UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

The role of electrostatics in the mechanism of ATP/ADP carrier function: an *in silico* study

Nuno Filipe Baltazar Maia Costa de Oliveira

MESTRADO EM BIOQUÍMICA
Especialização em Bioquímica

Dissertação orientada por:
Doutor Miguel Ângelo dos Santos Machuqueiro

2020

Acknowledgements

Firstly, I would like to acknowledge my supervisor, Dr. Miguel Machuqueiro, for his great amount of patience and for being a crucial pillar in my path through the vast molecular modeling world. I would like to express my unfathomable gratitude for all the efforts in introducing the many details of such an interesting research path and for his readiness and availability in answering my many questions. A huge acknowledgement is also owed to all the members of the theoretical (bio)chemistry research group for the amazingly helpful, friendly and welcoming environment they provided in both the lab and on the many lunch/coffee breaks I had the pleasure to partake.

A special thanks is due to the rookies in the group who started their thesis at the same time i did, for all the lunches, moments and hardships together during our learning time under the wings of Miguel Machuqueiro, Paulo Costa and Bruno Victor who had the patience of teaching us all. Although, for some of us rookies, we only met in early September, all the moments we have passed during this year were all memorable and made the master's experience much better, from the pouring rains of Coimbra shared with Filipe Rodrigues, Bernardo Henriques, Sara Ferreira, Andreia Fortuna, to all the fun breaks, relaxing or doing raids (or in Sara's case only playing Sudoku apparently) in their company and of Diogo Reis. I would like to also acknowledge all my friends who have accompanied me through the years, friends I met at the university, in the scouts and from the aikido practices, who made this period so much fun and for always extending their support.

A special acknowledgement is imperative to Bárbara Cazemajou for being my emotional support during last few years, and hopefully beyond that, for always being by my side and pushing me to work, nagging me when I would slack off. It would've not been the same without all our trips, adventures, stressing work hours and all our many moments together.

Finally, a crucial acknowledgement to the two persons who have made all this possible, my parents Manuel Oliveira and Regina Oliveira, for their unshakable support and for always striving the hardest to allow me to follow the path I choose in my life. For all the education they have given me and for always being there no matter what.

Preface

The present Master thesis was performed in the Molecular Modelling and Simulation group, integrated in the Biosystems and Integrative Sciences Institute of FCUL, and, although only one year was spent on the making of this work, it is the culmination of three years of continuous learning under the supervision of Dr. Miguel Machuqueiro. Since joining the MMS group in 2017, I was able to learn many details of computational techniques and integrating projects under development in the research group. With the help of my supervisor, who has a vast experience in computational simulation methods, I was able to apply techniques such as Molecular Dynamics, Constant-pH Molecular Dynamics and Umbrella Sampling not only on the present work, but also on past/parallel projects. During these 3 years, as suggested by Miguel, I've dedicated a good portion of my time accomplishing an improvement in the charge sets of phosphorylated amino acids in order to accurately describe their pH-binding phenomena on a concealed pocket. Such work was not only essential in solidifying my knowledge of the CpHMD technique and on electrostatic events, but it also culminated in my first poster presentation at EJIBCE-02 in 2018 and the results from this work were compiled in a Journal of Chemical Theory and Computation publication in mid August:

N. F. B. Oliveira, I. D. S. Pires, and M.Machuqueiro, "New phosphorylated amino acid parametrization to correctly reproduce their acid/base equilibria, including in protein binding events", *Proceedings of MOL2NET 2018, International Conference on Multidisciplinary Sciences, 4th Edition, MDPI*,(2018), pp. 6082.

N. F. B. Oliveira, I. D. S. Pires, and M. Machuqueiro. "Improved GROMOS 54A7 Charge Sets for Phosphorylated Tyr, Ser, and Thr to Deal with pH-Dependent Binding Phenomena". J. Chem. Theory Comput. 16, 10 (2020), pp. 6368–6376.

This work had a significant importance to the present thesis since it enabled us to employ a charge set for both ADP and ATP that integrated the improvement achieved in the charge set of terminal phosphates in order to correctly describe the electrostatic of both molecules inside the cavity of the protein.

In parallel with the Master Thesis work, I have been granted a Bachelors R&D Grant by FCT integrated in the project "As bases físicas da doença: O caso da Amiloidose Relacionada com a Diálise" (PTDC/FIS-OUT/28210/2017). This work consisted in the studying, through Molecular Dynamics, the stability of β-2 Microglobulin dimers obtained from a Monte-Carlo Ensemble Docking approach developed in the research group lead by Dr. Patrícia Faísca. This big opportunity has enabled me to test my knowledge in molecular dynamic simulations and allowed me to further develop my understanding on biomolecular events and forces acting on proteins. The grant is scheduled to end in January 2021 and throughout its duration it has resulted in a poster presentation work and lead to a currently being written scientific paper:

F. E. P. Rodrigues, N. F. B. Oliveira, D. Vila-Viçosa, P. F. N. Faísca, and M. Machuqueiro. "An MD-based protocol to assess the stability of β2M dimers obtained from MC-ED" 7º Encontro Jovens Investigadores de Biologia Computacional Estrutural, (2019).

Finally, using the ATP/ADP carrier protein studied in this thesis, a PypKa protocol as been demonstrated with the objective of exemplifying the use of this novel methodology, currently in active development in our group by P. B. S. Reis, to obtain fast and reliable pK_a values for residues in membrane proteins. In the future, this protocol will be published as a chapter in a book compiling several protocols for computational methodologies.

Abstract

To accurately describe the electrostatic interactions with atomistic detail is an exceedingly useful tool, since they play a major role in all biomolecular processes. In the present work, we introduced a new computational protocol based on Constant pH Molecular Dynamics (CpHMD) coupled to an Umbrella Sampling scheme (US-CpHMD). Using these techniques, we modeled the conformational changes of ATP/ADP carrier (AAC) coupled to the most relevant protonation events and even extracted the equilibrium energetics involved in the transport of ATP and ADP across the inner mitochondrial membrane. The transport activity of this protein is deeply connected with electrostatic interactions involved in the binding of the highly negative charged substrates.

CpHMD simulations were performed on pre-equilibrated apo-AAC:POPC systems, allowing the extraction of titration curves and pK_a values of several AAC residues. We were able to reproduce the reported pK_a shift of Lys22 from the water soluble pK_a value (10.4) to ~ 8, which makes this residue the responsible for the sensitivity of this protein to basic pH values, modulating the activity of AAC. Three acid aminoacids, Glu29, Asp134 and Asp231, also showed shifted pK_a values which revealed their role in establishing electrostatic interactions in the bottom of AAC cavity, forming the matrix salt-bridge network responsible for closing the C-state cavity.

Taking a step further, US-CpHMD was used to mimic the transport of substrates while capturing conformational and electrostatic effects. Clear differences between the import and export process were detected in our analysis which are connected to the distinct evolutionary pressures of each transport. The import has developed a larger positive electrostatic potential in the cavity and a clear selectivity towards ADP in order to efficiently capture this molecule. In addition, we observed a substrate-induced shift in the protonation of the acid residues present on the matrix salt-bridge network. This protonation triggers the conformational transition from C- to M-state by weakening the matrix salt-bridges, leading to an acceleration of the import. These characteristics were of high importance for a quick and efficient import process, which is crucial due to the low abundance of the ADP near the C-state cavity. In contrast, from these features only a slightly weaker positive potential was present on the export process. The absence of the other traits is tied with the high abundance of ATP molecules near the M-state cavity and its almost unneeded selectivity over ADP, which is caused by a significantly lower magnitude of evolutionary pressure exerted in this process. Hence the M-state did not develop the same traits as the C-state.

The entirety of results showed the success of employing this computational protocol in the analysis of mechanistic details at the atomic level, enabling the extraction of key electrostatic interactions and energetic profiles of the biomolecular process. The application of this protocol to less studied proteins and processes may prove highly advantageous, possibly aiding in the elucidation of their mechanism and key electrostatic details.

Keywords: Molecular Dynamics, CpHMD, Umbrella Sampling, Electrostatic forces, Transmembrane Transporters

Resumo

As forças electrostáticas têm um papel fulcral numa extensa panóplia de processos biomoleculares desde expressão de genes, resposta imune, condução de moléculas até centros catalíticos, acoplamento de subunidades, entre outros. A grande ubiquidade destas prende-se com a sua habilidade de actuarem a longas distâncias. Assim sendo, a capacidade de descrever as forças electrostáticas associadas aos diversos processos é uma ferramenta extremamente útil, porém, infelizmente, o seu estudo com detalhe atomístico é bastante difícil de realizar através de métodos experimentais. Desta forma os métodos computacionais têm vindo a ganhar relevância tanto pelo grande aumento de poder computacional nas últimas décadas, como também pela sua capacidade de colmatar esta lacuna a nível experimental.

Com isto em mente, o presente trabalho pretende demonstrar um protocolo computacional efectivo no estudo das propriedades electrostáticas de sistemas moleculares de grandes dimensões e complexidade. O objecto de estudo escolhido para a validação deste protocolo foi a proteína transportadora de ATP/ADP (AAC), responsável por importar o ADP para a matriz mitocondrial e exportar ATP da matriz para o espaço inter-membranar do mitocôndrio. Sendo capaz de atrair e transportar moléculas altamente negativas aponta imediatamente para a presença de uma superfície electrostática complexa e essencial para que o processo ocorra eficientemente. A somar às características mencionadas anteriormente, o facto deste transportador estar bem caracterizado e estudado torna-o um bom candidato a usar na demonstração de um protocolo de estudo da electrostática associada a processos biomoleculares.

Para concretizar este estudo três técnicas computacionais foram utilizadas funcionando em complementaridade. Primeiramente simulações longas de dinâmica molecular foram realizadas, após a inserção do transportador de ATP/ADP em membranas de POPC. Com estas torna-se possível avaliar a estabilidade do nosso sistema proteína-membrana, verificando a viabilidade das condições escolhidas. Efectivamente foi observado que o uso do método de campo de força generalizado para o tratamento da electrostática a longa distância introduziu instabilidades estruturais nas simulações. Assim, foi necessário readaptar o nosso protocolo aplicando o método de somas de Ewald com malha de partículas para o tratamento da electrostática a longa distância.

Procedendo a esta alteração no tratamento das nossas simulações torna-se possível equilibrar o nosso sistema de modo a ser empregue na técnica de Dinâmica Molecular a pH Constante (CpHMD). Esta acrescenta um nível de detalhe à descrição do sistema ao permitir a ocorrência de eventos de (des)protonação numa lista definida de moléculas e/ou resíduos de aminoácido. Realizando simulações de CpHMD em ambos os estados, C e M, do AAC em forma apo e a quatro valores de pH, 4, 5, 6 e 7, torna-se possível obter curvas de titulação e valores de pK_a para diversos aminoácidos presentes no transportador de membrana. Analisando a presença de desvios nos pK_a detectados nas nossas simulações de CpHMD torna-se possível a discussão acerca do seu ambiente circundante e de possíveis interacções electrostáticas entre resíduos da proteína. Destas simulações fomos capazes de identificar um grande desvio no pK_a da lisina 22, trazendo o seu valor de 10.4, quando solvatado, para ~ 8. Este desvio fora já reportado em publicações anteriores e, à semelhança do nosso trabalho, demonstra-se relevante para a

actividade do transportador, uma vez que a forma protonada deste resíduo é essencial à actividade. Das nossas análises foi também possível identificar três resíduos ácidos na proteína cujo pK_a se encontrava desviado para valores inferiores no estado C do AAC. Este desvio indica-nos a presença de interacções electrostáticas que estabilizam o estado carregado dos resíduos glutamato 29, aspartato 134 e aspartato 231. Efectivamente estas interacções consistem na formação de pontes salinas que ajudam no fecho da cavidade do AAC no estado C. Com estes resultados é possível demonstrar a utilidade do estudo de uma proteína através da técnica de CpHMD, sendo possível identificar resíduos chave para a actividade do transportador e interações cruciais para o bom funcionamento da proteína.

Finalmente, pretendemos não só estudar a forma apo do AAC mas também o processo inteiro de transporte dos substratos através da proteína transmembranar com o intuito de obter uma descrição energética do processo e em simultâneo captar informações acerca das forças electrostáticas envolvidas neste. No entanto, as simulações tanto de MD como de CpHMD apenas são viáveis em pequenas escalas temporais, à volta de poucos microssegundos, já o tempo total do processo de transporte poderá chegar aos milissegundos. De forma a contornar este problema, será usada uma técnica de amostragem aumentada, Umbrella Sampling (US). Nesta, o processo de transporte é dividido em várias janelas segundo uma coordenada de reação e um potencial de enviesamento é aplicado ao substrato para o manter na posição definida. Desta forma, somos capazes de extrair a descrição energética do processo e os perfis de protonação não só do substrato como também dos resíduos do transportador de ATP/ADP. Conjugando os resultados dos dois tipos de transporte, importação de substrato para a matriz e exportação de substrato da matriz, para cada substrato, ATP e ADP, é possível comparar a totalidade de condições de transporte e esclarecer detalhes acerca do mecanismo executado pela proteína.

Com as nossas simulações foi possível concluir que o mecanismo de importação dos substratos beneficia de diversos detalhes que não se encontram presentes no processo de exportação. Em primeira instância, através dos resultados obtidos da força potencial média (PMF) e da protonação de ambos os substratos, identificou-se a presença de um maior potencial electrostático positivo na cavidade do estado C, responsável pela captação de substrato no processo de importação. Este potencial conseguirá atrair os substratos do citoplasma para a cavidade a maiores distâncias do que o estado M. Aquando da comparação dos perfis energéticos de importação, foi observada a preferência energética da importação de ADP pelo seu perfil energético inferior relativamente ao do ATP. Já no caso da exportação, ambos apresentam perfis bastante semelhantes, não havendo um favorecimento do transporte de um substrato em particular. Por fim, os dois processos de transporte distinguem-se entre si também ao nível das redes de pontes salinas que executam o fecho da cavidade do estado C e M. Considerando o processo de importação, onde o estado C é responsável pela atração das moéculas, verificou-se a existência da rede de pontes salinas directamente no fundo da cavidade. Esta é constituída pelos resíduos ácidos mencionados nos resultados do CpHMD, Glu29, Asp134 e Asp231. Foi verificado que, ao ocorrer a aproximação do substrato ao fundo da cavidade do estado C, estes ácidos têm tendência a protonarem, o que leva, por consequência, ao enfraquecimento das pontes salinas e a uma mais rápida transição conformacional do estado C para o estado M. Em contrapartida, no processo de exportação, onde o estado M é responsável pela atracção dos substratos, a localização das pontes salinas encontra-se na extremidade citoplasmática da proteína, sendo assim insensível à presença do substrato no fundo da cavidade. Estas características distintas entre o estado C e M, relevantes para o processo de importação e exportação respectivamente, desenham um panorama que evidencia maior dificuldade na importação de substratos. *In vivo*, este processo é realizado ao atrair o ADP de um meio vasto e escasso em substrato junto da abertura do AAC para o interior da cavidade do estado C. Desta forma, a existência de um maior potencial positivo na cavidade, capaz de captar efectivamente o substrato, a presença de uma armadilha energética que previne a libertação

natural do ADP de volta para o meio envolvente e uma mais rápida transição conformacional devido ao enfraquecimento das pontes salinas, aquando da aproximação do substrato, são características essenciais para que o processo de importação se dê com a eficiência e rapidez necessárias. Em contraste, no processo de exportação, o estado M atrai o ATP da matriz mitocôndrial, cujo volume é mais reduzido e onde existe uma grande abundância de substrato. Desta forma este estado não é submetido à pressão do ambiente para evoluir no sentido de aumentar a rapidez e eficiência do transporte, reflectindo-se na ausência das características observadas no estado C.

Concluindo este trabalho, foi-nos possível demonstrar como o uso de técnicas computacionais permite um estudo detalhado e intensivo de um sistema de grandes porporções e complexidade, fornecendo informações cruciais acerca das forças electrostáticas e obtendo perfis energéticos que auxiliam na construção do mecanismo de acção da proteína. No presente caso, o protocolo foi usado numa proteína bem estudada e muito do seu mecanismo já fora elucidado. Porém, a aplicação deste mesmo protocolo numa proteína menos estudada poderá vir a revelar importantes detalhes sobre o seu mecanismo e levar à identificação de detalhes estruturais essenciais à actividade da proteína.

Palavras-chave: Dinâmica Molecular, CpHMD, Umbrella Sampling, Força Electrostática, Transportadores de Membrana

Index

List of Figures

List of Tables

List of Abbreviations

AAC	ATP/ADP Carrier
ADP	Adenosine diphosphate
Arg	Arginine
Asp	Aspartic acid
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
CE	Continuum electrostatics
CpHMD	Constant-pH molecular dynamics
DOPE	Discrete optimized protein energy
Glu	Glutamic acid
GRF	Generalized reaction field
His	Histidine
HMM	Hidden markov models
l-BFGS	limited memory - Broyden–Fletcher–Goldfarb–Shanno
LPBE	Linearized Poisson-Boltzmann equation
Lys	Lysine
MC	Monte Carlo
MCF	Mitochondrial carrier family
MD	Molecular dynamics
MM	Molecular mechanics
NAD	Nicotinamide adenine dinucleotide
PB	Poisson-Boltzmann
pbc	Periodic boundary conditions
PDB	Protein data bank
PEF	Potential energy function
PME	Particle mesh ewald
PMF	Potential mean force
POPC	1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine
PTYR	Phosphorylated tyrosine
RMSD	Root mean square deviation
SMD	Steered molecular dynamics
SPC	Simple point charge
US	Umbrella sampling
US-CpHMD	Umbrella sampling - Constant-pH molecular dynamics
WHAM	Weighted histogram analysis method

Chapter 1

Introduction

1.1 The Role of Electrostatics in Protein Function

A living cell is a highly complex and dynamic entity that depends on thousands of molecular interactions to execute its functions and stay healthy. Changing the interplay between molecules can have a huge impact on a biological system which, in the worst case, can result in disease. Therefore, understanding the forces that drive such interactions is extremely important to fully understand the mechanisms and events occurring in a living organism.

From the vast molecules that exist within a cell, proteins are one of the most interesting groups to study due to their ubiquitous presence in living organisms and high importance on biological processes. They are involved in gene expression, metabolite conversion, immune response to pathogens, and many more [1]. To perform such roles, it is common for a protein to recognize their interaction partners, even amongst millions of other molecules [2]. This process is usually fast, which points to the existence of specific forces that guide the protein to their partners [3]. From all the forces interacting on proteins and other molecules, the electrostatics forces are the best candidate to drive molecular interactions [4, 5]. These forces are able to act in long distances, hence being able to play a role in the guidance of the partner to the protein where the interaction will occur [3].

The electrostatic forces are not only involved in guiding ligands to proteins, but they are also a key factor on the strength of binding, where it has been noticed that the shape of the ligand is usually enhanced by the complementarity in its electrostatic surface [6]. Events that involve charge changes have a significant impact on the electrostatic potential. A good example is the protein (de)phosphorylation as an activating/inactivating mechanism, which is essential for several critical signalling pathways [7]. A good description of the electrostatic forces acting on a protein can therefore be crucial for the enlightenment of the protein's specific partners and functions [3, 8, 9].

Taking a deeper look at the essence of electrostatic interactions acting on molecules, we conclude that these are forces directly tied with the dipoles and charges present on the molecular structure. For proteins, this can be attributed to to the presence of polar or charged amino acids. There are mainly 20 different amino acids as the building blocks of proteins in living organisms [1] and their distinct side-chains provide different properties. A very important property in amino acids for electrostatics is the presence of ionizable side chains at given pH values. The protein electrostatic surface can actually be modulated by changes in pH, which lead to protonation events that can ionize or neutralize the amino acid side-chains [8, 10, 11]. These changes in the protein electrostatic surface at different pH values can be a powerful driving force for the (in)stability of proteins and also responsible for the modulation of enzyme catalytic activities [12–15].

Being able to describe, with atomistic detail, the electrostatic forces that act on protein interactions and how they are connected to the solution pH is a powerful tool that can enable the understanding of a vast number of biomolecular processes [3, 8, 9]. Identifying the ionizable residues that are essential in binding processes brings knowledge on the protein mechanism of action, characterize possible imbalances connected to disease, and help pinpointing drug targeting sites [16].

1.2 Transmembrane Transporters and Solute Transport

Transmembrane transporters are a class of proteins responsible for facilitating the motion of molecules across biological membranes. Their functions are diverse and of high importance to the cell, being responsible for the transport of nutrients, regulation of metabolites/ions, removal of toxins and many other functions [17]. Consequently, the vast number of molecules that are translocated by these proteins have varied chemical compositions.

The lipid bilayer is a selectively permeable barrier and the crossing of some molecules is impaired, mostly due to their charge/polarity and/or molecule size [1]. The existence of transmembrane proteins that are able to translocate these substrates across the impermeable membranes is crucial for the well being of the cell. The transport in cells that is mediated by proteins can be divided in passive transport, where concentration gradients are the only force driving the diffusion of the molecules trough the membrane, and in active transport, where the transport of the molecules is done against their concentration gradient by expending energy [1]. If the main source of energy for active transport is provided by a chemical reaction, such as the conversion of ATP in ADP, this is called a primary active transport, if otherwise the energy comes from the transport of another molecule this is called secondary active transport [1]. Since a great number of molecules that cannot naturally diffuse the membrane are charged, the electrostatic interactions with the transmembrane proteins responsible for their transport becomes an extremely relevant property to study.

The transmembrane transporter proteins are usually integral proteins, meaning they are completely inserted into the lipid bilayer. Their insertion in the membrane heterogeneous environment greatly impacts the distribution of residues on its surface. Normally the regions of the protein in direct contact with the lipid hydrocarbon tails are populated by non-polar side chains, while the charged and polar residues are favoured near the headgroup region, interacting with carbonyl, phosphate and choline groups [18]. In this case, the electrostatic surface is not directly tied with the activity but instead is connected with the stability of the protein in the membrane environment.

Electrostatic interactions play a crucial role in the transmembrane protein functions. The presence of charged residues at the channel entrance of these proteins can be responsible for substrate guidance, strongly counteracting the entropic penalty associated with this step [1, 19]. Inside the transporter, the solute can establish stabilizing interactions with appropriate charged residues in the cavity, which can greatly accelerate the molecule diffusion rate across the channel. These electrostatic interactions can also modulate which solute can interact with each channel. The presence of certain amino acid side chains can generate an electrostatic environment that confers specificity to the channel and only allowing a substrate, or class of substrates, to interact with the protein channel and, therefore, cross it [19–21].

The channel cavity opening is another key process of the solute transport across membranes where the electrostatic interactions are the main modulators. A significant amount of transmembrane transporters are not simple channels, with both ends opened simultaneously. In fact, a lot of them are required to be selective hence possessing at least one end closed, allowing only alternative opening [1, 17]. When the substrate binds deep in the open end pocket, it induces conformational changes that trigger the closing

of the open side and the opening of the closed side, allowing the substrate to diffuse [19]. Due to their transient state, the gates that open/close the protein channel are commonly composed of ionic and hydrogen bonds, allowing for a fast and energetically efficient break and reforming process. The specific role of these electrostatic interactions highlights their importance in living cells.

Having presented all these properties of transmembrane transporter that are dependent on electrostatic effects, it is no surprise that changes to the electrostatic surface of these proteins can have a large impact on their function. Mutations of polar and charged residues can have a strong impact on the protein binding sites, changing their ability to recognize the substrate, hence, to execute their functions [22, 23].

A good description of the electrostatics involved in solute transport by a transmembrane protein, which includes a detailed analysis of protonation changes occurring withing the cavity, can help elucidate its mechanism at the atomic level. It can reveal key residues involved in substrate recognition or involved in the opening and closing of transporters. With this information, we can have a better understanding of many diseases caused by mutations in key residues of the protein cavity and help improve the development of treatments for such pathologies.

1.3 Computational Methods in the Study of Biomolecules

Studying biochemical systems and their properties can be highly challenging via macroscopical and often indirect methods. Detailing properties at a single-molecule level is an arduous and costly task for most experimental approaches, which leveraged in recent years the application of computational methods to study these complex systems [24]. The rapid increase of computational power in the last 5 decades has allowed the steady growth of the computational chemistry/biochemistry fields [25–27]. By employing a model of the real world, simulations can describe properties of the chosen system with different levels of detail. Typically, computer simulation studies consist of 3 major steps: the choice of the model to describe the system, where the description obtained from the model must be validated with experimental data; the calculation itself; and finally, the calculation must be analysed to extract the results and to perform a type of verification/validation procedure [26]. If all these steps are successful, computational methods can provide important results on several biomolecular processes, such as perturbations on structure due to residue change, ligand binding effects, protein folding, amongst others. [24] The ability to obtain atomic level analyses of these processes is extremely useful in guiding experimental works like drug development, structure resolution, mechanistic study of processes, etc.

The application of computational methods has three main limiting factors, the level of detail, the size of the system, and the length of the simulation. All these share a common dependence of the hardware available. With more powerful equipment, the bigger, longer and more detailed the system can be, otherwise simulations become excessively time costly. The availability of reliable models to describe the systems of interest can also limit the applicability of such methods. If no models have yet been established to study given properties on certain conditions, it becomes unfeasible to use computational simulations to describe such processes. Finally, there is also a need for a good initial structure, since in order to performs simulations of a given system (e.g. proteins), there must be some structural information available. This can be fulfilled by direct structural resolution of the molecules (NMR, X-ray, etc...) or indirectly, via homology modeling methods, to obtain a structural model that is approximated to an existing template.

A great number of techniques have been developed to study many system properties with different detail levels. Quantum mechanics can be used to simulate systems with electronic detail, but only small systems are viable to be addressed [25, 26]. Molecular Mechanics associated with Molecular Dynamics

take advantage of the Born-Oppenheimer approximation [25–27] to tackle larger systems, their motions and study the associated thermodynamic properties. More complex techniques can be used to further focus on different types of properties and details.

The main objective of this thesis is to use molecular dynamics simulations coupled to Poisson-Boltzmann/Monte Carlo calculations to describe the conformational and protonation spaces of our system of interest and study the pH/protonation effects on the conformation/function of the protein. One of the best methods to achieve this is the Constant-pH Molecular Dynamics method, originally developed in António Baptista's Lab [28, 29] and currently also under developement at the Machuqueiro Lab [30, 31]. This method can combine the use of Molecular Dynamics, where the motion of all atoms in the simulation are calculated over time using the Newtonian equations of motion [24, 25, 27] with Poisson-Boltzmann and Monte Carlo calculations that allow protonation changes to occur on a set pH value. This technique opens the possibility of studying the effects of pH in the conformational space of proteins, providing useful information about their pH-dependent electrostatic surfaces, and pinpoint pH-sensitive residues that can be important by modulating the protein function.

We are interested in the transport of charged molecules across transmembrane transporters and how this process can be mediated by electrostatic interactions. For this, a higher level of complexity, regarding the timescale of the process, is added to the problem, which cannot be solved simply by using CpHMD simulations. Typical MD simulations work on the level of several nanoseconds to few microseconds. However, molecular transport through membranes occurs in the several microseconds to millisecond range. Such difference results in the impossibility to only use MD methods to describe the mechanism. A workaround for the problem is to use enhanced sampling techniques, which allow the study of several states of our system in distinct points of the process of interest in a way that all states combined provide a description of the entire process. The basis for these simulations is the application of a bias potential on one or more degree of freedom of the system. Examples of this type of treatment are umbrella sampling (US) simulations, conformational flooding, adaptive force bias and steered MD.[32] Using these methods the sampling is improved allowing the study of processes that would take extremely long MD simulations by discretizing the biomolecular process and executing MD simulations on each state.[32] Finally, these methods can reconstruct the free energy of the process, using the independent free energies of each state of the process. [32–34] Performing a combination of the potential bias with CpHMD simulations provides a very powerful technique which is able to address structural properties along the slow biomolecular process and still capture the pH effects on the electrostatics environment of our system.

1.4 The ATP/ADP Carrier as a Transmembrane Protein Model System

As discussed previously, the accurate description of the electrostatic interactions between the solute and the transporter can be essential to capture the process details. Using computational methods, we capture the complete picture of protonation changes that occur during the substrate transport. To proceed, we selected a transmembrane protein with a significant pH-dependence on its activity and with strong structural data already available. The ATP/ADP carrier (AAC) fitted these parameters adequately, since it is a crucial transmembrane protein involved in the transport of highly charged substrates (ATP and ADP) and it was discovered in the early sixties [35–37], possessing long years of accumulated knowledge. Another important requirement fulfilled by this protein was the existence of experimentally resolved structures. The AAC possesses two different states (which will be debated below), with several high quality structures available for one of the states (C-state).[38] Furthermore, for the second state of the

protein (M-state) a recent structure has been solved for a thermophilic fungus organism.[39] Although this is the only structure available for the M-state of AAC, it opens the possibility of using homology modelling to obtain a reliable structure of the mammals M-state AAC.

1.4.1 AAC Structural Details

The AAC is a transmembrane protein that is abundantly present in the inner mitochondria membrane [35, 39, 40]. This protein belongs to the mitochondrial carrier family (MCF) and it functions as a monomeric protein responsible for the exchange of both ATP and ADP across the impermeable inner mitochondria membrane.[39–41] The transport of these two substrates through the inner mitochondria membrane is crucial. ATP is the main source of chemical energy of the cell and it is produced from ADP in the matrix of the mitochondria, hence the maintenance of these substrates concentration is essential for the cell. [1, 39, 40, 42] Nevertheless, this is not an easy feat due to the impermeability of the inner mitochondria membrane. Unlike the outer mitochondria membrane, which is permeable to molecules with mass inferior to 4-5 kDa, the inner mitochondria membrane is highly impermeable to large and/or polar solutes, requiring proteins of the MCF to perform their transport. [40–42]

AAC has two distinct interconvertible states: a C-state conformation where the cavity is opened to the mitochondrial intermembranar space (the cytoplasmic side) and an M-state conformation, whose cavity is opened to the mitochondria matrix side (the matrix side) [42]. The *in vivo* C-state is usually responsible for the attraction of ADP to the bottom of its cavity. Once there, it undergoes a conformational change to the M-state, releasing the ADP to the mitochondrial matrix in favour of the concentration gradient (Figure 1.1A). On the other hand, the M-state can attract the ATP present in higher concentrations in the mitochondrial matrix, which upon reaching the bottom of the M-state cavity, can cross to the intermembrane space via a conformational shift to the C-state and consequent release (Figure 1.1B). The crossing of ATP to the mitochondria exterior takes advantage of both the concentration gradient and the membrane potential which favours the process. To study experimentally each state, inhibitors that can lock the conformation of the protein have been found. The AAC C-state is commonly blocked by atractylosides [43–45], while the M-state is blocked by bongkrekic acid [46, 47].

In terms of structure organization, AAC shares the common arrangement of the mitochondrial carrier family which is highly conserved in eukaryotes. This protein is composed of three homologous domains composed by two α-helices, summing to a total of 6 transmembrane helices. [38–40, 42, 48] The *Bos Taurus* AAC structure has been well-resolved and it shares 90% sequence similarity with the Human protein.[38] Due to their similarities and the overall conservation of structure of AAC in eukaryotes, the study of *Bos Taurus* AAC will give important information for a large amount of species, specially Humans. Looking specifically at *Bos Taurus* AAC (Figure 1.2), each homologous domain is composed by an odd numbered helix (H1, H3, H5), a loop with a short α-helix located in the matrix side (h1-2,h3-4,h5-6) and an even numbered helix (H2,H3,H4).

The odd-numbered helices possess a sequence motif that is conserved in all MCF, P-X-D/E-X-X-K/R-X-K/R and D/E-G-X-X-X-X-W/Y/F-K/R-G usually separated by 20 to 30 amino acid residues. The presence of the proline residue in this conserved sequence confers a characteristic kink to these helices which is responsible by the formation of the conic cavity. The charged residues present on these sequence motifs are also involved in the formation of inter-domain salt bridges that block the access to the cavity from the mitochondrial matrix, being named matrix salt-bridge network.[39, 48] Across species, AAC possesses another highly conserved sequence motif located in the even-numbered helices, F/Y-D/E-X-X-K/R. Similarly to the previous motif, the charged residues involved in this sequence have been

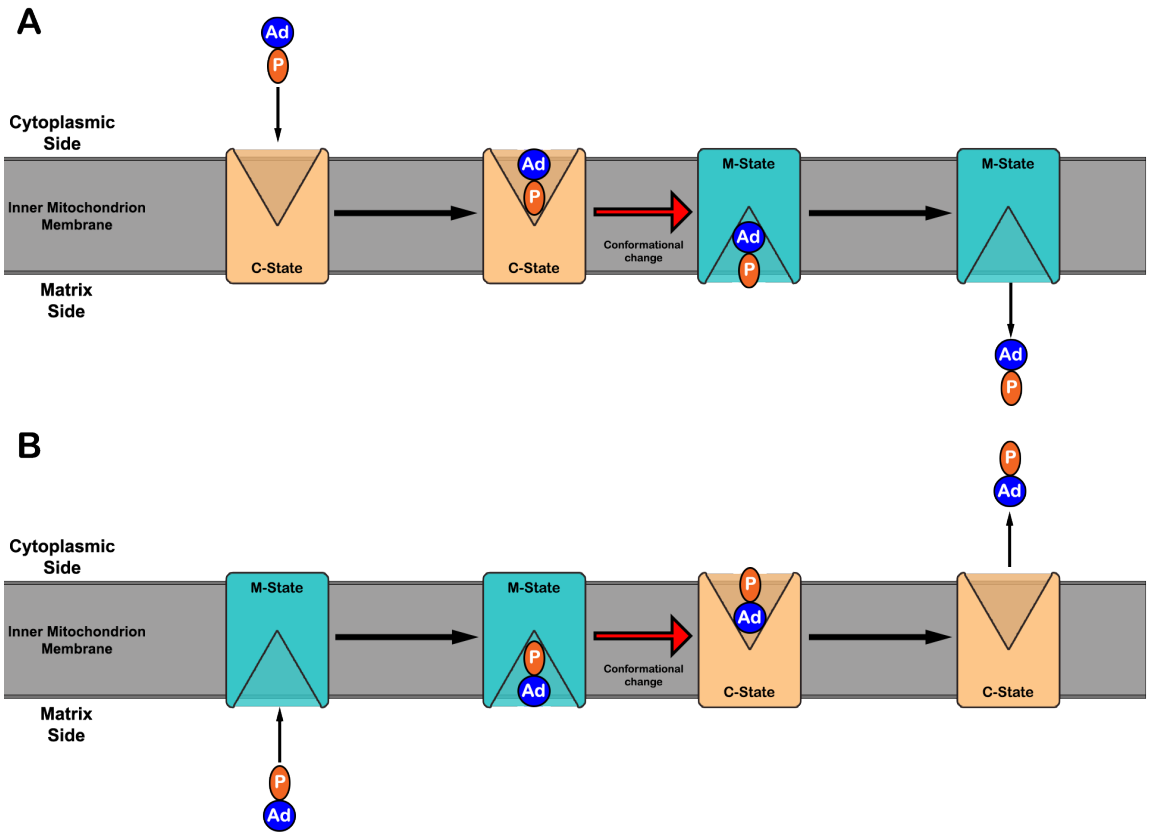

Figure 1.1: Schematic representation of substrate import from the cytoplasmic side to the mitochondrial matrix (A) and substrate export from the mitochondrial matrix back to the cytoplasmic side (B). The inner mitochondrial membrane is represented by a grey rectangular area, AAC's C-state is represented in light orange with the cavity opening pointed to the cytoplasmic side of the membrane, while the M-state is represented in light blue with the cavity opening pointed to the matrix side. The substrate is shown as a blue sphere representing the adenosine region and an orange ellipse representing the phosphate region.

proposed to form salt-bridge interactions that close the access to the cavity from intermembrane space, being named cytoplasmic salt-bridge network. [39, 42, 48, 49] There is still another sequence motif that has been regarded as an AAC signature, RRRMMM.[40] Mutations on the arginines that compose this consensus motif abolish the protein's activity, hinting for their involvement in substrate recognition and transport. [38]

The final structural detail to be pointed out about this protein relates to the lipids that surround it. Several crystal structures have pointed to the existence of 3 cardiolipin lipids that surround the protein on the inner leaflet of the membrane and are bound to the small grooves originated by the change from the even numbered helices to the matrix helices (h1-2,h3-4,h5-6).[41, 42] The presence of these particular lipids on this groove of the protein has been proposed to stabilize the protein structure due to the presence of a helix-to-loop transition between the major helices and the matrix small helices. [41, 42]

1.4.2 The Role of Electrostatics in AAC Function

AAC is responsible for the transport of the highly negative charged substrates ATP and ADP, which, at physiological pH, exhibit a charge of $-4\,e$ and $-3\,e$, respectively. There are evidences that this protein transports the substrates without the presence of divalent cations (Ca^{2+} and Mg^{2+}), which are normally complexed with ATP and ADP, partially attenuating their charge. [41, 50] Hence, the AAC cavity deals with the fully charged substrates during the transport, meaning that it most likely is abundant in positive

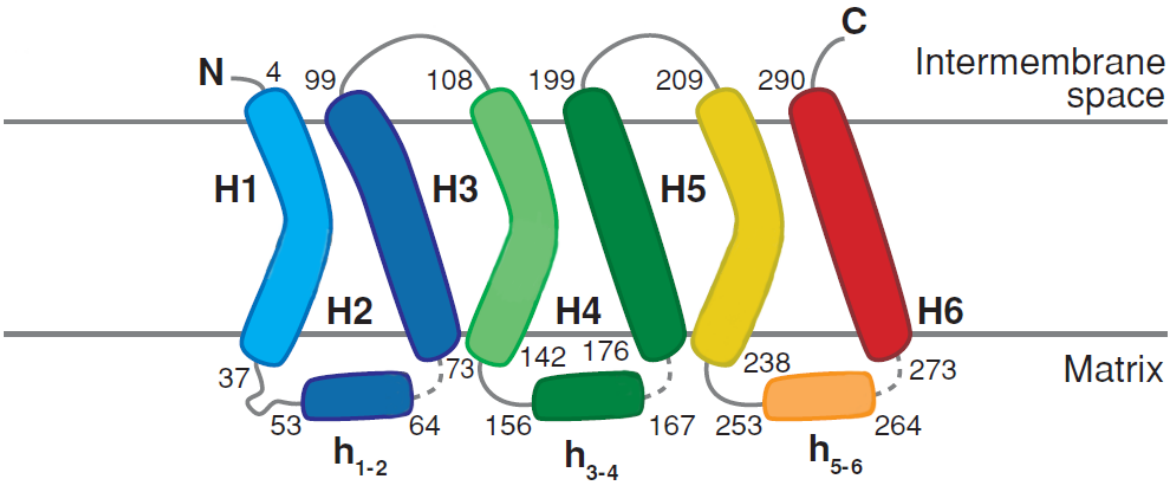

Figure 1.2: Representation of AAC secondary structure arrangement and domains in the membrane. Figure adapted from Nury et al. [41].

residues. In fact, the electrostatic surface of the AAC cavity is highly positively charged (Figure 1.3), which is extremely beneficial to attract the negative charged substrates [40, 41, 51]. Most data regarding the binding of the substrates to the cavity has been obtained with the C-state. Results for the M-state are scarce probably also due to the lack of structural information until recently. It has been found that AAC has a strict selectivity for ADP, ATP and their respective deoxy variants [40–42]. Surprisingly, even AMP, which is only 1 phosphate short from the structure of ADP, has a significantly lower affinity. Due to the similarity in their structures, the absence of transport for AMP must be hindered by other factors, including electrostatic interactions.

