UNIVERSITAT DE BARCELONA

Computational Modelling of the pH Effect on Intrinsically Disordered Proteins

Cristian Privat Contreras

Computational Modelling of the pH Effect on Intrinsically Disordered Proteins

Cristian Privat Contreras

UNIVERSITAT DE BARCELONA

A thesis submitted by

Cristian Privat Contreras

for the degree of Doctor in Philosophy in

Theoretical Chemistry and Computational Modelling

at the Universitat of Barcelona

Computational Modelling of the pH Effect on Intrinsically Disordered Proteins

March 2023

Director:

Dr. Jaime Rubio Martínez
(Universitat de Barcelona)

Director:

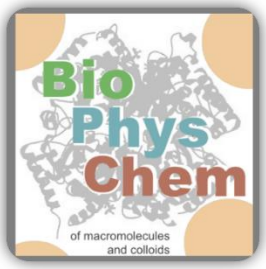
Dr. Sergio Madurga Díez
(Universitat de Barcelona)

Tutor:

Dr. Francesc Mas Pujadas
(Universitat de Barcelona)

Author:

Cristian Privat Contreras
(Universitat de Barcelona)



This doctoral thesis has been conducted in the research groups *Modelling of Biological Systems and Drug Design* of Dr. Jaime Rubio Martínez and *Biophysical Chemistry of Macromolecules and Colloids* of Dr. Francesc Mas Pujadas and Dr. Sergio Madurga Díez of the Department of Materials Science and Physical Chemistry of the *Universitat de Barcelona* in collaboration with the *Institute of Theoretical and Computational Chemistry* (IQTC). The research was supported by the computational resources of the IQTC and the financial support of the *Generalitat de Catalunya* (grant number 2017SGR1033) and the Spanish Structures and Excellence Programme María de Maeztu of the Ministry of Universities of Spain (grant number MDM-2017-0767).

Institut de Química Teòrica
i Computacional
UNIVERSITAT DE BARCELONA

EXCELENCIA
MARÍA
DE MAEZTU

Abstract

Intrinsically disordered proteins (IDPs) landed on the molecular biology framework at the turn of the 20th century to challenge the established protein function-structure paradigm. Due to their inherent flexibility and disorder-to-order transitions, IDPs play an important role in the adaptive regulation and mediation of biological responses within cells. However, the intrinsic disorder makes IDPs difficult to characterise by experimental techniques, hindering the elucidation of their mechanisms of action in biological functions. Molecular dynamics simulations can capture the conformational ensembles of macromolecules, but several issues need to be when simulating IDPs, such as proper parameterisation to reproduce the intrinsic disorder, improvement of the conformational sampling, or factors related to the cellular environment such as ionic strength, pH, molecular crowding, etc.

With the recent introduction of these proteins in the scientific landscape, this thesis is presented as a contribution to provide further insight into simulations of IDPs, especially on the effect of pH. Due to the high abundance of ionisable amino acids in IDPs, the incorporation of charge-conformation coupling into in-silico modelling is critical. Therefore, the effect of the dynamic change of protonation states depending on the solvent pH to generate conformational ensembles of IDPs is investigated using the constant pH Molecular Dynamics method. During the study, some shortcomings of this method were identified, which led to a detailed assessment of this approach implemented in AMBER. On the other hand, new force fields and water models designed for IDP simulation, as well as coarse-grained models or sampling techniques, are evaluated on the model IDP peptide, histatin-5, with one of the most extensive simulations of this peptide.

Finally, we focus on the IDP α-synuclein (αS), which is implicated in Parkinson's disease through its fibrillation and oligomerisation leading to deposition in Lewis bodies. Using the IDP-specific force field ff14IDPSFF, β-sheet-rich intermediates are detected in a fragment of αS. In addition, we provide a first insight into the effect of pH on αS and β-synuclein, and plan to continue this study in the future, using the knowledge gained in this thesis to unravel the mechanism of fibrillogenesis of these proteins.

Resumen

Las proteínas intrínsecamente desordenadas (IDPs) se popularizaron en el marco de la biología molecular a principios del siglo XX para cuestionar el paradigma de la función-estructura de las proteínas. Debido a su flexibilidad intrínseca y sus transiciones de desorden-a-orden, las IDPs desempeñan un papel clave en la regulación adaptativa y la mediación de respuestas biológicas en las células. No obstante, el desorden intrínseco provoca que las IDPs sean difíciles de caracterizar mediante técnicas experimentales, lo que dificulta la elucidación de sus mecanismos de acción en las funciones biológicas. Las simulaciones de Dinámica Molecular pueden captar los conjuntos conformacionales de las macromoléculas, sin embargo, en las simulaciones de IDPs es necesario abordar previamente varias cuestiones, como una parametrización adecuada para reproducir el desorden intrínseco, mejorar el muestreo conformacional o algunos factores del entorno celular tal como la fuerza iónica, el pH, el *molecular crowding*, etc.

Con la reciente introducción de estas proteínas en el panorama científico, esta tesis se presenta como una contribución para proporcionar una mayor comprensión de las simulaciones de IDPs, especialmente sobre el efecto del pH. Debido a la gran abundancia de aminoácidos ionizables en las IDPs, la incorporación del acoplamiento carga-conformación en las simulaciones es fundamental. Por lo tanto, se investiga el efecto del cambio dinámico de los estados de protonación en función del pH sobre la generación de conjuntos conformacionales de IDPs utilizando el método de Dinámica Molecular a pH constante. Durante el estudio, se identificaron algunas deficiencias del método, lo que nos impulsó a realizar una evaluación en profundidad del mismo. Por otro lado, también ponemos a prueba nuevos campos de fuerza o modelos de agua diseñados para la simulación de IDP, así como modelos de grano grueso o técnicas de muestreo, en el péptido modelo IDP, histatin-5, con una de las simulaciones más exhaustivas del péptido.

Por último, nos centramos en la IDP α-sinucleína (αS), implicada en la enfermedad de Parkinson a través de su fibrilación y oligomerización hasta depositarse en los cuerpos de Lewis. Utilizando el campo de fuerza específico para IDP ff14IDPSFF, se detectan intermedios ricos en láminas β en un fragmento de αS. Adicionalmente, proporcionamos unas pinceladas del efecto del pH sobre αS y β-sinucleína, con la intención de continuar este estudio en el futuro, utilizando los conocimientos adquiridos en esta tesis para entender el mecanismo de fibrilogénesis de estas proteínas.

Resum

Les proteïnes intrínsecament desordenades (IDPs) es van popularitzar en el marc de la biologia molecular a principis del segle XX per a qüestionar el paradigma de funció-estructura de les proteïnes. A causa de la seva flexibilitat intrínseca i les transicions de desordre-a-ordre, les IDPs exerceixen un paper clau en la regulació adaptativa i la mediació de respostes biològiques en les cèl·lules. El desordre intrínsec provoca que les IDPs siguin difícils de caracteritzar mitjançant tècniques experimentals, la qual cosa dificulta l'elucidació dels seus mecanismes d'acció en les funcions biològiques. Les simulacions de Dinàmica Molecular poden captar els conjunts de conformacions de les macromolècules, però en les simulacions de IDPs és necessari abordar prèviament diverses qüestions, com una parametrització adequada per a reproduir el desordre intrínsec, millorar el mostreig conformacional o factors relacionats amb l'entorn cel·lular com la força iònica, el pH, el *molecular crowding*, etc.

Amb la recent introducció d'aquestes proteïnes en el panorama científic, aquesta tesi es presenta com una contribució per a proporcionar una major comprensió de les simulacions de IDPs, especialment sobre l'efecte del pH. A causa de la gran abundància d'aminoàcids ionitzables en les IDPs, la incorporació de l'acoblament carrega-conformació en les simulacions és fonamental. Per tant, investiguem l'efecte del canvi dinàmic dels estats de protonació en funció del pH sobre la generació de conjunts de conformacions de les IDPs utilitzant el mètode de Dinàmica Molecular a pH constant. Durant l'estudi, es van identificar algunes deficiències en el mètode, la qual cosa ens va impulsar a realitzar una avaluació en profunditat d'aquest. D'altra banda, també posem a prova nous camps de força o models d'aigua dissenyats per a la simulació de IDP, així com models de gra gruixut o tècniques de mostreig, en la IDP model, histatin-5, amb una de les simulacions més exhaustives del pèptid.

Finalment, ens centrem en la IDP α-sinucleïna (αS), implicada en la malaltia de Parkinson a través de la fibril·lació i oligomerització que condueixen al seu dipòsit en els cossos de Lewis. Utilitzant el camp de força específic per IDPs, ff14IDPSFF, es detecten intermedis rics en fulles β en un fragment de αS. A més a més, proporcionem una primera pinzellada de l'efecte del pH sobre αS i β-sinucleïna, i planegem continuar aquest estudi en el futur, utilitzant els coneixements adquirits en aquesta tesi per a desentranyar el mecanisme de fibril·logènesi d'aquestes proteïnes.

Agraïments

A la primera persona que cal agrair en majúscules aquesta tesi és el Dr. Jaime Rubio, qui s'ha convertit en un mentor i referent acadèmic durant aquests últims vuit anys. El destí va voler que m'incorporés en el seu grup de recerca per fer unes pràctiques a l'estiu de 2016 i, afortunadament, gràcies a ell vaig poder emprendre aquesta trajectòria acadèmica en el món de la modelització molecular. A part del seu admirable i ampli coneixement, el que el fa especial sota el meu parer és que continua a primera línia en la recerca, implicant-se com qualsevol altre en les tasques d'investigació. La meva gratitud va més enllà de la ciència, ja que també li agrairé la humanitat, les rialles i les preocupacions que hem compartit tot aquest temps, mostrant-se sempre disposat a ajudar-te en el que faci falta. Per mi ha estat un plaer i una sort tenir-lo com a mentor. La seva filosofia de vida és tot un exemple, de la qual vull creure que m'enduc un tros. Del que estic segur és que sempre recordaré la seva referència que fa al acabar el dia, treta de El Gran Wyoming: *Mañana más, pero no mejor, porque es imposible*.

Qui també ha fet possible aquesta tesi han estat el Dr. Sergio Madurga i el Dr. Francesc Mas, sempre disposats a reunir-nos per debatre sobre la recerca feta durant aquest temps. Sense ells, no hagués conegut tot el món de las proteïnes intrínsecament desordenades o les simulacions a pH constant que resulta summament fascinant. Els hi agrairé sempre el bon ambient, la seva dedicació i totes les possibilites brindades en aquests últims anys. Ha estat un immens plaer posar en comú les diferents perspectives en la investigació, així com aprendre d'ells com a magnífics mentors acadèmics.

Per altra banda, cal mencionar a les companyes i companys del grup de recerca, especialment en aquesta última etapa. Gràcies pel bon rotllo, els *coffee break* i les cerveses en el bar, Guillem, Natàlia i Núria, així com les xerrades i la bona avinença que hem compartit durant tot aquest temps. Dins del departament no hagués estat el mateix sense la Bea o en *Cabra*, que m'han acompanyat en les frustracions i en els èxits acadèmics amb bromes i somriures ja des d'abans de la tesi, i també més enllà durant les barbacoes i moltes altres quedades amb els ara ja *exiliats*, l'Edgar i la Neus. També vull esmentar a la Susanna, amb qui vaig tenir la sort de trobar-me al primer any de carrera, i des de llavors hem compartit una infinitat de moments. Vull posar en especial valor la seva força de convicció per no desistir en allò que creu, a contracorrent si fa falta, doncs per a mi és tot un exemple com a persona i amiga. Però res hagués estat igual sense la

Laia, amb qui he viscut la experiència de la tesi tot i trobar-nos a 1,800 km de distància a través de videotrucades i missatges de veu, però també amb escapades i concerts que acabaven en catarsi i llàgrimes. Gràcies per les paraules i cançons compartides, per convertir-te en confident i per ser-hi sempre allà, gràcies perquè *las palabras que no existen nos pueden salvar*.

