_{i+1/2} \approx \Delta s^2 R_{i+1,i} \left(\frac{P_i}{P_{i+1}}\right)^{1/2} \tag{2.9}$$

where $D_{i+1/2}$ represents the arithmetic mean $(D(s_i) + D(s_{i+1}))/2$, Δs is the discretization step and P_i is the equilibrium population of the i-th region (note that $P_i = \exp(-\Delta F(s_i)/k_\mathrm{B}T)$ and is readily obtained from MD or MetaD simulations). In practice, the propagator (eq. 2.8) is constructed from the observed local transitions (from i to j) during the MD simulations, given a fixed lag time Δt. The rate matrix $\mathbf{R}_{ji}$ is obtained through the routine linalg.logm [188] of Scipy (v. 1.5.4) and, as a result, the position-dependent diffusion coefficient, $D(s)$, was determined from eq. 2.9. Note that the discrete regions do not correspond to the previous coordination states, but they are the result of a finer partitioning of the

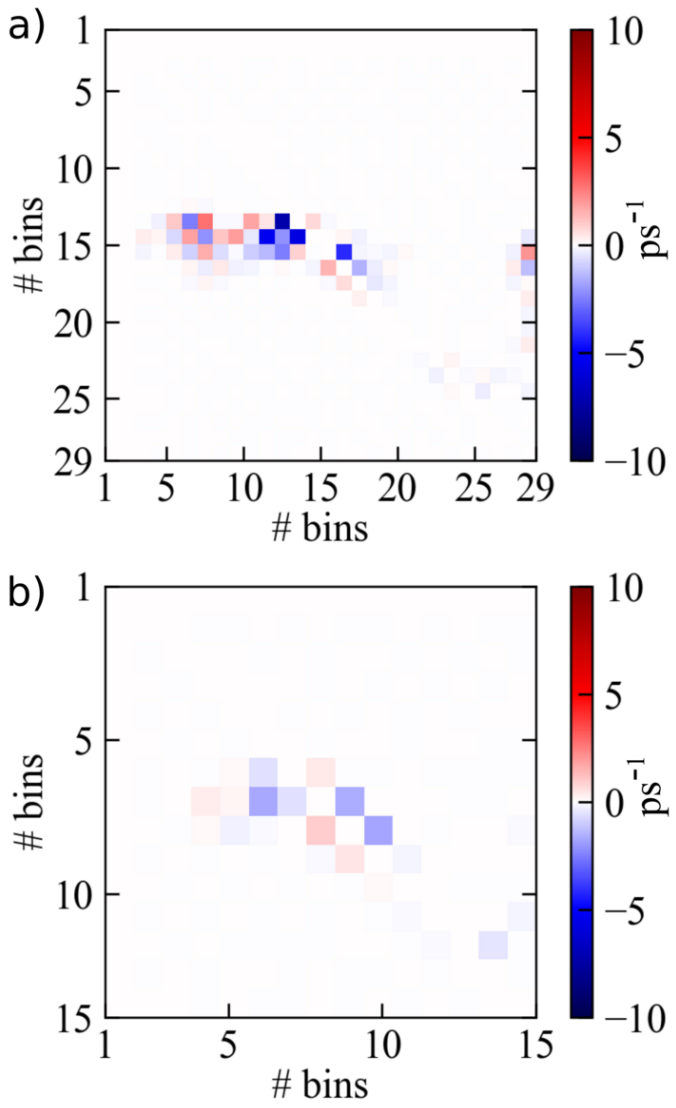

Figure 2.6: Map of detailed balance deviations ($P_i R_{ji} - P_j R_{ij} \neq 0$). The map highlights how far from the ideal detailed balance condition is the computed transition rate matrix. a) Example of a poor discretization (29 bins along the s coordinate for Hg^{2+}) showing a rough approximation to a birth-death process, around the 13-th bin. b) With 15 bins, the discrepancies are drastically reduced to less than 1 ps^{-1} and then the rate matrix can be accepted to construct $D(s)$. This example shows the importance of a correct discretization of the coordinate s.

coordination number space. In case of high free energy barriers and rare transition events, our approach takes advantage of the molecular sampling obtained during the MetaD simulations to extract starting configurations throughout all the coordinate space in order to run short MD runs (250 ps) aiming at determining the required local transition probabilities used to define the propagator, $(e^{t\mathbf{R}})_{ji}$. For each starting configuration, many MD replicas (>100) were carried out by randomly resampling the momenta. By tuning the discretization (Δs) and lag time (Δt) parameters, the transition probability matrix becomes essentially tridiagonal (Fig. 2.6). Moreover, it is possible to set up a simple validation test by exploiting the detailed balance condition. Assuming the process is Markovian and reversible, the detailed balance requires $P_i R_{ji} - P_j R_{ij} = 0$ at equilibrium. Then, the extent by which this relation differs from zero provides an uncertainty measure of the diffusion. In particular, taking into account the detailed balance, eq. 2.9 can be rewritten as:

$$D_{i+1/2} \approx \frac{D_1 + D_2}{2} = \frac{\Delta s^2}{2}\left[R_{i+1,i}\left(\frac{P_i}{P_{i+1}}\right)^{1/2} + R_{i,i+1}\left(\frac{P_{i+1}}{P_i}\right)^{1/2}\right] \tag{2.10}$$

In the ideal scenario in which the detailed balance strictly holds true, the first and the second term of the r.h.s. of eq. 2.10 do correspond exactly and eq. 2.9 is retrieved. In real cases, however, the small observed difference between the two terms, purposely renamed D_1 and D_2, is used to provide an estimate of the error of D, as $\delta D = \frac{|D_1 - D_2|}{2}$.

2.2.5 Committor analysis

In order to assess the reliability of the s collective variable (Eq. 2.1) as a proper reaction coordinate for describing the ion-water coordination dynamics, we carried out the analysis of the committor as originally proposed in ref. [109] and tested in various subsequent works [189, 190, 191, 192] (see also ref. [182] for a detailed discussion on the significance and reliability of a reaction coordinate). The committor, $\pi_i(s_0)$, is defined as the probability for the system to end up in state s_i while starting from a given coordinate s_0, which is usually considered at an intermediate "transition state" point between two or more thermodynamic states. In our mono-dimensional stochastic model, this function can be expressed as the probability for the system to reach first, at a later time, the state located on the right side (R) or the left side (L) of the starting coordinate s_0 (the exact time not being relevant):

$$\pi_R(s_0) = \frac{\int_{s_L}^{s_0} exp[\beta F(s) - lnD(s)]}{\int_{s_L}^{s_R} exp[\beta F(s) - lnD(s)]} \tag{2.11}$$

where $\pi_R(s_0)$ is the right committor, that is the probability of a trajectory to reach the state on the right (R) before the one on the left (L) when starting at the top of the dividing barrier s_0. The analysis of the distribution of the committor values ($p(\pi_R)$), typically constructed as an histogram, was evaluated from multiple MD simulations starting from system configurations lying at the separatrix (i.e., $\pi_R(s_0) = 0.5$) between two adjacent coordination states. In practice, the starting configurations (about 1200 selected configurations) in close proximity to a given energy barrier top, s_0 (i.e., $s_i < s_0 < s_{i+1}$), were generated by the MetaD simulation. From each of these configurations, 100 replica MD simulations were carried out by resampling randomly the system velocities for about 20 ps (a time interval sufficient to reach the bottom of either left or right coordination states). The obtained collection of ending states (i.e, s_L or s_R) was then used to estimate the committor probability distribution.

2.2.6 Simulation details

MD and MetaD simulations of the five ion-water systems (Ca^{2+}, Zn^{2+}, Hg^{2+} and Cd^{2+}) were carried out to estimate the free energy $\Delta F(s)$ along the coordinate s. In each case, a divalent cation was initially placed in a cubic box (40 Å x 40 Å x 40 Å, 2160 water molecules) and solvated with either the TIP3P [193] (Zn^{2+}, Hg^{2+} and Cd^{2+}) or SPC/E [194] (Ca^{2+}) water model. In case of Hg^{2+}, a solution at higher (0.5M) concentration was also investigated. Every system was neutralized with Cl^- counter-ions. The CHARMM27 [193] force field was used for Hg^{2+}, Cd^{2+}, and Zn^{2+}, while GROMOS35A6 [195] was adopted for Ca^{2+}. For Zn^{2+}, Hg^{2+} and Cd^{2+}, the non-bonded Lennard-Jones potential was modified by adding a $1/r^4$ term (i.e., using the so-called 12-6-4 potential developed by Merz and collaborators [54]) to better estimate the charge-induced dipole interactions in M(II) ions. In the dilute solution models, a distance restraint potential was applied between the cation and the counter-ions to avoid the formation of ionic clusters during the MD simulations, so to reproduce correctly ion-oxygen distances in the first solvation shell and average coordination number as reported in previous studies without counter-ions [196, 48]. The GROMACS [197] software package was used to perform a 1000 step minimization, followed by an equilibration (1 ns) in the NpT ensemble (at 300 K and 1 atm) to correctly resize the box volume. 1 microsecond MD production runs were performed according to the NVT ensemble. Metadynamic s[120] was employed to efficiently obtained the free energy profile, $\Delta F(s)$, of ion coordination (as described in ref. [169]). As a further test, the latter was compared with the one obtained from the corresponding pure MD simulation. The gaussian kernels were added every 5 ps with $\sigma = 0.01$ and $h = 0.1$ kJ/mol. The coordinate s was recorded at every timestep during both pure MD and metadynamics and the free energy profile successively reconstructed as $F(s) = -k_B T \ln P(s)$, with $P(s)$ the observed probability distribution. Standard deviation for F(s) computed through MetaD simulations is 1 kJ/mol. Metadynamics simulations were carried out using the open-source, community-developed PLUMED library (ver. 2.6) [135]. Langevin dynamics simulations were carried out by numerical integration of Eq. 2.3 with the Euler-Maruyama algorithm [198]. The integration timestep was set to 2 fs and, for each system, about 1000 replica simulations were performed starting from each state configuration, so to collect enough statistics for the evaluation of the MFPT.

2.3 Results and Discussion

2.3.1 Assessment of the kinetic model

The stochastic kinetic model and the proposed computational procedure to evaluate water exchange rates in the first solvation shell around hydrated ions were tested on a number of different systems, namely Ca^{2+}, Zn^{2+}, Hg^{2+} and Cd^{2+}. First, we considered the calcium ion since it is known that water exchange is relatively fast around Ca^{2+} and, then, readily accessible to standard MD simulations. The free energy profile of ion coordination was obtained by both MD and MetaD simulations, where the latter was carried out following the methodology originally proposed in ref. [169] (see details in the Methods section). Results are reported in Fig. 2.4 showing a very good agreement between pure MD and MetaD, in line with our previous study [169], thus supporting the use of MetaD to obtain the free energy as a function of the coordination number. In particular, Ca^{2+} displays three ion-water configurations within 15 kJ/mol (Fig. 2.7a), with coordination number 7, 8 and 9. The free energy barrier from coordination 8, which is the most favorable configuration, to 7 or 9 is about 15 kJ/mol, while the barrier to go back to 8 from the other coordination numbers

Table 2.1: MFPT for ion coordination in water, computed from MD simulation (MD), Langevin dynamics (LD) and Fokker-Planck integration (FP) (see Sec. 2.2.3 for details).

Ion	Transition	MD (ps)	LD (ps)	FP (ps)
Ca^{2+}	$7 \rightarrow 8$	1.58 ± 0.08	1.7 ± 0.1	1.65 ± 0.25
	$8 \rightarrow 9$	76 ± 2	88 ± 15	84 ± 11
	$8 \rightarrow 7$	120 ± 3	135 ± 12	131 ± 10
	$9 \rightarrow 8$	0.40 ± 0.05	0.5 ± 0.1	0.50 ± 0.05
Zn^{2+}	$6 \rightarrow 7$	304 ± 10	294 ± 25	287 ± 30
	$7 \rightarrow 6$	1.2 ± 0.3	1.5 ± 0.3	1.3 ± 0.3
Hg^{2+}	$7 \rightarrow 8$	0.5 ± 0.2	0.7 ± 0.1	0.7 ± 0.1
	$8 \rightarrow 9$	16.8 ± 0.5	21 ± 2	20 ± 3
	$8 \rightarrow 7$	$27 \pm 10 \cdot 10^3$	$20 \pm 3 \cdot 10^3$	$18 \pm 3 \cdot 10^3$
	$9 \rightarrow 8$	1.70 ± 0.15	1.5 ± 0.3	1.7 ± 0.2
Cd^{2+}	$6 \rightarrow 7$	1.6^*	1.8 ± 0.4	2.0 ± 0.5
	$7 \rightarrow 8$	2.6 ± 0.2	2.9 ± 0.3	2.8 ± 0.4
	$7 \rightarrow 6$	$\sim 10^{3*}$	$14 \pm 4 \cdot 10^3$	$14 \pm 3 \cdot 10^3$
	$8 \rightarrow 7$	17.7 ± 1.2	18.5 ± 2.0	17 ± 2

* Estimate obtained from the average of the observed transition times.

is significantly smaller ($<$ 2 kJ/mol). Then, we set out to evaluate the MFPTs for the corresponding coordination state transitions. The position-dependent diffusion coefficient was computed using the computational procedure described in Sec. 2.2.4, as depicted in Fig. 2.7b. $D(s)$ fluctuates between 0.33 and 0.06 ps^{-1} in the relevant space interval ($s = 7 - 9$) and the corresponding statistical error is on average rather small (0.04 ps^{-1}). The resulting MFPTs obtained from our kinetic model using either Langevin dynamics (LD) or Fokker-Planck integration are in very good agreement with the ones from long MD simulations, as shown in Table 2.1. Overall, the MFPTs reflected the observed $F(s)$ profile (Fig. 2.7a), with $\tau_{7/9 \rightarrow 8}$ being about one ps and $\tau_{8 \rightarrow 7/9} = 80 - 130$ ps, but the kinetic model captured fairly well the existing difference in the average transition times between $8 \rightarrow 7$ and $8 \rightarrow 9$ ($\Delta\tau \approx 45ps$). The latter finding could not have been predicted from the free energy profile alone and, therefore, highlights the beneficial use of such a kinetic analysis to unravel subtle differences in water exchange dynamics in the first solvation shell. Note that the relatively easy water exchange observed in the case of Ca^{2+} is well in line with both previous quantum mechanical calculations, X-ray and neutron diffraction experiments on $CaCl_2$ solutions and X-ray crystal structures reporting large variations in the coordination number, with values ranging from 6 to 10 (see, e.g., ref. [199]).

Going to Zn^{2+}, we found two free energy minima (Fig. 2.8a) corresponding to coordination number 6 and 7, separated by a relatively low energy barrier (16 kJ/mol) that allowed the sampling of numerous coordination state transitions from standard 1 μs MD simulation. Note that in this case the free energy profile, for the chosen force field, clearly pointed towards an associative mechanism as the preferred one for water exchange around the zinc ion. The corresponding MFPTs provided $\tau_{6 \rightarrow 7} \approx 300ps$ and $\tau_{7 \rightarrow 6} \approx 1ps$, again showing a nice match between our kinetic model and pure MD results (Table 2.1).

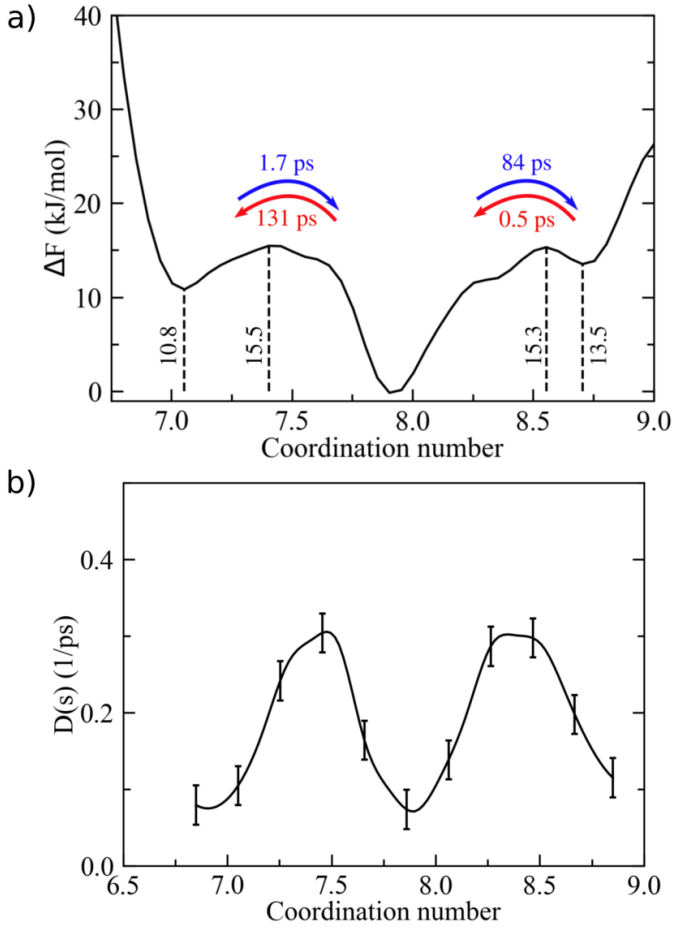

Figure 2.7: a) Free energy landscape of Ca^{2+} coordination in water. ΔF values at relevant points (i.e., local minima/maxima) are reported explicitly. Mean first-passage times corresponding to transitions between adjacent states are also reported as computed from the integration of the FPE. Standard deviation on F(s) is 1 kJ/mol. b) Position-dependent diffusion coefficient as a function of the coordination number. Error bars correspond to $\pm\delta D$ (see Sec. 2.2.4).

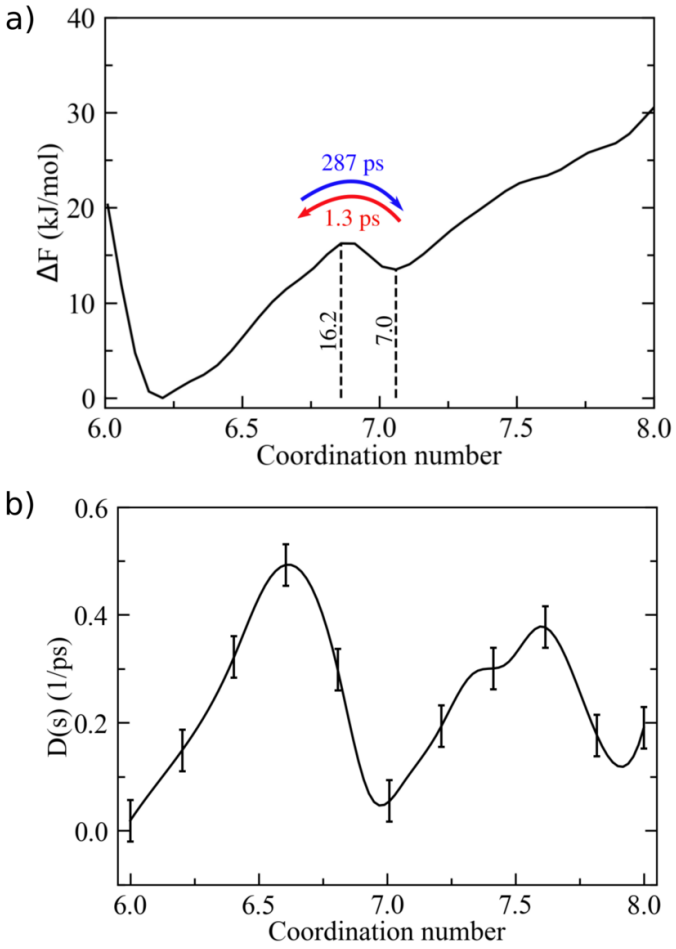

Figure 2.8: a) Free energy landscape of Zn^{2+} coordination in water. ΔF values at relevant points (i.e., local minima/maxima) are reported explicitly. Mean first-passage times corresponding to transitions between adjacent states are also reported as computed from the integration of the FP equation. Standard deviation on F(s) is 1 kJ/mol. b) Position-dependent diffusion coefficient as a function of the coordination number. Error bars correspond to $\pm\delta D$ (see Sec. 2.2.4).

Table 2.2: MFPT for Hg^{2+} coordination in water, computed from long MD simulation (MD), Langevin dynamics (LD), Fokker-Planck integration (FP), Kramers and backward-Kolmogorov (bwKLG) equation (see Methods for details).

	$7 \to 8$ (ps)	$8 \to 9$ (ps)	$8 \to 7$ (ps)	$9 \to 8$ (ps)
MD	0.5 ± 0.2	16.8 ± 0.5	$27 \pm 10 \cdot 10^3$	1.70 ± 0.15
LD	0.7 ± 0.1	21 ± 2	$20 \pm 3 \cdot 10^3$	1.5 ± 0.3
FP	0.7 ± 0.1	20 ± 3	$18 \pm 3 \cdot 10^3$	1.7 ± 0.2
Kramers	0.31 ± 0.06	7.0 ± 0.9	$12.8 \pm 2.0 \cdot 10^3$	0.48 ± 0.08
bwKLG	0.60 ± 0.05	20 ± 2	$17 \pm 3 \cdot 10^3$	1.6 ± 0.3

2.3.2 Predicting water exchange rates

The proposed kinetic approach was then applied to a few cations showing high free energy barriers (> 25 kJ/mol) and, hence, "rare" water exchange events not readily accessible to pure MD simulations. For Hg^{2+}, we observed three main coordination states, namely 7, 8 and 9, where the former was rather unfavorable being less stable by about 27 kJ/mol with respect to state 8 (Fig. 2.9a). In this case, of the two possible routes leading to water exchange in the first coordination shell ($8 \rightleftharpoons 7$ and $8 \rightleftharpoons 9$) only the one based on the associative mechanism appeared feasible. Accordingly, from our 1 μs MD simulation only 32 transitions from the most probable configuration (i.e., 8) to state 7 were observed, while the number of $8 \to 9$ transitions was three orders of magnitude greater. As a result of the poor statistics, the MFPT of the $8 \to 7$ transition could not be reliably obtained from the standard MD simulation (i.e., $\sigma(\tau) = 10 \cdot 10^3 ps$, Table 2.1). On the other hand, upon evaluation of the position-dependent diffusion coefficient (Fig. 2.10) from multiple short MD simulations according to our stochastic model, it was possible to estimate satisfactorily $\tau(8 \to 7)$ at an affordable computational cost (note that accuracy can be systematically improved if required). In particular, we compared favorably the result issuing from the direct backward Kolmogorov-Chapman equation (eq. 2.5), which is in our view the method of choice, to the alternative methods provided by the integration of the FP equation and Langevin dynamics, as reported in Table 2.2. As expected, all stochastic approaches provided consistent results, ($\tau = 18 \pm 3 ns$). The MFPT evaluated via Kramers equation (Table 2.2) for the same transition, however, appeared somewhat underestimated ($\tau = 12.8 \pm 2 ns$), likely due to the underlying approximations discussed above (Sec. 2.2.3). For all other $\tau' s$, easily evaluated by the pure MD simulation (i.e., $\tau \approx 1 - 20 ps$), results matched well the ones of the present kinetic model (Table 2.2), as for the previously considered cations.
Similarly, we tested the predictive capability of our kinetic model towards Cd^{2+}. Three distinct coordination number states emerged from our MetaD simulation (i.e., 6, 7 and 8 in Fig. 2.9b), among which the octa-coordinated water configuration resulted the most thermodynamically stable and the hexa-coordinated one the least populated with a separating barrier of about 29 kJ/mol. As a consequence, the observed number of transitions to the latter state was extremely small (i.e., 12) and a direct estimate of the MFPT from the MD simulation was rather problematic providing roughly the order of magnitude of τ ($\sim ns$, Table 2.1). In this case, the advantage of the stochastic approach proposed in this work was apparent in comparison to the poor statistics affecting the extended MD simulations. For the challenging $7 \to 6$ transition, the kinetic model provided a τ of about 14 ns, while the other transition times resulted at least three order of magnitude smaller and in very

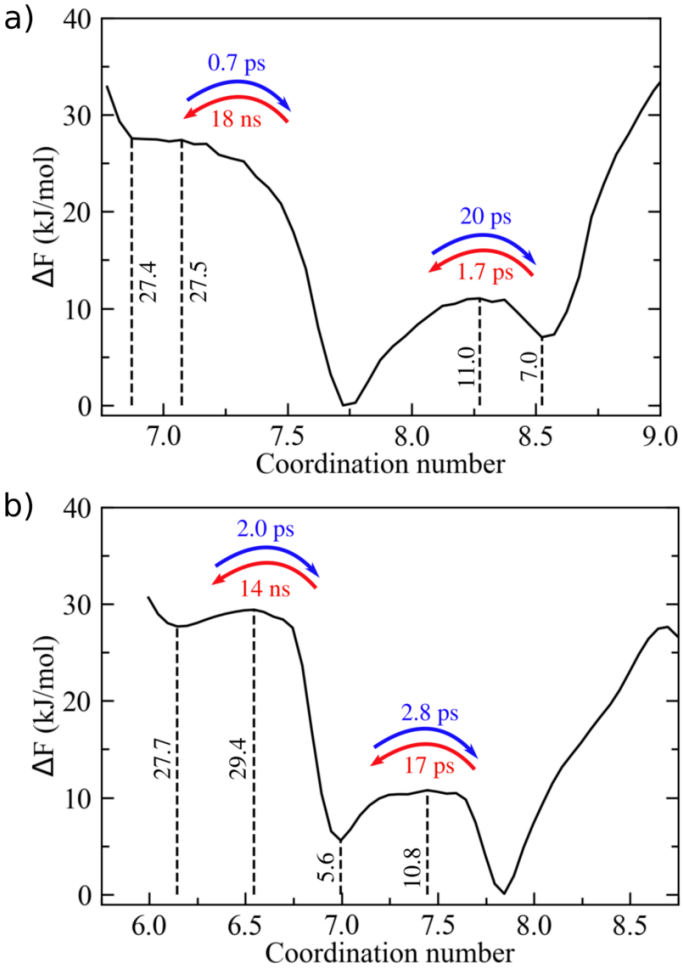

Figure 2.9: Free energy landscape of (a) Hg^{2+} and (b) Cd^{2+} coordination in aqueous solution. Vertical dashed lines indicate energy barriers (local maxima) and stable states (local minima) of interest. MFPTs computed from the integration of the Fokker-Planck equation are also reported as insets. Standard deviation on F(s) is 1 kJ/mol.

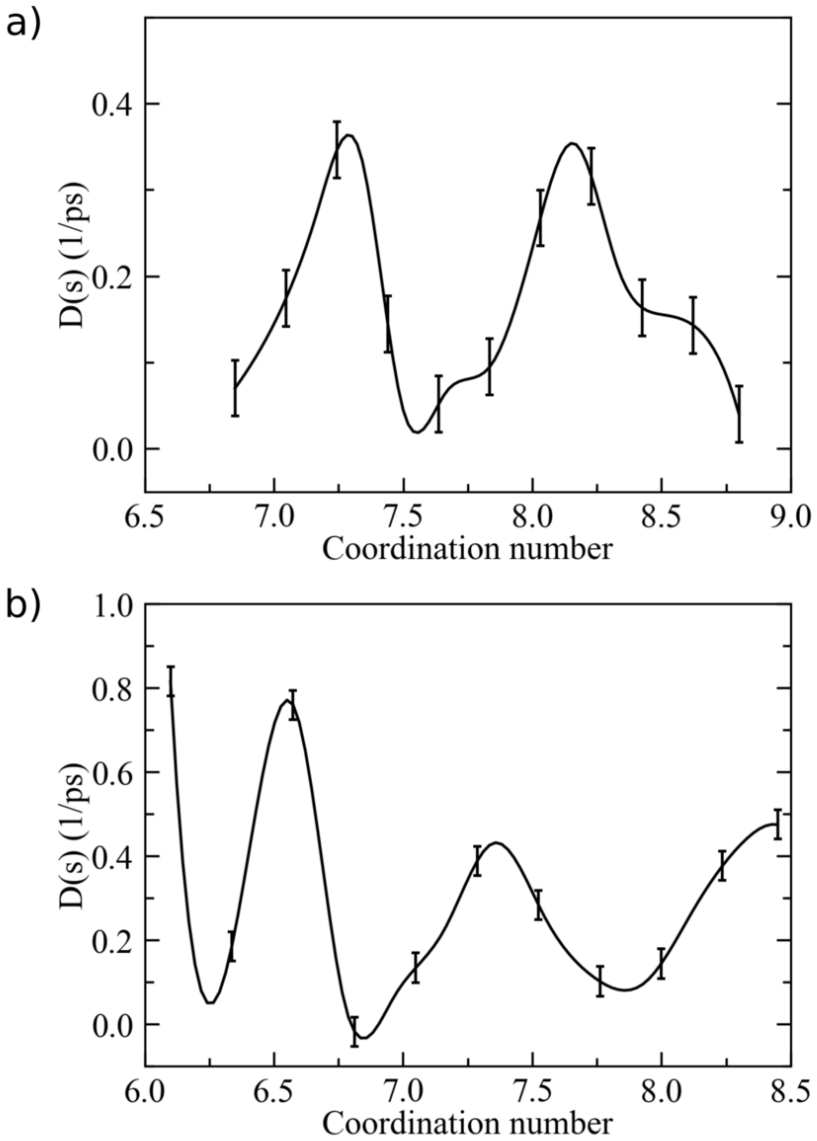

Figure 2.10: Position-dependent diffusion coefficient, $D(s)$, of a) Hg^{2+} and b) Cd^{2+} ion coordination as computed through the method proposed in Sec.2.2.4. Error bars are $\pm\delta D$.

good agreement with directly observed MD results (Table 2.1). It should be pointed out that a close comparison with experiments was not carried out in the present study since this would require a careful consideration of the variety of systems and physico-chemical conditions (e.g., ionic concentration, use of other ligands, temperature, etc..) at which the experiments are typically performed. Nevertheless, we observe that the range of the computed water exchange times for the systems under scrutiny (from $\approx ps$ to $\approx ns$) is well within the findings issuing from past NMR relaxation experiments [175, 200].

2.3.3 On the relationship between diffusion and free energy

While the position-dependent diffusion coefficient ($D(s)$) and the free energy function ($F(s)$) do appear as distinct terms of the present stochastic model and, from the computational viewpoint, are independently obtained before being plugged into the FP equation, it is worth noting that their mutual relationship in a real physical system is significant and should not be overlooked. To better investigate this point, we performed a test simulation of the Hg^{2+} system, as seen above, by applying a bias potential equivalent to the one computed from the MetaD simulation (i.e., the negative of the free energy profile along the s coordinate, $-\Delta F(s)$, see Fig. A.2), so to effectively obtain a barrier-less water exchange process. The idea was to inspect the change in $D(s)$ as a consequence of a significant modification of $\Delta F(s)$, thus highlighting the existing relation between the two ingredients of the kinetic model. In particular, upon applying the bias potential, the system was set free to move between different coordination states (Fig. A.3a). Under such artificial conditions, the resulting diffusion became basically constant ($\sim 0.1 ps^{-1}$) throughout the coordinate number space (Fig. A.3b), a signature of a purely diffusive regime, in stark contrast to the original unbiased system. This finding, in our view, represents a useful warning for those methodologies aiming at obtaining dynamical information from purposely biased systems.

2.3.4 Validation of the coordination number as a reaction coordinate

The Hg^{2+} in water system was also adopted to validate the use of the coordination number, as defined in eq. 2.1, as a suitable reaction coordinate for the description of the water exchange process. As thoroughly discussed by Peters in a thematic review [182], a given collective variable, e.g. based on physical considerations or chemical intuition, could prove useful for describing the kinetics of a dynamical transition between two well-defined molecular states without necessarily being an appropriate "reaction coordinate" for the same molecular process, that is not corresponding to the definition of a minimum free energy pathway and/or not including other relevant coordinates for a proper mechanicistic interpretation of the reaction under examination. However, an effective test to assess the quality of a putative coordinate is represented by the committor analysis, as originally proposed in ref. [109] (see Sec. 2.2.5 for more details). A bell shape distribution of $p(\pi_R)$ as a function of $\pi(s_0)$ and peaked around the separatrix region (i.e., $\pi(s_0) = 0.5$) is regarded as a positive test for a trial reaction coordinate [182]. Tests for the committor analysis of the Hg^{2+} system, when considering both free energy maxima ($p(s = 7.03)$ and $p(s = 8.35)$), were carried out and results are depicted in Fig. 2.11. The obtained distributions favorably support the choice of the present coordinate to follow the water exchange process in the first solvation shell around aqua ions.

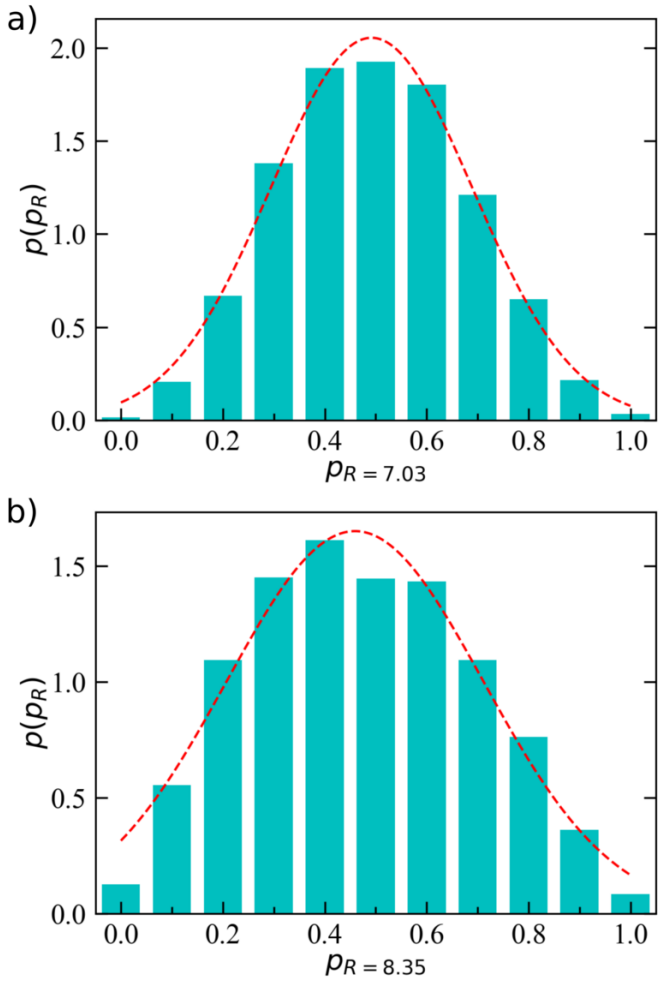

Figure 2.11: Committor probability distribution for Hg^{2+} coordination in water computed from an ensemble of short MD simulations. 1200 starting configurations were taken at the a) $s = 7.03$ and b) $s = 8.35$ barrier top. Then, 100 replica simulations were carried out for each configuration. A Gaussian fit of the probability distribution is also provided (red dashed line).

2.3.5 High ionic concentration

As a further test, we considered a relatively higher concentration (0.5M) of mercury ions in aqueous solution to assess the robustness of the proposed computational approach under such conditions. First, we observed a noticeable change of the main ion-water configurations in the first solvation shell, since a much larger range of coordination numbers around each Hg^{2+} became available (i.e., from 1 to 9, see Fig. 2.12a). In fact, at 0.5M concentration, ions compete each other more effectively for acquiring coordinating water molecules, which are now much less abundant with respect to the previous dilute solution. In particular, the effect of the counterions (i.e., Cl^-) on the first water shell of Hg^{2+} is also greatly enhanced, since ionic couples can form (and break apart) more easily at this concentration. As a result, the free energy landscape of ion coordination showed a noticeable rough surface characterized by multiple local minima (i.e., 9 coordination states) within a limited range of energy (about 10 kJ/mol). Also, dividing energy barriers were significantly reduced to about 10-15 kJ/mol between adjacent coordination states. As a consequence, a single preferential coordination state could not be identified at this concentration. Nonetheless, we again analyzed water exchange dynamics from both pure MD simulations and from the kinetic model. In the latter case, we obtained the position-dependent diffusion constant, as depicted in Fig. 2.12b, which overall reflected the same oscillating trend of the $F(s)$. As reported in Table A.1, water exchange was observed to occur rather frequently among all states, with MFPTs ranging from ~ 10 ps to ~ 80 ps. Once more, the transition times issuing from the stochastic approach revealed, overall, a good agreement with the direct MD estimates, taking into account statistical noise. This finding supported the use of the present computational method for studying ionic solutions at variable concentrations.

2.4 Conclusions

In this work, we presented a computational protocol (as sketched in Fig. 2.13) rooted into molecular dynamics, enhanced sampling and stochastic methods to obtain a comprehensive picture of solvent coordination and exchange around ions in solution. Our strategy starts from the evaluation of the free energy landscape as a function of the ion coordination number treated as a continuous collective variable. The free energy profile provides a "fingerprint" of ion coordination in solution by showing quantitatively the existing complex equilibrium between different solvent coordination states. As a result, the most probable first-shell ion-water configurations, the relative free energy stability and the corresponding transition barriers are determined. In a second step, the transition rate matrix describing the dynamical interchange of ion coordination is built up and the position-dependent diffusion constant is evaluated from multiple short MD simulations along the coordination number. At this point, it is worth noting that such a task, the most computationally intensive of our procedure, can benefit from fully independent and parallel MD runs. Then, the computed free energy and diffusion functions are plugged into a Fokker-Planck model to derive the (long-term) time evolution of ion coordination and solvent exchange at timescales not easily accessible to standard MD techniques. Solvent exchange rates are obtained in terms of mean-first-passage times between coordination states, thus providing a further important observable of ion microsolvation to be compared with available experiments. The computed rates are generally affected by a reasonably small error (within 10-20%), especially in view of the extremely wide range of timescales known from the literature (from 10^{-12} to $> 10^6$ s). Note, however, that the accuracy of the exchange rate estimates can be improved systematically within the present protocol, while the reliability of the results is

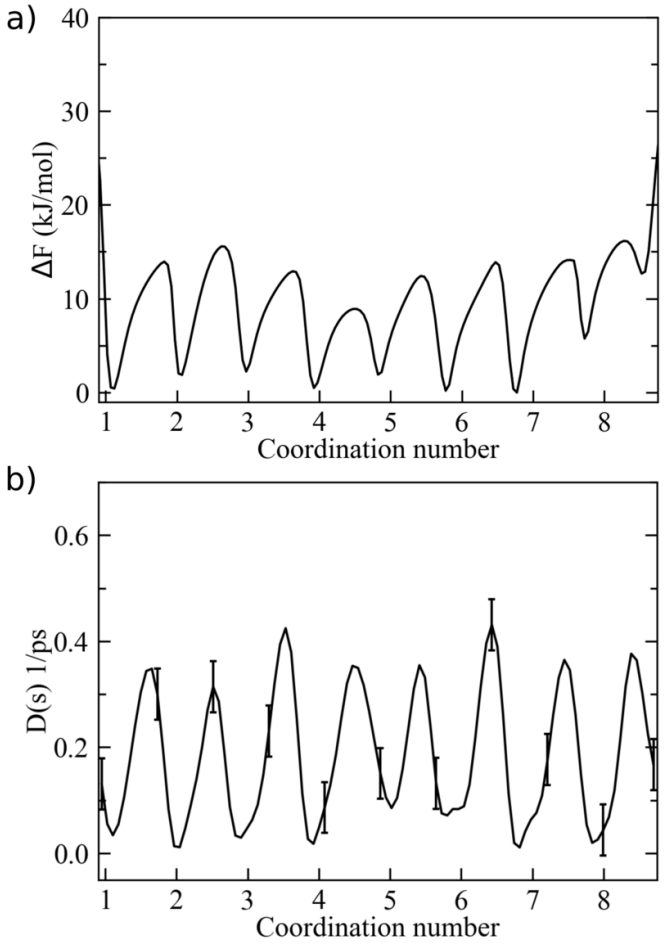

Figure 2.12: a) Free energy landscape of Hg^{2+} coordination in water from 0.5M $HgCl_2$ aqueous solution, as issuing from pure MD simulations. b) Position-dependent diffusion coefficient as a function of the coordination number. Error bars correspond to $\pm\delta D$ (see Sec. 2.2.4).

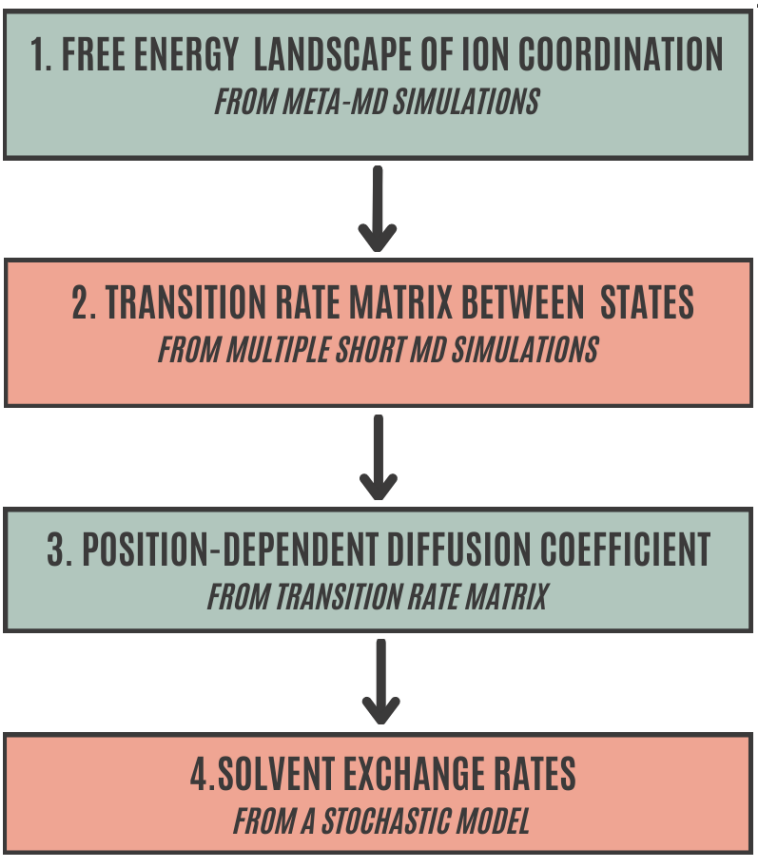

Figure 2.13: Workflow of the proposed computational protocol to effectively compute ion-water coordination and exchange rates in ionic solutions. A detailed description of the protocol is provided in the text.

closely related to the underlying force field employed. On this regard, we believe that our computational approach can be fruitfully exploited to investigate the agreement between current molecular models and experiments (e.g., NMR relaxation measurements) at an affordable cost. Note that comparison with experiments were not explicitly considered in the present methodological study, but will be investigated in future applications. Eventually, this approach can be also employed in force field development, so to optimize ion-solvent intermolecular potentials towards an additional, usually overlooked, parameter (i.e., solvent exchange rate).

Notably, the coordination number, adopted in this work as an effective coordinate for monitoring the ion-water coordination, passed successfully the committor analysis test and, therefore, can be regarded as a suitable and physically sound reaction coordinate for the process [182]. Besides, another advantage of this coordinate is that it is unbiased towards any specific water exchange mechanism, in contrast to other coordinates (e.g., the ion-water distance) typically employed in previous computational studies.

A further consideration that deserves some comments concerns the assumption of Markovianity. Here, the dynamical process is defined as Markovian given a suitable discretization of the selected coordinate (i.e., the coordination number), according to the general principle that even a non-Markovian process can turn Markovian at some coarse-grained description (that is, whenever there is a timescale gap between the relevant coordinate and the other degrees of freedom of the system). In this context, this seems justified by the fact that molecular collisions occur at a much faster timescales ($\sim$fs) than first solvation shell changes (at least $\sim$ps). Moreover, it is remarkable that exact MFPTs (and rates) can be computed from average transition rates, as obtained using approximate Markovian models, irrespective of the actual distribution of the lifetimes (i.e., the exact non-Markovian trajectory), as discussed in ref. [184]. In other words, the long time evolution of the (approximate) stochastic trajectory nicely corresponds, on average, to the detailed MD trajectory, as projected onto the same reaction coordinate.

Chapter 3

Thermodynamics of Metal-Acetate Interactions

In this Chapter, the development of parameters for the non-bonded potential between metal ions and acetate, as detailed in section 1.1.2, is discussed. As highlighted in the Introduction, accurately reproducing interactions between metal ions and ligands is a significant challenge. To utilize classical force fields effectively, it is crucial to verify their accuracy before making any quantitative assumptions. This project aims to improve current classical force fields between metal ions and acidic residues, driven by the need for greater accuracy in these types of interactions that frequently occur in biomolecules. The parameterization of the 12-6-4 LJ non-bonded potential, as described in section 1.1.2, adopted a top-down approach. This involved fitting the parameters to experimentally available binding free energy values for metal-acetate complexes. This Chapter is based on an article already published by Jafari et al. [201] that I coauthored.

3.1 Introduction

A significant number of metalloproteins are available in the Protein Data Bank (PDB)[32] with at least one acidic residues (i.e., aspartate and glutamate) coordinating a metal ion. Indeed, negatively charged residues are also essential for the function and stability of metalloenzymes and metalloproteins [202, 203]. However, there is a lack of parameters available for negatively charged residues while chelating to metal ions within the 12-6-4 LJ nonbonded model. Here, we used acetate ($CH3COO^-$) as a minimal molecular model representing the sidechains of these residues. Our results demonstrate that the standard 12-6 LJ and 12-6-4 LJ models cannot reproduce the experimental interaction energies of eleven metal ions with the acetate molecule. In the present study, we provide the adjusted C_4 parameters for the 12-6-4 LJ nonbonded model that can be used to accurately model the interactions between the COO-group and various metal ions in aqueous solutions, using three different water models (namely, TIP3P, SPC/E, and OPC).