Looking at the C-state, the binding of ADP is mainly performed by two different regions. A basic patch at the bottom of the cavity is responsible for the electrostatic interactions with the phosphate groups of the substrate, the elements of this patch are highly conserved amongst AACs from different species and mutations on them impairs the transport [41]. Some of the residues involved in this basic patch are Lys22 and Arg79, 187, 234, 235 and 279 (Figure 1.4 A) [38, 41]. This consensus binding site can neutralize the charge of ADP upon binding, which is highly advantageous since the import of ADP to the mitochondrial matrix is done against the membrane potential. The binding of ATP to this site may not be able to completely neutralize this substrate charge, hence impairing its transport and hinting at the specificity of the C-state to ADP [39, 49]. The second region is related with the adenine moiety present in these molecules. ATP and ADP have an adenine ring in their structure that interacts in the pocket comprised of Y186, G182, I183 and S227 [38, 41, 42, 48]. In this region, the tyrosine may be the most important residue since it is a part of a tyrosine ladder (Figure 1.4 B), that can help the adenine ring to slide upon and enter the cavity [41]. The aromatic ring stacking between the adenine group and the tyrosine side chains may be the source of the AAC selectivity to the adenine nucleotides. This tyrosine ladder has also been observed in other adenosine nucleotide transporters, highlighting its importance on the guidance of these substrates through the cavity [41]. The two mentioned binding regions help understand the lower binding affinities observed for AMP. Besides being less negative, this substrate is also shorter, hence, it cannot fully bind to both regions of the binding site simultaneously [42].

It is clear that electrostatics play an important role in the binding of these substrates to AAC. Mutations to some of these charged residues can result in transport impairment and to disease. From the indicated residues, Lys22 has been pinpointed as being pH-sensitive, implying that changes in the pH

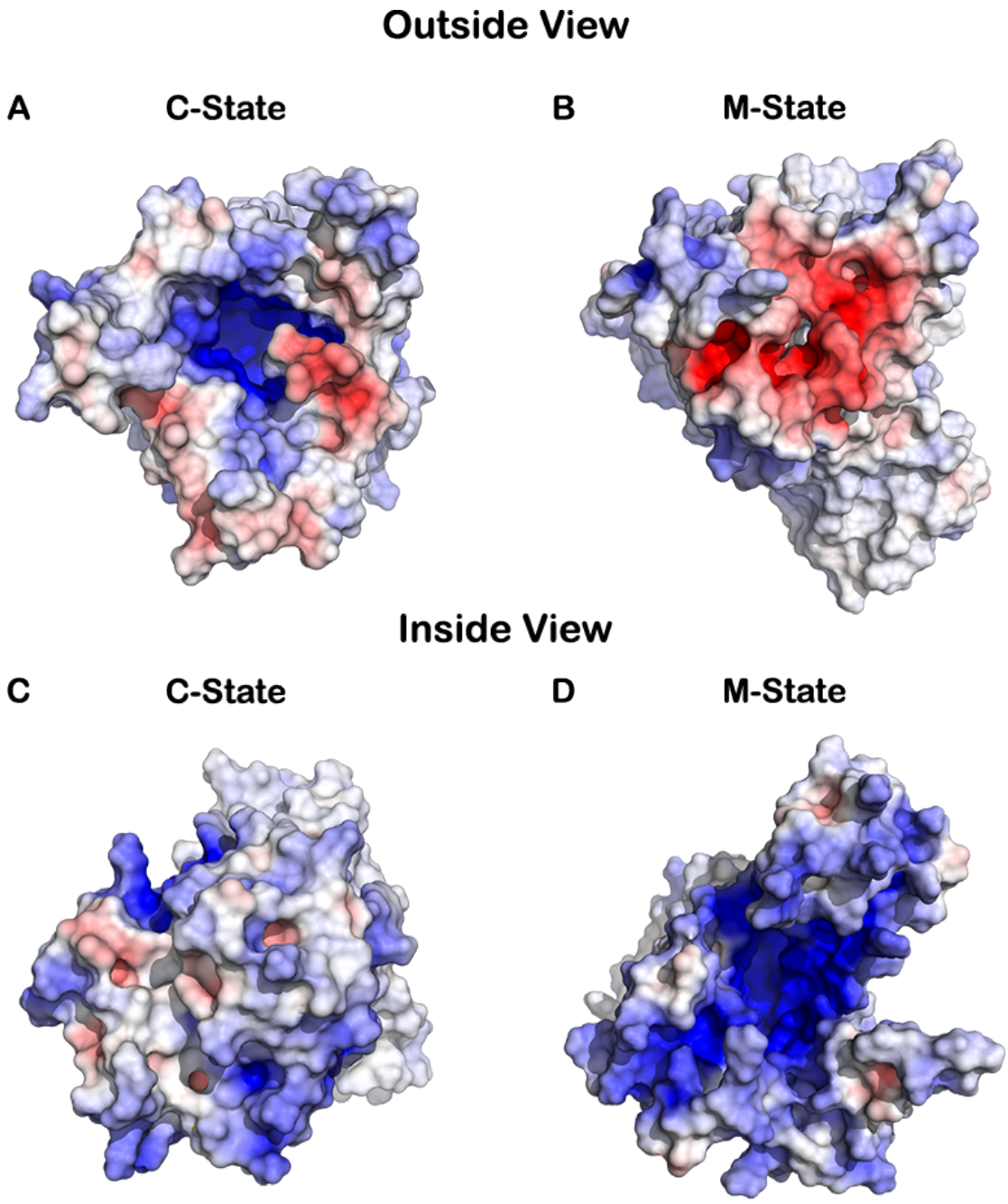

Figure 1.3: Electrostatic surface of AAC in both C-state (A and C) and M-state (B and D), viewed from outside the mitochondrial matrix (A and B) and from inside the mitochondrial matrix (C and D). Figures were rendered with PYMOL and the electrostatic surface was calculated using the APBS plugin (default parameters). A gradient of colors was used to identify the electrostatic charged regions ranging from positive (colored in blue) to negative (colored in red).

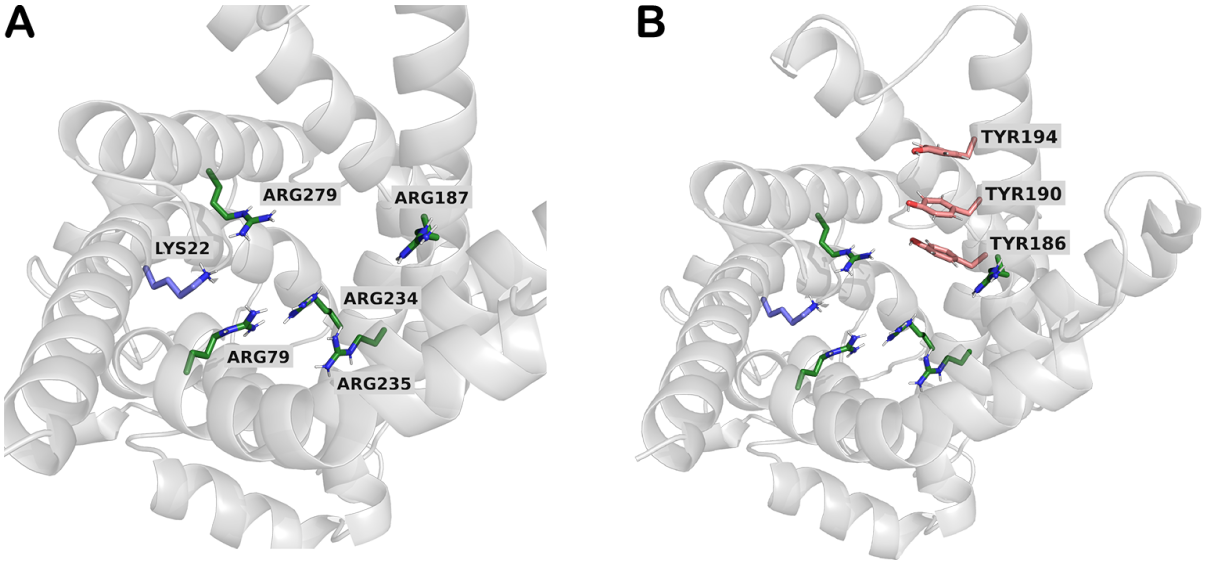

Figure 1.4: Close up view to the cationic amino acids (A) and tyrosine ladder (B) in the C-state cavity composing the binding site. Arginine, Lysine and Tyrosine carbon atoms are represented in green, blue and salmon, respectively. The AAC is shown as gray cartoon.

lead to changes in the protonation state of this residue [52]. Using H++ and QMPB/GMCT methods to predict protonation probabilities, Bidon-Chanal *et. al* calculated a pK_a value of ∼6.5 for Lys22, even though these proton binding affinities can change significantly during the transport process.

Electrostatic interactions are involved in another crucial event of this protein, its conformational change between the M- and the C-state. As discussed previously, the closing of the cavity from both sides of the membrane is accomplished by a network of salt bridges that involve the conserved MCF motifs [39, 41, 42]. In the AAC apo form, without substrate, the strength of these electrostatic interactions prevents the conformational transition between states. However, upon substrate binding, the electrostatic environment is changed and the disruption of these salt-bridges occurs, triggering the conformational change and the transport of the substrate [42, 53]. Hence, electrostatic interactions play an important role in the transport mechanism, being present in substrate attraction, selectivity and conformational change. This has been observed for AAC, but similar behaviors may also be present in a large variety of membrane transporters.

1.5 Objectives and Workflow

As discussed in the previous sections, the study of electrostatic interactions at the molecular detail is extremely relevant when discussing complex biomolecular processes. In the present work, our main goal is to establish a computational protocol to study these processes with increased realism. For this, we will combine the CpHMD method, which allow for protonation changes in the MD simulations, and an Umbrella Sampling scheme to describe the transport process, which is very slow and would otherwise be impractical to simulate with conventional methods. This way it is possible to obtain structural and protonation information of the protein in an apo form (using CpHMD simulations) and compare it with new information obtained when the substrate is being transported along the AAC channel (combining Umbrella Sampling with CpHMD methods).

In order to achieve an accurate description of ADP/ATP transport by AAC, we have devised a parallel workflow that aims at comparing data from both the resting and active protein (Figure 1.5). The first step

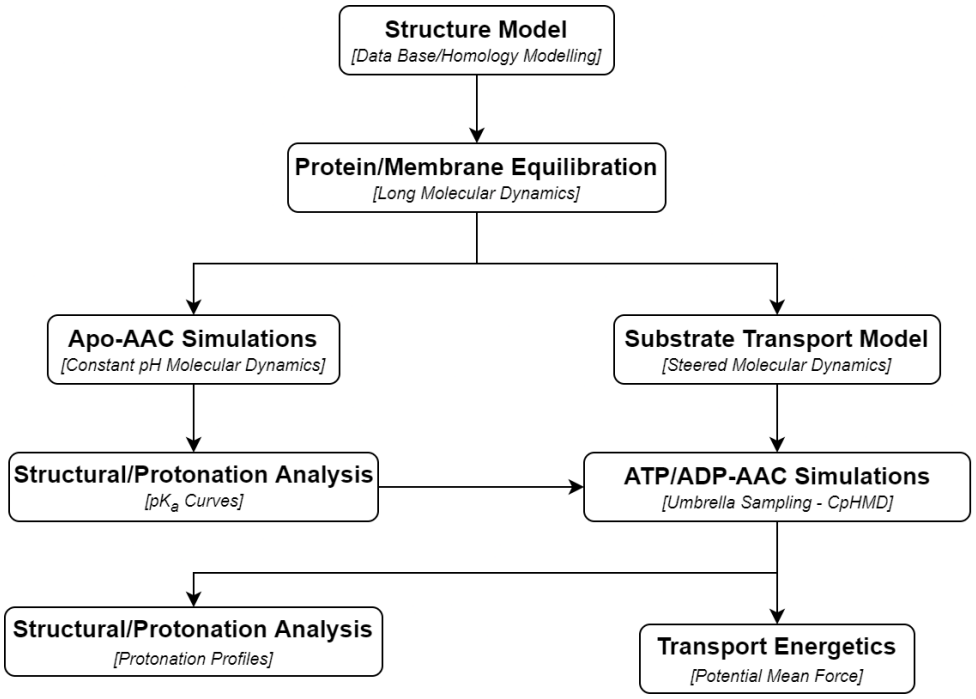

Figure 1.5: Schematic representation of the thesis workflow.

of our work consisted in gathering resolved structures of AAC from available databases. The *Bos Taurus* species has been selected to be used due to the availability of reliable experimental data and its long years of study. However, the M-state has not yet been resolved for this specie, which prompted us to use homology modelling to obtain a model of this conformational state. Following this step, our protein was inserted in a POPC lipid bilayer and equilibrated using long MD simulations. After convergence of the membrane biophysical properties and the protein/membrane interface, we started the CpHMD simulations of the apo forms of our protein, which are crucial for the identification of all pH-sensitive residues. These calculations were performed at different pH values, in the slightly acidic pH range, allowing the calculations of pK_a values of many residues, including those located in the AAC cavity. These pK_a values give us some information about the environment of the residue and the existence of intramolecular interactions in each AAC conformational state. The most relevant residues are then allowed to titrate in the CpHMD simulations coupled to the Umbrella Sampling (US-CpHMD) scheme.

In parallel to the study of the AAC apo form, we devised a steered MD protocol to obtain snapshots of our substrates in different positions along the transport process, which are the starting conformations used in the Umbrella Sampling simulations. In the steered MD simulations, the substrates are artificially pulled through the transmembrane channel mimicking the transport process across the inner mitochondrial membrane. Combining the starting configurations obtained from the steered MD protocol with the relevant residues identified from the apo-AAC CpHMD simulations, we setup an US protocol to study the transport process at pH 7. From these simulations, potential of mean force profiles of the substrates along the transport vector can be calculated and several other structural/protonation properties can also be correlated with the process.

Chapter 2

Theory and Methods

The scope of this thesis is the *in silico* study of the protein ATP/ADP carrier structure and function. This work is mainly focused on computational techniques that can be used to describe a system at the atomic level, therefore, it is necessary to describe the methods that will be used, exposing their strengths and limitations. Throughout our work, techniques such as homology modelling, Molecular Dynamics, Constant-pH Molecular Dynamics and Umbrella Sampling, will be used with the goal of obtaining relevant data for the system in study. In the following sections, the fundamental theory behind the different methods used and their relevant aspects will be presented.

2.1 Homology Modelling

We have decided to introduce first the homology modelling technique since it was the first step taken in our study. Currently, there is a large number of proteins which are of extreme medical relevance, being often related with drug transport and metabolism, signal transduction, permeability through membranes and commonly associated with pathologies. Such a high interest comes with the increased necessity of resolved structures to allow the identification of important domains. However, the number of protein sequences discovered largely exceeds the number of experimentally resolved structures in the available databases [54]. ATP/ADP Carrier is one of these cases, possessing an experimentally resolved C-state structure for many species and only a single resolved M-state structure. The main reasons for this yet scarce number of experimentally resolved structures for most proteins with already known sequence ranges from the increased financial cost, slowness of the process, and the difficulty in stabilizing certain proteins in order to allow their structure resolution. This is specially true in the case of membrane proteins, where achieving a stable crystallization is a challenge relatively to other small well solvated molecules [55].

To counter the absence of these structures, molecular biology turned to computer science in order to use the available structure data banks to predict 3D features using only the amino acid sequence of proteins and their resemblance to other available structures (homology modelling). Two observations are at the basis of this methodology: firstly, the 3D structure is derived from the amino acid sequence; and secondly, the 3D structure is significantly conserved, hence, changes during evolution occur much slower in structure than in sequence [56]. Both observations lead to the conclusion that similar sequences fold into identical structures [57].

Several steps are taken in homology modelling to derive the unknown 3D structure for our protein of interest (Figure 2.1). In first place, our target sequence is used to search for template structures in the available databases. The most common tool used in the first step is the Basic Local Alignment Search

Tool (BLAST) [58], but other approaches, such has Hidden Markov Models (HMMs) [59], can also be used with the same goal. After finding suitable templates, it is necessary to choose the one that better suits our target sequence. Usually the selection is based on the highest structure similarity, however, other factors can be considered in the selection, such has phylogenetic proximity or environmental factors (pH, solvent type, bound ligand, among others). This step is crucial since choosing an appropriate template structure will naturally wield a good model for the 3D structure of our target sequence.

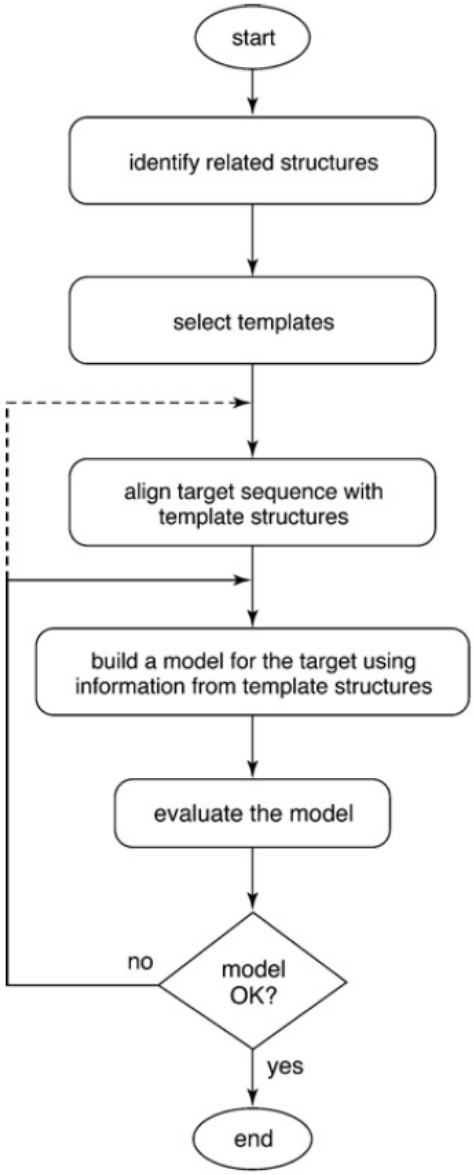

Figure 2.1: Different steps executed in a homology modelling protocol. Figure adapted from [60].

Following up, our 3D model will be built using the selected template as basis. Several approaches can be taken in model building: rigid-body assembly methods, segment matching methods, spatial restraint methods and artificial evolution methods [56]. In this work, we used the MODELLER package to perform homology modelling, which relies on spatial restrains to build the model structure. From the template, stereochemical restraints are defined on properties such as bond length, bond angle, dihedral angles and van der Waals contact distances. These restraints are then framed on our target sequence, considering its alignment with the template obtained previously [56]. Several models are constructed using these restraints, which leads to the final step of the homology modelling, choosing the best 3D

model.

The quality of the model obtained from MODELLER runs can be evaluated with different score criteria. MODELLER itself has an objective function that is able to score the quality of the structures obtained based on violations of the defined restraints and penalties of symmetry, amongst others. External methods to assess the quality of our model can also be used, such as the Discrete Optimized Protein Energy (DOPE). Using an ensemble of known native protein structures and applying probability theory, an atomic distance dependent potential is derived and used to rank our models in terms of quality [61].

Lastly, the best models that are obtained from the homology modelling process are submitted to a final quality verification. Bond lengths, torsion angles and rotational angles can be verified for violations, as well as performing a ramachandran plot to evaluate the stereochemistry of the aminoacids. WHATCHECK, Molprobity and PROCHECK are examples of software that can be used to perform the validation of the final model. Passing all these steps, a model for the unknown structure can be obtained and used in further studies.

Homology modelling is a technique that has a large number of applications, ranging from assisting in structure-based drug design, to the identification of active sites, amongst others. The derivation of high quality models opens the door to a plethora of *in silico* studies, which were being held by the lack of resolved protein structures.

2.2 Molecular Mechanics and Molecular Dynamics

2.2.1 Molecular Mechanics

Many biomolecular systems, that are important to study using computational simulations, are far too large to be tackled by quantum mechanics calculations. The computational power available currently is not enough to treat a protein structure in solution at the electronic level. Dealing with the entirety of the electrons present in a system containing a large protein, solvent molecules and, sometimes, a surrounding membrane is simply too much too handle with the current computational power available, especially if one requires to introduce motion and explore its conformational space. Therefore, molecular mechanics (MM), also called force field methods, are employed to bring down the computational load [25, 26].

In order to remove electrons from the calculations of these systems, the Born Oppenheimer approximation is used. This approximation is based on the large difference of mass between nuclei and the electrons. Due to their small size, electrons motions can be separated from the nuclear motions by assuming that the electrons can rapidly adjust to a change of position by the nuclei. As such, the energy of a molecule is characterized by an energy function that is only dependent of the nuclear coordinates [8, 25, 26]. Without the need to simulate the electrons of each atom, the computing time is greatly decreased, however, properties related to electronic distribution cannot be simulated.

MM methods aim to apply classical mechanical concepts to the molecular level. To achieve this, two crucial elements are required: first a potential energy function (PEF) which describes the energy of the system; and, secondly, an empirical force field that compiles the parameters to be used in the PEF.

2.2.2 Potential energy function

When using MM in a simulation, our system's energy will be described by a potential energy function (PEF) [25, 26]. There are several variations of PEFs, since terms can be added, removed or changed in order to better describe the properties intended. Thus, the choice of a PEF should be done according to the type of system in study [25].

A PEF attempts to express the intramolecular motions and interactions mathematically. If this is correctly achieved, it can be applied to a large variety of molecules, from small ones to large polymers or even proteins, since it is assumed that these molecules obey the same set of physical laws. The PEF has two major components: bonded interactions, such as bond stretching, angle bending, proper and improprial dihedral torsion; and nonbonded interactions, namely, van der Waals forces and electrostatic interactions (Equation 2.1).

$$V(\vec{r}^{\,N}) = V_b(\vec{r}^{\,N}) + V_a(\vec{r}^{\,N}) + V_{pd}(\vec{r}^{\,N}) + V_{id}(\vec{r}^{\,N}) + V_{vdW}(\vec{r}^{\,N}) + V_{elec}(\vec{r}^{\,N}) \tag{2.1}$$

In this equation $\vec{r}^{\,N}$ represents the positions of all atoms in their vector form. The first term in equation 2.1 is the sum of potential energy from bond length stretching. Usually this potential energy is represented by Morse's potential [62]. However, this equation is not computationally efficient and since it is uncommon in computer simulations to allow bonds to deviate significantly from their equilibrium values, a simpler equation can be used. Considering that the bond stretching has a similar behaviour to a spring, Hooke's law is usually used in computational simulations to calculate this potential, as shown in equation 2.2 [26].

$$V_b(\vec{r}^{\,N}) = \sum_{bonds}^{N_b} \frac{1}{2} k_{bn}(b_n - b_{0n})^2 \tag{2.2}$$

k_{bn} is the force constant between two atoms, b_n is the length of the bond and b_{0n} is the reference length of such bond. This way, equation 2.2 increases the energy of the system when the bond length (b_n) deviates from the reference bond length (b_{0n}).

In the second term of the PEF, we have the potential energy of angle bending, which is also presented as a harmonic potential function derived from Hooke's Law. Similarly to the previous term, this equation represents the energy penalization associated with the sum of the deviation in all angles of the molecule from their reference values (Equation 2.3).

$$V_a(\vec{r}^{\,N}) = \sum_{angles}^{N_\theta} \frac{1}{2} k_{\theta_n}(\theta_n - \theta_{0_n})^2 \tag{2.3}$$

The third and forth terms in the PEF account for the potential energy contributions of torsions. However, these differ slightly from one another. Equation 2.4 is the sum of potential energy of torsion in proper dihedrals, which are dihedrals constituted by consecutive atoms (1-2-3-4). This term is a cosine series expansion, having minimum and maximum at regular intervals.

$$V_{pd}(\vec{r}^{\,N}) = \sum_{torsion_{pd}}^{N_\varphi} k_{\varphi_n}[1 + cos(m_n \varphi_n - \delta_n)] \tag{2.4}$$

Equation 2.5 describes improper dihedral torsion via a harmonic potential. These type of dihedrals are necessary to use in order to maintain stereochemistry at chiral centers and also to prevent deformations when hydrogen atoms are treated implicitly. The GROMOS force fields use an united-atom approach, to reduce computational times, which treats implicitly the apolar hydrogens, correcting the mass and charge of the carbon they are attached/collapsed in [63] This simplification creates a problem in the degrees of freedom of some dihedrals, the absence of apolar hydrogens allows the surrounding atoms to greatly move their angles taking the place where the hydrogen should be. To prevent this effect, improper dihedrals are used to keep the position of these atoms in the correct location. Usually, contrary to proper

dihedrals, the improper dihedrals are defined by non-sequential atoms (1-4-3-2).

$$V_{id}(\vec{r}^{\,N}) = \sum_{torsion_{id}}^{N_\xi} \frac{1}{2} k_{\xi n}(\xi_n - \xi_{0n})^2 \tag{2.5}$$

The fifth and sixth terms in the PEF are the contributions of the nonbonded interactions in the system. These depend on atom proximity and not on bonds between atoms, therefore, their strength is inversely proportional with the distance between the atoms. Equation 2.6 describes the van der Waals forces (vdW) using a Lennard-Jones 12-6 potential [26]. r_{ij} is the distance between two atoms, $C_{12_{ij}}$ and $C_{6_{ij}}$ are interaction parameters referring to repulsion or attraction between the atoms, respectively.

$$V_{vwd}(\vec{r}^{\,N}) = \sum_{i=1}^{N} \sum_{j>i}^{N} \left(\frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \right) \tag{2.6}$$

The sixth equation in the PEF takes into account electrostatic interactions (Equation 2.7). In molecules, electrons are more attracted to more electronegative atoms, leading to an unequal charge distribution and polarity. In order to mimic the charge distribution of molecules in a simulation, point charges can be assigned to their atoms, which are referred to as partial charges. With these charges, electrostatic interaction can be calculated using Coulomb's law, as a sum of point charge pairs interacting. q_i and q_j are the partial atomic charges of atoms i and j, N is the number of atoms in the system, like before r_{ij} is the distance between two atoms, ε_0 is the permittivity in vaccum and finally ε_r is the relative dielectric constant of the medium where the atoms are placed.

$$V_{elec}(\vec{r}^{\,N}) = \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} \tag{2.7}$$

2.2.3 Force Fields

In order for the PEF to describe the energy of a system, it requires adequate parameters, which are compiled in a force field (FF). It usually contains optimal bond lengths, angles, dihedrals, Lennard-Jones parameters, the atomic partial charges, etc. All these parameters are obtained empirically, or with the help of quantum mechanics calculations, in order to accurately describe the experimental data available. Typically, quantum methods are used first-hand and then some manual adjustments can be made to adjust the simulation to correctly describe the experimental data. Therefore, the use of a force field must take into account the purpose they were parameterized for, and this choice is paramount to obtain accurate simulations.[25, 27]

An advantage of using force fields is their transferability. The same parameters can be used on a series of related molecules without the need to develop a new set of parameters. This allows for the force field parametrization to be done on small molecules and then extrapolated to more complex systems.[26] Nevertheless, some systems might still require reparameterization of certain molecules to account for different environments.

Two groups of force fields are worth to notice: the all-atom force fields, where all atoms are taken into account in the PEF calculations; and the united-atom force fields, where apolar hydrogens are treated implicitly by their contribution being included in the heavy atoms to which they would be bonded to (collapsed hydrogens). The advantage of the united-atom force fields is the lower computational resources required since a great amount of hydrogens are not treated explicitly. Currently, CHARMM and AMBER are two major families of all-atom force fields, while GROMOS is probably the major set of united-atom

force fields.

2.2.4 Molecular Dynamics

Molecular Dynamic simulations aim to sample the conformational space of the system using Newton's Laws of motion while integrating such motion with the potential energy described by the molecular mechanics potential energy function. Using these simulations, it is possible to obtain a description of the positions and velocities of all particles along the simulation time. Solving Newton's second law equation (Equation 2.8) leads to the acceleration of each particle in the instant t.

$$\frac{d^2\vec{r}_i(t)}{dt^2} = \frac{\vec{F}_i(t)}{m_i} \tag{2.8}$$

$\vec{r}_i$ is the coordinate vector of the particle i at instant t, m_i is the mass of that particle. $\vec{F}_i$ is the force acting on the atom at the t instant and it is described by a potential gradient acting on the atom i in the position r (Equation 2.9).

$$\vec{F}_i = -\nabla_{r_i} V \tag{2.9}$$

Integrating the Equation 2.8 leads to new positions and velocities for all the atoms in the system. In a realistic model, the forces acting on atoms will change with its different positions and with changes in its neighbors, hence, all motions are coupled together. The sheer amount of particles interacting in the system makes MD simulations become chaotic and the integration of the motion law's results in a many-body problem that cannot be solved analytically. A common approach to circumvent this limitation is to integrate the equations of motion using the finite difference method [26]. With this method, the integration over time is performed in many small steps (Δt), where the force is considered constant. As a consequence, Δt must be smaller than the faster degree of freedom in the system, otherwise the description of the system will not be accurate. We can use the properties of time t to calculate the positions and velocities at $t + \Delta t$. Several algorithms can be used to integrate the equations of motion using the finite difference method and most of them share the assumption that the positions and dynamic properties of the system can be approximated as a Taylor series expansion.

In this work, we used the leap-frog algorithm [64], which is a variation of the Verlet algorithm [65], a common finite difference method used in MD simulations. The leap-frog algorithm, developed by Hockney in 1970, uses the following relationships:

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \Delta t \vec{v}(t + \frac{1}{2}\Delta t) \tag{2.10}$$

$$\vec{v}(t + \frac{1}{2}\Delta t) = \vec{v}(t - \frac{1}{2}\Delta t) + \Delta t a(t) \tag{2.11}$$

The algorithm calculates the velocities at the previous time $t - \frac{1}{2}\Delta t$ and the current acceleration at t to obtain the velocity at time $t + \frac{1}{2}\Delta t$. Then the new positions at $t + \Delta t$ can be obtained from previous positions at t (Equation 2.10) and its velocities (Equation 2.11).

2.2.5 Periodic boundary conditions (pbc)

Ideally, MD simulation systems need to be relatively small in order to cut down simulation times. Therefore, it is common for these systems to use a finite sized box containing a limited number of molecules. This setup can lead to a problem at the box boundary. If no correction is applied, molecules

that approach the boundary of the box will encounter a non-interacting physical wall and be submitted to unrealistic forces. To deal with this discontinuity, MD simulations employ a correction method under the form of periodic boundary conditions (pbc). By using a periodic space-filling shape box in the MD simulation, the simulation box can be treated as being surrounded by identical copies of itself [25]. This ensures the continuity of the system where an atom that leaves the box at one point, will re-enter the opposite side of the box with identical velocity (Figure 2.2).

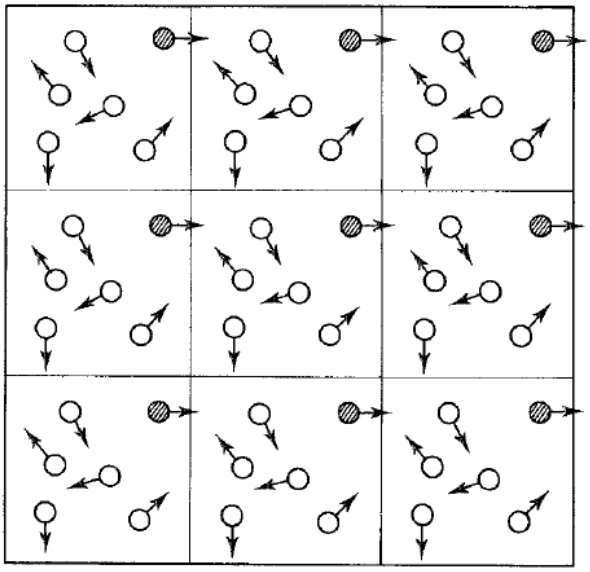

Figure 2.2: Representation of periodic boundary conditions in 2D. The center box is surrounded by copies of itself. Molecules of the system are represented as circles and their motion directions in a vector form. A grey filled circle represents a molecule that would leave the center box and reenter on the opposite side. Figure adapted from [26].

Since periodic copies of the same system are being used, it is required that the simulation box is large enough to avoid artifacts in the simulation. A typical example of this problem is when the solute interacts, even transiently, with its own periodic copy, creating a model of nonphysically high concentrations. To avoid this, we should increase the distance between the solute and the box boundaries, usually larger than half the cut-off radius used in the non-bonded interactions [25].

2.2.6 Temperature and Pressure Conditions

The goal of molecular simulations is to achieve a realistic description of the system *in vivo*. Therefore, we need to mimic physiological conditions as best as possible, this includes maintaining a constant temperature and pressure (NPT conditions). To obtain this, we need to apply algorithms that can model these conditions in our simulations. Temperature is modulated by tweaking the kinetic energy of the particles in the system, while pressure is usually kept constant by adjusting the system volume.

Temperature

Temperature is a macroscopic property and an important factor in nature and in living cells, since it can influence the structure and activity of its biomolecules. Usually, *in vivo* systems are subject to constant temperatures with small fluctuations. In MD simulations, the systems are microscopic, which means that the temperature needs to be imposed from an external thermostat. In this work, we used v-rescale [66], which is a variation of the Berendsen thermostat [67], to keep the temperature constant

by coupling our system with an external heat bath. The correction to the deviation from the desired temperature T_0 is achieved by Equation 2.12.

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.12}$$

This will lead to a time-exponential decay of the deviation in temperature (λ), which is given by Equation 2.13.

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left\{\frac{T_0}{T(t - \frac{\Delta t}{2})}\right\}\right]^{\frac{1}{2}} \tag{2.13}$$

In this equation τ_T is a time constant that will influence the decay of our temperature deviation. Small values of τ_T will wield tighter coupling of our system to the heat bath and will result in a fast temperature decay to the desired temperature, while larger τ_T values represent the opposite.

Both equations define the Berendsen thermostat algorithm [67]. However, when used, a canonical ensemble is not being sampled correctly, due to suppression of the system kinetic energy. This has been corrected by the v-rescale thermostat [66] which applies a stochastic term (dW) that ensures a correct kinetic energy (K) distribution (Equation 2.14).

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{dW}{\sqrt{\tau_T}} \tag{2.14}$$

Pressure

Similarly to temperature, pressure must also be kept constant throughout the simulation to correctly model an *in vivo* system. Two options are common and have been used successfully in MD simulations: a Berendsen barostat [67], which has a similar mathematical formalism as the Berendsen thermostat; and the Parrinello-Rahman barostat [68, 69]. The latter algorithm has been adopted in this work and it keeps the pressure constant by changing the volume of the simulation box according to Equation 2.15, where b is a representation of the box vectors in form of a matrix and V is the box volume.

$$\frac{d^2b}{dt^2} = V\ W^{-1}b'^{-1}(P - P_{ref}) \tag{2.15}$$

In this equation the coupling strength is determined by the inverse mass parameter matrix (W^{-1}) and is dependent of the size of the box. By providing an isothermal compressibility (β) and a pressure time constant (τ_p), the correct value for W^{-1} is calculated automatically according to Equation 2.16. The variable L in this equation is the largest box matrix element.

$$(W^{-1})_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L} \tag{2.16}$$

2.2.7 Energy Minimization

MD simulations are one of the most useful tool used to sample conformational states of our system. However, this method is inherently chaotic and, if not treated correctly, could lead our system to high energy conformations that are not representative. Hence, it is always advised to perform an energy minimization of the system and search the conformational space for a minimum solution to the PEF function [26]. Since the function that describes the energy of the system on Molecular mechanics is multidimen-

sional, there will exist many local energy minima making the energy landscape highly complex, which, in turn, makes it almost impossible to reach the global minimum. Reaching a local minimum is probably enough to achieve conformations with high stability.

The most common method to perform the energy minimization steps is the steepest descent [70], which is very robust and easy to implement while lacking efficiency in sampling the more stable minima [71]. This method searches for energy minima in the direction of the negative gradient, which can be understood as in the direction of the maximum component of the force. Therefore, the method is only able to search stable conformations downhill of the energy potential surface, being unable to perform jumps between local minima. The steepest descent algorithm uses a vector $\vec{r}$, that contains the Cartesian coordinates of the N atoms present in the system, and an initial maximum displacement h_0. The positions of the new conformations (r_{n+1}) obtained through the algorithm can be calculated by Equation 2.17.

$$\vec{r}_{n+1} = \vec{r}_n + \frac{F_n}{max(|F_n|)} h_n \tag{2.17}$$

In this equation, h_n is the maximum displacement allowed on conformation r_{n+1} when compared to the conformation r_n. F_n is the force applied on the atoms and the $max(|F_n|)$ refers to the maximum value for all the force components. The potential energy of the new state is calculated and compared to the previous one. If $V_{n+1} < V_n$ then the new coordinates are accepted and the maximum displacement value is increased 20% on the next step, allowing the positions of the atoms to move more on the next step. If $V_{n+1} > V_n$, the opposite occurs, the energy did not decrease and the jump that was given was probably too large. Therefore, the h_n is decreased to 20% of the initial value, leading to smaller atom movements in an attempt to find the local minimum in the energy landscape. This process will end after a user defined number of steps or if the maximum force component value reaches a user defined threshold.

As pointed before, the steepest descent algorithm can only search the local minima that sit nearby and directly downhill of the energy landscape, and being unable to perform jumps between these minima. More complex energy minimization methods, such as conjugate gradient and l-BFGS [72, 73], can overcome some of these limitations. These methods are slower and more complex than the steepest descent but allow for the sampling of other local minima, further away, and usually more stable. A detailed analysis of all these methods is outside the scope of this thesis.