També vull mencionar a la Maria Bel, l'Edgar, la Cris, en Pepe i en Pau, per haver-me acompanyat des d'aquell pis al barri de Collblanc fins el dia d'avui i convertir-se en una xarxa d'amics on desinhibir-se i passar una bona estona, ja sigui amb sopars, (molts) jocs de taules o *escape rooms*, i les escapades a Mallorca, és clar! Per altra banda, també l'Anna, la Laia F., la Maria, l'Aleix i, des de fa poc, la Lucía, qui han patit de primera mà tot el procés de la tesi durant la nostra convivència. Els hi vull agrair especialment el fet de donar-me un espai on riure, desfogar-me i, alhora, compartir les històries, bromes i anècdotes en el dia a dia. A tots ells, gràcies per ser-hi allà, tant en els petits instants com en els moments més difícils, perquè tots vosaltres també formeu part d'aquesta tesi.

Li dec una gratitud absoluta a la meva família pel seu suport i afecte. *Mama*, gràcies per ensenyar-me què és l'amor incondicional i sincer. I, *papa*, pel teu exemple i esforç que m'ha permès ser l'home que soc avui. A tots dos, per posar-me sempre per davant i estar sempre al meu costat. També vull especialment agrair a la meva àvia les trucades setmanals que acaben en rondalles i dites populars que m'han fet sentir com si fos a casa. I al meu avi, que sempre s'ha preocupat perquè no em faltés res. A tota la família, ja que em deixo molts per esmentar, gràcies per tots els consells, celebracions i ànims que ens hem compartit al llarg de tots aquests anys.

Per últim, a l'Edu per iniciar un camí junts i sempre recolzar-me en els moments més difícils d'aquesta aventura. Gràcies per tot, *carinyo*. Podria dir-te moltes coses, però prefereixo citar un fragment que m'ha acompanyat íntimament en aquests últims mesos tal i com ho has fet tu en aquests últims tres anys i mig.

I en un precís instant, sentir-la tota vibrar,
la vida que hem tingut, i també la que vindrà.
"Un Lloc", de Guillem Roma

Aquesta tesi està dedicada a la meva àvia Mercè
i a la memòria del meu millor amic de la infància
David Castro Rodà

Es tornarà ocell per un dia,
i d'entre les cendres podrà volar.
"El Vol de l'Home Ocell", Sangtraït

Table of Contents

Abbreviations

αS	α-Synuclein
βS	β-Synuclein
CG	Coarse-Grained
CpHMD	Constant pH Molecular Dynamics
CS	Chemical Shifts
γS	γ-Synuclein
DSSP	Dictionary of Secondary Structure of Proteins
DBI	Davies-Bouldin Index
E-αSNAC	Extended α-Synuclein NAC fragment
fpSS	fractions of Secondary Structure propensity
GB	Generalised-Born
HH	Henderson-Hasselbalch
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
LF	Langmuir-Freundlich
NMR	Nuclear Magnetic Resonance
MC	Monte Carlo
MD	Molecular Dynamics
MM	Molecular Mechanics
NAC	Non-Amyloid-β Component
PB	Poisson-Boltzmann
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PD	Parkinson's Disease
PME	Particle Mesh Ewald
psF	pseudo-statistics F
REMD	Replica Exchange Molecular Dynamics
RDF	Radial Distribution Function
R_g	Radius of Gyration
RMSD	Root Mean Square Deviation

RMSF	Root Mean Square Fluctuation
SAXS	Small-Angle X-Ray Scattering
SD	Steepest Descent
vdW	van der Waals

Chapter 1

Introduction

1.1. Motivation and Outline of the Thesis

This thesis was born from the motivation to understand the behaviour of molecular structures that are essential for the evolution of life itself: peptides and proteins. However, we must go back to my childhood when, in a completely naive and ignorant way, I was already curious about one of the simplest forms of life: the cell. Fortunately, today, after many years and efforts in my academic training, I have managed to give this interest a satisfactory place in my daily life by means of computational methods. I remember that at the beginning of the Chemistry degree, around 2016, I was fascinated by the applications of molecular modelling in the early stages of drug design, and thanks to this first contact with computational biochemistry, I jumped into the adventure of the academic research. After my master thesis on the selective inhibition of the Bcl-2 family proteins, my intention was to continue this research as a PhD student. However, life changes and you must adapt, and sometimes it changes for the better, because I was offered the opportunity to learn about an extremely interesting group of proteins *recently* discovered in the scientific landscape: the intrinsically disordered proteins (IDPs). Due to the high flexibility and disorder-to-order transitions between several conformational ensembles, IDPs pose a challenge for molecular modelling. Many novel approaches are now being developed to capture the properties of these proteins from an in-silico perspective. Among them, the dynamic regulation of the protonation states of the ionisable amino acids depending on the solvent pH, which is tightly coupled to the conformations of IDPs, is increasingly being introduced into the computational simulations. Unfortunately, there is still a lack of work in the literature on the approaches and solutions that have been developed for the successful modelling of these biochemical systems. Therefore, this thesis is presented as a modest contribution in the framework of pH-responsive IDPs simulations with the main motivation to enrich the understanding of IDP modelling and the strong charge-conformational coupling shown in these proteins.

This doctoral thesis is divided into two main topics: the assessment of the Constant pH Molecular Dynamics (CpHMD) method implemented in AMBER to model the charge regulation of ionisable amino acids, and the evaluation of the simulation methods, force fields and water models on the histatin-5 peptide. Based on the conclusions drawn from these studies, we initiate an investigation into the fibrillogenesis of the synuclein protein family, in particular the α-synuclein protein, which is involved in neurodegenerative diseases.

For the first topic of the thesis, Chapter 3 introduces the first contact with the CpHMD method using a test system: the polyaspartic acid decapeptide. This chapter, which initially aims to provide insight into the charge-conformation coupling over a range of pH values and the potential applicability of complexation isotherms for pKa prediction, also evaluates the performance of CpHMD in implicit and explicit solvation models. The conclusions drawn from the evaluation of the method raise some concerns about the conformational sampling of the polyaspartic acid, compelling us to continue with the testing of the CpHMD method implemented in AMBER. Thus, in Chapter 4, a series of tripeptides are studied in detail with simulations at constant pH in implicit solvent conditions in order to analyse the impact of the method in the conformational space by means of the Ramachandran maps. The shortcomings of the CpHMD implemented in AMBER are revealed, and the major reason for the failure to reproduce the conformational space is due to a poor definition of the partial charges of the backbone atoms, which remain fixed during the simulation. Motivated by the importance of the charge regulation in simulations at constant pH, Chapter 5 provides an overview of the CpHMD capabilities with the tripeptides in explicit water molecules, and, more importantly, assesses the extent of the limitations identified in the previous chapter for oligopeptides with a small number of ionisable amino acids. The results show that oligopeptide simulations with a few ionisable amino acids at a considerable distance show a good performance when compared to the conventional simulations.

On the other hand, the second topic of this thesis consists of two chapters on the relevance of using IDP-specific force fields to model these proteins. In Chapter 6 we use the IDP model histatin-5 to perform an exhaustive conformational sampling through a battery of simulations using various simulation methods, force fields, water models or sampling strategies. In addition, Chapter 7 evaluates the ability of the ff14SB and ff14IDPSFF force fields on the α-synuclein protein to capture intermediate conformations

prone to fibrillation, a potential cause of the onset of Parkinson's disease. At the end of the thesis, Chapter 8 brings together the knowledge gained in previous chapters on IDP modelling at constant pH to provide some insights into charge-structure coupling of the α- and β-synuclein proteins. This future study will involve an ambitious examination of the conformational ensembles of these two proteins under different pH conditions, but unfortunately it will not be completed within the time frame of the thesis.

At the beginning of the thesis, Chapter 2 contains a description of the theoretical background on which all the molecular modelling, simulation methods and analysis tools used in the chapters are based. The final conclusions on the two topics covered in this thesis are presented in Chapter 9.

1.2. A Brief Introduction to the Intrinsically Disordered Proteins

For more than a hundred years the protein function-structure paradigm had been rooted in the scientific community. For proteins to fulfil their biological function, it was thought that a well-defined 3D protein structure was imperative to enable the interaction between the molecular partners involved. The "lock-and-key" model, in which the receptor and the substrate require a specific configuration to fit together, or the "induced fit" model, in which a flexible receptor has an initial conformation in the active site that eventually switches and changes conformation upon the interaction with the substrate, were very popular for understanding protein functionality. However, a number of papers in the mid-20th century began to contradict and challenge the protein structure-function paradigm by pointing out the existence of disordered and flexible proteins that were capable of being biologically active[1]. In particular, with the advance of genomics in the protein identification, many protein sequences that were not expected to be folded were discovered, thus introducing the intrinsically disordered proteins at the turn of the century. These particular proteins have been assigned many names, such as foldable, floppy, mobile, chameleon, dancing proteins, partially folded, protein clouds, natively disordered, etc., to denote their unusual behaviour, and over time they have become more and more prevalent within the study of the proteome[2]. Fortunately, there is now a consensus on the definition of IDPs, which can be summarised as flexible proteins that exist as dynamic ensembles of interconverting conformations, similar to clouds of proteins, and that do not show any long-term stability for secondary or tertiary structures.

The amino acid composition is fundamental to understand the properties of IDPs. Several papers have compiled the frequency of occurrence of the amino acids in the composition of IDPs, finally classifying them into "order-promoting" amino acids, such as Ile, Leu, Val, Trp, Tyr, Phe, Cys and Asn, and "disorder-promoting" amino acids, such as Ala, Arg, Gly, Gln, Ser, Glu, Lys, and Pro[3,4]. The latter group is abundant in the sequences of IDPs. In addition, these amino acids are associated with low hydrophobicity and high net charge on proteins, which are critical for (i) preventing proteins from compacting in polar solvents and (ii) generating extended conformations due to the repulsion by means of electrostatic interactions. For this reason, IDPs are termed “intrinsically" disordered because the amino acid sequence and properties inherently confers this *protein disorder*.

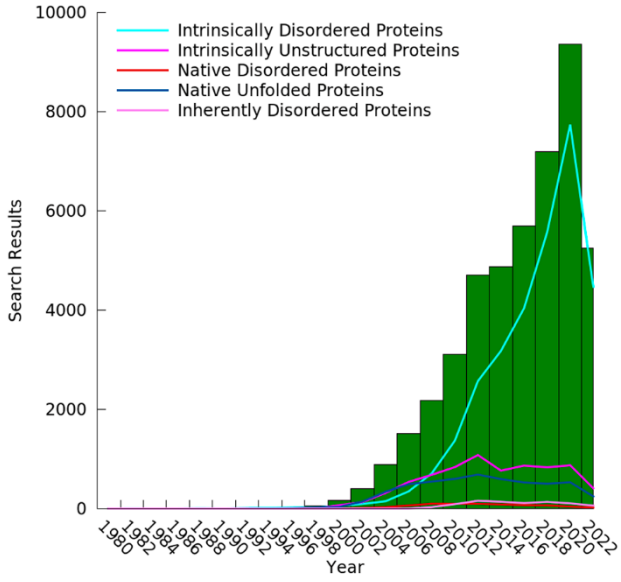


Figure 1. Number of results in the Google Scholar search tool for *intrinsically disordered proteins* or similar terms (including any combination of *intrinsically/natively/inherently* + *disordered/unfolded/unstructured/flexible* + *protein/proteins*) in the green boxes of the histogram. Each box covers the period of the indicated year and the following year, e.g., 1980-1981. The lines indicate the number of search results for the terms in the legend.