3.2 Methods

3.2.1 Refining the C4 terms in the 12-6-4 LJ-type nonbonded model

Due to the popularity of nonbonded models and the recent advances in the accuracy of the 12-6-4 model, we conducted a fine-tuning of the C_4 term parameters specifically for acidic residues and metal ions. With that in mind, to achieve the experimental binding free

energies for the complex of a metal ion with an acetate molecule, we modified the chelator atom polarizability, specifically the polarizability of the acetate oxygen atoms (referred to as α_0 in the following equation, the angle generated by the metal ion between the induced dipole and electronic field is defined by θ)

$$C_4 \approx \frac{1}{2}\alpha_0 \left(\frac{q}{4\pi\epsilon_0\epsilon_r}\right)^2 cos\theta \tag{3.1}$$

The equation below represents the 12-6-4 LJ model in the AMBER force field [54, 204, 82]. In this equation, r_{ij} represents the distance between particles i and j. The $\frac{C_4^{ij}}{r_{ij}^4}$ part represents the C4 term, which describes the ion-induced dipole interaction and is proportional to r^{-4}. The final part of the equation represents the coulomb pair potential, where e represents the proton charge and Q_i and Q_j denote the partial charges of atom i and j, respectively

$$\begin{aligned} U_{ij}(r_{ij}) &= \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6} - \frac{C_4^{ij}}{r_{ij}^4} \\ &= 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^6 - \kappa(\sigma_{ij})^2\left(\frac{\sigma_{ij}}{r_{ij}}\right)^4\right] \end{aligned} \tag{3.2}$$

where to reach the right hand side of equation 3.2 we have used the fact that $C_4^{ij} = \kappa C_6^{ij}$. Initially, we constructed a topology file by employing classical 12-6 LJ nonbonded parameters. We then used this topology file to generate a new one, incorporating default C_4 terms. Since no optimized C_4 parameters were available for the acetate-metal ion interaction, we generated a revised topology file using select polarizability values for the acetate target atoms. We conducted an unbiased MD simulation using the newly generated topology file to determine the equilibrium distance between acetate and the metal ion. This equilibrium distance was the primary distance for undertaking US simulations (see section 1.2.2). The umbrella sampling method, coupled with the weighted histogram analysis algorithm (WHAM) [205] has been widely used in computational studies to determine the potential of mean force (PMF) energy of metal ions in various environments [83, 206, 207, 208, 209]. We systematically examined various C_4 values and compared the calculated interaction energies with experimental interaction energies as our benchmark.

3.2.2 Potential of Mean Force and Molecular Dynamics Simulations

The PMF obtains the free energy profile of a system with respect to generalized coordinates such as Ramachandran angles or the distance between two atoms. The PMF free energy profile serves as a visual representation that reveals the stable conformations of a given molecule, as well as the barriers the molecule needs to overcome when transitioning from one state to another along a defined reaction coordinate.

To assess the convergence of PMF calculations, we initially performed umbrella sampling (US) using a 2 ns simulation length for each window. Subsequently, upon obtaining the experimental binding free energy, we conducted the US method for a given metal ion with simulation lengths of 2 ns, 4 ns, 6 ns, 8 ns, and 10 ns for each window. While we opted for a 4 ns simulation length for each window in SPC/E water [194], we discovered that the optimal simulation length for each window in TIP3P [210] and OPC [211] water is 6 ns. The reaction coordinate used was the distance between the carbon atom of the acetate carboxyl group and the metal ion. The sampling window width ranged from 0.05 Å($<$ 5Å) to 0.1 Å($>$ 5Å) , depending on the metal ion's proximity to acetate. The starting point

for sampling ranged from 1.8 to 2.5 Å, determined by the metal ion's radius. To assess the precision of our computed binding free energies, three replicas of PMF calculations were performed. The replicas showed an acceptable margin of error, measuring less than 0.35 kcal/mol. To further validate the obtained results using the US algorithm as implemented within the AMBER software package, we also carried out additional PMF calculations using the US method as implemented in PLUMED software [212, 213]. The latter binding free energies were consistent with those acquired from the standard implementation in AMBER. Note that in the following we will refer to either AMBER or PLUMED PMF data according to the procedure described above. All the MD simulations were performed using the AMBER 20 package [214] and the PLUMED package version 2.8.2. A four-step method of minimization was employed to minimize the system's energy. Only water molecules were minimized in the first step, while the acetate molecule and the metal ion were restrained. In the second step, another minimization was performed on the acetate hydrogen atoms and water molecules. The third step involved a minimization of acetate non-hydrogen atoms and water molecules. Finally, in the last step, a minimization was performed on the entire system. In each stage, 10,000 steps of the steepest descent were used. Afterward, the system was gradually heated from 0 to 300 K over 1 ns in the NVT ensemble. The heated system was then used for a 4 ns simulation in the NpT ensemble. The Berendsen barostat,[215] employing a pressure relaxation time of 5 ps, was applied to maintain the system's pressure. The Langevin thermostat was used as the temperature coupling algorithm. To constrain bond lengths, the SHAKE algorithm [84] was employed, and periodic boundary conditions were applied in all three dimensions (x, y, and z). A cutoff distance of 10 Åwas applied to handle non-bonded interactions. The data analysis and visualization in the present work were performed using the Visual Molecular Dynamics (VMD) [216] and Python packages [217].

3.2.3 MD Simulation of the Bacterial Glyoxalase I (Glx I) Metalloprotein

The initial protein structure was obtained from the PDB (PDB ID 1F9Z) [218]. This structure was used to generate a simulation system comprising of the OPC water model and the metaloprotein. A five-stage minimization process was conducted, with 10000 steps of steepest descent employed in each stage to optimize the simulation system. Afterward, a 5 ns NVT ensemble was performed to gradually raise the temperature from 0 to 300 K. Following this, the system underwent equilibration for 5 ns using the NpT ensemble. The equilibrated system was followed by 200 ns MD simulation using AMBER 20. The SHAKE algorithm was applied to constrain all bonds, while the Langevin dynamics algorithm was used to maintain the temperature at 300 K. Isotropic pressure scaling with the Berendsen barostat was utilized to control pressure.

3.3 Results and Discussion

In recent years, researchers have dedicated efforts to developing and optimizing polarizable force fields for transition metal ions [219, 220, 221]. In some instances, they have successfully reproduced experimental and ab initio calculation results. Despite polarizable force fields being more complicated and accurate than fixed-charge force fields, there are cases where the results from polarizable force fields are not as good as those from fixed-charge force fields [222, 223, 224]. One potential reason for this discrepancy could be the challenging nature of optimizing polarizable force fields [224].
A recent investigation into the binding of acetate to divalent metal ions, such as Mg(II),

Table 3.1: The experimental binding free energy (Exp. ΔG) of each metal ion with acetate, the ion electron configuration (Elec. Conf.), and the ion radius (r) in picometers.

Ion	Exp. ΔG (kcal/mol) [233, 234, 235]	Elec. Conf.	r(pm) [236]
Ni(II)	-1.95	[Ar] $3d^8$	69
Mg(II)	-1.33	[Ne]	72
Cu(II)	-3.01	[Ar] $3d^9$	73
Zn(II)	-2.16	[Ar] $3d^{10}$	74
Co(II)	-1.88	[Ar] $3d^7$	75
Cu(I)	-2.78	[Ar] $3d^{10}$	77
Fe(II)	-1.91	[Ar] $3d^6$	78
Mn(II)	-1.91	[Ar] $3d^5$	83
Cd(II)	-2.63	[Kr] $4d^{10}$	95
Ca(II)	-1.27	[Ar]	100
Ag(I)	-1.00	[Kr] $4d^{10}$	115

Ca(II), and Zn(II), indicates that the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) force field [225] has the potential to enhance the prediction of binding constants for acetate-cation complexes. However, it appears to overestimate the binding strength of the metal ions, leading to results that do not align quantitatively with experimental data [226] Zhifeng Jing et al.[227] used the AMOEBA force field to investigate the interactions between metal ions and protein compounds. The results showed that although AMOEBA predicts the solvation free energy of acetate within a satisfactory range, some discrepancies are noted for acetate-metal ion interaction energies, particularly for the polarization energy at short intermolecular distances.
Researchers also evaluated the variation of interaction energies of different protein zinc binding sites [228] and calcium-acetate [229] using QM calculations as the benchmark to evaluate polarizable force field accuracy. Their results showed that the CHARMM-Drude polarizable model [230, 231, 232] agrees with the QM benchmark, with errors around 10-80 kcal/mol; however, there were some remaining discrepancies. In the current work, we applied the AMBER ff14SB nonpolarizable force field to predict the binding free energy of acetate with eleven metal ions. Our results indicated that the re-parametrized 12-6-4 LJ model can reproduce the experimental interaction energy of the acetate-metal ion complexes within a margin of error of less than 0.35 kcal/mol.

3.3.1 Effect of Water Model and Polarizability on Binding Free Energy Calculations

Table 3.1 gives the experimental free energies of association between 11 divalent metal ions and acetate [233, 234, 235] along with their electronic configuration and ionic radii[236]. Overall, there appears to be no consistent trend between the polarizability value and the binding free energy. This phenomenon is also observed with metal ion-imidazole complexes[83], implying that other interactions, such water-water interactions surrounding the ions, play a significant role in the binding free energy. It is important to note that the selection of water molecules employed in simulations also holds a crucial influence on the interaction free energy. Table B.1 and Table B.3 gives the computed free energies between the 11 metal ions and acetate in the three-point TIP3P and SPC/E water models, respectively and the results for the OPC water model are given in Table 3.2. Interestingly, our findings indicate

Table 3.2: The calculated binding free energy of each metal ion – carboxylate complex in OPC water. Average column show the results from three replicas performed using AMBER US technique. The free energy values are in kcal/mol.

Ion	Exp. ΔG	Average*	Standard 12-6-4	12-6	PLUMED†
Ni(II)	-1.95	-1.72±0.02	-10.54	-9.27	-2.10
Mg(II)	-1.33	-1.67±0.08	-4.38	-7.68	-1.44
Cu(II)	-3.01	-3.16±0.24	0.10	-8.07	-2.60
Zn(II)	-2.16	-2.08±0.27	-6.45	-8.5	-2.08
Co(II)	-1.88	-1.86±0.28	-4.93	-8.64	-2.30
Cu(I)	-2.78	-2.99±0.28	-1.81	-2.03	-2.72
Fe(II)	-1.91	-1.92±0.39	-4.77	-8.43	-2.19
Mn(II)	-1.91	-1.82±0.16	-2.50	-6.79	-1.82
Cd(II)	-2.63	-2.78±0.02	-0.23	-5.81	-2.65
Ca(II)	-1.27	-1.21±0.04	-2.11	-3.84	-1.44
Ag(I)	-1.00	-0.90±0.14	0.97	0.34	-1.11

* The error is the std deviation calculated over 3 replicas of US performed in AMBER software

† Estimated binding free energy using US coded in PLUMED software.

that reproducing the experimental interaction energy requires employing lower polarizability values for the chelator atom within the OPC water model compared to the TIP3P and SPC/E water models (compare Tables B.2 and B.4, with 3.3).
The OPC water model was designed for improved accuracy in mimicking the properties of liquid water compared to TIP3P and SPC/E, which was achieved through the incorporation of additional sites on the water molecule. This improvement includes a stronger dipole and quadrupole, enabling water molecules to form stronger interactions with other compounds than the weaker dipole and quadrupole present in the three-point water models [211]. The heightened polarization effects of the OPC water model can significantly impact the prediction of binding free energies by influencing hydrogen bonding, dipole-dipole interactions, and even van der Waals forces. Therefore, when utilizing OPC in systems where water molecules play a pivotal role in mediating interactions, we have found that lower polarizability values for the chelator atoms become essential to accurately replicate experimental values.

3.3.2 Acetate binding mode in different water models

When studying metal-acetate systems we have two binding modes that could play a role in the coordination of the metal ion by acetate. In the monodentate binding mode only one of the carboxylate oxygen atoms coordinate to the metal ion, while in the bidentate mode both carboxylate oxygen atoms coordinate the metal ion. Importantly, the binding mode of metal ions in a complex with carboxylate or acetate varies depending on the specific ligands involved and the binding site composition. The structures obtained/predicted by X-ray crystallography, FT-IR spectroscopy, and UV-vis spectroscopy show that acetate ligands are bonded to Zn(II) in a monodentate binding mode [237]. Structural studies on zinc-containing proteins have also revealed that the carboxyl group coordinates with zinc in bidentate, monodentate, and even intermediate between monodentate and bidentate fashions [238]. However, DFT calculations indicate that the preference between monodentate and bidentate modes for Zn(II) depends on the properties of the binding site coordination

Table 3.3: Polarization α_0(Pol.) applied to equation 3.1 for each metal ion and related computed C_4 values used to reach experimental binding energies of ion-carboxylate complex for OPC water model.

Ion	Pol. (Å^3)	C_4 kcal/(mol Å^4)
Ni(II)	-0.169	-24.81
Mg(II)	-1.000	-87.95
Cu(II)	0.894	180.16
Zn(II)	0.432	67.31
Co(II)	0.357	50.43
Cu(I)	1.469	16.28
Fe(II)	0.369	39.35
Mn(II)	0.469	56.84
Cd(II)	0.819	124.21
Ca(II)	0.419	24.95
Ag(I)	1.650	94.84

environment. For instance, in a four-coordinate complex that is net positively charged, both mono- and bidentate coordination of carboxylate have similar energy. However, when additional ligands, with negative charges are introduced, the preference shifts towards the monodentate binding mode [238].
Previous studies show that both Ca(II) and Mg(II) prefer to coordinate with carboxylates in a monodentate binding fashion [239]. In contrast, bidentate binding becomes the preferred option when the attraction between the metal cation and the ligand's second oxygen is stronger than the interactions with nearby ligands [239]. In a combined experimental and computational approach it was suggested that Zn(II)-acetate binding likely involves both monodentate and bidentate geometries, while Mg(II)-acetate is predominantly in a monodentate binding mode [226].
A survey of Protein Data Bank (PDB) structures [240, 241] shows that Mg(II) binding sites have more monodentate carboxylates stabilized by coordinating water molecules while Ca(II) sites have more bidentate carboxylates stabilized by backbone interactions. In other words, water molecules stabilize monodentate binding in Mg(II) complexes, while backbone groups destabilize monodentate binding in Ca(II) complexes [241]. An X-ray crystallography study on different cobalt and nickel complexes in the presence of metal ion ligands, including tetraazamacrocycles, shows that smaller metal ions such as Co(III) bind to acetate in a bidentate mode, while larger metal ions such as Co(II) and Ni(II) in cross-bridged ligands tend to coordinate in a monodentate fashion. However, acetate prefers to have a bidentate binding mode with Co(II) and Ni(II) in the presence of unbridged ligands [242]. Overall, both experimental and QM studies consistently support that acetate and carboxylate ligands do not exhibit a unique monolithic behavior when binding to metal ions [237, 238, 239, 226, 240, 241, 242, 243, 244, 245, 246, 247]. Various factors, including the coordination number, type of metal ion, its valence, charge, size, and the composition of the binding site, can impact the tendency towards either bidentate or monodentate binding [239, 243, 244]. Despite that, an examination of structures in the PDB database and the Cambridge Structure Database (CSD) reveals that the carboxylate group more commonly binds to various metal ions through a monodentate binding fashion in both small molecules and metalloproteins [239, 240, 241, 248, 249, 250]. The type of water model can impact

Table 3.4: Preferred binding mode for each metal ion in the OPC water models. The two columns indicate the percentage of monodentate and bidentate binding modes in the acetate-metal ion complex calculated using the PMF profile for each acetate-metal ion complex and applying the Boltzmann distribution based on the minimum energy of each binding mode state.

Ion	Acetate Binding	
	Monodentate	Bidentate
Ni(II)	100.00%	0.00%
Mg(II)	100.00%	0.00%
Cu(II)	99.81%	0.19%
Zn(II)	100.00%	0.00%
Co(II)	100.00%	0.00%
Cu(I)	98.38%	1.62%
Fe(II)	100.00%	0.00%
Mn(II)	88.96%	11.04%
Cd(II)	14.29%	85.71%
Ca(II)	41.08%	58.92%
Ag(I)	29.20%	70.80%

both the calculated binding free energy and the binding mode of the acetate molecule in MD simulations. This impact arises from a combination of factors, including the electron configuration of the metal ion, its ionic radius, and the characteristics of the water models employed in the simulations. The properties of the selected water model can influence interactions between ions and solvents, as well as between ions and ligands. The variation in these interactions is why different polarizability values are needed to reproduce the experimental binding free energy in the OPC, TIP3P, and SPC/E water models.
Our findings demonstrate that the behavior of acetate, particularly in terms of monodentate and bidentate coordination, depends on both the nature of the metal ion and the specific water model used (Tables 3.4 and B.5).

We find that certain metal ions, including Cd(II) and Ca(II) exhibit consistent behavior, displaying a bidentate binding mode regardless of the water model employed. In TIP3P, the calcium ion exhibits a bidentate binding mode, but the monodentate form is nearly isoenergetic. While in SPC/E and OPC, it displays a bidentate preference with a slightly higher lying monodentate form (monodentate $\approx +0.75$ kcal/mol higher). On the other hand, Cd(II) demonstrates a bidentate (monodentate $\approx +1.5$ kcal/mol higher) mode consistently across all water models. Figure 3.1 shows the detailed free energy profile of the Cd(II) – acetate complex defined by the interaction between the acetate oxygen atoms and the metal ion using the 12-6-4 LJ nonbonded model. An examination of PDB structures uncovers that Cd(II) participates in both bidentate and monodentate binding modes with the carboxylate groups of metalloproteins [240]. Our results demonstrate that this metal ion displays a $\approx$ 1.5 kcal/mol preference for bidentate binding with the acetate molecule. Cu(II) exhibits different behaviors depending on the water model used. In three-point water models (TIP3P and SPC/E) it prefers a bidentate interaction (with a distinct monodentate minimum $\approx$ 0.1-0.5 kcal/mol higher in energy), while in OPC (4-point) water it prefers monodentate binding and there is only a slight shoulder at the location where the bidentate complex would appear. The reason for this is not entirely clear but one can hypothesize that differential

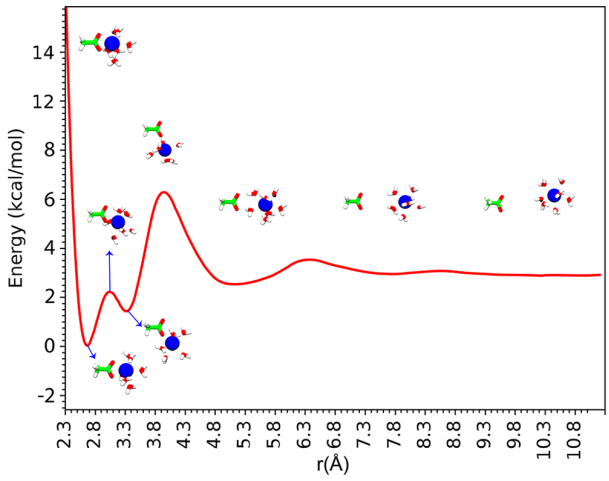

Figure 3.1: Depiction of the PMF energy profile of the Cd(II) and acetate complex in the TIP3P water model, accompanied by snapshots at various points along the profile.

water-water interactions around the ions may play a role. This is discussed further below.

Acetate binding mode in water models with three interaction sites

Overall, we noticed consistent binding behavior of the acetate molecule in both the TIP3P and SPC/E water models (Table B.5). This outcome was not surprising since both water models are characterized by three-point configurations with nearly identical O-H bond lengths and H-O-H angles [83]. Both predict the acetate molecule to prefer to coordinate with Ni(II), Mg(II), Zn(II), Co(II), Fe(II), and Mn(II) ions in a monodentate fashion (Figure B.1). Previous experimental studies have demonstrated that the binding mode of the carboxylic group with Ag (I) can vary from bidentate to monodentate depending on the ligands [251, 252]. Despite that, our results predict that acetate shows a ≈ 0.75 kcal/mol preference for bidentate binding mode with Ag (I), featuring a shoulder at the monodentate position (Figure B.1). Cu(I) prefers monodentate coordination but has a clear shoulder for the bidentate coordination mode (bidentate ≈ 1.5 kcal/mol higher in energy shoulder). Overall, we find that most divalent metal ions, including Ni(II), Mg(II), Zn(II), Co(II), Fe(II), and Mn(II) as well as Cu(I), tend to adopt a monodentate binding configuration with the ligand in all three water models, exhibiting a preference of approximately 1.5 to 3 kcal/mol (≈ 0.3 and ≈ 0.1 kcal/mol preference for Mn(II) in TIP3P and SPC/E, respectively). This trend aligns with the findings from statistical analyses conducted on structures from the CSD and PDB databases [239, 240, 241, 248, 249, 250, 253]. However, for Ag (I), Ca(II), and Cd(II) the bidentate binding mode is preferred in both TIP3P and SPC/E water models. While Cu(II) in these water models prefers the bidentate binding mode with acetate (Figures B.1 and B.2), in the OPC water model this metal ion shows a preference for monodentate binding (bidentate shoulder ≈ 3 kcal/mol higher in energy). Experimental studies indicate that copper can establish diverse complexes with carboxylate ligands, commonly serving as bidentate ligand [247]. However, there has also been the observation of an ongoing shift between monodentate and bidentate bonding modes in carboxylate [246]. Additionally, in the Cu(II) complex with Hmbm (1-methyl-1H-benzo[d]imidazol-2-yl) methanol, acetate coordinates with the metal ion in a monodentate fashion [237]. In fact, amongst divalent metal ions in the TIP3P and SPC/E water models, the size of the ion plays an important role in their binding mode with acetate. In comparison to other

metal ions, specifically Ca(II) and Cd(II) the larger radii of these ions result in bidentate coordination, facilitated by the available space for engaging with multiple oxygen atoms from the acetate. Conversely, the smaller ions, excluding Cu(II), exhibit a higher degree of selectivity, potentially favoring monodentate binding configurations (Table B.5).

Acetate binding mode in water models with four interaction sites

In the OPC water model, acetate binds to the majority of metal ions, including Ni(II), Mg(II), Cu(II), Zn(II), Co(II), Fe(II), Mn(II), and Cu(I) through monodentate coordination, showing a preference of approximately 1 to 1.5 kcal/mol for monodentate in the case of Cu(II), Mn(II), and Cu(I) over the bidentate coordination mode (Figure 3.2). As discussed, statistical analyses of structures from both the PDB and CSD databases consistently show that most metal ions often favor a monodentate binding mode with carboxyl groups. Remarkably, when using the OPC water model, which more accurately mimics liquid water properties, the results are closer to the experimental data, suggesting a tendency towards monodentate acetate binding with metal ions. As noted previously the binding behavior of metal ions depends on various factors [231, 232].
As shown in Tables 3.4, B.5 and Figures 3.2, B.1, B.2, acetate exhibits a change in behavior when forming complexes with Cu(II). This change is evident as the ligand alters its coordination mode from bidentate to monodentate. In the case of Zn(II), Co(II), and Fe(II), the shoulder at the bidentate position disappears. This suggests that in the presence of OPC water model, these metal ions shift from their initial monodentate (bidentate $\approx$ shoulder 2-3 kcal/mol higher in energy) binding mode to a monodentate binding mode. Although carboxyl groups typically exhibit monodentate coordination with metal ions, [239, 240, 241, 248, 249, 250, 253] it is important to note that there is no consistent trend in the binding modes of carboxyl groups with metal ions. Both bidentate and monodentate coordination have been observed in various acetate-metal ion complexes, [237, 238, 239, 226, 240, 241, 242, 243, 244, 245, 246, 247] and no experimental data is available for the preferences of a single acetate coordinating with metal ions in aqueous solution.

3.3.3 Binding free energy and the chelator atoms polarizability

The polarizability of the chelator atom plays a pivotal role in accurately reproducing the experimental binding energy. This parameter is influenced by various factors, including the ion charge, the number of unpaired electrons in the ion's outermost subshell, solvent effects, and ligand structure. In our study, we obtained nearly identical polarizability values for the closely related water models, TIP3P and SPC/E. Overall, there was a slight increase in the applied polarizability for the chelating atom within the SPC/E model compared to that of the TIP3P water model. Significantly, in order to effectively replicate the experimental binding energy between metal ions and acetate, it was necessary to employ a reduced polarizability value for the ligand chelator atoms in the OPC water model. When considering alkaline earth metal ions, it becomes evident that ions with larger radii require higher polarizability for the chelator atom to accurately reproduce the experimental binding energy within all three water models (see Tables 3.3, B.2 and B.4). Another significant implication is observed for Ca(II) and Mg(II), where the chelator atom with a bidentate binding mode needs greater polarizability. To observe this effect, one can compare the Ca(II) and Mg(II) PMF energy profiles presented in Figures 3.2, B.1, B.2 with the corresponding polarizability values listed in Tables 3.3, B.2 and B.4. This trend in ionic radius holds true for ion families sharing similar chemical properties, such as Zn(II) and Cd(II), Cu(I) and Ag(I),

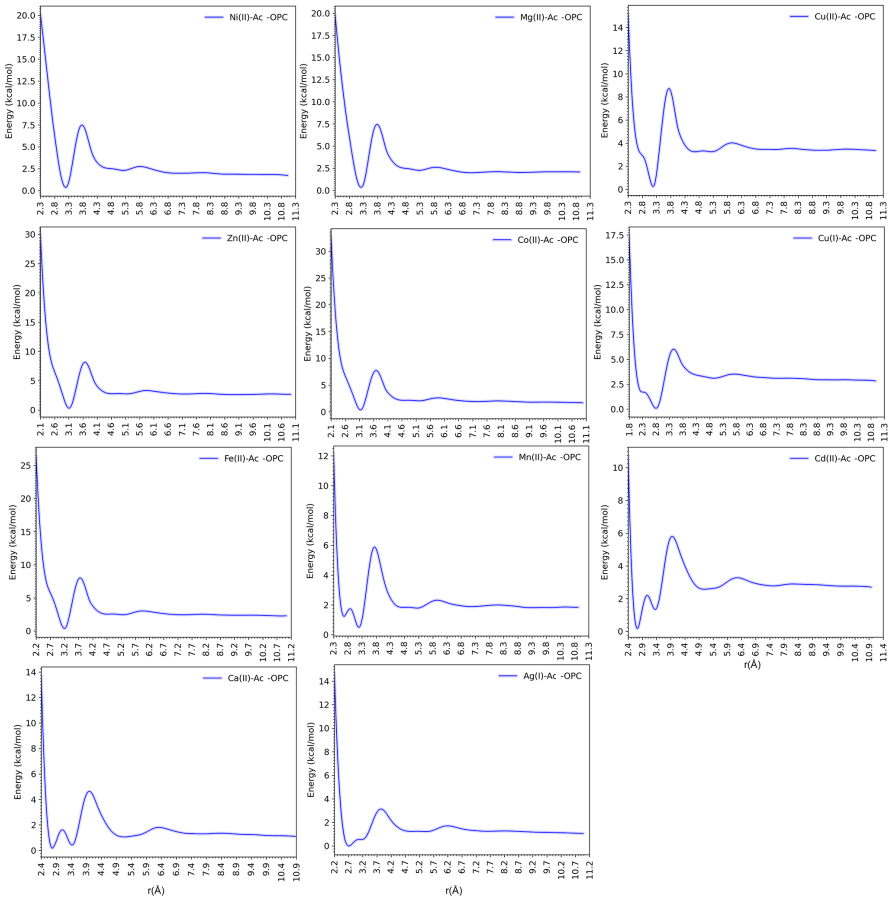

Figure 3.2: The PMF free energy profiles of metal ion-acetate complexes in the OPC water model. Ac stands for the acetate molecule. The first local minimum, occurring at approximately 2.8 Å (2.3 Å for Cu(I)), corresponds to the bidentate binding mode. The second local minimum, observed at around 3-3.5 Å (2.8 Å for Cu(I)), shows the monodentate binding mode.

and Mg(II) and Ca(II). Essentially, ions with larger radii lead to higher polarizability on the chelator atoms. In a previous work, this behavior was observed for the nitrogen atom of imidazole [83]. The polarizability of the chelator atom in the acetate ligand was found to be directly proportional to the radius of the ions within the same family. This relationship was applicable to divalent metal ions in three-point water models. In the case of OPC water, this correlation was evident only for monovalent metal ions, implying larger monovalent metal ions required higher polarizability. However, in systems with three interaction site water models, a lower polarizability on the oxygen atoms was required with larger metal ions to replicate the experimental binding free energy (as observed in Cu(I) and Ag(I)). In the case of monovalent metal ions, the behavior of the chelating atoms in acetate mirrors that of the chelating atoms in the imidazole molecule used in previous studies [83]. This observation indicates that in systems involving monovalent metal ions, the current standard 12-6 LJ and 12-6-4 LJ nonbonded models generally underestimate induced dipole interactions. As a result, chelator atoms require higher polarizability to effectively reproduce the experimental binding free energy, compared to systems with divalent metal ions. This holds true for divalent metal ions like Cu(II) and Cd(II) in both the TIP3P and SPC/E water models, which are involved in bidentate coordination with acetate.
In the context of the smaller ions such as Ni(II) and Mg(II) in the OPC water model, we found it necessary to employ negative polarizability values to accurately reproduce the experimental binding free energy (see Table 3.3). This phenomenon could be attributed to two potential reasons. First, with regards to the nature of the metal ions, the smaller ones have a higher charge density, leading to a more tightly bound electron cloud around them. Accordingly, they tend to form stronger interactions with the chelator atoms. Secondly, in terms of the water model OPC water models are optimized in terms of dipole and quadrupole moments to provide a more precise representation of liquid water's dipole moments and polarization behavior. Thus, we speculate that the utilization of negative polarizability values became necessary in order to correct the overestimation observed in the 12-6 LJ and standard 12-6-4 LJ models for these smaller ions.

3.3.4 Comparison of 12-6 LJ and 12-6-4 LJ models with the Modified 12-6-4 LJ Model

Figure 3.3 compares the experimental and calculated binding energies of eleven metal ions. Various parameter sets were employed to assess the binding energies of eleven different metal ions. Specifically, binding energies were computed utilizing three distinct models: the default 12-6 LJ model, the default 12-6-4 LJ model, and the modified 12-6-4 LJ model, extracted from AMBER PMF studies. Furthermore, to enhance the robustness of the findings, the binding energies associated to the modified 12-6-4 LJ model were additionally validated using PLUMED PMF simulations. The standard 12-6 LJ model underestimates the binding interactions with Ag(I). However, it seems that the interactions of Cu(I) with the acetate molecule are more favorable. This is because the Cu(I) ion binds to acetate in a monodentate binding mode with a shoulder at the bidentate position, allowing the ligand to have a greater space for establishing multiple interactions with the metal ion. For the other metal ions studied, this model tends to overestimate the binding interactions of acetate, regardless of the water model employed (indicated by the gray bars in Figure 3.3). Another significant implication arises when using the standard 12-6 LJ nonbonded model. That is, as the radius of the metal ion increases, the degree of overestimation generally decreases. Incorporating the default C_4 terms into the 12-6 LJ model (the 12-6-4 LJ nonbonded model) causes a significant adjustment in the binding free energy for most metal ions with

acetate. However, the resulting free energy values deviate from the experimental binding free energies (shown by the black bars in Figure 3.3). Compared to the OPC water model, the default 12-6-4 LJ parameters underestimate the binding free energy of the majority of metal ions in water models with three interaction sites. However, most of the metal ions exhibit stronger interactions with acetate when the default 12-6-4 LJ parameters are used in conjunction with the OPC water model. By implementing the modified C_4 terms in this study within the newly developed 12-6-4 LJ nonbonded model, we successfully reproduced the experimental binding free energies of metal ions with acetate, achieving an acceptable margin of error of approximately 0.35 kcal/mol (depicted as the green bars in Figure 3.3). Our findings from the AMBER PMF simulations align closely with those obtained using PLUMED as an external validation to compute the PMF, with a margin of error of around 0.5 kcal/mol.

To assess our parameters beyond a system composed exclusively of acetate and a metal ion, we applied them to a system containing the Escherichia coli Glyoxalase I metalloprotein (PDB ID: 1F9Z) [218]. A Previous study on this protein indicates that the z12-6 parameters fail to maintain the coordination modes of both Ni(II) binding sites after 80 ns of MD simulations [254]. This metalloprotein is a homodimer comprising two metal-binding sites, with each binding site including His5, His74, Glu56, Glu122, and two water molecules coordinating with one nickel (see Figure 3.4). The results from this study [254] show that two histidine residues contributing to the metal coordination sites lose their interactions with the metal ion. Two oxygen atoms from the negatively charged residues in the binding site then take over and coordinate with the metal ion in a bidentate mode.

Our results demonstrate that after a 200 ns MD simulation using our optimized 12-6-4 parameter set for both negatively charged residues and histidine [83], the two histidine residues in both metal binding sites maintain their interactions with the metal ions (Figure 3.4). Meanwhile, the negatively charged residues (Glu56 and Glu122) coordinate with the metal ion in a monodentate mode, consistent with the crystal structure of the metalloprotein. In the crystal structure, two water molecules coordinate with the metal ions, as Figure 3.4 (panel (b)) shows the two water molecules remain in both Ni(II) binding sites of the metalloprotein during the simulation. This outcome confirms the transferability of our optimized 12-6-4 LJ parameters to other systems containing proteins.

3.4 Conclusions

In this work, we optimized the 12-6-4 LJ nonbonded model to accurately reproduce the experimental binding free energy between acetate and different metal ions. These parameters can be further extended to metalloproteins and metal transporters, facilitating the successful prediction of metal ion interactions with acidic residues (aspartate and glutamate). As previously demonstrated [54], the 12-6-4 LJ parameters enhance the metal ion interactions in aqueous solutions via the inclusion of dipole-induced interactions between the metal ions and water. However, our findings in this study reveal that while the standard 12-6-4 LJ parameters improve metal ion interactions with acetate, they still deviate from the experimental binding free energy values. This work effectively addresses this issue by precisely reproducing the experimental binding free energy of various metal ions with acetate by parametrizing the 12-6-4 model specifically for acetate. Our investigation demonstrates that optimizing the experimental binding free energy across different solvent models depends on several factors, including the metal ion's nature (e.g., the count of unpaired electrons in its outermost subshell), the solvent's characteristics, and the ligand. Successfully accounting for these factors can be achieved by adjusting the C_4 terms. The

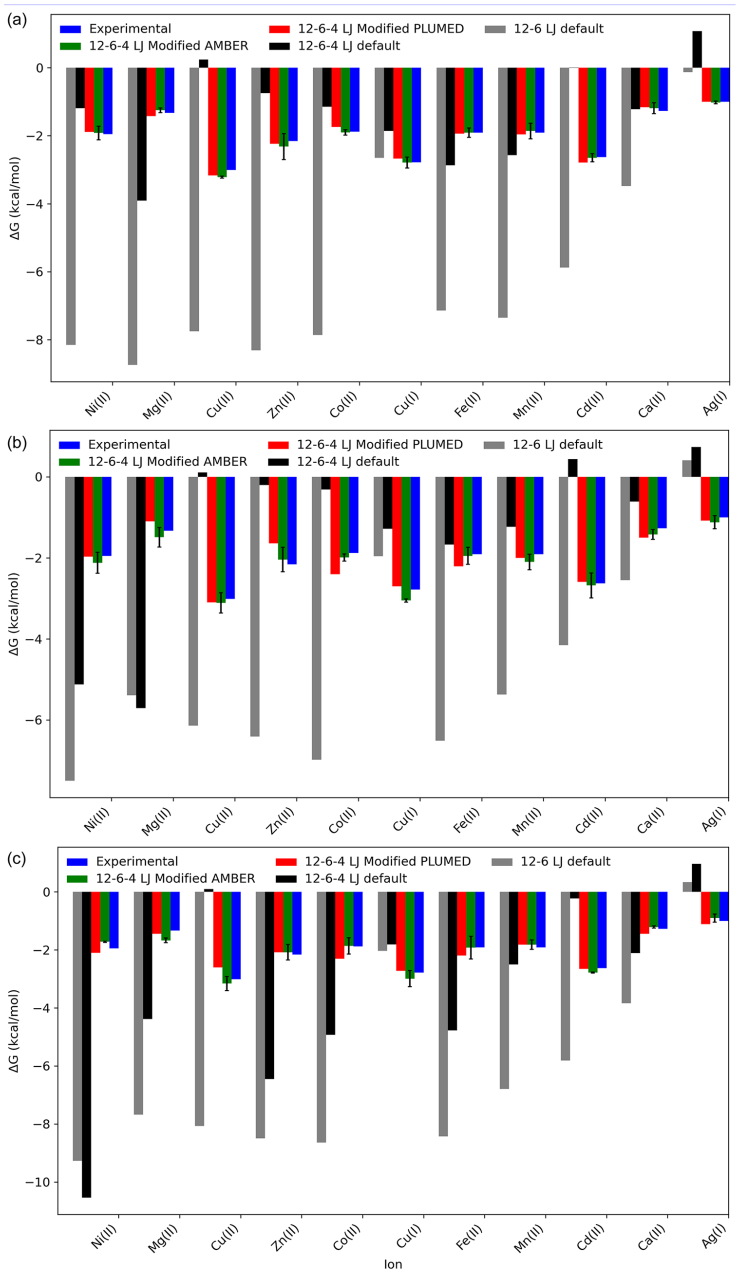

Figure 3.3: Obtained binding free energies using the default 12-6 LJ nonbonded model (gray bars), the default 12-6-4 LJ nonbonded model (black bars), an optimized 12-6-4 LJ model as obtained using AMBER (green bars, capped lines indicate standard deviations), and a modified 12-6-4 LJ model as obtained using PLUMED (red bars). The blue bars show the experimental binding free energy. (a), (b), and (c) represent the results obtained using the TIP3P, SPC/E, and OPC water models, respectively.

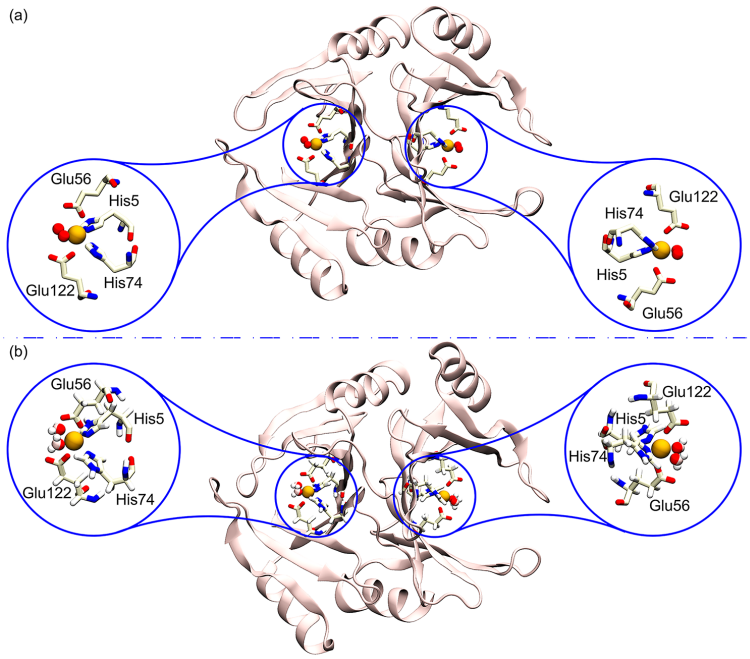

Figure 3.4: Panel (a) shows the crystal structure of the Glx 1 protein, with the residues (pale yellow) coordinating with nickel (gold-colored balls). Panel (b) displays the last snapshot of the MD simulations, illustrating the residues coordinating with the metal ion. As depicted in panel (b), the metal ion coordination is maintained using our optimized 12-6-4 LJ parameter set after 200 ns of MD simulations. The protein structure, water molecules (panel (a))/OPC water molecules (panel (b)) in the metal binding site, and the binding site residues are depicted as Cartoon, VDW, and Licorice models.

obtained parameters are broadly applicable in a large number of systems involving metal ions.

Chapter 4

Simulating the chemical equilibria of metal complexes: Insights into the ligand exchange mechanism

In this Chapter, a computational methodology well-suited for evaluating stability constants, dissociation and association rates, and for uncovering exchange mechanisms of metal ion-ligand complexes is presented. A general methodology is proposed for assessing the thermodynamic and kinetic parameters of ion-ligand complexes, such as stability constants and ligand exchange rates, as well as for hinting at the underlying mechanisms of complex formation through the investigation of solvent (water) molecules in the first hydration shell. Results align well with the binding energies used for the fitting of the underlying force field and show a good agreement with available experimental exchange times. Demonstrated to be computationally efficient, this methodology can be applied to exploring various metal-ligand complexes with a feasible effort. Its robustness has been shown even in more stable complexes or those with longer exchange times, such as those involving Ni^{2+}, suggesting promising potential for widespread applications.

4.1 Introduction

Metal ion-ligand complexes play pivotal roles across diverse domains of chemistry, encompassing catalysis [255], biochemistry [256], materials science [257], and applications in analytical and environmental chemistry [258]. These complexes, formed through ligand coordination with metal ions, have contributed significantly to the development of practical solutions. Notably, the design of polymer-bounded chelating ligands has expanded the possibilities of forming complexes with various transition metals, broadening their applications [259, 260, 261, 262, 263]. Transition metal complexes have found utility in molecular biology, serving as tools for probing nucleic acid structure [264] and facilitating specific or nonspecific cleavage of nucleic acids [265, 266]. In nanotechnology, these complexes have been extensively studied due to their potential as structural and electron-transfer probes [267, 268], as well as candidates for anticancer drugs [269, 270, 271, 272]. A systematic methodology is imperative for the comprehensive study of metal-ligand complexes and their properties, given their extensive applications.
However, a review of the literature over the past five decades reveals a significant disparity between experimental and computational studies in this field. Experimental studies have delved into the stability, preferred geometries, and behavior of metal-ligand complexes

under different conditions. In contrast, computational studies face challenges in accuracy, particularly for rare earth metals and transition metal ions, necessitating high-level theory and computationally intensive approaches [44, 45, 46, 47]. Additionally, addressing the variety of binding modes and timescales between different ligands requires sampling configurational space efficiently [273, 234]. While higher-level theory approaches and *ab initio* calculations [51, 52] have commonly been employed to address the accuracy issue, a systematic and computationally affordable quantitative procedure for computing thermodynamic and kinetic quantities of metal ion-ligand complexes is yet to be fully established.
This study introduces a methodology that combines enhanced sampling methods and Markov State Models (MSM) to systematically evaluate the stability and rate constants of metal ion-ligand complexes. The method's efficacy is demonstrated through tests on the solution of cadmium cation (Cd^{2+}) and ethylenediamine (en) at various concentrations. To showcase the approach's flexibility, a different cation with greater stability and longer exchange times (Ni^{2+}) is also investigated, along with amines equipped with different coordinating groups, namely methylamine (nme), diethylenetriamine (dien), and putrescine (put). The investigation includes a dissection of the enthalpic and entropic contributions to the stability of complexes, thus shedding some light on the chelate effect. A further important outcome of this study, achieved through the application of a coordination number for the metal ion center and the explicit solvent, is the ability to probe the molecular phenomena underlying the formation and dynamics of metal complexes, within the intrinsic approximations of the chosen level of theory, in this case, classical force fields.

4.2 Methods

4.2.1 Simulation methods

Taking as a reference a recent work [274], Cd^{2+} cations, en ligands and counterions were solvated in a truncated octahedron box with total volume of 310 nm^3 and 9600 water molecules. Different concentrations were tested for the cadmium-en complex, namely 0.225 M, 0.15 M, 0.1 M, and 0.05 M of en, with 45, 30, 20 and 10 ligands, respectively. To further assess the model, two additional 0.08 M solutions were prepared keeping the same stoichiometry between metal ion and ligands (1:3). The first one with 1 Cd^{2+} and 3 en and the second one with 2 Cd^{2+} and 6 en, in both cases adjusting the volume to enforce the 0.08 M concentration solution. For nme a 0.08 M solution with 1 Cd^{2+} and 6 nme and a 0.15 M solution with 10 Cd^{2+} and 60 nme were prepared. For dien and putrescine ligands the only concentration tested was 0.15 M with 10 Cd^{2+} and 20 dien and 30 put respectively. In all these simulations the number of counterions employed ensured neutrality being twice the number of the cadmium cations in each system. A 0.08 M solution with 1 Ni^{2+} and 3 en ligands was prepared to probe the methodology against a system that experimentally has shown slower exchange times. The force field employed for metal cations and water interactions was the 12-6-4 Lennard-Jones (LJ) to correctly take into account the ion-induced dipole interaction that can be significant in highly charged systems [54]. Regarding en, a recently published optimized m12-6-4 LJ [274] was used to correctly reproduce the binding affinity of a single en ligand with different metal cations simulated. This parametrization was kept also for the metal-nitrogen interaction for the dien ligand. Regarding nme, its metal-nitrogen interaction has been slightly boosted from the paper of [274] to a polarizability value of 3.35 $Å^3$ (from a value of 3.16 $Å^3$) to try to reproduce the experimental fact that a three-fold and four-fold Cd-nme complex should form around cadmium cation at high concentrations [275]. The nme parametrization was kept also for the metal-nitrogen

interaction for the putrescine ligand. It is important to note that the force field used for the cadmium cation overestimates its coordination at 8 instead of the experimentally reported 6 [276]. However, this parameterization has been employed in other studies [277][83] as it accurately reproduces experimental data such as solvation free energy and ion-water distance [54].
The water model used was TIP3P [278]. Chloride ions were treated with a customized 12-6 LJ nonbonded model to prevent counterions from interacting strongly with the metal cations, following what had been experimentally seen [279] where a stable ionic medium is utilized to keep the activity coefficients of a particular ion constant. This is achieved experimentally by introducing a high concentration of a specific anion, such as ClO_4^- or NO_3^-, which is intended to be unreactive and not form any complexes with the ions under study. By doing so, the activity coefficients of the ions being studied can be considered to remain constant in all the solutions [280].
All the simulations were performed with Amber22 [281] employing periodic boundary conditions (PBC) and PME to treat long-range interactions with a 12 Å cutoff. Minimization was performed using 20000 steps of steepest descent followed by 10000 steps of conjugate gradient. A 250 ps NPT heating procedure was performed to heat the system from 0 K to 300 K followed by a 1 ns equilibration at 300K with constant NPT conditions using a Langevin thermostat at 1 atm. The equilibrated geometries were read for the NPT production simulation runs each of 200 ns. An integration time step of 1 fs was used in the heating step and 2 fs in the production runs. Langevin dynamics temperature control was employed in the heating and the production runs with a collision rate equal to 1.0 ps. SHAKE algorithm was used for water molecules for all simulations.