2.2.8 Long-range Interactions

The treatment of non-bonded interactions is pivotal for the realism and efficiency of MD simulations. When dealing with large systems it becomes unbearable to compute all van der Waals and Coulombic interactions between every atom of the system [26, 27]. This is specially true when pbc is applied, since it makes the system virtually infinite. A simple solution to solve this problem is the use of cut-offs to restrict the number of interactions computed. Typically, a twin-range cut-off scheme is used with the GROMOS force fields, where two spheres are considered, a short distance sphere with radius R_c^1 centered on atom i and a long distance sphere with R_c^2 radius also centered on atom i (Figure 2.3). With this scheme, we can apply two different restrictions, one based on distance, that is accomplished by setting both radii distances R_c^1 and R_c^2, and secondly a frequency restriction, where it is defined how often are the neighbour lists updated. The short distance sphere is usually updated every MD step while the list of atoms that fall within the R_c^1 and R_c^2 region is updated less frequently (normally this second list is updated every 5 steps). Both restrictions were set to reduce the computational load of updating the neighbour atom list, which is often a rate-limiting step. The lower update frequency on the long distance

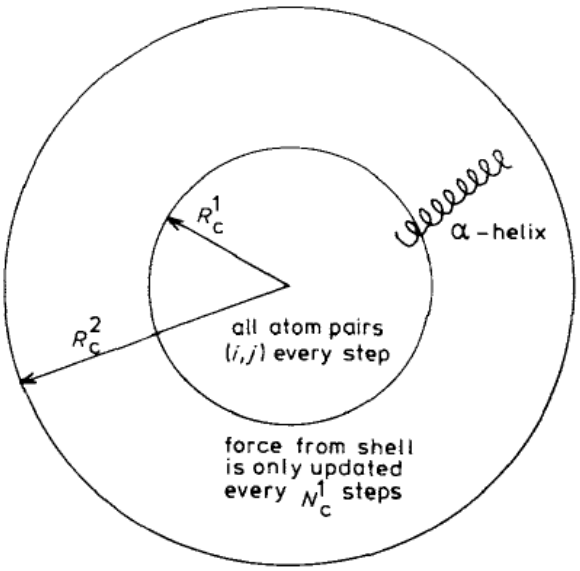

Figure 2.3: Twin-Range schematic representation. R_c^1 is the cutoff raddi centered on particle i delimiting the inner sphere where the atom pairs are updated every step. R_c^2 is the cutoff raddi centered on particle i delimits the outer sphere where the atom pairs are only updated every N_c^1 steps. Figure adapted from [25].

atom list assumes that the long range atoms have a negligible impact on the non-bonded terms of the potential energy. With the current state of computing power, the twin-range cutoff scheme does not pose a significant performance upgrade when compared with the single cut-off scheme (a single R_c radius sphere). Furthermore, it has been shown to create system instability in many systems [74]. Therefore, in this work, the twin-range cutoff scheme was only used in some preliminary testing, while the single cut-off was adopted for all production runs.

For van der Waals calculations, the application of a cut-off is relatively straight forward. Lennard-Jones potential rapidly tend to zero when the distance increases ($1/r^6$), therefore a cut-off truncation can be applied to define which atoms should be included. A sphere of radius R_c is centered on atom i and only the atom pairs ij within this sphere will be considered for the calculation. Unfortunately, this approach cannot be directly applied to the electrostatic interactions since these forces decay less abruptly with the distance ($1/r$). Applying a simple cut-off truncation without correcting for long range effects will most likely result in severe problems on the molecular properties of the system. Therefore, it is mandatory to use methods than can perform this correction and deal with these long-range electrostatic forces. In this work, we use the Generalized Reaction-Field (GRF) [75] and the Particle Mesh Ewald (PME) [76, 77] methods to deal with long-range electrostatics.

The GRF is a continuum method based on Reaction-Field theory [78] that aims at solving the Poisson-Boltzmann equation. A cut-off with radius R_c is defined centered on atom i dividing the system in two. Inside the cut-off sphere with radius R_c the atomic charges are treated explicitly with a dielectric constant of ε_1. Beyond the cut-off distance, the environment is assumed to be homogeneous and therefore can be treated as a continuum medium with a dielectric constant of ε_2 and ionic strength I. The electrostatic potential inside the inner region of the cut-off is calculated using a direct Coulomb term that uses the atomic charges within the inner sphere and the reaction field potential dependent of ε_2 and I. This way the electrostatic potential of the inner region is dependent of the outer region medium. An attenuating effect is felt on the electrostatic interactions of atom pairs inside the cut-off which is caused by the reaction field potential of the outer region. The GRF method is a conceptually simple method and is computationally efficient. However, a major caveat is the use of a homogeneous continuum to non-

isotropic systems. Indeed, when membrane systems are involved, this method can create some artifacts that are difficult to pinpoint and to quantify.

The PME method [76, 77] treats the long range electrostatic energy of atom *i* using Ewald summation [79], taking into account the interactions with all other particles, including the ones on the surrounding periodic system images. Although it derives from Ewald summations, PME improves the convergence speed by assigning charges to a grid that is placed on the system and then applying a Fourier transformation algorithm. Hence, the system is divided in smaller parts where the potential is calculated much faster than with typical Ewald sums. Obtaining a highly accurate description of the electrostatic potential of the system comes at high computational cost, since an increased number of grid points would be required. Ewald sum methods, including PME, were developed to calculate electrostatic potential energy of ionic crystals, which are periodic and neutral [79]. This raises a problem when non-neutral systems are being studied. Some implementations try to circumvent this by including a background neutralizing charge when non-neutral systems are studied. Neutralization is obtained by spreading indiscriminately counter-charge on the grid in order to compensate for the net charge present. Applying the background charge correction to systems with small net charge is usually acceptable and without major consequences, since the charge correction that is spread throughout the grid is small. However, when the net charge of the system is high, the background correction charges distributed are not negligible to the calculation, which can produce significant artifacts [80]. Therefore, with PME and when possible, it is highly advised to use counter-ions to neutralize the system box.

2.3 Poisson–Boltzmann and Monte Carlo

2.3.1 Poisson–Boltzmann

As highlighted in section 1.1 the electrostatic potential is a key factor in many molecular processes, including protein functions. Electrostatic interactions in proteins are usually regarded as a major driving force for substrate affinity and binding, catalysis, aggregation, and protein folding. As such, a correct study of these forces is highly valuable to understand a great panoply of processes.

In biochemical processes involving large systems, such as a system composed of a membrane protein embedded in a lipid bilayer and water, it becomes impractical, due to the high computational load, to calculate the complete electrostatic potential of every simulation step. To circumvent this, we can resort to simplified methods like continuum electrostatics (CE) [8]. Instead of accounting for every particle in the system, two distinct dielectric regions are defined. The solvent is seen as one region that has a large dielectric constant (ε_s), due to the dipole reorganization freedom present in this region, and an ionic strength of I, resulting from counter-ions shielding effects. The solute is defined as the second region, having a lower dielectric constant (ε_p) since it has a large number of constrained dipoles, such as peptide bonds in a protein. Therefore, our CE model is described by a small dielectric constant region that is surrounded in a high dielectric region. The separation between these two regions is called the solvent accessible surface. Due to the presence of an ionic strength in the solvent region, an ion exclusion boundary is also present and is defined by the minimum distance a counterion in solution can approach the solute. With these conditions, the electrostatic potential can be calculated as a continuous function in space.

A common approach to obtain the electrostatic potential is to use the linearized form of Poisson–Boltzmann (LPBE) equation for a rigid body [8], which in our case will be the solute (Equation 2.18).

$$\nabla \cdot [\varepsilon(\vec{r})\nabla\phi(\vec{r})] - \kappa^2(\vec{r})\varepsilon(\vec{r})\phi(\vec{r}) = -4\pi\rho_p(\vec{r}) \tag{2.18}$$

$\vec{r}$ represents the position of the charge point in the system, $\varepsilon(\vec{r})$ is the dielectric constant, $\phi(\vec{r})$ is the electrostatic potential and $\rho(\vec{r})$ is the charge density. An extra term is present on the equation due to the existence of ionic strength on the solvent medium. This will have an effect on the electrostatic potential that is calculated and is given by the reciprocal Debye length (κ) that is derived from the Debye–Huckel theory [81]. It determines the limit length a charge q_i can be shielded by the ionic effect (Equation 2.19).

$$\kappa(\vec{r}) = \begin{cases} \left(\frac{8\pi e^2 I}{\varepsilon_s k_B T}\right)^{1/2}, \text{if } \vec{r} \text{ is accessible to ions} \\ 0, \text{if } \vec{r} \text{ is not accessible to ions} \end{cases} \tag{2.19}$$

Where k_B is the Boltzmann Constant, T is the absolute temperature and I and ε_s are the ionic strength and the solvent dielectric constant, respectively. In this equation, we determine two different conditions. κ will have its normal form if the $\vec{r}$ is in a region accessible to ions. However, if the region is inaccessible to ions, this variable will be changed to zero and the Poisson equation (Equation 2.20) will be used to calculate the electrostatic potential.

$$\nabla[\varepsilon(\vec{r})\nabla\phi(\vec{r})] = -4\pi\rho_p(\vec{r}) \tag{2.20}$$

For systems with a large number of particles and with complex geometries, obtaining an analytical solution to the LPBE equation is difficult. Hence, a numerical method, such as the finite difference method, is applied in order to achieve a solution [82]. Commonly, cubic grids are used in the finite differences method to discretize the system, allowing iterative calculations on each grid point. This is done in two steps, first a coarser grid is used to solve the LPBE equation which gives a good initial guess of the potential at the boundaries, then a finer grid will be applied (focusing) and calculations are executed in this smaller grids until a convergence criterion is reached [82]. A converged electrostatic potential of a given grid allows the calculation electrostatic energies (W) which represent the work required to bring the charges in question from zero to their actual value (Equation 2.21).

$$W = \frac{1}{2}\sum_i q_i\phi(\vec{r}_i) \tag{2.21}$$

By comparing the electrostatic energy between two different protonation states, we are able to quantify these electrostatic energy differences (ΔG_{elect}), which will be pivotal to estimate p$K_a$values.

2.3.2 Protonation free energy and pK_a calculations

As explained in the previous section, the electrostatic potential can be used to obtain the energy difference between two distinct protonation states. However, since most of the systems in study are proteins with multiple pH titrable groups, this is not a simple straightforward calculation. To deal with this problem, we can use a thermodynamic cycle involving the titrable sites in solution and on the final protein environment (Figure 2.4) [8]. The use of a thermodynamic cycle allows the concentration of many effects related to the protonation step, including electronic/quantum effects, into the solution part of the cycle. These can be calibrated in solution and should hold in the protein part of the thermodynamic cycle.

From the thermodynamic cycle, we can deduce the free energy for a protonation event in protein

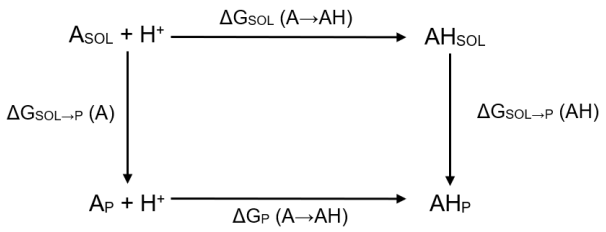

Figure 2.4: Thermodynamic Cycle of a titration event in solvent (A_{SOL}) and in protein environment (A_P).

environment, which is given by the free energy of the protonation in solution and the transitions of each state (protonated and deprotonated) from solvent to protein (Equations 2.22 and 2.23).

$$\Delta G_P(A \rightarrow AH) = \Delta G_{SOL}(A \rightarrow AH) + \Delta G_{SOL \rightarrow P}(AH) - \Delta G_{SOL \rightarrow P}(A) \tag{2.22}$$

$$\Delta G_P(A \rightarrow AH) = \Delta G_{SOL}(A \rightarrow AH) + \Delta\Delta G_{SOL \rightarrow P}(A \rightarrow AH) \tag{2.23}$$

We can convert the free energy of the protonation of a residue in a protein to a pK_a value according to equation 2.24.

$$\mathrm{p}K_a = \frac{\Delta G}{2.3 k_B T} \tag{2.24}$$

Therefore, if we use the LPBE equation to obtain the value of ΔG, we can compute the pK_a of our residue. To achieve this, we need to split the pK_a energy contributions in two parts, an intrinsic part, which only depends on the protein environment (pK^{int}) and all the interactions between all residues (W_{ij}) (Equation 2.25).

$$\mathrm{p}K_a = \mathrm{p}K^{int} + W_{ij} \tag{2.25}$$

In the first contribution, we calculate the pK^{int} value of the residue, assuming that all other titrating residues are in a reference state (Equation 2.26).

$$\mathrm{p}K^{int} = \mathrm{p}K^{mod} + \frac{\Delta\Delta G^{env}}{2.3 k_B T} \tag{2.26}$$

Here, the pK^{int} is the sum of the pK of the model residue in solution (pK^{mod}) and the difference of free energy of protonation between solution and in the protein environment ($\Delta\Delta G^{env}$).

Protein structures commonly exhibit several titrating residues, hence, the pK_a of the interest residue must also include a contribution reflecting interactions between titrable sites (W_{ij}). These interaction terms are dependent on the pH, therefore the p$K_a$of our residue of interest is also tightly dependent on the pH of the solution. To evaluate the probability of all protonation states, a vector $\vec{a}$ can be used, which includes every combination of possible protonation states for each site in the protein. Rewriting equation 2.24, we obtain an equation for the energy difference between the reference protonation state ($\vec{a}_0$) and a protonation state $\vec{a}_i$ (Equation 2.27).

$$\Delta G_{\vec{a}_0 \rightarrow \vec{a}_i} = -2.3 k_B T \sum \vec{a}_i \gamma_i \mathrm{p}K^{int}{}_i + \sum_i \sum_{j \neq i} \vec{a}_i \vec{a}_j \Delta W_{ij} \tag{2.27}$$

In this equation, γ_i is the charge of the titrable site when ionized and ΔW_{ij} is the free energy of the interaction between the titrable sites i and j. All these terms can be calculated from the LPBE-derived electrostatic potential, providing a function from where the protonation states (and the pK_a value) of each

residue of the protein, can be calculated.

2.3.3 Monte Carlo

The approach established in the previous section to estimate the protonation state of a site has a noticeable limitation. Since the protonation state is dependent of the environment electrostatics and usually several titrable sites exist on a protein, the total number of combinations of protonation state can become too computationally expensive [83]. To tackle this problem, instead of calculating every possible protonation combination for the titrable sites of our protein, a Monte Carlo (MC) simulation using a Metropolis criterion can be performed [83–85]. MC is a sampling method that can easily be applied to sample the most relevant protonation states in a protein [83].

In MC, each new protonation state generated, like changing from state $\vec{a}_0$ to $\vec{a}_1$, at a given pH, can be written in terms of free energy (Equation 2.28), similarly to what was done in equation 2.27.

$$\Delta G_{\vec{a}_0 \to \vec{a}_1}(pH) = -2.3 k_B T \sum a_i \gamma_i (\mathrm{p}K^{\mathrm{int}}{}_i - pH) + \sum_i \sum_{j \neq i} a_i a_j \Delta W_{ij} \tag{2.28}$$

The generation of a new protonation state ($\vec{a}_2$) will allow its comparison with the previous state ($\vec{a}_1$) through the difference of their free energies when compared with the reference state ($\vec{a}_0$) (Equation 2.29).

$$\Delta\Delta G_{\vec{a}_1 \to \vec{a}_2}(pH) = \Delta G_{\vec{a}_0 \to \vec{a}_2} - \Delta G_{\vec{a}_0 \to \vec{a}_1} \tag{2.29}$$

This free energy difference can be given to the Metropolis criterion to decide if the trial is accepted (Equation 2.30).

$$\text{Protonation state } \vec{a}_2 \text{ accepted if } \begin{cases} \Delta\Delta G_{\vec{a}_1 \to \vec{a}_2}(pH) < 0 \\ \Delta\Delta G_{\vec{a}_1 \to \vec{a}_2}(pH) \geqslant 0, \text{with a probability of } e^{\frac{-\Delta\Delta G_{\vec{a}_1 \to \vec{a}_2}(pH)}{k_B T}} \end{cases} \tag{2.30}$$

The newly generated state will be accepted if the energy difference is negative or conditionally accepted if the energy difference is otherwise. If the new state is accepted, the process is repeated with the generation of a new MC move and evaluation, until a determined number of steps is reached [25]. Choosing an appropriate high number of MC steps is important to guarantee a vast ensemble of stable protonation states with enough sampling to generate accurate probabilities. With the probabilities of the relevant protonation states, we can extract pK_{a} values for all residues in our protein.

2.4 Constant-pH Molecular Dynamics

The interest in accurately study protonation states in biomolecular systems, which are tied to pH, has peaked the interest of the modelling community. Achieving a correct description of the electrostatics of a biomolecular system is highly advantageous to better understand the interactions that happen in the systems. Two different methodologies have been discussed so far. The MM/MD simulations which treat the system using a force field to describe the energy of the system and integrate the equations of motion in order to sample conformations. This approach is limited when it comes to pH effects since in MM/MD the system titrating sites are fixed by setting a user-defined representative protonation state. The electrostatic environment of the protein is, therefore, biased by that choice, lacking the flexibility of changing protonation states that could be very relevant in some systems. On the other end, a combination of Continuum Electrostatics with Monte Carlo sampling (PB/MC) has been presented as a way to

quantify the electrostatic energy of a system when protonation changes occur. Nevertheless, this method lacks the structural flexibility and the conformational sampling provided by MM/MD, due to the solute being treated as a rigid entity in the LPBE calculations.

The ideal method would combine these two approaches and achieve both the conformational and protonation sampling in a coupled way. The Constant-pH Molecular Dynamics (CpHMD) method, originally developed by A. M. Baptista[28, 29] is a great example of such method. A small schematic is provided to better describe the flow of this method (Figure 2.5). In CpHMD, a PB/MC step starts the

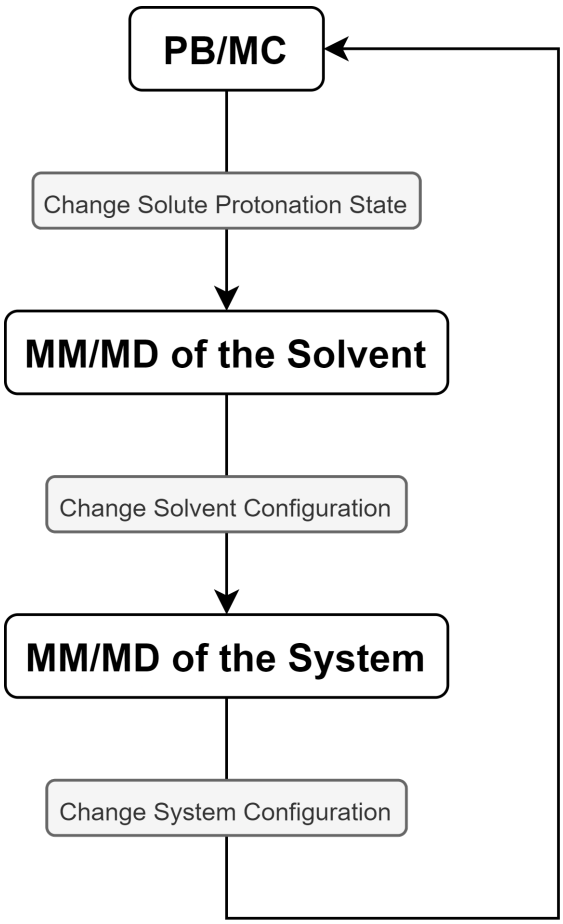

Figure 2.5: Schematic representation of the steps taken in a CpHMD simulation.

simulation by calculating a representative protonation state (the last of the MC run) at the specified pH value. This will define the protonation states of the titrating sites in the system for a defined MD segment. The first very short MD segment, where the solute is freezed, allows the solvent to adapt to the new charges obtained from the PB/MC, eliminating possible unfavourable interactions between the solvent and a titrable site. A longer unrestrained MM/MD step is then performed in the system, which constitutes our real MD production. Finally, the last conformation obtained will subjected to a new PB/MC calculation to estimate new protonation states and the cycle repeats until the user-defined simulation time is reached. With this method, we can sample both conformation and protonation states, at a given pH value, which can be used to estimate several thermodynamic properties and average protonations for all sites. If more than one pH value is simulated using this method, titration curves and pK_a values can be obtained for each titrable amino acid.

2.5 Umbrella Sampling

The main goal of this thesis is to study the effects of substrate transport (ATP and ADP) on our AAC membrane protein. For this, we will calculate the titration curves of several key residues in the protein cavity and use them as sensors to quantify the effects of the substrate presence in the channel. Also, with a correct description of the electrostatic balance during transport, we can estimate the energy barrier of the whole import/export processes. We certainly can extract a lot of information with pK_a values of the apo form of AAC, obtained from CpHMD simulations. However, in a typical MM/MD simulation it is virtually impossible to simulate the transport of a large substrate through a membrane protein channel. This is mainly due to the time scale accessible to MM/MD and CpHMD simulations (sub-microsecond range) which is very far away from the timescale required to capture a complete membrane crossing process (probably approaching the milisecond timescale). A way to deal with this problem is to use enhanced sampling techniques where a biased potential is applied to a MD simulation to force a certain event in the system.

In our work, we use the Umbrella Sampling (US) technique which aims at sampling the transition between two states by splitting it into short segments (umbrellas) [33]. This requires the choice of a reaction coordinate (ξ) which is a parameter that can be used to distinguish two different thermodynamic states. The most common choice for the reaction coordinates is to use geometric properties such as distances, torsions, variation of root mean square between states, and so on [33]. Depending on the value of the reaction coordinate, distinct thermodynamic states can be sampled and studied energetically. In order to prevent the reaction coordinate value to drift from the reference ξ, in each umbrella, an additional energy term is added to the system, this term is called a bias potential (ω_i). This term guarantees an efficient sampling around the defined reaction coordinate. One of the most effective ways of implementing this is by using a harmonic bias potential with a certain force constant (K), as shown in equation 2.31.

$$\omega_i(\xi) = \frac{K_i}{2}(\xi - \xi_i^{ref})^2 \tag{2.31}$$

This bias is very appealing due to the simplicity of design and the low number of parameters. The choice of K is key, since higher values of K will lead to narrower distributions of the sampling window, which is due to the higher energy penalties applied to the larger deviations from the ξ^{ref}. Unless the energetic gain of the deviated state can overcome the penalization of the potential bias, the system will have a tendency to stay near the ξ^{ref}.

The correct parameters of the bias potential are crucial for US simulations and to obtain an unbiased free energy of the system along the reaction coordinate, which can also be called the potential of mean force (PMF) [34]. The strength of the bias (K) and the number of windows must be chosen to allow the best possible sampling and to cover all the range of ξ that is intended to study. In order to obtain the PMF on window i ($A_i(\xi)$), an unbiased distribution is required. However, the MD simulation using the bias potential results in a biased distribution along the reaction coordinate. If the sampling on each reaction coordinate is adequate, the biased distribution obtained from the MD simulation in each reaction coordinate window can be converted into an unbiased distribution and the unbiased free energy can be given by the following equation: [34]

$$A_i(\xi) = -k_B T ln P_i^b(\xi) - \omega_i(\xi) + F_i \tag{2.32}$$

P_i^b is the notation of the biased distribution on window i, while F_i is the bias potential term associated with window i and it is required to connect the free energy curves $A_i(\xi)$ from the different windows.

The combination of the free energy curves of each window and, by consequence the calculation of F_i, to obtain the PMF over all the ξ range, is done with the Weighted Histogram Analysis Method (WHAM) [86] which will allow the calculation of the unbiased PMF from our umbrella sampling system.

2.5.1 Steered MD

Umbrella Sampling uses a fixed bias potential in each defined window, which requires individual simulations for each defined window. To obtain starting conformations along the US pathway, we can apply a variable bias potential in a single simulation with the objective of driving the system from one state to another. This is called the Steered molecular dynamics (SMD) method [87], which is intuitive due to the steering nature of the simulation from state A to state B. This type of simulation can be used, for example, to force a molecule into crossing an energy barrier. However, one should evaluate carefully if such process is performed in an equilibrium or non-equilibrium way. If the variation of the bias potential is smaller than the relaxation time of the system, the simulation can be considered as an equilibrium SMD, otherwise we are in the presence of a non-equilibrium SMD [33]. The extraction of equilibrium properties, such as the free energy, could still be extrapolated from non-equilibrium SMD, even though it is uncommon and inefficient due to the creation of artefacts. Nevertheless, this approach is very useful to sample the US pathway and generate its initial conformations that are still going to equilibrated under each individual bias potential window, which will correct the SMD artifacts.

2.6 Simulation Settings

2.6.1 AAC System Setup

The first step of this work is to build the ATP/ADP carrier system embedded in the membrane. The two AAC structures were inserted in a pre-equilibrated 256 POPC lipid bilayer. Due to the large size of the protein, simply moving lipids and water molecules to accommodate the presence of the trans-membrane protein will not be enough. Hence the PyMOL software [88] was used to remove lipids from the membrane with the objective of opening the pore where our proteins were to be inserted. This was achieved by removing a total 56 lipids from both layers, resulting in a final membrane with 200 lipids (100 in each bilayer) solvated with around 11300 SPC water molecules [89].

To describe the complete transport process, we need the two conformational states of our protein (the C- and the M-state) from the *Bos Taurus* species. For the C-state, this is straightforward since a structure has been resolved by X-Ray Diffraction with a resolution of 2.2 Å (PDBid: 1OKC [38]). In this structure, the carboxyatractyloside and the cardiolipins were manually removed with PyMOL and the structure was used directly in the membrane embedding procedure. The AAC M-state is known to be less stable, specially using detergent solutions, which has set an obstacle for structure solving [48]. In early 2019, an M-state structure has been obtained for the organism *Thermothelomyces thermophila* with a single mutation to increase thermal stability (PDBid:6GCI [48]). Although the thermophila's AAC sequence only has 51% sequence identity with *Bos Taurus*, it is known that this protein structure is highly conserved across species. Therefore, homology modelling was employed using the thermophile organism as a template and the sequence of *Bos Taurus* AAC as a target. The MODELLER package [90] was used to achieve the sequence alignment and to obtain a model structure of the *Bos Taurus* M-state. The best model was chosen based on the best DOPE score and was then verified with procheck software, where ramachandran plots were drawn to validate the quality of our model [91]. After a visual inspection

on the overlapping structures (Figure 2.6) and confirming that no residues were sampling the forbidden regions of the ramachandran plot, we proceeded to the POPC membrane insertion protocol.

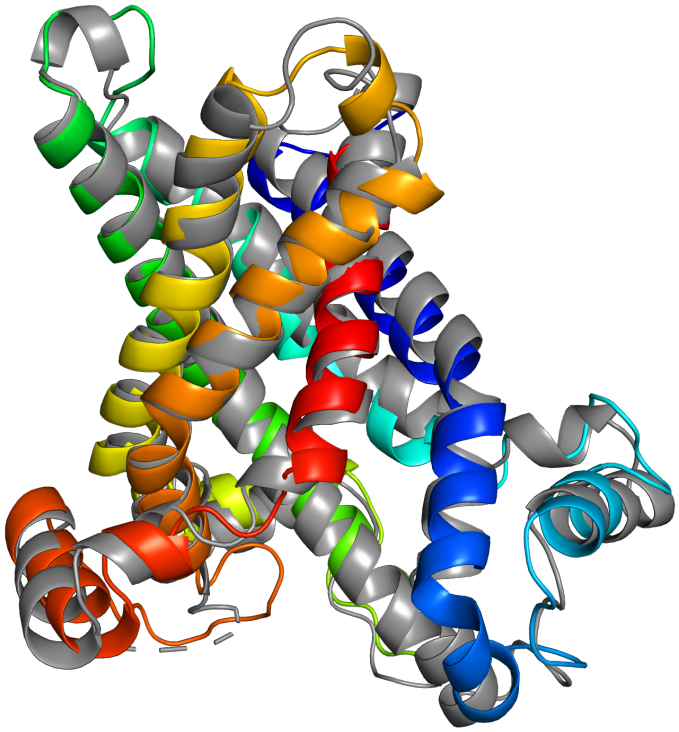

Figure 2.6: Structure alignment between AAC M-state template structure from *Thermothelomyces thermophila* (colored in gray) and the *Bos Taurus* M-state model obtained from homology modelling (colored with PYMOL Cα rainbow coloring).

2.6.2 MD settings

With both AAC states inserted into the membrane it is essential to execute MM/MD equilibration steps to achieve stable protein/membrane complexes. All MD simulations were ran using the GROMACS 2018.6 software [92, 93] and the GROMOS 54A7 force field [94]. Two different long range electrostatic treatments were applied, firstly Generalized Reaction-Field was used to achieve equilibration, however, this procedure led to structural instabilities in the simulations. These artifacts were avoided using the Particle Mesh Ewald (PME) method (see section below). The GRF/PME simulations used a cut-off of 1.4/1.0 nm for the long-range electrostatic and van der Waals interactions, respectively. The GRF was used with a group-based cutoff scheme, while PME relied on the Verlet scheme, which is atomistic. All bonds in AAC and POPC were constrained using the parallel linear constraint solver p-LINCS [95], while in water, the SETTLE algorithm [96] was used.

Due to the system complexity, we have developed a careful procedure to start the MM/MD simulations with stable protein/membrane system. This was achieved via three protocols that are as follows: first an energy minimization protocol of the system; second, a complex system initialization step to achieve a stable protein/membrane interface; and third, a long system equilibration step. With this procedure, we obtain stable system configurations ready for production.

The energy minimization procedure was performed in two steps using the steepest descent algorithm. The first step consisted on 20 000 steps with maximum step size of 0.001 nm without the application of

LINCS constraints. A second stage followed consisting of 10 000 steps with the same maximum step size. In this second stage constraints were applied to all bonds of the system.

After the minimization steps, we need to initialize the temperature and the pressure in our systems. However, the opening of a pore in the POPC membrane to insert the protein had a large impact on the system, which still needs to adapt to the presence of the protein. Also, the AAC cavity is empty without the presence of solvating water molecules. If a standard initialization step was performed right after minimization the system would crash instantly. Hence, a more complex system relaxation procedure was developed consisting of 5 different steps with different degrees of restraints. The first relaxation step had a duration of 50 ps with a time step of 0.001 ps. Velocities for the molecules were generated using the v-rescale thermostat and following a Maxwell distribution at 310 K. A temperature coupling strength of 0.01 ps was chosen. This step has the particularity of using freeze groups in the entire AAC, restricting any movement of the protein and allowing adjustments of the water molecules and lipids around the protein. Constraints were applied to the system by using the p-LINCS algorithm.

A second relaxation step was run for 200 ps with a time step of 0.0001 ps. Freeze groups were no longer applied on the AAC, being replaced by position restraints on the backbone of the protein with a strength of 10 000 kJ mol^{-1} nm^{-2}. Temperature coupling was changed from 0.01 ps to 0.1 ps. Pressure was applied on the system and it was kept semi-isotropically at 1 bar by the Parinello-Rahman barostat. A pressure coupling strength of 1 ps and a compressibility of $4.5e^{-5}$ bar^{-1} were used. Constraints remained the same as on the first step. The small time step used in this second relaxation of the system has the purpose of allowing very little time integrations to occur, permitting the system to slowly adapt to the new conditions. The pressure barostat will start to push water molecules inside the vacuum cavity of AAC, hence small time steps are ideal to prevent the simulation to crash.

Following the second relaxation step, which was critical, three more steps were used to progressively release the restraints applied on AAC. These last relaxation steps reverted the time step to its normal value (0.001 ps) and changed the pressure coupling to 2.0 ps. The only difference between them is the simulation time, position restrains and constraints applied. The third relaxation step was run for 100 ns with position restraints of 100 kJ mol^{-1} nm^{-2} for AAC backbone. No constraints were applied. The fourth relaxation step and fifth relaxation steps had a higher duration, 500 ps and 1 ns respectively, and p-LINCS constraint were applied once again. The position restraints were reduced and finally removed, being 10 kJ mol^{-1} nm^{-2} for the forth step and absent for the fifth step of relaxation, allowing for the entire system to be freed.

With the completion of this long initialization protocol, our system is ready for equilibration. Three replicates were done for each system (C- and M-state) and the MD simulations were ran for 500 ns with a time step of 0.002 ps. The long duration of our MM/MD simulations is required due to the long equilibration times expected for our complex system, in particular, to allow for convergence in the protein positioning along the membrane normal and in the cavity volume. This was further supported by the fact that the GRF-induced structural deformations were identified in these runs (see below).

2.6.3 GRF-induced Structural Instabilities

In the previous sections, we mentioned that two different methods were used to treat long range electrostatic energies, GRF and PME. In this thesis, we have started by using the GRF method to treat the long range electrostatics in all MM/MD simulations, which is the best and most compatible approach to be used with CpHMD simulations using GROMOS force fields [74, 97]. Using GRF would be ideal for CpHMD simulations since it can deal with a non neutral systems, which is highly advantageous where

the titrable sites can change protonation states and therefore cause fluctuations on the system charge.

In our system, a large membrane protein ATP/ADP Carrier was embedded in a POPC membrane, which, due to its size and complexity, should make users take extra attention on the course of the MM/MD simulations. The theoretical basis of RF and GRF methods is the approximation of the region outside the defined cut-off to a continuum medium. This should work well on solvated proteins, however, when simulating non-isotropic systems, like a protein inserted in a membrane, this becomes a large approximation. During our MD simulations with GRF, after 50–70 ns, we observed some high energy conformational transitions where the lipid head groups were sliding in between protein helices and approaching the AAC cavity (Figure 2.7). This effect could be explained by an inaccurate description of the electrostatic energies that can bring the negative portion of the lipids to have a high tendency to enter the highly positive AAC channel.

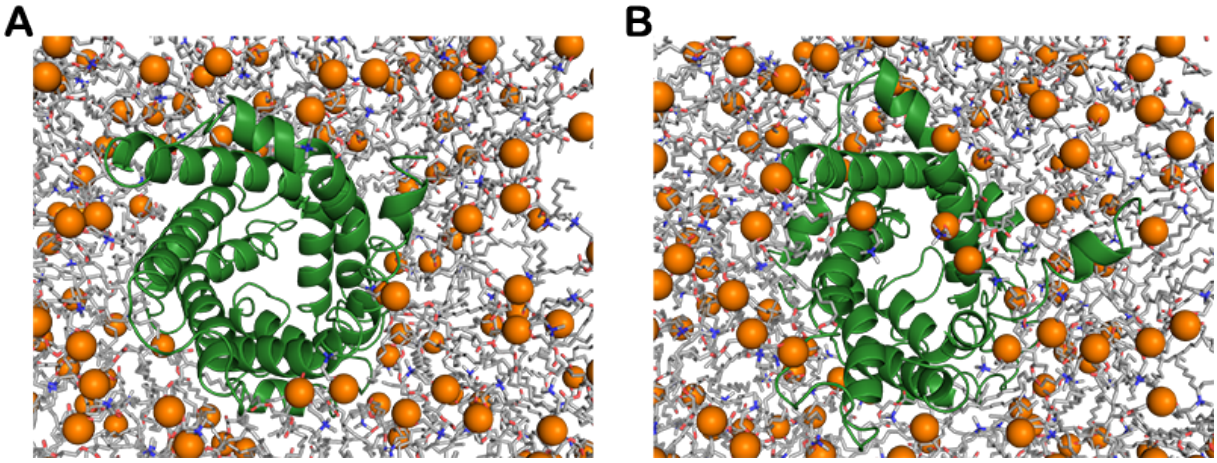

Figure 2.7: Structural representation of a simulation with GRF, at the start (A) and end (B) of simulation time, where lipid molecules are pushing into the ATP/ADP Carrier C-state cavity. AAC is represented as cartoon and colored in forest green, the phosphorous atoms of POPC lipids are shown as orange spheres while the remainder part of the lipids is shown as sticks with the carbons colored in gray.

On another MM/MD replicate, we observed a different effect where lipids were not driving towards the AAC cavity channel. Instead, the AAC C-state undergoes a drastic conformational transition where the channel opens, leading to a pore like configuration (Figure 2.8). This type of pore-like open conformation is still considered as a mechanistic possibility for this protein [98], even though it is highly unlikely due to the consequences it would have to the mitochondria.

The application of a cut-off method to decrease calculation times can introduce energetic artifacts [99]. If several charged groups are near the cut-off boundary, during the MD simulation they can be either treated implicitly or explicitly, depending on their position. That discontinuity is known to generate kinetic energy peaks [74] which can be tolling for an extra stable system. Our system possesses a high number of charged residues, leading to a net charge of $\sim +17\ e$, which increases the probability of finding charge groups at the cutoff boundaries and trigger these mentioned instabilities.

Due to the problems observed when using GRF in the treatment of electrostatics, we adopted the PME method in our simulations, which added an extra computational cost, as explained in the theory section 2.2.8. Furthermore, in CpHMD simulations it is important to bring the system charge close to neutrality. This means that in our preparation procedure for CpHMD, smaller preliminary CpHMD simulations are required in order to evaluate the total charge of the system at each pH value. Once the net charge of the system at each pH is estimated, it needs to be neutralized using counter ions so that, as stated before, the background charge distributed in the PME grids is minimal. Therefore, the implementation of the PME

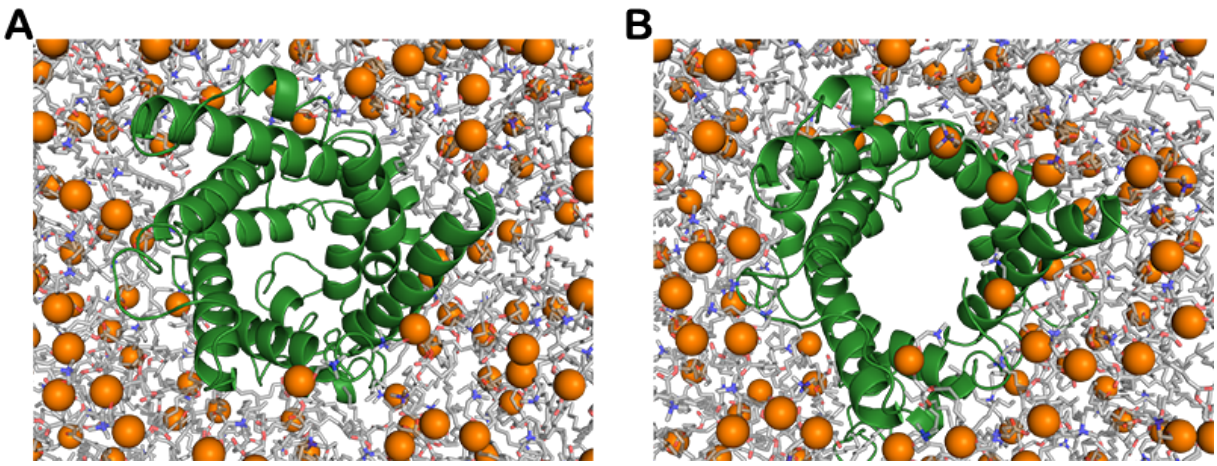

Figure 2.8: Structural representation of a simulation with GRF, at the start (A) and end (B) of simulation time, where an unnatural opening of the C-state AAC channel was observed. AAC is represented as cartoon and colored in forest green, the phosphorous atoms of POPC lipids are shown as orange spheres while the remainder part of the lipids is shown as sticks with the carbons colored in gray.

method comes with both a computational cost and an increased user management cost. However, in all PME simulations no system deformations nor artifacts have been observed, which supported our choice of long-range electrostatics treatment method.

2.6.4 Apo-AAC CpHMD settings

CpHMD simulations can be used to study protonation changes in AAC titrable residues at different pH values. All CpHMD simulations performed in this work used GROMACS 5.1.5 [93] with the GROMOS 54A7 force field [94]. Most MM/MD settings, including the long-range electrostatic treatment used in these simulations, were similar to the ones described in the MM/MD simulations (Section 2.6.2).

For the Poisson-Boltzmann calculations, required in the CpHMD method, the software Delphi v5.1 was used [100]. The atomic charges were obtained directly from the GROMOS 54A7 force field and the atom radii were derived from the available Lennard-Jones parameters of the respective atom type [101–104]. A dielectric constant of 2 was chosen for the solute while 80 was used on the solvent [28, 29]. To generate the molecular surface a probe of 1.4 Å radius was used. The simulations were executed with 0.1 M of ionic strength and with an ion exclusion layer of 2 Å. A two-step focusing procedure was performed with 1 and 0.25 Å for the large and focus grids, respectively. Relaxation values of 0.20 and 0.75 were used for the linear and nonlinear iteration steps with a convergence threshold of 0.01 kT e^{-1}.