This intrinsic disorder allows IDPs to carry out their biological functions, which escape the traditional mechanisms of the globular proteins. In general, IDPs participate in protein-protein interaction (PPI) networks through the following mechanisms: (i) the one-to-many mechanism in which one IDP can bind to many partners, and (ii) the many-to-one mechanism in which many IDPs can bind to the same, usually ordered, partner[5,6]. *Chameleon* behaviour is important in these mechanisms because it dictates that an amino

acid sequence can be conformationally modulated to have a different secondary structure and side chain contribution in order to interact in different ways with the same or a different partner[7,8], thus highlighting the role of the protein disorder in this ability. Furthermore, IDPs typically undergo disorder-to-order transitions towards preferred conformations when they bind to partners according to the principles of the induced fit model. Indeed, some studies suggest that IDPs can adopt a preformed conformational state prior to the binding, which acts as a driving force[9,10]. These bindings processes have high specificity and low affinity, thereby ensuring the reversibility of IDP interactions and enabling them to bind to different partners[11–16]. However, IDPs do not only fulfil their biological functions through these disorder-to-order transitions and folding, but also through their flexibility, pliability, and plasticity in dynamic complexes[17].

Therefore, IDPs are promiscuous binders that play an important part in the adaptive regulation and mediation of the biological responses of the cells due to their high flexibility. Nowadays, they are recognised in a broad spectrum of functions and are classified as chaperones, effectors, entropic chains, scavengers, display sites and assemblers[12,17,18]. Depending on their function, they can participate in many processes including transcriptional and translational regulation, cellular signalling, small molecule storage, protein phosphorylation, self-assembly regulation, and molecular recognition[13,18]. These biological processes are essential for the proper cell cycle as they are responsible for cell differentiation, cell-cell communication, cell cycle progression, apoptosis, and so on.[19–22]. Furthermore, in recent years, bioinformatic studies of the genome sequences have reported that IDPs are very abundant in eukaryotes, exhibiting long intrinsically disordered regions (IDRs) in more than half of the functional proteins, especially for the proteins involved in signalling processes (~70%)[23–25]. In the human proteome, 32% of the proteins are identified as IDPs (those ones with more than 30% of disordered residues) and, in fact, a 34% of biological functions reported in the Uniprot/Swissprot database are related to IDPs[26–28]. In addition, they have also been identified in several human diseases[29–31], such as cancer (AFP, p53, and BRCA-1)[25,32], neurodegenerative diseases[33] (Alzheimer's and Parkinson's diseases, involving amyloid-β and α-synuclein, respectively), cardiovascular diseases[34] (hirudin, thrombin), diabetes[35] (amylin) or in pathogenic viruses and microbes[36,37].

More and more of studies on IDPs are being published due to their fundamental activity in several biological processes or their therapeutic role in the treatment of human

diseases. In this thesis, two IDPs have been examined. In particular, the human salivary peptide histatin-5, which is generally used as a model IDP for the assessment of force fields and simulation methods[38,39], and the α-synuclein protein, which is implicated in neurodegenerative diseases and other synucleinopathies[40].

1.2.1. Synuclein Protein Family

Synucleins are small, highly conserved, intrinsically disordered proteins found primarily in the neurons of vertebrate animals. The number of proteins belonging to this family varies between species, but the α-, β-, and γ-synuclein (αS, βS, and γS, respectively) are present in birds and mammals[41–43]. αS has been the most investigated in recent years because of its involvement in Parkinson's disease (PD), but the three synuclein proteins have been implicated in neurodegenerative diseases, synucleopathies and/or cancer[44]. Unfortunately, there is a great deal of uncertainty about the mechanism by which these proteins exert either the biological functions or the pathological activity, so further studies on this family of proteins are needed.

αS (140 aa), βS (132 aa) and γS (127 aa) have a similar amino acid sequence, in particular αS and βS which share 60% sequence similarity. The structure of synucleins is divided into three domains: (i) the N-terminal domain, an amphipathic region that interacts with lipid bilayers and is therefore involved in membrane association, (ii) the non-amyloid-beta component (NAC) domain, characterised by abundant hydrophobic amino acids, and (iii) the C-terminal domain, a region rich in negatively charged acidic and proline amino acids. The N-terminal domain is highly conserved among the three synucleins and is consists of seven imperfect sequences of 11 amino acid repeats that adopt a helix structure upon binding to vesicles or micelles[51,52]. The NAC domain plays a pivotal role in the formation of αS fibril aggregates through a hydrophobic effect and hydrogen bonding within a β-sheet-rich structure[53]. In fact, Giasson et al.[54] reported that amino acids 71-82 are critical for the formation of αS fibrils. In contrast, there is a deletion of 11 amino acids in the NAC domain of βS, whereas this region is not highly conserved in γS compared to αS. Given the importance of this domain for aggregation and the modifications present in βS and γS, the variation in the fibrillation behaviour present in αS with respect to the βS and γS homologs is reasonable[55]. Finally, the C-terminal domain is distinguished by an acidic tail that regulates the solubility of the proteins according to the charge and length[56]. In γS, the amino acid sequence does not have the two repeats

found in αS and βS due to the shorter number of amino acids in the C-domain. All these synucleins are disordered under physiological conditions.

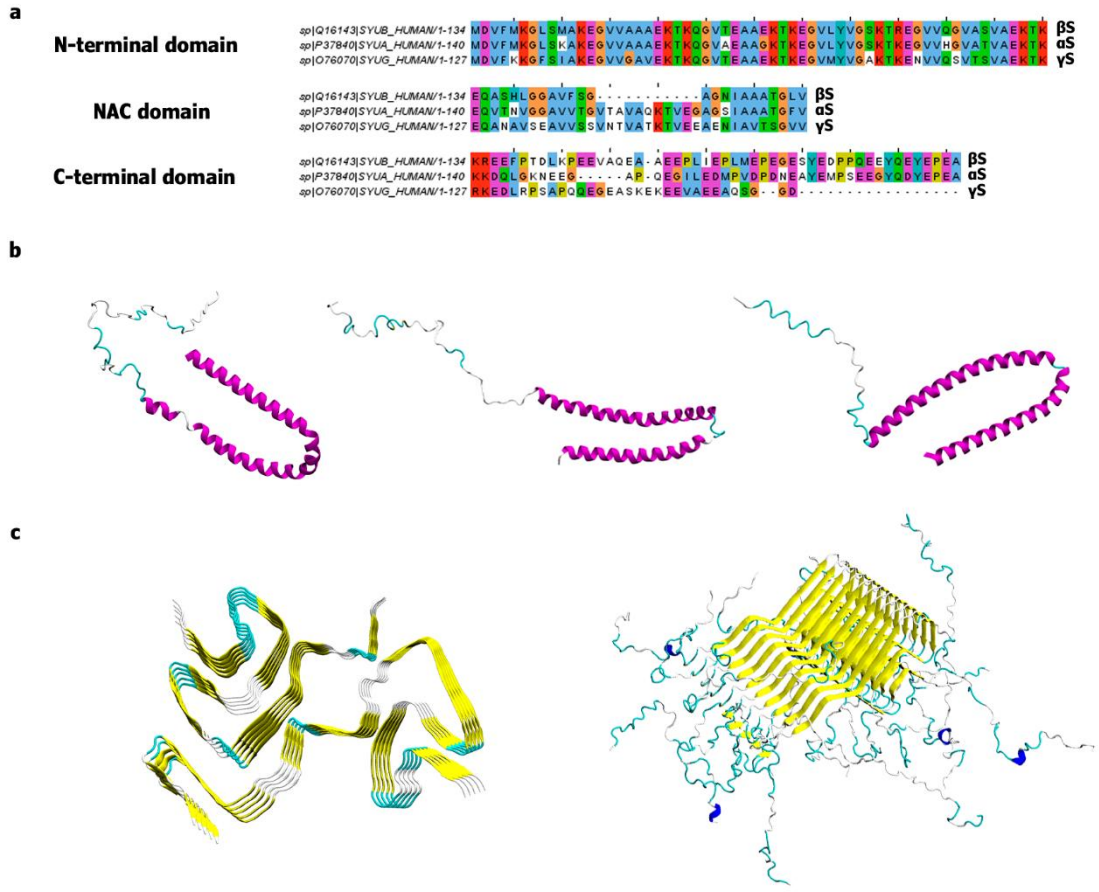


Figure 2. (a) Alignment of the amino acid sequences of human αS, βS and γS using the ClustalO tool[45] with the program Jalview[46]. (b) Structure of human αS (middle, PDB-code: 1XQ8[47]), βS (left) and γS (right) monomers. The structures of βS (UniProt-code: Q16143) and γS (UniProt-code: Q6FHG5) are predicted by the AlphaFold Monomer v2.0 pipeline[48]. (c) Structures of αS fibrils in a multiple system atrophy type II-2 (left, PDB-code: 6XYQ[49]) and decameric form (right, PDB-code: 2N0A[50]).

The biological roles of the synucleins are not fully understood, but some functions in which αS is involved have been discovered to date, such as promoting SNARE complex assembly, synaptic vesicle regulation, and membrane remodelling[57,58]. On the other hand, βS and γS play a modulating role in the αS-synaptic vesicle binding through the formation of oligomers of these two proteins with αS (since βS and γS have a lower binding affinity to synaptic vesicles in comparison to αS)[59]. In fact, some studies have demonstrated the neuroprotective nature of βS in synucleinopathies caused by αS aggregation[60,61]. It has also been found that βS is an important biomarker for early

Alzheimer's disease[62]. In addition, high levels of γS and low levels of αS in plasma have recently been associated with autism spectrum disorder in children[63]. Abnormal expression of γ-synuclein in stage I and II has also been observed in several human cancers and may therefore be useful as a biomarker in the detection of tumourigenesis[64]. Other biological functions involving αS, βS and γS can be found in the literature[65].

Some neurodegenerative diseases are caused by the aggregation of proteins into amyloid-like fibrils. This category includes αS, whose aggregation in Lewy bodies and neuritis is associated with the onset of PD[66,67]. Under physiological conditions, αS is present in the monomeric or oligomeric form, the latter facilitated by prior binding to lipid membranes[68,69]. Some mutations (A53T[70], A30P[71], E46K[72], H50Q[73], among others) and gene multiplication[74,75] or triplication[76] have been recognised as potential triggers of αS aggregation and toxicity. Above a certain concentration of αS deposition, there is an intrinsic toxic gain-of-function in the nature of the protein, ultimately causing in the origin of the synucleopathies. However, there is also the hypothesis that these diseases are driven by the loss of αS function when it is sequestered in the aggregations. Interestingly, fibrillogenesis is not observed in βS and γS under physiological conditions. Instead, γS forms fibrils only under aggregation-promoting conditions with a larger lag phase compared to αS, whereas βS normally acts as an inhibitor or retardant of the αS aggregate formation[77]. However, βS is not exempt from being implicated in neurodegenerative diseases, and certain mutations (P123H and V70M) have been reported to be related to dementia through Lewy bodies[78]. In addition, γS inclusions in motor neurons are associated with the amyotrophic lateral sclerosis disease[79]. Therefore, all these diseases caused by the aggregation of any of the three proteins of the synuclein family can be classified as synucleinopathies[40,80].