4.2.2 Thermodynamic derivation

The equilibrium constant (or association constant) is directly related to the equilibrium concentration of bound ([ML]) and unbound ligand ([L]) and metal ([M]) through $K = [ML]/[M][L]$. If more than one ligand can bind to the metal center this can be generalized for the i-th equilibrium constant as

$$K_i = \frac{[ML_i]}{[ML_{i-1}][L]} \tag{4.1}$$

where the logarithm of this quantity (pK_i) is the one usually experimentally measured as equilibrium constants in metal complexes [279].
To define these metal complexes states it has been used a simple and intuitive collective variable, namely the coordination number [169, 160], as it has been proven to be effective for studying the local environment of ions. In the present work, both a water coordination number and a ligand coordination number have been implemented, the former monitors the number of water-oxygens within 3.2 Å from the metal ion (2.9 Å for Ni^{+2}) while the latter records the number of nitrogen atoms of the different amine ligands within 3.4 Å from the metal ion.
Enhanced sampling was necessary to sample the equilibrium populations of different ligand-cation binding states $[ML_i]$ in equation 4.1. A well-tempered metadynamics [120, 124] variant was adopted in this work. In fact, in order to be able to study realistic concentrations of metal ions and ligands, all the possible states $[ML_i]$ of the metal ions, intended as a subset of metal ions coordinated with i ligands, must be sampled. To this point, parallel-bias metadynamics (PBMetaD) [130] altogether with partition family setup [132] was used to bias alternatively one of the N metal ions ligands or water oxygen coordination state.

Together with 16 multiple-walkers (MW) [282] a converged equilibrium population of coordination states was reached in a reasonable amount of time depending on the system (systems containing Cd^{2+} were simulated with this setup for 40 ns each MW and the system with Ni^{2+} for 80 ns each MW). The deposited Gaussians had initial height, width, and deposition stride equal to 1 kJ/mol, 0.1, and 1 ps, respectively; the bias factor was set to 10. Metadynamics simulations were carried out using the open-source, community-developed PLUMED library (ver. 2.8)[213] [135]. The minimum free energy path of each 2D free energy profile has been extracted using the open-source software MEPSA (ver. 1.6)[283].
From the converged free energy profile an estimate of the equilibrium population of each state can be done assuming Boltzmann population

$$\frac{n_i}{n_j} = exp^{-\Delta F_{ij}/k_{\mathrm{B}}T} \tag{4.2}$$

where n_i is the population at equilibrium associated to the $[ML_i]$ coordination state, $\Delta \mathrm{F}_{ij}$ is the difference in free energy between the i-th and j-th coordination states, k_{B} is the Boltzmann constant and T the temperature of the system. If equation 4.2 is substituted into equation 4.1 the equilibrium constant can be rewritten as

$$K_i = \frac{exp\left(-\Delta F_{ij}/k_{\mathrm{B}}T\right)}{[L]} \tag{4.3}$$

where the unknown is the unbound ligand concentration $[L]$, however it can be computed by subtraction since the initial concentration of ligand $[L_0]$ is known as

$$[L] = [L_0] - \sum_{i=1}^{N_s} i[ML_i] \tag{4.4}$$

where N_s are the total number of possible coordination states.
Binding entropy for the i-th state of the metal ion-ligand complex $\mathrm{T}\Delta \mathrm{S}_i^{bind}$ was computed by subtraction from the Gibbs free energy of binding $\Delta \mathrm{G}_i^{bind}$=-RT $\log \mathrm{K}_i$ and the enthalpy of binding $\Delta \mathrm{H}_i^{bind}$. To compute $\Delta\ \mathrm{H}_i^{bind}$, 100 ns of unbiased MD was run for each metal ion-ligand coordination state both in the bound and unbound state. The running average of the total energy was taken to extract $\Delta \mathrm{H}_i^{bind}=\mathrm{H}_i^{bound}$-$\mathrm{H}_i^{unbound}$.

4.2.3 Kinetic modeling through the Markov State Model

The characterization of the configurational space of the metal complexes described by the coordination number allowed an easy construction of a Markov State Model (MSM) to compute the rate constants between different coordination states. Initial structures were extracted every 0.5 step from the metal-water and metal-ligands coordination maps. From each of these structures 200 unbiased MD replicas were simulated randomly resampling the momenta. It is important to note that a similar approach using metadynamics to help build a reliable MSM has been used previously to explore the dynamics of the helical peptide Aib_9[284]. Each replica was 2 ns long and the coordination state for ion-ligand and ion-water was recorded every 100 fs. These data were processed using an in-house Python v.3.8 script with the help of the deeptime [285] python library. K-Means++ [146] was used to find the lowest number of MSM microstates that satisfied the implied timescale (see Eq.1.115 and Fig. C.5) and Chapman-Kolmogorov analysis (see Eq. 1.120 and Fig. C.6) [143]. Successively PCCA+ [150] was applied to reduce the MSM microstates to the actual experimental measurable metal complex states. The transition matrix W_{ij} associated with the MSM

was also reduced accordingly and mean first passage times (MFPT) were computed using transition path theory (TPT) [286] and the errors associated were computed using a full Bayesian approach as described in [139, 143].
To compare with experimental rate constants computed MFPT needed to be transformed accordingly taking into account the free ligand concentration. From reaction law theory for the i-th complex state, it can be written a kinetic equation of the form

$$\frac{d[ML_i]}{dt} = -k_{-i}[ML_i] + k_i[ML_i][L] - k_{i+1}[ML_i][L] + k_{i+1}[ML_{i+1}] \tag{4.5}$$

A similar kinetic equation can be written for the populations of the reduced MSM of the form

$$\frac{dn_i}{dt} = -k_{-i}n_i + k_i n_i - k_{i+1}n_i + k_{i+1}n_{i+1} \tag{4.6}$$

To have compatibility between the experimental and the computed kinetic models a one-to-one correspondence must be superimposed. Thus, the forward rate constants of the MSM must be multiplied by a factor γ^2/n_{lig}^{eq} where n_{lig}^{eq} is the number of unbounded ligand at equilibrium (eq. 4.4) and $\gamma = N_{Av}V = 0.6023 * V(nm^3)$, with N_{Av} being the Avogadro's number and V the volume of the simulation box. Regarding the backward rate constants, the inverse of the MFPT calculated from the MSM must be multiplied just by a factor γ.

4.2.4 Mechanism of complex formation

To analyze the mechanism of ion-ligand binding, several unbiased trajectories were generated for the 10 Cd^{2+} and 30 en (10 ns each) and 10 Cd^{2+} and 30 nme (5 ns each), saving atomic positions every 0.1 ps. The distances between nitrogen atoms of the amines and cadmium ions were computed. A ligand was considered bound when its nitrogen atom(s) entered the ion first coordination shell (distance < 3.4 Å) and remained for at least 20 ps, to ignore spurious binding/unbinding events.
The binding events were classified by the change in cadmium-ligand coordination number (0-1, 1-2, 2-3, 3-4 bindings). The ion-ligand and ion-water coordination numbers were computed for 20 ps before and after the binding events. The water/ligand coordination number average and evolution during the binding events were used to classify the mechanism of different binding events (dissociative and associative).
Water-ligand exchange mechanism was also analyzed by monitoring the distances and the angle between entering nitrogen, exiting water, and the ion during each binding event, as done in a recent work on water exchange kinetics in magnesium ions [287].

4.3 Results and Discussion

4.3.1 Cadmium ethylenediamine

A typical system for studying the stability of metal complexes is a solution of Cd^{2+} and ethylenediamine ($C_2H_4(NH_2)_2$). This system was used to validate the proposed computational procedure. To test the generalization of such a methodology, different concentrations (changing volume, number of ions, and number of ligands) were prepared and analysed. From here on, the states of the metal complexes will be expressed by omitting the coordinating waters for simplicity.

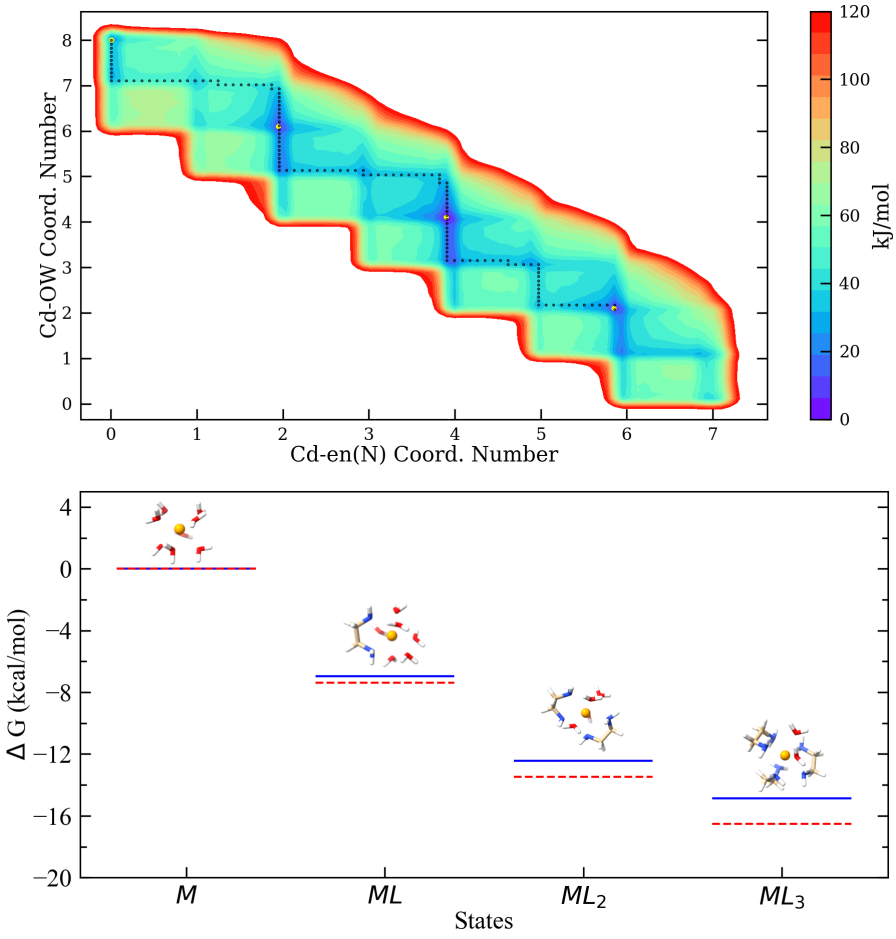

Figure 4.1: On top, free energy map showing different metal ion-ligand coordination states ($Cd(en)_i$) and metal ion-water ($Cd(OW)_i$) coordination states. Yellow points are the minima usually measured by experiments. Dotted black line is the minimum free energy pathway. At the bottom, computed (solid blue) and experimental (dashed red) stability constants for different coordination states.

Thermodynamics of ligand exchange

Various free energy maps, like the one presented in Fig. 4.1, have been computed, and different stability constants have been derived for various system conditions, as summarized in Table C.1. The methodology demonstrates robustness and coherence when there is an excess of ligand in the solution, specifically at a ratio of 3/1 or greater compared to the number of cations. This aligns with a well-established experimental setup, given that ion concentrations are typically a few orders of magnitude lower than ligand concentrations [288] [280]. It is noteworthy that the values of ΔF_{ij} in Table C.2, as expressed in eq. 4.3, computed using the proposed methodology closely aligns with those derived from experimental stability constants, even when dealing with uncommon ligand concentrations where an excess of ligands is not evident. This behavior can be understood by examining equations 4.3 and 4.4. When the ligand concentration drops below a certain threshold, the concentration of free ligands at equilibrium approaches zero. Consequently, even a minor error of a few kJ/mol in estimating ΔF_{ij} can lead to a significant error in estimating the stability constant pK_i. This observation underscores the recommendation of using an excess of ligand when applying the proposed methodology, mirroring the approach used in experimental evaluations of stability constants. Conversely, when comparing data at different concentrations, especially in systems with a ligand deficiency, it is advisable to compare the different maps by analyzing their energy minima for the coordination states.
Another notable aspect of this procedure is that, by introducing metal-water coordination as a second variable, it becomes possible to suggest different molecular exchange mechanisms for the formation of metal complexes. Analytical chemists find it vital to determine whether this mechanism is associative (the ligand enters the coordination shell before a water molecule exits) or dissociative (a water molecule exits the coordination shell before a ligand enters). Analyzing the minimum free energy path in these maps qualitatively reveals the preferred mechanism for the cation. In Figure 4.1, for Cd^{2+}, the mechanism clearly demonstrates a dissociative nature consistent with experimental findings [289].
Lastly, the most compelling aspect of this approach is its ability to assess the underlying force fields employed for specific systems. In this context, the specially modified 12-6-4 potential for Cd^{2+} and ethylenediamine [274] performed exceptionally well. It not only accurately predicted all three stability constants, as evident in Figure 4.1, despite being tuned solely against experimental pK_1, but also effectively predicted the dissociative-type mechanism of Cd^{2+} complexes.

Kinetics of ligand exchange

To obtain a more profound comprehension of the dynamics governing ligand exchange, the computation of rate constants for formation and dissociation is of utmost importance. Experimentally measured rate constants exhibit significant variation among complexes, contingent on the central metal ion [290]. Moreover, rate constants for formation and dissociation with amine complexes may diverge by up to six orders of magnitude [291] [292], with the latter often exceeding the temporal scale accessible through molecular dynamics. To address this, a Markov State Model (MSM) was employed, established from the thermodynamic map as described in Section 4.2.3. Coarsely graining the MSM microstates to represent experimental metal-ligand coordination states (see Fig.4.2 and Section 4.2.3), mean first passage times (MFPTs) were extracted, enabling the determination of rate constants for all formation and dissociation steps, as illustrated in Figure 4.2.
For validation, the model considered the same metal-to-ligand ratio (1:3) at varying ligand concentrations (0.08 M and 0.15 M) and ion concentrations (0.026 M and 0.05 M), detailed

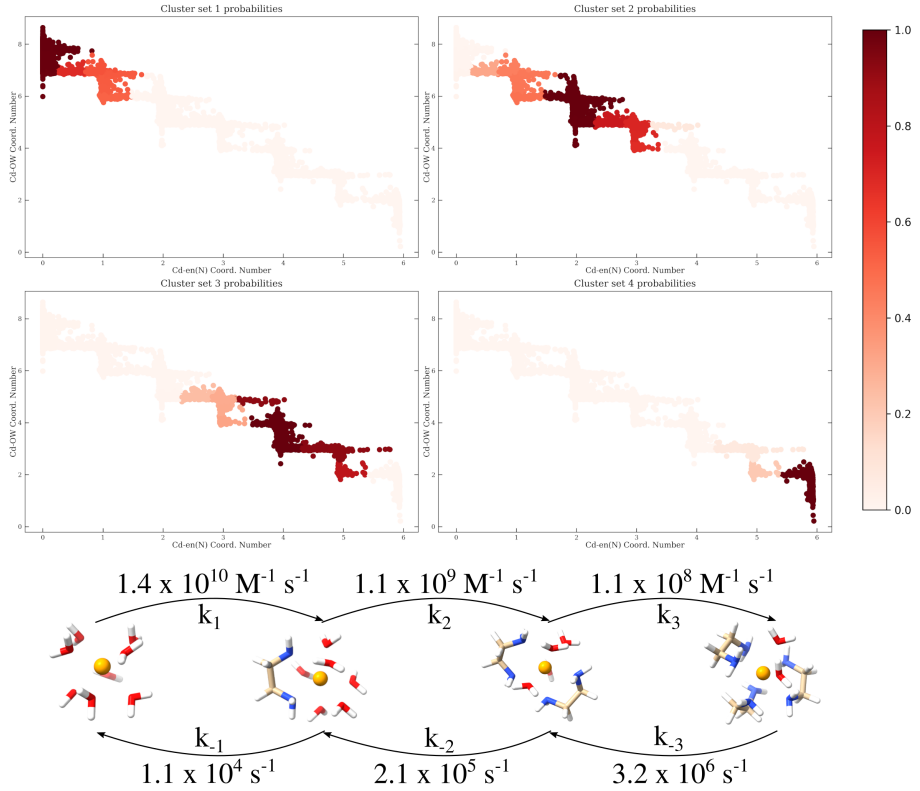

Figure 4.2: On top, PCCA+ on the MSM for the Cd^{2+}-en system where the 4 clusters found correspond to the experimentally measured coordination states. At the bottom, kinetic formation and dissociation rate constants extracted from the reduced MSM.

in Tables C.3 and Table C.4. While the MFPTs varied due to concentration differences, the model consistently produced rate constants of the same order of magnitude, highlighting its robustness when extended to more realistic concentrations and multiple metal centers. Following coarse-graining of the initial MSM, the resulting submodel closely represented the experimental metal coordination states. This facilitated the estimation of stability constants pK_i, computed by the ratio of the forward k_i and backward k_{-i} rate constants, as demonstrated in Table C.5. The stability constants calculated through this method aligned with results from the free energy map, providing an additional test of the MSM's correctness, particularly concerning pK_2 and pK_3. pK_1 appeared slightly overestimated by the MSM compared to values from the thermodynamic map. This discrepancy could be attributed to the fact that the formation k_1 and dissociation constant k_{-1} exhibited the greatest relative error. The former, being the fastest process of the model and closer to the lag time of the MSM model, might be prone to a higher relative error. The latter, representing the slowest unbinding time and being the most challenging to predict, could consequently have the greatest error.

It is essential to note that experimental data on exchange times of metal complexes involving cadmium are absent in the literature. Nonetheless, works suggest that, in the absence of such data, the first formation rate k_1 is proportional to the water exchange kinetics of the same ion k^{M-H_2O} [293, 294]. Reviewing the data in Table C.4, k_1 is on the order of 10^{10} M^{-1} s^{-1}, approximately two orders of magnitude higher than experimental data available on water exchange times in ionic solutions containing the cadmium cation [295]. In a previous study employing the same force field for cadmium and water (12-6-4) [160], the exchange time was estimated to be a few tens of picoseconds, translating to an exchange rate of approximately $5 * 10^{10}$ M^{-1} s^{-1}, aligning with the estimate of the MSM for the formation rate of the $Cd(en)_1$ complex. In subsequent studies [296], attempts were made to model the formation rate k_1 of metal complexes as the product of the metal-water exchange rate k^{M-H_2O} (also referred to as metal-water bond rupture in the literature) and a geometric factor that accounts for the average distance between the metal center and the ligand in the second solvation shell [297]. Taking this model and inserting 4 Angstroms as the distance, this factor is estimated to be approximately 0.15. The water exchange rate remains the rate-determining step; however, in this case, it would be about 10^7 M^{-1} s^{-1} using experimental rates and $7.5 * 10^9$ M^{-1} s^{-1} if the rate estimated for cadmium with the same force field is inserted, confirming the validity of the estimate.

Moreover, examining Table C.4, a decreasing trend is noticeable for subsequent formation or dissociation rates. This is a curious observation not present in the literature, at least to our knowledge. Qualitatively, this trend could be explained as a progressively decreasing ease for ligands to bind to the metal center when another ligand is already present, and vice versa for the dissociation rate. However, this interpretation should be approached with caution, as the literature lacks data and models for formation constants other than the first. While this methodology may not be used to blindly extract exchange times, it can definitely be employed to explore intricate kinetic mechanisms underlying the force field in play.

Water-Ligand exchange mechanism

As mentioned in Section 4.3.1, the investigation of the free energy maps reveals potential binding mechanisms between cadmium ions and ethylenediamine in the formation of Cd(en) complexes. The formation of metal complexes in aqueous solutions is usually seen as a substitution reaction between the first coordination shell water molecules and the

entering ligands [17]. According to Langford and Stengle [298], the ligand exchange mechanism can be classified into four categories: associative (A), dissociative (D), interchange associative (I_a), and interchange dissociative (I_d). The former two exhibit a detectable intermediate with decreased (D) or increased (A) coordination numbers, while the latter two lack detectable intermediates. Distinguishing between these mechanisms is not always experimentally accessible [17], prompting the use of computational methods to elucidate ligand exchange in metal complexes [299, 300, 301].
Experimental studies on the formation of mono and polyamine complexes with transition divalent metal ions, inspired by Eigen's proposed mechanism [289], indicate a dissociative mechanism. Kinetic measurements highlight water loss as the rate-determining step in the metal complex formation reaction [293, 296, 302, 303]. Strong correlations exist between ion-water exchange rates and complex formation rate constants [293, 304]. While experiments on first-row divalent transition metal ions provide water exchange rates from 10^4 to 10^7 s^{-1} [17, 305, 304], there is a lack of literature data on the formation mechanism of cadmium-amine complexes, where cadmium(II) exhibits a faster water exchange rate ($\approx 10^8$ s^{-1}) [306].
In Figure 4.1, the minimum free energy path suggests a dissociative mechanism for the Cd-en system. During the formation of the Cd(en) complex, a water molecule exits the cadmium first coordination shell before the entry of the first en nitrogen. Subsequently, the second nitrogen enters the first coordination shell along an associative path. However, the minimal energy difference between associative and dissociative paths impedes a clear definition of the binding mechanism.
To validate these findings, we analyzed the binding mechanism between Cd-en from a series of unbiased simulations (see sec. 4.2.4). When the first ethylenediamine (en) nitrogen binds to a cadmium ion, the evolution of water and en coordination number reveals a dissociative mechanism, with a water molecule leaving the first coordination shell of the cadmium ion before the nitrogen atom enters (Figure 4.3a). In the second binding event, both associative and dissociative paths are possible (Figure 4.3c). This aligns well with the minimum energy paths displayed in Figure 4.1.
Inspired by the analysis proposed by Falkner and Schwierz in the study of water exchange around a solvated magnesium ion [287], probability distributions of leaving water positions in the first coordination shell of cadmium provide additional insights into Cd(en) complex formation (Figure 4.3b, d, e). For the first binding (Figure 4.3b), water molecules tend to leave the first coordination shell in a *trans* position (angle $> 90°$). The distance probability distribution (Figure C.8) and the decrease in water coordination number (Figure 4.3a) suggest that the exchange occurs outside the first coordination shell, as expected for a dissociative mechanism.
Figures 4.3c, d, e display coordination number and leaving water position probability distributions for the ethylenediamine chelating ring closure. Here, an associative pathway is also present, as it can be seen from the water and nitrogen-ion distance probability distribution (Figure C.8). Coordination number and distance probability distributions (Figure 4.3c, Figure C.8) suggest that the water molecule leaves the first coordination shell after the entrance of the second en nitrogen. In both associative and dissociative mechanisms, the water leaves the coordination shell in a *trans* position relative to the entering nitrogen (Figure 4.3d, e). For the associative mechanism, the leaving water is also in a *trans* position relative to the first bound nitrogen (Figure 4.3d). In the dissociative one, instead, the water near the bound nitrogen has a similar probability of leaving the coordination shell. The difference in the leaving water position between the two mechanisms could be explained by a different coordination shell rearrangement. In the dissociative mechanism, the coordina-

tion shell can rearrange before the entrance of the second nitrogen, thus no main difference is present between the water near or in a *trans* position relative to the bound en. In the associative mechanism, instead, the temporary over-coordination forces a coordination shell rearrangement that pushes away the water molecule in a *trans* position relative to both en nitrogen atoms. Similar consideration can be done when second en molecule binds a cadmium ion (Figure C.7).
In summary, the binding of an en molecule to a cadmium ion unfolds in two distinct events. The first en nitrogen binds through a dissociative path, following the loss of a water molecule from the cadmium first coordination shell. The second binding can occur through both associative and dissociative paths. In both bindings, exiting water molecules leave the first coordination shell from the opposite side with respect to the entering nitrogen. For the chelating ring formation, the associative mechanism leads to a temporary cadmium over-coordination. The next coordination shell rearrangement forces a water molecule in *trans* position to leave.
It is crucial to note that the ligand exchange mechanism strongly depends on the force field used in simulations [287]. A more accurate estimation of ligand exchange and binding mechanisms could be achieved with higher-level calculations or transition path sampling techniques. Moreover, mechanisms (A, D, I_a, I_d) are usually classified by the activation volume [307], but such measurements are beyond the scope of this work. Nevertheless, the agreement between the results from unbiased simulations (Figure 4.3a,c) and the minimum free energy path in Figure 4.1 underscores the procedure's ability to predict and estimate the binding mechanism between metal ions and ligands.

4.3.2 Ni^{2+}-amine complexes

Regarding the study of metal complexes with simple amines, one of the most extensively researched cations is undoubtedly nickel [308, 296, 309]. In the literature, stability constants as well as formation and dissociation rate constants can be found, as Ni^{2+} is easier to be studied experimentally with techniques extracting kinetic information, such as stopped-flow and temperature jump experiments [310, 311]. However, from a computational modeling perspective, studying complexes with a metal center like nickel is highly challenging, compared to cadmium. This is due to its capability to form more stable complexes and to have water exchange times on the order of microseconds [290]. Both of these aspects can be seen as a huge step further into testing the methodology just presented in situations involving very rare events usually deeply inaccessible even in long unbiased MD simulations.
By employing a specially modified force field for the Ni^{2+}-en interaction [274] and preparing a relatively small system (one Ni^{2+}, 3 en in a 64 nm^3 box), an attempt was made to gather thermodynamic and kinetic information for these types of complexes, similarly to what has been done for cadmium complexes. Examining Figure C.1 and Table C.6, it is evident that the estimated first stability constant pK_1 closely aligns with both the experimental value and the one obtained through the evaluation of Potential of Mean Force (PMF) from umbrella sampling as computed in the study where the force field was calibrated [274]. However, the subsequent stability constants seem to degrade with an increasing number of ligands. In particular, pK_2 appears to be underestimated by approximately 0.7, and pK_3 by about 1.2 compared to experimental values [292]. Lastly, another property that can be extracted from the thermodynamic maps is that the Ni^{2+}-en complex, contrary to cadmium, exhibits a preference for an associative mechanism. This is evident from the free energy pathway where the loss of water from the first solvation shell occurs after the entry of an amino group (see Fig. C.1). However, this direction contradicts what has been ob-

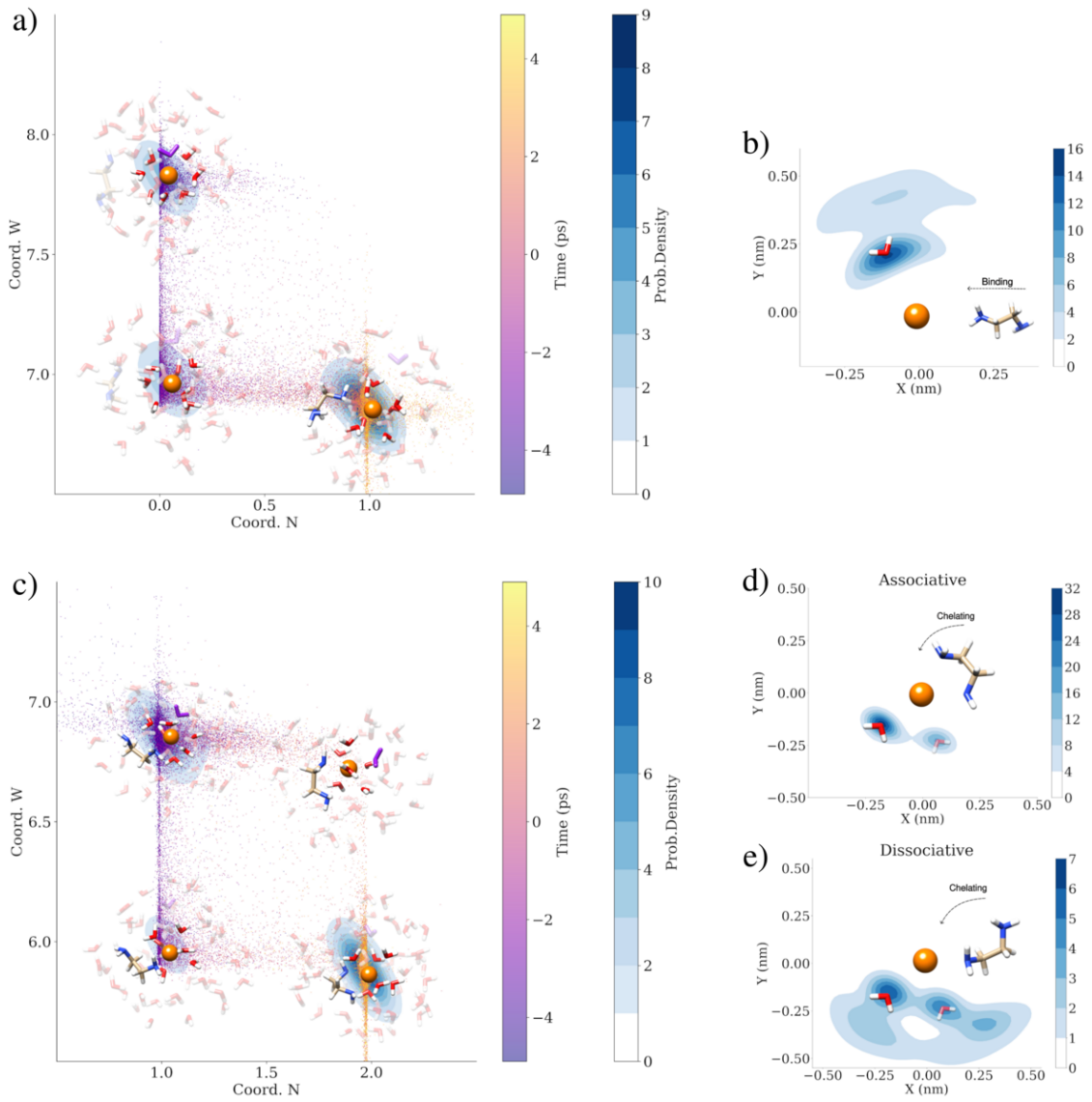

Figure 4.3: a) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the en first nitrogen binding. Scattered plot represents the time evolution and the probability distribution identifies the main steps of the mechanism. b), Probability distribution of the leaving water position for the en first nitrogen binding. c) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the en chelating ring closure. d), e) Probability distribution of the leaving water position for the en chelating ring closure with associative (d) and dissociative (e) mechanism.

served through High-Pressure Multinuclear Magnetic Resonance experiments, which have demonstrated solvent exchange occurring in a dissociative manner [312]. This incongruity potentially provides further opportunities for refining the nichel parametrization in this direction.
If one wishes to explore this system from a kinetic perspective, it can be observed from Table C.7 that the formation rate of the first Ni(en) complex is on the order of 10^6 M^{-1} s^{-1}, which closely aligns with the experimental rate of $3.5*10^5$ M^{-1} s^{-1} [292]. Additionally, the water exchange time for nickel was calculated employing the 12-6-4 force field for Ni^{2+} and TIP3P water model, yielding a τ^{Ni-H_2O} of 5.5 μs (equivalent to an exchange rate of $1.8*10^5$ s^{-1}) that closely follows what was measured experimentally through NMR experiments of $3.2*10^4$ s^{-1}, confirming on one hand that the ligand exchange is somewhat related to the water exchange rate and on the other hand a coherence on what has been seen experimentally regarding the exchange rate being faster when ligands are present in solution. It is noteworthy that, in this case, the kinetics are better reproduced by the underlying force field compared to that of cadmium, despite both accurately reproducing complex stabilities. Furthermore, a similar kinetic trend to that of cadmium can be observed, where the formation rate appears to decrease with an increasing number of coordinated ligands, and vice versa for dissociation rates.
Information on the dissociation rate k_{-1} can be extracted from the same experimental study [292]. Through stopped-flow techniques it has been measured to be $8 * 10^{-3}$ s^{-1}. Once again, this aligns closely with the value calculated through MSM of $2.6 * 10^{-1}$ s^{-1} (see Table C.7), further validating the power of this computational approach in attempting to estimate kinetic rates on the order of seconds, which are typical of experiments but far from simulation timescales.

4.3.3 Insights into the chelate effect

The versatility of the procedure allows for exploring different ligands, and the first choice was to focus on the differences in terms of thermodynamics and kinetics between the bidentate ethylenediamine and its monodentate counterpart methylamine (CH_3NH_2). This made it possible to compare simulated results with experimental values present in the literature and to investigate at a molecular level the so-called chelate effect and the underlying causes differentiating between en and nme complexes.
Looking through Table C.8 the stability constants for cadmium-nme complexes are well reproduced. The major disagreement is with pK_4 and it was expected since the original parametrization has been done on reproducing the experimental binding energy of Cd^{2+} with just one methylamine and an attempt has been made to match better also the other experimental binding energies (see sec. 4.2.1 and Table C.8). In any case, the interaction between cadmium and nme was not raised more, so as not to make coordination states 5 and 6 too likely as they were reported as improbable to happen in [275].
To make possible a direct comparison between methylamine and ethylenediamine complexes, configurations showing an even number of ligands of the former were compared with the corresponding ones made with the latter ligand, thus ensuring a fair comparison between an equal number of amino groups. From Table 4.1 and Figures 4.4 and 4.5, it can be observed that the binding affinity for two methylamines and a single ethylenediamine is rather similar. However, the difference becomes significant (approximately 4.5 kcal/mol) when comparing the next term, i.e. 4 nme and 2 en. Such an increasing difference is well-documented in the literature [313, 234]. Upon subtracting the computed enthalpic terms (note that enthalpy is similar between nme and en for an equal number of amino groups), it

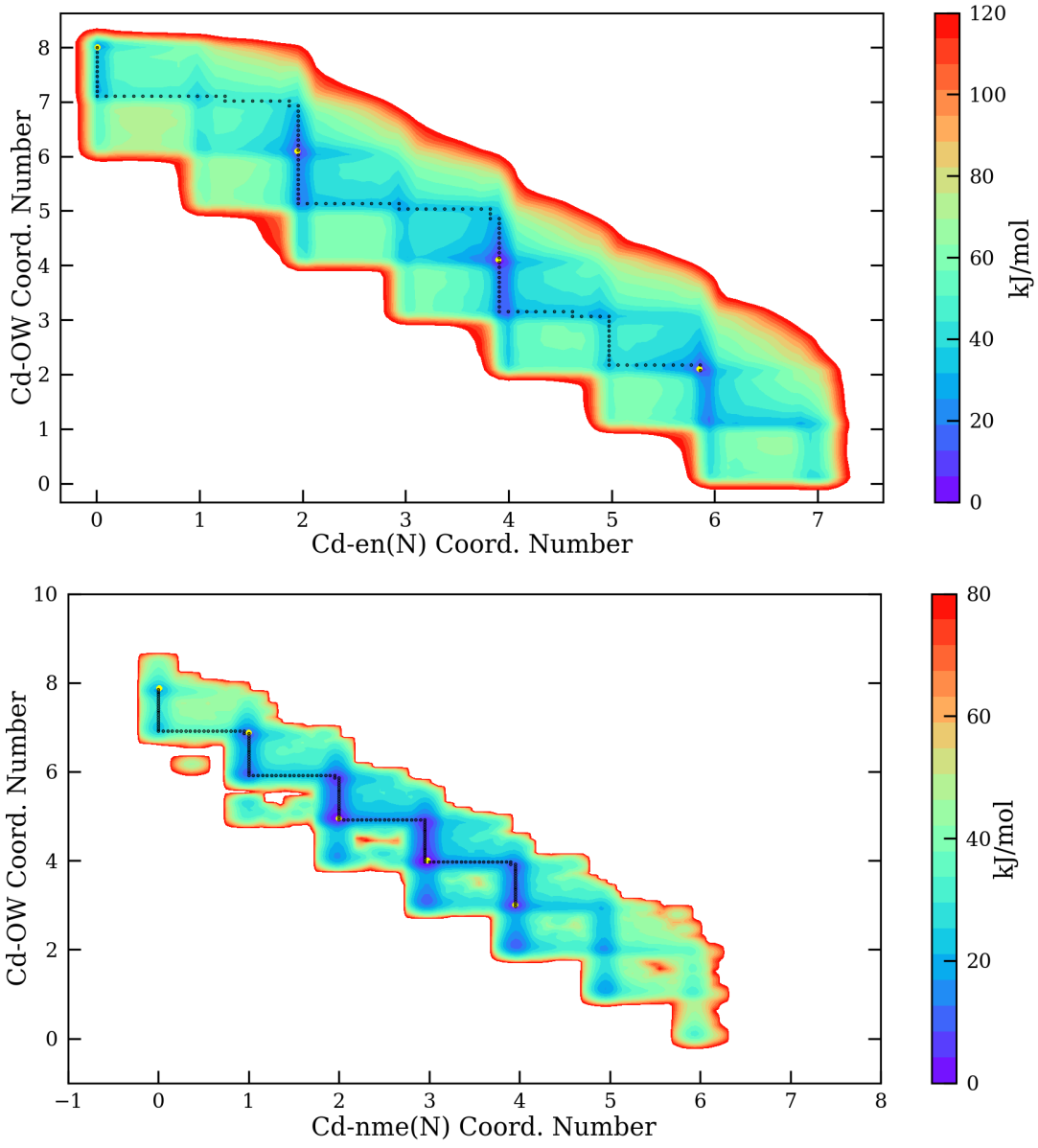

Figure 4.4: Free energy maps showing different metal-ligand and metal-water coordination states for Cd^{2+}-en (on the top) and Cd^{2+}-nme (on the bottom). Yellow points are the minima usually measured by experiments. The dotted black line is the minimum free energy pathway.

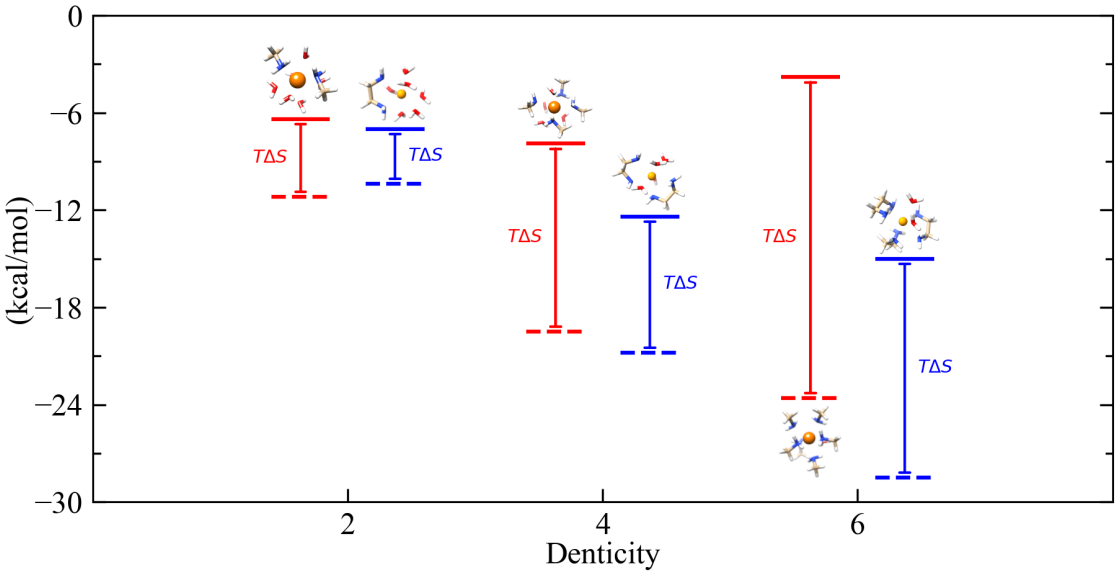

Figure 4.5: Computed ΔG (solid bars) and ΔH (dashed bars) for different coordination states for nme (red) and en (blue). Vertical lines depict -TΔS.

is possible to compare them in terms of entropy. From this perspective, it seems clear that over 75% of the difference in binding energy between 4 nme and 2 en can be attributed to the entropic gain of the bidentate ligand compared to the monodentate ligand.
Numerous experimental studies have focused on dissecting the energetic contributions leading to chelate ring formation, differentiating between enthalpic and entropic contributions. Experimentally, this is achieved by subtracting the enthalpic contribution obtained from fitting stability constants against temperature (Van 't Hoff equation) from the free energy to determine the entropic contribution. Therefore, it is pertinent to compare the obtained results with experimental data that closely resemble the simulation conditions. In this case, Ref. [313] seems to be a suitable study, noting that the ionic strength was 0.15 M, and the temperature was kept at 25 °C. Taking this experimental work as a reference for the Cd-en system, despite the free energies being similar (ΔG_1 is underestimated by 0.5 kcal/mol, and ΔG_2 by 1.0 kcal/mol), the enthalpic and entropic contributions differ significantly. Observing Table 4.1, it becomes apparent that the enthalpy seems to be much higher in absolute terms compared to experimental values (almost double when comparing the same number of amino groups). Similarly, entropy also appears to be overestimated because in no simulated case does it contribute to the lowering of free energy. From this, it can be inferred that the parametrization, aimed at reproducing stability constants and, consequently, free energies accurately, had to compensate for the excessive overestimation of the system's entropy with a greater non-bonded term interaction between cadmium and nitrogen of ethylenediamine.
However, while enthalpic and entropic contributions seem to be overestimated in absolute terms, it is noteworthy that the trend of enthalpy remaining more or less constant (at most decreasing slightly) for successive coordinated ligands is also maintained in the simulated system. This observation holds true for the entropic contribution as well, which, both experimentally and computationally (see Table 4.1 and Figure 4.5), tends to contribute progressively more to the complex destabilization. These characteristics are also evident in the Cd-nme system, where it is additionally observed that the entropic contribution to

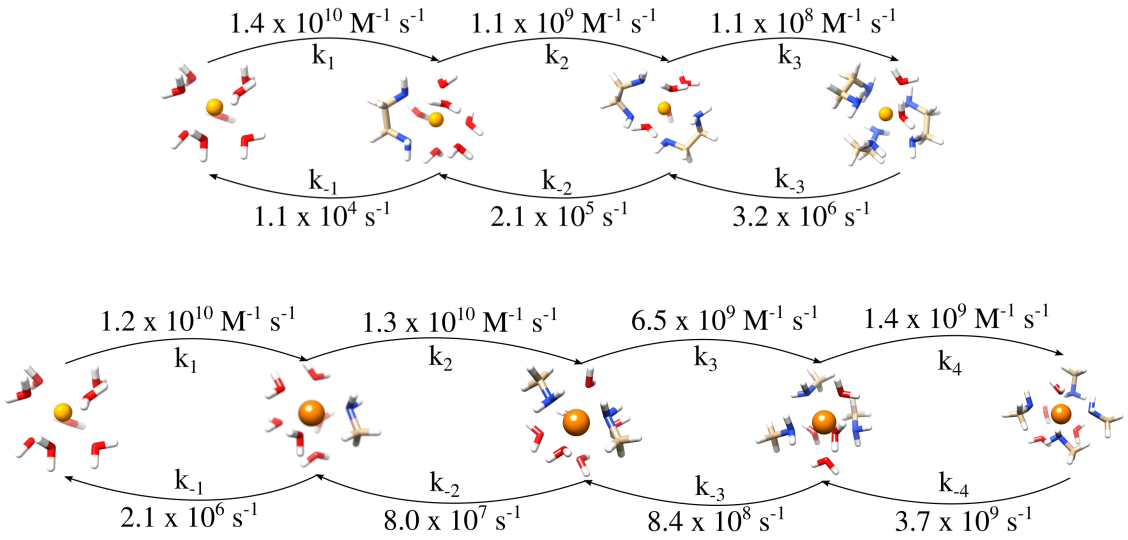

Figure 4.6: Kinetic formation and dissociation rate constants extracted from the reduced MSM for en (top) and nme (bottom).

the destabilization of the complexes does increase more rapidly with the number of amino groups as compared to the bidentate counterpart, as also reported in the work of Spike et al. [275].

It is believed that the possibility to disentangle the various energetic contributions of the complex stability is crucial for further applications. On one hand, it can be used to improve upon existing metal ion-ligand force fields, and on the other hand, it provides an effective tool for scrutinizing the molecular mechanisms driving the formation of metal complexes in solution. Through energy maps, it is possible to estimate the binding affinity of different coordination states, however, the barriers that can be extracted from a free energy pathway of thermodynamic profiles cannot be used to explain kinetically what is happening. Using the formation and dissociation rates for nme and en as in Figure 4.6, it appears clear that the formation rate is comparable for the same amino groups (nme_2 and en_1, approximately 10^{10} M^{-1} s^{-1}, and nme_4 and en_2, approximately 10^9 M^{-1} s^{-1}). However, the major difference (several orders of magnitude) lies in the dissociation rates, which seem to be the primary indicators of the chelating effect from a kinetic perspective). This fact is corroborated also experimentally where slower dissociation rates are usually considered to be responsible for the high stability constant of chelated metal complexes [294]. This can also be qualitatively interpreted through the free energy pathways of nme and en, where the fact that odd coordination states for the bidentate are at higher energies makes the disassembly of the complex with ethylenediamine more challenging from a kinetic standpoint. Given the paucity of experimental data regarding the kinetics of metal complexes, the approach presented becomes relevant, particularly due to its intrinsic flexibility and the fact that it is computationally convenient. This allows for the systematic study of exchange times for various metal centers with different ligands. Moreover, it can be especially informative for ions where conducting such experiments is currently unfeasible, as is the case with cadmium.