Monte Carlo calculations were performed to sample the protonation states in our system using the free energy terms derived from the Poisson-Boltzmann calculations. The PETIT software was used in the MC calculations with the inclusion of proton tautomerism [105, 106]. 10^5 MC cycles were executed on each conformation, each corresponding to an attempt of changing each individual site and pairs of sites whose interaction is larger than 2 pK units.

In order to obtain pK_a curves of several AAC residues on both C- and M-state apo forms, CpHMD simulations of 100 ns at four different pH values (4, 5, 6 and 7) were performed. For each of these pH values, 3 replicates of the system were simulated in order to improve the sampling in such pH. Temperature and pressure conditions were the same as described in the MD settings (Section 2.6.2). The use of PME to treat the long range interactions added an extra step in our procedure. While using PME, it is advisable to keep the total charge of the system near its neutrality, which can be challenging

when using CpHMD and allowing several residues to freely change their protonation state. In order to determine how many counter-ions were necessary to add in each pH value, small CpHMD simulations of 20 ns were performed with the sole objective of estimating the number of counter-ions needed to bring our system near to neutral charge values. Even after the counter-ions were added to the system, the total charge was monitored through the whole simulation to ensure that no large deviations occur. We observed in some cases that 20 ns were not enough to determine the correct number of counter-ions, which required us to repeat the process until it converged. Table 2.1 summarizes the average charges in each simulation and the counter ions (Cl^-) added to bring the total charge near neutrality.

Table 2.1: Average total charge and number of counter-ions added to the CpHMD simulations of the apo-AAC system in both C-state and M-state at the four different pK_a values (4, 5, 6 and 7).

pH	Total Charge		Counter Ions
	C-state	M-state	(Cl^-)
7	+17	+18	17
6	+20	+20	20
5	+23	+25	24
4	+30	+32	30

2.6.5 ATP and ADP Parameters and pK^{mod} Calibration

We are interested in studying the transport process in AAC, hence, the ADP/ATP substrates are needed. Our initial problem comes from the fact that in the GROMOS 54A7 force-field only the protonated terminal phosphate ATP species has been parameterized. Therefore, in order to use this molecule in our CpHMD approach it would be needed to parameterize the fully deprotonated ATP species and then adapt the ATP parameters to build ADP by removing one phosphate group. Several approaches can be taken to obtain these parameters, but the most common is to rely on the ATB repository [107–109] and sometimes refine the charge sets using RESP procedures [102, 103]. Usually, the partial charges are obtained using well-solvated residues, which can be inadequate when the compounds are interacting in a poor solvated environment, like the AAC binding cavity. This has been exemplified in our recent work, published in mid August [110], which showed that RESP partial charges, designed for well solvated molecules, do fail when used in CpHMD simulations describing binding events. We showed that only a curated charge set for phosphorylated Tyrosine (pTyr) was able to predict the protonation equilibrium in the pY1021/PLC-γ1 complex [111]. Considering that our compounds also involve a phosphate group undergoing binding phenomena, we can adapt the atomic partial charges from the pTyr residue [110] into ADP/ATP.

We avoided the parameterization of the deprotonated ATP from scratch by using as a basis the ATP parameters present in the GROMOS 54A7 force-field. A minor adjustment was performed in the charges of the phosphate groups in order to circumscribe their charge (-1 *e* per unit) to the atoms assigned to the phosphate charge group. This process, simplified the removal of a phosphate group from ATP, building the ADP molecule. Relying on the transferability properties of our GROMOS force field, the charge set obtained for pTyr in our published work [110] can now be adapted and replace the charges of the last phosphate group in both ATP and ADP molecules. These parameters are available for the -1 (protonated) and -2 (deprotonated) terminal phosphate of these compounds since this group has a pK_a value of 6.3 and 6.5 for ADP and ATP, respectively [112]. The remaining phosphate groups have too low pK_a values, which at physiological pH, should remain ionized.

With our ADP and ATP parameterization complete, a pK_a calibration step is required, which consists in simulating the substrates free in solution using CpHMD and to calculate their pK_a shifts. The pK^{mod} values are the ones that when added to these shifts, it leads to the solution experimental pK_a values. In the ADP calibration procedure, simulations of 50 ns at pH 0.5, 1, 1.5, 2 and 3 were performed, while for ATP, simulations of 50 ns were performed at pH 5, 6 and 7. pK_a shift of 1.48 and 5.97 were obtained for ADP and ATP, respectively, which lead the pK^{mod} values of 4.78 and 0.53 for ADP and ATP, repectively. With the completion of the parameterization and calibration of the substrate molecules, they are ready to be used on the US-CpHMD simulations that would follow.

2.6.6 Enhanced Sampling Simulations

In the present work, we are interested in studying, via computational methods, the transport of the ATP and ADP through the inner mitochondria membrane mediated by the AAC protein. However, as previously discussed the time scale of regular atomistic MM/MD simulations is too small to encompass this entire process, which prompt us to resort to enhanced sampling techniques. We will be using the US method dividing our system in several windows covering the complete transport process. The first step to use these methodologies is the establishment of a reaction coordinate, which in our case, was the position of the susbtrate in the z axis of the simulation box, also corresponding to the membrane normal. By building US windows, covering the entire box height, we will be able to describe the passing of the molecule from one side to the other. For each US window, representative starting configurations of the system are required. The best method to obtain these starting frames is Steered Molecular Dynamics which can drive our substrate molecule along the complete reaction coordinate.

2.6.6.1 Steered Molecular Dynamics

Steered MD is a type of enhanced sampling simulation that applies a variable bias to drive the system from one state to the next state of the reaction coordinate. Using this technique, we can insert the substrate in our system and carefully pull it from one side of the membrane to the other. In a simple study of crossing a molecule through a lipid bilayer, this process would be easy and straight forward. However, since our substrate needs to cross a protein channel, it is more complex to achieve this without introducing conformational bias. To avoid this, we have inserted the substrate molecule (ATP was used in this procedure) near the bottom of the cavity in each AAC state. Then, instead of doing only one steered MD spanning the entire z axis of the box, we performed separated SMD simulations, always pulling the substrate from the bottom of the cavity to the water phases (either cytoplasm or matrix sides). Joining the sampled configurations obtained with the two halves of the system, we can span the complete reaction coordinate.

This approach allows to pull ATP from both the terminal phosphate and the Adenine side, resulting in configurations where ATP is either entering or leaving the mitochondria. The substrates were inserted into the AAC cavity with either the phosphate pointing to the center of the cavity, mimicking the entry of the substrate into the pocket, or with the adenine group pointing to the center of the cavity, which mimics the exiting process. The complete process results in four sets of configurations, substrate with phosphate pointed to the cavity on both C- and M-state (Figure 2.9A and B) and substrate with adenine pointed to the cavity in both C- and M- state (Figure 2.9C and D). These provide the necessary halves of the system which can be coupled together with the US-CpHMD methodology and recover the complete transport process. To simplify and to reduce the computational cost, we only perform these SMD simulations for ATP, which could later be converted into ADP by editing the gro files.

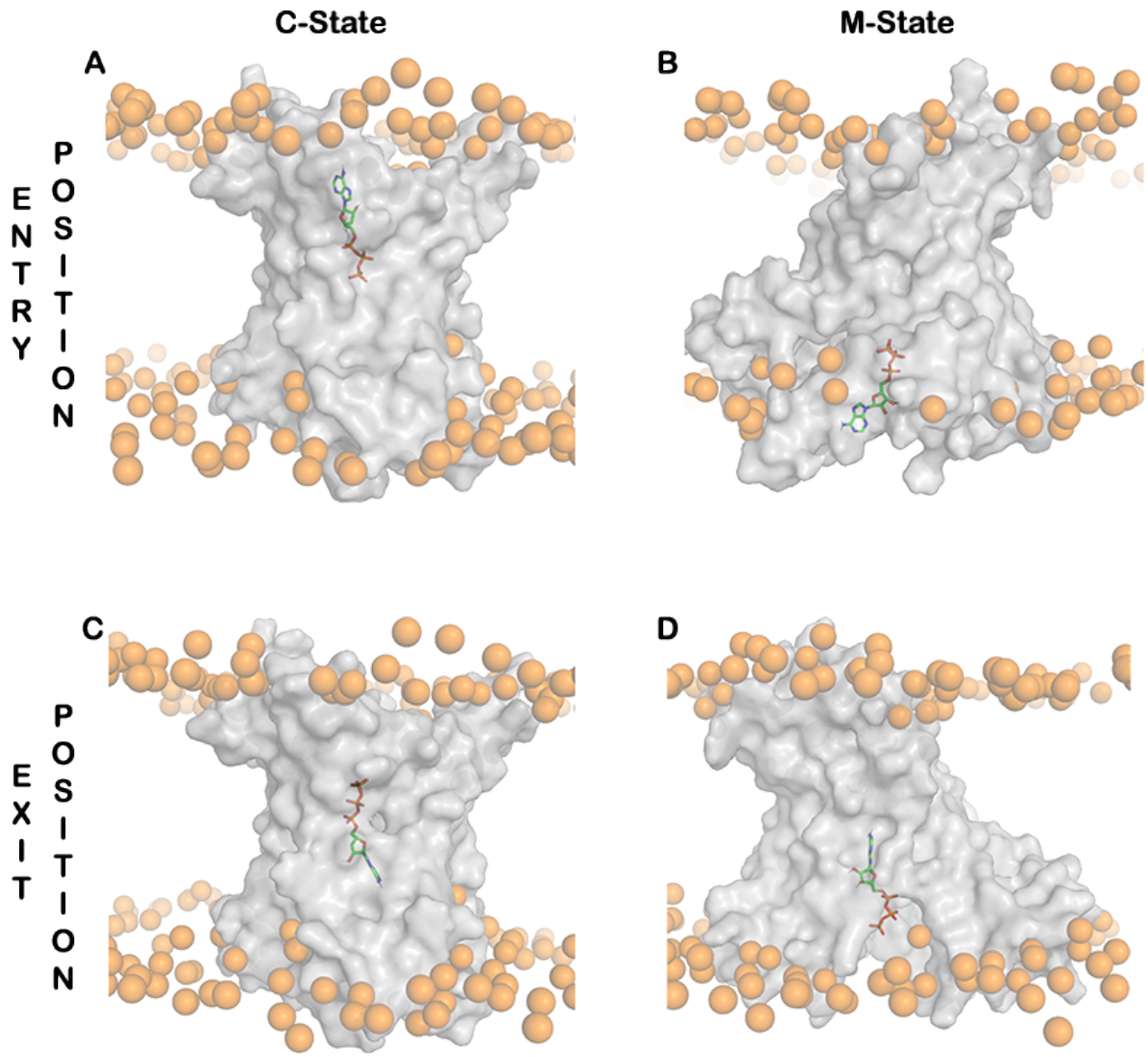

Figure 2.9: Starting configurations of each Steered MD run. Panels A and B represent AAC in the C- and M-states, respectively, with the substrate in entry position (pointing the phosphate group to the bottom of the cavity). panels C and D show the C- and M-state of AAC, respectively, with the substrate in exiting position (having the adenine group pointing to the bottom of the cavity). AAC is depicted as a grey transparent surface and the lipids phosphorus atoms are represented as orange spheres.

The parameters used are very similar to the ones in the MM/MD simulations, with the exception of the simulation time and the pull code section. In SMD, the simulation time is coupled to the pull code and should be chosen to define the distance travelled along the reaction coordinate. Since the simulation stops when our substrate reaches the end of the simulation box in the z axis, we can simply define a large time. In the pull code, we have used two groups, the $C\alpha$ atoms used to define the center of our protein channel (Figure 2.10), which we conventionally set to zero in the transport coordinate, and one of the previously mentioned substrate pulling regions (last phosphate or adenine group), which are chosen depending on the direction. We defined the pull force type as umbrella and set the "direction" geometry with a force constant of 1000 kJ mol^{-1} nm^{-2}. The rate of moving the substrate group was 0.025 nm ns^{-1}, which was crucial in these simulations. If the pull rate is too large our molecule will move quickly in the z axis and could drag portions of the channel, creating deformations in the AAC protein. Choosing a high enough rate that is still able to avoid such artifacts is very important in this methodology.

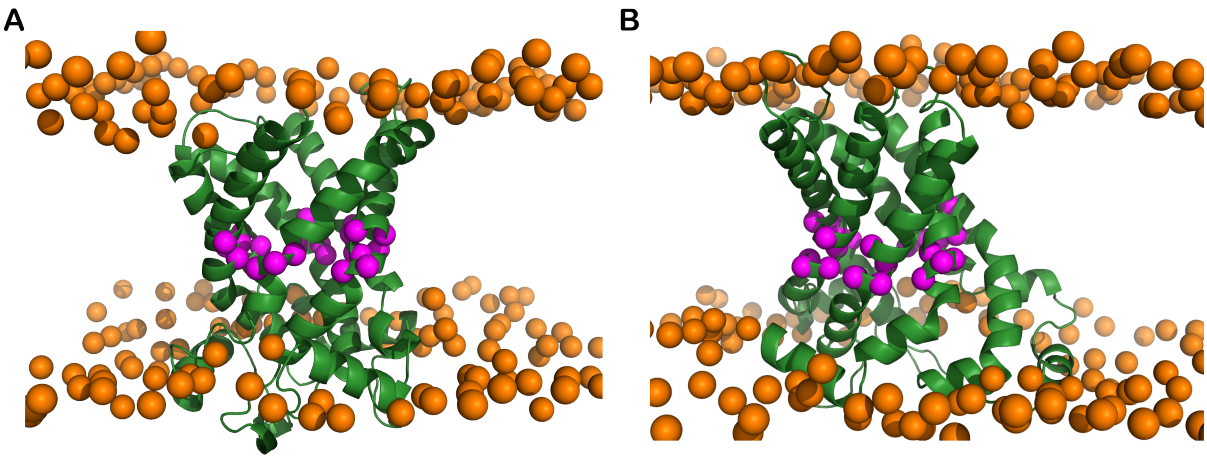

Figure 2.10: Structural representation of the $C\alpha$ atoms of the C- (A) and M-state (B) of AAC, selected to define the protein center. The $C\alpha$ atoms are colored in magenta and represented as spheres, the AAC structure is shown as a cartoon and colored in forest green and the phosphorus atoms of the lipids are represented as orange spheres to identify the membrane region.

2.6.6.2 Umbrella Sampling – CpHMD

The initial step in setting up a US protocol is the definition of the umbrella windows positions and the force constants used in each. Unlike SMD, in the US simulations, we apply a potential bias to keep our molecule in the region of the defined umbrella along the reaction coordinate. Throughout our US simulations the bias potential was always applied to the last phosphate of each substrate, ATP and ADP. The user defined AAC center is located at the zero of the transport coordinate and our simulation box spans distances between~ -4 nm and $\sim +4$ nm. Therefore, building US windows at 0.2 nm intervals would lead to 40 independent simulations, which is very computationally expensive. To circumvent this, in both extremes of the reaction coordinate (-4 nm to -2.4 nm and $+2.4$ nm to $+4$ nm), we increased the space between windows (Figure 2.11). Due to the asymmetry of the AAC, in the cytoplasmic side it resulted in 4 windows (-3.9, -3.5, -3.0 and -2.7 nm), while in the matrix side, it resulted 6 windows ($+2.6$, $+2.8$, $+3.0$, $+3.3$, $+3.6$ and $+3.9$ nm).

The force constants applied in each US window should be carefully chosen to obtain a good sampling along the transport coordinate vector. A high force constant of 1000 $kJmol^{-1}nm^{-2}$ was applied to our US windows in the range -2.7 to +3.0 nm, which is the region more prone to deformations due to the stronger interactions between the substrates and the AAC protein. The application of a larger force prevents the substrate position to deviate too much, while the narrower sampling histograms can be brought together by decreasing the separation between windows (0.2 nm). At the two ends of the transport vector, the interactions are usually milder and often with water. This allows the separation of the US windows coupled to smaller force constant values to still favor the overlap between windows. The substrate can sample conformations further from the reference reaction coordinate of that window, filling all gaps between US windows. The complete set of force constant values used in the US windows is available in Table 2.2. During our US simulations of the extremity windows (-3.9 and $+3.9$ nm), there were cases in which the substrate moved beyond the simulation box limits causing the simulation to crash. To prevent this, the simulation force constant of those windows was increased to 500, which decreased these events.

A very important objective in this work is to sample the import and export of both ADP and ATP, making a total of 4 different processes. In each of these processes, we need to use umbrellas obtained from the two AAC states (the M- and the C-state), to describe correctly the conformational transition that is coupled to the transport process. To obtain the PMF energetic profile two US-CpHMD simulation sets

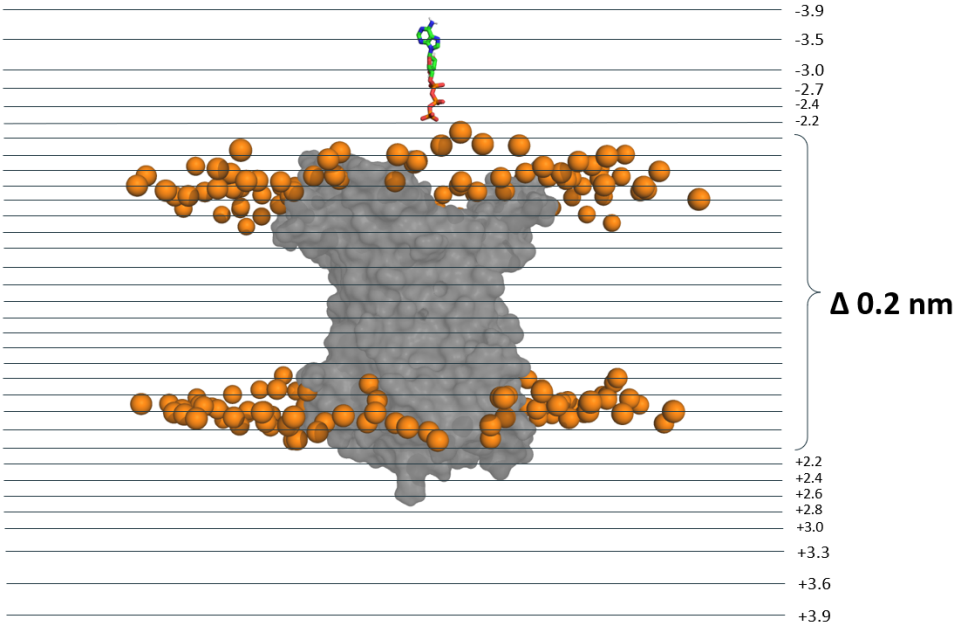

Figure 2.11: Umbrella Sampling windows used in our work. The AAC protein is shown as a gray surface, the lipid phosphorous atoms are in orange spheres, and ATP is shown as sticks.

Table 2.2: Force constant values assignment to each US window along the substrate transport vector.

Position (nm)	**Force constant** ($kJmol^{-1}nm^{-2}$)
-3.9	250 or 500
-3.5	250
-3.0	500
-2.7	500
-2.4	1000
-2.2	1000
all windows spaced at 0.2 intervals	1000
+2.2	1000
+2.4	1000
+2.6	1000
+2.8	1000
+3.0	1000
+3.3	500
+3.6	250
+3.9	250 or 500

were performed, a first one that described the transport of the substrate without the occurrence of conformational transition (Figure 2.12A), and secondly a set of simulations that included the conformational change that occurs when the substrate reaches the bottom of the cavity (Figure 2.12B). In regards of the first set, the complete US scheme applied to the C- and M-states will provide US windows that are realistic (the substrate will be on the open side of AAC) and others that probably are not (the substrate will be on the closed side of AAC). Although these later windows do not represent the *in vivo* process, they are very useful to identify which windows define bottom of the cavity and which are high energy, requiring the conformational transition to be correctly sampled. This information is crucial to help us define which windows from each state should be used when building the complete substrate import/export process.

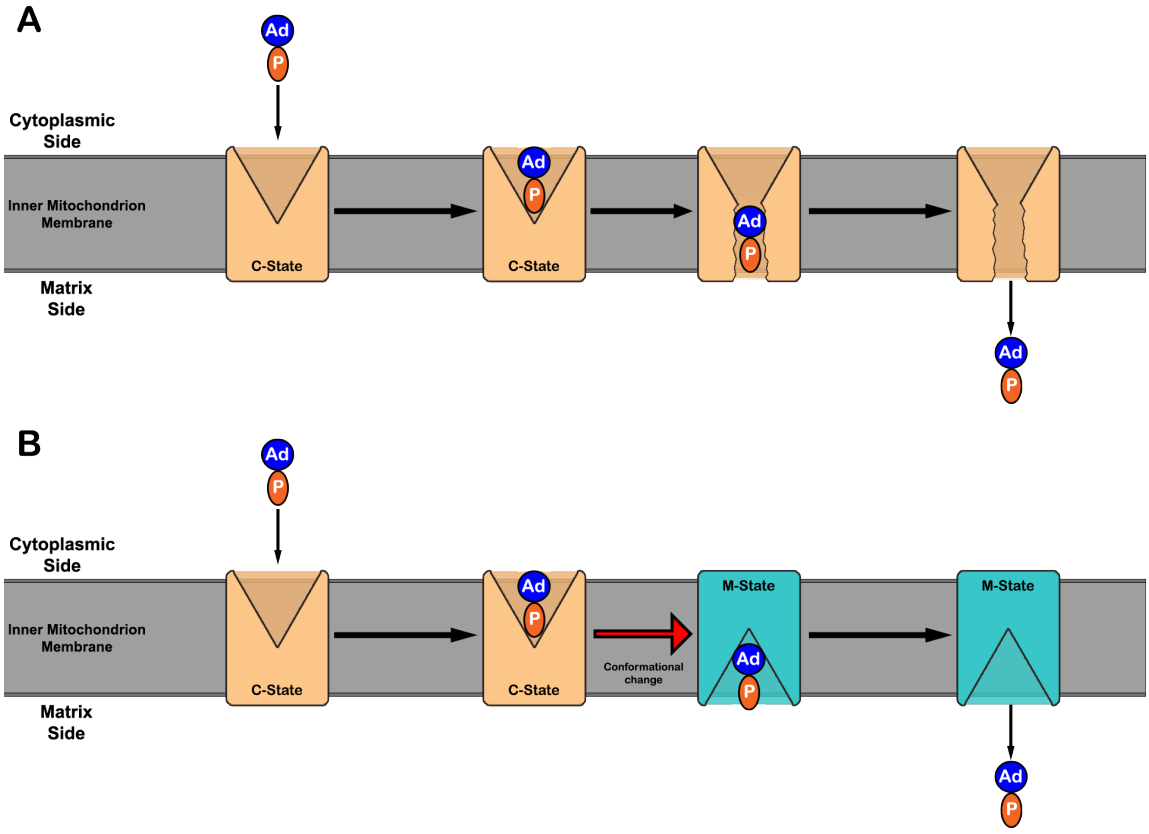

Figure 2.12: Schematic representation of substrate transport, shown for the import process only, without conformational transition of AAC (A) and with conformational transition (B). The inner mitochondrial membrane is represented by a gray rectangular area, AAC's C-state and M-state are represented as an orange and light blue rectangle respectively, the substrate is shown as a blue sphere for adenosine and an orange ellipse for the phosphates, representing both ATP and ADP.

After identifying which US windows lead to an energy increase, we substitute all these and the next windows by the equivalent in the other AAC state. However, we need conformations where the substrate is in an exiting position, meaning it is pointing the adenine region to the center of the cavity and the phosphate groups to the solvent. Our preliminary results indicate that the position +0.2 nm was the bottom of the cavity. Therefore, to obtain the exiting paths starting from the center of the cavity, we only need to simulate the US windows with the substrate in exiting position from +0.2 nm to -3.9 or +3.9 nm in the export and import processes, respectively. Also, since after $\sim$2.0 nm from the center (-2.0 nm for export and $+2.0$ nm for import) our substrate should have enough rotation freedom, then we can use the previously umbrellas, independent of the orientation. This way, we can skip 5 of the 17 US simulations, per process and per substrate, which corresponds to a significant computational saving.

All US simulations were performed with the CpHMD method at pH 7.0 for 150 ns using the same settings as described for the apo-AAC study (Section 2.6.4). From the apo-AAC simulations, we obtained titration curves and pK_a values of all pH sensitive residues in the protein. Using this information, and to reduce the simulation time, we have restricted the number of residues that are allowed to titrate in the US protocol (Figure 2.13). Only aspartate, glutamate, and histidine residues were allowed to titrate, since all tyrosine and lysine residues will have their standard charge (neutral Tyr and charged Lys) stabilized by the presence of the substrate (ADP/ATP).

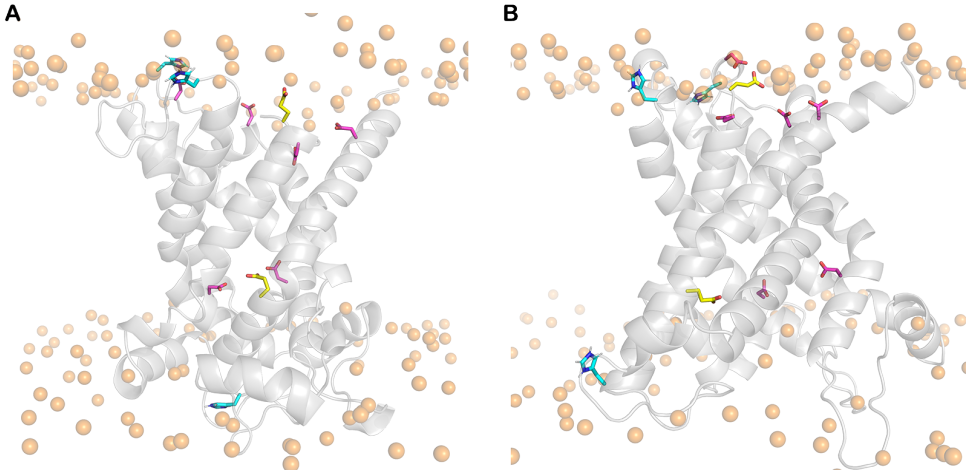

Figure 2.13: Structural representation of the C- (A) and the M-state (B) AAC. The titrable residues are shown in sticks. Aspartates are shown in yellow, glutamates are shown in pink and histidines are shown in cyan.

2.6.7 Simulation analysis

All analysis of the results were performed using tools available in the GROMACS package [93] or in-house tools. Only the equilibrated portion of both CpHMD and US-CpHMD simulations were used when extracting equilibrium properties. On the latter analysis, when calculating the PMF for each simulation scheme, the Weighted Histogram Analysis Method (WHAM) [86] was used via the integrated GROMACS 5.1.5 package tools [93]. When obtaining structural representations to aid the molecular visualization of our discussion the PyMOL software [88] was used and in order to plot the several graphics presented in our work the gnuplot software [113] was employed.

When errors are presented throughout this work, two approaches have been taken. In the CpHMD simulations, since we have made independent replicates, the errors were calculated using the standard error of the mean between the three performed replicates. In the US-CpHMD simulations, we have not yet performed replicates of each window due to the high computational cost it entails. As a consequence, the properties extracted from these simulations are time-correlated. Therefore, the calculations of the standard error of the mean required the use of an autocorrelation function to determine the number of independent blocks in the simulation. In this approach, we have considered the data to be uncorrelated once the autocorrelation function reached a value of 0.1.

Chapter 3

Results

3.1 Simulation Equilibration

3.1.1 Molecular Dynamics Equilibration

Studying large systems like the AAC transmembrane protein inserted into a POPC membrane requires extra attention on its equilibration step. Since the embedding of our protein into the membrane was executed manually with the PYMOL software, it is imperative to follow several equilibration properties throughout our long MD runs to carefully verify the stability of the system. If we do not achieve an acceptable equilibration of our membrane system, we might need to reconsider the approach used to study the transport performed by ATP/ADP carrier. Such was the case when simulations were ran using the GRF treatment for long range electrostatic interactions. In those simulations, the lipid molecules entered the AAC cavity disrupting its integrity, hence it triggered the change from the GRF treatment to the PME (Section 2.6.3).

To evaluate the equilibration state of our system during MD simulations the typical analyses were performed (RMSD, secondary structure analysis, system x/y area, etc). Looking at the RMSD of both AAC states, we can notice that in the C-state (Figure 3.1A) all three replicates maintained their RMSD values low and considerably stable throughout the simulation. On the M-state, although the values are also small, it is noticeable that the first replicate has deviated its conformation from the other two replicates (Figure 3.1B).

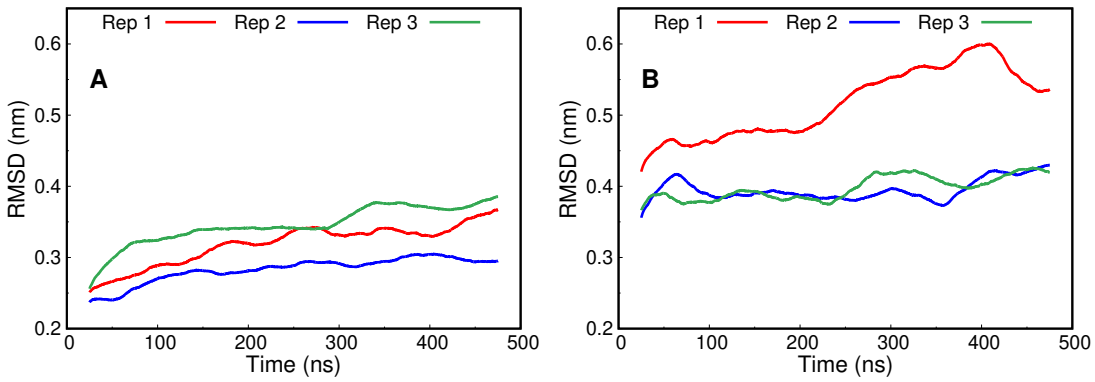

Figure 3.1: RMSD of the long MD runs. Each step is structurally compared with the structure obtained in the initialization procedure of the respective replicate. The RMSD of each replicate in its respective AAC state, C-state (A) and M-state (B), was plotted using gnuplot [113] and a floating window of 50 ns was used in order to reduce local fluctuations of each replicate curve.

The slight deviation observed in replicate 1 of the M-state RMSD profile is also present when plotting the percentage of helix content along the simulation. In this replicate, a decrease of around 4% in helix content can be seen, which seems to be coupled to the increase of RMSD (Figure 3.2B) The remainder of long MD simulations performed had only slight fluctuations on their values, giving us some confidence on the equilibration of our simulations.

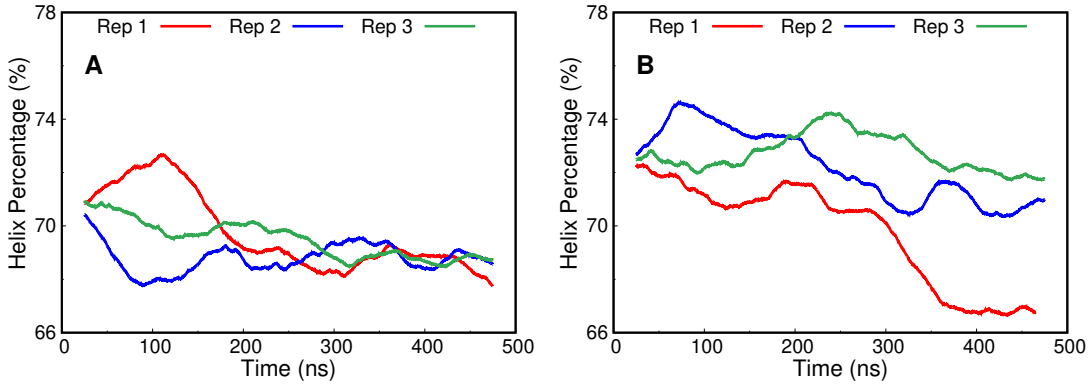

Figure 3.2: Percentage of helix content of the C-state (A) and M-state (B) during the MD runs, obtained using DSSP calculations [114, 115]. On both plots a floating window of 50 ns was used in order to reduce local fluctuations.

To investigate the reasons that lead the M-state first replicate to deviate from the other two replicates, we looked at the AAC structure during the simulation time. Comparing the M-state replicates, the h1-2 matrix α-helix (residues 53-64) and the loop preceding the even-numbered H2 helix (residues 65-73) seem to be the main culprits. Loosing secondary structure in a region consisting of the loop between the matrix helix h1-2 and the even-numbered helix H2 might not be that impactful, since such connecting region is highly flexible. The same cannot be said about the loss of secondary structure on the matrix helix h1-2. In fact, in the first M-state replicate it was seen that helix h1-2 (cyan region of Figure 3.3A) unfolded significantly at the end of the simulation (cyan region of Figure 3.3B). Looking at all three replicates in their last conformations, it becomes clear that the observed deviation in helical content comes from these regions. In replicates 2 and 3, the residues 65–73 retain their overall helical fold (Figure 3.4), which will be often accounted as α-Helix by DSSP. The presence of relatively unstable and possibly transient helical structure in this loop region might be related with the use of homology modelling to obtain our structure instead of experimental methods. Either way, despite the slight loss of structure of the matrix helix h1-2 in the first M-state replicate, we have opted to proceed with its analysis since the affected region is not directly related with the binding cavity. However, this replicate will be kept under scrutiny when its structure is used in following simulations.

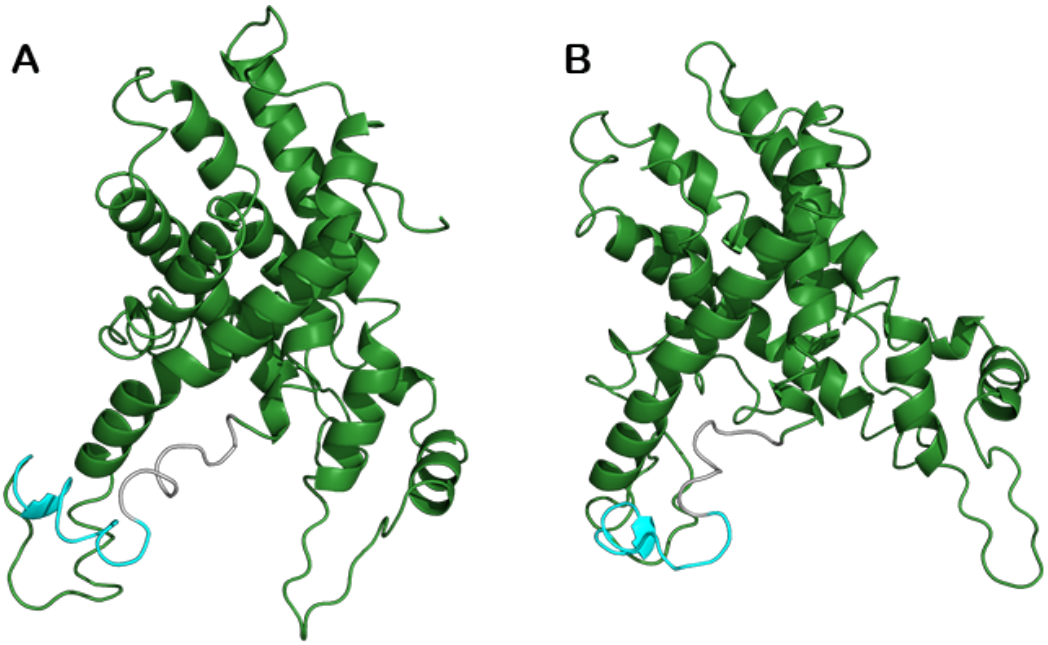

Figure 3.3: Structural representation of ATP/ADP Carrier M-state structure of the first long MD run replicate. Structures at the start (A) and end (B) of the simulation with h1-2 matrix helix highlighted in cyan and the connecting region of h1-2 to H2 in gray. Both structures were drawn on PYMOL software [88].

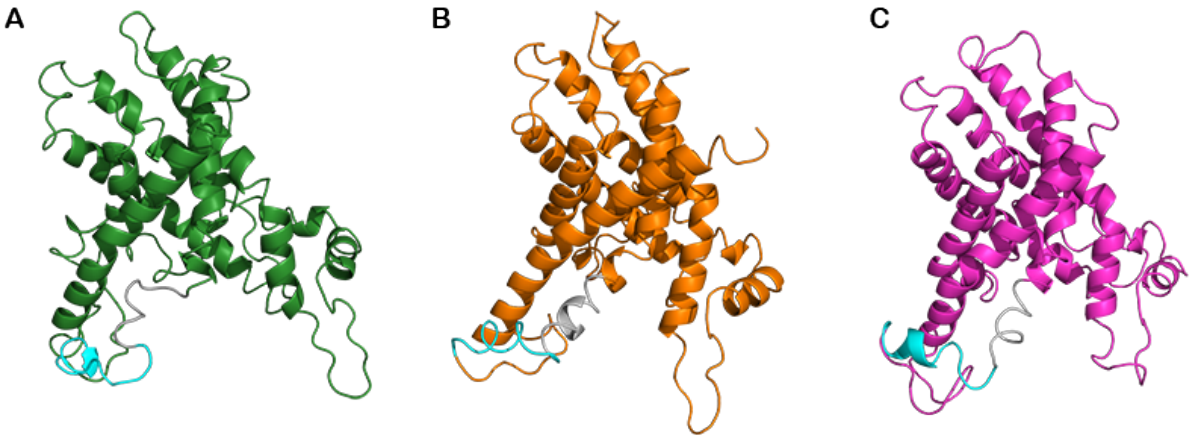

Figure 3.4: Structural representation of ATP/ADP Carrier M-state last conformation of each replicate. The first (A), second (B) and third (C) replicate had the AAC protein represented as a cartoon colored in green, orange and magenta, respectively. The h1-2 matrix helix was highlighted in cyan and the connecting region of h1-2 to H2 in gray. All structures were drawn on PYMOL software [88].

When dealing with membrane systems an important measure to follow is the variation in the system *xy* area. With this approach, we can follow any deviation in the area values and, upon convergence, it confirms that the lipids have been able to accommodate to the transmembrane protein. All our simulations show only small deviations in the initial nanoseconds (Figure 3.5), which suggests that the protein/lipid interface was already near equilibrium. We observe some variations on the property in some replicates, which are probably coupled with more pronounced conformational transitions in the protein, like the case observed for replicate 1 of the M-state (Figure 3.5B). Nevertheless, all these fluctuations seem transient, since at the end time the area has returned to the normal values.