1.3. The Role of Solvent pH in the Charge Regulation and the Charge-Conformation Coupling

Among the various environmental factors that can potentially influence IDPs, some studies have reported that αS and βS fibril formation is pH dependent. At mildly acidic pH, αS fibrillates more rapidly while βS gains the capacity to form fibrils, in contrast to to physiological pH conditions[81]. Given the high presence of ionisable acidic amino acids of the C-terminal domain, it is reasonable to expect that pH may modulate the conformation and hence the fibrillation propensity of the synucleins. Indeed, Santos J. et al.[82] suggest that aggregation in amyloid-like fibrils is pH-dependent through modulation

of the hydrophobic effect, electrostatic interactions, and the degree of protonation of ionisable amino acids. On the other hand, Pálmadóttir et al.[83] show that charge regulation during the αS fibrillation leads to a significant increase in pH, shifts in the pKa of acidic amino acids in the C-terminal domain and in the proton binding capacitance. Thus, pH-dependent charge regulation plays an important part in the mechanism of the fibril formation.

Charge regulation is defined as the ability of a macromolecule to modulate its ionisation state when subjected to external physicochemical perturbations in order to adapt to a new environment. In proteins or peptides, charge regulation generally occurs through the migration of protons from or to ionisable amino acids, thereby affecting the acid-base equilibrium, or ion binding. The mechanism of charge regulation was originally described by Linderstrom-Lang in 1920[84], and later Kirkwood and Shumaker[85] demonstrated the correlation between the charge distribution in the intermolecular interactions of two proteins using the perturbation theory of statistical mechanics, which was confirmed by light scattering[86]. Since then, many papers have been published demonstrating the effect of charge regulation on protein-protein[87–89], protein-polyelectrolyte[90–93] or protein-surface interactions[94–96], ligand-receptor binding[89,97], protein folding[98,99], and many other processes[100–103].

By means of a statistical mechanics description of the charge, z, the effect of charge regulation can be quantified by the binding capacitance, C. The capacitance is then merely the measure of the variation of the charge in response to an external electric potential, φ, as defined in Eq. 1, where $\beta = 1/k_B T$ is the inverse thermal energy and e is the electron unit charge.

$$C = \langle z^2 \rangle - \langle z \rangle^2 = -\frac{\partial \langle z \rangle}{\beta e \partial \varphi} \qquad \textbf{Eq. 1.1}$$

Interestingly, the capacitance is strongly dependent on the solvent pH and shows a charge response function to small perturbations of the solvent pH in Eq. 2. for a macromolecule with *N* ionisable amino acids and a total charge number, $Z = \sum_{i=1}^{N} z_i$. For a protein, the capacitance can be obtained from the slope of the experimental titration curves, if possible. Otherwise, atomistic simulations at constant pH are a suitable option for determining the binding capacitance, from which the $\langle Z^2 \rangle$ and $\langle Z \rangle^2$ can be calculated directly.

$$C = \langle Z^2 \rangle - \langle Z \rangle^2 = -\frac{1}{ln10}\frac{\partial \langle Z \rangle}{\partial pH} \qquad \textbf{Eq. 1.2}$$

Charge regulation is essential for some protein mechanisms and interactions, as we have already mentioned, but it is likely to be even more important in IDPs due to the inherent flexibility of ionisable amino acids. In fact, the electrostatic interactions between ionisable amino acids cause the macromolecules to modulate their conformation in order to minimise the electrostatic repulsion or increase the electrostatic attraction, and simultaneously the change in the structure of the macromolecule also affects the interactions between the ionisable amino acids, thus causing potential changes in the ionisation states. It is therefore clear that charge and the conformation of a protein are tightly coupled. Furthermore, charge fluctuations also come into play in this charge-conformation coupling since the ionisable states do not remain fixed over time, but rather vary within a probability distribution. Several simulation and experimental studies have shown that charge fluctuations are a fundamental phenomenon for some protein-protein, protein-ligand or protein-membrane interactions[104].

Consequently, the inclusion of the solvent pH as well as the charge regulation and the fluctuation in molecular modelling is crucial for a correct description of the pH-responsive proteins, especially for flexible or small macromolecules such as polyelectrolytes, peptides and IDPs.

1.4. Bibliography

1. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59 (2001).
2. Uversky, V. N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chem Rev* **114**, 6557–6560 (2014).
3. Williams, R. M. *et al.* The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pacific Symp Biocomput Pacific Symp* **2001**, 89–100 (2000).
4. Campen, A. *et al.* TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept Lett* **15**, 956–963 (2008).
5. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS J* **272**, 5129–5148 (2005).
6. Uversky, V. N. Intrinsic Disorder-based Protein Interactions and their Modulators. *Curr Pharm Des* **19**, 4191–4213 (2013).
7. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
8. Minor, D. L. & Kim, P. S. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730–734 (1996).

9. Fuxreiter, M., Simon, I. N., Friedrich, P. & Tompa, P. Preformed Structural Elements Feature in Partner Recognition by Intrinsically Unstructured Proteins. *J Mol Biol* **338**, 1015–1026 (2004).

10. Choi, U. B., Sanabria, H., Smirnova, T., Bowen, M. E. & Weninger, K. R. Spontaneous Switching among Conformational Ensembles in Intrinsically Disordered Proteins. *Biomolecules* **9**, 114 (2019).

11. Dunker, A. K. *et al.* Intrinsically disordered protein. *J Mol Graph Model* **19**, 26–59 (2001).

12. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527–533 (2002).

13. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol* **6**, 197–208 (2005).

14. Uversky, V. N. & Dunker, A. K. Understanding protein non-folding. *Biochim Biophys Acta* **1804**, 1231–1264 (2010).

15. Zhou, H.-X. Intrinsic disorder: signaling via highly specific but short-lived association. *Trends Biochem Sci* **37**, 43–48 (2012).

16. Mollica, L. *et al.* Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Front Mol Biosci* **3**, 1–18 (2016).

17. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**, 3346–3354 (2005).

18. Handa, T., Kundu, D. & Dubey, V. K. Perspectives on evolutionary and functional importance of intrinsically disordered proteins. *Int J Biol Macromol* **224**, 243–255 (2023).

19. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* **18**, 343–384 (2005).

20. Garza, A. S., Ahmad, N. & Kumar, R. Role of intrinsically disordered protein regions/domains in transcriptional regulation. *Life Sci* **84**, 189–193 (2008).

21. Zhou, J., Zhao, S. & Dunker, A. K. Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. *J Mol Biol* **430**, 2342–2359 (2018).

22. Ferreon, C., Chris, A., Ferreon, M., Trivedi, R. & Nagarajaram, H. A. Intrinsically Disordered Proteins: An Overview. *Int J Mol Sci* **23**, 14050 (2022).

23. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genom Inform* **11**, 161–71 (2000).

24. Oldfield, C. J. *et al.* Comparing and combining predictors of mostly disordered proteins. *Biochemistry* **44**, 1989–2000 (2005).

25. Iakoucheva, L. M. *et al.* Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *J Mol Biol* **323**, 573–584 (2002).

26. Dunker, A. K. *et al.* The unfoldomics decade: An update on intrinsically disordered proteins. *BMC Genom* **9**, 1–26 (2008).

27. Colak, R., Kim, T., Michaut, M., Sun, M. & Irimia, M. Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLoS Comput Biol* **9**, 1003030 (2013).

28. Deiana, A., Forcelloni, S., Porrello, A. & Giansanti, A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One* **14**, e0217889 (2019).

29. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annu Rev Biophys* **37**, 215–246 (2008).

30. Uversky, V. N. *et al.* Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. *BMC Genom* **10**, 1–17 (2009).

31. Uversky, V. N. *et al.* Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases. *Chem Rev* **114**, 6844–6879 (2014).

32. Santofimia-Castaño, P. *et al.* Targeting intrinsically disordered proteins involved in cancer. *Cell Mol Life Sci* **77**, 1695–1707 (2020).

33. Coskuner-Weber, O., Mirzanli, O. & Uversky, V. N. Intrinsically disordered proteins and proteins with intrinsically disordered regions in neurodegenerative diseases. *Biophys Rev* **14**, 679–707 (2022).

34. Cheng, Y., Legall, T., Oldfield, C. J., Dunker, A. K. & Uversky, V. N. Abundance of Intrinsic Disorder in Protein Associated with Cardiovascular Disease. *Biochemistry* **45**, 10448–10460 (2006).

35. Du, Z. & Uversky, V. N. A Comprehensive Survey of the Roles of Highly Disordered Proteins in Type 2 Diabetes. *Int J Mol Sci* **18**, 2010 (2017).

36. Goh, G. K. M., Dunker, A. K. & Uversky, V. N. A comparative analysis of viral matrix proteins using disorder predictors. *Virol J* **5**, 1–10 (2008).

37. Xue, B. *et al.* Viral Disorder or Disordered Viruses: Do Viral Proteins Possess Unique Features? *Protein Pept Lett* **17**, 932–951 (2010).

38. Henriques, J. O., Cragnell, C. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput* **11**, 3420–3431 (2015).

39. Cragnell, C., Rieloff, E. & Skepö, M. Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions. *J Mol Biol* **430**, 2478–2492 (2018).

40. Goedert, M., Jakes, R. & Spillantini, M. G. The Synucleinopathies: Twenty Years On. *J Parkinsons Dis* **7**, 51–69 (2017).

41. Jakes, R., Spillantini, M. G. & Goedert, M. Identification of two distinct synucleins from human brain. *FEBS Lett* **345**, 27–32 (1994).

42. Lavedan, C. *et al.* Identification, localization and characterization of the human γ-synuclein gene. *Hum Genet* **103**, 106–112 (1998).

43. George, J. M. The synucleins. *Genome Biol* **3**, 1–6 (2002).

44. Lavedan, C. The Synuclein Family. *Genome Res* **8**, 871–880 (1998).

45. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276–W279 (2022).

46. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

47. Ulmer, T. S., Bax, A., Cole, N. B. & Nussbaum, R. L. Structure and dynamics of micelle-bound human α-synuclein. *J Biol Chem* **280**, 9595–9603 (2005).

48. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439–D444 (2022).

49. Schweighauser, M. *et al.* Structures of α-synuclein filaments from multiple system atrophy. *Nature* **585**, 464–469 (2020).

50. Tuttle, M. D. *et al.* Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat Struct Mol Biol* **23**, 409–415 (2016).

51. Fusco, G. *et al.* Direct observation of the three regions in α-synuclein that determine its membrane-bound behaviour. *Nat Commun* **5**, 1–8 (2014).

52. Bussell, R. & Eliezer, D. A Structural and Functional Role for 11-mer Repeats in α-Synuclein and Other Exchangeable Lipid Binding Proteins. *J Mol Biol* **329**, 763–778 (2003).

53. Kim, T. E. *et al.* Excess membrane binding of monomeric alpha-, beta- and gamma-synuclein is invariably associated with inclusion formation and toxicity. *Hum Mol Genet* **30**, 2332–2346 (2021).

54. Giasson, B. I., Murray, I. V. J., Trojanowski, J. Q. & Lee, V. M. Y. A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of α-Synuclein Is Essential for Filament Assembly. *J Bio Chem* **276**, 2380–2386 (2001).

55. Ducas, V. C. & Rhoades, E. Investigation of Intramolecular Dynamics and Conformations of α-, β- and γ-Synuclein. *PLoS One* **9**, e86983 (2014).

56. Kim, T. D., Paik, S. R. & Yang, C. H. Structural and Functional Implications of C-Terminal Regions of α-Synuclein†. *Biochemistry* **41**, 13782–13790 (2002).

57. Bisaglia, M., Mammi, S. & Bubacco, L. Structural insights on physiological functions and pathological effects of ±-synuclein. *FASEB J* **23**, 329–340 (2009).

58. Bendor, J. T., Logan, T. P. & Edwards, R. H. The Function of α-Synuclein. *Neuron* **79**, 1044–1066 (2013).

59. Carnazza, K. E. *et al.* Synaptic vesicle binding of α-synuclein is modulated by β- and γ-synucleins. *Cell Rep* **39**, (2022).