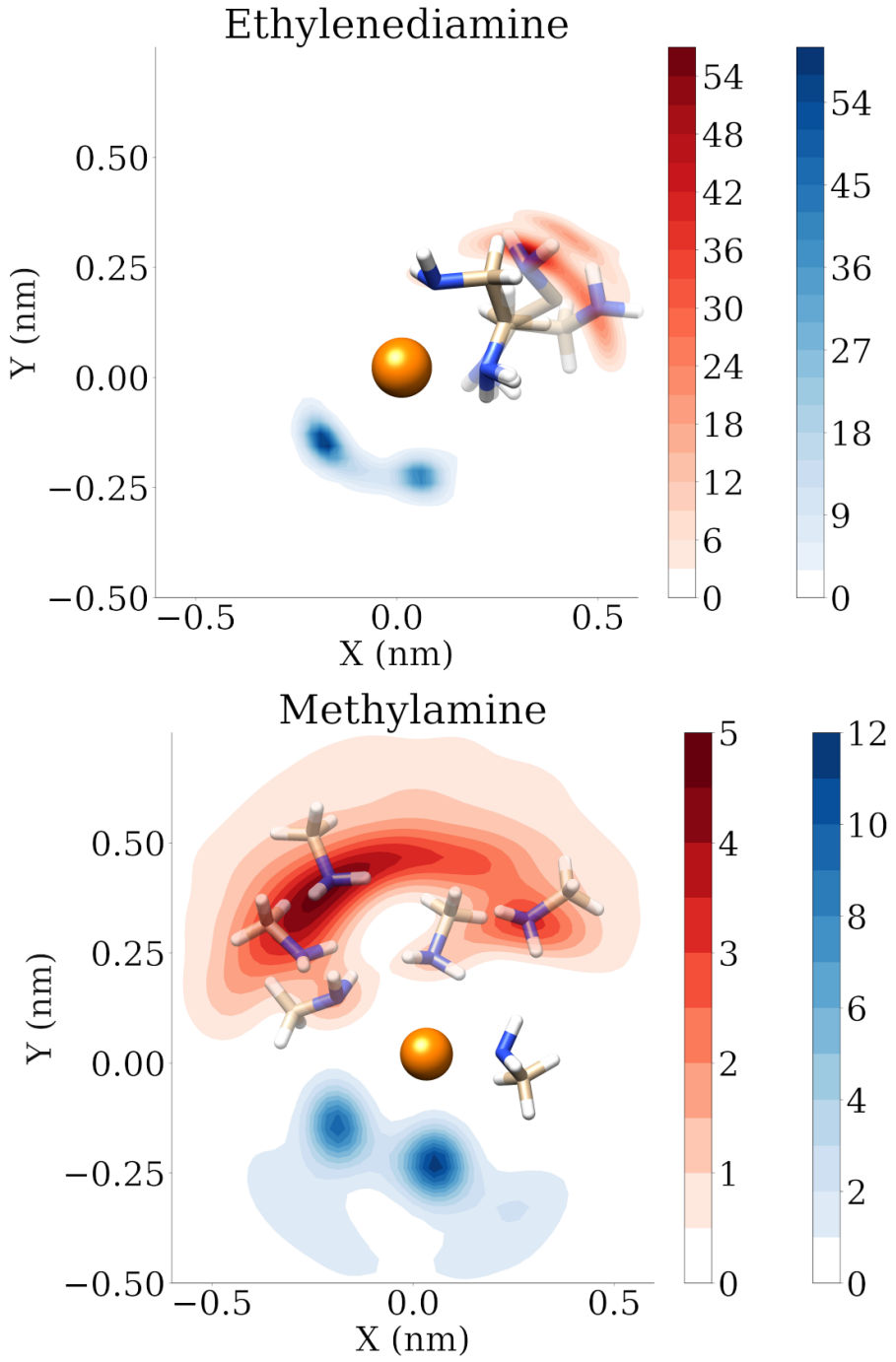

Figure 4.7: Different probability distributions of the second nitrogen entering in the first coordination shell (in red) and the water leaving the coordination sphere (in blue) for en (on the top) and nme (on the bottom) ligand after a first amino group has been coordinated to the cadmium cation.

Table 4.1: Thermodynamic quantities computed for Cd^{2+}-L_i with different ligands. ΔH^* and $-T\Delta S^*$ are ethalpies and entropies measured per amino group, dividing the corresponding quantity by the associated denticity. Thermodynamic quantities are in kcal/mol

L_i	denticity	ΔG	ΔH	$-T\Delta S$	ΔH^*	$-T\Delta S^*$
nme_1	1	-3.4	-5.5	2.1	-5.5	2.1
nme_2	2	-6.4	-11.2	4.8	-5.6	2.4
nme_3	3	-7.8	-15.1	7.3	-5.0	2.4
nme_4	4	-7.9	-19.5	11.6	-4.9	2.9
nme_6[1]	6	-3.8	-23.6	19.8	-3.9	3.3
en_1	2	-7.0	-10.4	3.4	-5.2	1.7
en_2	4	-12.4	-20.8	8.4	-5.2	2.1
en_3	6	-15.0	-28.5	13.5	-4.8	2.3
$dien_1$	3	-9.3	-13.0	3.7	-4.3	1.2
$dien_2$	6	-14.6	-24.0	9.4	-4.0	1.6
put_1	2	-4.3	-10.9	6.6	-5.5	3.3
put_2	4	-7.6	-21.3	13.7	-5.4	3.4

Mechanisms of ligand exchange and the chelate effect

The comparison between the minimum free energy pathways as obtained from the ethylenediamine and methylamine free energy maps (Figure 4.4) reveals some differences in the binding mechanism between the two ligand molecules. Both amines follow a dissociative mechanism for the first binding events, when the first en nitrogen atoms or the first nme molecule enters the cadmium first coordination shell. For both ligands, the cadmium coordination number probability distributions indicate that the first binding occurs after a water molecule has left the first coordination shell (Figure 4.3a, Figure C.9a). The second binding event, instead, shows a different mechanism between en and nme. As mentioned before in Section 4.3.1, the bidentate ligand can close its chelating ring with an associative mechanism, with a water molecule leaving upon binding of the second amino group. On the other hand, the second nme follows a dissociative mechanism. The probability distribution of the position of the leaving water molecule (Figure 4.7, Figure C.9d) is more spread for the nme complex, suggesting a higher tendency to lose a water molecule before the second nme binding in contrast to the en complex. The probability distribution of nme and water coordination number during the binding event also shows a higher tendency to lose another water molecule after the second binding (Figure C.9b). This higher liability of the first coordination shell is more evident with the increasing of coordinating nme molecules (Figure C.10a-f). The second nme molecule can enter in *cis* or *trans* position relative to the first nme, with the latter more probable. The en, instead, closes its chelating ring with a fixed configuration geometry (Figure 4.7). In both cases, the leaving water is in *trans* position relative to the entering nitrogen. Another water molecule leaves the first coordination shell, with a preferred exit angle $> 90°$ respect to the second coordinating nme. The probability distribution shows also that the water molecule between the two coordinating nme has a high probability of leaving, as if its position is destabilized by the presence of two ligand molecules. If the second nme molecule coordinates the cadmium ion in *cis* position (as the second en nitrogen atom), the leaving water position is similar to the ethylenediamine case.

Overall, the main difference in the binding mechanism between en and nme is the second binding event (1-2 binding). The presence of an associative pathway in the chelating ring closure of the Cd(en) complex formation shows the ability of the cadmium ion to temporary sustain over-coordination to achieve a better stability due to chelation. This behaviour is not present in the second nme binding, where a water molecule must leave before the entrance of a second nme, following a dissociative mechanism. Furthermore, the $Cd(nme)_2$ complex shows a more liable first coordination shell, with an high tendency to loose another water molecule after the coordination of a second nme. In particular, the water between the two nme molecules in *trans* position is more prone to leave, showing the tendency of the complex to achieve the same coordination geometry as the Cd(en) complex.

4.3.4 Entropic effects of pluridentate and chain length

An additional strategy for observing the chelating effect involves studying the variation in the stability of complexes with n-dentate polyamines. In this regard, the behavior of diethylenetriamine ($HN(CH_2CH_2NH_2)_2$) was examined and compared with its bidentate and monodentate counterparts. As detailed in the methods section 4.2.1, a new force field for dien was not developed; rather, the m12-6-4 used for en was transferred to dien. Although the calculated values do not perfectly align with experimental ones (as highlighted in Table C.10), they can still be considered meaningful for comparison with the bidentate and monodentate ligands.
The observation that the stability of tridentate complexes is underestimated emerges from Table C.10, where the enthalpy for denticity is lower compared to that estimated for en and nme, suggesting the need to increase the interaction between dien and cadmium to align with experimental pK values, in order to obtain an enthalpic contribution per amino group similar to other n-dentate amines or even higher as measured experimentally [234], or the need of a better parametrization for the dihedral angles of the alkyl chain taking into account the meaningful conformations (cis-cis, cis-trans or trans-trans) [314]. However, this enthalpic deviation does not obscure the appreciation of the distinctly entropic factor guiding the chelating effect, even for the tridentate ligand. It is noted that, considering a denticity of 3 in Table 4.1, the entropic contribution to the destabilization of the complex for dien is halved compared to that for nme. The most comprehensive comparison is at denticity 6 of Table 4.1, where the entropic contribution to complex formation follows the sequence dien > en > nme, a result consistent with various experiments [234, 273].
In literature, models have been proposed to explain trends in the stability of complexes with increasing denticity of polyamines. Here, an attempt is made to compare the results of the proposed approach with the one developed by Adamson [315], who tried to devise a model to calculate the pK_1 of n-dentate polyamines based on stability constants of ammonia for the same metal ion.

$$pK_1(\text{N-dentate}) = \sum_{i=1}^{N} pK_i(NH_3) + (N-1)log(55.5) \tag{4.7}$$

By substituting the stability constants of ammonia with those calculated for methylamine (assuming they are similar for Cd^{2+}, as experimentally measured [275]), it is observed that the estimated values using equation 4.7 generally seem to overestimate those calculated through the methodology presented here. For example, pK_1(en) is estimated to be around 6.1 compared to the measured value of 5.1 reported in Table C.10, and pK_1(dien) is estimated to be around 8.9 compared to the measured value of 6.8 in Table C.10. Other

subsequent models have been developed based on this, but the results do not deviate significantly from those mentioned above.
In examining monoamines or synthetic polyamines, the focus has been primarily on compounds of interest in inorganic chemistry. However, in this brief paragraph, consideration is given to putrescine ($1,4-Diaminobutane$) as a ligand, representing the class of biogenic polyamines, which holds special biological significance, together with spermine and spermidine. In this case, similar to dien, no *ad hoc* force field was formulated for putrescine as a ligand. Instead, the nitrogen-cation interaction used for nme (see sec. 4.2.1) was transferred assuming that this ligand behaves more like two monodentate amines separated by four methyl groups rather than classic bidentate ligand as en. The limitations proposed for dien, where better parametrization of non-bonded interactions and of the dihedral angles of the alkyl chain could improve agreement with experiments, are therefore also applicable here.
Table C.10 reveals that the estimated stability constants (pK_1 and pK_2) underestimate experimental values by about 0.8. However, this can still be considered a good parametrization, given what has been mentioned earlier. Analyzing the energetic contributions in Table 4.1 reveals that, on the one hand, the enthalpic contribution per amino group aligns with other synthetic amines, but the associated entropy is higher than measured in other systems. This observation may suggest that greater rotation freedom of the second amine, due to the longer chain separating the two amines, may contribute more to the entropy associated with the complex destabilization.
It is acknowledged that the investigation of polyamines with increasing alkyl chain lengths introduces various challenges, alongside the evident complexity associated with ligand parametrization. These include different protonation states with increasing methylene chain [316], hydrogen bonding between different ligands with nitrogen coordinating and non-coordinating nitrogen [317], and the formation of complexes in a bridging conformation [318], which may require different approaches for pK estimation. Nevertheless, we believe that this aspect does not diminish the value of the proposed methodology; on the contrary, it remains applicable for the exploration of significant phenomena. For example, experimentally establishing stability constants for putrescine with divalent cations is very challenging due to precipitation formation issues [319]. Therefore, the use of this method could be interesting as it allows the systematic study of biogenic polyamine complexes with various biologically relevant cations, for which using an approach of comparing relative stability constant of complexes rather than absolute ones could be very interesting in biological applications [320]. Alternatively, attempting to study how the ring-closing rate changes with the size of the methylene chain [321] or trying to justify why mercury prefers to form mainly linear rather than ring complexes [322] could be very interesting future directions.

4.4 Conclusions

This work has presented a comprehensive and computationally efficient methodology that allows for the study of the stability and kinetics of metal complexes in solution through molecular dynamics simulations.
The method discussed was tested on the Cd^{2+}-en complex and was subsequently applied to different cations (Ni^{2+}) and a series of amines as ligands (nme, dien, put). It has been demonstrated that a classical force field accurately tuned can predict fairly well the stability constants of metal ion-ligand complexes in solution, despite some notable differences with experimental values between enthalpic and entropic contributions, which are somewhat compensated in the case of the free energy. In particular, the role of entropy in the chelate effect has been investigated, revealing a decrease for ligands with increasing denticity con-

firming experimental findings [279, 275, 234, 273]. Another key aspect of this work was to elucidate the exchange mechanism occurring in metal complexes, initially determining the type of mechanism (associative or dissociative) and subsequently exploring molecular changes in the first solvation shells of the metal ions interacting with water molecules being released or incoming amine groups. Finally, starting from the thermodynamic states explored through enhanced sampling, it was possible to extract information about the formation and dissociation rates of metal complexes much larger than those explored with usual simulation methods. To the best of our knowledge, this is the first computational study in the literature to investigate exchange kinetics for metal complexes.
The methodology has some limitations, such as the necessity to work with an excess of ligands for a correct estimation of the number of free ligands, and it should be noted that the procedure reflects the underlying force field. However, having a systematic method for investigating both the thermodynamic functions and the rate constants of metal complexes can significantly aid the development of improved force fields thus striving to bridge the gap with experiments. Furthermore, while the study of more complex ligands (macrocycles or longer polyamines) may require more effort in terms of force field accuracy and defining new collective variables, the interest that biology and technology have in these systems should not deter from the use and improvement of computationally innovative approaches, such as the one presented here.

Chapter 5

Chanalyzer: a computational geometry approach for the analysis of protein channel shape and dynamics

In this Chapter a novel geometric algorithm to analyse the shape of a generic pore and provide an estimate of the translocation pathway is discussed. Morphological characterization, which includes identification of the main axis, the corresponding local radius, and the detailed description of the global shape of the cavity, is integrated with a physico-chemical description of the surface facing the pore lumen. Remarkably, the possible existence or temporary appearance of fenestrations from the channel interior towards the outer lipid matrix is also accounted for. As a test case, we applied the present approach to the analysis of an engineered protein channel, the mechanosensitive channel of large conductance (MscL). This chapter is part of a broader collaborative work in which my contribution focused on the molecular dynamics simulations and analysis performed on MscL, used for comparison with the Chanalyzer algorithm. Therefore, a concise introduction to the problem and a brief overview of the mathematical and geometrical aspects of the project will be provided before delving into the sections I coauthored. This Chapter is based on an article already published by Raffo et al. [323].

5.1 Introduction

As discussed in the Introduction, ion channels play key roles in biology and for this reason they are the target of over 20% of the drugs on the market. Therefore, characterizing the structural and dynamic features of ion channels can significantly improve our understanding of their functioning and unveil the more subtle details of their mechanism, which is often elusive to experimental observations.
Here, we propose a novel computational tool for the automatic recognition and structural characterization of a protein channel from a given MD trajectory, rooted into the alpha-shape complex analysis and the notion of discrete-flow as described in [324]. Alpha shapes theory is also at the basis of how the NanoShaper software builds the protein Solvent Excluded Surface (SES) and finds cavities and pockets [325]. This method enriches the capabilities of NanoShaper to identify and characterize cavities in molecular structures [326] and is focused more specifically on the identification of permanent or transient channels

formed within a protein, from which it was dubbed Chanalyzer. Notably, it does not require predefined user parameters for channel identification, such as the notion of membrane plane, it is numerically robust and well-suited for analyzing large collection of molecular configurations as issuing from extended MD simulations of biological systems. In addition, the method supports a detailed geometric analysis based on the concepts of skeleton and centerline and identifies both channel ends through the use of graph-based techniques. The identification of the channels is computationally efficient and a proximity strategy has been adopted to accelerate the calculation of channel entrance and exit from one MD frame to the following one. Interestingly, its extensive geometric characterization allows for the identification of the channel main axis, but also of possible ancillary tunnels and potential fenestrations facing the lipid membrane. It also provides a geometric approximation of the local section orthogonal to the central axis of the lumen, to more easily identify symmetry breaking configurations and anomalies. To validate the present approach and to show the complete set of geometrical information that it provides, we analyzed a number of MD trajectories of an engineered protein channel, the mechanosensitive channel of large conductance (MscL). Overall, results show a very good agreement with previous calculations of channel local radius, as well as the provisioning of new information thus supporting the routinely use of Chanalyzer for the detailed study of protein channels in a large variety of biological systems.

5.2 Methods

5.2.1 Chanalyzer and geometrical approach

The MD trajectory of the channel is converted from the original *dcd* to the multiple *pdb* format using the VMD tool, [216] after excluding water molecules and ions. Then it is split into individual frames and finally annotated with the atomic radii information. Finally, only the information on atom centers and radii is retained in a .xyzr file, which is the standard input of NanoShaper [325].
Firstly, the channel is identified using alpha shape theory [327]. This is followed by a geometric characterization to identify and prune the "skeleton" of the channel. Subsequently, the visible contour of the channel, as shown in Figure 5.1, is extracted. Lastly, the centerline, approximating the translocation pathway along the pore, is computed using the Vascular Modelling ToolKit (VMTK) [328]. Its overall complexity is estimated to be $O(n^2 log(n))$. For further details on this algorithmic section, more information are provided in [323].

5.2.2 Molecular dynamics simulations

The studied molecular systems are four variants of the engineered MscL channel, a pentameric protein channel, originally investigated in [329] (PDB code 2OAR). Each variant differs from the others in the way residue 21 is functionalized among the five subunits. In this functionalization, a photo activating ligand, namely the 6-nitroveratryl alcohol, which splits into 6-nitrosoveratyl aldehyde and a free acid upon light irradiation, was attached through a Cysteine-selective alkylating reagent to the residue 21 of each protein monomer (more details can be found in [329]). Concerning the MD simulations, the adopted force field is CHARMM (v.27 [330]) for the protein and (v.36 [331]) for the lipid. The parameters for the photo-activating ligand have been computed through QM calculations at the DFT level of theory. The water model is TIP3P, while the ionic concentration of K^+ and Cl^- is set to 1 M. Equilibration is carried out for about 10 ns in the NpT ensemble, followed by the production run according to the NVT ensemble under normal conditions (T=300K).

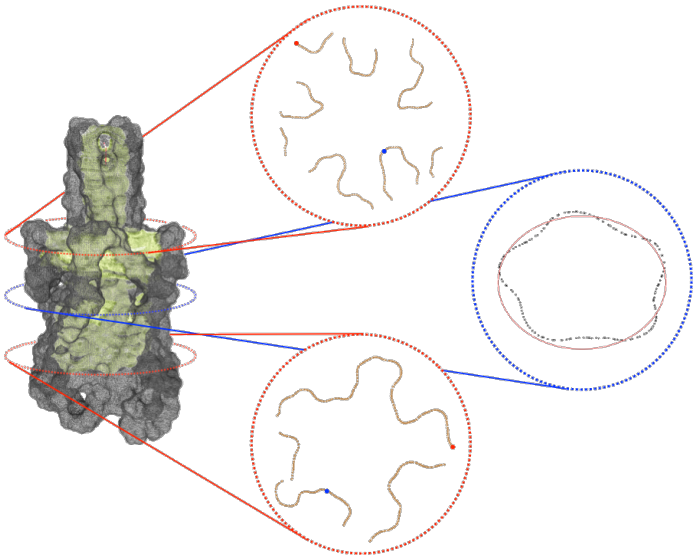

Figure 5.1: Visible contour of a channel. On the left, the visible contour of the channel (in light green) and the molecular surface (in grey). On the right, three sections of the visible contour: on the top, the section coinciding with the bifurcation of the skeleton; in the middle, the central section of the channel; at the bottom, a section clearly revealing the pentalobated nature of the channel. In each section, we spotlight some of the geometric features provided by the proposed approach. Specifically, in the top and in the bottom sections, the closest and the farthest points to the centerline are represented in blue and red, respectively. In the middle section, the ellipse (depicted in red) that best fits it is shown. Its knowledge allows to retrieve further information about the local channel shape, such as its eccentricity.

For non-bonded interactions a cutoff of 12Å is used. All simulations are performed with periodic boundary conditions, treating the long-range electrostatic interactions with the Particle Mesh Ewald (PME) algorithm. [332].

5.3 Results and Discussion

The Chanalyzer approach is applied to the analysis of the MscL system in order to compare the present analysis to the one originally carried out in [329]. In that work, a variable number of modifications of the MscL channel was applied to generate corresponding molecular models by attaching a photo-activating ligand at residue 21 to the five monomers of the protein, as experimentally tested by Feringa and collaborators [333]. We focused on four of these functionalized systems, namely the NL, 1L, 3L and 5L, having 0, 1, 3 and 5 photo-activating ligands, respectively. In the analysis reported in [329], snapshots of the MD trajectory were superimposed and fed to the HOLE software [334] after removing the side chains, to measure the radius of the channel of the differently generated models. This was aimed at highlighting the symmetry breakage and at confirming the progressive engineered expansion of the channel radius with sequential addition of negative charges upon photo-ligand removal. This effect is apparent in Figure 5.2 where the computed average radii evaluated along the longitudinal channel axis (i.e., at different z values) by the present approach and by the HOLE software [334] are shown. Averages are performed over

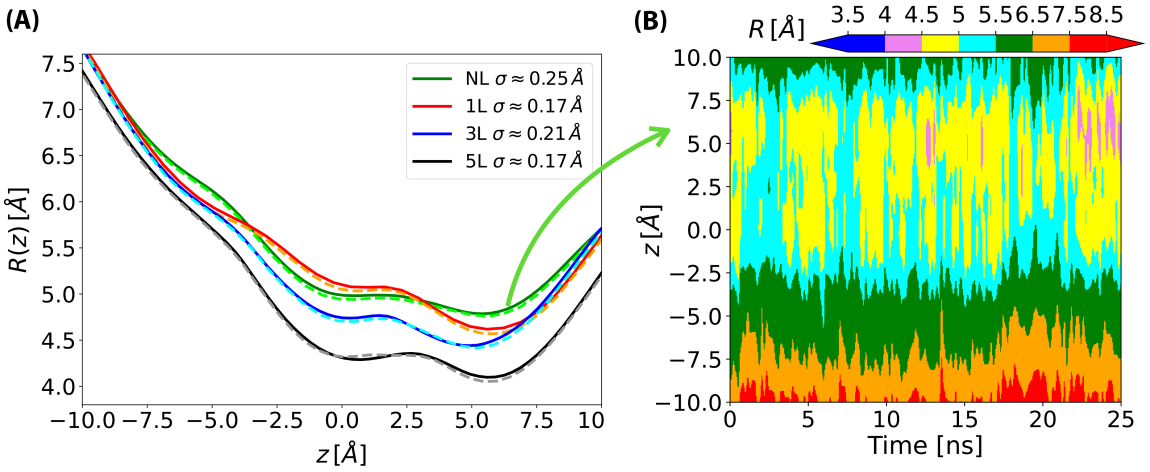

Figure 5.2: (A) Solid lines, time-averaged channel radius along the axial z position for each of the considered systems as obtained by Chanalyzer with associated standard deviation (in the legend). Dashed lines, the same radius derived via the HOLE software. (B) Example of the dynamical behavior for the no-ligand system. The colormap is associated to the instantaneous value of the radius, as returned by Chanalyzer.

about 150 MD configurations for each considered system and a re-binning procedure is used along the z-axis. The agreement between Chanalyzer (solid lines) and HOLE (dashed lines) data is extremely good, with results basically indistinguishable in most regions and abundantly within the standard deviation in correspondence of local minima and saddle points. The observed systematic shift in pore radius (from 5L to NL) supports an increasing role of the residual charges, thus confirming the previously observed charge-mediated MscL gating [335].

In two of the simulated systems, namely NL and 1L, potassium ion percolation was also observed. It is therefore interesting to compare how close are ion permeation paths to the centerlines of the corresponding MscL channels. Note, however, that ion translocation pathways can be affected by local and specific steric or electrostatic effects induced by residue side chains, not accounted for in our evaluation of the centerline which is based only on backbone atoms. In Figure 5.3, the centerlines, as returned by Chanalyzer, and the average trajectories followed by the K^+ ions in the NL and 1L systems are depicted. In the latter case, the average is carried over the multiple MD configurations and a re-binning is performed along the z-axis. The general trend is that in those locations where the radius is smaller, as expected, ions are more constrained towards the centerline. However, around $z = 10$Åthe ions tend to be more displaced towards the channel walls as compared to other locations with similar radii. This is likely a consequence of the net electrostatic attraction of the charged residues, upon photo-ligand removal.

Interestingly, this suggests that a systematic comparison between centerlines and ion trajectories could be used as an indirect way to probe the local interactions between ions and residues along the channel and can be helpful in suggesting preferential mechanisms affecting translocation.

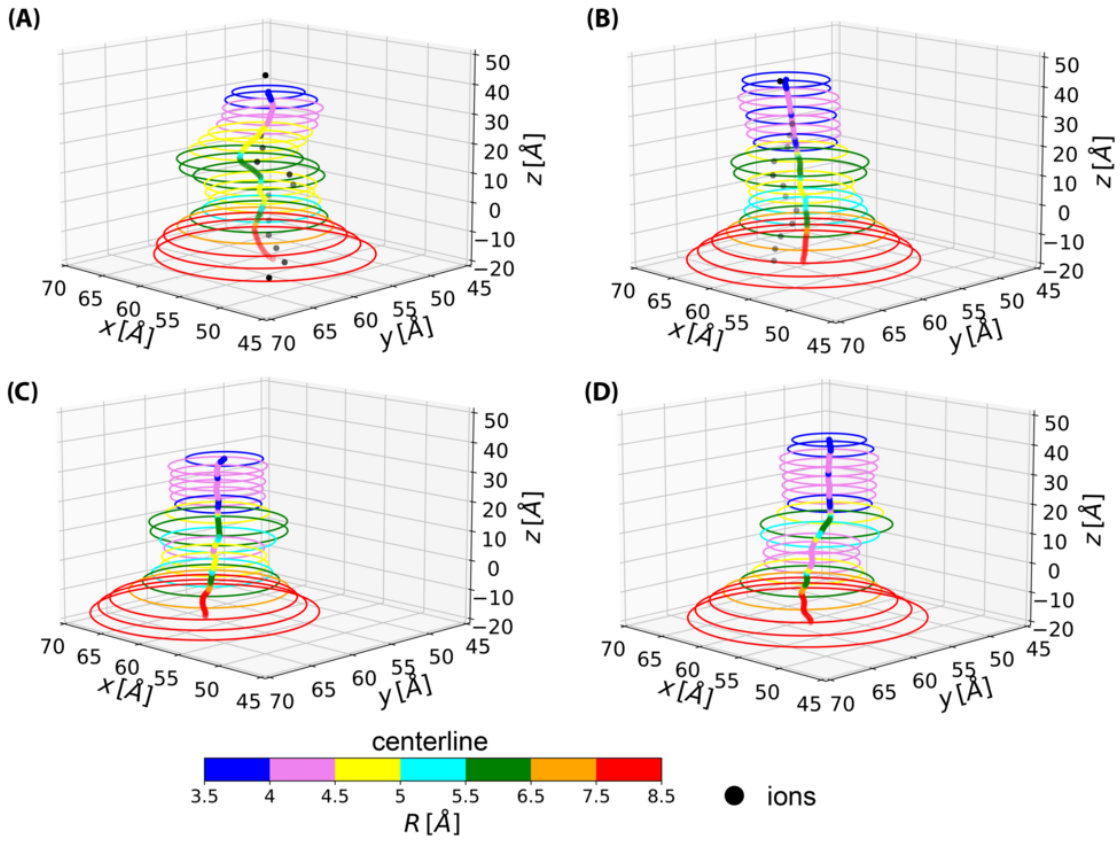

Figure 5.3: Average centerlines. Colors code for the size of the associated radius. Black dots are average ion positions for the permeating configurations. From top to bottom and left to right: (A) NL, (B) 1L, (C) 3L, and (D) 5L.

5.4 Conclusions

We presented here the application of the Chanalyzer geometric approach to the analysis of the channel morphology and dynamics of four differently functionalized forms of the MscL system, as a test case.
Computational geometry has been already exploited to study the details of molecular structures. For instance, cavities and tunnels arising at the molecular surface of a protein have been studied with the NanoShaper software [325, 336]. In the field of protein channels, a popular analysis tool is provided by the HOLE software [334], which finds the maximum radius of a sphere centered within the channel starting from a given point (provided by the user to be inside the channel), so as that it does not overlap with the van der Waals interior surface of the pore and makes that sphere proceed and adapt its size throughout the channel, assumed to be nearly rectilinear.
Successively, two tools have been developed by the same group, namely CAVER [337] and its improved version, MOLE [338], to explore routes between protein clefts and cavities. CAVER's underlying algorithm is based on a skeleton search using a three-dimensional grid. Finally, MolAxis [339], a more recent tool also based on alpha shape theory, was successfully applied to the 5HT3 receptor [340] to identify lateral ion channels besides the central longitudinal one. However, MolAxis still strongly relies on user parametrization and

can suffer from method specific artifacts and approximations.
While further improvements of the Chanalyzer project are still needed and its development is currently ongoing, it already sets up a framework that enables the accurate evaluation of several channel features that start from the purely geometric analysis but can easily integrate other relevant physico-chemical information, e.g. the chemical nature (i.e., atom identification) of the pore lumen as it inherits the properties of the SES calculated by NanoShaper.
Remarkably, the fact that Chanalyzer does not need user-specific parameterization, such as a predetermined direction of the channel axis, does represent a clear advantage in the treatment of those cases where the geometrical shape is not predominantly tubular and may present bi- or multi-furcations, as well as ancillary pathways towards the surrounding lipid matrix. These so-called fenestrations may have relevant biological or biophysical implications still not well known [341]. For such reasons, we believe that biophysical modelling can significantly benefit from user-friendly and versatile geometric approaches, such as Chanalyzer.

Chapter 6

Unraveling the molecular origin of an inherited channelopathy in the voltage-gated potassium channel Kv4.3

In this Chapter, a study that integrates experimental and computational methods is discussed, focusing on the voltage-gated potassium channel Kv4.3 and two of its potentially harmful mutations. As highlighted in the Introduction, potassium channels play a crucial role in human neuronal activity. Understanding the functioning of the wild-type channel and the impact of point mutations in different channel regions is vital for comprehending complex pathologies. This research was conducted in collaboration with an experimental group at the University of Twente, led by Prof. Armagan Kocer, which performed patch clamp experiments on voltage-gated Kv4.3 potassium channels expressed in Chinese hamster ovary cells.
Given the system's complexity and size, the computational effort was substantial. On one front, the focus was on molecular modeling and MD simulations to investigate how point mutations might molecularly affect channel conductance and the role of water in ion translocation. On the other, a Markov state model (MSM) was used to model the channel's current response to voltage excitations, as mentioned in section 1.3.2. This approach was necessary due to the timescale of the tetrameric protein channel's closing and opening being too large for standard MD simulations (approximately 10^{-1} s). The former approach shed light on the crucial role of hydration in the selectivity filter and the potential harm caused by mutations introducing more hydrophobic residues in that channel region. The latter highlighted the increased inertia of the mutants' inactive states compared to the wild-type.
This Chapter is based on work that is being prepared for submission.

6.1 Introduction

Kv4.3 is a voltage-gated potassium channel that plays a crucial role in the electrical signaling of the brain and heart [342]. It forms a heteromeric complex with its auxiliary partners potassium channel interacting proteins (KChIPs) [343] and the dipeptidyl peptidase-like proteins [344, 345]. The channel generates somatodendritic so-called 'A-type' transient potassium currents. In the cerebellar Purkinje neurons, Kv4.3 activates at membrane potentials that are subthreshold to potentials required for generating an action potential (AP).

While active, by releasing K^+ ions out of the cell, these channels counterbalance membrane depolarization and temporarily resist the generation of an AP. However, the channel also inactivates very fast, and the neuron can fire an AP upon receiving excitatory signals until Kv4.3 recovers from inactivation. As AP metrics encode the neuronal information, any dysfunction or altered kinetics of Kv4.3 can significantly affect neuronal communication [346, 347, 348, 349, 350]. Indeed, pathogenic mutants of Kv4.3 in Purkinje neurons, which provide the sole electrical output of the cerebellar cortex for fine-tuning the voluntary motor and cognitive activities, were associated with the movement disorder spinocerebellar ataxia type 19/22 (SCA 19/22) [351, 352, 353]. Dysfunction of these channels degenerates Purkinje neurons, with an unknown mechanism, leading to neuronal death and cerebellar atrophy. The patients lose control over their walking, speech, body balance, and eye movements. Currently, there is no cure or treatment to reduce the symptoms. Furthermore, while hereditary mutants show their effect mostly later in life, recently, de novo mutations with devastating consequences at very early ages also started to emerge [354, 355], urging a better understanding of the (mal-)functioning of Kv4.3 channels.

Kv4.3 is a tetrameric protein with cytoplasmic N- and C-termini (Figure 6.1A). Each subunit has a transmembrane domain (TMD) comprised of six alpha helixes (S1–S6). While S1–S4 forms the modulatory and highly conserved voltage-sensing domain (VSD), S5–S6 helices form the innermost structures and line the pore domain (PD) (Figure 6.1B and 6.1C). Voltage-induced conformational changes in the VSD are mechanically coupled to the PD via the S4–S5 linker and potassium ions are coordinated at the selectivity filter (SF) at the outer end of the S5–S6 loop (Figure 6.1C).

Kv4.3 shares the key characteristic structural features of all voltage-gated potassium (Kv) channels in its core region (S1–S6), but it has specific and highly conserved structural aspects that define its different working mechanism [356]. It activates and, more importantly, inactivates rapidly in ways different from the conventional N- or P/C-type inactivation [356, 357, 358, 359]. Before opening with depolarizing voltages, the resting channels undergo multiple conformational changes and reach a closed-active state (CA) (Figure 6.1D). Upon further depolarization, the channels either open (O) or, more preferentially, turn into a closed-inactive state (I0-I4) [356, 357, 358, 359, 360]. Inactive channels recover rapidly at hyperpolarized membrane potentials [356], and unlike Shaker-type Kv channels, they do not re-open during recovery from inactivation [358].

In the case of pathological Kv4.3 mutants, previous studies focused on the localization of the channel and its activation and inactivation in the presence of the main auxiliary subunit potassium channel-interacting protein (KChIP2b), which will be referred to as KChIP for the rest of the article [352, 353, 361]. It has been shown that most of the mutants have trouble reaching the cell membrane in mammalian cells, and some display significant changes in the voltage-dependence of gating. However, due to the lack of a suitable crystal structure, it is not clear how such pathogenic mutations affect the Kv4.3 structure and, in turn, how these changes alter the gating mechanism, kinetics, and modulation by KChIP2 at the molecular level. In this study, we systematically investigated two hereditary mutations, M373I and S390N, to understand how the point mutation affected the structure and function of the Kv4.3 channel alone and in the presence of KChIP. These mutations are predicted to be in the selectivity filter and S6 helix, respectively, and both are well-conserved among wild-type (WT) Kv4.3 orthologues in various organisms. While M373I causes mild signs in patients, S390N causes severe symptoms, ranging from saccadic eye movements and hearing deficits to cognitive impairment (12). To investigate these mutations, we first generated a homology model of the human Kv4.3 (Figure 1A-C) to map the mutation sites onto the protein structure and allow structure-function studies. Then, we performed

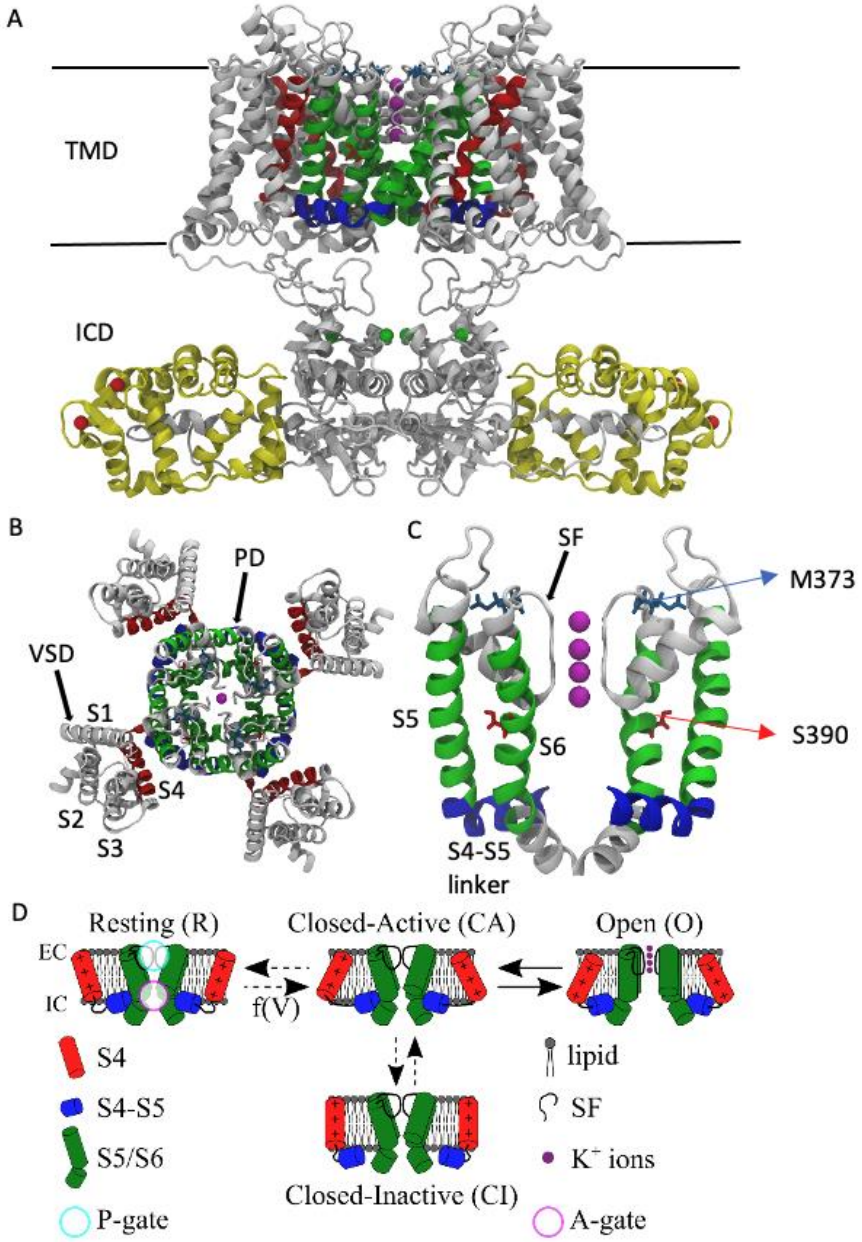

Figure 6.1: Atomistic model and current gating model of Kv4.3 channel. (A) Side view of the open conformation of the wild-type (WT) Kv4.3 full-length model in its tetrameric form showing transmembrane (TMD) and cytoplasmic intracellular (ICD) domains and KChIP1 auxiliary subunits. Black lines indicate the lipid bilayer. Crystallographic potassium, zinc and calcium ions are shown magenta, green and red, respectively. Kv4.3 and KChIP1 residues are shown in white and yellow, respectively. For clarity, only two of the four KChIP1 auxiliary subunits are presented. (B) Top view of the WT TMD with the location of voltage-sensing domains (VSDs) (S1–S4 residues 182–307) and the pore domain (PD) (S5/S6 residues 321–402). (C) Side view of the WT PD showing two monomers. The location of the SF (residues 367–372) is indicated by a black arrow. The sites of point mutations in WT Kv4.3 residues M373 and S390 are highlighted in sky blue and red licorice, respectively.

a systemic functional study on the electrophysiological activity of the WT Kv4.3 and two mutants at the single-cell level. Finally, we adopted a multi-state Markov model to describe our experimental findings on channel kinetics and to get further mechanistic insights into the activation/inactivation mechanism. Our data showed that, the functional consequences of the mutations on the channels structure and function and their differential modulation by the KChIP and presented mechanistic explanation for key functional findings.

6.2 Methods

6.2.1 Molecular model of the Kv4.3 channel

An atomistic model of the human WT Kv4.3 channel was built up using a combination of homology modeling and X-ray crystal structures since there is a high sequence similarity among Kv channels of all families. Since there is a ≈47% sequence identity in the transmembrane domain (TMD: helices S1-S6) between the Kv4.3 channel and the Kv2.1 paddle-Kv1.2 chimera channel [362], the structural model of the Kv4.3 channel was generated using the latter high resolution X-ray crystal structure as a template (PDB ID: 2R9R) (Figure S2). Sequence alignments were performed with Clustal 2 [363, 364] using the graphical user interface ClustalX version 2.0 [365, 366]. The homology model of a single monomer of the Kv4.3 channel was obtained from the SWISS-MODEL server and Swiss-PdbViewer [367, 368, 369, 370, 371]. Then, a tetrameric form of the WT Kv4.3 channel in its open state was built up from the monomeric form of the homology model of the Kv4.3 TMD and the X-ray crystal structure of the Kv4.3 tetramerization (T1) domain, which constitutes the cytoplasmic intracellular domain (ICD) of the channel, complexed with its auxiliary subunits (PDB ID: 2NZ0) [372, 373] using the VMD 1.9.3 software package [216]. Then, the essentially nonfunctional T1 domain of the protein and its KChIP1 auxiliary subunits were omitted [374, 375, 376]. Moreover, the model of Kv4.3 TMD was validated with the Memoir server [377, 378] and the homology modeling algorithm designed for membrane proteins MEDELLER [379] by using the Kv4.3 TMD as a target sequence and the Kv2.1 paddle-Kv1.2 chimera channel TMD as a template. The structural alignments of Kv4.3 TMDs generated with two different methods gave Cα atoms RMSDs of the order of 0.2 Å. Four crystallographic K^+ ions were placed in the selectivity filter (SF: residues 367-372) using as a reference the position of the ions in the X- ray crystal structure reported by Long et al. in 2007 [362].

6.2.2 Molecular dynamics simulations

Initially, the structure of the wild type Kv4.3 TMD (residues 165-411) was embedded in a slightly asymmetric POPC lipid bilayer containing 443 lipids (225 and 218 POPC molecules were in upper and lower leaflets, respectively) and solvated in an aqueous solution with a 150 mM KCl salt concentration using the CHARMM-GUI Membrane Builder tool [380]. From the WT system, we generated two Kv4.3 mutant models: M373I, in the pore domain, next to the selectivity filter; S390N, in the S6 subunit. All MD simulations were performed with the NAMD 2.12 software package [381]. The CHARMM36 force field was used for the protein and lipids [331], and the TIP3P model for water [210]. The smooth Particle Mesh Ewald (PME) method was used to calculate the electrostatic interactions [332], with a short-range interaction cut-off of 12 Å using a switching function. A multiple time-step algorithm was used to integrate the equations of motion, with a time step of 4 and 2 fs for the long-range and short-range interactions, respectively, and 1 fs for the bonding interactions [382]. All bond lengths involving hydrogen atoms were held fixed using the

SHAKE algorithm [84].
Each system was subject to the following protocol:

- 5000 steps of energy minimization with the position of all protein and solvent atoms, ions and POPC polar headgroups fixed with a force constant of 1 kcal/molÅ^2.

- short NpT simulations with 0.5 fs to 2.0 fs timesteps were performed gradually releasing the harmonic positional restraints, at a constant physiological temperature of 310.15 K using a Langevin thermostat and at 1 atm pressure using a Langevin-Nosè-Hoover piston method [383, 384].

- 300 ns NpT equilibration simulation was carried out while keeping restrained the position of the crystallographic ions and the distances of opposing carbon atoms of the carbonyl moieties of the selectivity filter with a force constant of 10 kcal/molÅ^2 [385], followed by a 50 ns NpT equilibration releasing all restraints except those of the selectivity filter (0.25 kcal/molÅ^2, restraints are retained to prevent collapse and inactivation of the selectivity filter) and by further addition of KCl salt up to 500 mM (note that during this equilibration step some crystallographic ions are released in the solution, hence the production run had no memory of the initial arrangement of the crystallographic ions).

- NVT production runs were performed at 310.15 K applying a homogeneous external electric field (E_z) along the z-axis (L_z) perpendicular to the membrane and proportional to a voltage of 1 Volt (E_z = -1 V/L_z) [329].

Even though the applied voltage is relatively high compared to standard experiments, we note that similar voltages have been used in several previous computational studies [376, 386] to speed up ionic flow and potassium channels have shown good stability under such conditions, as observed in our work. Moreover, due to the comparative character of our study between WT and mutants, the main findings are not affected by this choice.
The channel pore radius was measured with the HOLE program [334]. Analysis of contact maps and residue-residue distances were performed with Gromacs software tools [197]. Images, movies and trajectory analyses were performed using VMD graphical tools, analytical plugins and Tcl scripts. The number of permeation events was measured with a homemade script, dividing the simulation box in three different regions, the first one at the intracellular side of the lipid membrane, the middle one at the selectivity filter level and the last one at the extracellular side of the membrane.