In a membrane-inserted protein system, we can also look at its position relative to the center of the membrane. This can provide a measure of the convergence along the membrane normal. To use as a

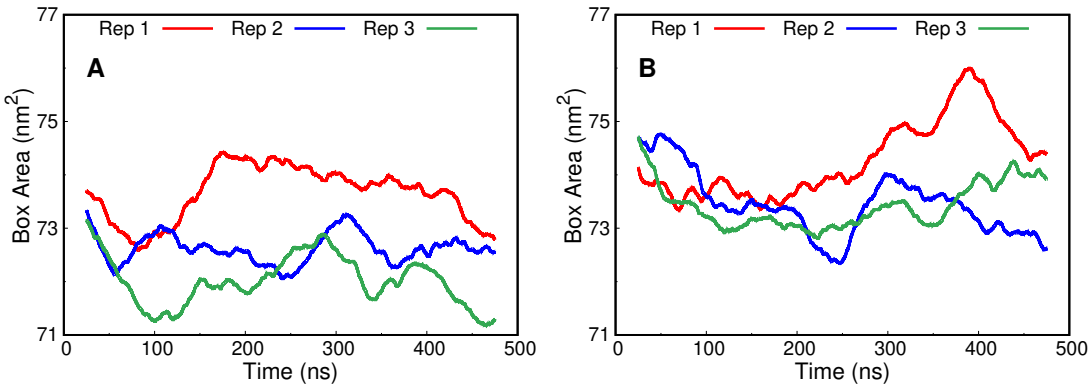

Figure 3.5: Area of the simulation system in the *xy* axis over time. Both C-state (A) and M-state (B) have each individual replicate plotted with a floating window of 50 ns in order to reduce local fluctuations.

reference in the protein, we have manually selected several Cα atoms near the mid point of AAC. This defines a hypothetical center of our protein to be used in the analysis and corresponds to the center used later on in many other approaches. However, note that this does not correspond to either the center of mass nor the geometrical center of our protein. Regarding the definition of the center of the membrane, this was achieved by using the average position of all z coordinates of the POPC phosphorus atoms. By comparing the z coordinate difference between the protein and membrane reference groups, we are able to follow the variations of the carrier position along the membrane normal. All simulations show that this structural property is well-converged (Figure 3.6) and that slow fluctuations exist, like in replicate 1 of C-state ($\sim$300 ns), but they are small in magnitude and tend to return to the original positions. All these results indicate a very good convergence in the protein-membrane insertion axis of system.

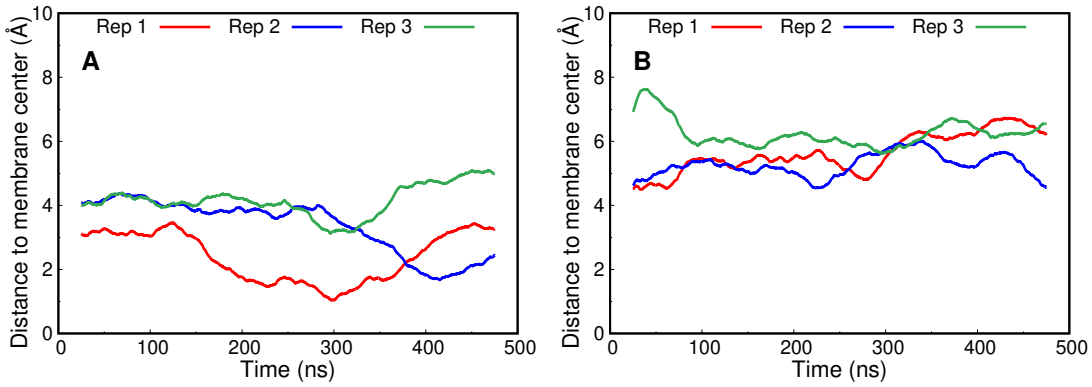

Figure 3.6: Distance to membrane center of the AAC center group over time in both the C- (A) and the M-state (B). The distance values were plotted using a floating window of 50 ns in order to reduce local fluctuations.

Overall these extensive MD runs have served the purpose of verifying the integrity of our protein in order to guarantee the successful insertion of the transmembrane protein into the POPC membrane. We observe only small conformational changes in the protein and small fluctuations in the membrane biophysical properties, which suggests good convergence and gives us confidence to use these structures on the following simulations.

3.1.2 CpHMD Equilibration

Adding pH effects using CpHMD simulations, can have a large impact on the structure of proteins. Our system is a large transmembrane protein abundant in protonable amino acids, so it is important

that we also confirm the CpHMD simulations convergence. Accessing the integrity and stability of our protein at each pH is important to make sure that the obtained protonations reflect a healthy apo form of AAC instead of a compromised one.

When examining the RMSD plots throughout the 100 ns of CpHMD simulation, we observe most systems promptly reaching equilibrium, exhibiting low RMSD values and without deviations (Figure 5.1). In some replicates, we see a slight increase in the RMSD values in different points of the simulations, but the deviations are still within 2 Å, suggesting that they correspond to helix or loops small movements. Regarding helicity, also no abnormal loss of structure was observed during the CpHMD simulations (Figure 5.2). As expected, the first replicate of the AAC M-state showed slightly lower helical contents, reflecting the conformational transition that took place in the MD equilibration. These results confirm a fast equilibration of our system in the CpHMD simulations and also highlights the success of the long 500 ns pre-equilibration procedure. We also performed the system area and the membrane insertion analyses, only to confirm the equilibrated state of our protein (Figures 5.3 and 5.4).

We used PME to deal with long-range electrostatics on our CpHMD simulations, for which box neutrality is essential. Therefore, the total charge calculation of our system at each pH value will confirm whether the number of counter-ions added to the simulation was sufficient to bring the system near neutrality (Figure 3.7). Furthermore, the total charge data over time also provide an important convergence measurement of the simulations. For all systems, equilibrium is achieved relatively fast ($\sim$20 ns) and only small charge fluctuations are observed. These results confirm that both the protonation and conformation spaces, which are coupled, are converged after these initial segments.

Overall, we confirm that all calculated properties of our system quickly converge and equilibrium is achieved. The 500 ns pre-equilibration MD run, which preceded the CpHMD simulations, may have had a positive impact on these results. The equilibration time of our system, which is the time our system takes to adapt to the new conditions, is $\sim$20 ns, hence for all the equilibrium properties extracted from these simulations, only the frames from 20 ns onwards will be used.

3.1.3 Umbrella Sampling Equilibration

The US method is distinct from the previous two in the number of simulations it entails. We have executed four different schemes of Umbrella Sampling, two following the substrate import (ATP and ADP) from the cytoplasm to the matrix and two following the export process. Each of these US schemes contains 35 simulations of 150 ns (one for each US window), which makes the analysis of equilibrium properties a bit challenging. The simplest approach is to perform the equilibration analysis on all US windows and plotting the data altogether, in an attempt to pinpoint outlier simulations. If we find simulations deviating significantly from the distribution, we could argue that they are deviating from equilibrium, which would require a deeper analysis of the events that lead to such variation. Keeping this in mind, we also need to address the difficulty of our large system to acquire equilibrium. Its large size and the presence of the substrate in the cavity can cause the system to slightly change its conformation. However, as long as these conformational changes are not significant, convergence should not be an issue.

As initial analysis of the RMSD data show that most US windows quickly converge to an RMSD value of 0.2–0.3 nm (Figure 5.5). There are a few deviations to this pattern that stand out from the most common profiles, with a slight prevalence in the ADP US simulations of both C and M-state (Figure 5.5B and D). However, these are very small deviations, reaching only 0.4 nm which are easily attributed to loop rearrangements in the protein, induced by interactions with the substrate (+2.6 nm US

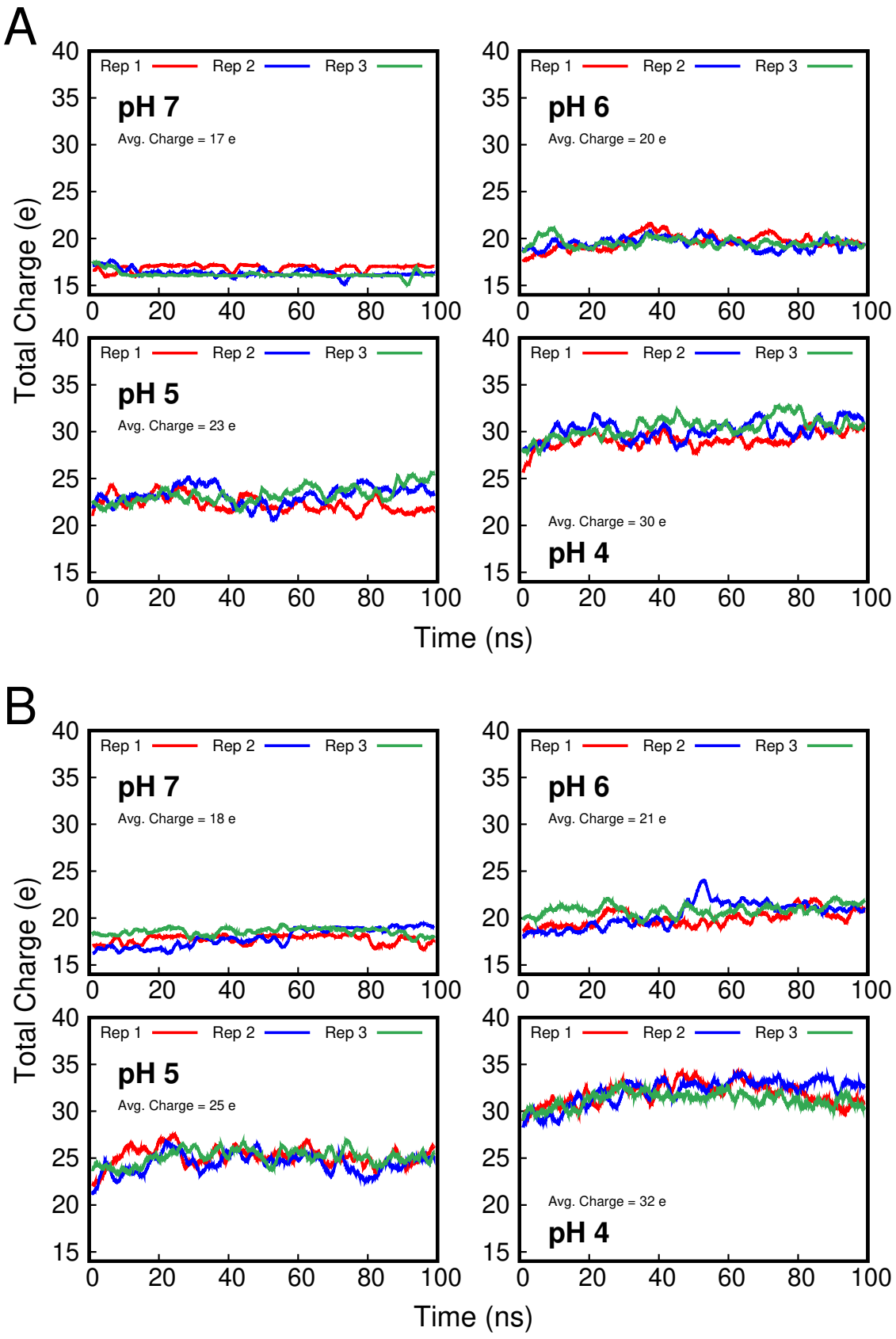

Figure 3.7: Total charge of the C- (A) and M-state (B) AAC during the CpHMD simulations at four different pH values. A floating window of 2 ns was used to remove local fluctuations.

windows in ADP, corresponding to the the green lines in Figure 5.5B and D), or to small changes in the protein cavity to accommodate the substrate molecules at its center.

The helical content of AAC in these simulations was also analysed and confirmed that the protein secondary structure, which is mostly α-helical, is very stable throughout the simulations (Figure 5.6). Even though there are only very small deviations during the simulations, with probably only one exception (see below), we observe two different clusters of helical content, one group with $\sim$71% helix and the other with $\sim$68% . This is a very small difference between the simulations and is due to the conformational change that we observed in the MD simulations (Figure 3.2). In those equilibration simulations, the C-state AAC converged to $\sim$69% of helix content, while the M-state (in replicates 2 and 3) converged to a higher value ($\sim$72%). Since in US scheme we are using both states sequentially, to mimic the conformational transition that happen in the transport process, it is expected that these two different regimes will stand out.

Our AAC protein is clearly very stable during the transport process. However, there is a small outlier in both ATP simulations (Figures 5.6A and C). In the -2.4 nm window in the cytoplasmic side (colored black in the plots), a decrease in helical content can be seen at $\sim$110 ns. Please note that this is the same simulation in both plots, since at this distance from the protein cavity, the substrate has full rotational freedom and there is no distinction between the initial steps of the import process and the final steps of the export. The small decrease in helicity observed is due to the interaction of ATP with the end portions of two helices on the cytoplasmic side of AAC. Hence, this loss of structure is very mild and does not seem to be related with structure unfolding or denaturing events.

Similar as previously done, we also followed the box area (Figure 5.7) and the distance of AAC to the membrane center (Figure 5.8) in order to evaluate the convergence and stability of the protein/membrane interface during the transport process. The system area provides a measure of the lipid packing with AAC, while the distance to the membrane center evaluates the protein fluctuations along the membrane normal. In all cases, we observe an almost perfect distribution of the US windows data over time, with only marginal and transient fluctuations out of the distributions. These results confirm that these biophysical properties, related with the AAC/POPC interface, converge very fast and are not significantly altered during the transport processes.

The main goal of using an US method to describe our process is the extraction of free energy profiles for the substrate over each step of the transport. To achieve this, a reaction coordinate is chosen to describe our process and is divided in several steps to allow the representation of the different states. To describe the transport of the substrate, we have chosen as our reaction coordinate the distance in the z axis between the substrate terminal phosphate group and the user defined center of the AAC cavity. This has allowed us to slice the entire route of the substrate from one side of the membrane to the other in 35 windows, enabling us to obtain free energy values along this reaction coordinate. In order to use the umbrella energies of each biased window and calculate the unbiased free energy profile, which is a potential of mean force (PMF), for the transport process, two conditions must be met. Firstly, the sampling in each US window must be good enough to describe correctly the conformational space of the interaction between solute and AAC at that position of the reaction coordinate. Secondly, the umbrellas chosen need to be close enough to assure the sampling continuity between umbrellas, which is pivotal to provide reliable PMF energy profiles of transport process.

For the first condition, focused on obtaining good sampling of the conformational space of the complex, the best solution is to perform long CpHMD simulations in each umbrella and, if possible, to perform replicate simulations of each scheme. We focused on increasing the simulations length (150 ns), since doing replicates, even if just triplicates, would be too computationally demanding within the 1 year

duration of this thesis. To verify the sampling between the different umbrellas and assure the necessary overlap for the PMF calculations, we can plot histogram distributions of the distance to the reference group (Figure 3.8 Left). In these plots, we can identify the regions where the sampling is reduced and make sure that there are no regions without overlap between adjacent histograms. If there were gaps between the data distributions, the energy profiles obtained in that region would not be as accurate and large error bars would appear in the PMF calculation. These problems are usually corrected by adding extra US windows covering that gap in the reaction coordinate. It is also very useful to observe the time series data from which the histograms were obtained (Figure 3.8 Right). In these plots, we can identify possible problems, like situations where the solute, at some point of the simulation, is dragged from its minimum umbrella position and stays there for a long period of time. This case would suggest a lack of sampling and that we would probably need longer simulations and/or different force constant values in the US scheme.

In our distance distributions of the water phase, corresponding to the ones on the far left and far right, the histograms cover a larger area of sampling. This is due to the smaller force constant used in these windows that allowed the substrate to sample regions further away from the fixed reaction coordinate. At these distances from the AAC it is assumed that the substrate is in solution and/or only interacting marginaly with AAC. Therefore, it is beneficial to allow it to cover a larger sampling region in order to reduce the number of US windows and save computational resources. The opposite is observed when the substrate enters the AAC cavity, where the histograms become thinner, due to an increased constant force, which counteracts the strong interactions between the solutes and AAC, and prevents large deviations from the reaction coordinate.

In overall, we do not observe very large deviations in our US simulations, even though some regions could still benefit from more sampling. One of such regions can be seen in the ADP import simulations at the 0.0 nm umbrella histogram, which deviated to positive values and reduced the sampling between this and the -0.2 nm umbrella (Figure 3.8B). This effect is not drastic and there seems to be enough histogram overlap to discard the need to add an extra umbrella at the -0.1 nm position. In the remaining regions, the simulations also provide good sampling of the reaction coordinates along the substrate transport, which should allow reliable PMF calculations.

In the US simulations, there are many equilibrium properties that can be computed and plotted along the substrate position vector. For this, we need to assure that the simulations are converged and that we are sampling equilibrium. Looking at all the analysis done for these simulations, we can safely consider that all US windows are converged after 20 ns. Therefore, only frames from that point onward were used in the analysis of equilibrium properties such as PMFs, distances and protonation states.

3.2 AAC pH Titration

From the equilibrated segments of our CpHMD simulations (Section 3.1.2), we extracted pH titration data for a list of residues in the AAC structure. Our simulations at pH values 4, 5, 6, and 7, bring us valuable information about amino acids that are more prone to protonate at these pH values. Such data will be very useful to select which residues will be allowed to (de)protonate in the US-CpHMD simulations. This choice is crucial to reduce the computational time spent in the US-CpHMD calculations, since the number of residues that require PB/MC calculations directly impacts the time needed to complete the CpHMD simulations.

From all pH-titrating aminoacids in AAC, only some were allowed to change their protonation state, since we simulated a narrow pH range (4–7). We included the two acidic residues, aspartic (Asp) and

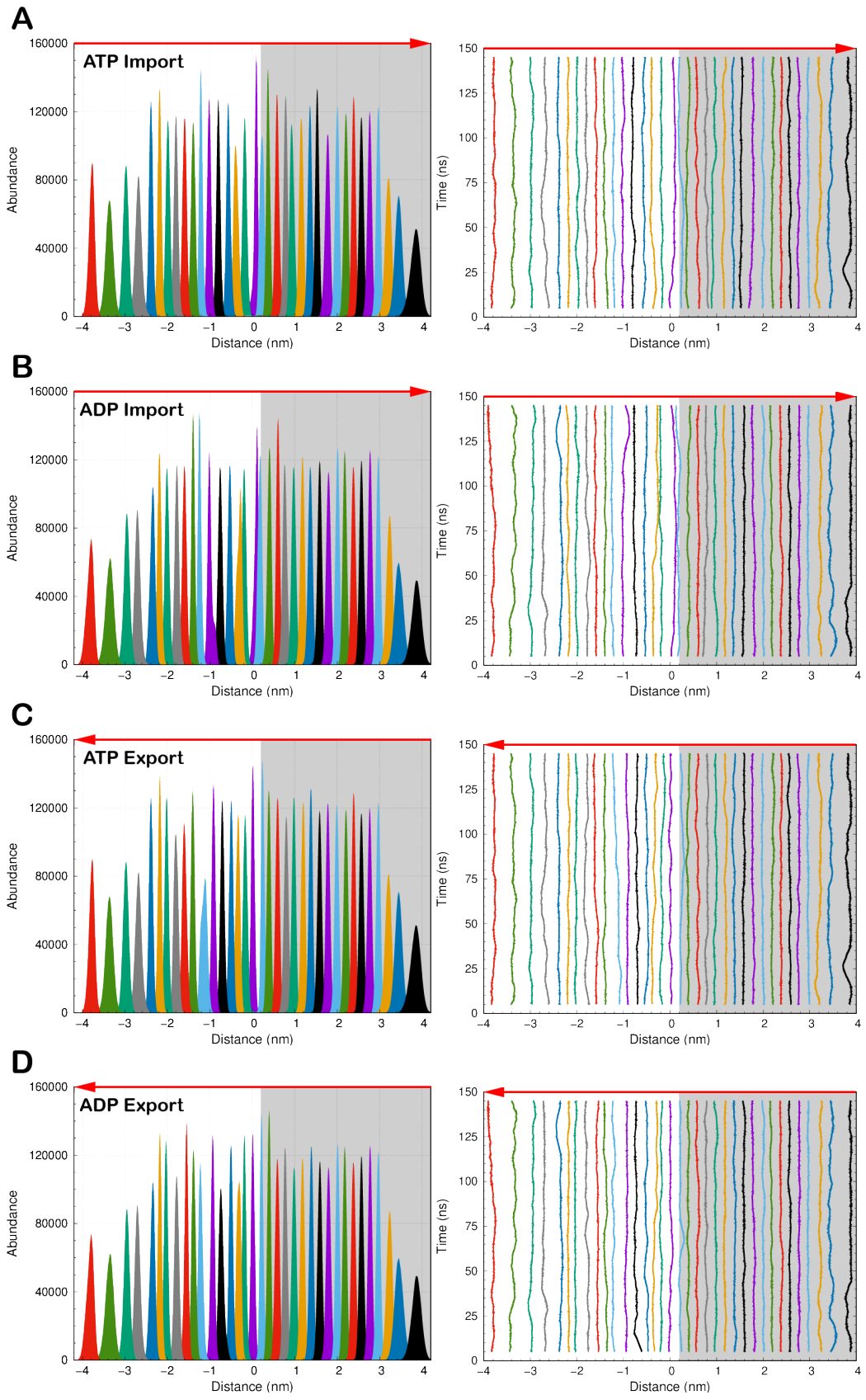

Figure 3.8: US distance distributions (Left) or time evolutions (Right) for the import process of ATP (A) and ADP (B), and also the export of ATP (C) and ADP (D). In the time series, a floating window of 10 ns was used to reduce local fluctuations.

glutamic (Glu) acids, and two basic residues, histidine (His) and lysine (Lys). Both acids and histidines are titrating in the studied pH range [116], while lysines, even though possessing a much larger pK_a values (10.4 in water [116]), may have their neutral forms stabilized due to desolvation, with concomitant decrease in the pK_a. Furthermore, in a previous study [52] two lysine residues have been proposed to undergo significant pK_a shifts, bringing their values closer to the physiological pH. This hypothesis prompted us to also include these residues in our pH titration study with the apo-AAC protein. Arginines (Arg) were kept protonated/charged and not allowed to change during the simulations, despite their important role in the interactions between AAC and the substrate. The main reason is that these residues have very high pK_a values (>12.0) and only in rare occasions can be stabilized in the neutral form.

From the individual titration curves of all residues included in our simulations, we calculated the pK_a values using a fit procedure of the data (including replicates) to a Hill equation (Table 3.1). The data was separated by amino acid type and are only shown for residues that titrated, at least in one state, within our pH range. The two lysine residues titrating in our pH range (Lys22 and 32) were in fact the ones identified previously to be pH sensitive [52]. With the pK_a values of the residues, valuable information can be extracted regarding the environment surrounding them and about interactions they might form with other residues.

Table 3.1: pK_a values for several residues present on the AAC in both the C- and M-state. The errors reported were obtained from the standard error given by gnuplot [113] when fitting our data to a Hill curve. Values for the typical water exposed pK_a values of each residue (Water) were obtained from Reference [116].

Residue		**Water**	**C-state**	**M-state**
Asp	2	**3.8**	3.4 ± 0.5	3.4 ± 1.3
	10		6.3 ± 0.1	>8
	55		4.3 ± 0.1	4.3 ± 0.1
	92		4.0 ± 0.1	5.4 ± 0.1
	103		3.1 ± 0.5	≈ 3
	134		<3	4.0 ± 0.3
	143		3.7 ± 3.5	3.9 ± 0.1
	167		4.5 ± 0.1	4.6 ± 0.1
	195		4.8 ± 0.3	4.4 ± 0.2
	203		3.9 ± 0.3	3.9 ± 0.1
	231		<3	3.1 ± 0.7
	247		3.8 ± 0.3	3.1 ± 0.6
	255		3.5 ± 0.2	4.2 ± 0.1
	263		3.9 ± 0.1	3.6 ± 0.0
	291		5.4 ± 0.3	4.1 ± 0.2
Glu	29	**4.3**	3.0 ± 0.7	3.2 ± 0.1
	47		4.0 ± 0.2	4.2 ± 0.2
	63		3.5 ± 0.1	5.7 ± 0.2
	152		4.1 ± 0.1	4.2 ± 0.1
	264		4.3 ± 0.1	4.3 ± 0.1
	292		4.8 ± 0.0	5.5 ± 0.3
His	39	**6.6**	5.3 ± 0.1	4.8 ± 0.4
	105		6.0 ± 0.1	5.8 ± 0.1
	208		5.3 ± 0.2	5.5 ± 0.1
Lys	22	**10.4**	7.9 ± 1.9	8.0 ± 0.7
	32		7.0 ± 0.3	>8

3.2.1 pK_a Shifts in Key Amino Acid Residues

As stated before, the electrostatic environment and the degree of solvent exposure are crucial in defining the pK_a values of protein residues. From well-solvated peptides, we have the pK_a values for all residues surrounded by water without any significant interaction with neighboring residues [116]. However, when a residue is concealed inside a protein internal domain, like a binding cavity, pK_a shifts are bound to happen due to desolvation and strong electrostatic interactions. These shifts are a consequence of the different stabilities between the neutral and the charged species in that specific conformational ensemble. An example of this effect is the case of several acidic groups in the pHLIP peptide which need to protonate (become neutral), even if only transiently, to allow this peptide to insert into a lipid bilayer [117]. There are also cases where the charged state gains prevalence over the neutral state even in desolvated environments. In this case, there must be an opposite charge in the vicinity that is capable of stabilizing the charged state of that residue through electrostatic interactions. An interesting example of this is the pY1021/PLC-γ1 complex where strong interactions with cationic residues allow a phopsho-tyrosine to remain fully charged, upon binding [110].

From the pK_a list (Table 3.1), the histidine residues do not seem to be involved in any major interactions, since all wielded similar results. In some cases, we observe small pK_a shifts to acidic values, suggesting a small desolvation effect, which is stabilizing their neutral forms (the two deprotonated tautomeric forms). Looking at the location of these residues, we can find two histidines in the top region of AAC, near the intermembrane space, and one histidine in the bottom region of AAC, near the mitochondrial matrix (Figure 3.9). These locations confirm that the histidine side chains are mostly

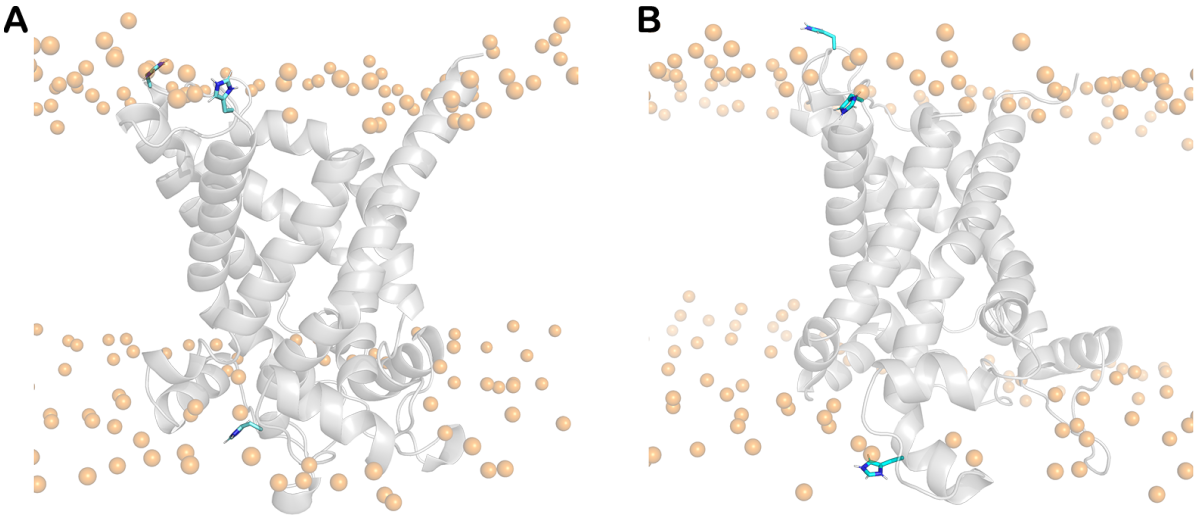

Figure 3.9: Structural representation highlighting the location of the histidine residues in AAC C- (A) and M-states (B). The histidine residues are shown with sticks in cyan carbon atoms. His39 is located in the bottom of AAC while His105 and 208 are on the top. AAC is represented as a cartoon in gray color and the phosphorous atoms of the lipids are represented as orange spheres to allow the identification of the bilayer region.

solvent-exposed and that only small interactions with their neighboring residues are leading to the observed pK_a shifts. It is also possible that in some configurations, the proximity to the POPC membrane can be playing a role. Surprisingly, His39 showed a lower pK_a in the M-state. Due to its location closer to the bottom of the AAC, which is tightly closed in the C-state and opened in the M-state, we would expect a more solvent exposed residue in the M-state. Nevertheless, His39 sits at the end of a transmembrane helix, which is brought closer to the membrane apolar environment with the conformational transition to the M-state, which explains the lower pK_a value in this state.

Overall, the acidic aminoacid residues are mostly located at the water-accessible regions of our protein (Figure 3.10). In these locations, we do not expect large pKa shifts due to desolvation effects, but

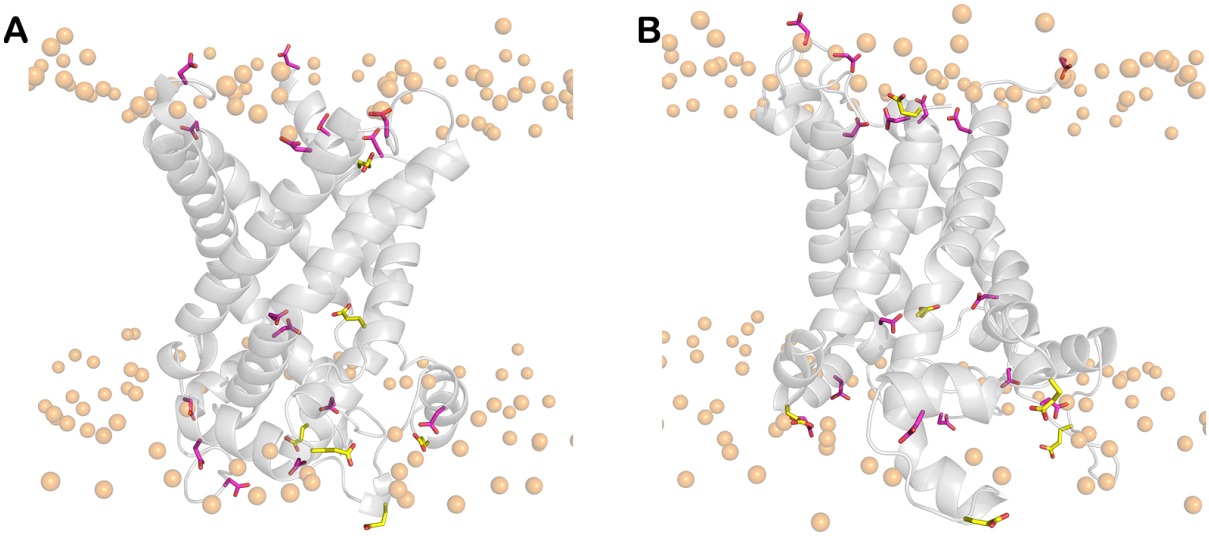

Figure 3.10: Structural representation highlighting the location of the acidic residues in AAC C- (A) and M-states (B). The aspartate (magenta) and glutamate (yellow) residues are shown in sticks with colored carbon atoms. AAC is represented as a cartoon in gray color and the phosphorous atoms of the lipids are represented as orange spheres to allow the identification of the bilayer region.

some heterogeneity in behavior is expected, since these regions are prone to change significantly between the C- and the M-state. This will inevitably lead to differences in pK_a values between states for some of these groups. Indeed, several acidic aminoacids show deviations to higher pK_a values caused mainly by desolvation. In some cases, the effect is present in one state, while in other cases, it is insensitive to the conformational state change. The small pK_a shifts can be explained by their side chains being usually facing the solvent, even if they are located near the protein channel cavity. Nevertheless amongst these residues, there are some that possess peculiar shifts. In this section, we will skip the discussion of residues Glu29, Asp134 and Asp231, which are involved in the electrostatic network at the core of the AAC channel and will be discussed in the following section.

We will pick a few residues to illustrate the capabilities of the CpHMD methodology and the choice is based mostly on the pK_a shifts between states. Residue Glu63, in the C-state, has its pK_a value shifted from typical 4.3 [116] to a more acidic value (3.5). This indicates that a stabilization of its charged state by a neighboring positive charge needs to occur. Interestingly, in the M-state, this pK_a value is increased to 5.7, which suggest a contrary effect. Examining the position of Glu63 in the structure, we noticed that it integrates the h1-2 matrix helix. When in the C-state, this helix is exposed to solvent and also in the vicinity of positive residues, which stabilize a packed conformation of the h1-2 helix close to the transmembrane helices. In the M-state, with the channel opening and the h1-2 helix adsorbing to the membrane, the environment surrounding Glu63 becomes less positive and more apolar, which leads to an increased stability of the neutral state and an increase in the residue pK_a value.

Another interesting residue is Asp10 with pK_a values markedly different from the remaining acidic groups (Table 3.1). We can actually estimate a pK_a value of ≈ 8.9, in the M-state, by extrapolating the Hill curve fit (Figure 5.9). Since this residue is at the beginning of the AAC sequence, it is located near the N-terminal region, sitting close to the top of our carrier (cytoplasmic side). In the C-state, this residue is sitting on top of the opened cavity and in the M-state it is on the closed part of our protein. The larger shift in the M-state is probably easier to interpret, since in this state there is an almost complete

insertion into the membrane environment. Indeed, this effect is not only due to desolvation of the apolar environment, but also due to strong interactions with the lipids phosphate groups (Figure 3.11A).

In the C-state, the cavity is open on the Asp10 side and the N-terminal helix gains extra flexibility which seems to destabilize the lock conformations with the lipid phosphate groups (Figure 3.11B). Consequently, Asp10 becomes surrounded by a more heterogeneous environment among lipids, which causes a milder shift in its calculated pK_a value.

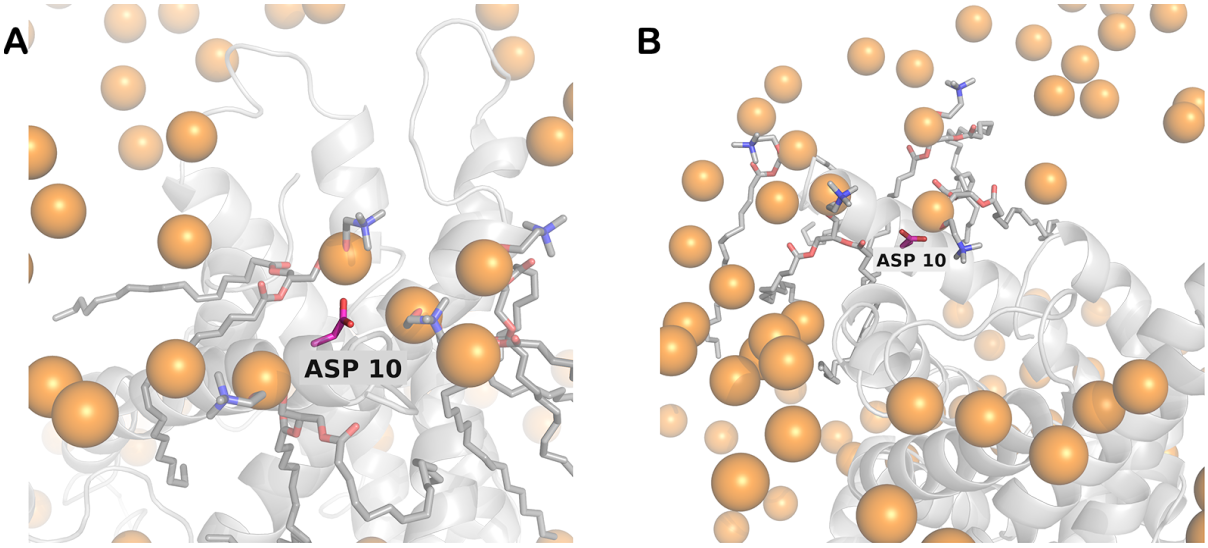

Figure 3.11: Structural representation highlighting the location of Asp10 in AAC M- (A) and C-states (B). Asp10 residue is shown in sticks with carbon atoms colored in magenta. AAC is shown as a cartoon and colored gray. The phosphorous atoms of the POPC lipids are shown as orange spheres. The lipids in the vicinity of the residue are shown with sticks and with carbons colored in gray.

Finally, most lysine residues remained charged throughout the CpHMD simulations, even at pH 7. However, both Lys22 and 32 showed at least a partial deprotonation at pH 7, which allowed us to estimate their pK_a values (Table 3.1). Previous studies have proposed that these residues would have decreased pK_a values, close to 7.0 [52]. Indeed, our CpHMD results are in very good agreement with these authors data. We captured these shifts in both residues even though the exact pK_a values are difficult to calculate due to the lack of sampling and the unavailability of data at higher pH values than 7.0 (Figure 5.10). The shifts observed for both lysine residues can be explained by desolvation and/or strong interactions with positive residues. These residues are located at the bottom of the C-state cavity, which indicates that desolvation effects may be playing an important role (Figure 3.12). Another factor to consider is the amino acid composition of the bottom of the AAC cavity, being composed by positive residues, which outnumber the acidic ones, and are responsible for the binding of the phosphate groups of the substrates. The result of these two additive effects is a decrease in the pK_a values of these lysine residues. Previous studies have revealed that the activity of AAC is pH-dependent [52]. Therefore, the identification of the residues responsible for this dependence may help us interpret the activity variation with pH and better understand the whole process. The protein exhibits a loss in activity at pH values lower than ∼6.0 [52]. The main culprit in this loss of activity is most likely the substrate, whom needs to be fully deprotonated to trigger the initial binding steps and for the transport to occur. Hence, at pH lower than 6 (near the pK_a values of ADP and ATP; 6.3 and 6.5, respectively [112]), the solutes tend to become less negative, which softens the electrostatic-driven recognition by the carrier. This protein also loses the activity at pH values higher than ∼7.5 [52]. In this case, the residues responsible for the pH dependence should come from AAC. It has been proposed that the lysine residues located at the channel cavity are responsible for

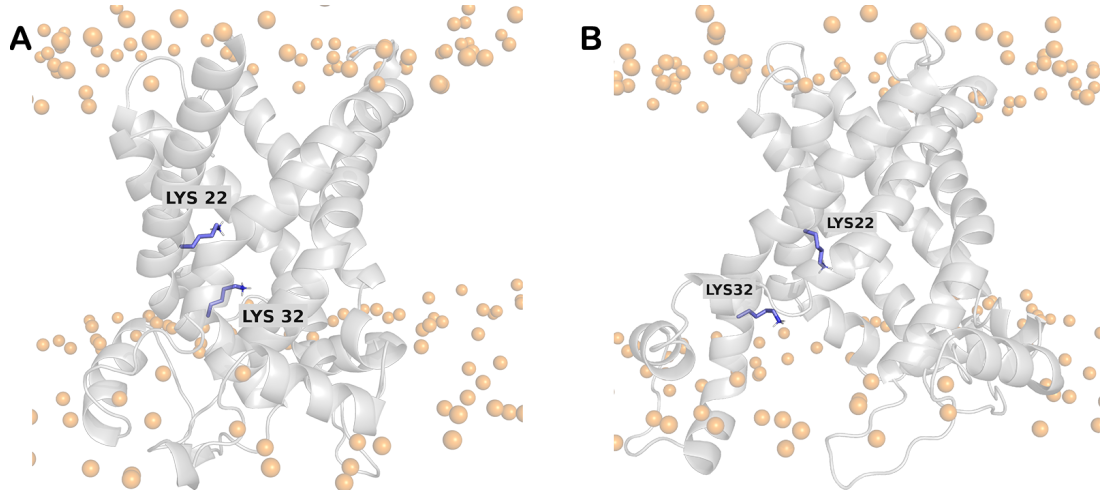

Figure 3.12: Structural representation highlighting the location of Lys22 and 32 in AAC M- (A) and C-states (B). Lysine residues are shown in sticks with carbon atoms colored slate blue. AAC is shown as a cartoon and colored gray. The phosphorous atoms of the POPC lipids are shown as orange spheres.

this effect [52]. In particular, Lys22, which has a similar shifted pK_a value in both the C- and M-states, is probably the one responsible for effect. In the work done by Bidon *et al.*, the activity of AAC was measured with mutations in Lys22 confirming that its presence the protein is essential for activity [52]. Integrating the pH dependence results with the activity results, we can conclude that the protonated form of lysine 22 is essential for the transport activity of AAC to happen.