60. Hashimoto, M., Rockenstein, E., Mante, M., Mallory, M. & Masliah, E. Synuclein Inhibits-Synuclein Aggregation: A Possible Role as an Anti-Parkinsonian Factor. *Neuron* **32**, 213–223 (2001).

61. Janowska, M. K., Wu, K.-P. & Baum, J. Unveiling transient protein-protein interactions that modulate inhibition of alpha-synuclein aggregation by beta-synuclein, a pre-synaptic protein that co-localizes with alpha-synuclein. *Sci Rep* **5**, 213–223 (2015).

62. Barba, L. *et al.* Cerebrospinal fluid β-synuclein as a synaptic biomarker for preclinical Alzheimer's disease. *J Neurol Neurosurg Psychiatry* **94**, 83–86 (2023).

63. Al-Mazidi, S. & Al-Ayadhi, L. Y. Plasma Levels of Alpha and Gamma Synucleins in Autism Spectrum Disorder: An Indicator of Severity Highlights of the Study. *Med Princ Pract* **30**, 160–167 (2021).

64. Liu, H. *et al.* Loss of Epigenetic Control of Synuclein-γ Gene as a Molecular Indicator of Metastasis in a Wide Range of Human Cancers. *Cancer Res* **65**, 7635–7643 (2005).

65. Surguchov, A. & Surguchev, A. Synucleins: New Data on Misfolding, Aggregation and Role in Diseases. *Biomedicines* **10**, 3241 (2022).

66. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).

67. Spillantini, M. G., Crowther, R. A., Jakes, R., Hasegawa, M. & Goedert, M. α-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc Natl Acad Sci USA* **95**, 6469 (1998).

68. Bartels, T., Choi, J. G. & Selkoe, D. J. α-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation. *Nature* **477**, 107–110 (2011).

69. Gould, N. *et al.* Evidence of native α-synuclein conformers in the human brain. *J Biol Chem* **289**, 7929–7934 (2014).

70. Conway, K. A. *et al.* Acceleration of oligomerization, not fibrillization, is a shared property of both α-synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy. *Proc Natl Acad Sci* **97**, 571–576 (2000).

71. Lashuel, H. A., Overk, C. R., Oueslati, A. & Masliah, E. The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nat Rev Neurosci* **14**, 38–48 (2013).

72. Zhao, K. *et al.* Parkinson's disease associated mutation E46K of α-synuclein triggers the formation of a distinct fibril structure. *Nat Commun* **11**, (2020).

73. Porcari, R. *et al.* The H50Q Mutation Induces a 10-fold Decrease in the Solubility of α-Synuclein. *J Biol Chem* **290**, 2395–2404 (2015).

74. Chiba-Falek, O. Structural variants in SNCA gene and the implication to synucleinopathies. *Curr Opin Genet Dev* **44**, 110–116 (2017).

75. Chartier-Harlin, M. C. *et al.* α-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167–1169 (2004).

76. Miller, D. W. *et al.* α-Synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology* **62**, 1835–1838 (2004).

77. Biere, A. L. *et al.* Parkinson's Disease-associated α-Synuclein Is More Fibrillogenic than β- and γ-Synuclein and Cannot Cross-seed Its Homologs. *J Biol Chem* **275**, 34574–34579 (2000).

78. Ohtake, H. *et al.* β-Synuclein gene alterations in dementia with Lewy bodies. *Neurology* **63**, 805–811 (2004).

79. Peters, O. M. *et al.* Gamma-synuclein pathology in amyotrophic lateral sclerosis. *Ann Clin Transl Neurol* **2**, 29–37 (2015).

80. Galvin, J. E., Lee, V. M. Y. & Trojanowski, J. Q. Synucleinopathies: Clinical and Pathological Implications. *Arch Neurol* **58**, 186–190 (2001).

81. Moriarty, G. M. *et al.* A pH-dependent switch promotes-synuclein fibril formation via glutamate residues. *J Biol Chem.* **292**, 16368–16379 (2017).

82. Santos, J. *et al.* pH-Dependent Aggregation in Intrinsically Disordered Proteins Is Determined by Charge and Lipophilicity. *Cells* **9**, (2020).

83. Pálmadóttir, T., Malmendal, A., Leiding, T., Lund, M. & Linse, S. Charge Regulation during Amyloid Formation of α-Synuclein. *J Am Chem Soc* **143**, 7777–7791 (2021).

84. Linderstrøm-Lang, K. On the ionization of proteins. *Cr. Trav. Lab. Carlsberg* **15**, 1–29 (1924).

85. Kirkwood, J. G. & Shumaker, J. B. Forces between Protein Molecules in Solution Arising from Fluctuations in Proton Charge and Configuration. *Proc Nat Acad Sci USA* **38**, 863–871 (1952).

86. Timasheff, S. N., Dintzist, H. M., Kirkwood, J. G. & Coleman, B. D. Studies of Molecular Interaction in Isoionic Protein Solutions by Light-Scattering. *Proc Nat Acad Sci USA* **41**, 710–714 (1955).

87. Elcock, A. H. & Mccammon, J. A. Calculation of Weak Protein-Protein Interactions: The pH Dependence of the Second Virial Coefficient. *Biophys J* **80**, 613–625 (2001).

88. Lund, M. & Jönsson, B. On the Charge Regulation of Proteins. *Biochemistry* **44**, 5722–5727 (2005).

89. Aguilar, B., Anandakrishnan, R., Ruscio, J. Z. & Onufriev, A. v. Statistics and Physical Origins of pK and Ionization State Changes upon Protein-Ligand Binding. *Biophys J* **98**, 872–880 (2010).

90. Ullner, M., Jönsson, B. & Widmark, P. O. Conformational properties and apparent dissociation constants of titrating polyelectrolytes: Monte Carlo simulation and scaling arguments. *J Chem Phys* **100**, 3365 (1998).

91. Torres, P. *et al.* Protonation of β-lactoglobulin in the presence of strong polyelectrolyte chains: a study using Monte Carlo simulation. *Colloids Surf B Biointerfaces* **160**, 161–168 (2017).

92. Landsgesell, J. *et al.* Simulations of ionization equilibria in weak polyelectrolyte solutions and gels. *Soft Matter* **15**, 1155–1185 (2019).

93. Blanco, P. M., Madurga, S., Narambuena, C. F., Mas, F. & Garcés, J. L. Role of Charge Regulation and Fluctuations in the Conformational and Mechanical Properties of Weak Flexible Polyelectrolytes. *Polymers* **11**, 1962 (2019).

94. Tsao, H.-K. The Electrostatic Interaction of an Assemblage of Charges with a Charged Surface: The Charge-Regulation Effect. *Langmuir* **16**, 7200–7209 (2000).

95. Hartvig, R. A., Van De Weert, M., Østergaard, J., Jorgensen, L. & Jensen, H. Protein adsorption at charged surfaces: The role of electrostatic interactions and interfacial charge regulation. *Langmuir* **27**, 2634–2643 (2011).

96. Narambuena, C. F., Longo, G. S. & Szleifer, I. Lysozyme adsorption in pH-responsive hydrogel thin-films: the non-trivial role of acid-base equilibrium. *Soft Matter* **11**, 6669 (2015).

97. Cera, E. Di. Stochastic linkage: Effect of random fluctuations on a two-state process. *J Chem Phys* **95**, 5082 (1991).

98. Whitten, S. T., García-Moreno E., B. & Hilser, V. J. Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins. *Proc Nat Acad Sci USA* **102**, 4282–4287 (2005).

99. Di Russo, N. V., Estrin, D. A., Martí, M. A. & Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput Biol* **8**, e1002761 (2012).

100. Gitlin, I., Mayer, M. & Whitesides, G. M. Significance of Charge Regulation in the Analysis of Protein Charge Ladders. *J Phys Chem B* **107**, 1466–1472 (2003).

101. Hartig, S. M., Greene, R. R., Dikov, M. M., Prokop, A. & Davidson, J. M. Multifunctional Nanoparticulate Polyelectrolyte Complexes. *Pharm Res* **24**, 2353–2369 (2007).

102. Popa, I. *et al*. Importance of charge regulation in attractive double-layer forces between dissimilar surfaces. *Phys Rev Lett* **104**, 228301 (2010).

103. Trefalt, G., Behrens, S. H. & Borkovec, M. Charge Regulation in the Electrical Double Layer: Ion Adsorption and Surface Interactions. *Langmuir* **32**, 380–400 (2016).

104. Lund, M. & Jö, B. Charge regulation in biomolecular solution. *Q Rev Biophys* **46**, 265–281 (2013).

Chapter 2

Theoretical Background

This chapter presents the theoretical background on which the research for this thesis was based. For this purpose, the content is divided into four topics: (i) the fundamentals of molecular mechanics and dynamics, (ii) the introduction of the pH effect using the constant pH molecular dynamics method, (iii) some considerations on IDPs in molecular dynamics simulations and (iv) techniques for the analysis of the generated conformational ensembles. In each of these topics, we explain the theory, approaches, and applications of the various methods in a general but comprehensive manner, so that the reader can understand the performance, capabilities and limitations of the simulations. If the reader is interested in a particular topic, we recommend that they refer to the bibliography provided throughout the chapter.

2.1. Fundamentals of Molecular Mechanics and Molecular Dynamics

Molecular mechanics (MM) focuses on modelling three-dimensional molecular structures using potential functions based on the principles of classical mechanics (such as the harmonic oscillator, Lennard Jones, or Coulombic potential). These molecules are formed by atoms, defined by the position of their nucleus according to the Born-Oppenheimer approximation (i.e., the wave function of the electron and the nucleus can be separated), which are treated as spheres of a given radius connected by bonds typically described as harmonic springs. From this model of spheres and bonds, other terms such as bond angles, dihedral or improper dihedrals, all of which are included in the bonded interactions, and the non-bonded interactions, such as electrostatic or van der Waals forces, are derived to define the potential energy surface of the molecular system. All the mathematical functions and their associated constant parameters are collected within the force fields, a concept that was born in the first half of the 20th century from vibrational spectroscopy and later extended by D. H. Andrews in molecular mechanics[1], and constitute the so-called interatomic potential of the system.