6.2.3 Kinetic model of the Kv4.3 channel

In this work, a recently proposed multi-state Markov model [159] was adopted to describe the voltage-gated dynamics of the potassium channel Kv4.3 in the presence of KChIP2b, as observed in our experiments. The electrophysiological measurements of the channels in the presence and in absence of KChIP on the kinetics of the steady state activation, inactivation, closed-state inactivation and recovery from inactivation were used to fit the model rate constant parameters to investigate the differences between WT and two single point mutations, namely M373I and S390N. A system of ordinary differential equations associated with the kinetic model was implemented using Python (version 3.8.3) and solved numerically using the Runge-Kutta algorithm as implemented in the Scipy library with an integration step of 0.0001 ms. Moreover, the voltage clamp (VC) step protocols simulated

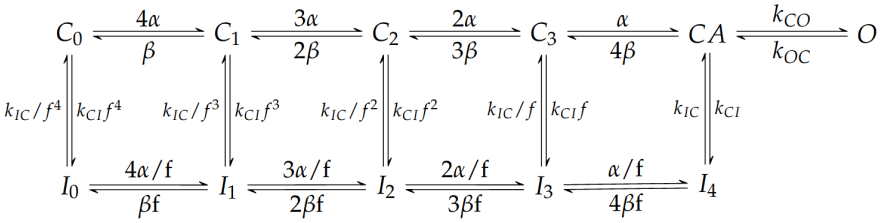

Figure 6.2: Gating scheme with five closed and five parallel inactivated states and a single open state. Voltage-dependent rate constants have been defined as $\alpha(\mathrm{V}) = \alpha_0 e^{(\alpha_1 \frac{VF}{RT})}$, $\beta(\mathrm{V}) = \beta_0 e^{(-\beta_1 \frac{VF}{RT})}$, $k_{CO}(\mathrm{V}) = k_{CO_0} e^{(k_{CO_1} \frac{VF}{RT})}$ and $k_{OC}(\mathrm{V}) = k_{OC_0} e^{(-k_{OC_1} \frac{VF}{RT})}$. Where F is the Faraday constant, R the gas constant and T the temperature at which the experiment was conducted. k_{CI} , k_{IC} and f are constants and hence voltage independent.

with the kinetic model were implemented to reproduce closely the corresponding voltage-time functions adopted in the experimental procedures for activation, inactivation, CS-inactivation and recovery from inactivation data. The model parameters were optimized simultaneously on four different datasets each for WT and single point mutations M373I and S390N.
The four datasets were from experimental steady state data of activation, inactivation, CS-inactivation and recovery from inactivation VC protocols. The total cost function has been implemented as described in [158], where the total cost function (F) is the sum of the n partial cost functions (f_i) and N=4 are the number of VC protocols taken into account as expressed by

$$F(x_1, \ldots, x_k) = \sum_{i=1}^{N} w_i f_i(x_1, \ldots, x_k) \tag{6.1}$$

where x_i are the parameters to be adjusted, k is the number of parameters to be optimized and w_i are the weights for each protocol. The partial cost functions have been computed as the mean-square deviation of the simulated and experimental data. Given the number of data points (M) and the absolute experimental or simulated value considered (y_j^{exp}, y_j^{sim})

$$f_i(x_1, \ldots, x_k) = \frac{1}{M} \sum_{j=1}^{M} f_i(y_j^{exp} - y_j^{sim})^2 \tag{6.2}$$

The rate constants of the model employed by Lin et al. [159] were used as a starting point for the optimization algorithm. The optimization algorithm used was the differential evolution algorithm [387] as implemented in Scipy (v.1.9.3), this global optimization algorithm allowed us to find the best possible solution in the parameter space, that was restricted to physically sound boundaries (i.e. no negative values or greater of the inverse of the integration step 10000 s^{-1}).
A second global optimization was performed on the two mutants keeping fixed the parameters α_0, α_1, β_0 and β_1 of the model Figure 6.2 as the ones found for the wild type just for the system in the presence of KChIP. Finally, the simulated steady states curves for activation and inactivation were fitted with a Boltzmann function to extract the computed hemi potential $V_{1/2}^{act}$, $V_{1/2}^{inact}$ and the computed slope k^{act}, k^{inact}. The CS-inactivation and recovery from inactivation steady states curves were fitted by a simple exponential to extract the computed τ_{csi} and τ_{rec}.
This procedure has been done for the WT, M373I and S390N both in absence and in presence of KChIP. The parameter set for the model presented are compared in Tables D.1 and

D.2.
Furthermore, this methodology has been expanded to develop a software called pyChanneLab, featuring a user interface (UI) entirely written in Python language and available under the MIT License on GitHub at https://github.com/lianctrl/pyChanneLab. The workflow is illustrated in Figure D.1. Initially, the MSM object is created using an MSM Editor as shown in Figure D.2, and then exported. Next, the experimental protocol object is generated via the Experiment Builder as depicted in Figure D.3, and exported. After creating both objects, experimental data from voltage clamping experiments can be loaded, ensuring they match the number of protocols generated previously. The initial parameters of the constructed MSM can be selected, along with various options for the global optimization algorithm, such as integration timestep, weight of each protocol, boundaries for parameter search, and stopping criteria. Upon completion of the optimization, the optimized parameters are displayed and can be utilized to compare *in silico* experiments with experimental data.

6.3 Results and Discussion

6.3.1 A molecular model of the Kv4.3 channel and its mutants

To gain a mechanistic understanding of the main structural and functional differences induced by point mutations, an atomistic model of the Kv4.3 channel in its open state was generated using a combination of X-ray crystal structures and homology modeling techniques, as described in the Material and Methods section. Kv4.3 TMD (residues 165–411) reflects closely the typical architecture observed in other Kv channels, where each monomer is equipped with a VSD (S1-S4, residues 182–307) linked to the central pore-forming S5 and S6 helices, the latter embedding the highly conserved ion SF (residues 367–372) of the channel (Figure 6.1A and 6.1B). From the model, it can be observed that both mutations, namely 373 and 390, are within the PD, but in different spatial regions of the channel (Figure 6.1B and 6.1C). Position 373 is located on the H5 loop connecting S6 with S5 immediately above the SF segment. The mutation introduces a more hydrophobic residue isoleucine in place of methionine in this interfacial region of the protein facing the extracellular environment (Figure 6.1B and 6.1C). On the other hand, position 390 is approximately located in the middle of the S6 helix (residues 382–402) within the pore lumen and below the SF, and it is oriented away from the channel central axis and towards the S5 helix (Figure 6.1C). S390N introduces a bulkier side chain (i.e., asparagine) in place of a serine in a tightly packed interfacial region between S6 and S5. Our model suggests that M373I might affect the potassium ion conductance and the channel inactivation via altering the hydration properties and the rigidity of the SF, whereas S390N might have more severe consequences for the channel function.

6.3.2 Mutations affect potassium ion translocation in MD simulations

To shed some light on the molecular determinants causing the observed decrease in single-channel conductance, a comparative in silico investigation between the mutants and the WT Kv4.3 channel was carried out. Here, the main results highlighting the structural and functional differences issued from one microsecond MD simulations are presented. Kv4.3 TMD was embedded in a 1-palmitoyl-2-oleoylphosphatidylcholine (POPC) lipid bilayer for both WT and mutants and exposed to a 500 mM KCl aqueous solution, while an applied voltage of 1 V was maintained throughout the simulation to study the channel during its active state at exercise conditions. In the case of M373I, only minor changes in the protein

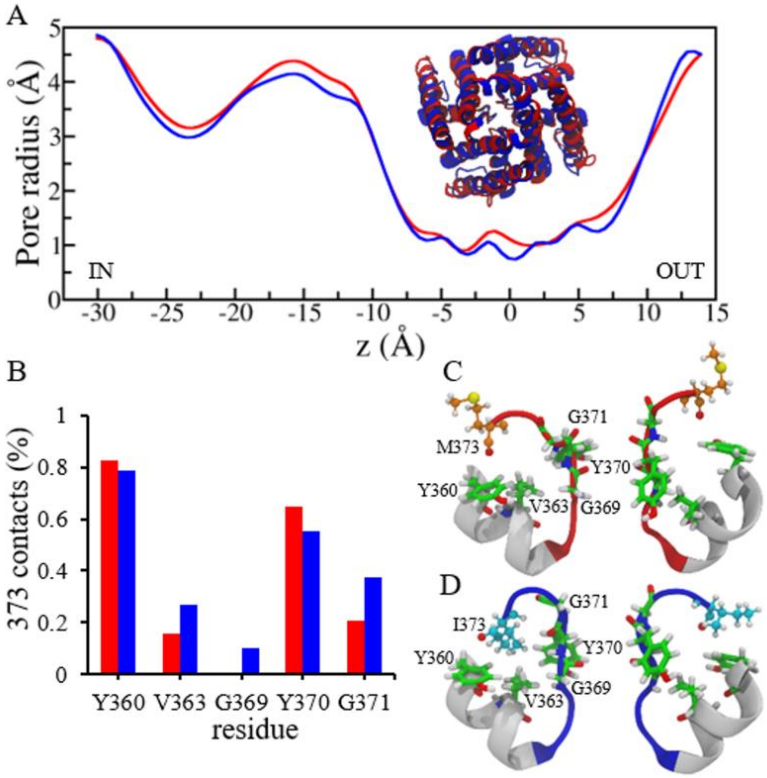

Figure 6.3: (A) Average pore radius along the channel axial position (z-coordinate) for human WT (red) and M373I (blue) Kv4.3 TMDs embedded in a POPC lipid bilayer and simulated with an applied voltage of 1 V. Inset shows a bottom view of the structural alignment of PDs for WT (red) and M373I (blue). (B) The average number of contacts of residue 373 of human WT and M373I Kv4.3 TMDs with the PD region defined by residues 350–372. (C, D) Side views of WT and M373I SFs show the different interaction of M373 and I373 with Y360 and SF residues. SF residues for WT and M373I are highlighted in red and blue, respectively. Residues 360–366 are shown in white. M373 and I373 are shown in orange and sky-blue CPK representations, respectively. Residues in contact with M373 and I373 are shown in licorice representations. The cutoff for considering a residue in contact with residue 373 was 3.0 Å. The analysis was performed over the whole trajectory.

average structure with respect to the WT were noticed, as shown by the alignment of the equilibrated structures of both systems (Figure 6.3A, inset). Moreover, the average pore size measured along the channel longitudinal axis also showed no significant differences; both channels displayed a pore radius of about 4 Å in the so-called cavity region of the TMD, which immediately precedes the narrow SF stretch where the pore radius drops to 1 Å (Figure 6.3A). Despite such similarities in global protein structure, the estimated ion conductance in MD simulations, evaluated from the observed K^+ permeation events, showed a noticeable decrease in M373I, about 40% less (from 45 to 24). Among others, contacts with SF residues V363 and G369 were negligible in WT but noticeable in M373I (Figure 6.3B), suggesting a somewhat higher interaction of mutant isoleucine with the SF region with respect to the wild-type methionine, as depicted in Figure 6.3C and 6.3D, a result due to the increased hydrophobicity of the former residue. In turn, residue 373 was less exposed to the solvent in the mutant than in the WT channel, as shown by the large reduction of solvent accessible surface area (WT: 303 ± 49 $Å^2$; M373I: 171 ± 43 $Å^2$) and

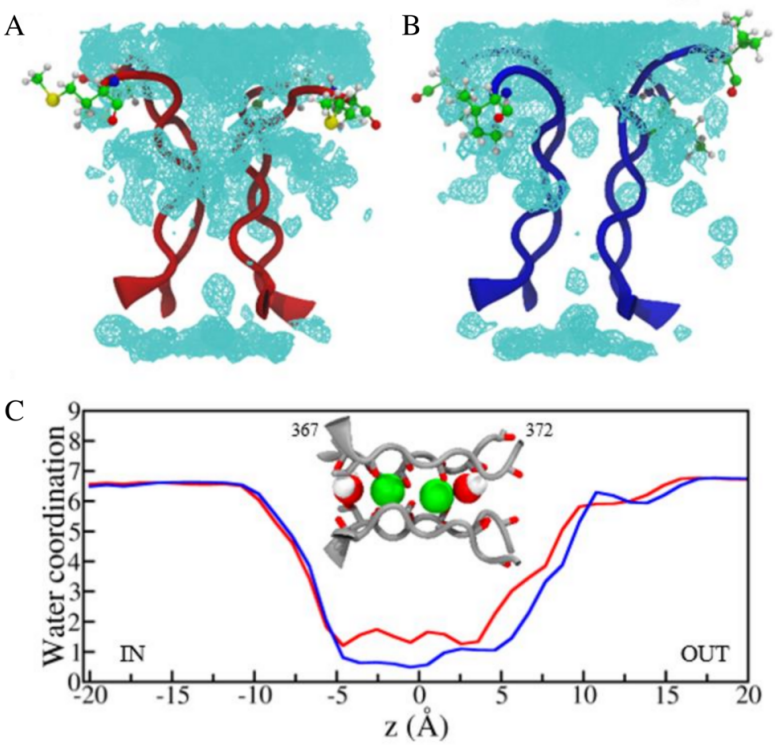

Figure 6.4: (A, B) Water density volumetric maps of human WT (red) and M373I (blue) Kv4.3 SFs. Residue 373 is shown with a CPK representation in both models. The water density of the WT SF, which is represented showing only one hydration layer, is much larger than that of M373I. The volumetric maps were measured over the whole trajectory. (C) Water coordination number of K^+ ions of human WT and M373I Kv4.3 TMDs. The SF of WT is shown in silver with highlighted carbonyl groups and T367 side chains in licorice representation. Water molecules and permeating K^+ ions (green) are shown in space filling representation. The analysis was performed over the whole trajectory.

Table 6.1: Gating parameters for steady state activation protocol on WT, M373I and S390N in absence of KChIP. Parameters extracted from fitting the steady state activation curve through a Boltzmann function with $V_{1/2}$ and k_{act} as fitting parameters.

Protocol	Parameters	WT	M373I	S390N
Activation	$V_{1/2}^{exp}$ (mV)	15.3 ± 1.8	18.6 ± 2.0	10.0 ± 2.1
	$V_{1/2}^{sim}$ (mV)	15.4 ± 0.1	18.8 ± 1.0	10.5 ± 1.1
	k_{act}^{exp} (mV)	16.5 ± 1.3	16.6 ± 1.3	19.4 ± 1.5
	k_{act}^{sim} (mV)	16.4 ± 0.9	15.7 ± 0.9	19.5 ± 1.0

the average number of water molecules in contact with residue 373 (WT: 11 ± 1; M373I: 9 ± 1).
However, the most remarkable effect of the M373I mutation was observed in the change of water hydration within the SF region. From the analysis of the local water density in the SF, the relevant dehydration effect caused by M373I mutation in this sensible portion of the channel was apparent (Figure 6.4A and 6.4B). Remarkably, this effect was significant not only when considering the average hydration level of the SF throughout the whole MD simulation but also when we specifically considered ion solvation during K^+ translocation events, as displayed by the decrease in water coordination number along the channel axis whenever a potassium ion approached the SF region (Figure 6.4C). The loss of coordinating water molecules appeared more significant towards the exit of the SF (about 2 water molecules), where K^+ becomes again fully hydrated and is released into the extracellular environment, thus indicating a less favorable ion translocation pathway in M373I. Irrespective of the specific details on the conduction mechanism and the role of water during K^+ translocation within the SF region, which is still a matter of debate in the literature, our findings indicate, overall, that non-negligible local effects take place upon M373I mutation, which in turn could affect channel electrophysiology as observed experimentally.
In the case of S390N, on the other hand, the functional studies revealed that the open probability of this mutant is very low, but once it is opened, it also has lower conductance compared to the WT. In the MD simulation, this channel variant turned rapidly into a closed state, thus preventing any ionic current. Consequently, no ion conductance was estimated in this case. Nevertheless, further analyses of the S390N model provided valuable insight into the origin of Kv4.3 channelopathy caused by this single mutation. To discover any further functional effects of each mutation, we systematically studied channel gating in vitro at each step of channel gating from closed-inactive to open forms and with the help of the related MSM.

6.3.3 Steady-state activation

The effects of the mutations on the voltage-dependence and kinetics of the steady-state activation were investigated by activating the channels through a voltage protocol (see Introduction section on voltage clamping) made of a series of 50 ms depolarizing test potentials from -90 mV holding potential to +60 mV (in 10 mV steps). The activation curves were plotted as normalized mean maximum currents versus test potentials and fitted with a Boltzmann function (Figure 6.5 and Table 6.1). When expressed alone, the WT channel had a half-activation potential ($V_{1/2}$) of 15.3 ± 1.8 mV and the voltage dependence, i.e. the slope factor k_{act}, of 16.5 ± 1.3 mV. (Figure 6.5 and Table 6.1). The M373I mutation did not affect too much the voltage-dependence of activation and had $V_{1/2}$ of 18.6 ± 2.0 mV and

Table 6.2: Gating parameters for steady state activation protocol on WT, M373I and S390N in presence of KChIP. Parameters extracted from fitting the steady state activation curve through a Boltzmann function with $V_{1/2}$ and k_{act} as fitting parameters.

Protocol	Parameters	WT-KChIP	M373I-KChIP	S390N-KChIP
Activation	$V_{1/2}^{exp}$ (mV)	9.6 ± 2.5	11.0 ± 2.1	13.9 ± 2.2
	$V_{1/2}^{sim}$ (mV)	9.6 ± 1.1	11.0 ± 1.0	14.0 ± 0.9
	k_{act}^{exp} (mV)	18.7 ± 2.0	17.6 ± 1.7	17.9 ± 1.4
	k_{act}^{sim} (mV)	18.6 ± 1.0	17.5 ± 0.9	17.6 ± 0.8

k_{act} of 16.6 ± 1.3 mV. However, while M373I was similar to WT the S390N had lower $V_{1/2}$ of 10.0 ± 2.1 mV and k_{act} significantly higher (19.4 ± 1.5 mV, p= 0.004), showing that SS-activation of S390N channels were relatively less responsive to voltage changes. When co-expressed with KChIP, the half-activation voltages of WT and M373I channels revealed a similar significant hyperpolarizing shift as compared to WT and M373I expressed alone (Figure 6.5A-D) with $V_{1/2}$ of 9.6 ± 2.5 mV (p=0.017) and 11.0 ± 2.1 mV (p=0.0012), respectively. The only difference of M373I-KChIP from the WT-KChIP was that, while KChIP increased the k_{act} for WT, it did not affect that of M373I, hence there appeared no significant differences for M373I relative to WT with k_{act} of 18.7 ± 1.7 mV (WT) versus 17.6 ± 2.0 mV (p<0.05) (M373I), showing slightly higher voltage dependence of M373I (Table 6.2).

For S390N, unlike the WT, KChIP could not alter $V_{1/2}$ of activation. The $V_{1/2}$ in the absence and presence of KChIP were (Figure 6.5E-F and Table 6.1 and Table 6.2) 10.0 ± 2.1 mV and 13.9 ± 2.2 mV, respectively, with no statistically significant difference. However, as compared to WT-KChIP, $V_{1/2}$ of S390N-KChIP was significantly different, with a $V_{1/2}$ of 13.9 ± 2.2 mV for S390N-KChIP versus 9.6 ± 2.5 mV for WT-KChIP (p=0.0003), as only the WT activation was modulated by KChIP. Unlike WT, KChIP also reduced the slope factor for S390N, hence, S390N became similar to WT on its k_{act}, i.e., 17.9 ± 1.4 mV versus 18.7 ± 2.0 mV, respectively.

The overall results show that the steady-state activation of WT current is modulated by KChIP towards earlier activation (hyperpolarized shift in $V_{1/2}$) with a lower response to voltage changes (higher k_{act}). The main effect of M373I mutation on the steady-state activation is that KChIP could not increase the channel response to voltage changes (the same k_{act} as in the absence of KChIP), as it did to WT. S390N mutation, on the other hand, increased the channel's response to voltage changes when expressed alone and had different modulation by KChIP. That is, in the presence of KChIP, the channel activated faster but at later stages of depolarization (depolarized shift in $V_{1/2}$), opposite of the WT channels, and the response of the S390N channels to voltage changes did not increase.

6.3.4 Steady-state inactivation

The effect of the mutations on the voltage-dependence of inactivation was elucidated by the conventional double pulse protocol, in which 1 s depolarizing pre-pulse potentials from -90 mV holding potential to +60 mV (in 10 mV steps) inactivates the channels and the following constant test pulse at +50 mV for another 1 s activates the non-inactivated channels. When expressed alone, the WT Kv4.3 demonstrated a mean $V_{1/2}$ of inactivation and a k_{inact} of -56.6 ± 0.3 and 8.2 ± 0.2 mV, respectively. The $V_{1/2}$ of inactivation and a k_{inact} of M373I were also like the WT, i.e. -54.6 ± 0.7 and 9.2 ± 0.6 mV (Figure 6.6 and

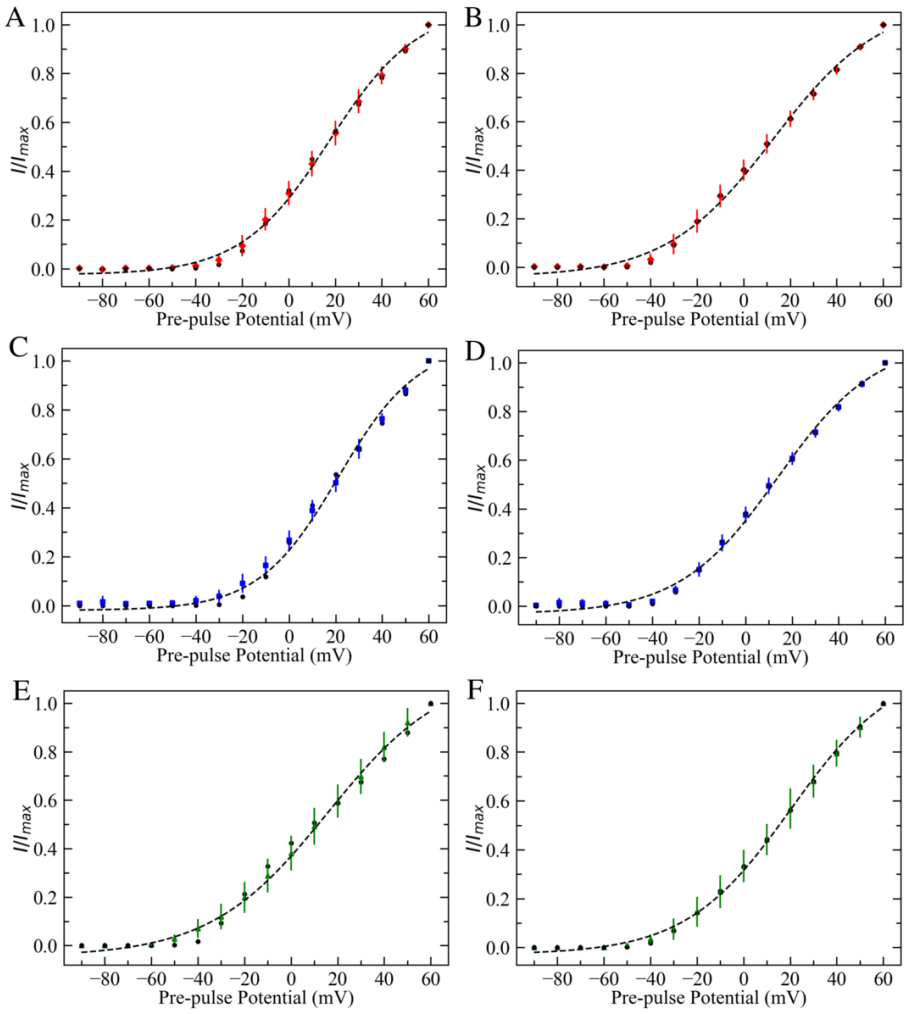

Figure 6.5: Steady state activation curves for: WT in absence of KChIP (A) and with KChIP (B); M373I in absence of KChIP (C) and with KChIP (D); S390N in absence of KChIP (E) and with KChIP (F). Experimental points are represented as scatter dotted plot. Simulated experiments are represented as black dots and black dashed curves.

Table 6.3: Gating parameters for steady state inactivation protocol on WT, M373I and S390N in absence of KChIP. Parameters extracted from fitting the steady state inactivation curve through an inverse Boltzmann function with $V_{1/2}$ and k_{inact} as fitting parameters.

Protocol	Parameters	WT	M373I	S390N
Inactivation	$V_{1/2}^{exp}$ (mV)	-56.6 ± 0.3	-54.6 ± 0.7	-66.1 ± 0.4
	$V_{1/2}^{sim}$ (mV)	-56.9 ± 0.2	-53.0 ± 0.3	-66.6 ± 0.4
	k_{inact}^{exp} (mV)	8.2 ± 0.2	9.2 ± 0.6	6.0 ± 0.4
	k_{inact}^{sim} (mV)	7.9 ± 0.2	7.9 ± 0.2	5.5 ± 0.4

Table 6.4: Gating parameters for steady state inactivation protocol on WT, M373I and S390N in presence of KChIP. Parameters extracted from fitting the steady state inactivation curve through an inverse Boltzmann function with $V_{1/2}$ and k_{inact} as fitting parameters.

Protocol	Parameters	WT-KChIP	M373I-KChIP	S390N-KChIP
Inactivation	$V_{1/2}^{exp}$ (mV)	-59.1 ± 0.1	-56.3 ± 0.2	-57.9 ± 0.2
	$V_{1/2}^{sim}$ (mV)	-59.1 ± 0.1	-56.3 ± 0.2	-59.0 ± 0.2
	k_{inact}^{exp} (mV)	4.4 ± 0.1	5.0 ± 0.1	6.5 ± 0.1
	k_{inact}^{sim} (mV)	4.1 ± 0.1	4.2 ± 0.2	5.1 ± 0.2

Table 6.5: Gating parameters for closed-state inactivation protocol on WT, M373I and S390N in absence of KChIP. Parameters extracted from fitting the closed-state inactivation curve through an exponential function with τ_{csi} as fitted parameter.

Protocol	Parameters	WT	M373I	S390N
CS-Inact	τ_{csi}^{exp} (ms)	385 ± 23	489 ± 27	120 ± 3
	τ_{csi}^{sim} (ms)	398 ± 16	567 ± 45	118 ± 13

Table 6.3). However, the mutant channels inactivated significantly faster than the WT at all tested voltages (Figure 6.6A,C,E). The S390N mutation, on the other hand, affected all steady-state inactivation properties of the channel. $V_{1/2}$ of inactivation exhibited an 10 mV significant hyperpolarizing shift to -66.1 ± 0.4 mV (p=0.043), k_{inact} decreased to 6.0 ± 0.4 mV (p=0.042), and channels inactivated significantly faster than the WT at all analyzed potentials (Figure 6.6).

When co-expressed, KChIP did not affect the $V_{1/2}$ of inactivation of the WT and the M373I mutant. In contrast, it shifted the voltage-dependence of S390N 8 mV to the more depolarized voltages to -57.9 ± 0.2 mV (p=0.027) (Figure 6.6F and Table 6.4), thereby correcting the hyperpolarizing shift in the $V_{1/2}$ induced by the mutation on the channel alone (Tables 6.3 and 6.4). KChIP also significantly decreased the k_{inact} of WT and M373I (Table 6.4) as compared to its absence. The k_{inact} became 4.4 ± 0.1 mV for WT (p<0.0001), 5.0 ± 0.1 mV M373I (p=0.0024) for M373I. Leaving the S390N unaltered k_{inact} and 6.5 ± 0.1 mV for S390N (p=0.019). The overall results show that KChIP does not change $V_{1/2}$ inactivation but enhances the sensitivity of the WT channels to voltage changes (smaller k_{inact}). KChIP modulates M373I mutants as it does WT channels and compensates for the effect of the mutant's influence on inactivation kinetics of the channel. However, as in the activation, the steady-state inactivation of S390N is modulated differently by KChIP; $V_{1/2}$ of inactivation shifted toward WT $V_{1/2}$ range. However, KChIP-modulated channels still inactivates faster than the WT.

6.3.5 Closed-state inactivation

Upon reaching the closed-activatable state, native Kv4.3 channels preferentially turn into a closed-inactive rather than an open state [388, 389]. A double-pulse protocol was used to test the effect of mutations on the closed-state inactivation (CSI). The Kv4.3 channels were inactivated from the closed-active (CA) state with a pre-pulse at -50 mV for a variable interval from 10 to 550 ms. A following test pulse at +50 mV for 1 s reported the current originating from the fraction of channels that did not enter the closed-inactive state during

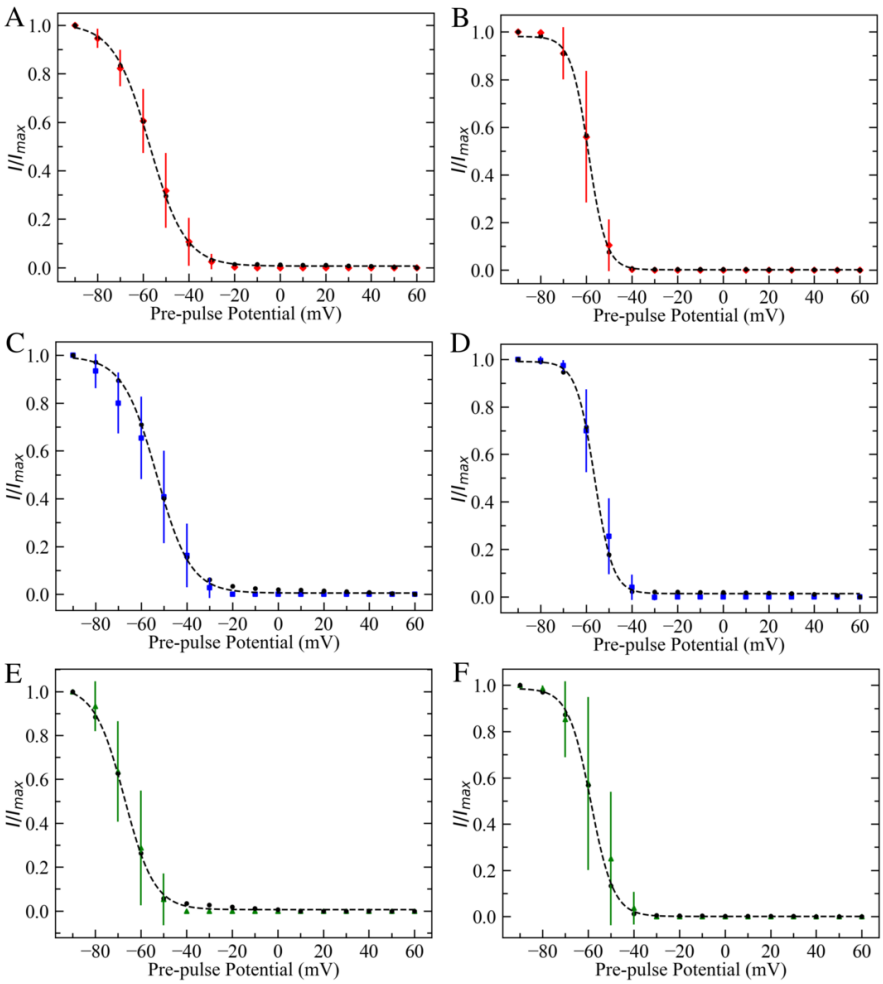

Figure 6.6: Steady state inactivation experiments for: WT in absence of KChIP (A) and with KChIP (B); M373I in absence of KChIP (C) and with KChIP (D); S390N in absence of KChIP (E) and with KChIP (F). Experimental points are represented as scatter dotted plot. Simulated experiments are represented as black dots and black dashed curves.

Table 6.6: Gating parameters for closed-state inactivation protocol on WT, M373I and S390N in presence of KChIP. Parameters extracted from fitting the closed-state inactivation curve through an exponential function with τ_{csi} as fitted parameter.

Protocol	Parameters	WT-KChIP	M373I-KChIP	S390N-KChIP
CS-Inact	τ_{csi}^{exp} (ms)	157 ± 4	331 ± 6	152 ± 5
	τ_{csi}^{sim} (ms)	163 ± 6	336 ± 9	207 ± 23

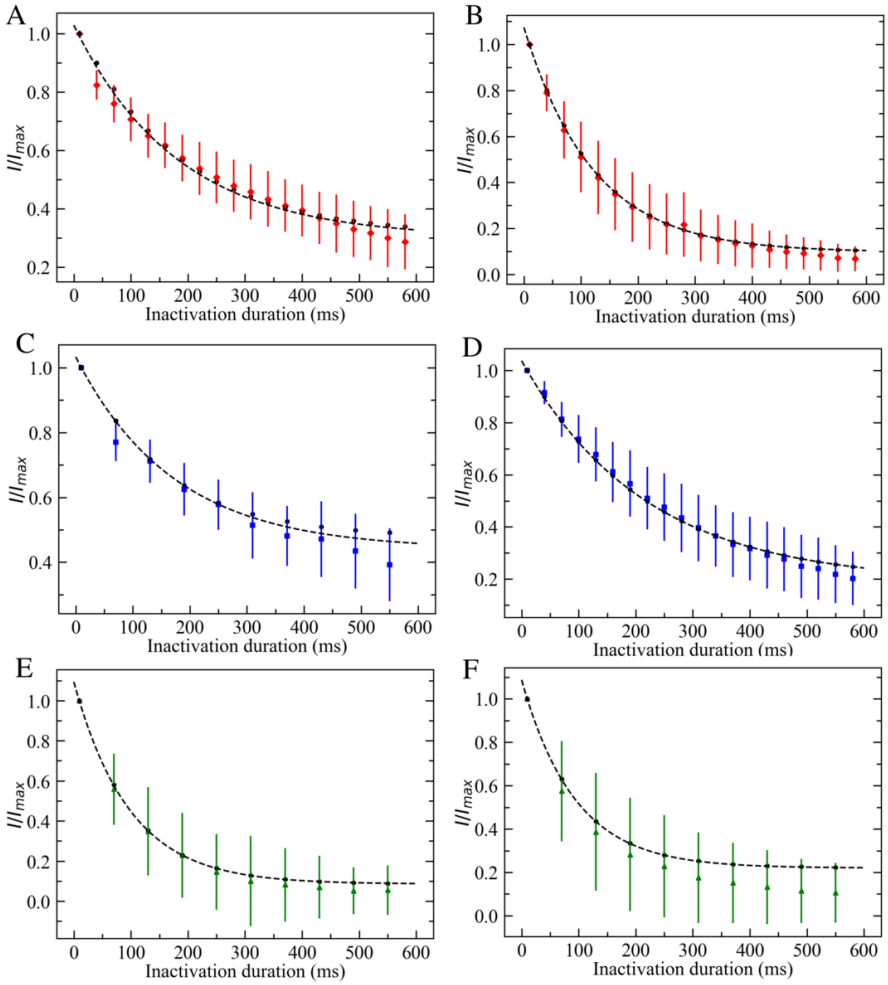

Figure 6.7: Steady state closed-state inactivation experiments for: WT in absence of KChIP (A) and with KChIP (B); M373I in absence of KChIP (C) and with KChIP (D); S390N in absence of KChIP (E) and with KChIP (F). Experimental points are represented as scatter dotted plot. Simulated experiments are represented as black dots and black dashed curves.

the pre-pulse. The mean closed-state inactivation curves and the inactivation time constants were derived by fitting single exponential functions of normalized pick currents after each inactivation duration (Figure 6.7). As shown in (Figure 6.7 and Table 6.5, when expressed alone, WT channels entered CSI with a τ_{csi} of 385 ± 23 ms. KChIP modulation significantly speeded up the inactivation (τ_{csi} 157 ± 4 ms, $p<0.05$). KChIP (Table 6.6), significantly speeded up the CSI of M373I channels (τ_{csi} 489 ± 27 ms) to a τ_{csi} of 331 ± 6 ms ($p<0.05$). The effect of mutation on the channel CSI was even more dramatic for S390N. When alone, the channels were closing very fast (τ_{csi} 120 ± 3 ms, $p<0.0005$). However, unlike in the case of WT, KChIP modulation slowed them down and eventually they behaved like WT with a τ_{csi} of 152 ± 5 ms.

6.3.6 Recovery from inactivation

The effect of M373I and S390N on the time course of recovery from inactivation was evaluated using a double pulse protocol at -90 mV. The channels were activated and completely inactivated with a first pulse at +50 mV for 1 s. Before the second pulse at +50 mV was

Table 6.7: Gating parameters for recovery from inactivation protocol on WT, M373I and S390N in absence of KChIP. Parameters extracted from fitting the recovery from inactivation curve through an exponential function with τ_{rec} as fitted parameter.

Protocol	Parameters	WT	M373I	S390N
Recovery	τ_{rec}^{exp} (ms)	213 ± 26	280 ± 33	222 ± 16
	τ_{rec}^{sim} (ms)	214 ± 12	286 ± 25	225 ± 15

Table 6.8: Gating parameters for recovery from inactivation protocol on WT, M373I and S390N in presence of KChIP. Parameters extracted from fitting the recovery from inactivation through an exponential function with τ_{rec} as fitted parameter.

Protocol	Parameters	WT-KChIP	M373I-KChIP	S390N-KChIP
Recovery	τ_{rec}^{exp} (ms)	83 ± 3	90 ± 8	48 ± 1
	τ_{rec}^{sim} (ms)	83.3 ± 0.3	91 ± 2	47 ± 1

applied, the channels were exposed to -90 mV of interpulse potential for a variable interval from 0 to 540 ms for channel recovery. At the end of each interval, a second pulse at +50 mV was applied, reporting the channels recovered during the interpulse interval, i.e., the fraction of channels that exit the closed-inactive state. The mean recovery from inactivation curves and recovery time constants were derived by fitting single exponential functions.
When expressed alone, WT channels recovered from inactivation slowly with a τ_{rec} of 213 ± 26 ms (Figure 6.8 and Table 6.7). KChIP significantly speeded it up (τ_{rec} 83 ± 3, p=0.0003). M373I mutation affected the recovery of the channel alone from inactivation (τ_{rec} 280 ± 33 ms), when co-expressed with KChIP the recovery was accelerated thus, the mutant channel recovery was very similar to the WT (τ_{rec} 90 ± 8, p<0.05) under the same conditions. Surprisingly, the severe mutant S390N was similar to the WT when expressed alone with a τ_{rec} of 222 ± 16 ms. However, the KChIP modulation of the recovery of S390N channels was very different with respect to the WT (τ_{rec} 48 ± 1 ms) (Figure 6.8 Tables 6.7 and 6.8).

6.3.7 Kinetic interpretation of the Kv4.3 channel and its mutants

The MSMs developed for the Kv4.3 channel wild type (WT) and its two mutants, both in the absence and presence of KChIP, successfully replicated experimental steady-state curves (Figures 6.5, 6.6, 6.7, 6.8). The experimental gating parameters, particularly the half-activation potential ($V_{1/2}^{act}$) and the slope factor of activation (k^{act}), were astonishingly well reproduced, aligning within experimental errors for the WT, M373I, and S390N, both with and without KChIP (Tables 6.1 and 6.2). For the inactivation gating parameters, all half-inactivation potentials ($V_{1/2}^{inact}$) were accurately predicted by the MSMs. The slope factor of inactivation (k^{inact}) for the WT, in both the absence and presence of KChIP, fell within experimental errors. However, for the mutants M373I and S390N, under all conditions, there was a consistent observation that the MSMs slightly anticipated the voltage dependence of inactivation (k^{inact}) by approximately 10% (Tables 6.3 and 6.4). Regarding closed-state inactivation times, the MSMs accurately replicated these times within experimental error for the WT, both with and without KChIP, S390N without KChIP, and M373I with KChIP. For M373I alone and S390N with KChIP, MSMs predicted τ^{csi} just

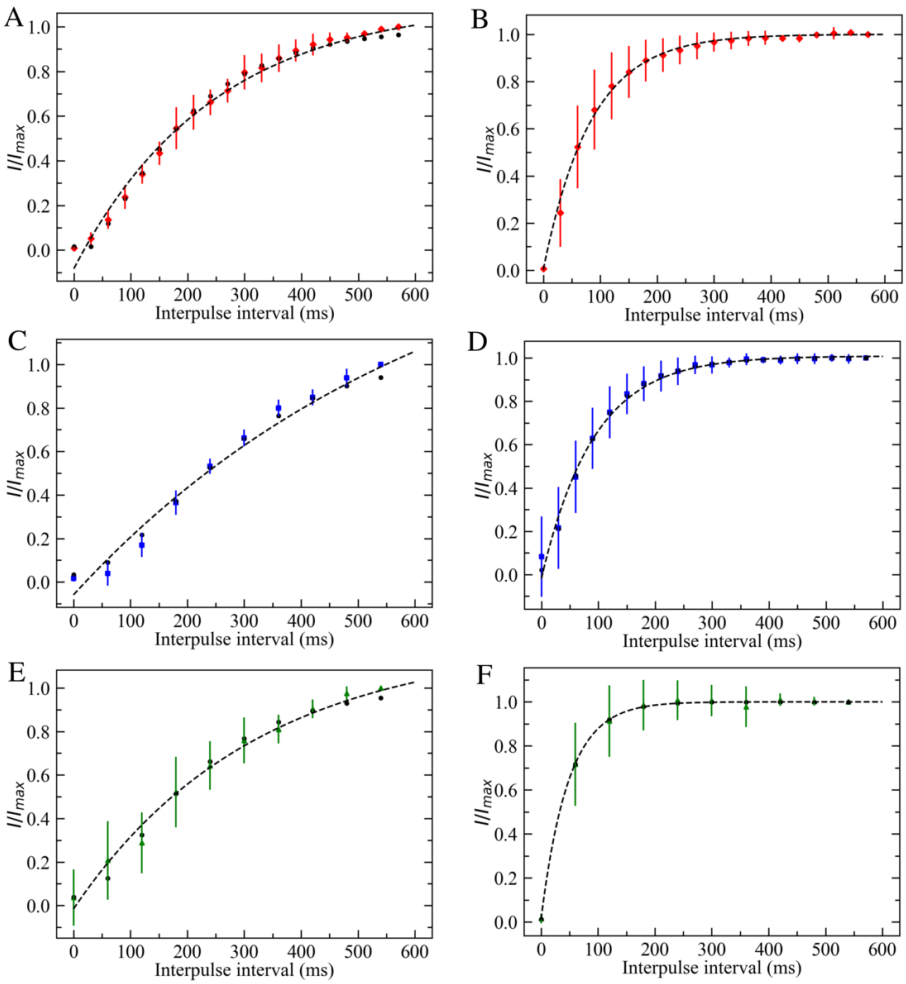

Figure 6.8: Steady state recovery from inactivation experiments for: WT in absence of KChIP (A) and with KChIP (B); M373I in absence of KChIP (C) and with KChIP (D); S390N in absence of KChIP (E) and with KChIP (F). Experimental points are represented as scatter dotted plot. Simulated experiments are represented as black dots and black dashed curves.

outside the experimental error (Tables 6.5 and 6.6), yet the simulated closed-state inactivation curves matched well within experimental error bars for these systems (Figures 6.7C and 6.7F). This discrepancy might be attributed to those experimental curves having fewer data points for fitting, thus receiving a lower "relative weight" in the MSMs (Equation 6.2), which consequently favored alignment with the other three experimental protocols. For the recovery from inactivation times (τ^{rec}), the MSMs accurately reproduced these times within the experimental standard deviation for WT, M373I, and S390N, both in the absence and presence of KChIP (Tables 6.7, 6.8).
Additional insights can be extracted from MSMs after their construction based on experimental data. One approach involves analyzing the channel populations across the MSM states (Figure 6.2). During the activation protocol, the populations of the simulated channels in the presence of KChIP were recorded at +60 mV (Figure D.4) and -10 mV (Figure D.5), revealing that the M373I mutation exhibits less inertia compared to the WT (showing ease in transitioning from the inactive state at -10 mV to the open state at +60 mV). Conversely, the S390N mutant demonstrated greater inertia compared to the WT (with fewer channels transitioning from the inactive state at -10 mV to the open state at +60 mV). Further analysis on channel populations during closed-state inactivation protocols without KChIP (Figure 6.9) indicated that while the WT and M373I had approximately 30% of their channels in closed states before transitioning to inactive states, the S390N mutation appeared more severe, with less than 10% of channels in closed states before moving to inactive states, possibly explaining the experimentally suggested tendencies of this point mutation to inactivate more rapidly.
Additionally, rate constants from the kinetic scheme (Tables D.1 and D.2) can be utilized to interpret experimental findings. The ratio k_{CO_0}/k_{OC_0} estimates the predisposition of channels to transition to the open state from a closed active state at 0 V. When co-expressed with KChIP, this ratio for WT and M373I is ≈ 1.4 and ≈ 2.0, respectively, while for S390N, the ratio dramatically decreases to ≈ 0.05, further emphasizing the mutation's severity in hindering the channel's transition to an open state.