3.2.2 The Salt-Bridge Networks in AAC Channel

Looking at the results obtained in our simulations (Table 3.1), we have deliberately skipped the analysis of 3 key residues that wielded marked shifts in their pK_a values. These are Glu29, Asp134, and Asp231, which in some cases, and contrasting with the majority of the remaining acidic residues, they barely have any protonation in our simulations. This is particularly true in the C-state, where our fitting procedure struggled to produce accurate pK_a values. In some cases, we opted to just provide an indication that these pK_a values were lower than 3 (Figure 5.11).

Looking at the sequence of the *Bos Taurus* AAC, we can locate these three residues, 29, 134 and 231, in the P-X-D/E-X-X-R/K MCF sequence motif (Figure 5.12) which has been proposed to take a role in forming the matrix salt-bridge network responsible for the closure of the cavity when the protein is in its C-state. Considering the involvement of these residues in the formation of a salt bridge, the pK_a shifts observed are indeed expected, since the establishment of an ionic bond with a positive residue would stabilize the charged state (deprotonated form of the carboxylic acid) overcoming the desolvation effect of the bottom of the cavity. By capturing this type of interactions in our simulations, we can search for the basic residues that are in close proximity of these 3 acids, to identify their partners constituting the matrix salt-bridge network.

By calculating the distance separating the different acidic and basic aminoacids located in the AAC channel, we were able to identify close salt-bridge interactions between: Glu29 and Arg137 (Figure 3.13A); Asp134 and Arg234 (Figure 3.13B); Asp231 and Lys32 (Figure 3.13C). These amino acid pairs have been previously identified in the literature to be inserted in the MCF sequence motif (Figure 3.14 and 5.12) [41, 42]. As it can be seen in the sequence of *Bos Taurus* AAC, there is a possibility of other interactions between residues within the same MCF sequence motif, in this case, Glu29 could interact with Lys32, Asp134 could interact with Arg137, and Asp231 could interact either with Arg234,

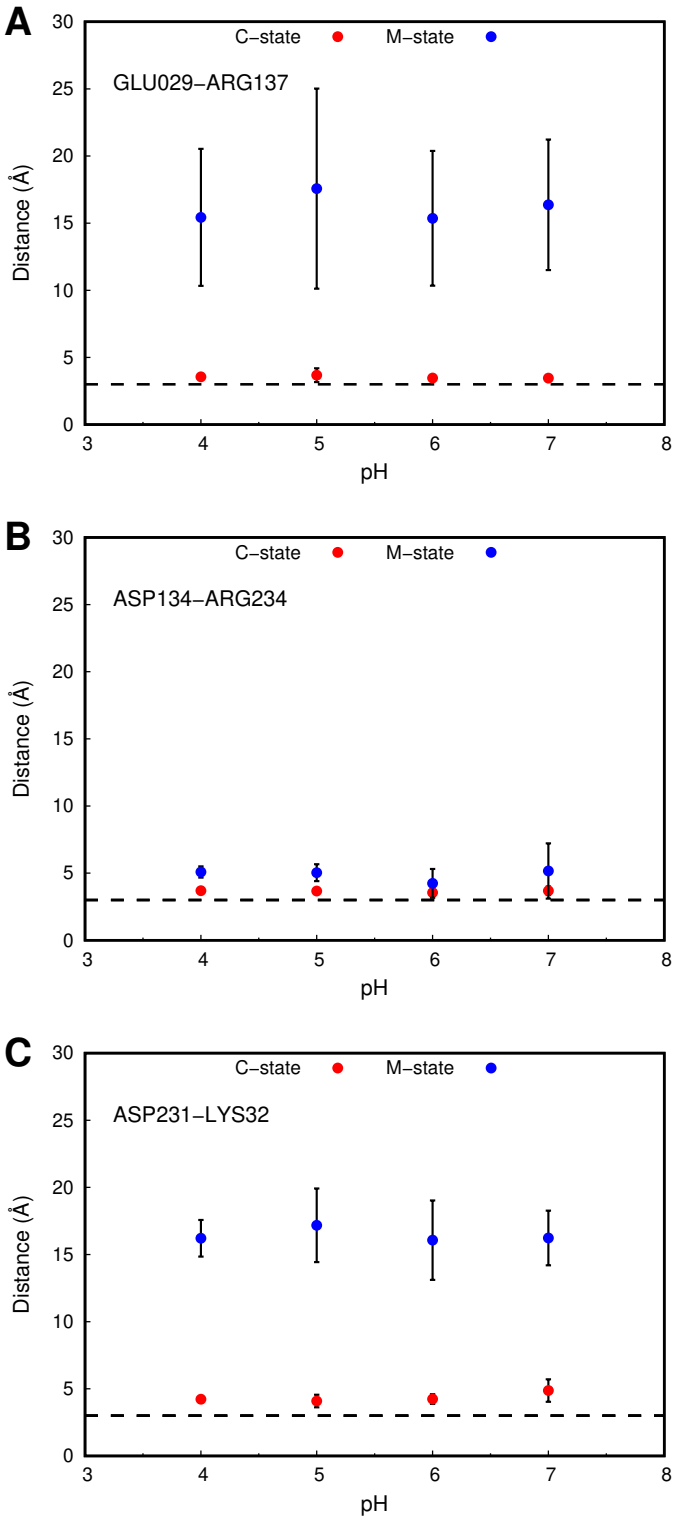

Figure 3.13: Average distance between the residues forming the matrix salt-bridge network, at all pH values and both AAC conformational states. The Glu29–Arg137 (A), Asp134–Arg234 (B), and the Asp231–Lys32 (C) pairs are shown. The data for the C- and M-state of AAC are shown in red and blue, respectively. A dashed line was added marking the distance of 3 Å to be used as a reference distance where ionic bonds are more likely to be formed.

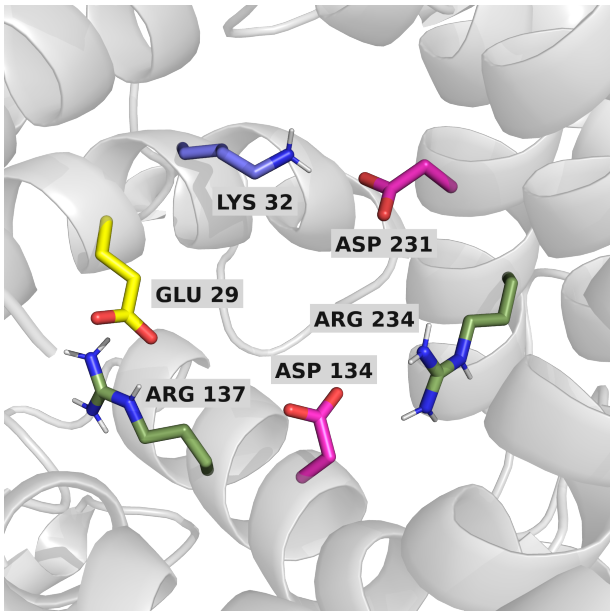

Figure 3.14: Structural representation of the matrix salt-bridge network formed on the bottom of the AAC C-state cavity. AAC is represented as gray colored cartoon, Asp, Glu, Lys, and Arg residues have their carbon atoms colored in magenta, yellow, forest green, and slate blue, respectively.

Arg235, or Arg236. These interactions are very scarce, since they could only happen when their normal ionic bond is broken. From our distance analysis we are also able to access the differences in these salt-bridge pairs between both AAC states. Interestingly, the Glu29–Arg137 and Asp231–Lys32 distances change significantly between the C- and M-state. In the C-state, these acids are in close proximity with their respective basic pair, creating a very stable salt bridge, which gets disrupted when the structure changes to the M-state. On the contrary, the Asp134–Arg234 salt bridge is not fully disrupted in the M-state, showing only a small increase in the average distances, probably because these two regions of the protein do not move apart significantly when the channel opens.

In the previous section, we have discussed Lys32, whose shifted pK_a value in the AAC C-state may play an important role in the pH-sensitivity of AAC. Note that this residue is paired with Asp231 in the matrix salt-bridge network (Figure 3.14). A basic residue with a shift to lower pK_a values (Table 3.1), indicates that a destabilization of the charged state is occurring. This suggests that this specific ionic interaction may not be very strong, allowing the Lys32 residue to move away from its partner. The average distance at pH 7 for this pair is $\sim$5Å (Figure 3.13C), which already suggests that the interaction is not very strong. We calculated the distance distributions for this salt-bridge and it is clear that Lys32 is only locked to Asp231 in a salt bridge in a small part of the simulations, specially at higher pH values (Figure 3.15).

After looking in detail to the matrix salt-bridge network of AAC C-state, we can also study the salt-bridge network responsible for closing the cytoplasmic access to the cavity when AAC is in the M-state, the so-called cytoplasmic salt-bridge network. Once again, if a similar network is formed when the protein is in the M-state, it should result in pK_a shifts on some of the acidic residues involved in those salt-bridges. However, none of the residues present significant decreases in pK_a values when transitioning to the M-state.

To pinpoint the residues that might be involved in a possible formation of a salt-bridge network when the protein is in the M-state, we relied on the available literature. Several authors used bioinformatic analysis to identify residues of a conserved domain in the even-numbered helices, Y/F-D/E-x-x-K/R [42,

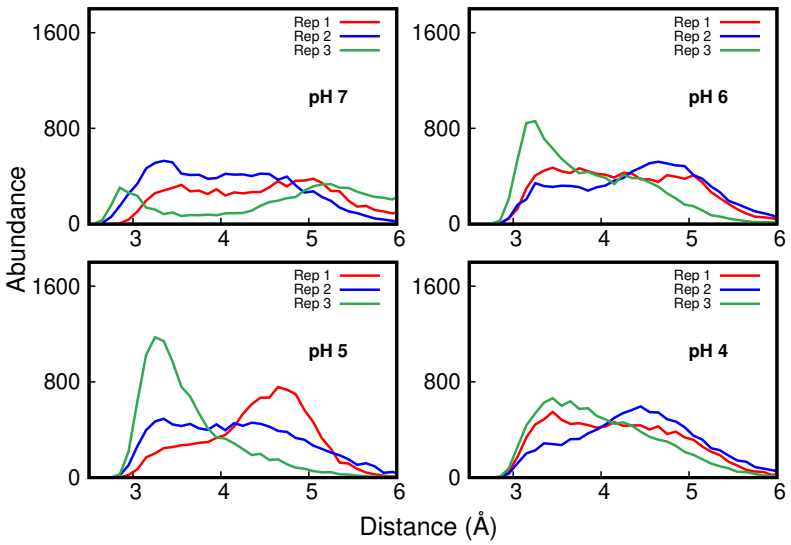

Figure 3.15: Distribution of distances between the Asp231 and Lys32 in AAC C-state. Each pH value examined is presented in an individual plot, which is identified by label.

48]. In the *Bos Taurus* sequence, the residue pairs from this conserved sequence domain probable to be in ionic bond would be Asp92 interacting with Lys294 or 295, Asp195 with Lys95, and Asp291 with Lys198 (Figure 5.13). In our AAC structure, we did not include residues beyond 292, since they were considered unstructured and were not resolved in the *Bos Taurus* structure. This will limit our analysis of the first putative salt bridge, but we can perform it for the remaining cases.

After identifying the residues most likely to establish ionic interactions, we can look at their pK_a values. In the titration curves of Asp92, 195 and 291 (Figure 5.14), we observe several different behaviours. For Asp92 an unexpectedly higher pK_a value was observed for the M-state, but this only reflects the packed environment in which this residue is inserted and the absence (not included in the simulation) of the basic residue partners which could have stabilized its charged state. Therefore, the pH titration curve and the pK_a value for this residue in the M-state are probably not reliable.

For Asp195 and 291, in the M-state, we observe small shifts in the pK_a values towards lower values, which probably cannot be assigned to the formation of a matrix salt-bridge network, but indicates that these acidic residues are interacting with slightly more positively charged regions. The most promising partners for Asp195 were both Lys91 and 95. In fact, Lys91 was often the closest partner, but since both Lys residues are just 4 residues apart, we should not discard Lys95 as a partner (Figure 3.16). In the average distance data, we confirmed that both Lys residues are stabilized by the Asp195. The interactions are significant in the M-state, but the matrix salt-bridge network does not seem to be formed strongly, specially at physiological pH values.

For Asp291, the pK_a value in the C-state is already high (5.4), which is most likely related with the position of this residue in the C-terminal of AAC, which brings it often to positions close to the membrane residues where it loses its charge due to desolvation. Looking at this residue in the M-state, we observe a 1.3 pK units shift to lower values (Figure 5.14C). This shift can be triggered by a change in the environment, with more solvation than in the C-state, or due to interactions with an amine partner. Similarly, we analysed the distance between Asp291 and all neighboring Lys and Arg residues of AAC and identified Lys198 as the closest partner that could establish an ionic interaction (Figure 3.17).

In summary, we have obtained significant results from the CpHMD simulations in the AAC apo form

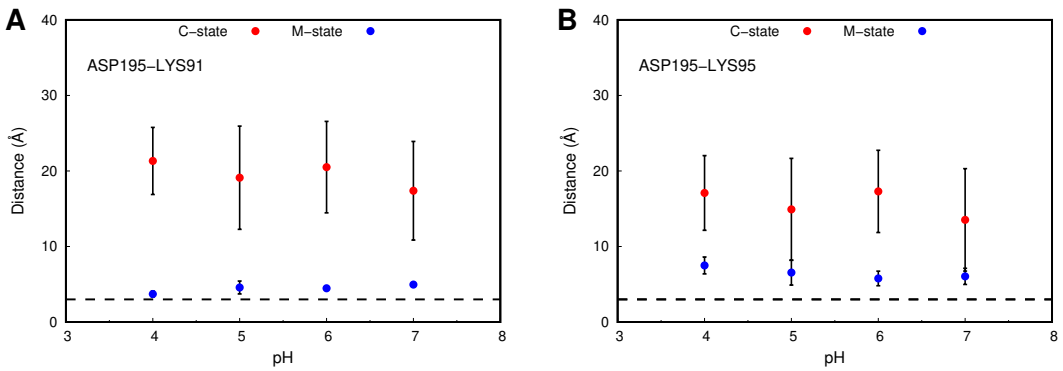

Figure 3.16: Average distance, at each pH value, between Asp195 and the closest basic residues, Lys91 (A) and Lys95 (B). Lys91 was the residue found to be closest to Asp195 to form an ionic bond. The data for the C- and M-state of AAC are shown in red and blue, respectively. The error bars were calculated using the standard error of the mean between the 3 replicates at each pH value. A dashed line was added marking the distance of 3 Å to be used as a reference distance where ionic bonds are more likely to be formed.

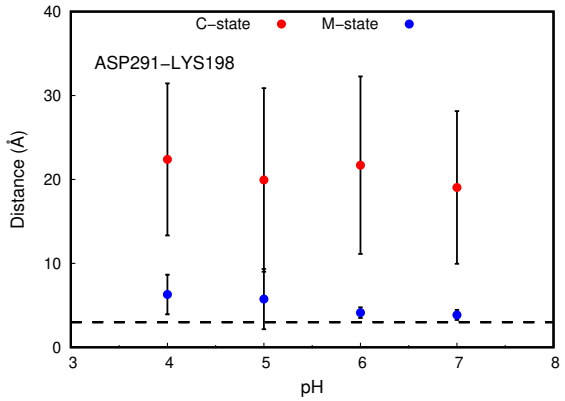

Figure 3.17: Average distance, at each pH value, between Asp291 and the closest basic residue, Lys198. The data for the C- and M-state of AAC are shown in red and blue, respectively. The error bars were calculated using the standard error of the mean between the 3 replicates at each pH value. A dashed line was added marking the distance of 3 Å to be used as a reference distance where ionic bonds are more likely to be formed.

regarding both the structural aspects and their coupled protonation states. The sheer size of transmembrane proteins and their inclusion in membrane patches increases the difficulty of studying in detail these complex systems. However, as presented in both present and previous sections, there is a huge amount of information and molecular detail on the environment of each amino acid residue that can be extracted from looking at conformational ensembles and pK_a values. For AAC, the residues involved in the formation of the salt-bridge networks have been characterized in detail and this information can help us interpret other even more complex results about this protein.

3.3 Umbrella Sampling

In the present thesis, we are interested in studying the transport of a substrate (ATP or ADP) through a transmembrane transporter (AAC). As stated in section 2.6.6, to describe this transport process through the mitochondrial membrane, we have sliced our system along the membrane normal, which is the z axis coordinate of our box. Both the import and the export processes will be analysed in our simulations, independent of their *in vivo* relevance. To capture all processes, we need to perform several US-CpHMD

simulations using the C- and M-states and with the substrates phosphate groups pointing in both directions. This is particularly important in the protein channel, since the solute molecules cannot freely rotate. All US-CpHMD schemes were built on the AAC structure in each state, which means that the conformational transition, which is essential for transport, needs to be added to the final scheme. Indeed, the realistic process requires the use of a given number of umbrellas for the substrate entrance into an open conformation, complemented by another set of umbrellas with the other conformational state, where the substrate is allowed do diffuse to the other side (Figure 2.12B). We also computed a few extra US windows where the conformational transition is absent, to investigate the magnitude of the energy barriers encountered by the substrates when going through a closed channel (Figure 2.12A).

3.3.1 Potential of Mean Force (PMF)

An important result obtained from the US schemes is the energetic profiles of the different processes, the PMFs. In the first set of simulations, without accounting for any conformational transition, the description of the transport from the bottom of the cavity onward, in both C- and M-states of AAC, is not representative of what would naturally happen in the mitochondria. By not performing the conformational transition, the substrate will have to open the closed portion of the AAC, resulting in a higher energy penalty, probably without physical significance *in vivo* (Figure 3.18). However, these plots are valuable to pinpoint after which US windows the conformational transition needs to take place. These should correspond to the first umbrella that significantly increases the PMF energy, hinting that it would require a change in the conformational landscape to allow motion in that direction. In the import process, performed with AAC in the C-state, an increase in the substrate energy is captured around $+0.5$ nm (Figure 3.18A), while in the export process, with AAC in the M-state, this energetic increase is present around 0.0 nm (Figure 3.18B). From these results, we propose that the US umbrella at $+0.2$ nm is the ideal window to perform the conformational transition which is coupled to the transport processes.

Using US windows from different schemes, we can reproduce the complete profiles (Figure 3.19). As expected, the new profile portions computed (red and blue lines) are of lower energies than the ones that assumed a closed channel. Looking at the energies obtained at the starting positions of each transport process, ~ -4.0 nm and $\sim +4.0$ nm for the import and export, respectively, it can be noted that both plots quickly acquire a negative slope. The absence of a plateau, more pronounced in the import process (Figure 3.19), suggests that our defined zero energy was assigned to umbrellas where the substrate molecules are still slightly influenced by AAC or the membrane. Nevertheless, since our analysis will mainly be focused on comparisons between different substrate profiles in each transport process, most of these approximations should cancel out. A possible workaround to investigate this approximation would be to increase the height of our simulation box, allowing sampling of conformations where the substrate would be even further from the AAC cavity. However, these would come at a significant computational cost.

All PMF profiles show that the AAC cavities, on both sides, have an attractive role in guiding the substrate to the open channel. This is illustrated by a marked continuous decrease in energy for both molecules right from the start, until reaching the pocket bottom ($+0.2$ nm). This attraction can most likely be attributed to the electrostatic interactions between the phosphate groups and the cationic groups in the AAC cavity pocket. Interestingly, we note a clear difference in the energy profiles of the import process at ~ -1.0 nm position, where ADP is more easily stabilized than ATP (Figure 3.19A). This difference gets even more pronounced when the substrates go through the exit pathway. The electrostatic interactions, which are smaller in ADP, can help rationalize the differences in this last exit step. However,

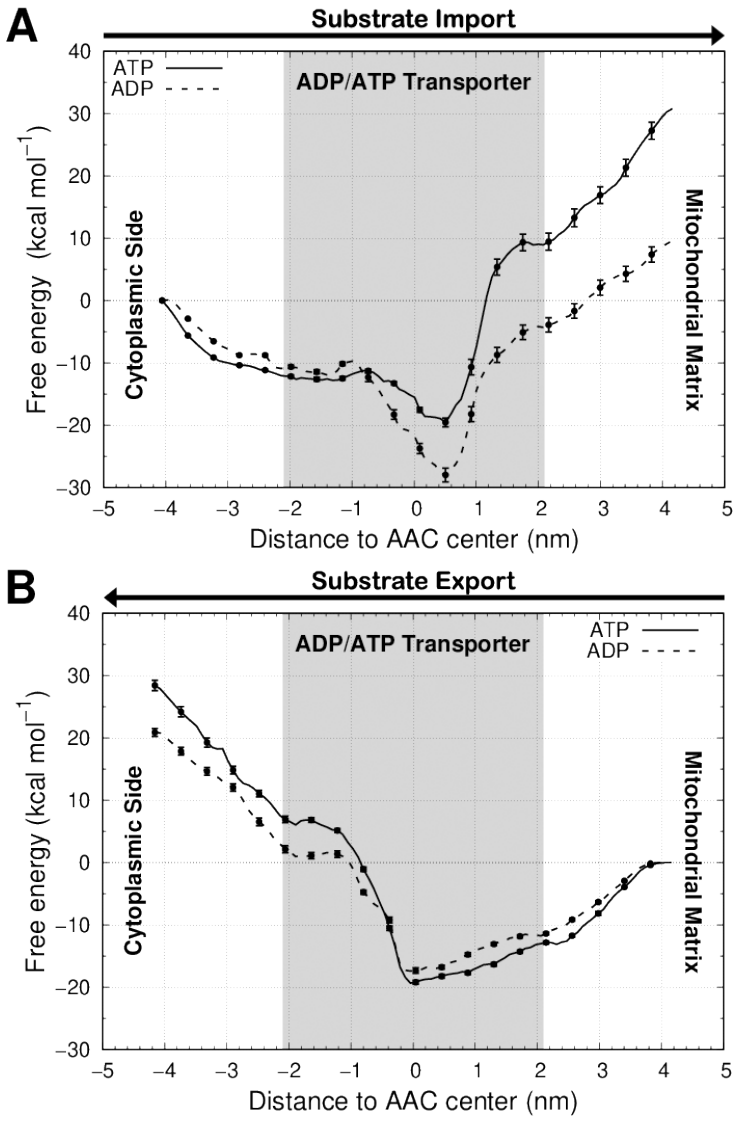

Figure 3.18: PMF energy profiles of both substrates, ATP (full line) and ADP (dashed line), in the import (A) and export (B) transport processes. These processes assume that no conformational transition in the AAC occurred, hence, the C-state was exclusively used in the import (A), while the M-state was used for the export (B). A gray background area was used to represent the positions occupied by the AAC. The PMF error bars are shown only every 5 points for clarity.

in the C-state, the over-stabilization of the ADP defies the electrostatic potential and may be taking advantage of some size exclusion mechanism.

These results points at the existence of selectivity for ADP over ATP in the import transport process. In contrast, in the export process, there are no significant differences between the energy profiles of ADP and ATP (Figure 3.19B). In order to interpret some of these distinct characteristics being captured in the PMF profiles, we need physiological context. In the import process, ADP molecules need to be captured from the cytoplasmic side of the membrane. This full transport process is done against the membrane potential, which is formed due to the proton gradient generated by the electron transport chain (Figure 3.20). In terms of compartments, the cytoplasm is much larger than the mitochondrial matrix, suggesting that substrate concentrations should be significantly lower on the cytoplasmic side. As a consequence, AAC needs to be much more effective in capturing substrate molecules from the cytoplasmic side and, preferably, with some selectivity. Evolution may have pressured AAC to adopt

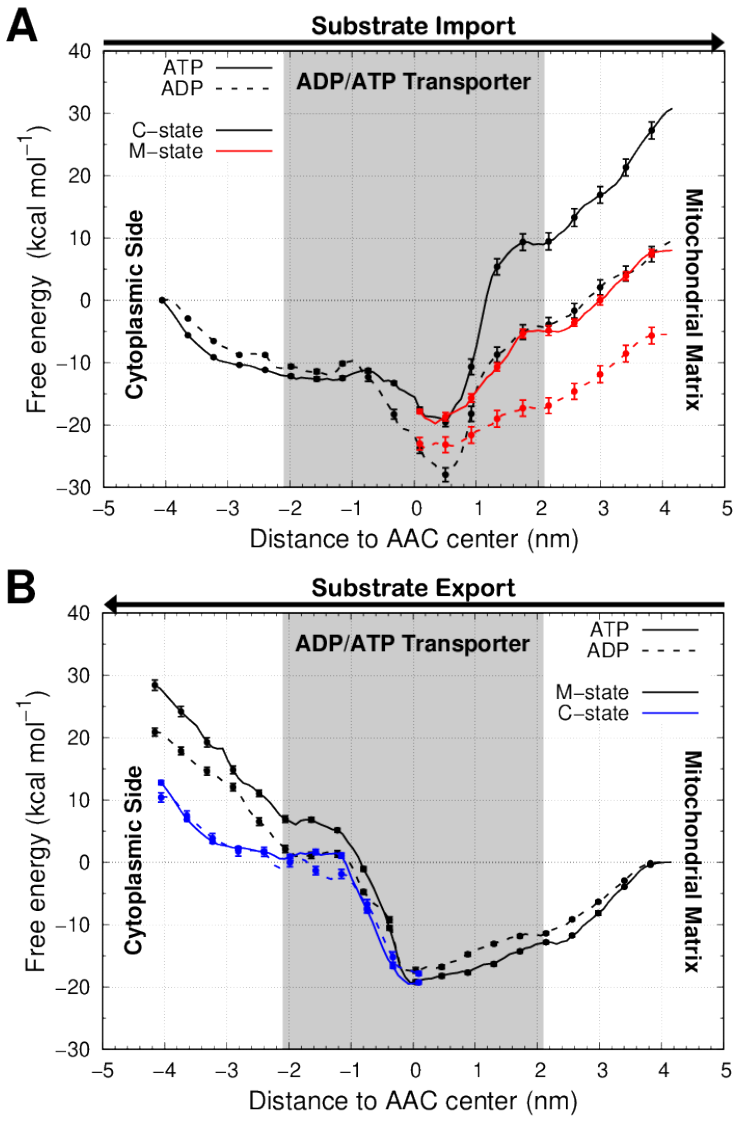

Figure 3.19: PMF energy profiles of both substrates, ATP (full line) and ADP (dashed line), in the import (A) and export (B) transport processes. The processes assuming a conformational transition in the AAC are shown in red and blue lines. A gray background area was used to represent the positions occupied by the AAC. The PMF error bars are shown only every 5 points for clarity.

traits in the C-state that improve the attraction and retention of ADP inside its cavity in order to guarantee its efficient transport. The PMF profiles seem to capture these electrostatic attraction funnel from distance and also estimate a ∼10 kcal/mol difference favoring ADP over ATP.

On the other hand, the export process *in vivo* is responsible for transporting ATP from the mitochondrial matrix back to the cytoplasmic side. In the matrix, there is a higher concentration of ATP molecules compared to ADP, since they are being converted by ATP synthase [1, 118, 119], and the export process is performed in favor of the membrane potential. Therefore, there is only evolutionary pressure for AAC in the M-state to become efficient, with almost no pressure for exerting selectivity. Our PMF data is in excellent agreement with this reasoning.

The lack of selectivity for the substrate in the AAC export process and only a partial selectivity in the import counterpart, indicates that we cannot discard the possibility of reverse transport in this carrier. In fact, it has been shown that upon reversing the membrane potential *in vitro*, the favoured

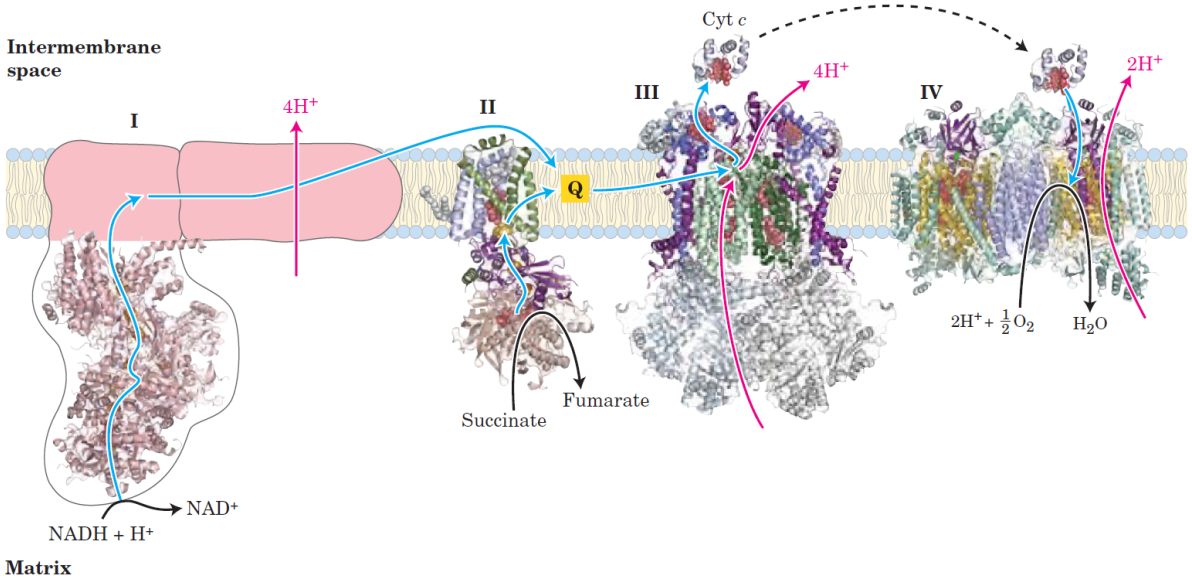

Figure 3.20: Summary of both electron and proton flows through complexes I-IV of the electron transport chain. The flows of electrons and protons are marked by blue and pink arrows, respectively. Image adapted from [1].

transported molecules in each process would be inverted [36, 40, 50, 120]. Therefore, both ATP and ADP can be imported to the matrix or exported from the matrix, depending on the conditions presented in the environment. A similar mechanism is found in the ATP synthase protein when the membrane potential is abolished, it inverts the process and becomes an ATPase, exporting protons from the matrix by hydrolyzing ATP molecules [118, 119].

3.3.2 ATP/ADP Protonation Profiles

Unlike typical US simulations, which are usually performed using conventional MD, we used CpHMD simulations in each umbrella window. With this methodology, all key residues in AAC and substrates are allowed to change their protonation states during simulations, according to the pH chosen (pH 7.0). We have allowed (de)protonation events to occur in the substrates and in histidines and a reduced group of acidic amino acids of AAC. Even though, we observed a decrease in the pK_a values of Lys22 and 32 of apo-AAC, we expect that the presence of the substrate will introduce a strong shift towards higher values, which make them highly positive. By reducing the number of residues that require PB/MC calculations to define protonation states during the CpHMD, we are able to reduce the computational cost of performing the large number of US simulations.

We opted to uncouple the protonation data from their US window to discuss the results extracted from the US scheme, mimicking the substrate transport. We used the value of the reaction coordinate of each conformation to perform a moving average protonation (0.1 nm windows) along the entire reaction coordinate span. Due to the presence of the biasing potential in each US window, taking this approach to obtain the protonation profile during the transport raises the question of using statistical re-weighting. Since values from different windows can be used to obtain an average protonation value, their contributions may not have the same weight. As an example, a protonation value referent to a conformation of reaction coordinate of $+2.0$ nm in the $+2.2$ nm window ensemble should have higher weight than a protonation with the same reaction coordinate but on the $+2.0$ nm window ensemble. When performing biased MD simulations, we should always remove the original bias in order to extract equilibrium properties. Examples of this are the PMF calculations shown above, which use the WHAM method to

perform the re-weighting steps [86]. However, when calculating properties along a reaction coordinate, which are not real free energies, the re-weighting step may not be required. Note that the US simulations can be used just to generate new conformations typical of that specific region/umbrella. Performing an uncorrected average of that data should provide a very good guess of the conditional free energy at that point of the reaction coordinate. In all tests performed in our research group, the differences between the re-weighted and the uncorrected protonation profiles are just marginal and, in the future, we will also investigate it for this system.

Using the equilibrated portion of the US-CpHMD simulations, we extracted protonation profiles for the terminal phosphate group of each substrate during the transport processes. This group was allowed to titrate in our simulations since its pK_a values are close to our pH 7.0 (6.3 and 6.5 for ADP and ATP, respectively) [112]. The first observation obtained from the protonation profiles is that, at the starting positions of each transport (~ -4.0 and $\sim +4.0$ nm for the import and export respectively), the protonation values observed (~ 0.2) are in agreement with a well-solvated substrate at pH 7 (Figure 3.21). Comparing these results with the PMF profiles, in which a small plateau could be perceived in the start of

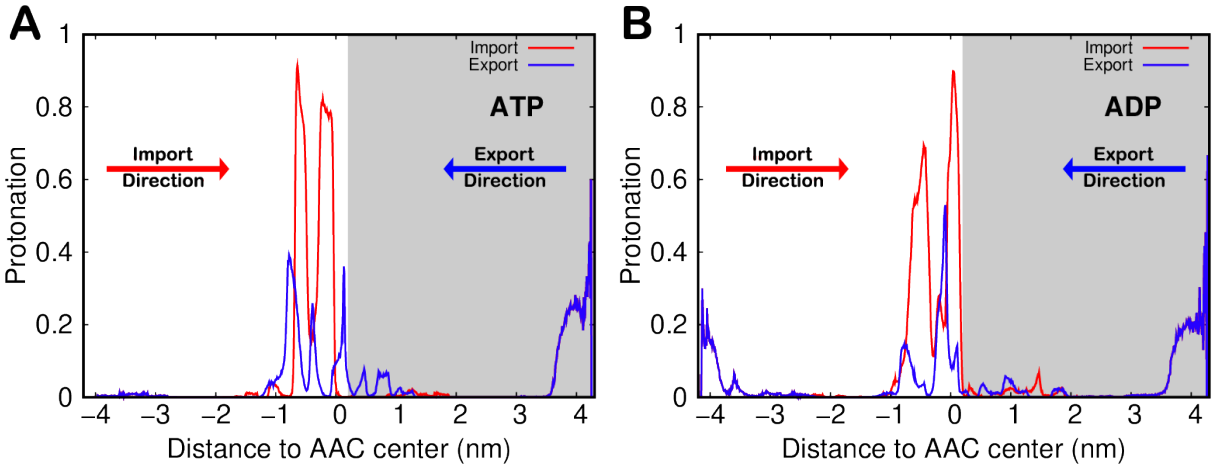

Figure 3.21: Protonation profiles of ATP (A) and ADP (B) during both the import (red lines) and export (blue lines) processes obtained from the US-CpHMD simulations. A white and gray backgrounds were used for umbrellas obtained with AAC C- and M-states, respectively. In the more solvated edge regions, both import (red) and export (blue) lines are superimposed, since the same umbrella windows were used.

the export process (Figure 3.19B), we are now confident that, in these windows, we were able to sample a solvated substrate without significant electrostatic effects from AAC or the membrane. Nevertheless, the substrate profile obtained at the start of the import process showed a fully ionized ATP (Figure 3.21A), indicating that this substrate, at ~ 4.0 nm distance from the channel pocket, is already sensing its positive electrostatic potential. This is in line with the observation, also from the PMF analysis, that the C-state form of AAC is able to exert a stronger attraction on the substrates than the M-state. This reinforces our conclusions regarding the AAC asymmetry and that the C-state cavity has a higher positive electrostatic potential than the M-state. This is also in agreement with the smaller availability of substrate in the *in vivo* import process, requiring a stronger and more efficient transport trigger.

Both our substrates have a propensity to protonate their terminal phosphate group when reaching distances close to the center of the cavity. However, these protonation events seem to take place mostly on the cytoplasmic side and when the protein is in the C-state. This observation is directly tied with a balance between two different counteracting effects: the desolvation and the direct interactions with basic amino acids. When inserting ADP or ATP in the protein channel, there are strong desolvation effects that stabilize the protonated form. However, when in the close vicinity of many basic amino

acids, charged positively, this effect can be counter-balanced and the ionized phosphate is again the most stable species. Since most amino groups are located in the matrix side of the channel, this explains why substrate molecules protonate less in this section of the reaction coordinate. We also observe that there is a higher tendency to protonate the substrate in the import process comparing to the export. As we mentioned, this protonation comes mainly from desolvation effects, which can be modulated by the orientation of the adenine base of the substrates. In the import, this base groups are located between phosphates and solvent in the C-state open cavity, which maximizes the desolvation penalty. On the contrary, in the export process, the terminal phosphate will be significantly more solvent exposed which will minimize desolvation and lead to lower protonation values.

The higher abundance of protonated substrate species, in the import process when compared with the export process, is an interesting result that can be contextualized with the biological process. As stated previously, the import is performed against the membrane potential and there is a prevalence of protons outside the inner mitochondrial membrane. Therefore, importing a highly charged negative substrate, such as ADP and ATP, comes at an energetic cost. To minimize this, AAC could have been evolutionary pressured to help reduce the charge of the substrate. On the other hand, in the export process, which occurs in favor of the membrane potential, the presence of a high negative charge is beneficial for the transport.

3.3.3 AAC Structure and Protonation During Transport

Similar protonation profiles can be calculated for residues of the AAC cavity, which may undergo (de)protonation events during the substrate transport process. However, we should first verify which AAC residues change their protonation state when we force the conformational transition. We have calculated the average protonation of the equilibrated portion of the windows prior (0.0 and +0.4 nm for import and export, respectively) and post (+0.2 nm in both transports) conformation change. With this analysis, we will identify residues which undergo (de)protonation events during the artificially induced conformation transition when the substrate is present in the bottom of the cavity. We have applied a protonation change cutoff of 0.2 to select only the most relevant residues and, unsurprisingly, the three acids involved in the formation of the matrix salt-bridge network are captured in our analysis (Table 3.2). These three acids presented low pK_a values in both forms of apo-AAC, being below 3 on the apo C-

Table 3.2: Residues with protonations affected by the conformation change in the US-CpHMD simulations. Only residues where the protonation prior to conformational change (pre-CC) differ more than 0.2 from the protonation post conformational change (pos-CC) are presented. The presented protonation values were averaged using only the equilibration portion of each US window.

Transport Process	**Substrate**	**Residue**	**Protonation**	
			Pre-CC	**Pos-CC**
Import	ATP	Glu29	1.00 ± 0.00	0.00 ± 0.00
		Asp231	0.99 ± 0.01	0.00 ± 0.00
	ADP	Asp231	0.92 ± 0.09	0.00 ± 0.00
Export	ATP	Glu29	0.20 ± 0.14	0.99 ± 0.01
		Asp231	0.03 ± 0.02	0.99 ± 0.01
	ADP	Asp134	0.00 ± 0.00	0.57 ± 0.17

state and around 3.1/3.2 on the apo M-state (see section 3.2). Their low pK_a values, especially in the C-state, were a consequence of strong interactions with their cationic group partners, which efficiently counteracted all desolvation effects present. However, in the presence of the substrate, the amino groups

are recruited to stabilize the charged phosphates, leaving the acidic residues prone to desolvation effects that result in their protonation (Table 3.2). This effect is more pronounced at the conformational change, since this structure alteration leads to disruption or formation of the matrix salt-bridge network.