Initially, MM was also named as empirical force field method, and even some of its predecessors were popular at the time, such as the Westheimer method, which analysed

the hidden conformations of biphenyls by modelling of shifts of each atom relative to the equilibrium positions including attractive and repulsive non-bonded terms. Other notable work in the development of the force fields was the Hill force field, which included the Lennard-Jones potential 6-12[2], or the work of Dostrovsky, Hughes and Ingold on non-bonded interaction terms for substitution and elimination reactions[3]. In fact, the Hill force field is very similar to modern formulations of force fields, which are based on an expression of simple additive functions to describe intra- and intermolecular interactions:

$$U(r_1, \dots, r_N) = U_{bonds} + U_{angles} + U_{dihedrals} + U_{improper} + U_{Coulomb} + U_{van\ der\ Waals} \quad \textbf{Eq. 2.1}$$

The *potential energy function* includes the bond stretching, angle bending, dihedral or torsional or improper dihedral terms within the bonded interactions. To reproduce the simple vibrations of bond distance and angle, the U_{bonds} and U_{angles} terms are approximated by harmonic oscillators, for which a force constant (k_b, k_θ) and an equilibrium or reference value (r_0 and θ_0) is defined. This model defines the local covalent structure of the molecule since huge energy is required to significantly deform the bond or angle relative to the equilibrium values. Next, the torsion or dihedral angle, that is, the angle of rotation around the longitudinal axis of a chemical bond, is expressed by the $U_{dihedral}$ term and consists of a sum of cosine functions with multiplicity n, phase δ (typically restricted to 0 or 180º), and a torsional energy barrier k_φ. Finally, for complex molecular geometries, improper dihedrals are defined to preserve planar structures from out-of-plane distortions or to avoid mirror images and thus retaining the chirality of a molecule. This $U_{improper}$ term is also defined as a harmonic potential that depends on a force constant k_ω and an equilibrium or reference dihedral ω_0.

$$U(\boldsymbol{r_1}, \dots, \boldsymbol{r_N}) = \sum_{bonds} k_b\left(\boldsymbol{r_{ij}} - r_0\right)^2 + \sum_{angles} k_\theta\left(\boldsymbol{\theta_{ijk}} - \theta_0\right)^2 + \sum_{dihedrals} k_\varphi\left[1 + \cos\left(n\boldsymbol{\varphi_{ijkl}} - \delta\right)\right] + \sum_{impropers} k_\omega\left(\boldsymbol{\omega_{ijkl}} - \omega_0\right)^2 \quad \textbf{Eq. 2.2}$$

Non-bonded interactions are commonly defined by the Coulomb potential for electrostatic interactions and the Lennard-Jones potential for non-polar interactions. The Coulomb potential states that the electrostatic force is directly proportional to the product of the partial charges between pairs of atoms, q_i and q_j, and inversely proportional to the

quadratic distance between them, r_{ij}, and the relative permittivity of the medium, ε_r. The Lennard-Jones potential, on the other hand, is a 12-6 potential composed of a repulsive term $1/{r_{ij}}^{12}$ arising from the overlapping of electronic orbitals according to the Pauli repulsion, and an attractive term $1/{r_{ij}}^{6}$ derived from the dispersion forces or van der Waals interactions. The sum of the attractive and repulsive forces between pairs of atoms gives a potential model with an energy well defined by a depth $\epsilon_{min,ij}$ at a distance $R_{min,ij}$.

$$U(\boldsymbol{r_1}, \dots, \boldsymbol{r_N}) = \sum_{Coulomb} \frac{q_i q_j}{\varepsilon_r \boldsymbol{r_{ij}}} + \sum_{LJ} \epsilon_{min,ij} \left[\left(\frac{R_{min,ij}}{\boldsymbol{r_{ij}}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{\boldsymbol{r_{ij}}} \right)^{6} \right] \qquad \textbf{Eq. 2.3}$$

The applications of 3D models based on molecular mechanics are many and varied, highlighting Molecular Dynamics (MD) or Monte Carlo (MC) simulations, but also including other notable applications such as energy minimisation, molecular structure refinement or ligand-protein docking for drug design.

2.2. Molecular Dynamics

Molecular Dynamics (MD) simulations have become a powerful and popular tool in recent decades to gain insight into biomolecular structure, recognition and function of biological processes. MD simulations can explore the potential energy surface described by molecular mechanics, provide a microscopic interpretation of the phenomena, and even predict thermodynamic, kinetic, and structural properties of the molecular systems by modelling the motions and interactions of the atoms of a macromolecule over time. Combined with experimental structural biology techniques such as X-ray crystallography, nuclear magnetic resonance (NMR), small angle X-ray scattering (SAXS), Förster resonance energy transfer (FRET), etc., atomistic MD simulations are widely used to study enzymatic reaction mechanisms, optimise drug design projects, reveal pathologies related to protein misfolding, and many other applications in biophysics, materials science, molecular biology, pharmaceutical chemistry, and so on.

These methods emerged in the 1950s from the theoretical physics community. The first MD simulation was performed on a simple gas system at the end of the 1950s, using a model of rigid spheres with perfect collisions[4], but it was not until 1976 that the first simulation of a protein was completed, with a simulation time of 9.2 ps of the BPTI protein[5]. In fact, the first μs-length simulation of MD was not reported until 1998[6], whereas simulations of hundreds of nanoseconds with much larger number of atoms,

around 1000-10,000, are now standard. Fortunately, advances in computational power and the development and optimisation of methods and algorithms have contributed significantly to the performance of the simulations. In particular, the introduction of graphical processing units (GPUs) has meant a significant improvement in computational power, and software usability has also been refined over the years to become more user-friendly.

The MD approach is based on the iterative solution of the Newton's second law of motion within the framework of classical mechanics. From the interactions and the potential energy function defined by the MM models, it is possible to calculate the force acting on each atom and to propagate the motion of these particles in time through iterative algorithms, ultimately generating a trajectory of the molecules within the simulation system. By applying analysis techniques to these trajectories, the conformational ensembles and the intermolecular interactions of the molecular systems can be captured in MD simulations.

In more detail, if we consider a molecular system of N atoms with Cartesian coordinates $\boldsymbol{r}_i$ interacting with a potential $U(r_1, \dots, r_N)$ and apply Newton's second law of motion to the system, we can deduce that the force acting on an atom i is directly proportional to the mass, m_i, and acceleration, $\boldsymbol{a}_i$, of that particle.

$$\boldsymbol{F_i} = m_i \cdot \boldsymbol{a_i} = m_i \frac{\partial^2 \boldsymbol{r_i}}{\partial t^2} \qquad \textbf{Eq. 2.4}$$

From the Lagrange function, $\mathcal{L}$, a formulation of classical mechanics that is defined as the difference between the kinetic and potential energy to obtain the time evolution of a dynamic system, $\mathcal{L} = K - U$, and the gradient of the position of each atom, ∇_r, we can determine the force acting on each atom, $\boldsymbol{F_i}$, so that we can establish a connection between Newton's second law of motion and the potential energy function constructed by the MM model. By directly relating the derivative of the potential energy, the position of the particles and time is now defined as:

$$\boldsymbol{F_i} = \nabla_r = -\frac{\partial U}{\partial \boldsymbol{r_i}} \rightarrow -\frac{\partial U}{\partial \boldsymbol{r_i}} = m_i \frac{\partial^2 \boldsymbol{r_i}}{\partial t^2} \qquad \textbf{Eq. 2.5}$$

To solve this equation, which couples the particle motion and potential energy functions, it is necessary to apply iterative numerical methods. In MD simulations, finite difference methods are typically used, which discretise time into small time intervals, Δt,

in order to integrate the equations of motion. For this purpose, it is assumed that the motion can be approximated by standard Taylor series expansions by generally using the position $\boldsymbol{r_i}(t)$, the velocity $\partial \boldsymbol{r_i}(t)/\partial t$ and the acceleration $\partial^2 \boldsymbol{r_i}(t)/\partial t^2$ for the propagation of the position of each atom in the molecular system:

$$\boldsymbol{r_i}(t+\Delta t) = \boldsymbol{r_i}(t) + \frac{\partial \boldsymbol{r_i}(t)}{\partial t}\Delta t + \frac{\partial^2 \boldsymbol{r_i}(t)}{dt^2}\frac{\Delta t^2}{2} + \cdots \qquad \textbf{Eq. 2.6}$$

Normally this approximation is truncated at the second derivative and ignores the upper terms of the Taylor expansion. In fact, this truncation is quite crude and can lead to fluctuations and drifts in the total energy of the molecular system at long simulation times. Fortunately, numerical algorithms and other improvements implemented in MD simulations mitigate the errors associated with the integration of the equations of motion. Among the most popular integrators are the simple Verlet[7], the leapfrog[8] and the velocity Verlet[9]. The simple Verlet algorithm calculates the positions $\boldsymbol{r_i}(t+\Delta t)$ from the positions of the previous time step $\boldsymbol{r_i}(t-\Delta t)$ and the accelerations at time t, $\boldsymbol{a_i}(t)$.

$$\boldsymbol{r_i}(t+\Delta t) = 2\boldsymbol{r_i}(t) - \boldsymbol{r_i}(t-\Delta t) + \frac{\partial^2 \boldsymbol{r_i}(t)}{dt^2}\Delta t^2 \qquad \textbf{Eq. 2.7}$$

On the other hand, the leapfrog algorithm is a variation of the simple Verlet integrator commonly used in MD, which provides both positions and velocities during the simulation. In comparison with its predecessor, this algorithm is more efficient and minimises the numerical error. The leapfrog integrator is therefore based on calculating the velocities at time $t + 1/2\,\Delta t$, Eq. 8, and then updating the positions at time $t + \Delta t$, Eq. 9. If the velocities at time t are required, a simple calculation is performed using the velocities at time $t + 1/2\,\Delta t$ and $t - 1/2\,\Delta t$, as detailed in Eq. 10.

$$\boldsymbol{v_i}\left(t+\frac{1}{2}\Delta t\right) = \boldsymbol{v_i}\left(t-\frac{1}{2}\Delta t\right) + \frac{\partial^2 \boldsymbol{r_i}(t)}{dt^2}\Delta t \qquad \textbf{Eq. 2.8}$$

$$\boldsymbol{r_i}(t+\Delta t) = \boldsymbol{r_i}(t) + \boldsymbol{v_i}\left(t+\frac{1}{2}\Delta t\right)\Delta t \qquad \textbf{Eq. 2.9}$$

$$\boldsymbol{v_i}(t) = \frac{1}{2}\left[\boldsymbol{v_i}\left(t+\frac{1}{2}\Delta t\right) + \boldsymbol{v_i}\left(t-\frac{1}{2}\Delta t\right)\right] \qquad \textbf{Eq. 2.10}$$

The time evolution of positions and velocities can be calculated for each atom of the molecular system by iteratively following these steps and solving the potential energy function for the acceleration calculation. Other alternative integrators are also available, such as the velocity Verlet integrator, a descendant of the Verlet algorithm with

similarities to the leapfrog algorithm, or the Beeman integrator[10], which uses a more precise expression for the velocity calculation. All these integration algorithms are time reversible, i.e., we can return to the initial starting point if we start from another point of the simulation. The time reversibility is due to the symmetry in the evaluation of the derivatives in these integrators, which also guarantees the conservation of energy and momentum in many cases.

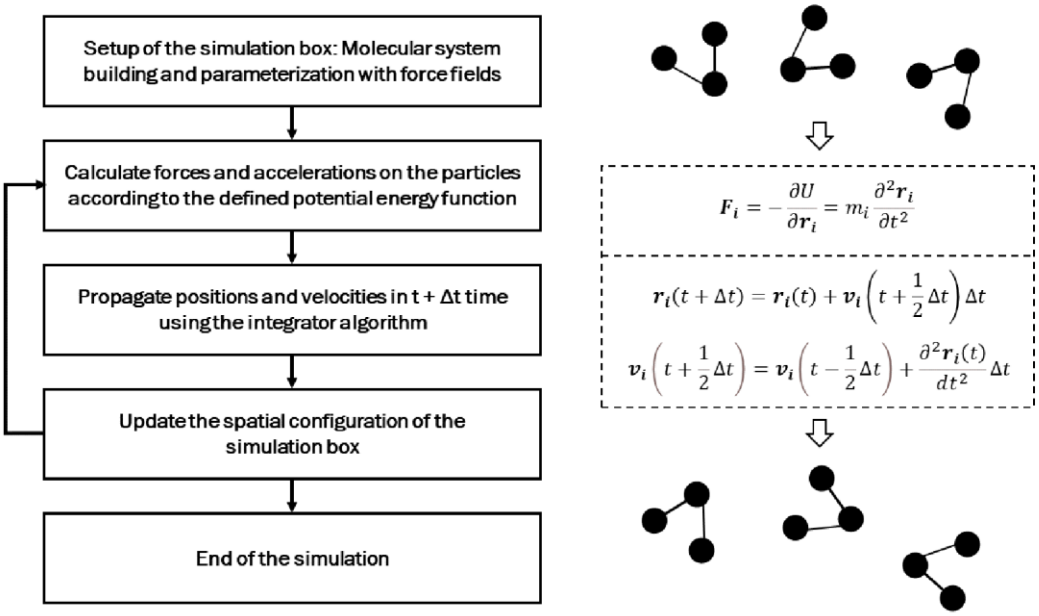


Figure 1. A schematic outline of the steps involved in the traditional Molecular Dynamics method.