6.4 Conclusions

Taken together, the experimental results showed that, the main functional effects of the mild M373I mutation on the channel alone were slightly lower single-channel conductance, faster steady-state inactivation, and slower recovery from inactivation. Co-expression with the main auxiliary partner, KChIP, enhanced their surface expression, despite still lower than the WT. KChIP modulation corrected the channels' steady-state inactivation. It also significantly accelerated the recovery from inactivation, but the mutant channels were still significantly slower than the WT. However, KChIP could not modulate its voltage-dependence of activation, causing significantly higher voltage dependence of activation (lower k_{act}) than WT. At the molecular level, the constructed MD model and full atomistic simulations showed that M373I mutation introduces a bulk and hydrophobic amino acid at the exit of the SF of the pore and as compared to the native methionine, isoleucine generates tighter packing and reduced mobility of the SF and it reduces the water hydration at the top of the SF, explaining the slightly lower conductance of single channels.
The functional effects of the S390N mutation were severe. The channel expression was dramatically affected; there were no detectable channel currents on the cell membrane under physiologically relevant growth conditions. When expressed at lower temperatures, active channels could be expressed, despite significantly lower densities. As compared to the WT channel, the mutation significantly lowered the voltage-dependence of steady-state activa-

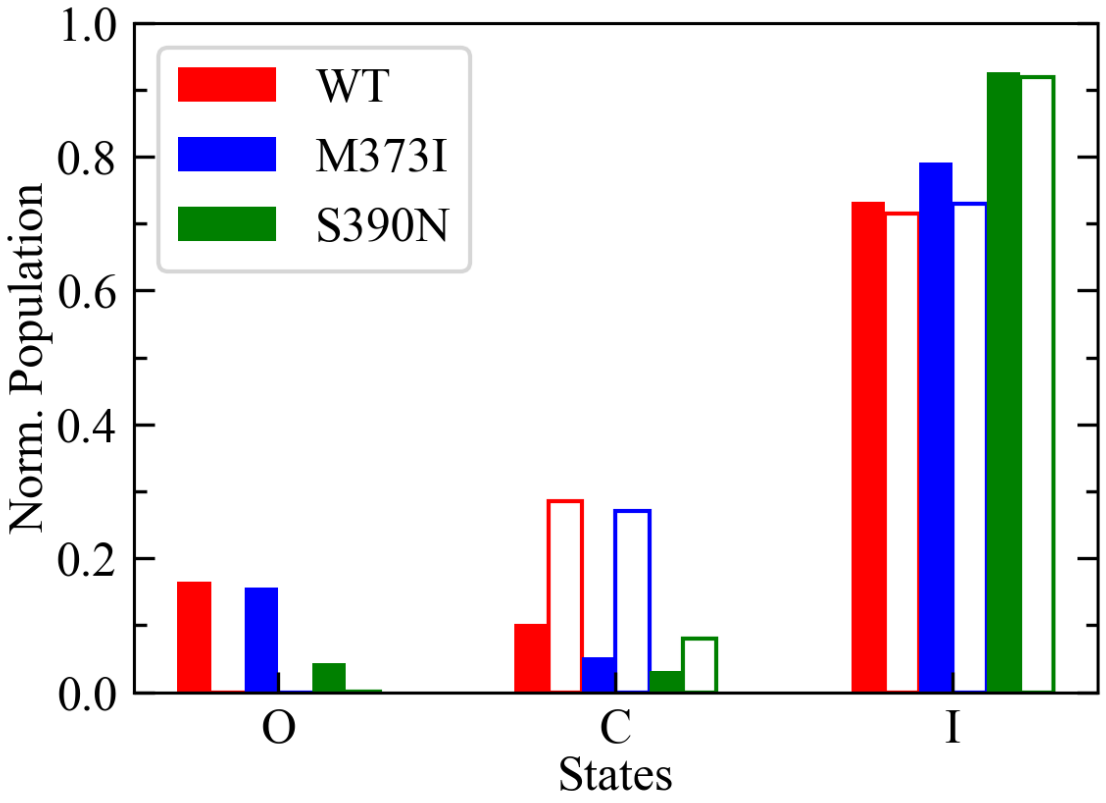

Figure 6.9: Population analysis of the kinetic model (Figure 6.2 and Table D.1) in absence of KChIP as obtained from simulation of WT (red), M373I (blue) and S390N (green) upon keeping the voltage at -50mV for 310 ms (unfilled bars) and successive peak activation with pulse at +60mV (filled bars). For clarity, the population of all closed states (C0 to CA) and inactive states (I0 to I4) are collectively summed up as C and I in the bar graph. The number of ion channels found in each state was normalized with respect to the total number of simulated channels.

tion (higher k_{act}), significantly shifted the steady-state inactivation toward less depolarized membrane potentials with higher voltage dependence (lower k_{inact}), and very significantly accelerated the entry into the closed-inactive state. Co-expression with KChIP improved the expression level with still a need for assisted, non-physiological growth conditions. Surprisingly, KChIP modulations were different in S390N than in WT. KChIP significantly shifted the $V_{1/2}$ of activation to more in place of less depolarized voltages. It corrected the steady-state inactivation by shifting the $V_{1/2}$ toward more instead of less depolarized voltages, and corrected the voltage-dependence by lower k_{inact} towards the WT value. KChIP also corrected the kinetics of entering into the closed-inactive state by increasing the τ_{csi} (Tables 6.5 and 6.6). Full atomistic simulations were unable to detect any channel opening in the S390N mutation. In this context, the MSMs used were notably successful in complementing patch clamp experiments to elucidate the severity of the S390N mutation. Specifically, they highlighted the mutation's significant difficulty in maintaining the closed-active state, which is essential for transitioning to the open state, and the pronounced inertia of S390N in moving from inactive to open states.

This study demonstrates how a combination of experimental and computational methodologies can effectively identify potentially detrimental mutations in the Kv4.3 channel. On one hand, MD simulations shed light on molecular reasons behind the M373I mutation's behavior, while on the other, MSMs provided insight into the S390N mutation's severity by illustrating its reduced tendency to adopt the closed-active state.

Chapter 7

Conclusions and Future Perspectives

In this dissertation, various computational approaches have been thoroughly discussed and presented, focusing on (i) ions in aqueous solutions, (ii) ion-ligand complexes, and (iii) ion transport in channels. The Introduction set forth the thesis's objective: to develop computational methodologies for the aforementioned topics that are as general as possible. Additionally, aiming to design these methodologies to be flexible for application to different scenarios when altering parts of the studied system and robust enough to explore timescales close to those observed in experiments.

For (i), in Chapter 2 a combined approach involving enhanced sampling MD simulations and stochastic equations was introduced to model the thermodynamics and kinetics of the microsolvation of aqua ions. This method was tested on various ions (e.g., Ca^{2+}, Zn^{2+}, Hg^{2+}, Cd^{2+}) with different solvation shells, solvation energies, and water exchange times. The computational protocol, when compared to extended unbiased MD simulations, demonstrated close agreement for events within standard MD timescales and provided greater statistical accuracy for rare events at significantly lower computational costs.

For the second point (ii), in Chapter 3, the use of enhanced sampling MD simulations, specifically umbrella sampling, was demonstrated for developing a reliable force field for metal ion-carboxylate interactions. The 12-6-4 Lennard-Jones (LJ) nonbonded model was fine-tuned against experimental binding energies of 11 metal ions with acetate, using three different water models (TIP3P, SPC/E, and OPC). Initially successful in describing metal ion-water interactions more accurately [54], this 12-6-4 nonbonded model has been extended to metal ion-imidazole [83] and, in this work, to metal ion-acetate interactions. Building an accurate force field for these interactions is crucial when comparing simulated properties with experimental data, as shown in the same chapter for the Escherichia coli Glyoxalase I metalloprotein. Improved parametrization of Ni^{2+}-carboxylate interactions led to better alignment with the crystal structure, demonstrating promising transferability to other metalloproteins, enzymes, ion transporters, and ion channels containing negatively charged residues.

Again regarding the second point (ii), in Chapter 4, an integrated approach combining enhanced sampling MD simulations and Markov State Models (MSMs), inspired by previous work in Chapter 2, was proposed to calculate stability constants and exchange rates of ion-ligand complexes. This methodology was tested with cadmium cations and amines of varying denticity (nme, en, dien, put) in water solutions. The stability constants of different cadmium complexes were systematically calculated for various metal ions and ligand concentrations, closely aligning with the experimental ones. Additionally, thermodynamic

quantities like enthalpy and entropy of ligand exchange were calculated, highlighting areas for force field improvement in accurately representing the experimentally well-known but computationally elusive chelate effect. By simultaneously employing coordination numbers of waters and ligands, the minimum free energy pathway was extracted, revealing the exchange mechanism (associative or dissociative) of the metal ion-ligand complex. Rates of formation and dissociation were also computed for more stable complexes involving Ni^{2+} cations (exchange times on the order of seconds). This procedure proved flexible and effective for complexes with various amines, cations, and concentrations of ions and ligands. The ability of MSMs to bridge the gap between experimental and computationally accessible timescales was particularly promising.

Concerning the third topic (iii), Chapter 5 discussed a software designed to measure the geometry of a pore. The proposed algorithm is capable of producing both the morphology of the pore and its centerline (geometric path). This software, tested on the MscL channel, is adaptable to any pore or channel structure. In the context of ion transport in channels, it holds the potential to identify hollow regions within the pore structure that may lead to fenestration events. Additionally, it allows for the comparison of the centerlines with the actual pathways followed by the ions, shedding light on whether ion translocation through the channel is influenced more by steric or electrostatic factors.

Still on the third point (iii), in Chapter 6, the focus was on examining potential changes in ion transport within a voltage-gated potassium channel. The study concentrated on the Kv4.3 channel, both in its wild-type form and two mutants, M373I and S390N. MD simulations revealed a significant impact of the more hydrophobic mutation at the selectivity filter (SF) in the M373I variant, leading to reduced conductance. Concurrently, a MSM of Kv4.3 elucidated the increased inertia of the mutants, particularly S390N, towards the inactive state. Additionally, MSMs of Kv4.3 corroborated experimental evidence of the mitigating effects that the auxiliary protein KChIP has on channel mutants. The capability to correlate channel dynamics with a MSM, and the flexibility of MSMs to emulate various experimental patch clamp protocols, aligns this computational approach closely with the overarching aim of the thesis.

In general terms, it can be stated that the objective set out in the Introduction, and reiterated at the beginning of this chapter, has been satisfactorily met. It was highlighted that the computational approaches presented were developed with the aim of being flexible and applicable to various contexts. On this note, it is indeed feasible to attempt extending these methodologies to explore systems of greater complexity than those previously presented.

A logical extension in relation to the first point (i) would be to examine the effects of varying electrolyte concentrations in aqueous solutions. In Chapter 2, a 0.5 M $HgCl_2$ solution was analyzed, indicating a potential competitive interaction among ions for coordinating water molecules. The emergence of these nonlinear effects, whether from the same cations or an increasing role of counterions, remains unclear. A viable approach could involve studying different cations at various concentrations to understand why ion pairings and ion clusters appear more favorable for certain ions compared to others. Such an investigation might uncover molecular mechanisms that play a role in ion translocation in synthetic channels [390].

Regarding the second point (ii), there are multiple possible extensions. As discussed in Chapter 4, one direction involves investigating complexes with biogenic amines, which are crucial for human health. The primary effort here would be to develop a force field model capable of accurately simulating amines with long alkyl chains. A specific focus on extending the work from Chapters 3 and 4 has recently commenced in our research group. NMR studies have shown that an intrinsically disordered protein (IDP) named α-synuclein,

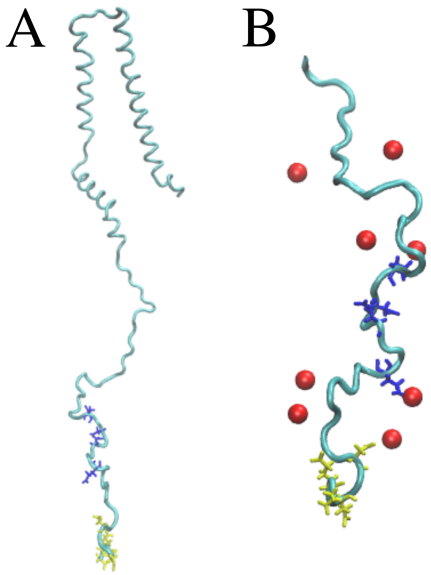

Figure 7.1: A) α-synuclein protein represented in cartoon with residues Asp119, Asp121, Glu123, and Glu126 in licorice blue and Asp135, Glu137, Glu139, and Ala140 in licorice yellow. B) Last section of α-synuclein (residues 100-140) with same highlighted residues and Ca^{2+} ions interacting with negatively charged residues.

which is implicated in Parkinson's neurodegenerative disease, is sensitive to jumps in Ca^{2+} concentration, particularly in disordered regions with a high presence of negatively charged residues [391]. The computational approach would utilize the accurate force field developed in Chapter 3 for calcium-carboxylate interactions, alongside the methodology proposed in Chapter 4, to investigate the atomic-level interactions occurring when calcium ions bind to and release from Glu and Asp residues in the disordered region of α-synuclein (refer to Figure 7.1).

Finally, regarding the third point (iii), the computational method based on MSMs presented in Chapter 6 can be applied to other voltage-gated ion channels, particularly for systematic investigations of channel mutants. The motivation behind developing a software with a Graphical User Interface (GUI) was to provide a user-friendly computational tool for electrophysiology experts who may not be familiar with MSMs and algorithms. The software aims to encompass as many experimental setups as possible. To achieve this, it would be necessary to incorporate a stochastic integration algorithm for the equations underlying the MSM, enabling fitting against data from single-channel recordings. Additionally, the experimental protocol builder in the software requires significant modifications to allow users to input protocols for raw current curves and kinetic data. Currently, it only facilitates setting up protocols that lead to steady-state curves in voltage clamping experiments. Experimental measurements of ions in water solutions, ion-ligand complexes, and ion transport in ion channels have been available for decades. With technological advancements, these measurements have become increasingly accurate and offer more detailed spatial and temporal resolution. However, discerning molecular mechanisms from these measurements remains challenging, often involving timescales difficult to achieve by theoretical models that aim to account for these mechanisms. In the context just described, this thesis is believed to serve as a valuable bridge between experimental observations and theoretical interpretations.

Appendices

Appendix A

Supporting Data for Chapter 2

Analysis of the normalized populations for discrete states

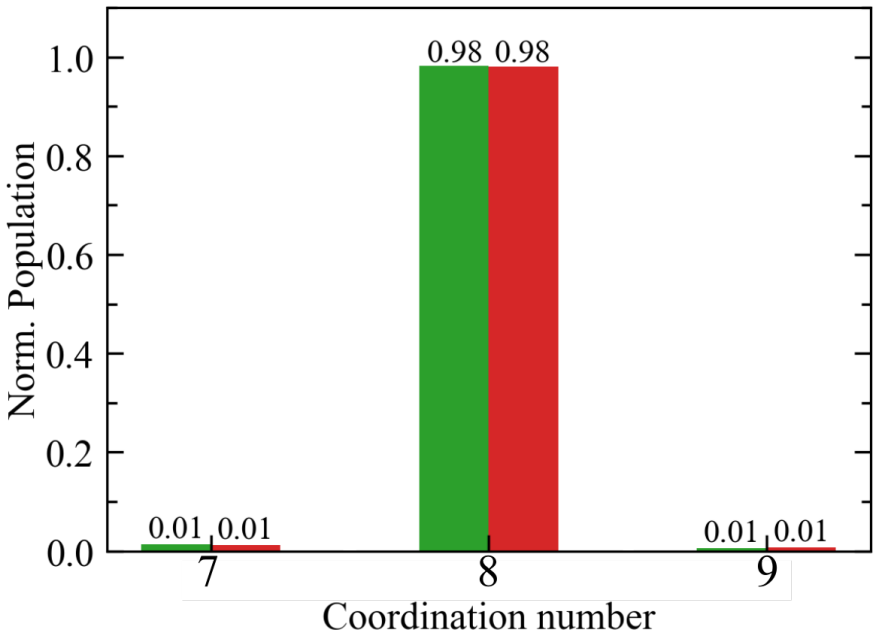

Figure A.1: Analysis of the normalized population for each of the three discrete coordination states of Ca^{2+}. Red bars, results computed from the history-based algorithm described in Sec. 2.2.3 Green bars, results obtained by direct assignment to a coordination state of each sampled MD configuration.

Hg^{2+} system with counter bias applied

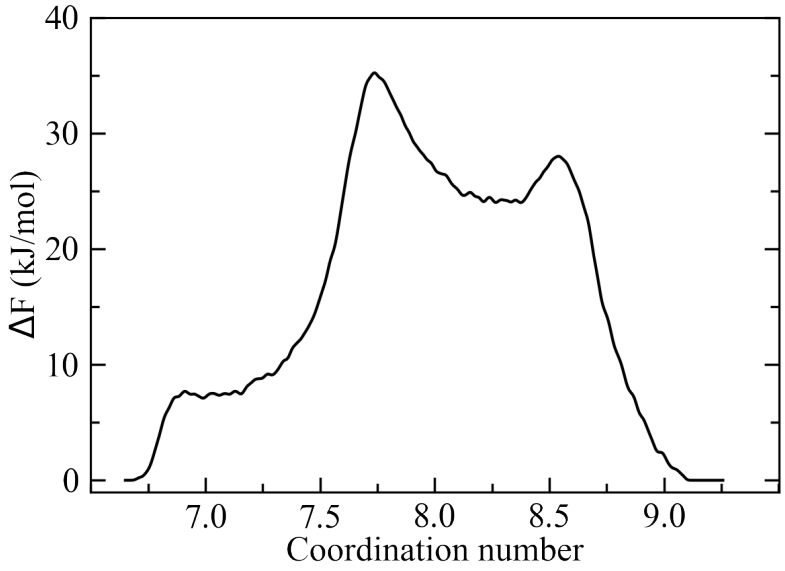

Figure A.2: Profile of the bias potential applied to the Hg^{2+} system in a test MD simulation to neutralize any free energy barrier along the water coordination variable, s.

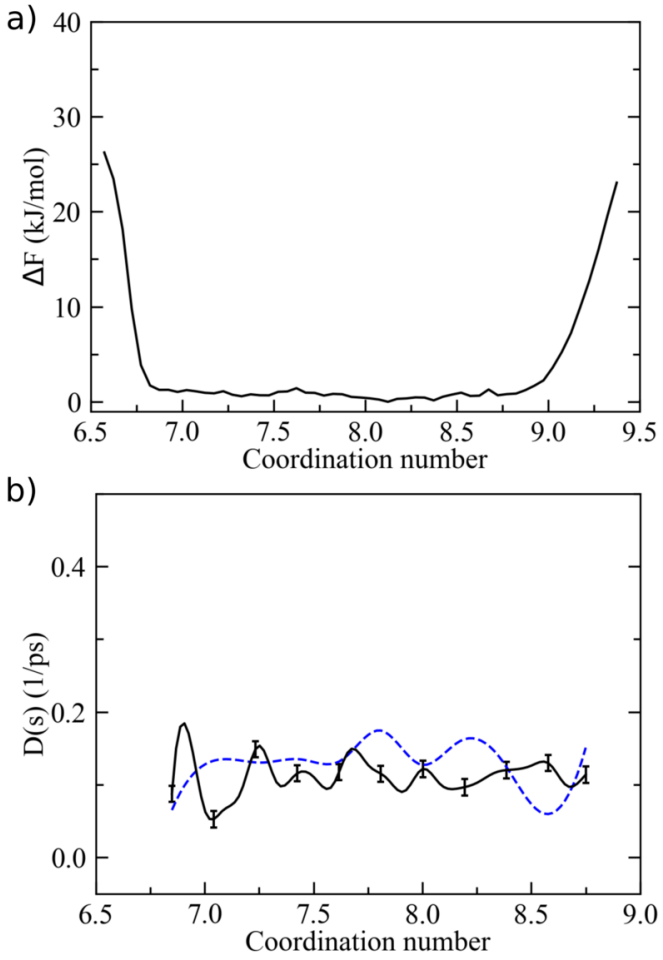

Figure A.3: a) Barrier-less free energy profile of Hg^{2+} coordination in water with an applied counteracting potential. b) Position-dependent diffusion coefficient as a function of the coordination number. Error bars correspond to $\pm\delta D$ (see Sec. 2.4). Blue dashed line is the diffusion computed from local mean squared displacements. In both cases, D oscillates slightly around the average value of 0.1 ps^{-1}.

Hg^{2+} high molarity free energy and MFPTs

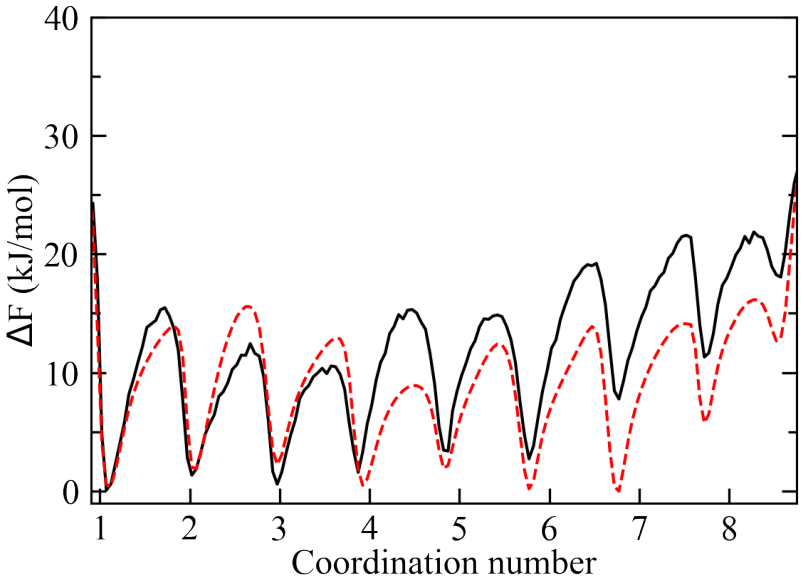

Figure A.4: Free energy landscape of Hg^{2+} coordination in water from a 0.5M $HgCl_2$ aqueous solution, as issuing from MetaD (black solid line) and pure MD (red dashed line) simulations. In this case, the observed deviations should be ascribed to a poorer statistics of the MetaD simulation, since only one Hg^{2+} ion (out of 20) was considered when computing the bias potential. This can be easily improved using a different implementation of the algorithm, but we preferred to keep the same protocol for consistency with the other MetaD simulations.

Table A.1: MFPT for 0.5M of Hg^{2+} computed from pure MD simulations and FP integration. In the latter case, MFPTs were estimated using $\Delta F(s)$ from both pure MD (FP) and from MetaD (FP*) (see Fig. A.4).

Transition	MD (ps)	FP (ps)	FP* (ps)
$1 \to 2$	45 ± 5	71 ± 30	92 ± 38
$2 \to 1$	17 ± 5	50 ± 30	69 ± 35
$2 \to 3$	66 ± 20	69 ± 6	36 ± 5
$3 \to 2$	44 ± 18	67 ± 7	40 ± 5
$3 \to 4$	19 ± 3	18 ± 4	18 ± 4
$4 \to 3$	21 ± 9	41 ± 8	9 ± 2
$4 \to 5$	12 ± 2	11 ± 3	49 ± 18
$5 \to 4$	11 ± 2	8 ± 2	32 ± 9
$5 \to 6$	10 ± 2	20 ± 5	30 ± 8
$6 \to 5$	28 ± 3	29 ± 7	33 ± 8
$6 \to 7$	19 ± 8	38 ± 10	147 ± 42
$7 \to 6$	35 ± 15	36 ± 8	21 ± 6
$7 \to 8$	92 ± 30	71 ± 9	58 ± 14
$8 \to 7$	2.9 ± 0.4	9 ± 1	15 ± 3
$8 \to 9$	18 ± 4	32 ± 11	33 ± 8
$9 \to 8$	1.8 ± 0.4	3.3 ± 0.8	4.0 ± 1.2

Appendix B

Supporting Data for Chapter 3

Table B.1: The calculated binding free energy of each metal ion – carboxylate complex in TIP3P water. Average column show the results from three replicas performed using AMBER US technique. The free energy values are in kcal/mol.

Ion	Exp. ΔG	Average*	Standard 12-6-4	12-6	PLUMED†
Ni(II)	-1.95	-1.92±0.2	-1.19	-8.15	-1.89
Mg(II)	-1.33	-1.25±0.07	-3.91	-8.74	-1.42
Cu(II)	-3.01	-3.22±0.03	0.24	-7.75	-3.17
Zn(II)	-2.16	-2.32±0.38	-0.75	-8.31	-2.24
Co(II)	-1.88	-1.90±0.08	-1.15	-7.86	-1.74
Cu(I)	-2.78	-2.79±0.16	-1.86	-2.65	-2.67
Fe(II)	-1.91	-1.91±0.14	-2.87	-7.14	-1.94
Mn(II)	-1.91	-1.86±0.23	-2.57	-7.35	-1.96
Cd(II)	-2.63	-2.65±0.12	0.01	-5.88	-2.79
Ca(II)	-1.27	-1.19±0.16	-1.22	-3.48	-1.16
Ag(I)	-1.00	-1.02±0.04	1.08	-0.13	-1.00

* The error is the std deviation calculated over 3 replicas of US performed in AMBER software

† Estimated binding free energy using US coded in PLUMED software.

Table B.2: Polarization α_0(Pol.) applied to equation 3.1 for each metal ion and related computed C_4 values used to reach experimental binding energies of ion-carboxylate complex for TIP3P water model.

Ion	Pol (Å^3)	C_4 kcal/(mol Å^4)
Ni(II)	0.644	94.91
Mg(II)	0.125	11.50
Cu(II)	0.944	190.17
Zn(II)	0.730	117.08
Co(II)	0.657	95.41
Cu(I)	1.700	8.24
Fe(II)	0.432	48.76
Mn(II)	0.469	47.45
Cd(II)	0.894	114.91
Ca(II)	0.600	36.27
Ag(I)	1.500	86.22

Table B.3: The calculated binding free energy of each metal ion – carboxylate complex in SPC/E water. Average column show the results from three replicas performed using AMBER US technique. The free energy values are in kcal/mol.

Ion	Exp. ΔG	Average*	Standard 12-6-4	12-6	PLUMED†
Ni(II)	-1.95	-2.12±0.26	-5.12	-7.5	-1.97
Mg(II)	-1.33	-1.49±0.24	-5.71	-5.39	-1.10
Cu(II)	-3.01	-3.11±0.25	0.11	-6.14	-3.10
Zn(II)	-2.16	-2.04±0.30	-0.2	-6.41	-1.64
Co(II)	-1.88	-1.99±0.09	-0.31	-6.98	-2.4
Cu(I)	-2.78	-3.05±0.04	-1.28	-1.96	-2.7
Fe(II)	-1.91	-1.95±0.21	-1.67	-6.51	-2.21
Mn(II)	-1.91	-2.10±0.19	-1.23	-5.37	-2.00
Cd(II)	-2.63	-2.68±0.31	0.44	-4.15	-2.59
Ca(II)	-1.27	-1.42±0.12	-0.61	-2.55	-1.5
Ag(I)	-1.00	-1.12±0.16	0.74	0.41	-1.08

* The error is the std deviation calculated over 3 replicas of US performed in AMBER software

† Estimated binding free energy using US coded in PLUMED software.

Table B.4: Polarization α_0(Pol.) applied to equation 3.1 for each metal ion and related computed C_4 values used to reach experimental binding energies of ion-carboxylate complex for SPC/E water model.

Ion	Pol (Å^3)	C_4 kcal/(mol Å^4)
Ni(II)	0.450	63.95
Mg(II)	0.125	10.58
Cu(II)	1.069	225.35
Zn(II)	0.869	139.14
Co(II)	0.819	118.65
Cu(I)	2.500	15.58
Fe(II)	0.657	201.10
Mn(II)	0.694	74.45
Cd(II)	1.000	137.67
Ca(II)	0.869	53.56
Ag(I)	1.550	98.75

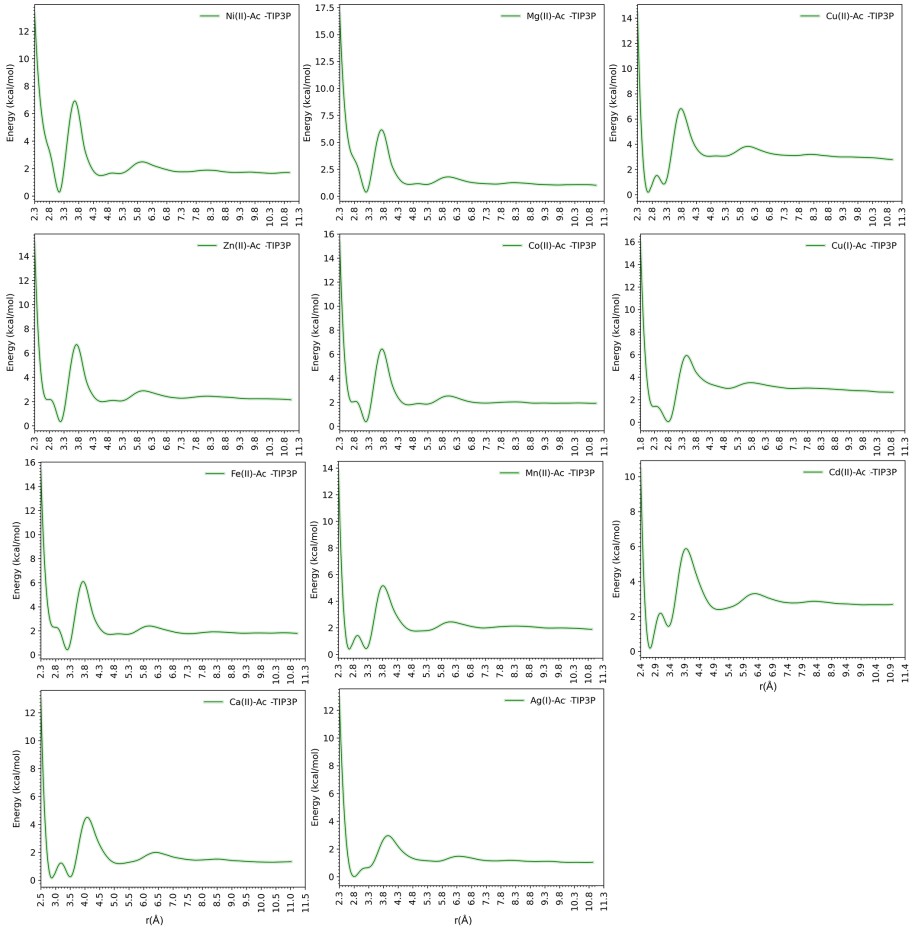

Figure B.1: The PMF free energy profiles of metal ion-acetate complexes in the TIP3P water model. Ac stands for the acetate molecule. The first local minimum, occurring at approximately 2.8 Å (2.3 Å for Cu(I)), corresponds to the bidentate binding mode. The second local minimum, observed at around 3-3.5 Å (2.8 Å for Cu(I)), shows the monodentate binding mode.

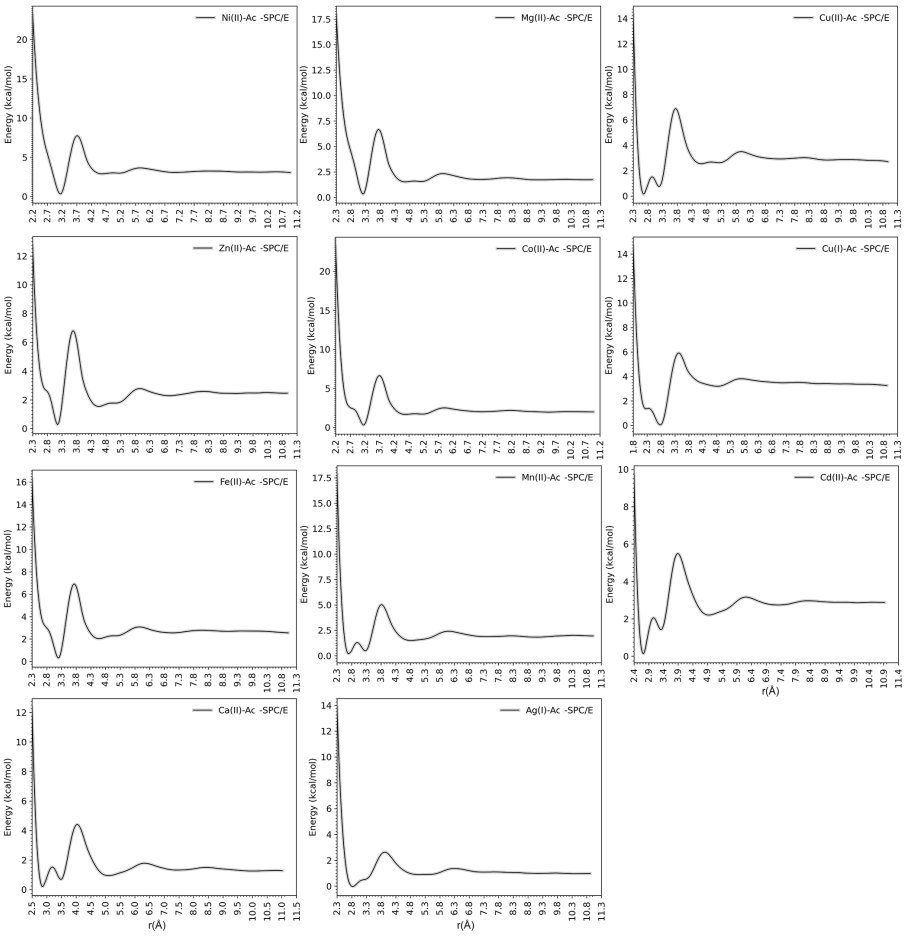

Figure B.2: The PMF free energy profiles of metal ion-acetate complexes in the SPC/E water model. Ac stands for the acetate molecule. The first local minimum, occurring at approximately 2.8 Å (2.3 Å for Cu(I)), corresponds to the bidentate binding mode. The second local minimum, observed at around 3-3.5 Å (2.8 Å for Cu(I)), shows the monodentate binding mode.

Table B.5: Preferred binding mode for each metal ion in the TIP3P (left) and SPC/E (right) water models. The two columns indicate the percentage of monodentate and bidentate binding modes in the acetate-metal ion complex[a].

Ion	TIP3P Binding		SPC/E Binding	
	Monodentate	Bidentate	Monodentate	Bidentate
Ni(II)	99.98%	0.02%	100.00%	0.00%
Mg(II)	99.95%	0.05%	100.00%	0.00%
Cu(II)	30.96%	69.04%	31.79%	68.21%
Zn(II)	98.47%	1.53%	99.59%	0.41%
Co(II)	97.71%	2.29%	99.34%	0.66%
Cu(I)	96.40%	3.60%	96.76%	3.24%
Fe(II)	98.73%	1.27%	99.82%	0.18%
Mn(II)	64.62%	35.38%	53.89%	46.11%
Cd(II)	14.49%	85.51%	13.16%	86.84%
Ca(II)	48.07%	51.93%	28.82%	71.18%
Ag(I)	24.37%	75.63%	30.43%	69.57%

[a] The percentages were calculated by using the PMF profile for each acetate-metal ion complex and applying the Boltzmann distribution based on the minimum energy of each binding mode state.

Appendix C

Supporting Data for Chapter 4

Table C.1: Stability constants (pK_i) between different metal ligand coordination states as computed following section 4.2 at different cadmium and ethylenediamine concentrations

#Cd^{2+} #en	pK_1	pK_2	pK_3
exp[1]	5.4	4.47	2.1
1-3	5.2	4.0	1.8
2-6	5.1	4.1	1.8
10-45	5.1	3.8	1.7
10-30	5.1	3.9	1.9
10-20	4.4	3.5	1.0
10-10	3.7	2.7	not av.

Table C.2: Energy difference from free energy profile (ΔF_{ij}) between different metal ligand coordination states at different cadmium and ethylenediamine concentrations. Between parenthesis the value that it would be expected if taking experimental stability constants as starting point to reverse equation 4.3 in section 4.2

#Cd^{2+} #en	ΔF_{01} (kJ/mol)	ΔF_{12} (kJ/mol)	ΔF_{23} kJ/mol
1-3	18.5 (19.5)	12.8 (14.2)	-0.3 (1.3)
2-6	18.4 (19.4)	12.7 (14.1)	-0.3 (1.2)
10-45	23.7 (24.9)	15.6 (19.5)	3.6 (6.7)
10-30	17.2 (20.4)	13.3 (15.1)	1.6 (2.4)
10-20	15.4 (12.9)	3.8 (7.0)	-5.2 (-7.4)
10-10	5.0 (2.7)	-5.2 (-2.7)	-26.1 (-15.5)

Table C.3: Mean first passage times (τ_i) between different Cd(en)$_i$ coordination states. These results are from the MSM using 60 centers and 60 ps lagtime for 1-3, 60 centers and 60 ps lagtime for 10-30

Ligand conc.	τ_1 (ns)	τ_2 (ns)	τ_3 (ns)
0.08 M	2.2± 0.6	13.4± 0.2	228 ± 5
$0.15^{\dagger}M$	2.7± 0.4	31.6± 0.6	331± 7
	τ_{-3} (ns)	τ_{-2} (ns)	τ_{-1} (ns)
	313± 7	2359± 37	(2.8± 0.5)*10^4
	316± 7	4878± 98	(9.5± 1.2)*10^4

Table C.4: Reaction rates (k_i) between different Cd(en)$_i$ coordination states.

Ligand conc.	k_1 (M^{-1} s^{-1})	k_2 (M^{-1} s^{-1})	k_3 (M^{-1} s^{-1})
0.08 M	(4.1± 1.1)*10^{10}	(6.6± 0.1)*10^9	(3.89± 0.08)*10^8
$0.15^{\dagger}M$	(1.4± 0.2)*10^{10}	(1.14± 0.02)*10^9	(1.09± 0.02)*10^8
	k_{-3} (s^{-1})	k_{-2} (s^{-1})	k_{-1} (s^{-1})
	(3.17± 0.07)*10^6	(4.2± 0.7)*10^5	(3.5± 0.6)*10^4
	(3.16± 0.06)*10^6	(2.05± 0.04)*10^5	(1.1± 0.1)*10^4

Table C.5: Stability constants (pK$_i$) between different Cd(en)$_i$ coordination states as computed from the ratio of the formation and dissociation rate constants of table C.4.

Ligand conc.	pK_1	pK_2	pK_3
0.08 M	6.1± 0.8	4.2± 0.1	2.09± 0.05
$0.15^{\dagger}M$	6.1± 0.5	3.8± 0.1	1.55± 0.01

Table C.6: Stability constants (pK$_i$) between different Ni(en)$_i$ coordination states.

#Ni^{2+} #en	pK_1	pK_2	pK_3
exp[1]	7.35	6.21	4.15
1-3	7.12	5.53	2.96

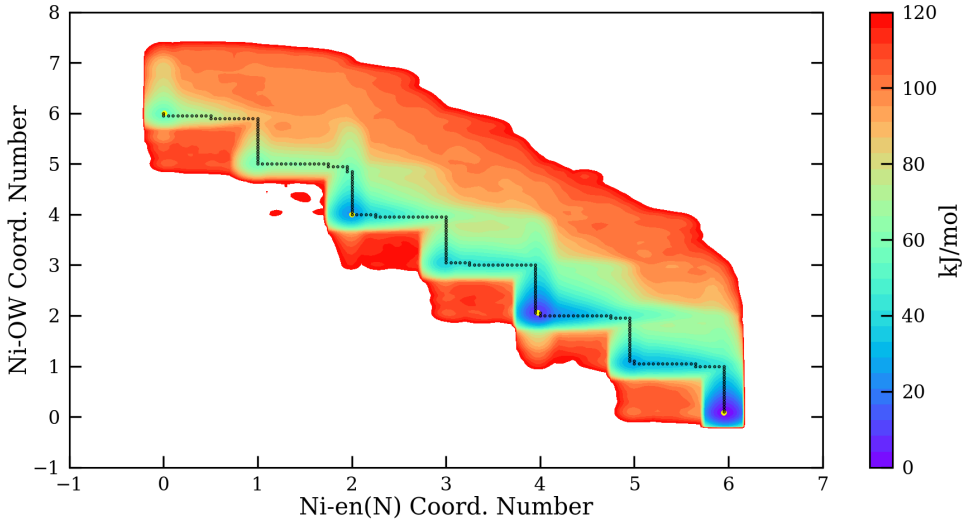

Figure C.1: Free energy map showing different $Ni(en)_i$ and $Ni(OW)_i$ (nichel-water) coordination states. Yellow points are the minima usually measured by experiments. Dotted black line is the minimum free energy pathway.

Table C.7: Mean first passage times (τ_i) and reaction rates (k_i) between different $Ni(en)_i$ coordination states. These results are from the MSM using 32 centers and 200 ps lagtime

τ_1 (μs) $(3.2 \pm 0.3)*10^{-2}$	τ_2 (μs) $(4.4 \pm 0.6)*10^{-1}$	τ_3 (μs) 8.8 ± 1.1
τ_{-1} (s) 3.8 ± 0.4	τ_{-2} (s) $(4.7 \pm 0.7)*10^{-1}$	τ_{-3} (s) $(3.3 \pm 0.4)*10^{-2}$
k_1 (M^{-1} s^{-1}) $(2.8 \pm 0.3)*10^{6}$	k_2 (M^{-1} s^{-1}) $(2.1 \pm 0.3)*10^{5}$	k_3 (M^{-1} s^{-1}) $(1.0 \pm 0.1)*10^{4}$
k_{-1} (s^{-1}) 0.26 ± 0.03	k_{-2} (s^{-1}) 2.1 ± 0.3	k_{-3} (s^{-1}) 31 ± 4

Table C.8: Stability constants (pK_i) between different $Cd(nme)_i$ coordination states as computed following section 4.2 at different cadmium and methylamine concentrations and using different polarizability values.

#Cd^{2+} #nme	pK_1	pK_2	pK_3	pK_4
exp[1]	2.745	2.063	1.131	0.611
1-6[2]	2.52	1.82	0.92	0.05
10-60[2]	2.51	1.99	0.98	0.13
10-60[3]	2.10	1.21	0.22	-0.81

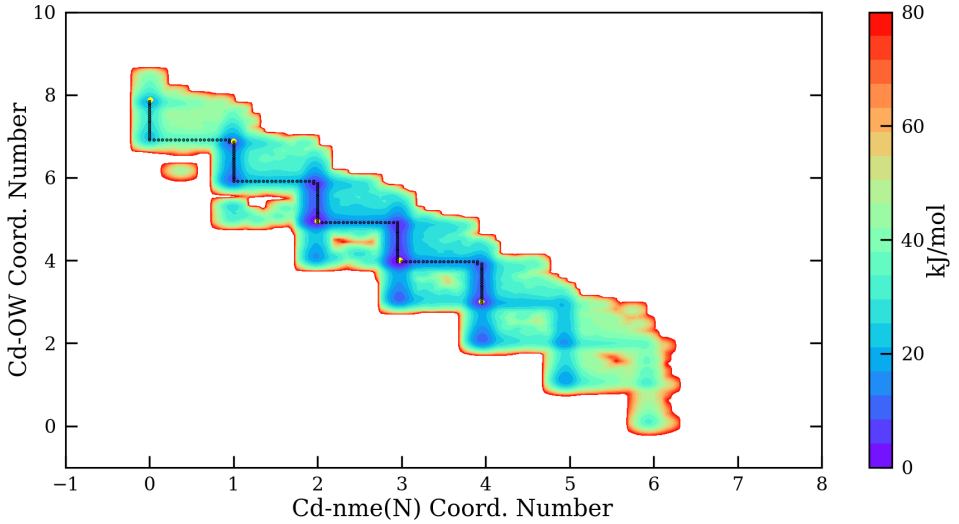

Figure C.2: Free energy map showing different $Cd(nme)_i$ and $Cd(OW)_i$ (cadmium-water) coordination states. Yellow points are the minima usually measured by experiments. Dotted black line is the minimum free energy pathway.

Table C.9: Mean first passage times (τ_i) and reaction rates (k_i) between different $Cd(nme)_i$ coordination states. These results are from the MSM using 40 centers and 20 ps lagtime for 1-3

τ_1 (ns) 0.82± 0.01	τ_2 (ns) 0.72± 0.01	τ_3 (ns) 1.49± 0.01	τ_4 (ns) 7.03± 0.04
τ_{-1} (ns) 483 ± 7	τ_{-2} (ns) 12.47± 0.07	τ_{-3} (ns) 1.19± 0.01	τ_{-4} (ns) 0.27± 0.01
k_1 (M^{-1} s^{-1}) (1.17± 0.01)*10^{10}	k_2 (M^{-1} s^{-1}) (1.34± 0.01)*10^{10}	k_3 (M^{-1} s^{-1}) (6.46± 0.02)*10^{9}	k_4 (s^{-1}) (1.37± 0.01)*10^{9}
k_{-1} (s^{-1}) (2.07± 0.03)*10^{6}	k_{-2} (s^{-1}) (8.02± 0.05)*10^{7}	k_{-3} (s^{-1}) (8.39± 0.02)*10^{8}	k_{-4} (s^{-1}) (3.65± 0.02)*10^{9}

Table C.10: Stability constants (pK_i) between different metal ligand coordination states of Cd^{2+} with diethylenetriamine (dien) and putrescine (put).

Ligand	pK_1	pK_2
dien[1]	8.2	5.5
dien	6.8	3.9
put[2]	3.98	3.2
put	3.11	2.45

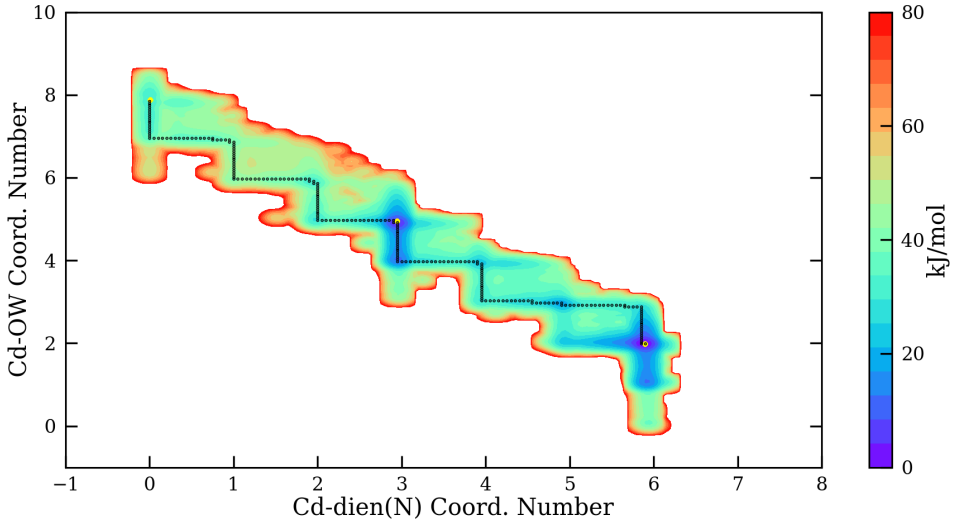

Figure C.3: Free energy map showing different $\mathrm{Cd(dien)}_i$ and $\mathrm{Cd(OW)}_i$ (cadmium-water) coordination states. Yellow points are the minima usually measured by experiments. Dotted black line is the minimum free energy pathway.