Looking at the complete protonation profiles of these three acidic aminoacids, it becomes evident that the presence of the highly charged substrates leads to their protonation (Figure 3.22). As we men-

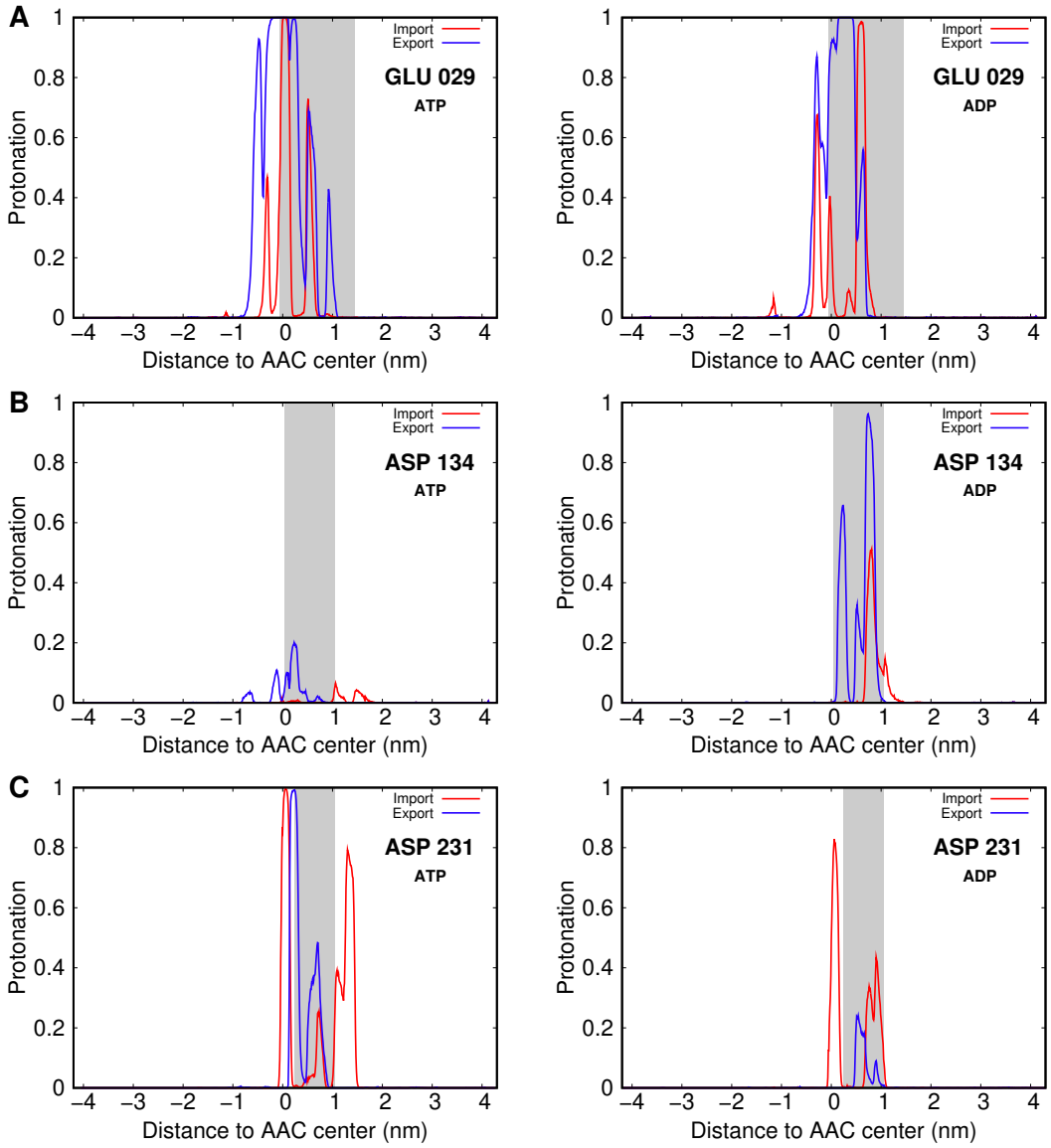

Figure 3.22: Protonation profiles of Glu29 (A), Asp134 (B), and Asp231 (C), during the transport of both ATP (left) and ADP (right) in the import process (blue line) and in the export process (red line). A background gray area highlights the region sampled by each amino acid during the simulations.

tioned before, these protonations are a consequence of the close approximation of the negatively charged phosphates to the acidic amino acids. Even if the phosphate groups do not interact directly with these residues, their highly negative charge will sequester all cationic groups, dampening the positive electrostatic environment. This will leave all acidic residues subject to desolvation effects, which increase their pK_a values to the point that these groups protonate at pH 7.0.

To bring biological context to these protonation events, we must take in consideration their role in the matrix salt-bridge network. As pointed in the discussion of the pK_a shifts of these residues (Section 3.2.2), they form salt bridge interactions with basic residues that exist in abundance on the bottom

of the AAC cavity. The formation of this network has been pointed as a key feature in the stabilization of the transmembrane protein in the C-state. When the import process introduces the substrate in the bottom of the cavity, and by consequence leading to the protonation of these three residues, the aforementioned salt-bridges become weaker, probably triggering its disruption. With the absence of these stabilizing interactions, the energy barrier of the conformational change from C- to M-state is lowered, allowing the presence of the substrate to induce the transition and then proceed to be imported to the mitochondrial matrix.

We investigated further this possibility of weakening in the matrix salt-bridge network induced by the substrate insertion into the C-state cavity. We evaluated the distance between the ionic pairs that compose this network throughout our umbrella sampling simulations aiming at identifying salt-bridge disruption events along the transport process. Plotting the average distance between the salt-bridge pair in each US window, we are unable to capture all modifications to the salt-bridge network pairs during the import process of both substrates. As expected, at the $+0.2$ nm window and with the conformational transition, we observed the disruption of some ionic pairs and their average distances converging to the values calculated in the apo M-state analysis (Figure 3.23). Indeed, only the Asp134–Arg234 ionic pair seems to not break completely, which was already observed in the apo-AAC calculations. Our data shows that the presence of a substrate at the bottom of the cavity was not sufficient to induce the separation of any of the two ionic pairs expected to be weakened, the Glu-29–Arg137 and the Asp231–Lys32. Although, we are able to capture the weakening of this matrix salt-bridge network, the large conformational transition that it will trigger probably takes longer than the 150 ns that we were able to simulate. These major rearrangements of membrane proteins which include movements of several transmembrane helices, are usually happening in time scales inaccessible to our computational protocols.

In light of the interesting results obtained for the import process, we also analysed the protonation behavior of the cytoplasmic salt-bridge, aiming to identify similarities for the export process. However, the residues that form this salt-bridge are located near the cytoplasmic end of AAC. Their long distances to the bottom of the AAC cavity makes them insensible to the inserted substrates and only reacting to their presence during the exiting portion of the export process. Looking at the protonation profiles for residues Asp92 and Asp291, we observe that only Asp92 is protonating significantly (Figure 3.24. This difference in behaviour is probably due to direct interactions with the substrate and would require a better sampling with more replicates to evaluate its importance. The fact that none of these residues significantly change protonation with the conformational transition in the export process, suggests that the presence of the substrate does not have a weakening effect on the cytoplasmic salt-bridge network. In line to what we have observed in this work, it seems that most evolution pressure exerted on AAC was focused on the import process, while not many tweaks are present to optimize the substrate export from the mitochondria.

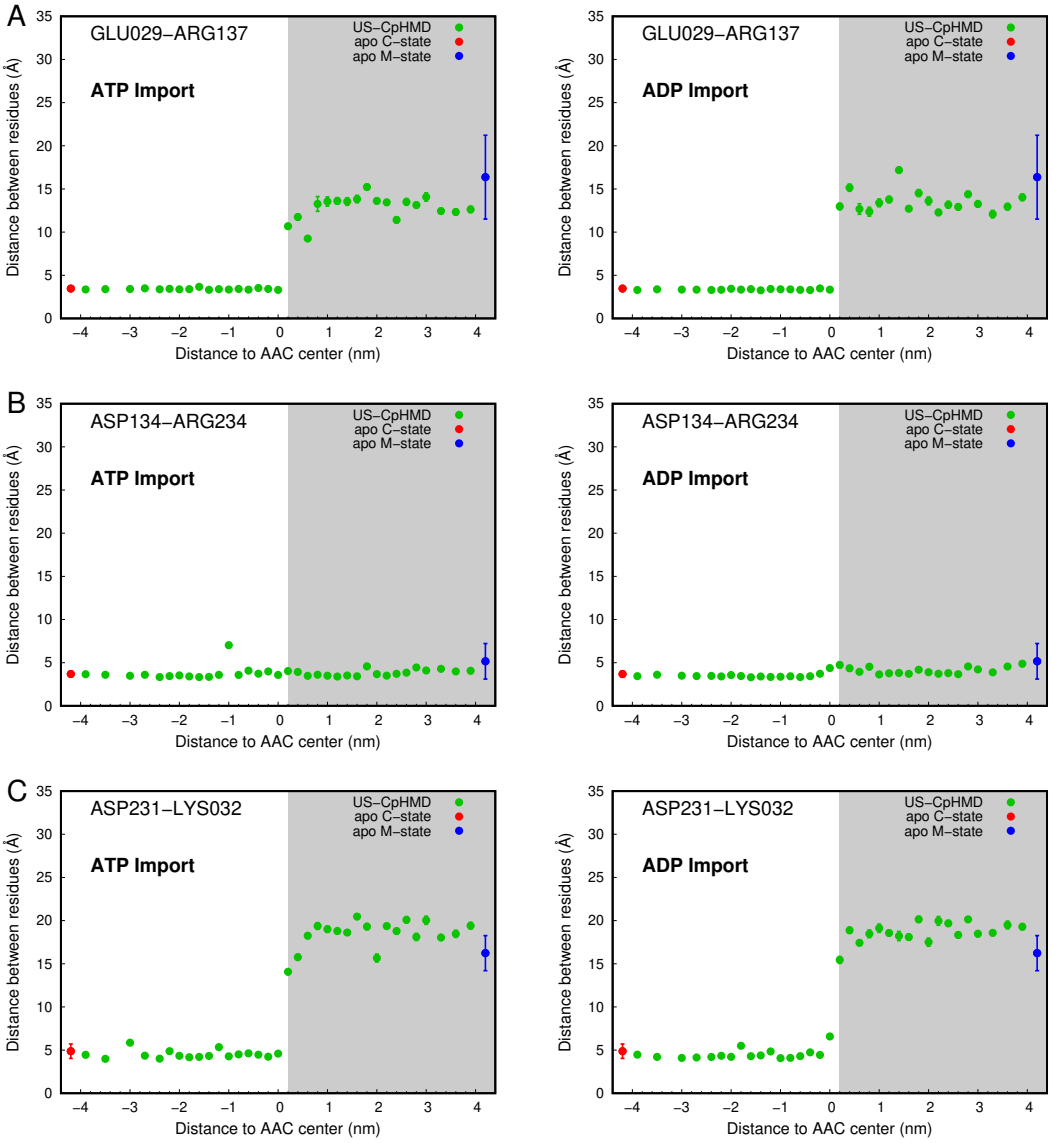

Figure 3.23: Average distances between the salt-bridge pairs: Glu29–Arg137 (A), Asp134–Arg234 (B), and Asp231–Lys32 (C), in each US-CpHMD window for the import process (green points) of both ATP (left) and ADP (right). For comparison, the average distance of each pair in the apo-AAC pH 7 simulations of C- (red point) and M-state (blue point) is also shown. The US windows obtained from the C- and M-state have white and gray backgrounds, respectively.

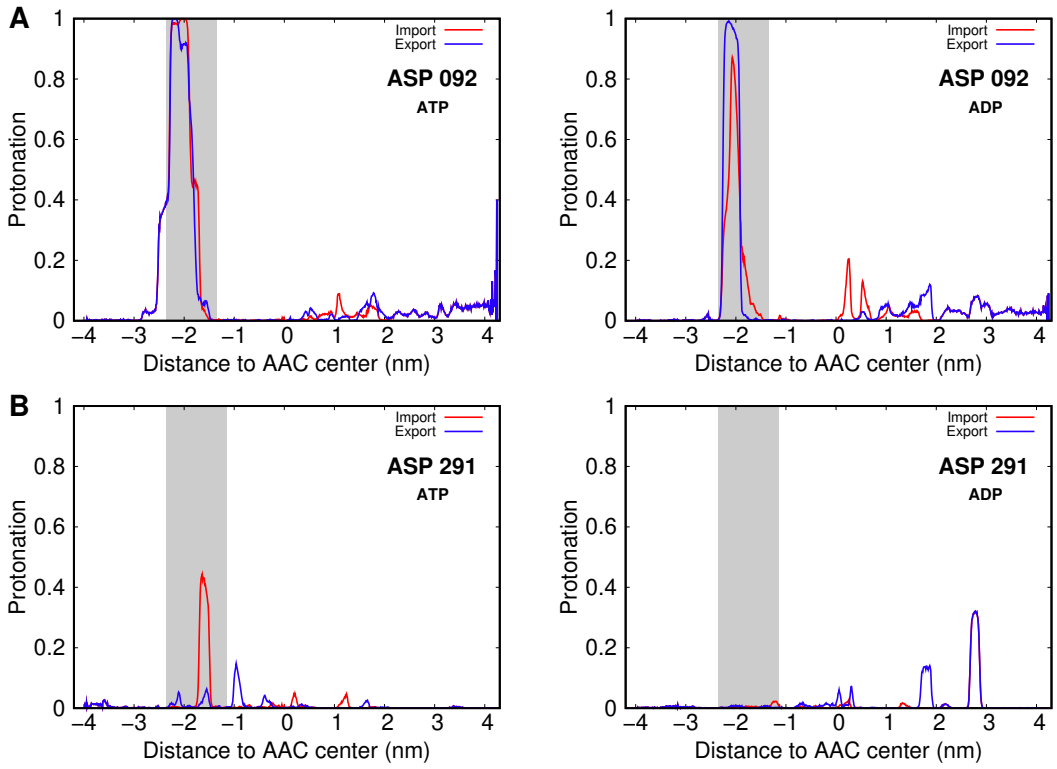

Figure 3.24: Protonation profiles of two acidic amino acids in the cytoplasmic salt-bridge network, Asp92 (A) and 291 (B), during the import (red line) and export (blue line) of both ATP (left) and ADP (right). A background gray area highlights the region sampled by each amino acid during the simulations.

Chapter 4

Concluding Remarks

In the present work, we propose an effective atomistic computational protocol to study the electrostatic details involved in the transport of ATP and ADP by AAC across the inner mitochondrial membrane. Throughout this work, three computational techniques were employed with distinct goals in mind: firstly, long MD simulations were performed to equilibrate our protein in the correct membrane environment; secondly, CpHMD simulations allowed the study of the apo-AAC at different pH values; and lastly, US-CpHMD simulations were performed to mimic the complete transport processes (import and export), providing us information about the energetics and the structural/protonation changes associated with the substrate transport.

Dealing with a system composed by a large transmembrane protein embedded in a lipid bilayer patch is itself a considerable challenge. At the start of our work, in the equilibrium MD simulations we noticed that using a GRF methodology to treat the long-range electrostatic interactions led to structural instabilities. To pursue the causes of these instabilities was out of the scope of this thesis. Alternatively, we employed the PME method, which resulted in very stable protein/mmebrane complexes throughout the simulations. Unfortunately, using PME and CpHMD simulations leads to a slight increase in computational cost and the need to bring our system close to neutrality.

After equilibration of the apo-AAC:POPC system, we performed CpHMD simulations on both apo-AAC states (C- and M-state). This allowed us to extract titration curves and pK_a values for several AAC residues at the two endpoint configurations of the transport process. From these results, we reproduced the important pK_a shift of Lys22, from the water soluble pK_a value (10.4) to ~ 8, which had been reported in the literature [52]. In fact, the deprotonation of this residue has been proposed as the cause for an activity decrease for this carrier [52]. Our data also identified Glu29, Asp134, and Asp231 residues due to their shifted pK_a values. These are concealed in the protein cavity and establish several electrostatic interactions with basic aminoacids, forming what is called the matrix salt-bridge network. Although this interaction network is well known in the literature, it is remarkable that our methodology captured them with a lot of electrostatics information and atomistic detail.

Going a step further, we devised a strategy to mimic the transport of both our substrates (ATP and ADP) while capturing the conformational and electrostatic events coupled to AAC activity. The use of enhanced sampling methodologies, in this case Umbrella Sampling, coupled with CpHMD was crucial for this step. From the complex analysis of these results, we identified clear differences in the import and export processes, allowing the construction of a step-by-step mechanism for the AAC transport. We observed a larger positive electrostatic potential in the cavity of the C-state, at the start of the import process, than in the M-state, at the start of the export process. Due to this, the C-state attracts substrates at larger distances, compared with the M-state, which we have correlated with the biological context of

this system. We also observed a clear selectivity towards ADP in the import process. In contrast, in the export process there was no difference in the PMF profiles of both substrates, hence, no selectivity. Finally, we also observe a substrate-induced shift in the protonation of key acidic residues of the matrix salt-bridge. These results indicate that the substrate can weaken this salt bridge network and trigger the needed conformational transition to complete the import process.

Our results confirmed several characteristics already known in the literature and provided new data regarding the selectivity of the process and the protonation changes which are coupled to the electrostatic forces driving the transport. We illustrate how the import process was evolutionary pressured to become more efficient and selective towards ADP. This step needs to overcome the membrane potential and deal with the fact that ADP concentration will be significantly lower. Therefore, whenever ADP is available nearby the mitochondria inner membrane, it needs to be funneled to the AAC protein and effectively transported. AAC achieved this by developing a strong positive electrostatic potential on the cavity capable of attracting and trapping ADP inside and prevent its unbinding back to the vast cytoplasmic side. The biological context for the export process is very different from the import due to the small volume of the mitochondrial matrix compartment and high abundance of ATP, which is being actively synthesized by ATP synthase. Consequently, a larger number of ATP molecules will be readily available near the AAC M-state cavity resulting in almost no need to positively select it over ADP. Therefore, the magnitude of the mentioned evolutionary pressure on AAC export process has most probably been significantly lower.

Taking all our extensive work into account, the obtained results show the success of using a computational protocol to analyse key mechanistic details at the atomic level. Employing the presented techniques allowed us not only to pinpoint key electrostatic interactions and environment effects, which are crucial for the activity of the protein, but also to extract energetic profiles on the transport process, aiding in the construction of a detailed view of the events that take place during the activity of the protein. Application of this protocol to other less studied proteins and biomolecular processes may help us elucidate their molecular electrostatic details, which are usually very difficult to capture using wet-lab experiments.

4.1 Future Work

Taking the extensive work performed in this thesis, it is our utmost intention to compile both the description of our protocol and the results obtained in a scientific paper. However, to complete the analysis and results it is yet necessary to evaluate the impact of performing a re-weight on the protonation profiles obtained in the US-CpHMD simulations. From previous tests done in our research group, the use of re-weighting wields only small differences, nevertheless it is important to verify if such is the case in our data. Upon completing this task and since we have only performed CpHMD simulations of pH 7.0 on the US-CpHMD, we might be interested in testing the transport process in low pH, probably at the value of 5.0, in order to evaluate differences in the energy of the substrates and in the electrostatic interactions during the activity of the protein. However, due to the time these will take to perform, they will most likely be absent in the paper publication. Finally, and probably most importantly, we will seek to make plans on using the presented protocol in less-studied proteins with biological relevance, both to further validate our work but also to study mechanistic and electrostatic details yet to be revealed in such systems.

Bibliography

[1] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*. New York, New York: WH Freeman and Company, 2005.

[2] P. Carbonell, R. Nussinov, and A. del Sol. "Energetic determinants of protein binding specificity: insights into protein interaction networks". *Proteomics*, 9, 7 (2009), pp. 1744–1753.

[3] Z. Zhang, S. Witham, and E. Alexov. "On the role of electrostatics in protein-protein interactions". *Phys. Biol.* 8, 13 (2011), p. 035001.

[4] A. H. Elcock et al. "Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin". *J. Mol. Biol.* 291, 1 (1999), pp. 149 –162.

[5] Z. Radic et al. "Electrostatic Influence on the Kinetics of Ligand Binding to Acetylcholinesterase: Distinctions between Active Center Ligands and Fasciculin". *J. Biol. Chem.* 272 (1997), pp. 23265–77.

[6] J. Guenot, R. J. Fletterick, and P. A. Kollman. "A negative electrostatic determinant mediates the association between the *Escherichia coli* trp repressor and its operator DNA". *Protein Sci.* 3, 8 (1994), pp. 1276–1285.

[7] L. Johnson and D. Barford. "Electrostatic effects in the control of glycogen phosphorylase by phosphorylation". *Protein Sci.* 3, 10 (1994), pp. 1726–1730.

[8] P. J. Martel, A. Baptista, and S. B. Petersen. "Protein electrostatics". *Biotechnology annual review*. Vol. 2. Elsevier, 1996, pp. 315–372.

[9] P. Kukić and J. E. Nielsen. "Electrostatics in proteins and protein–ligand complexes". *Future Med. Chem.* 2, 4 (2010), pp. 647–666.

[10] A. D. MacKerell, M. Sommer, and M. Karplus. "pH dependence of binding reactions from free energy simulations and macroscopic continuum electrostatic calculations: application to 2'GMP/3'GMP binding to ribonuclease T1 and implications for catalysis." *J. Mol. Biol.* 247 4 (1995), pp. 774–807.

[11] A.-S. Yang and B. Honig. "On the pH Dependence of Protein Stability". *J. Mol. Biol.* 231, 2 (1993), pp. 459–474.

[12] E. Bailey, F. Stirpe, and C. B. Taylor. "Regulation of rat liver pyruvate kinase. The effect of preincubation, pH, copper ions, fructose 1,6-diphosphate and dietary changes on enzyme activity". *Biochem. J.* 108, 3 (1968), pp. 427–436.

[13] F. Leprince and H. Quiquampoix. "Extracellular enzyme activity in soil: effect of pH and ionic strength on the interaction with montmorillonite of two acid phosphatases secreted by the ectomycorrhizal fungus Hebeloma cylindrosporum". *Eur. J. Soil Sci.* 47, 4 (1996), pp. 511–522.

[14] G. Schreiber and A. R. Fersht. “Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering”. *Biochemistry*, 32, 19 (1993), pp. 5145–5150.

[15] M. F. García-Mayoral et al. “pH-Dependent Conformational Stability of the Ribotoxin α-Sarcin and Four Active Site Charge Substitution Variants”. *Biochemistry*, 45, 46 (2006), pp. 13705–13718.

[16] D. T. Bravo, N. G. Kolmakova, and S. M. Parsons. “Mutational and pH Analysis of Ionic Residues in Transmembrane Domains of Vesicular Acetylcholine Transporter.” *Biochemistry*, 44, 22 (2005), pp. 7955–7966.

[17] M. Saier. *Families of Transporters and Their Classification*. John Wiley and Sons Ltd, 2003. Chap. 1, pp. 1–17.

[18] H.-X. Zhou and X. Pang. “Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation”. *Chem. Rev.* 118, 4 (2018), pp. 1691–1741.

[19] D. Deng et al. “Crystal structure of the human glucose transporter GLUT1”. *Nature*, 510, 7503 (2014), pp. 121–125.

[20] A. F. Kintzer and R. M. Stroud. “Structure, inhibition and regulation of two-pore channel TPC1 from Arabidopsis thaliana”. *Nature*, 531, 7593 (2016), pp. 258–262.

[21] D. A. Doyle. “The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity”. *Science*, 280, 5360 (1998), pp. 69–77.

[22] C. G. Bailey et al. “Loss-of-function mutations in the glutamate transporter SLC1A1 cause human dicarboxylic aminoaciduria”. *J. Clin. Investig.* 121, 1 (2011), pp. 446–453.

[23] R. M. Ryan et al. “The position of an arginine residue influences substrate affinity and K+ coupling in the human glutamate transporter, EAAT1”. *J. Neurochem.* 114, 2 (2010), pp. 565–575.

[24] S. A. Hollingsworth and R. O. Dror. “Molecular Dynamics Simulation for All”. *Neuron*, 99, 6 (2018), pp. 1129 –1143.

[25] W. F. van Gunsteren and H. J. C. Berendsen. “Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry”. *Angew Chem. Int. Ed.* 29, 9 (1990), pp. 992–1023.

[26] R. L. Andrew. *Molecular modeling principles and applications*. Prentice Hall, 2001.

[27] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, 2017.

[28] A. M. Baptista, V. H. Teixeira, and C. M. Soares. “Constant-pH molecular dynamics using stochastic titration”. *J. Chem. Phys.* 117, 9 (2002), pp. 4184–4200.

[29] M. Machuqueiro and A. M. Baptista. “Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation–Conformation Coupling in Decalysine”. *J. Phys. Chem. B*, 110 (2006), pp. 2927–2933.

[30] V. H. Teixeira, A. S. C. Capacho, and M. Machuqueiro. “The role of electrostatics in TrxR electron transfer mechanism: A computational approach”. *Proteins Struct. Funct. Bioinf.* 84, 12 (2016), pp. 1836–1843.

[31] D. Vila-Viçosa et al. “A pH Replica Exchange Scheme in the Stochastic Titration Constant-pH MD Method”. *J. Chem. Theory Comput.* 15, 5 (2019), pp. 3108–3116.

[32] A. Barducci, M. Bonomi, and M. Parrinello. “Metadynamics”. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1, 5 (2011), pp. 826–843.

[33] J. Kästner. “Umbrella sampling”. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 1, 6 (2011), pp. 932–942.

[34] B. Roux. “The calculation of the potential of mean force using computer simulations”. *Comput. Phys. Commun.* 91, 1 (1995), pp. 275 –282.

[35] H. Heldt, H. Jacobs, and M. Klingenberg. “Endogenous ADP of mitochondria, an early phosphate acceptor of oxidative phosphorylation as disclosed by kinetic studies with C^14 labelled ADP and ATP and with atractyloside”. *Biochem. Biophys. Res. Commun.* 18, 2 (1965), pp. 174 –179.

[36] E. Pfaff, M. Klingenberg, and H. Heldt. “Unspecific permeation and specific exchange of adenine nucleotides in liver mitochondria”. *Biochem. Biophys. Acta*, 104, 1 (1965), pp. 312 –315.

[37] E. Duee and P. Vignais. “Éxchange entre adenine-nucleotides extra- et intramitochondriaux”. *Biochem. Biophys. Acta*, 107, 1 (1965), pp. 184 –188.

[38] E. Pebay-Peyroula et al. “Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside”. *Nature*, 426, 6962 (2003), pp. 39–44.

[39] J. J. Ruprecht et al. “The Molecular Mechanism of Transport by the Mitochondrial ADP/ATP Carrier”. *Cell*, 176, 3 (2019), 435–447.e15.

[40] M. Klingenberg. “The ADP and ATP transport in mitochondria and its carrier”. *Biochem. Biophys. Acta*, 1778, 10 (2008), pp. 1978 –2021.

[41] H Nury et al. “Relations between structure and function of the mitochondrial ADP/ATP carrier”. *Annu. Rev. Biochem.* 75 (2006), pp. 713–741.

[42] E. R. Kunji et al. “The transport mechanism of the mitochondrial ADP/ATP carrier”. *Biochem. Biophys. Acta*, 1863, 10 (2016), pp. 2379 –2393.

[43] A. Kemp and E. Slater. “The site of action of atractyloside”. *Biochem. Biophys. Acta*, 92, 1 (1964), pp. 178–180.

[44] J. B. Chappell and A. R. Crofts. “The effect of atractylate and oligomycin on the behaviour of mitochondria towards adenine nucleotides”. *Biochem. J.* 95, 3 (1965), p. 707.

[45] P. Vignais, P. Vignais, and G. Defaye. “Gummiferin, an inhibitor of the adenine-nucleotide translocation. Study of its binding properties to mitochondria”. *FEBS Letters*, 17, 2 (1971), pp. 281–288.

[46] P. J. Henderson and H. A. Lardy. “Bongkrekic acid. An inhibitor of the adenine nucleotide translocase of mitochondria”. *J. Biol. Chem.* 245, 6 (1970), pp. 1319–1326.

[47] G. J. M. Lauquin et al. “Isobongkrekic acid, a new inhibitor of mitochondrial ADP-ATP transport: radioactive labeling and chemical and biological properties”. *Biochemistry*, 15, 11 (1976), pp. 2323–2327.

[48] J. J. Ruprecht and E. R. Kunji. “Structural changes in the transport cycle of the mitochondrial ADP/ATP carrier”. *Curr. Opin. Struct. Biol.* 57 (2019), pp. 135 –144.

[49] A. J. Robinson, C. Overy, and E. R. S. Kunji. “The mechanism of transport by mitochondrial carriers based on analysis of symmetry”. *Proc. Natl. Acad. Sci. USA*, 105, 46 (2008), pp. 17766–17771.

[50] E. Pfaff and M. Klingenberg. “Adenine Nucleotide Translocation of Mitochondria”. *Eur. J. Biochem.* 6, 1 (1968), pp. 66–79.

[51] Y. Wang and E. Tajkhorshid. “Electrostatic funneling of substrate in mitochondrial inner membrane carriers”. *Proc. Natl. Acad. Sci. USA*, 105, 28 (2008), pp. 9598–9603.

[52] A. Bidon-Chanal et al. “How Do Membrane Transporters Sense pH? The Case of the Mitochondrial ADP–ATP Carrier”. *J. Phys. Chem. Lett.* 4, 21 (2013), pp. 3787–3791.

[53] M. S. King et al. “Formation of a cytoplasmic salt bridge network in the matrix state is a fundamental step in the transport mechanism of the mitochondrial ADP/ATP carrier”. *Biochem. Biophys. Acta*, 1857 (2016), pp. 14–22.

[54] M. Wiltgen and G. P. Tilz. “Homology modelling: a review about the method on hand of the diabetic antigen GAD 65 structure prediction”. *Wiener Medizinische Wochenschrift*, 159, 5 (2009), pp. 112–125.

[55] M. Baker. “Making membrane proteins for structures: a trillion tiny tweaks”. *Nat. Methods*, 7, 6 (2010), pp. 429–434.

[56] M. T. Muhammed and E. Aki-Yalcin. “Homology modeling in drug discovery: Overview, current applications, and future perspectives”. *Chem. Biol. Drug Des.* 93, 1 (2019), pp. 12–20.

[57] E. Krieger, S. Nabuurs, and G. Vriend. *Structural Bioinformatics*. Wiley-Liss, 2012, pp. 507–520.

[58] S. F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” *Nucleic Acids Res.* 25, 17 (1997), pp. 3389–3402.

[59] J. Söding. “Protein homology detection by HMM-HMM comparison”. *Bioinformatics*, 21, 7 (2005), pp. 951–960.

[60] N. Eswar et al. “Comparative Protein Structure Modeling Using Modeller”. *Curr. Protoc. Bioinformatics*, 15, 1 (2006), pp. 5.6.1–5.6.30.

[61] M.-y. Shen and A. Sali. “Statistical potential for assessment and prediction of protein structures”. *Protein Sci.* 15, 11 (2006), pp. 2507–2524.

[62] P. M. Morse. “Diatomic Molecules According to the Wave Mechanics. II. Vibrational Levels”. *Phys. Rev.* 34 (1 1929), pp. 57–64.

[63] X. Daura, A. E. Mark, and W. F. Van Gunsteren. “Parametrization of aliphatic CHn united atoms of GROMOS96 force field”. *J. Comput. Chem.* 19, 5 (1998), pp. 535–547.

[64] R. W. Hockney. “POTENTIAL CALCULATION AND SOME APPLICATIONS.” *Methods Comput. Phys.* 9 (1970), p. 136.

[65] W. C. Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. *J. Chem. Phys.* 76, 1 (1982), pp. 637–649.

[66] G. Bussi, D. Donadio, and M. Parrinello. “Canonical sampling through velocity rescaling”. *J. Chem. Phys.* 126, 1 (2007), p. 014101.

[67] H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath”. *J. Chem. Phys.* 81, 8 (1984), pp. 3684–3690.

[68] M. Parrinello and A. Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. *J. Appl. Phys.* 52, 12 (1981), pp. 7182–7190.

[69] S. Nosé and M. Klein. “Constant pressure molecular dynamics for molecular systems”. *Mol. Phys.* 50, 5 (1983), pp. 1055–1076.

[70] R. Fletcher and M. J. D. Powell. “A Rapidly Convergent Descent Method for Minimization”. *Comput. J.* 6, 2 (1963), pp. 163–168.

[71] M. Abraham et al. *GROMACS User Manual version 5.1.5*. www.gromacs.org, 2017.

[72] R. H. Byrd et al. “A limited memory algorithm for bound constrained optimization”. *SIAM J. Sci. Comput.* 16 (1995), pp. 1190–1208.

[73] C. Zhu et al. “Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization”. *ACM Trans. Math. Soft.* 23, 4 (1997), 550–560.

[74] T. F. D. Silva et al. “The Impact of Using Single Atomistic Long-Range Cutoff Schemes with the GROMOS 54A7 Force Field”. *J. Chem. Theory Comput.* 14, 11 (2018), pp. 5823–5833.

[75] I. G. Tironi et al. “A generalized reaction field method for molecular dynamics simulations”. *J. Chem. Phys.* 102 (1995), pp. 5451–5459.

[76] T. Darden, D. York, and L. Pedersen. “Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems”. *J. Chem. Phys.* 98, 12 (1993), pp. 10089–10092.

[77] U. Essmann et al. “A smooth particle mesh Ewald method”. *J. Chem. Phys.* 103, 19 (1995), pp. 8577–8593.

[78] L. Onsager. “Electric Moments of Molecules in Liquids”. *J. Am. Chem. Soc.* 58, 8 (1936), pp. 1486–1493.

[79] P. P. Ewald. “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. *Annalen der Physik*, 369, 3 (1921), pp. 253–287.

[80] J. S. Hub et al. “Quantifying Artifacts in Ewald Simulations of Inhomogeneous Systems with a Net Charge”. *J. Chem. Theory Comput.* 10, 1 (2014), pp. 381–390.

[81] P. Debye and E. Hückel. “Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen”. *Physikalische Zeitschrift*, 24, 185 (1923), p. 305.

[82] M. K. Gilson, K. A. Sharp, and B. H. Honig. “Calculating the electrostatic potential of molecules in solution: Method and error assessment”. *J. Comput. Chem.* 9, 4 (1988), pp. 327–335.

[83] P Beroza et al. “Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides.” *Proc. Natl. Acad. Sci. USA*, 88, 13 (1991), pp. 5804–5808.

[84] N. Metropolis and S. Ulam. “The Monte Carlo Method”. *J. Am. Stat. Assoc.* 44, 247 (1949), pp. 335–341.

[85] N. Metropolis et al. “Equation of State Calculations by Fast Computing Machines”. *J. Chem. Phys.* 21, 6 (1953), pp. 1087–1092.

[86] S. Kumar et al. “The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method”. *J. Comput. Chem.* 13, 8 (1992), pp. 1011–1021.

[87] S. Izrailev et al. “Steered molecular dynamics”. *Computational molecular dynamics: challenges, methods, ideas*. Springer, 1999, pp. 39–65.

[88] Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8*. 2015.

[89] J. Hermans et al. "A Consistent Empirical Potential for Water-Protein Interactions". *Biopolymers*, 23 (1984), pp. 1513–1518.

[90] B. Webb and A. Sali. "Comparative Protein Structure Modeling Using MODELLER". *Curr. Protoc. Bioinformatics*, 54, 1 (2016), pp. 5.6.1–5.6.37.

[91] R. A. Laskowski et al. "PROCHECK: a program to check the stereochemical quality of protein structures". *J. Appl. Crystallogr.* 26, 2 (1993), pp. 283–291.

[92] D. Van Der Spoel et al. "GROMACS: Fast, flexible, and free". *J. Comput. Chem.* 26, 16 (2005), pp. 1701–1718.

[93] M. J. Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". *SoftwareX*, 1-2 (2015), pp. 19 –25.

[94] N. Schmid et al. "Definition and testing of the GROMOS force-field versions 54A7 and 54B7." *Eur. Biophys. J.* 40 (2011), p. 843.

[95] B. Hess. "P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation". *J. Chem. Theory Comput.* 4, 1 (2008), pp. 116–122.

[96] S. Miyamoto and P. A. Kollman. "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models". 13, J. Comput. Chem. (1992), pp. 952–962.

[97] P. B.P. S. Reis et al. "Role of Counterions in Constant-pH Molecular Dynamics Simulations of PAMAM Dendrimers". *ACS Omega*, 3, 2 (2018), pp. 2001–2009.

[98] C. Pereira et al. "ADP/ATP carrier is required for mitochondrial outer membrane permeabilization and cytochrome c release in yeast apoptosis". *Molecular Microbiology*, 66, 3 (2007), pp. 571–582.

[99] S. Canzar et al. "Charge Group Partitioning in Biomolecular Simulation". *J. Comput. Biol.* 20, 3 (2013), pp. 188–198.

[100] W. Rocchia et al. "Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: Applications to the molecular systems and geometric objects". *J. Comput. Chem.* 23, 1 (2002), pp. 128–137.

[101] V. H. Teixeira et al. "On the Use of Different Dielectric Constants for Computing Individual and Pairwise Terms in Poisson-Boltzmann Studies of Protein Ionization Equilibrium". *J. Phys. Chem. B*, 109 (2005), pp. 14691–14706.

[102] D. Vila-Viçosa et al. "On the treatment of ionic strength in biomolecular simulations of charged lipid bilayers". *J. Chem. Theory Comput.* 10 (2014), pp. 5483–5492.

[103] V. H. Teixeira et al. "Protonation of DMPC in a Bilayer Environment Using a Linear Response Approximation". *J. Chem. Theory Comput.* 10 (2014), pp. 2176–2184.

[104] H. A. Santos et al. "Constant-pH MD simulations of DMPA/DMPC lipid bilayers". *J. Chem. Theory Comput.* 11, 12 (2015), pp. 5973–5979.

[105] A. M. Baptista and C. M. Soares. "Some Theoretical and Computational Aspects of the Inclusion of Proton Isomerism in the Protonation Equilibrium of Proteins". *J. Phys. Chem. B*, 105 (2001), pp. 293–309.

[106] A. M. Baptista, P. J. Martel, and C. M. Soares. “Simulation of Electron-Proton Coupling with a Monte Carlo Method: Application to Cytochrome c_3 Using Continuum Electrostatics”. *Biophys. J.* 76 (1999), pp. 2978–2998.

[107] A. K. Malde et al. “An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0”. *J. Chem. Theory Comput.* 7, 12 (2011), pp. 4026–4037.

[108] K. B. Koziara et al. “Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies”. *J. Computer-Aided Mol. Design*, 28, 3 (2014), pp. 221–233.

[109] M. Stroet et al. “Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane”. *J. Chem. Theory Comput.* 14, 11 (2018), pp. 5834–5845.