2.2.1. Bond and Angle Constraint Algorithms

A notable limitation of the integrators in atomistic simulations is the length of the time step, Δt. When the molecular system is defined at all-atom resolution, the time step is limited to the fastest motion between pairs of atoms, which is the bond vibration of any hydrogen-involving bond. Thus, the efficiency of sampling the potential energy surface of a molecular system through time integration is slowed down by this time step limitation. Nevertheless, certain protocols have been implemented to increase Δt, such as the constraint algorithms or an increase in the mass of the hydrogen atom (and thus *freezing* the vibration of the H-involving bonds) by repartitioning the atomic masses within a molecule. As a result of these methods, the time step of all-atom MD simulations typically ranges from the standard values of 0.5-1 fs to 2-4 fs, depending on the approximations used. In fact, by sacrificing the accuracy of the atomic structure, either

by neglecting the hydrogens in the structure or by simplifying the molecular models with coarse-grained resolution, simulations can be performed with time steps of 10-40 fs, thus facilitating modelling on the micro- or millisecond time scale, which would normally not be feasible in conventional all-atom simulations.

Constraint algorithms fix distances and angles of covalent bonds in order to freeze the atomic vibrations. By imposing a constraint, such as an equilibrium distance between two pairs of atoms, the molecular system is forced to satisfy this condition during each integration step of the simulation. For the reasons given above, in simulations of biochemical systems these constraints are typically applied to H-involving bonds, leaving the rest of the molecular system free. Consequently, the constraint algorithms allow to increase the time step of the simulations by reducing the vibrations of the H-involving bonds, and thus increasing the simulation performance without compromising the trajectory (most important motions in such studies transcend the *frozen* bond vibrations). Among the most popular bond length-fixing algorithms are SHAKE[11], which modifies the Verlet integrator, RATTLE[12], which operates on the velocities of the velocity Verlet integrator, and LINCS[13], which resets the bond angles to the correct distance.

Here we detail the SHAKE algorithm, which has been used in the simulations performed of this thesis. This algorithm consists in imposing fixed interatomic distances with a constraint σ_k on the internal coordinates between two atoms forming a covalent bond. Then the constraint σ_k on the distance between atoms i and j is defined as:

$$\sigma_k = \boldsymbol{r_k^2} - d_k^2 = 0 \qquad \textbf{Eq. 2.11}$$

where $\boldsymbol{r_k} = \left(\boldsymbol{r_j} - \boldsymbol{r_i}\right)^2$ is the bond length vector and d_k is the desired equilibrium bond length between atoms i and j. Thus, for a system with K constraints that must satisfy $\sigma_k(r_1, \ldots, r_N) = 0$, the N atoms of a molecular system are subject to the equations of motion redefined as in Eq. 12, where λ_k are Lagrange multipliers that must be solved to satisfy the constraints.

$$\frac{\partial^2 \boldsymbol{r_i}(t)}{\partial t^2} m_i = -\frac{\partial}{\partial \boldsymbol{r_i}} \left[U(\boldsymbol{r_i}(t) - \sum_{k=1}^{K} \lambda_k \sigma_k(t) \right] \qquad \textbf{Eq. 2.12}$$

The resolution of the Lagrange multipliers is performed iteratively through coupled quadratic equations until the constraint satisfies a threshold, usually defined as ϵ / d_k^2 where ϵ is a constant to ensure an accuracy from 10^{-4} up to 10^{-8} Å. Thus, the algorithm

procedure consists of (i) the motion of the atoms through the integration algorithm without applying any constraint, (ii) the calculation of the deviation of the bond length and application of the constraint forces to correct it, and (iii) checking the deviations again to determine if they are below the desired threshold, if not, the second and third steps are repeated until the constraint is satisfied.

SHAKE is a numerically stable, simple and time-reversible algorithm. In addition, it allows the time step to be tripled in comparison with the original integrator algorithm. The combination of these constraints with RESPA[14] allows the time step to be further increased at the cost of not being able to find solutions for large bond length shifts. The LINCS algorithm, on the other hand, resets the constrained bond lengths after integrating the motion. It has some advantages over the SHAKE algorithm, such as greater stability and speed, but can only be applied to bond constraints and isolated angle constraints.

2.2.2. Periodic Boundary Conditions and Truncation of Interactions

To realistically represent a biochemical system, e.g., a protein in solution, using MD simulations one would need to model a simulation box containing a large number of proteins, water molecules and ions. This would ensure that the proteins in the simulation box were adequately solvated, and would also include the short- and long-range interactions between all the molecules that can play a part in the sampling of the potential energy surface would be taken into account. However, besides the prohibitive computational cost of integrating the equations of motion of all these molecules and calculating their interactions, we would have a second problem: the edge effect of the atoms close to the boundaries of the simulation box. Fortunately, an optimal solution to these drawbacks has been devised, the so-called periodic boundary conditions (PBCs).

The periodic boundary conditions define a periodic unit box (in the context of MD simulations, the simulation box) that is periodically repeated in all directions in the space to surround the central simulation box, formally representing an infinite bulk system. The atoms in the periodic images behave identically to the simulation box, and when an atom leaves the boundaries of the unit box and enters an image, the identical atom in the opposite image enters the unit box. In this way, the number of atoms in the unit box is always preserved. By using PBCs, the number of atoms needed to reproduce a bulk system is drastically reduced and, in addition, the introduction of periodic images into the molecular system eliminates edge effects. In fact, a minimum solvation distance of ~10-

15 Å around the protein has been found to satisfy bulk conditions. Depending on the molecular system, PBCs can be applied in simulation boxes with different geometries. The cubic box is commonly used in MD simulations, although rhombic dodecahedrons or truncated octahedrons have also gained popularity, especially for globular proteins since they reduce the number of water molecules needed in the system compared to a cubic box.

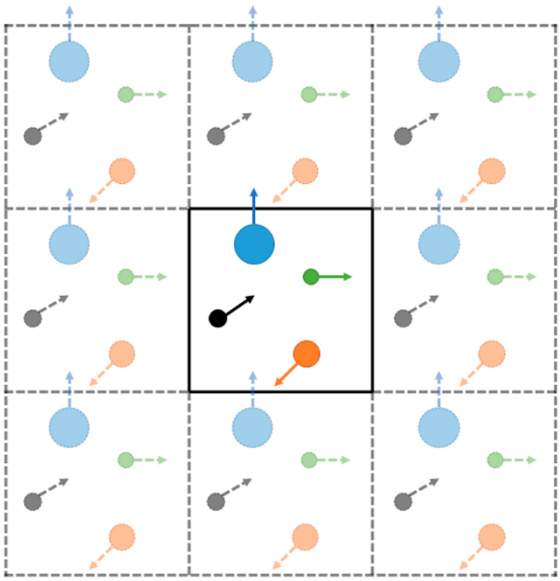


Figure 2. Periodic boundary conditions in a system of four coloured particles. The simulation box is in the middle, defined by solid lines. The periodic boxes or images, also virtual particles, are defined by dashed lines.

On the other hand, PBCs require some considerations regarding pairwise interactions, since the images around the simulation box would cause an infinite sum of interactions between atoms. Therefore, the so-called minimum image convention is applied, which states that an atom i can only interact with the nearest atom j, regardless of whether atom j is in the simulation (or unit) box or is an image of a periodic box. In other words, each atom can only interact with the atom j only once. To achieve this convention, it is necessary to define a spherical cut-off of radius r_c around each particle, which cannot be larger than half the shortest side of the simulation box. Otherwise, the interactions would be duplicated, and the convention could not be applied. Thus, interactions within the cut-off are considered, i.e., the short-range interactions, and any interaction above r_c is neglected, thus truncating the potential energy, and allowing the calculation of interactions in a computationally feasible manner.

The spherical truncation[15] causes a discontinuity in the energy calculated around an atom, and an abrupt increase in the potential can destabilise the simulation. In some situations, a smoothing function is applied to avoid crashes or other pitfalls during the simulations. Non-polar vdW interactions, usually modelled by the Lennard-Jones 12-6 potential, can be properly reproduced within the radius r_c (~10-12 Å) because the LJ potential decays significantly with the increase of the distance between two interacting atoms increases. However, this does not occur with long-range electrostatic interactions, which often play a paramount role in the first steps of protein-protein association or ligand-receptor binding, and other approaches must be considered. Several modifications of the spherical truncation have been reported to improve the estimation of these interactions, although they are limited by the computational cost of N^2. Of note is the generalised reaction field method[16], which proposes an explicit calculation of the electrostatic interactions inside the spherical cut-off, while outside the truncation sphere establishes a uniform dielectric continuum dependent on the ionic strength is established. This method has proved to be efficient and simple, enabling the electrostatic interactions to be approximated at a reasonable computational cost. However, it does not conserve energy well and requires prior knowledge of the external dielectric potential, although it has been shown to be consistent with other methods in some particular systems.

The most popular method currently used in MD simulations to calculate electrostatic interactions is the particle mesh Ewald[17] (PME). This method separates the electrostatic potential between atoms i and j into a short-range contribution, E_{sr}, which is calculated in real space by a direct sum, and the long-range interactions, E_{lr}, which are summed in reciprocal space by Fourier transforms.

$$E = \sum_{i,j} \varphi(\boldsymbol{r_j} - \boldsymbol{r_i}) = E_{sr} + E_{lr} = \sum_{i,j} \varphi_{sr}(\boldsymbol{r_j} - \boldsymbol{r_i}) + \sum_{k} \widetilde{\Phi}_{lr}(\boldsymbol{k})|\tilde{\rho}(\boldsymbol{k})| \qquad \textbf{Eq. 2.13}$$

In contrast to the Ewald summation method[18], which was originally developed to estimate the electrostatic energy in ionic crystals, the PME method uses fast Fourier transforms, $\widetilde{\Phi}_{lr}$, for the summation in reciprocal space, which evaluates the charge density field, $\tilde{\rho}(\boldsymbol{k})$ after discretising it into a mesh in space to reduce the computational cost to $N \cdot log(N)$. As a result, the algorithm scales faster and shows efficient performance in the calculation of the electrostatic potential. However, to apply the PME method, it is necessary to take into account the periodicity assumption implicit in the Ewald summation and requires PBCs in a molecular system with a neutral net charge.

2.2.3. Statistical Ensembles and Experimental Conditions

A statistical ensemble is a collection of all those microstates (i.e., the microscopic configuration of a system) that reflect the same macroscopic state of a system and are therefore described by a set of macroscopic observable variables. This means that if we perform several experiments under the same conditions and repeatedly observe the thermodynamic properties, this does not imply that the microscopic or molecular states in the different measurements are identical. Thermodynamic variables typically observed in biochemical systems include pressure p, temperature T, volume V, number of particles N and chemical potential μ. In early MD simulations, molecular systems were integrated in the microscopic collective, i.e., the number of atoms N, volume V and energy E were fixed during the simulation in order to reproduce a macrostate described by these three thermodynamic properties (N, V, E). In fact, in the case of the microcanonical ensemble, conserving the energy of the molecular system is only possible in an ideal simulation in which the equations of motion are integrated without errors (i.e., by expanding the Taylor series to infinity) and the potential terms fully considered. However, to make the simulations computationally feasible, a number of approximations are required, as explained in the previous sections, but prevent the microscopic ensemble from being correctly simulated by MD. In addition, biochemical experiments in the laboratory are usually carried out under conditions of constant temperature or pressure, if not both, so reproducing these conditions at the microscopic level is more likely to yield results that are consistent with the experiment. For these reasons, most biochemical simulations are performed in the canonical (N, V, T) or isobaric-isothermal (N, p, T) ensembles, in which the total energy of the system can fluctuate. There are other interesting collectives, such as the grand-canonical collective (μ, V, T), in which the chemical potential μ remains fixed, but the particles can be exchanged with an external bath such that the number of particles N varies with time.