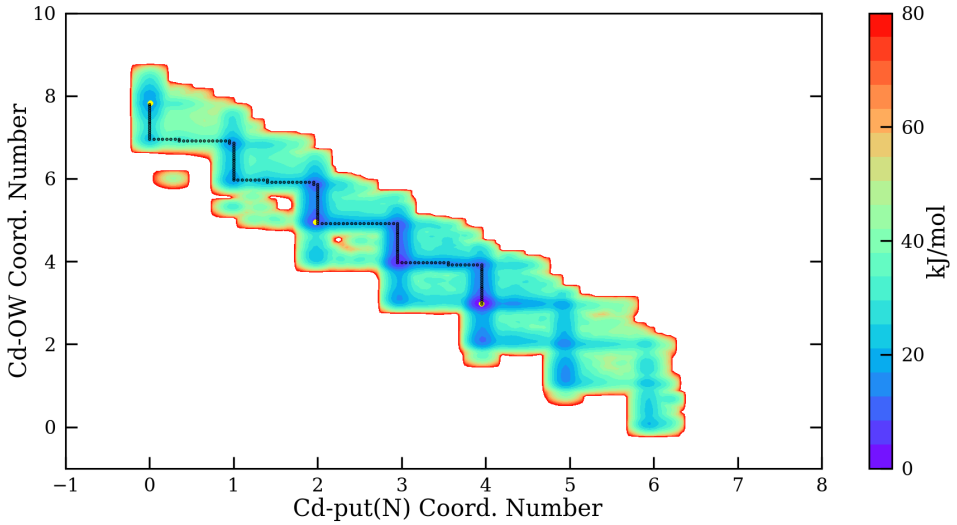

Figure C.4: Free energy map showing different $\mathrm{Cd(put)}_i$ and $\mathrm{Cd(OW)}_i$ (cadmium-water) coordination states. Yellow points are the minima usually measured by experiments. Dotted black line is the minimum free energy pathway.

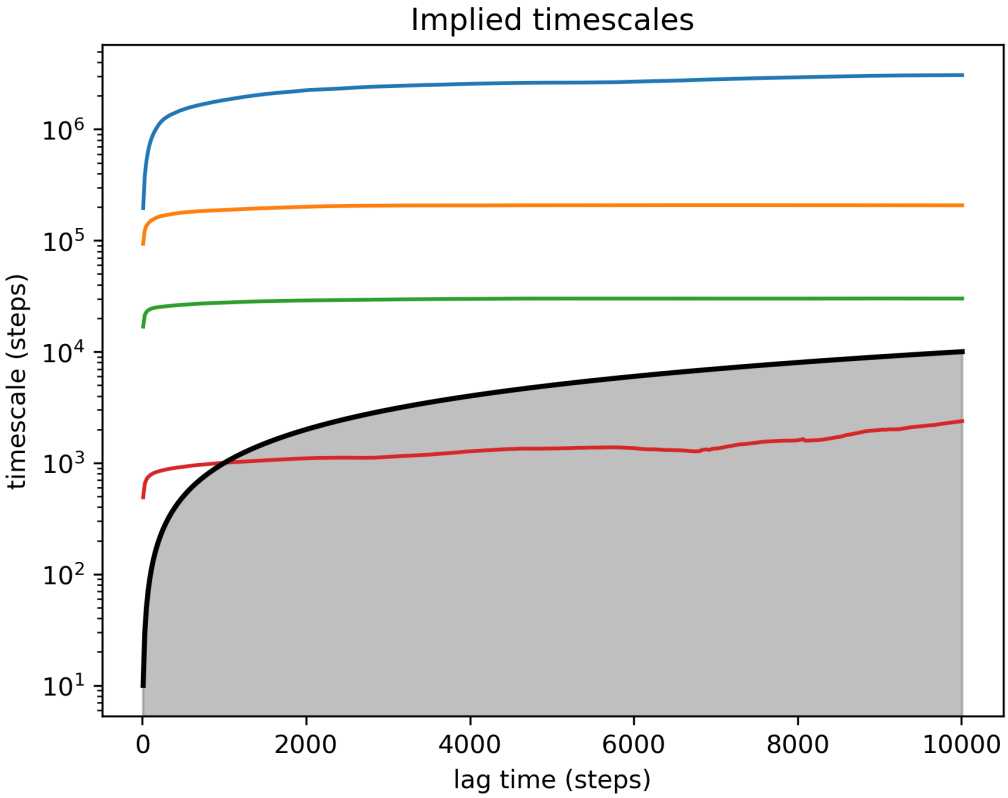

Figure C.5: Implied timescales plot for the 4 slowest eigenvalues associated to the MSM of the cadmium-ethylenediamine system. After 600 steps (60 ps) the three slowest implied timescales are nicely approximated even for longer lagtimes.

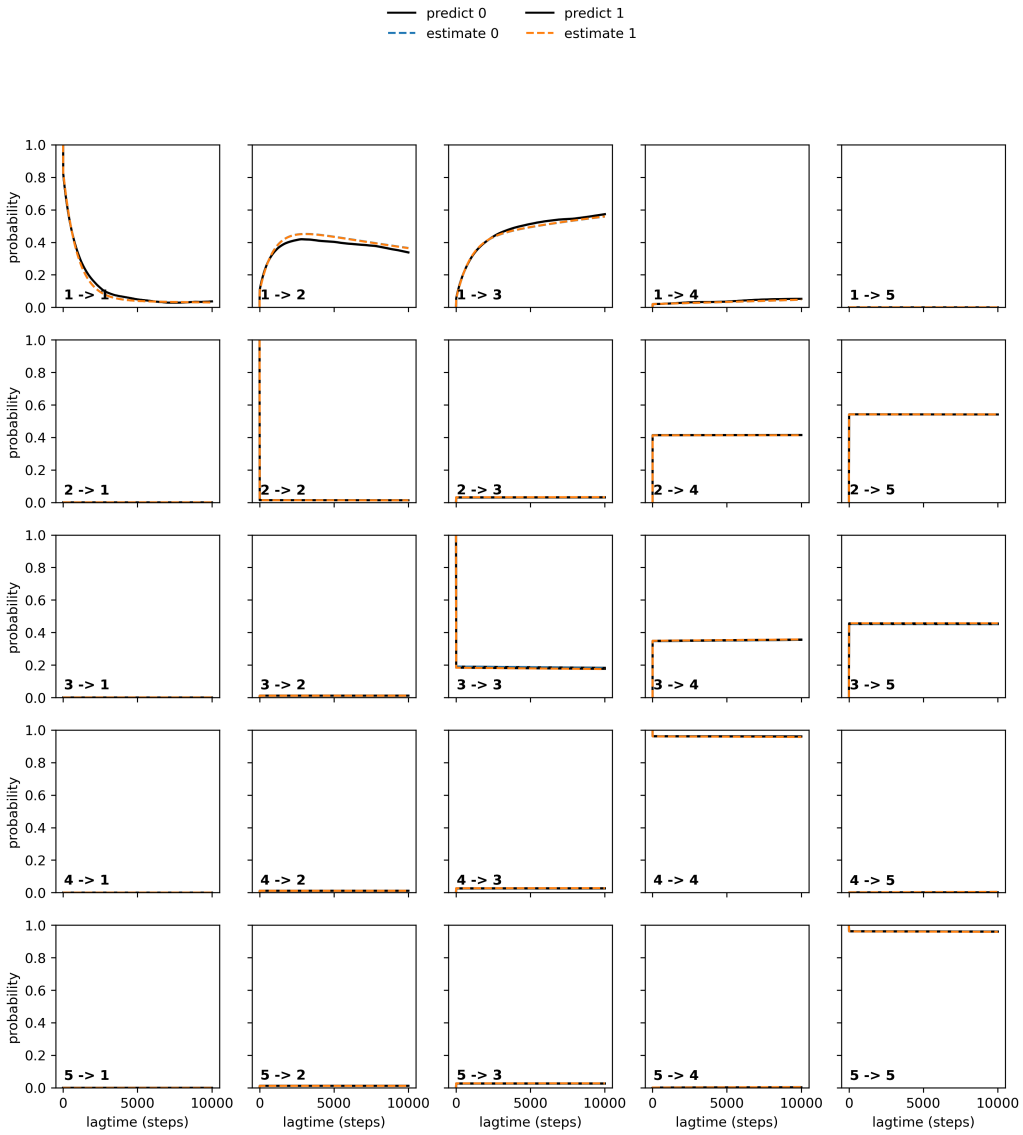

Figure C.6: Chapman-Kolmogorov test for the 60 ps lagtime. The dynamics of the 5 metastable states at logner lagtimes are well reproduced for the lagtime chosen (60 ps).

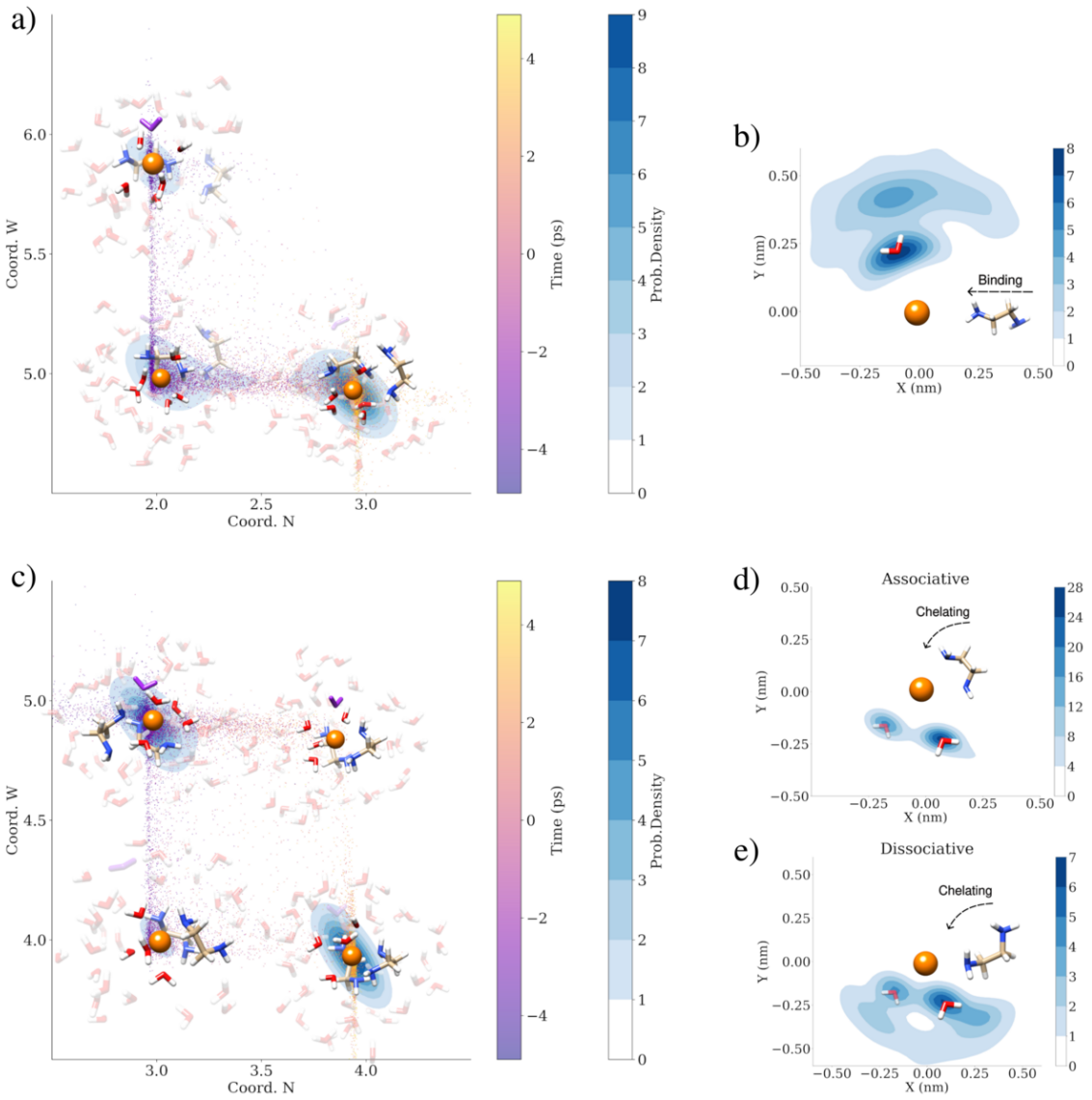

Figure C.7: a) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the second en first nitrogen binding. Scattered plot represents the time evolution and the probability distribution identifies the main steps of the mechanism. b), Probability distribution of the leaving water position for the second en first nitrogen binding. c) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the second en chelating ring closure. d), e) Probability distribution of the leaving water position for the second en chelating ring closure with associative (d) and dissociative (e) mechanism.

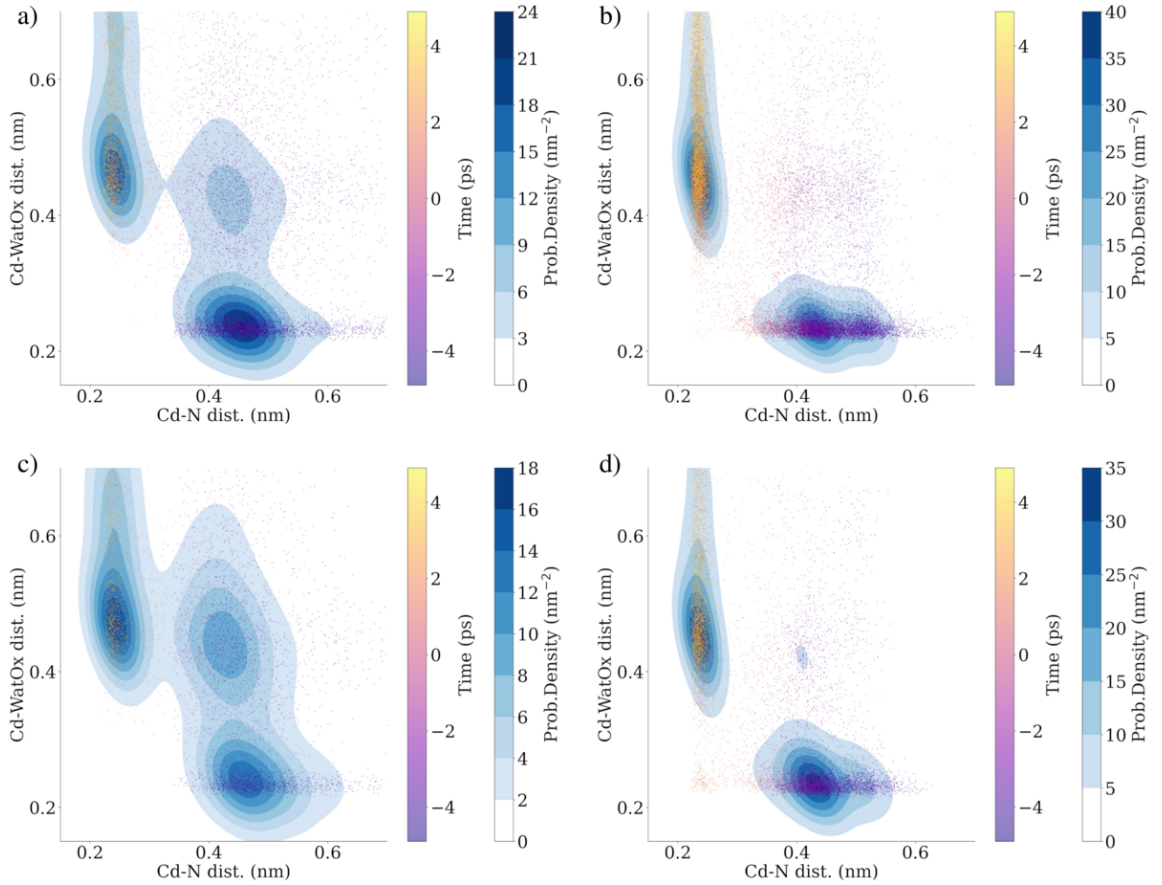

Figure C.8: a), c), Evolution and probability distribution of water and nitrogen-ion distances 5 ps before and after the first and second en first nitrogen binding. Scattered plot represents the time evolution and the probability distribution identifies the main steps of the mechanism. b), d) Evolution and probability distributions of water and nitrogen-ion distances 5 ps before and after the first and second en chelating ring closure.

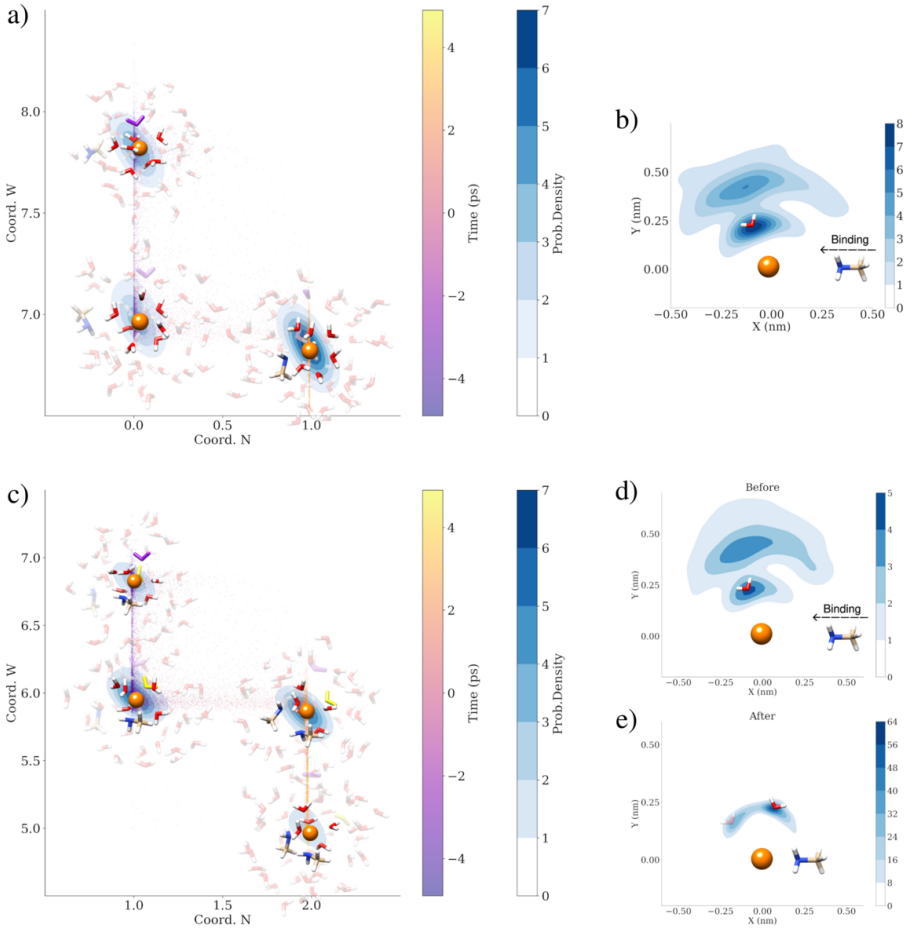

Figure C.9: a), Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the first nme binding. Scattered plot represents the time evolution and the probability distribution identifies the main steps of the mechanism. b), Probability distribution of the leaving water position for the first nme binding. c) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the second nme binding. d), e) Probability distribution of the leaving water position for the second nme binding before (d) and after (e) the binding event.

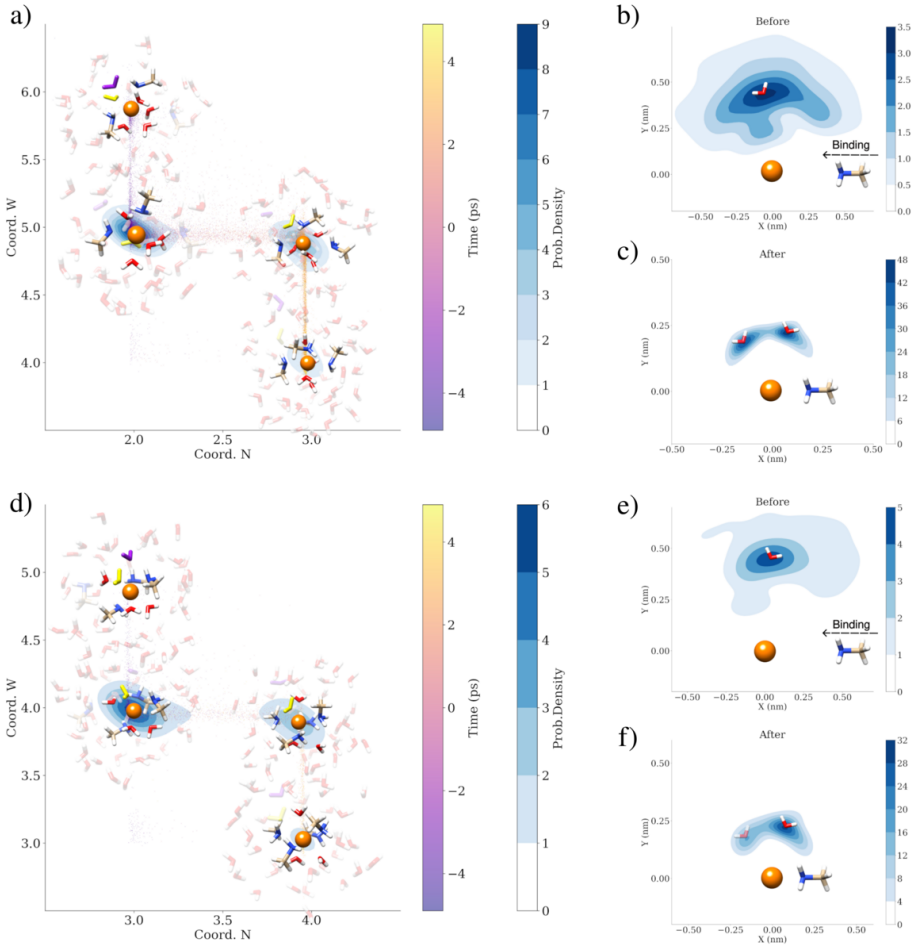

Figure C.10: a) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the third nme binding. b), c) Probability distribution of the leaving water position for the third nme binding before (b) and after (c) the binding event. d) Evolution and probability distribution of water and nitrogen coordination number 5 ps before and after the second nme binding. e), f) Probability distribution of the leaving water position for the second nme binding before (e) and after (f) the binding event.

Appendix D

Supporting Data for Chapter 6

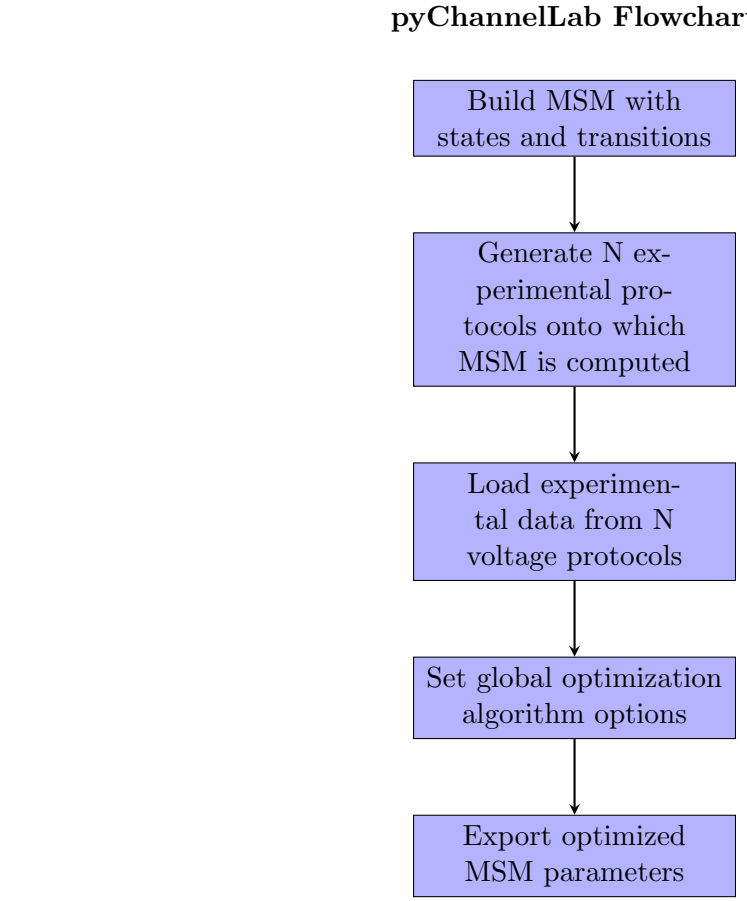

Figure D.1: Workflow of the software pyChanneLab

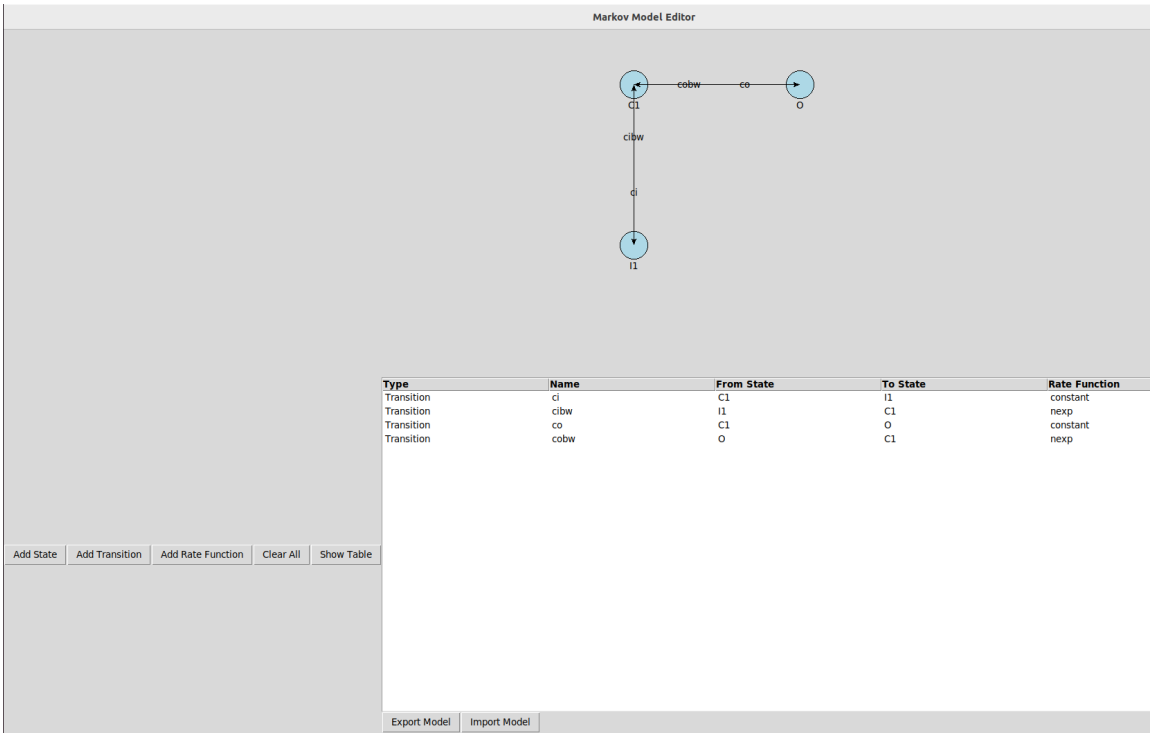

Figure D.2: Screenshot of the MSM Editor User Interface

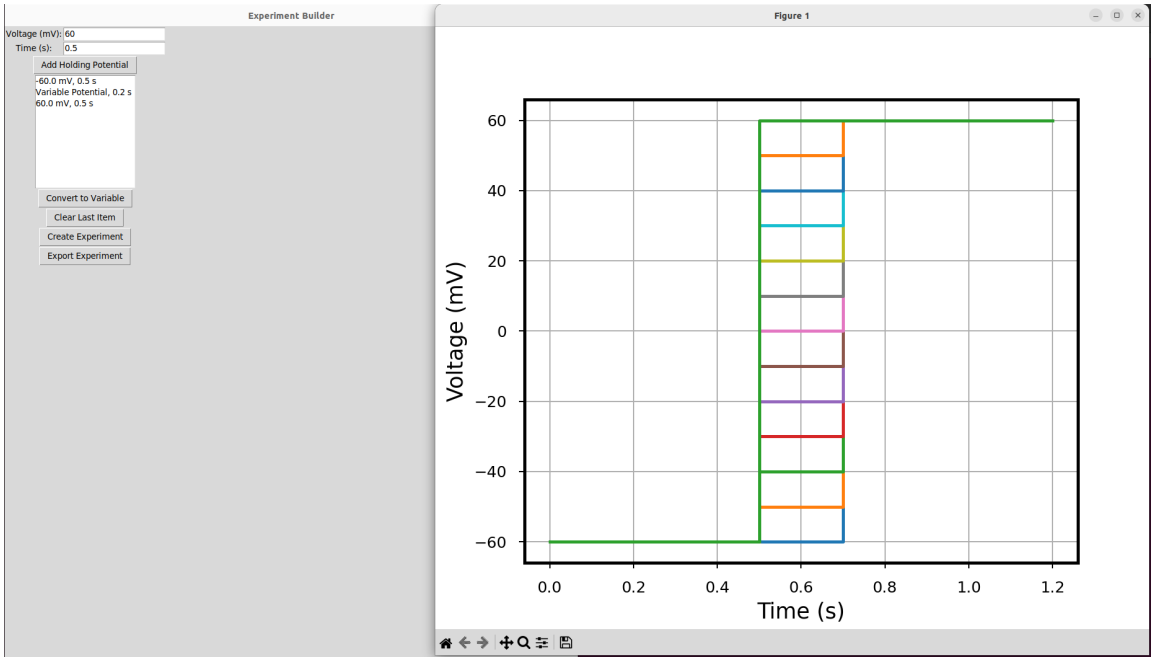

Figure D.3: Screenshot of the experimental protocol builder User Interface

Table D.1: Kinetic model rate constants for the system with KChIP. The first three columns are the parameters found through global optimization algorithms. The last two columns are the kinetic model parameters found through global optimization algorithms for the two mutants keeping the first four parameters fixed (F) to the ones of the wild-type.

Parameters	WT	M373I	S390N	M373IF	S390NF
$\alpha_0(s^{-1})$	870	1645	133	870	870
α_1	1.16	2.17	0.79	1.16	1.16
$\beta_0(s^{-1})$	2.53	5.11	4.10	2.53	2.53
β_1	2.36	2.18	1.95	2.36	2.36
$\mathrm{k}_{CO_0}(s^{-1})$	327	495	370	265	477
k_{CO_1}	0.48	0.33	0.056	0.27	0.71
$\mathrm{k}_{OC_0}(s^{-1})$	448	992	21	87	903
k_{OC_1}	0.14	0.56	0.024	1.44	0.03
k_{CI}	161	51	710	1988	124
k_{IC}	0.01	0.19	0.092	0.01	0.29
f	0.31	0.35	0.26	0.01	0.34

Table D.2: Kinetic model rate constants for the system without KChIP. The parameters have been found through global optimization algorithms starting from the parameters found in table D.1

Parameters	WT	M373I	S390N
$\alpha_0(s^{-1})$	939	341	1518
α_1	2.13	2.35	2.02
$\beta_0(s^{-1})$	56	27	14.6
β_1	3.4e-4	3.3e-4	0.01
$k_{CO_0}(s^{-1})$	305	78	338
k_{CO_1}	0.585	0.64	0.33
$k_{OC_0}(s^{-1})$	822	119	962
k_{OC_1}	0.03	9.6e-3	0.25
k_{CI}	49.7	95.6	70
k_{IC}	0.2	0.15	1.0
f	0.45	0.43	0.25

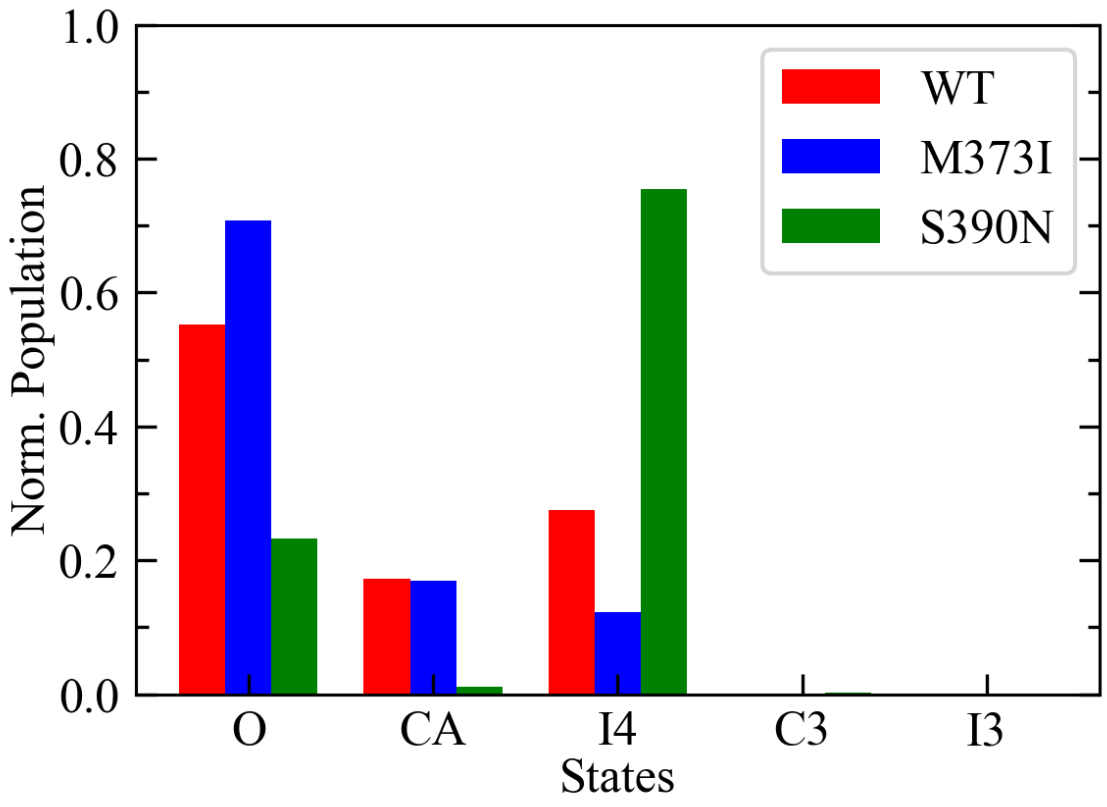

Figure D.4: Population analysis of the kinetic model (Figure 6.2 and Table D.1) as obtained from simulation of WT (red), M373I (blue) and S390N (green) in presence of KChIP upon activation at +60mV and at maximum peak conductance (i.e., maximum open state population). The number of ion channels found in each state was normalized with respect to the total number of simulated channels.

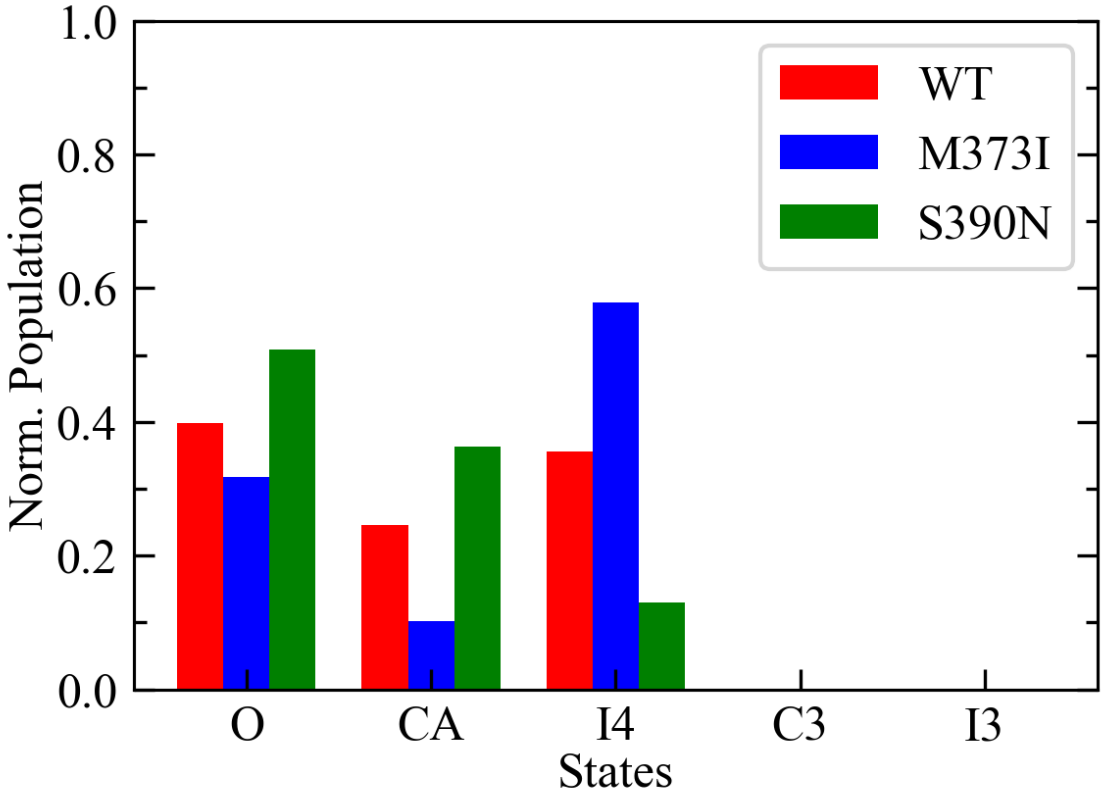

Figure D.5: Population analysis of the kinetic model (Figure 6.2 and Table D.1) as obtained from simulation of WT (red), M373I (blue) and S390N (green) in absence of KChIP upon activation at -10mV. The number of ion channels found in each state was normalized with respect to the total number of simulated channels.

Bibliography

[1] Y.Marcus, *Ions in Solution and their Solvation*, John Wiley & Sons, Ltd (2015).

[2] A.Ben-Naim and Y.Marcus, *The Journal of Chemical Physics* **81**, 4 (1984).

[3] O. Y.Samoilov, *Discuss. Faraday Soc.* **24**, (1957).

[4] T. J.Gilligan and G.Atkinson, *The Journal of Physical Chemistry* **84**, 2 (1980-01).

[5] E.Goldammerv., *Modern Aspects of Electrochemistry*, Springer US (1975).

[6] S. F.Lincoln, *Helvetica Chimica Acta* **88**, 3 (2005).

[7] Y.Marcus, *Ion Solvation*, Wiley, Chichester (1985).

[8] Y.Marcus, *The Journal of Physical Chemistry B* **118**, 35 (2014).

[9] W.Liptay, *Zeitschrift für Elektrochemie, Berichte der Bunsengesellschaft für physikalische Chemie* **66**, 3 (1962).

[10] P.Atkins, P. W.Atkins and J. d.Paula, *Atkins' Physical Chemistry*, OUP Oxford (2014).

[11] M. T.Beck, *Pure and Applied Chemistry* **59**, 12 (1987).

[12] M.Tanaka and M.Tabata, *Bulletin of the Chemical Society of Japan* **82**, 10 (2009).

[13] G. H.Nancollas and M. B.Tomson, *Pure and Applied Chemistry* **54**, 12 (1982).

[14] C. W.Davies, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **68**, 0 (1972).

[15] O.Exner, *Correlation Analysis in Chemistry: Recent Advances*, Springer US (1978).

[16] A.Ringbon and L.Eriksson, *Acta chem. scand* **7**, 7 (1953).

[17] J. E.Huheey, E. A.Keiter, R. L.Keiter and O. K.Medhi, *Inorganic chemistry: principles of structure and reactivity*, Pearson Education India (2006).

[18] J. A.Pople, W. G.Schneider, W. G.Schneider, H. J.Bernstein and H. J.Bernstein, *High-resolution Nuclear Magnetic Resonance*, McGraw-Hill (1959).

[19] D. L.Rabenstein and R. J.Kula, *Journal of the American Chemical Society* **91**, 10 (1969).

[20] S. E.Harding and B. Z.Chowdhry, *Protein-ligand Interactions, Structure and Spectroscopy: A Practical Approach*, Oxford University Press (2001).

[21] H. L.Schläfer, G.Gliemann and G.Gliemann, *Basic Principles of Ligand Field Theory*, Wiley-Interscience (1969).

[22] A. B. P.Lever, *Inorganic Electronic Spectroscopy*, Elsevier (1984).

[23] I. B.Bersuker, *Chemical Reviews* **101**, 4 (2001).

[24] C. J.Suckling, *Enzyme Chemistry*, Springer Netherlands (1990).

[25] D. J.Klein, P. B.Moore and T. A.Steitz, *RNA (New York, N.Y.)* **10**, 9 (2004).

[26] B.Hille, *Ion Channels of Excitable Membranes*, Sinauer Associates Inc. (2001).

[27] S.Parthasarathy and M.Ravishankar, *J Anesth Clin Pharmacology* (2007).

[28] T.Shinoda, H.Ogawa, F.Cornelius and C.Toyoshima, *Nature* **459**, 7245 (2009).

[29] J. E.Sirois, Q.Lei, E. M.Talley, C. L.III and D. A.Bayliss, *Journal of Neuroscience* **20**, 17 (2000).

[30] H.Hibino, A.Inanobe, K.Furutani, S.Murakami, I.Findlay and Y.Kurachi, *Physiological Reviews* **90**, 1 (2010).

[31] C.-C.Shieh, M.Coghlan, J. P.Sullivan and M.Gopalakrishnan, *Pharmacological Reviews* **52**, 4 (2000).

[32] C.Andreini, G.Cavallaro, S.Lorenzini and A.Rosato, *Nucleic Acids Research* **41**, D1 (2012).

[33] C. E.Outten, and T. V.O'Halloran, *Science* **292**, 5526 (2001).

[34] D.Osman, M. A.Martini, A. W.Foster, J.Chen, A. J. P.Scott, R. J.Morton, J. W.Steed, E.Lurie-Luke, T. G.Huggins, A. D.Lawrence et al., *Nature Chemical Biology* **15**, 3 (2019).

[35] T. M.Fyles, *Chem. Soc. Rev.* **36**, (2007).

[36] D.Sterratt, B.Graham, A.Gillies and D.Willshaw, *Principles of Computational Modelling in Neuroscience*, Cambridge University Press (2011).

[37] S.-W. O.Jun Wang and Y.-J.Wang, *Channels* **11**, 6 (2017).

[38] E.Carafoli, *Nature Reviews Molecular Cell Biology* **4**, 4 (2003).

[39] A.Mironenko, U.Zachariae, B. L.de Groot and W.Kopec, *Journal of Molecular Biology* **433**, 17 (2021).

[40] B.Corry and M.Thomas, *Journal of the American Chemical Society* **134**, 3 (2012).

[41] M. J.Ryan, L.Gao, F. I.Valiyaveetil, A. A.Kananenka and M. T.Zanni, *Journal of the American Chemical Society* **0**, 0 (2024).

[42] H.-b.Yu, M.Li, W.-p.Wang and X.-l.Wang, *Acta Pharmacologica Sinica* **37**, 1 (2016).

[43] Y.Zhao, S.Inayat, D. A.Dikin, J. H.Singer, R. S.Ruoff and J. B.Troy, *Proceedings of the Institution of Mechanical Engineers, Part N: Journal of Nanoengineering and Nanosystems* **222**, 1 (2008-03-01).

[44] M. N.Weaver, K. M. J.Merz, D.Ma, H. J.Kim and L.Gagliardi, *Journal of Chemical Theory and Computation* **9**, 12 (2013).

[45] I.Hyla-Kryspin and S.Grimme, *Organometallics* **23**, 23 (2004).

[46] N. J.DeYonker, T. R.Cundari and A. K.Wilson, *The Journal of Chemical Physics* **124**, 11 (2006).

[47] W.Jiang, N. J.DeYonker, J. J.Determan and A. K.Wilson, *The Journal of Physical Chemistry A* **116**, 2 (2012).

[48] P.Li and K. M.Merz, *Chem. Rev.* **117**, 3 (2017).

[49] K. E.Riley and K. M.Merz, *The Journal of Physical Chemistry A* **111**, 27 (2007).

[50] R.Gilson and M. C.Durrant, *Dalton Trans.* (2009).

[51] V.Vallet, U.Wahlgren and I.Grenthe, *Journal of the American Chemical Society* **125**, 48 (2003).

[52] M.Uudsemaa and T.Tamm, *The Journal of Physical Chemistry A* **107**, 46 (2003).

[53] M.Bruschi, L.Bertini, V.Bonačić-Koutecký, L.De Gioia, R.Mitrić, G.Zampella and P.Fantucci, *The Journal of Physical Chemistry B* **116**, 22 (2012).

[54] P.Li and K. M.Merz, *J. Chem. Theory Comput.* **10**, 1 (2014).

[55] I. V.Leontyev and A. A.Stuchebrukhov, *The Journal of Chemical Physics* **130**, 8 (2009).

[56] R. T.Sanderson, *Science* **114**, 2973 (1951).

[57] P.Drude, *The Theory of Optics*, Dover Publications Inc.: New York (1902).

[58] A.Warshel and M.Levitt, *Journal of Molecular Biology* **103**, 2 (1976).

[59] M.Souaille, H.Loirat, D.Borgis and M.Gaigeot, *Computer Physics Communications* **180**, 2 (2009).