[110] N. F. B. Oliveira, I. D. S. Pires, and M. Machuqueiro. “Improved GROMOS 54A7 Charge Sets for Phosphorylated Tyr, Ser, and Thr to Deal with pH-Dependent Binding Phenomena”. *J. Chem. Theory Comput.* 16, 10 (2020), pp. 6368–6376.

[111] A. U. Singer and J. D. Forman-Kay. “pH Titration studies of an SH2 domain-phosphopeptide complex: Unusual histidine and phosphate pK_a values”. *Protein Sci.* 6, 9 (1997), pp. 1910–1919.

[112] R. Phillips. “Adenosine and the Adenine Nucleotides. Ionization, Metal Complex Formation, and Conformation in Solution”. *Chem. Rev.* 66, 5 (1966), pp. 501–527.

[113] T. Williams, C. Kelley, and many others. *Gnuplot 4.6: an interactive plotting program*. `http://gnuplot.sourceforge.net/`. 2013.

[114] W. Kabsch and C. Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. *Biopolymers*, 22, 12 (1983), pp. 2577–2637.

[115] W. G. Touw et al. “A series of PDB-related databanks for everyday needs”. *Nucleic Acids Res.* 43, D1 (2014), pp. D364–D368.

[116] R. L. Thurlkill et al. “pK values of the ionizable groups of proteins”. *Protein Sci.* 15, 5 (2006), pp. 1214–1218.

[117] D. Vila-Viçosa et al. “Membrane-Induced p K $_{\mathrm{a}}$ Shifts in *wt* -pHLIP and Its L16H Variant”. *J. Chem. Theory Comput.* 14, 6 (2018), pp. 3289–3297.

[118] R. A. Capaldi and R. Aggeler. “Mechanism of the F1F0-type ATP synthase, a biological rotary motor”. *Trends Biochem. Sci.* 27, 3 (2002), pp. 154–160.

[119] D. G. Nicholls and S. J. Ferguson. “The Chemiosmotic Proton Circuit in Isolated Organelles”. *Bioenergetics*. Elsevier, 2013, pp. 53–87.

[120] R. Kraemer and M. Klingenberg. “Modulation of the reconstituted adenine nucleotide exchange by membrane potential”. *Biochemistry*, 19, 3 (1980), pp. 556–560.

Chapter 5

Annex 1: Figures

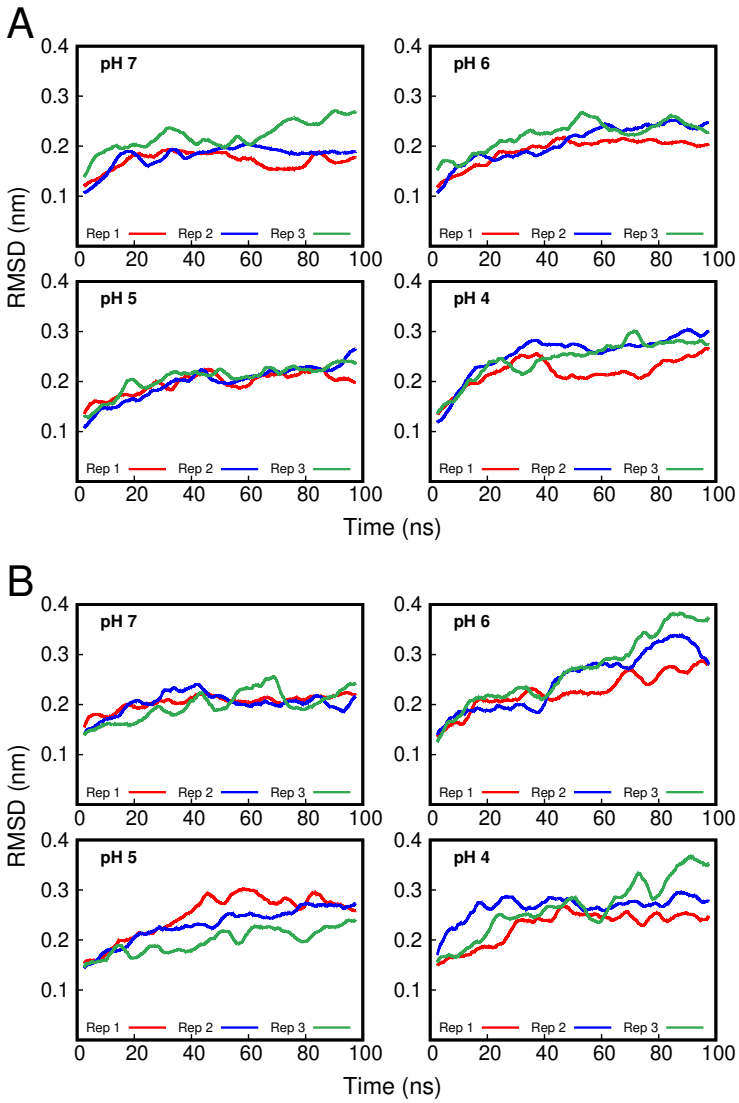

Figure 5.1: Root mean square deviation (RMSD) of the apo C- (A) and M-state (B) CpHMD simulations at the pH values of 7, 6, 5 and 4. Each step is structurally compared with the starting structure obtained from the long MD run of each replicate.The RMSD of each replicate in its respective AAC state was plotted with a floating window of 1 ns in order to reduce local fluctuation of each replicate curve.

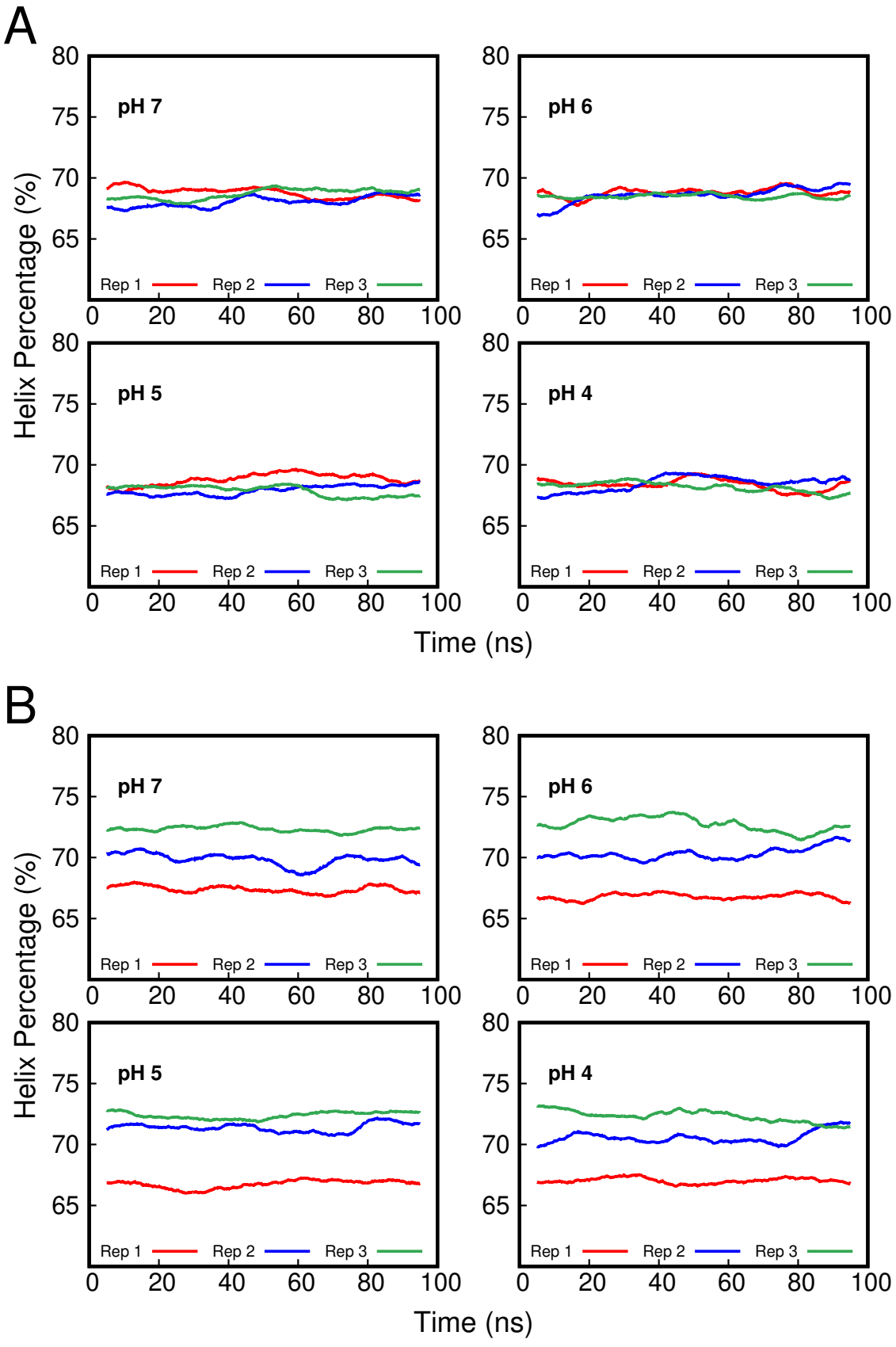

Figure 5.2: Percentage of helix content of the C-state (A) and M-state (B) during the CpHMD simulations at pH 7, 6, 5 and 4, obtained using DSSP calculations [114, 115]. On both plots a floating window of 5 ns was used in order to reduce local fluctuations.

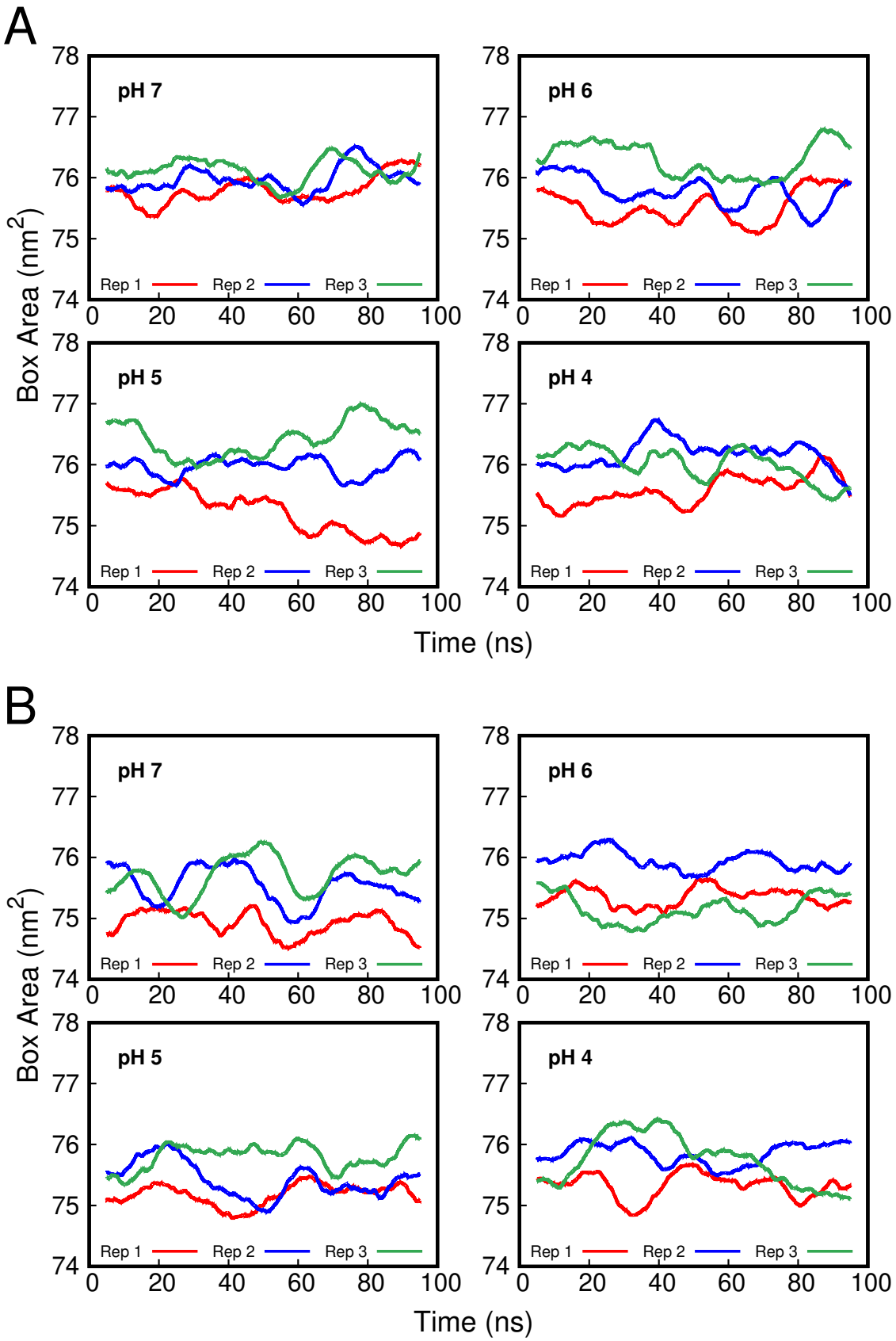

Figure 5.3: Area of the simulation box in the *xy* axis across the CpHMD run at four different pH values. Both C-state (A) and M-state (B) have each individual replicate plotted with a floating window of 10 ns in order to reduce local fluctuations.

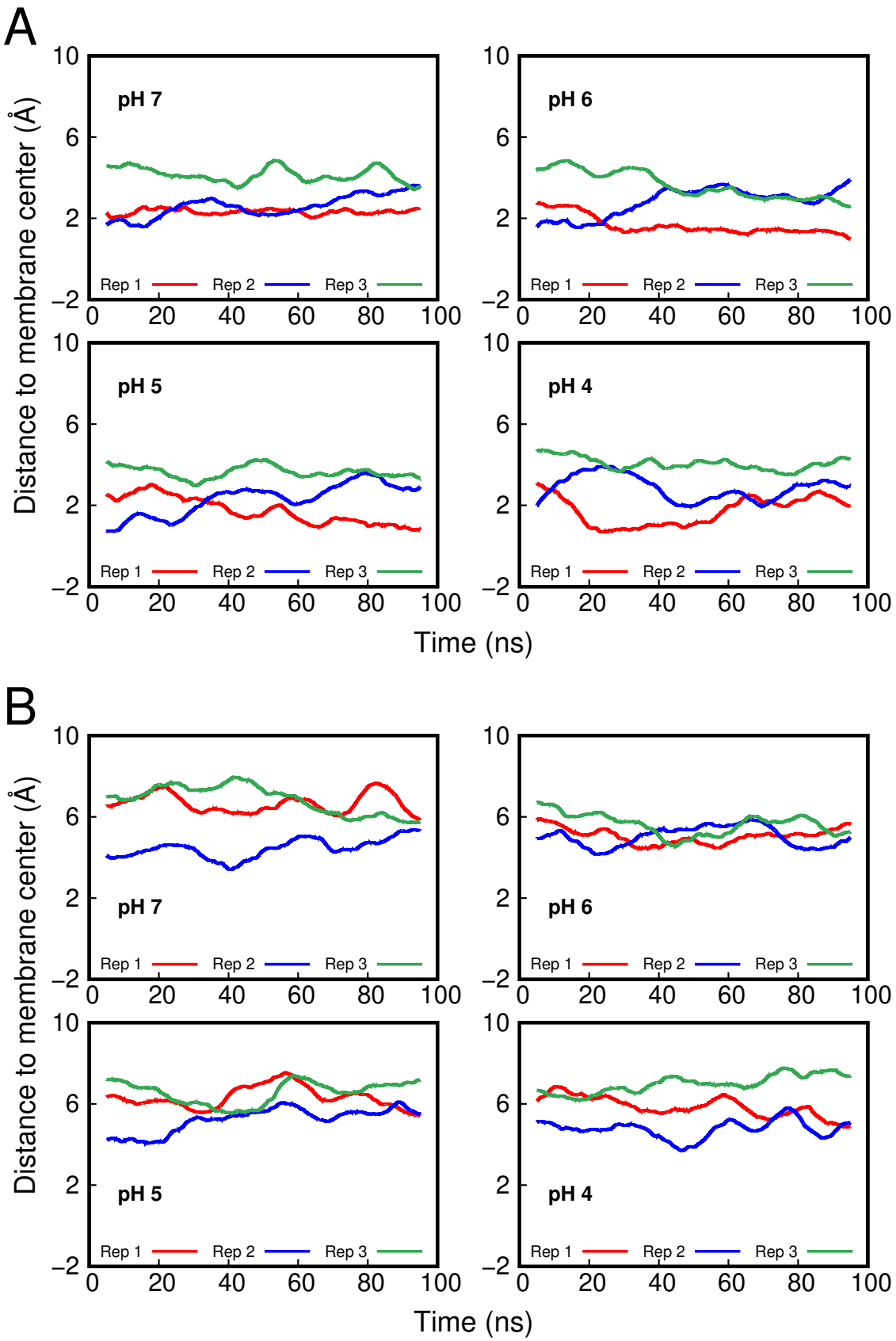

Figure 5.4: Distance to membrane center of the AAC center group in both states, C-state (A) and M-state (B) in the CpHMD at pH 7, 6, 5 and 4. The distance values were plotted using a floating window of 5 ns in order to reduce point fluctuation.

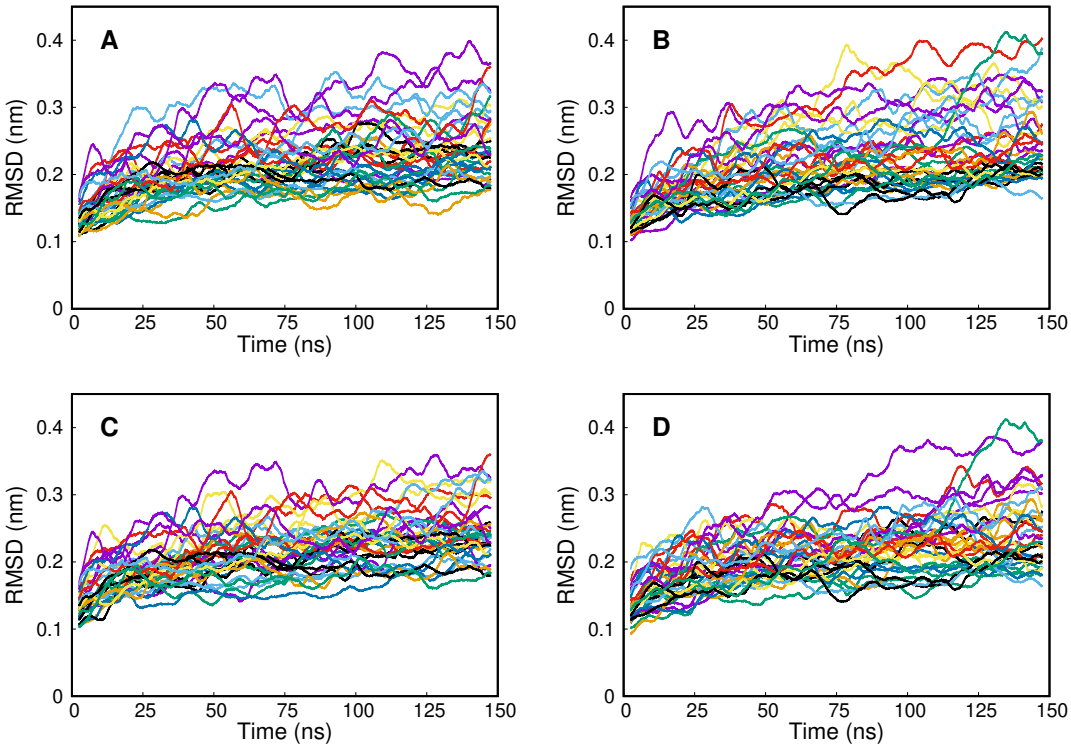

Figure 5.5: Root mean square deviation (RMSD) of each Umbrella Sampling window in the import of ATP (A) and ADP (B) from the cytoplasmic side to the mitochondrial membrane, and the opposite export process of ATP (C) and ADP (D). In all plots, a floating window of 5 ns was used to reduce local fluctuations.

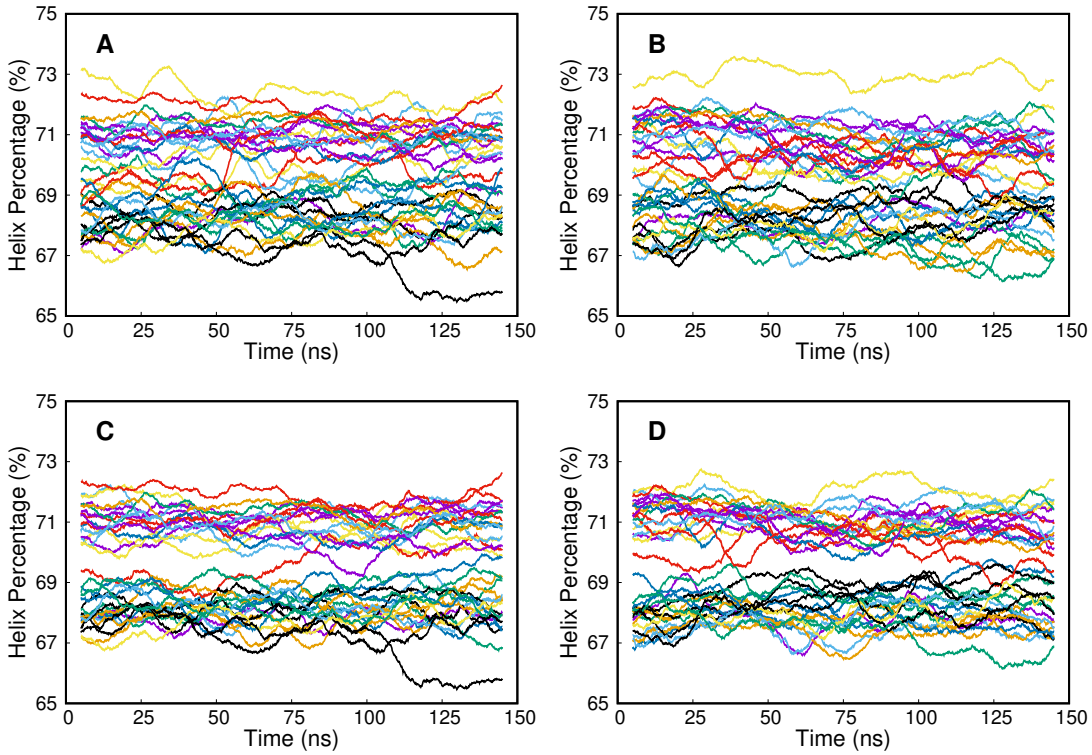

Figure 5.6: Percentage of Helix content of the import of ATP (A) and ADP (B) to the matrix, and also for the export of ATP (C) and ADP (D) from the matrix. A floating window of 10 ns was used to reduce local fluctuations.

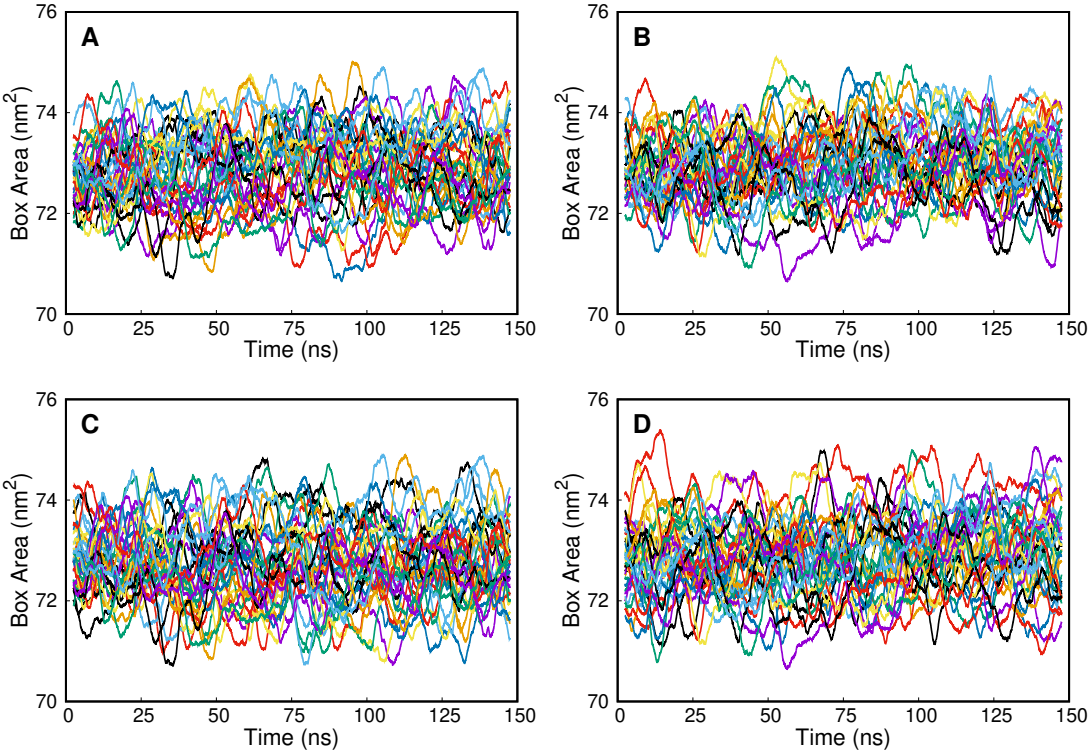

Figure 5.7: Area of the simulation box in the *xy* plane for the import of ATP (A) and ADP (B) to the matrix and the opposite export process of ATP (C) and ADP (D). A floating window of 5 ns was used to reduce local fluctuations.

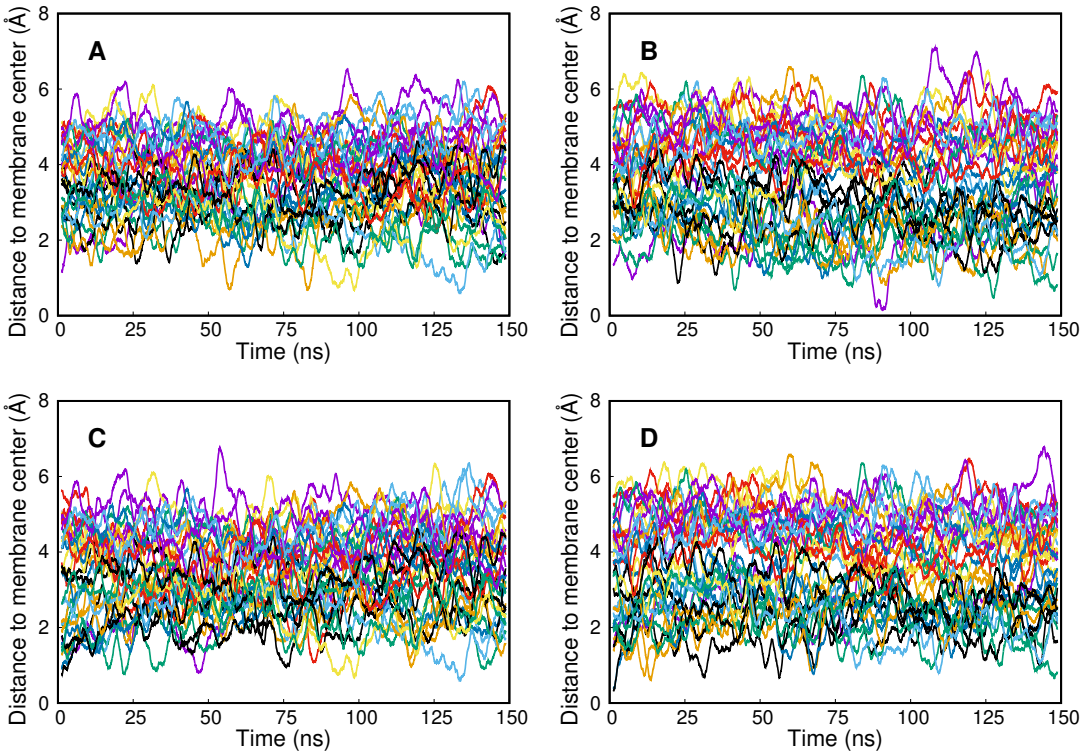

Figure 5.8: Distance of AAC center group to the center of the POPC membrane for the import of ATP (A) and ADP (B), and for the opposite export process of ATP (C) and ADP (D). A floating window of 2.5 ns was used to reduce local fluctuations.

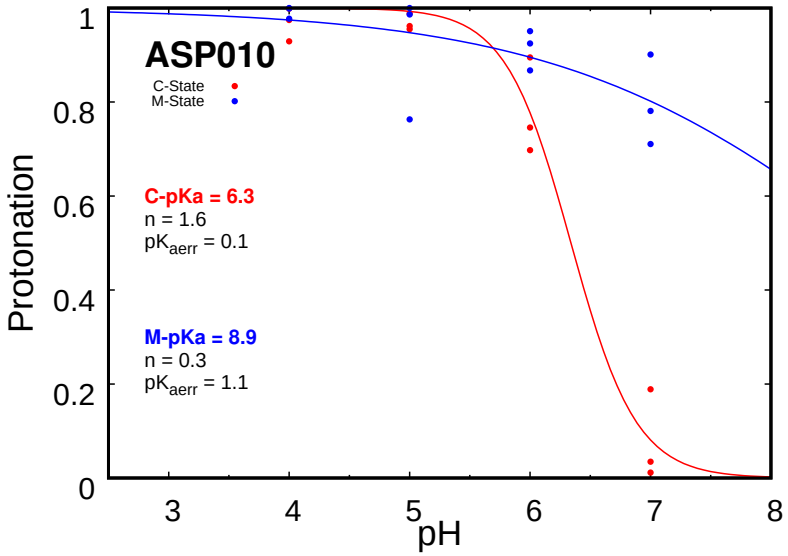

Figure 5.9: Titration curves of Asp10 in the C-state (red line) and in the M-state (blue line). The average protonation values of each replicate at each pH value are represented with colored points. The titration curves were obtained from Hill curve fits to the data points.

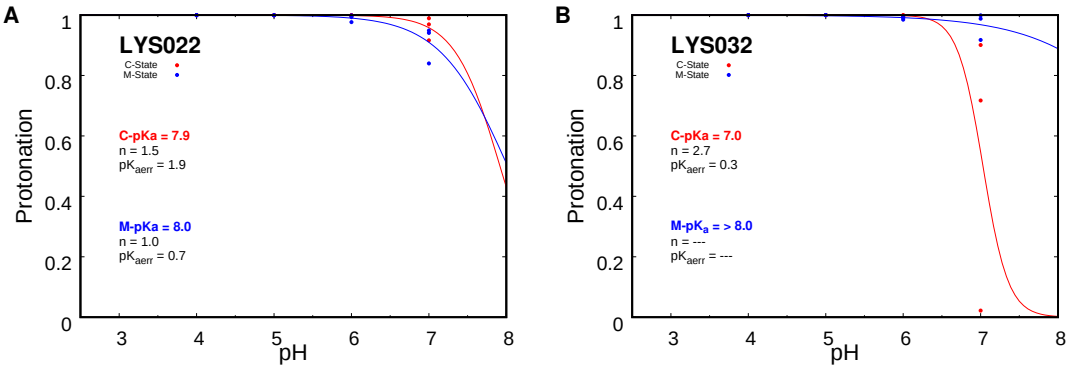

Figure 5.10: Titration curves of Lys22 (A) and 32 (B) in the C-state (red line) and in the M-state (blue line). The average protonation values of each replicate at each pH value are represented with colored points. The titration curves were obtained from Hill curve fits to the data points.

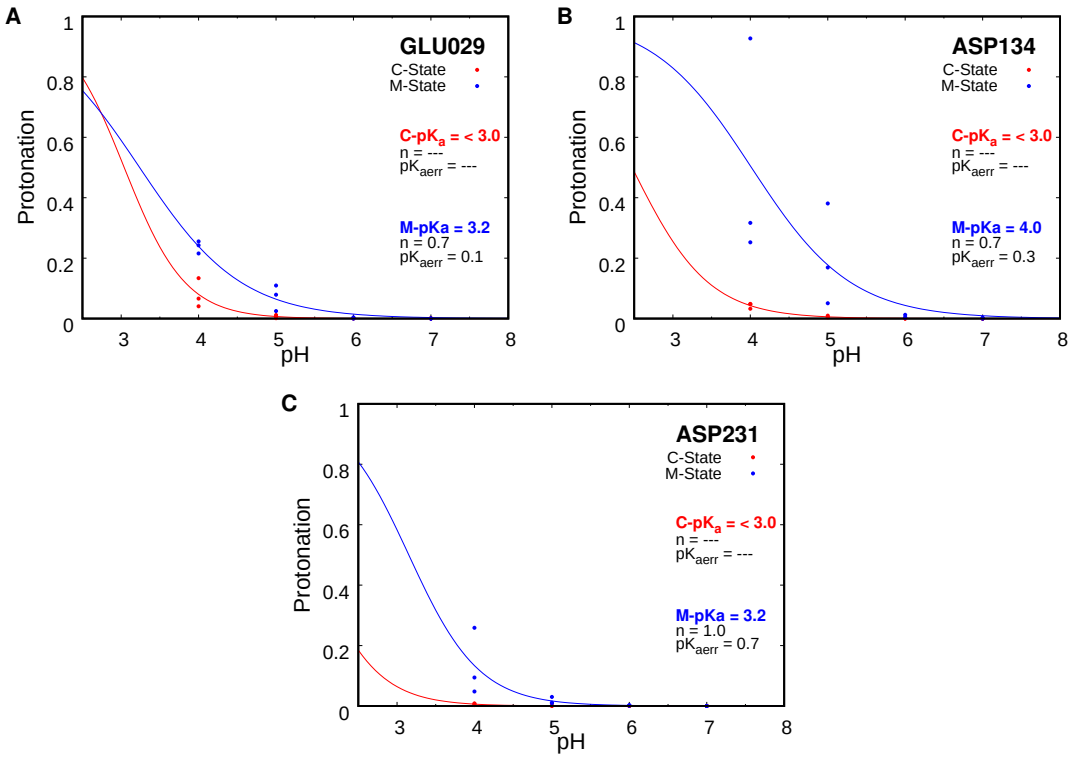

Figure 5.11: Titration curves of Glu29 (A), Asp134 (B), and Asp231 (C) in the C-state (red line) and in the M-state (blue line). The average protonation values of each replicate at each pH value are represented with colored points. The titration curves were obtained from Hill curve fits to the data points.

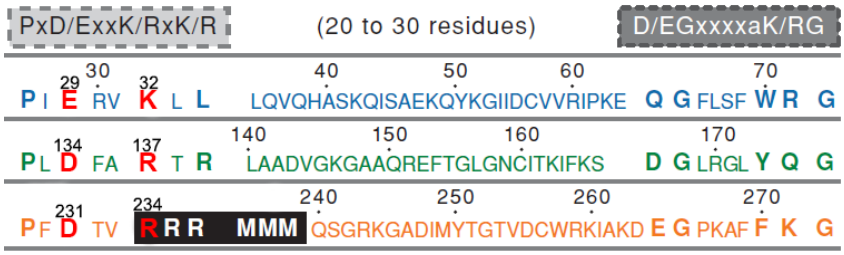

Figure 5.12: *Bos Taurus* MCF sequence motif (P-X-D/E-X-X-R/K-(20 to 30 residues)-D/E-G-X-X-X-X-W/Y/F-K/R-G) present in each odd-numbered trans-membrane helix (H1, H3 and H5). The AAC specific sequence domain (R-R-R-M-M-M) is highlighted in black. The residues, both acid and base, involved in the matrix salt-bridge network are marked in red. Adapted from [41].

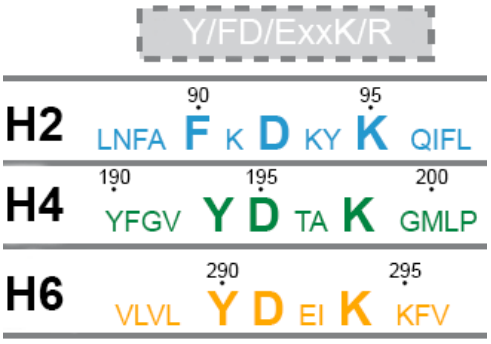

Figure 5.13: *Bos Taurus* Y/F-D/E-x-x-K/R conserved sequence domain present in the even numbered trans-membrane helices (H2, H4 and H6).

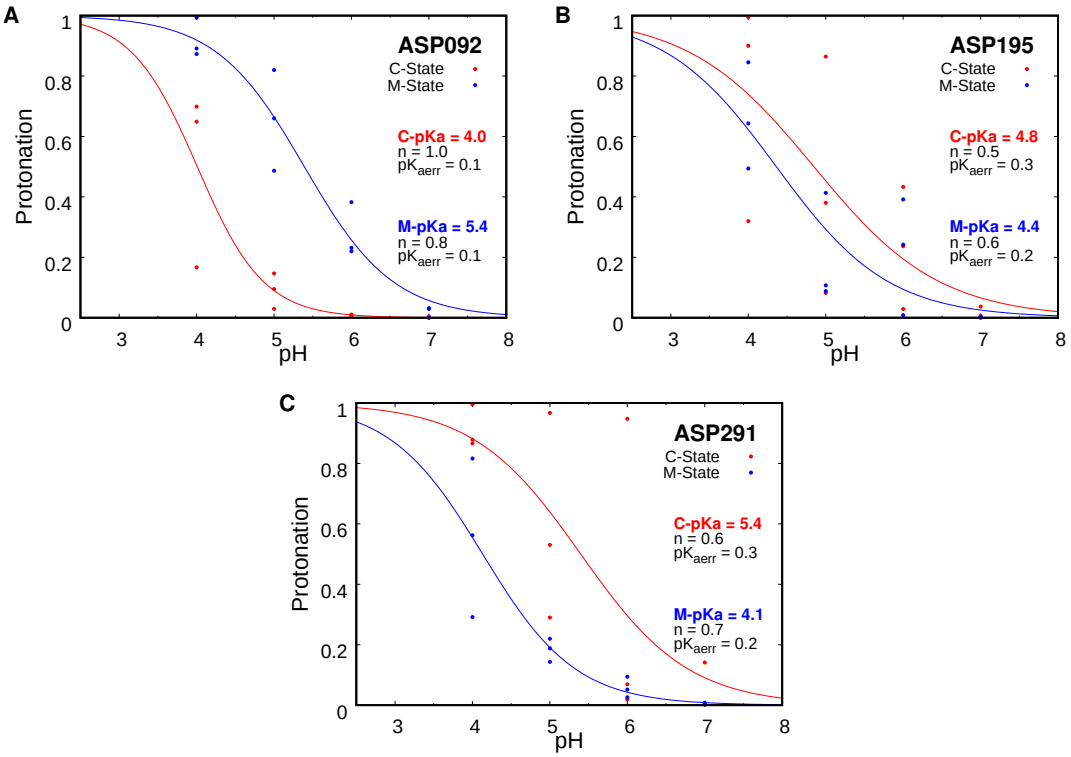

Figure 5.14: Titration curves of Asp92 (A), 195 (B), and 291 (C) in the C-state (red line) and in the M-state (blue line). The average protonation values of each replicate at each pH value are represented with colored points. The titration curves were obtained from Hill curve fits to the data points.