[60] C.Maffeo, S.Bhattacharya, J.Yoo, D.Wells and A.Aksimentiev, *Chemical Reviews* **112**, 12 (2012).

[61] D.Mackay, P.Berens, K.Wilson and A.Hagler, *Biophysical Journal* **46**, 2 (1984).

[62] A.Šali and T. L.Blundell, *Journal of Molecular Biology* **234**, 3 (1993).

[63] M.Steinegger and J.Söding, *Nature Communications* **9**, 1 (2018).

[64] P. E. M.Lopes, B.Roux and A. D.MacKerell, *Theoretical Chemistry Accounts* **124**, 1 (2009).

[65] S.Yefimov, E.van der Giessen, P. R.Onck and S. J.Marrink, *Biophysical Journal* **94**, 8 (2008).

[66] A. L.Hodgkin and A. F.Huxley, *The Journal of Physiology* **117**, 4 (1952).

[67] A.Hospital, J.Goni, M.Orozco and J.Gelpi, *Advances and applications in bioinformatics and chemistry* **8**, (2015).

[68] D.Roccatano, A.Barthel and M.Zacharias, *Biopolymers* **85**, (2007).

[69] S.Sharma, F.Ding and N.Dokholyan, *Biophysical Journal* **92**, (2007).

[70] I.Tinoco and J.Wen, *Biological Physics* **6**, (2009).

[71] R.Brandman, Y.Brandman and V.Pande, *PLoS One* **7**, (2012).

[72] D.Case, T.Darden and T.Cheatham, *Journal of Computational Chemistry* **2**, (2012).

[73] B.Brooks, C.Brooks and A.Mackerell, *Journal of Computational Chemistry* **2**, (2009).

[74] B.Hess, C.Kutzner, D.Spoelv. d. and L.E., *Journal of Chemical Theory and Computation* **3**, (2005).

[75] M.Nelson, W.Humphrey and A.Gursoy, *The International Journal of High Performance Computing Applications* **4**, (1996).

[76] D.Chandler, *Introduction to modern statistical mechanics*, New York: Oxford University Press (1987).

[77] M.Tuckerman, *Statistical mechanics: theory and molecular simulation*, New York: Oxford University Press (2010).

[78] N.Metropolis, A.Rosenbluth, M.Rosenbluth, M.Teller and E.Teller, *Journal of Chemical Physics* **21**, (1953).

[79] B.Alder and T.Wainwright, *Journal of Chemical Physics* **31**, (1959).

[80] A.Leach, *Molecular modelling: principles and applications*, Pearson Education Press (2001).

[81] T.Halgren, *Journal of American Chemical Society* **20**, (1991).

[82] A.Sengupta, Z.Li, L. F.Song, P.Li and K. M. J.Merz, *Journal of Chemical Information and Modeling* **61**, 2 (2021).

[83] Z.Li, L. F.Song, G.Sharma, B.Koca Fındık and K. M. J.Merz, *Journal of Chemical Theory and Computation* **19**, 2 (2023).

[84] J.-P.Ryckaert, G.Ciccotti and H. J. C.Berendsen, *Journal of Computational Physics* **23**, 3 (1977).

[85] M.Allen and D.Tildesley, *Computer simulation of liquids*, New York: Oxford University Press (1987).

[86] G.Grosso and G.Pastori Parravicini, *Solid state physics*, Academic Press (2005).

[87] B.Luty, I.Tironi and W.Gusterenv., *Journal of Chemical Physics* **103**, (1995).

[88] H.Friedman, *Molecular Physics* **29**, (1975).

[89] H.Ding, N.Karasawa and W.Goddard, *Chemical Physics Letters* **196**, (1992).

[90] T.Schneider and E.Stoll, *Physical Review B* **17**, (1978).

[91] N.Gronbech-Jensen and O.Farago, *Molecular Physics* **111**, (2013).

[92] G.Bussi and M.Parrinello, *Physical Review E* **75**, (2008).

[93] J.Hénin, T.Lelièvre, M. R.Shirts, O.Valsson and L.Delemotte, *Living Journal of Computational Molecular Science* **4**, 1 (2022).

[94] R. C.Bernardi, M. C.Melo and K.Schulten, *Biochimica et Biophysica Acta (BBA) - General Subjects* **1850**, 5 (2015-05).

[95] Y. I.Yang, Q.Shao, J.Zhang, L.Yang and Y. Q.Gao, *The Journal of Chemical Physics* **151**, 7 (2019-08-21).

[96] A. S.Kamenik, S. M.Linker and S.Riniker, *Physical Chemistry Chemical Physics* **24**, 3 (2022).

[97] C.Clementi, H.Nymeyer and J. N.Onuchic, *Journal of Molecular Biology* **298**, 5 (2000).

[98] P. G.Bolhuis, C.Dellago and D.Chandler, *Proceedings of the National Academy of Sciences* **97**, 11 (2000).

[99] S. S.Cho, Y.Levy and P. G.Wolynes, *Proceedings of the National Academy of Sciences* **103**, 3 (2006).

[100] M. A.Rohrdanz, W.Zheng, M.Maggioni and C.Clementi, *The Journal of Chemical Physics* **134**, 12 (2011).

[101] F.Noé and C.Clementi, *Current Opinion in Structural Biology* **43**, (2017-04).

[102] G.Pérez-Hernández, F.Paul, T.Giorgino, G.De Fabritiis and F.Noé, *The Journal of Chemical Physics* **139**, 1 (2013).

[103] M.Chen, *The European Physical Journal B* **94**, 10 (2021-10).

[104] M.Ceriotti, G. A.Tribello and M.Parrinello, *Proceedings of the National Academy of Sciences* **108**, 32 (2011).

[105] W.Chen, A. R.Tan and A. L.Ferguson, *The Journal of Chemical Physics* **149**, 7 (2018).

[106] L.Bonati, V.Rizzi and M.Parrinello, *The Journal of Physical Chemistry Letters* **11**, 8 (2020).

[107] S.Mehdi, Z.Smith, L.Herron, Z.Zou and P.Tiwary, arXiv:2306.09111 (2023-06-16).

[108] F.Pietrucci, *Reviews in Physics* **2**, (2017-11).

[109] C.Dellago, P. G.Bolhuis, F. S.Csajka and D.Chandler, *J. Chem. Phys.* **108**, 5 (1998).

[110] P. G.Bolhuis, D.Chandler, C.Dellago and P. L.Geissler, *Annu. Rev. Phys. Chem.* **53**, 1 (2002).

[111] J. P. M.Postma, H. J. C.Berendsen and J. R.Haak, *Faraday Symp. Chem. Soc.* **17**, (1982).

[112] R. W.Zwanzig, *The Journal of Chemical Physics* **22**, 8 (2004).

[113] C.Jarzynski, *Phys. Rev. E* **56**, (1997).

[114] J.Comer, J. C.Gumbart, J.Hénin, T.Lelièvre, A.Pohorille and C.Chipot, *The Journal of Physical Chemistry B* **119**, 3 (2015).

[115] O.Valsson and M.Parrinello, *Phys. Rev. Lett.* **113**, (2014).

[116] A.Mitsutake, Y.Sugita and Y.Okamoto, *Peptide Science* **60**, 2 (2001).

[117] Y.IBA, *International Journal of Modern Physics C* **12**, 05 (2001).

[118] Y.Sugita, A.Kitao and Y.Okamoto, *The Journal of Chemical Physics* **113**, 15 (2000).

[119] K.Hukushima and K.Nemoto, *Journal of the Physical Society of Japan* **65**, 6 (1996).

[120] A.Laio and M.Parrinello, *PNAS USA* **99**, 20 (2002).

[121] G.Bussi, A.Laio and P.Tiwary, *Metadynamics: A Unified Framework for Accelerating Rare Events and Sampling Thermodynamics and Kinetics*, Springer International Publishing (2018).

[122] A.Barducci, R.Chelli, P.Procacci, V.Schettino, F. L.Gervasio and M.Parrinello, *Journal of the American Chemical Society* **128**, 8 (2006).

[123] J.Pfaendtner, D.Branduardi, M.Parrinello, T. D.Pollard and G. A.Voth, *Proceedings of the National Academy of Sciences* **106**, 31 (2009).

[124] A.Barducci, G.Bussi and M.Parrinello, *Physical Review Letters* **100**, 2 (2008).

[125] A.Barducci, M.Bonomi and M.Parrinello, *WIREs Computational Molecular Science* **1**, 5 (2011).

[126] V.Babin, C.Roland and C.Sagui, *J. Chem. Phys.* (2008).

[127] D.Branduardi, G.Bussi and M.Parrinello, *J. Chem. Theory Comput.* (2012).

[128] F.Baftizadeh, P.Cossio, F.Pietrucci and A.Laio, *Current Physical Chemistry* **2**, 1 (2012).

[129] P.Raiteri, A.Laio, F. L.Gervasio, C.Micheletti and M.Parrinello, (2005).

[130] J.Pfaendtner and M.Bonomi, *J. Chem. Theory Comput.* (2015).

[131] A.Gil-Ley and G.Bussi, *Journal of Chemical Theory and Computation* **11**, 3 (2015).

[132] A.Prakash, C. D.Fu, M.Bonomi and J.Pfaendtner, *J. Chem. Theory Comput.* (2018).

[133] G.Bussi, F. L.Gervasio, A.Laio and M.Parrinello, (2006).

[134] P.Tiwary and M.Parrinello, *J. Phys. Chem. B* (2015).

[135] G. A.Tribello, M.Bonomi, D.Branduardi, C.Camilloni and G.Bussi, *Comput. Phys. Commun.* **185**, 2 (2014).

[136] C. W.Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer (2004).

[137] N. G.Van Kampen, *Stochastic processes in physics and chemistry*, North Holland (2007).

[138] J. E.Moyal, *Journal of the Royal Statistical Society. Series B (Methodological)* **11**, 2 (1949).

[139] F.Noé, *The Journal of Chemical Physics* **128**, 24 (2008).

[140] J. D.Chodera, N.Singhal, V. S.Pande, K. A.Dill and W. C.Swope, *The Journal of Chemical Physics* **126**, 15 (2007).

[141] G. R.Bowman, X.Huang and V. S.Pande, *Methods* **49**, 2 (2009).

[142] T. F.Gonzalez, *Theor. Comput. Sci.* **38**, (1985).

[143] F. N.Gregory R. Bowman, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Springer Dordrecht (2016).

[144] K. A.Beauchamp, G. R.Bowman, T. J.Lane, L.Maibaum, I. S.Haque and V. S.Pande, *Journal of Chemical Theory and Computation* **7**, 10 (2011).

[145] S.Lloyd, *IEEE Transactions on Information Theory* **28**, 2 (1982).

[146] D.Arthur and S.Vassilvitskii, *Technical Report, Stanford* (2006).

[147] W. C.Swope, J. W.Pitera, F.Suits, M.Pitman, M.Eleftheriou, B. G.Fitch, R. S.Germain, A.Rayshubski, T. J. C.Ward, Y.Zhestkov et al., *The Journal of Physical Chemistry B* **108**, 21 (2004).

[148] F.Noé, I.Horenko, C.Schütte and J. C.Smith, *The Journal of Chemical Physics* **126**, 15 (2007).

[149] C.Schütte, A.Fischer, W.Huisinga and P.Deuflhard, *Journal of Computational Physics* **151**, 1 (1999).

[150] R.Susanna and W.Marcus, *Advances in Data Analysis and Classification* **7**, (2013).

[151] F.Noé, C.Schütte, E.Vanden-Eijnden, L.Reich and T. R.Weikl, *Proceedings of the National Academy of Sciences* **106**, 45 (2009).

[152] A.Mardt, L.Pasquali, H.Wu and F.Noé, *Nature Communications* 5 (2018).

[153] T.Löhr, K.Kohlhoff, G. T.Heller, C.Camilloni and M.Vendruscolo, *ACS Chemical Neuroscience* **13**, 12 (2022).

[154] C.Koch, *Biophysics of Computation: Information Processing in Single Neurons* , Oxford University Press (1998).

[155] D. J.Wilkinson, *Stochastic modelling for systems biology*, CRC press, Taylor & Francis group (2019).

[156] A.Tveito and G. T.Lines, *Computing Characterizations of Drugs for Ion Channels and Receptors Using Markov Models*, Springer International Publishing (2016).

[157] D. T.Gillespie, *Annual Review of Physical Chemistry* **58**, 1 (2007).

[158] L. L.Perissinotti, J.Guo, P. M.De Biase, C. E.Clancy, H. J.Duff and S. Y.Noskov, *Biophysical Journal* **108**, 6 (2015).

[159] M.A. Linc., S. C.Cannon and D. M.Papazian, *Proceedings of the National Academy of Sciences* **115**, 15 (2018).

[160] L.Sagresti, L.Peri, G.Ceccarelli and G.Brancato, *Journal of Chemical Theory and Computation* **18**, 5 (2022).

[161] Y.Marcus, *Chem. Rev.* **109**, 3 (2009).

[162] D. T.Richens, *The Chemistry of Aqua Ions: Synthesis, Structure and Reactivity: ATour Through the Periodic Table of the Elements*, Wiley (1997).

[163] E.Gouaux and R.MacKinnon, *Science* **310**, 5753 (2005).

[164] E.Khare, N.Holten-Andersen and M. J.Buehler, *Nat. Rev. Mater.* **6**, 5 (2021).

[165] J.Blumberger, *J. Am. Chem. Soc.* **130**, 47 (2008).

[166] Y.Marcus, *J. Phys. Chem. B* **109**, 39 (2005).

[167] I.Persson, P.D'Angelo, S.De Panfilis, M.Sandström and L.Eriksson, *Chem. Eur. J.* **14**, 10 (2008).

[168] J.Blumberger and M.Sprik, *J. Phys. Chem. B* **108**, 21 (2004).

[169] G.Brancato and V.Barone, *J. Phys. Chem. B* **115**, 44 (2011).

[170] S.Chempath, L. R.Pratt and M. E.Paulaitis, *J. Chem. Phys.* **130**, 5 (2009).

[171] D.Asthagiri, P. D.Dixit, S.Merchant, M. E.Paulaitis, L. R.Pratt, S. B.Rempe and S.Varma, *Chem. Phys. Lett.* **485**, 1 (2010).

[172] F. P.Rotzinger, *J. Am. Chem. Soc.* **119**, 22 (1997).

[173] G.Mancini, G.Brancato and V.Barone, *J. Chem. Theory Comput.* **10**, 3 (2014).

[174] J.Neely and R.Connick, *J. Am. Chem. Soc.* **92**, 11 (1970).

[175] L.Helm and A. E.Merbach, *Coord. Chem. Rev.* **187**, 1 (1999).

[176] R.Rey and J. T.Hynes, *J. Phys. Chem.* **100**, 14 (1996).

[177] S.Kerisit and C.Liu, *J. Phys. Chem. A* **117**, 30 (2013).

[178] N.Schwierz, *J. Chem. Phys.* **152**, 22 (2020).

[179] D.Chandler, *J. Chem. Phys.* **68**, (1978).

[180] P.Hänggi, P.Talkner and M.Borkovec, *Rev. Mod. Phys.* **62**, (1990).

[181] G.Hummer, *New J. Phys.* **7**, (2005).

[182] B.Peters, P. G.Bolhuis, R. G.Mullen and J.-E.Shea, *J. Chem. Phys.* **138**, 5 (2013).

[183] H. C.Öttinger, *Stochastic Calculus*, Springer Berlin Heidelberg (1996).

[184] A. M.Berezhkovskii and A.Szabo, *J. Chem. Phys.* **150**, 5 (2019).

[185] R.Elber, A.Fathizadeh, P.Ma and H.Wang, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11**, 4 (2021).

[186] J.Crank and P.Nicolson, *Math. Proc. Camb. Philos. Soc.* **43**, 1 (1947).

[187] H.Kramers, *Physica* **7**, 4 (1940).

[188] A. H.Al-Mohy and N. J.Higham, *SIAM Journal on Scientific Computing* **34**, 4 (2012).

[189] G.Hummer, *J. Chem. Phys.* **120**, 2 (2004).

[190] Y. M.Rhee and V. S.Pande, *J. Phys. Chem. B* **109**, 14 (2005).

[191] J. D.Chodera and V. S.Pande, *Phys. Rev. Lett.* **107**, (2011).

[192] W.Lechner, J.Rogal, J.Juraszek, B.Ensing and P. G.Bolhuis, *J. Chem. Phys.* **133**, 17 (2010).

[193] A. D.MacKerell, D.Bashford, M.Bellott, R. L.Dunbrack, J. D.Evanseck, M. J.Field, S.Fischer, J.Gao, H.Guo, S.Ha et al., *J. Phys. Chem. B* **102**, 18 (1998).

[194] H. J. C.Berendsen, J. R.Grigera and T. P.Straatsma, *J. Phys. Chem.* **91**, 24 (1987).

[195] C.Oostenbrink, A.Villa, A. E.Mark and W. F.Van Gunsteren, *J. Comput. Chem.* **25**, 13 (2004).

[196] C. S.Babu and C.Lim, *J. Phys. Chem. A* **110**, 2 (2006).

[197] M. J.Abraham, T.Murtola, R.Schulz, S.Páll, J. C.Smith, B.Hess and E.Lindahl, *SoftwareX* **1-2**, (2015).

[198] G.Maruyama, *Rend. Circ. Mat. Palermo* **4**, (1955).

[199] A. K.Katz, J. P.Glusker, S. A.Beebe and C. W.Bock, *J. Am. Chem. Soc.* **118**, 24 (1996).

[200] L.Helm and A. E.Merbach, *Chem. Rev.* **105**, 6 (2005).

[201] M.Jafari, Z.Li, L. F.Song, L.Sagresti, G.Brancato and K. M. J.Merz, *The Journal of Physical Chemistry B* **0**, 0 (2024).

[202] M. J.Chalkley, S. I.Mann and W. F.DeGrado, *Nature Reviews Chemistry* **6**, 1 (2022).

[203] A. W.Maniccia, W.Yang, J. A.Johnson, S.Li, H.Tjong, H.-X.Zhou, L. A.Shaket and J. J.Yang, *PMC Biophysics* **2**, 1 (2009).

[204] Z.Li, L. F.Song, P.Li and K. M.Merz, *J. Chem. Theory Comput.* **16**, 7 (2020).

[205] S.Kumar, J. M.Rosenberg, D.Bouzida, R. H.Swendsen and P. A.Kollman, *Journal of Computational Chemistry* **13**, 8 (1992).

[206] H.Leontiadou, A. E.Mark and S. J.Marrink, *Biophysical Journal* **92**, 12 (2007).

[207] A. E.Ailenei and T. A.Beu, *Journal of Molecular Structure* **1245**, (2021).

[208] V.Ngo, H.Li, A. D.Mackerell, T. W.Allen, B.Roux and S.Noskov, *Journal of Chemical Theory and Computation* **17**, 3 (2021).

[209] L.Shen, Z.Xu, Z. W.Zhou and G. H.Hu, *Chinese Physics B* **23**, 11 (2014).

[210] W. L.Jorgensen, J.Chandrasekhar, J. D.Madura, R. W.Impey and M. L.Klein, *The Journal of Chemical Physics* **79**, 2 (1983).

[211] S.Izadi, R.Anandakrishnan and A. V.Onufriev, *The Journal of Physical Chemistry Letters* **5**, 21 (2014).

[212] G. A.Tribello, M.Bonomi, D.Branduardi, C.Camilloni and G.Bussi, *Computer Physics Communications* **185**, 2 (2014).

[213] The PLUMED consortium, *Nature Methods* **16**, 8 (2019).

[214] D. A.Case, K.Belfon, I. Y.Ben-Shalom, S. R.Brozell, D. S.Cerutti, I. I. I.T.E. Cheatham, V. W. D.Cruzeiro, T. A.Darden, R. E.Duke, G.Giambasu et al., *University of California, San Francisco* (2020).

[215] H. J. C.Berendsen, J. P. M.Postma, W. F.Gunsterenv., A.DiNola and J. R.Haak, *The Journal of Chemical Physics* **81**, 8 (1984).

[216] W.Humphrey, A.Dalke and K.Schulten, *Journal of Molecular Graphics* **14**, (1996).

[217] G.Van Rossum and F. L.Drake, *Python 3 Reference Manual*, CreateSpace (2009).

[218] M. M.He, S. L.Clugston, J. F.Honek and B. W.Matthews, *Biochemistry* **39**, 30 (2000).

[219] I. V.Kurnikov and M.Kurnikova, *The Journal of Physical Chemistry B* **119**, 32 (2015).

[220] J. Y.Xiang and J. W.Ponder, *Journal of Computational Chemistry* **34**, 9 (2013).

[221] D.Semrouni, W. C. I. I. I.Isley, C.Clavaguéra, J.-P.Dognon, C. J.Cramer and L.Gagliardi, *Journal of Chemical Theory and Computation* **9**, 7 (2013).

[222] T.Verstraelen, S.Vandenbrande and P. W.Ayers, *The Journal of Chemical Physics* **141**, 19 (2014).

[223] C. M.Baker, *WIREs Computational Molecular Science* **5**, 2 (2015).

[224] Z.Jing, C.Liu, S. Y.Cheng, R.Qi, B. D.Walker, J.-P.Piquemal and P.Ren, *Annu. Rev. Biophys* **48**, (2019).

[225] P.Ren and J. W.Ponder, *Journal of Computational Chemistry* **23**, 16 (2002).

[226] D. M.De Oliveira, S. R.Zukowski, V.Palivec, J.Hénin, H.Martinez-Seara, D.Ben-Amotz, P.Jungwirth and E.Duboué-Dijon, *Physical Chemistry Chemical Physics* **22**, 41 (2020).

[227] Z.Jing, R.Qi, C.Liu and P.Ren, *Journal of Chemical Physics* **147**, 16 (2017).

[228] E.Ahlstrand, K.Hermansson and R.Friedman, *Journal of Physical Chemistry A* **121**, 13 (2017).

[229] T.Martinek, E.Duboue-Dijon, S.Timr, P. E.Mason, K.Baxova, H. E.Fischer, B.Schmidt, E.Pluharova and P.Jungwirth, *Journal of Chemical Physics* **148**, 22 (2018).

[230] J. A.Lemkul, J.Huang, B.Roux and A. D.Mackerell, *Chemical Reviews* **116**, 9 (2016).

[231] H.Yu, T. W.Whitfield, E.Harder, G.Lamoureux, I.Vorobyov, V. M.Anisimov, A. D. J.MacKerell and B.Roux, *Journal of Chemical Theory and Computation* **6**, 3 (2010).

[232] P. E. M.Lopes, J.Huang, J.Shim, Y.Luo, H.Li, B.Roux and A. D. J.MacKerell, *Journal of Chemical Theory and Computation* **9**, 12 (2013).

[233] J. W.Bunting and K. M.Thong, *Canadian Journal of Chemistry* **48**, (1970).

[234] R. M.Smith and A. E.Martell, *Critical Stability Constants*, Springer US (1989).

[235] R. D.Hancock and F.Marsicano, *Inorganic Chemistry* **19**, 9 (1980).

[236] A.Frank Wells, *Structural Inorganic Chemistry*, Oxford University Press (1984).

[237] A.Benhassine, H.Boulebd, B.Anak, A.Bouraiou, S.Bouacida, M.Bencharif and A.Belfaitah, *Journal of Molecular Structure* **1160**, (2018).

[238] U.Ryde, *Biophysical Journal* **77**, 5 (1999).

[239] T.Dudev and C.Lim, *Accounts of Chemical Research* **40**, 1 (2007).

[240] T.Dudevt and C.Lim, *Journal of the American Chemical Society* **128**, 5 (2006).

[241] T.Dudev and C.Lim, *Journal of Physical Chemistry B* **108**, 14 (2004).

[242] A. N.Cain, T. R. N.Carder Freeman, K. D.Roewe, D. L.Cockriel, T. R.Hasley, R. D.Maples, E. M.Allbritton, T.D'Huys, T.Van Loy, B. P.Burke et al., *Dalton Transactions* **48**, 8 (2019).

[243] M. J.Stevens and S. L.Rempe, *Journal of Physical Chemistry B* **120**, 49 (2016).

[244] W. K. H.Ho, Z. Y.Bao, X.Gan, K. Y.Wong, J.Dai and D.Lei, *Journal of Physical Chemistry Letters* **10**, 16 (2019).

[245] W.Plazinski and M.Drach, *Journal of Physical Chemistry B* **117**, 40 (2013).

[246] F.Valach, M.Melník, G.Bernardinelli and K. M.Fromm, *Journal of Chemical Crystallography* **36**, 9 (2006).

[247] N.Muhammad, M.Ikram, F.Perveen, M.Ibrahim, M.Ibrahim, Abel, Viola, S.Rehman, S.Shujah, W.Khan et al., *Journal of Molecular Structure* **1196**, (2019).

[248] A.Sigel and H.Sigel, *Metal ions in biological systems*, CRC Press (1998).

[249] C. J.Carrell, H. L.Carrell, J.Erlebacher and J. P.Glusker, *Journal of the American Chemical Society* **110**, 26 (1988).

[250] P.Chakrabarti and P.Chakrabarti, *Protein Eng* **4**, 1 (1990).

[251] Effendy, F.Marchetti, C.Pettinari, R.Pettinari, B. W.Skelton and A. H.White, *Inorganica Chimica Acta* **360**, 5 (2007).

[252] A.Grodzicki, I.Łakomska, P.Piszczek, I.Szymańska and E.Szłyk, *Coordination Chemistry Reviews* **249**, 21 (2005).

[253] M. M.Harding, *Acta Crystallogr D Biol Crystallogr* **55**, (1999).

[254] M.Smith, Z.Li, L.Landry, K. M.Merz and P.Li, *Journal of Chemical Theory and Computation* **19**, 7 (2023).

[255] W. A.Herrmann and C. W.Kohlpaintner, *Angewandte Chemie International Edition in English* **32**, 11 (1993).

[256] J.Joseph and G. A.Rani, *Applied Biochemistry and Biotechnology* **172**, 2 (2014).

[257] V. A.Friese and D. G.Kurth, *Coordination Chemistry Reviews* **252**, 1 (2008).

[258] P.Cieśla, P.Kocot, P.Mytych and Z.Stasicka, *Journal of Molecular Catalysis A: Chemical* **224**, 1 (2004).

[259] M.Hruby, I. I. S.Martínez, H.Stephan, P.Pouckova, J.Benes and P.Stepanek, *Polymers* **13**, 22 (2021).

[260] C. A.Ghisalberti, E.Falletta, C.Lammi, G.Facchetti, R.Bucci, E.Erba and S.Pellegrino, *Polymer Testing* **90**, (2020).

[261] M. S.Diallo, W.Arasho, J. H. J.Johnson and W. A.Goddard III, *Environmental Science & Technology* **42**, 5 (2008).

[262] J.Guom., P.Makvandi, C.Weic., J.Chenh., H.Xuk., L.Breschi, D. H.Pashley, C.Huang, L.Niun. and F. R.Tay, *Acta Biomaterialia* **90**, (2019).

[263] S.Fortuna, F.Fogolari and G.Scoles, *Scientific Reports* **5**, 1 (2015).

[264] A. M.Pyle and J. K.Barton, *Probing Nucleic Acids with Transition Metal Complexes*, John Wiley & Sons, Ltd (1990).

[265] M. R.Rodríguez, M. J.Lavecchia, B. S.Parajón-Costa, A. C.González-Baró, M. R.González-Baró and E. R.Cattáneo, *Biochimie* **186**, (2021).

[266] Z.Yu and J.Cowan, *Current Opinion in Chemical Biology* **43**, (2018).

[267] A. T.Aron, K. M.Ramos-Torres, J. A. J.Cotruvo and C. J.Chang, *Accounts of Chemical Research* **48**, 8 (2015).

[268] J.Berrones Reyes, M. K.Kuimova and R.Vilar, *Current Opinion in Chemical Biology* **61**, (2021).

[269] M.Frezza, S.Hindo, D.Chen, A.Davenport, S.Schmitt, D.Tomco and Q.Ping Dou, *Current Pharmaceutical Design* **16**, 16 (2010).

[270] C. C.Konkankit, S. C.Marker, K. M.Knopf and J. J.Wilson, *Dalton Trans.* **47**, (2018).

[271] U.Ndagi, N.Mhlongo and M.Soliman, *Drug Des Devel Ther.* **11**, (2017).

[272] H.Hu, Q.Xu, Z.Mo, X.Hu, Q.He, Z.Zhang and Z.Xu, *Journal of Nanobiotechnology* **20**, 1 (2022).

[273] A. E.Martell and R. D.Hancock, *Metal Complexes in Aqueous Solutions*, Springer US (1996).

[274] A.Sengupta, A.Seitz and K. M. J.Merz, *Journal of the American Chemical Society* **140**, 45 (2018).

[275] C. G.Spike and R. W.Parry, *Journal of the American Chemical Society* **75**, 11 (1953).

[276] Y.Marcus, *Chemical Reviews* **88**, 8 (1988).

[277] M. T.Panteva, G. M.Giambaşu and D. M.York, *The Journal of Physical Chemistry B* **119**, 50 (2015).

[278] A. D. J.MacKerell, D.Bashford, M.Bellott, R. L. J.Dunbrack, J. D.Evanseck, M. J.Field, S.Fischer, J.Gao, H.Guo, S.Ha et al., *The Journal of Physical Chemistry B* **102**, 18 (1998).

[279] P.Paoletti, *Pure and Applied Chemistry* **56**, 4 (1984).

[280] R. S.Tobias, *Journal of Chemical Education* **35**, 12 (1958).

[281] D. A.Case, R. E.Duke, R. C.Walker, N. R.Skrynnikov, T. E.Cheatham III, O.Mikhailovskii, C.Simmerling, Y.Xue, A.Roitberg, S. A.Izmailov et al., (2022).

[282] P.Raiteri, A.Laio, F. L.Gervasio, C.Micheletti and M.Parrinello, *The Journal of Physical Chemistry B* **110**, 8 (2006).

[283] I.Marcos-Alcalde, J.Setoain, J. I.Mendieta-Moreno, J.Mendieta and P.Gómez-Puertas, *Bioinformatics* **31**, 23 (2015).

[284] M.Biswas, B.Lickert and G.Stock, *The Journal of Physical Chemistry B* **122**, 21 (2018).

[285] M.Hoffmann, M.Scherer, T.Hempel, A.Mardt, B.Silvad., B. E.Husic, S.Klus, H.Wu, N.Kutz, S. L.Brunton et al., *Machine Learning: Science and Technology* **3**, 1 (2021).

[286] P.Metzner, C.Schütte and E.Vanden-Eijnden, *Multiscale Modeling & Simulation* **7**, 3 (2009).

[287] S.Falkner and N.Schwierz, *The Journal of Chemical Physics* **155**, 8 (2021).

[288] B. E.Douglas, H. A.Laitinen and J. C. J.Bailar, *Journal of the American Chemical Society* **72**, 6 (1950).

[289] M.Eigen, *Advances in the Chemistry of the Coordination Compounds (Proceedings of the Sixth International Conference on Coordination Chemistry)*, MacMillan, New York (1961).

[290] J.Burgess, *Ions in Solution*, Woodhead Publishing (1999).

[291] L. J.Kirschenbaum and K.Kustin, *J. Chem. Soc. A* (1970).

[292] R. W.Taylor, H. K.Stepien and D. B.Rorabacher, *Inorganic Chemistry* **13**, 6 (1974).

[293] M.Eigen and R. G.Wilkins, *The Kinetics and Mechanism of Formation of Metal Complexes*, (1965).

[294] D. T.Richens, *Chemical Reviews* **105**, 6 (2005).

[295] H.Taube, *Chemical Reviews* **50**, 1 (1952).

[296] D.Rorabacher, *Inorganic Chemistry* **5**, 11 (1966).

[297] R. M.Fuoss, *Journal of the American Chemical Society* **80**, 19 (1958).

[298] C. H.Langford and T. R.Stengle, *Annual Review of Physical Chemistry* **19**, 1 (1968).

[299] B.Rode, C.Schwenk, T.Hofer and B.Randolf, *Coordination Chemistry Reviews* **249**, 24 (2005).

[300] P. T.Snee, J.Shanoski and C. B.Harris, *Journal of the American Chemical Society* **127**, 4 (2005).

[301] S.Doudou, K.Arumugam, D. J.Vaughan, F. R.Livens and N. A.Burton, *Phys. Chem. Chem. Phys.* **13**, (2011).

[302] D.Rorabacher and C.Melendez-Cepeda, *Journal of the American Chemical Society* **93**, 23 (1971).

[303] T. S.Roche and R. G.Wilkins, *Journal of the American Chemical Society* **96**, 16 (1974).

[304] S. F.Lincoln and A. E.Merbach, *Advances in inorganic chemistry* **42**, ARTICLE (1995).

[305] R. G.Wilkins, *Substitution Reactions*, John Wiley & Sons, Ltd (1991).

[306] A. E.Merbach, *Kinetics of Solvent Exchange Reactions at High Pressure*, Springer Netherlands (1987).

[307] R. G.Wilkins, *Accounts of Chemical Research* **3**, 12 (1970).

[308] D. W.Margerum, D. B.Rorabacher and J. F. G.Clarke, *Inorganic Chemistry* **2**, 4 (1963).

[309] F. R.Shu and D. B.Rorabacher, *Inorganic Chemistry* **11**, 7 (1972).

[310] T. S.Turan and D. B.Rorabacher, *Inorganic Chemistry* **11**, 2 (1972).

[311] T. S.Turan, *Inorganic Chemistry* **13**, 7 (1974).

[312] A. E.Merbach, *Pure and Applied Chemistry* **54**, 8 (1982).

[313] F. A.Cotton and F. E.Harris, *The Journal of Physical Chemistry* **59**, 12 (1955).

[314] M. J.Wilkinson G, *Comprehensive Coordination Chemistry*, Pergamon Press (1987).

[315] A. W.Adamson, *Journal of the American Chemical Society* **76**, 6 (1954).

[316] Y.Takeda, K.Samejima, K.Nagano, M.Watanabe, H.Sugeta and Y.Kyogoku, *European Journal of Biochemistry* **130**, 2 (1983).

[317] L.Lomozik, L.Bolewski and R.Dworczak, *Journal of Coordination Chemistry* **41**, (1997).

[318] M.Kojima, K.Morita and J.Fujita, *Bull Chem Soc Jpn* (1981).

[319] G. B.Hares, W. C.Fernelius and B. E.Douglas, *Journal of the American Chemical Society* **78**, 9 (1956).

[320] R.Jastrzab, L.Lomozik and B.Tylkowski, *Physical Sciences Reviews* **1**, 6 (2016).

[321] Z. L.Tesic, T. J.Janjic, M. J.Malinar and M. B.Celap, *Journal of chromatography* (1989).

[322] L.Lomozik and R.Bregier-Jarzebowska, *Polish J Chem* 73 (2016).

[323] A.Raffo, L.Gagliardi, U.Fugacci, L.Sagresti, S.Grandinetti, G.Brancato, S.Biasotti and W.Rocchia, *Frontiers in Molecular Biosciences* **9**, (2022).

[324] H.Edelsbrunner, M.Facello and J.Liang, *Discrete Applied Mathematics* **88**, 1 (1998).

[325] S.Decherchi and W.Rocchia, *PLoS ONE* **8**, 4 (2013).

[326] S.Decherchi, A.Spitaleri, J.Stone and W.Rocchia, *Bioinformatics* **35**, 7 (2018).

[327] H.Edelsbrunner and E. P.Mücke, *ACM Trans. Graph.* **13**, 1 (1994).

[328] M.Piccinelli, A.Veneziani, D. A.Steinman, A.Remuzzi and L.Antiga, *IEEE Transactions on Medical Imaging* **28**, 8 (2009).

[329] B.Chandramouli, D.Di Maio, G.Mancini, V.Barone and G.Brancato, *PLOS ONE* **10**, 3 (2015).

[330] A. D.MacKerell, D.Bashford, M.Bellott, R. L.Dunbrack, J. D.Evanseck, M. J.Field, S.Fischer, J.Gao, H.Guo, S.Ha et al., *The Journal of Physical Chemistry B* **102**, 18 (1998).

[331] J. B.Klauda, R. M.Venable, J. A.Freites, J. W.O'Connor, D. J.Tobias, C.Mondragon-Ramirez, I.Vorobyov, A. D.MacKerell and R. W.Pastor, *The Journal of Physical Chemistry B* **114**, 23 (2010).

[332] T.Darden, D.York and L.Pedersen, *The Journal of Chemical Physics* **98**, 12 (1993).

[333] A.Koçer, M.Walko, W.Meijberg and B. L.Feringa, *Science* **309**, 5735 (2005).

[334] O. S.Smart, J. G.Neduvelil, X.Wang, B.Wallace and M. S.Sansom, *Journal of Molecular Graphics* **14**, 6 (1996).

[335] J. P.Birkner, B.Poolman and A.Koçer, *Proceedings of the National Academy of Sciences* **109**, 32 (2012).

[336] S.Decherchi, J.Colmenares, C. E.Catalano, M.Spagnuolo, E.Alexov and W.Rocchia, *Communications in Computational Physics* **13**, 1 (2013).

[337] M.Petřek, M.Otyepka, P.Banas, P.Košinová, J.Koča and J.Damborský, *BMC Bioinformatics* **7**, 316 (2006).

[338] M.Petřek, P.Košinová, J.Koča and M.Otyepka, *Structure* **15**, 11 (2007).

[339] E.Yaffe, D.Fishelovitch, H. J.Wolfson, D.Halperin and R.Nussinov, *Proteins: Structure, Function, and Bioinformatics* **73**, 1 (2008).

[340] D.Di Maio, B.Chandramouli and G.Brancato, *PLOS ONE* **10**, 10 (2015).

[341] C.Jorgensen, L.Darré, V.Oakes, R.Torella, D.Pryde and C.Domene, *Molecular Pharmaceutics* **13**, 7 (2016).

[342] S. G.Birnbaum, A. W.Varga, L.-L.Yuan, A. E.Anderson, J. D.Sweatt and L. A.Schrader, **84**, 3 .

[343] K. J.Rhodes, K. I.Carroll, M. A.Sung, L. C.Doliveira, M. M.Monaghan, S. L.Burke, B. W.Strassle, L.Buchwalder, M.Menegola, J.Cao et al., **24**, 36 .

[344] P.Strop, A. J.Bankovich, K. C.Hansen, K.Christopher Garcia and A. T.Brunger, **343**, 4 .

[345] K.Turnow, K.Metzner, D.Cotella, M. J.Morales, M.Schaefer, T.Christ, U.Ravens, E.Wettwer and S.Kämmerer, **110**, 2 .

[346] D.Anderson, W. H.Mehaffey, M.Iftinca, R.Rehak, J. D. T.Engbers, S.Hameed, G. W.Zamponi and R. W.Turner, **13**, 3 .

[347] T.Sacco and F.Tempia, **543**, 2 .

[348] Y.Wang, J. C.Strahlendorf and H. K.Strahlendorf, **567**, 1 .

[349] S. G.Cull-Candy, C. G.Marshall and D.Ogden, **414**, 1 .

[350] R.Bardoni and O.Belluzzi, **69**, 6 .

[351] J. R.Giudicessi, D.Ye, D. J.Tester, L.Crotti, A.Mugione, V. V.Nesterenko, R. M.Albertson, C.Antzelevitch, P. J.Schwartz and M. J.Ackerman, **8**, 7 .

[352] Y.-C.Lee, A.Durr, K.Majczenko, Y.-H.Huang, Y.-C.Liu, C.-C.Lien, P.-C.Tsai, Y.Ichikawa, J.Goto, M.-L.Monin et al., **72**, 6 .

[353] A.Duarri, J.Jezierska, M.Fokkens, M.Meijer, H. J.Schelhaas, W. F. A.Dunnend., F.Dijkv., C.Verschuuren-Bemelmans, G.Hageman, P.Vliesv. d. et al., **72**, 6 .

[354] K.Smets, A.Duarri, T.Deconinck, B.Ceulemans, B. P.Warrenburgv. d., S.Züchner, M. A.Gonzalez, R.Schüle, M.Synofzik, N.AaV. d. et al., **16**, 1 .

[355] M.Kurihara, H.Ishiura, T.Sasaki, J.Otsuka, T.Hayashi, Y.Terao, T.Matsukawa, J.Mitsui, J.Kaneko, K.Nishiyama et al., **17**, 2 .

[356] H. H.Jerng, P. J.Pfaffinger and M.Covarrubias, **27**, 4 .

[357] S. P.Patel and D. L.Campbell, **569**, 1 .

[358] R.Bähring and M.Covarrubias, **589**, 3 .

[359] J. D.Fineberg, T. G.Szanto, G.Panyi and M.Covarrubias, **6**, 1 .

[360] M. R.Skerritt and D. L.Campbell, **3**, 11 .

[361] A.Duarri, M.-C. A.Lin, M. R.Fokkens, M.Meijer, C. J. L. M.Smeets, E. A. R.Nibbeling, E.Boddeke, R. J.Sinke, H. H.Kampinga, D. M.Papazian et al., **72**, 17 .

[362] S. B.Long, X.Tao, E. B.Campbell and R.MacKinnon, *Nature* **450**, 7168 (2007).

[363] R.Chenna, H.Sugawara, T.Koike, R.Lopez, T. J.Gibson, D. G.Higgins and J. D.Thompson, *Nucleic Acids Research* **31**, 13 (2003).

[364] M.Larkin, G.Blackshields, N.Brown, R.Chenna, P.McGettigan, H.McWilliam, F.Valentin, I.Wallace, A.Wilm, R.Lopez et al., *Bioinformatics* **23**, 21 (2007).

[365] F.Jeanmougin, J. D.Thompson, M.Gouy, D. G.Higgins and T. J.Gibson, *Trends in Biochemical Sciences* **23**, 10 (1998).

[366] J. D.Thompson, T. J.Gibson, F.Plewniak, F.Jeanmougin and D. G.Higgins, *Nucleic Acids Research* **25**, 24 (1997).

[367] M.Biasini, S.Bienert, A.Waterhouse, K.Arnold, G.Studer, T.Schmidt, F.Kiefer, T. G.Cassarino, M.Bertoni, L.Bordoli et al., *Nucleic Acids Research* **42**, W1 (2014).

[368] S.Bienert, A.Waterhouse, T. A. P.Beerd., G.Tauriello, G.Studer, L.Bordoli and T.Schwede, *Nucleic Acids Research* **45**, D1 (2017).

[369] N.Guex, M. C.Peitsch and T.Schwede, *ELECTROPHORESIS* **30**, S1 (2009).

[370] P.Benkert, M.Biasini and T.Schwede, *Bioinformatics* **27**, 3 (2011).

[371] M.Bertoni, F.Kiefer, M.Biasini, L.Bordoli and T.Schwede, *Scientific Reports* **7**, 1 (2017).

[372] M.Pioletti, F.Findeisen, G. L.Hura and . D. L.Minor, *Nature Structural & Molecular Biology* **13**, 11 (2006).

[373] H.Wang, Y.Yan, Q.Liu, Y.Huang, Y.Shen, L.Chen, Y.Chen, Q.Yang, Q.Hao, K.Wang et al., *Nature Neuroscience* **10**, 1 (2007).

[374] M. O.Jensen, D. W.Borhani, K.Lindorff-Larsen, P.Maragakis, V.Jogini, M. P.Eastwood, R. O.Dror and D. E.Shaw, *Proceedings of the National Academy of Sciences* **107**, 13 (2010).

[375] M. O.Jensen, V.Jogini, D. W.Borhani, A. E.Leffler, R. O.Dror and D. E.Shaw, *Science* **336**, 6078 (2012).

[376] M. O.Jensen, V.Jogini, M. P.Eastwood and D. E.Shaw, *Journal of General Physiology* **141**, 5 (2013).

[377] J.-P.Ebejer, J. R.Hill, S.Kelm, J.Shi and C. M.Deane, *Nucleic Acids Research* **41**, W1 (2013).

[378] J. R.Hill and C. M.Deane, *Bioinformatics* **29**, 1 (2012).

[379] S.Kelm, J.Shi and C. M.Deane, *Bioinformatics* **26**, 22 (2010).

[380] S.Jo, T.Kim, V. G.Iyer and W.Im, *Journal of Computational Chemistry* **29**, 11 (2008).

[381] J. C.Phillips, R.Braun, W.Wang, J.Gumbart, E.Tajkhorshid, E.Villa, C.Chipot, R. D.Skeel, L.Kalé and K.Schulten, *Journal of Computational Chemistry* **26**, 16 (2005).

[382] A. W.H. Grubmüller and K.Schulten, *Molecular Simulation* **6**, 1-3 (1991).

[383] S. E.Feller, Y.Zhang, R. W.Pastor and B. R.Brooks, *The Journal of Chemical Physics* **103**, 11 (1995).

[384] G.Martyna, D.Tobias and M.Klein, *Journal of Chemical Physics* **101**, (1994).

[385] G.Starek, J. A.Freites, S.Bernèche and D. J.Tobias, *Journal of Computational Chemistry* **38**, 16 (2017).

[386] K.Kasahara, M.Shirota and K.Kinoshita, *PLOS ONE* **8**, 2 (2013).

[387] R.Storn and K.Price, *Journal of Global Optimization* **11**, 4 (1997).

[388] M. R.Skerritt and D. L.Campbell, *American Journal of Physiology-Cell Physiology* **293**, 3 (2007).

[389] C.Xie, V. E.Bondarenko, M. J.Morales and H. C.Strauss, *American Journal of Physiology-Cell Physiology* **297**, 5 (2009).

[390] H.Amiri, K. L.Shepard, C.Nuckolls and R.Hernández Sánchez, *Nano Letters* **17**, 2 (2017).

[391] L.Pontoriero, M.Schiavina, M. G.Murrali, R.Pierattelli and I. C.Felli, *Angewandte Chemie International Edition* **59**, 42 (2020).