To perform the simulations in these statistical ensembles, it is necessary to apply a thermostat or a barostat to keep the temperature or pressure constant. In this thesis the simulations are mostly carried out in the canonical ensemble, therefore more emphasis is given in this section, while the isobaric-isothermal ensemble is mainly used to equilibrate the simulation box. As for the thermostats, the simplest algorithm is the Berendsen thermostat[19] which scales the velocities at each integration step based on a λ-scaling factor. This method couples an external thermal bath at the desired temperature to the

molecular system, thus adding or removing heat from the simulation box at an exchange rate that depends on the temperature difference between the external bath and the molecular system. Although the Berendsen thermostat is widely used in MD simulations, the lack of kinetic energy fluctuations within the molecular system prevents this thermostat from correctly representing the statistical ensemble. As a solution, a generalisation of the Berendsen thermostat has been proposed, introduced as the velocity-rescaling thermostat[20], which includes a stochastic term that allows for a correct distribution of the kinetic energy. This algorithm eventually replaced the Berendsen thermostat. Among the velocity-rescaling methods, there is also the Nosé-Hoover thermostat[21–23], which uses an extended system to relax the temperature of the molecular system by adding artificial terms (velocity and coordinates) to the Lagrangian function. Due to the cyclic fluctuations of this thermostat, a series of thermostats are chained to correct this problem and ensure the ergodicity of the system, which is essential for obtaining meaningful information from the simulation.

In contrast to the previous thermostats, we introduce the velocity randomising (or stochastic) Langevin thermostat[24]. This thermostat integrates directly on the Langevin equation of motion, in which dissipative forces are included in Newton's equation of motion through a friction term λ_i and a random force term R_i to reproduce random collisions between the atoms of the simulation box and random particles of an external thermal bath at the desired temperature T. The frequency of the collisions between the particles is determined by λ_i and the random force is related to this frequency parameter by $\langle \boldsymbol{R}(0)\boldsymbol{R}(t)\rangle = 2m_i k_b T \gamma_i \delta(t)$, where k_b is the Boltzmann's constant and $\delta(t)$ is the Dirac function. Therefore, the equation of motion when using Langevin thermostat is defined as follows:

$$m_i \frac{\partial^2 \boldsymbol{r_i}}{\partial t^2} = \boldsymbol{F_i} - m_i \gamma_i \frac{\partial \boldsymbol{r_i}}{\partial t} + \boldsymbol{R_i} \qquad \textbf{Eq. 2.14}$$

This thermostat has demonstrated good performance and the equation of motion is physically meaningful with a real friction parameter that can simulate solvent molecules despite the complexity of the stochastic fluctuation.

On the other hand, some barostats rely on the architecture of thermostat algorithms (or vice versa) to perform their task. Thus, they generally modify the dimensions of the simulation box and consequently the coordinates of the atoms inside the system to adjust the volume and thus maintain the pressure of the system. Among the various methods, it

is worth mentioning the Berendsen barostat[25], which presents the problem presented above, the Parrinello-Rahman pressure coupling barostat[26,27], which applies a system extension to the Lagrangian, the Langevin pistol method[28] or hybrid methods[29].

2.2.4. Solvation Models

Biomolecules, such as proteins, DNA, membranes, etc., are usually found in very crowded and complex environments, composed mostly of water with the presence of other macromolecules or ions. From a computational point of view, including all the components of a physiological environment in the simulation box is very expensive. In fact, the typical dimensions of the simulation boxes do not permit the inclusion of other macromolecules. Therefore, most simulations assume only an aqueous solvation medium, either pure or ionic, to fill the molecular system. The presence of water is very important because of its effect on the conformation of macromolecules, which makes them a critical factor in biological processes such as solvation and self-assembly phenomena. Therefore, an accurate representation of the water properties, solvent-solute and solvent-solvent interactions is essential for the simulation and study of biomolecules.

There are two major approaches to incorporate the properties of water into the environment of the simulations. One is the implicit (or continuous) solvation model, which assumes a continuous medium with electrostatic and non-polar contributions to mimic the properties of water. Most implicit solvation models treat the electrostatic or polar interactions and the non-polar interactions separately, as shown in Eq. 15.

$$\Delta G_{solv} = \Delta G_{el} + \Delta G_{nonpolar} \qquad \textbf{Eq. 2.15}$$

For the polar contribution, the electrostatic interactions are traditionally estimated by the Poisson-Boltzmann (PB) equation[30,31], in which the description of the electrochemical potential has proven to be robust but computationally expensive. According to this model, the solute is treated as a dielectric body with a shape determined by the atomic cavity radius, in which the point charges are distributed in the atomic centres. An electric force field is then generated corresponding to this dielectric body and the solute, and the PB equations are used to estimate the electrostatic interactions.

$$\boldsymbol{\nabla} \cdot [\epsilon(\boldsymbol{r_i})\boldsymbol{\nabla}\phi(\boldsymbol{r_i})] = -4\pi\rho(\boldsymbol{r_i}) - 4\pi\lambda(\boldsymbol{r_i}) \sum_i z_i c_i e^{\frac{-z_i\phi(\boldsymbol{r_i})}{kT}} \qquad \textbf{Eq. 2.16}$$

where $\epsilon(\boldsymbol{r_i})$ is the dielectric constant, $\phi(\boldsymbol{r_i})$ is the electrostatic potential, $\rho(\boldsymbol{r_i})$ is the solute charge density, $\rho(\boldsymbol{r_i})$ is the masking layer function of the Stern model, z_i is the ion charge and c_i is the ion charge concentration in the bulk.

Alternatively, the PB equation has been simplified by an analytical solution in which the solute is modelled as a set of spheres with radius R_i, charge q_i and filled with a dielectric constant of 1. Subsequently, the solute is surrounded by a continuum solvent with a specific ε, so that the electrostatic interactions can be solved with the following analytical equation:

$$\Delta G_{el} = -\frac{1}{2}\sum_{ij}^{N}\frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)}\left(1 - \frac{e^{-\mathcal{K} f_{GB}}}{\varepsilon}\right) \qquad \textbf{Eq. 2.17}$$

where $f_{GB}(r_{ij}, R_i, R_j)$ is a smooth function that depends on the distance between the atoms i and j and the associated radius assigned to each atom, also referred to as the effective Born radii, and $\mathcal{K}$ is the Debye-Huckel screening length. This approach, called generalized Born (GB) model[32–34], is an efficient and simple method to simulate implicit solvation in MD simulations. It also has the advantage of being parallelizable on computers, which is very convenient for computationally demanding studies such as protein folding, solvation free energy calculations, simulations at constant pH, etc.

In contrast, the non-polar contribution is usually addressed by the solvent accessible surface area (SASA)[35,36], a method that numerically estimates the molecular surface area exposed to the solvent using spherically distributed dots and a pair of parameters related to the surface tension, γ, and the free energy in vacuum, c.

$$\Delta G_{nonpolar} = \gamma SASA + c \qquad \textbf{Eq. 2.18}$$

Despite the advantages of implicit solvation models in terms of computational efficiency, as they significantly reduce the number of molecules and interactions within the simulation box (water molecules can account for up to 90% of the system), these models have important shortcomings in microscopic properties and solvent-solute interactions. The implicit solvation approach assumes the absence of explicit solute-solvent interactions, such as hydrogen bonds, and of entropic effects arising from the distribution of the solvent around the solute, in addition to the overstabilisation of intramolecular salt bridges or hydrogen bonds, resulting in improper conformational sampling of the molecular system.

On the other hand, in order to have a realistic description of the molecular system, explicit water models can be introduced into the simulation box. These explicit models consist of a specific geometric structure and a set of force field parameters (i.e., bond lengths, angles, partial charges, or Lennard Jones-related constants, among others) that reproduce the water molecule. Common water models of interest for this thesis are summarised in Table 1.

	TIP3P	**SPC**	**OPC3**	**TIP4P**	**TIP4P-D**	**OPC**
O-H bond (Å)	0.9572	1.0	0.9789	0.9572	0.9572	0.8724
H-O-H angle (°)	104.52	109.4667	109.47	104.52	104.52	103.6
O-M bond (Å)	-	-	-	0.15	0.1546	0.1594
$\mathbf{q_O}$ (e)	-0.834	-0.82	-0.8952	0.0	0.0	0.0
$\mathbf{q_H}$ (e)	0.417	0.41	0.4476	0.52	0.58	0.6791
$\mathbf{q_M}$ (e)	-	-	-	-1.04	-1.16	-1.358
$\boldsymbol{\sigma_O}$ (Å)	3.1506	3.166	3.17427	3.15365	3.165	3.1666
$\boldsymbol{\varepsilon_O}$ (kcal/mol)	0.1521	0.1554	0.1634	0.155	0.2238	0.2128

Table 1. Force field parameters (distances, angles, partial charges and van der Waals parameters) of the most relevant 3- and 4-point water models in this thesis.

In the 1980s, the most popular water models, the Transferable Interaction Potential[37] (TIP3P) and the Simple Point Charge[38] (SPC) water molecules, were developed based on the description of water by 3-point models. These models are still used today in combination with certain well-established force fields, although some limitations in the ability to reproduce certain properties of water have already been demonstrated. Fortunately, a collection of other versions has emerged over the years, based on modifications or the addition of new points to the original models. For example, 4- and 5-point water models (TIP4P[37], TIP5P[39]), with improved description of the intermolecular potential energy (SPC/E[40]), better compatibility with the Ewald summation method (TIP3P-Ew[41], TIP4P-Ew[42]), flexible versions of the rigid model (SPC/Fw[43], TIP3P/Fw[44]), 4-point polarisable models (SWM4-NDP[45]), with improved London dispersion force interactions (TIP4P-D[46]), updates from previous models (TIP5P-2018[47], TIP4P/2005[48]) and even *high-accuracy* models (OPC[49], OPC3[50]). Given the wide variety of models available, users are encouraged to read the specifications of each model carefully and choose according to the purpose of the simulation and the affordability of the computational performance.

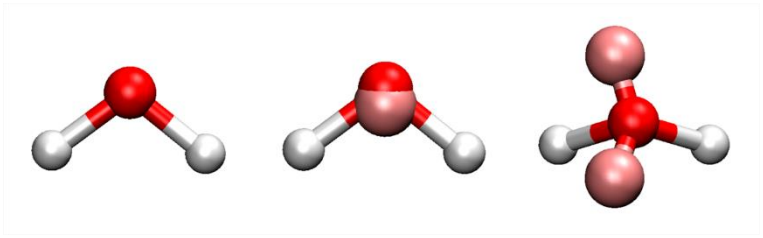


Figure 3. Water molecule models included in the review by Kadaoluwa Pathirannahalage et al.[51]. Partial charge points of 4- and 5-point models are coloured in pink and cyan, respectively.

2.2.5. Considerations for Simulation and Software

As well as the explicit water models, the parameters and terms used in the force fields for the description of biomolecules have also been developed over the years. An accurate parameterisation is crucial for a successful modelling of the potential energy surface and hence the molecular systems. However, the simulation setup should also be handled with care to ensure a correct description of the modelled system. Several procedures are usually required before running the simulation, such as the selection and preparation of 3D structures with missing atoms or regions, the prediction of the protonation state of the titratable amino acids, the addition of disulphide bridges, the solvation of the simulation box (water molecules, ions, ...), the parameterisation of organic molecules, the heating of the simulation box (i.e., by gradually increasing the temperature in the external bath of the thermostat), adjusting the density of the simulation box (with a barostat), equilibrating the system, and so on. As can be seen, all these processes can be difficult for non-experts in the field of molecular modelling, but fortunately there are several software packages available that can deal with them in a systematic, user-friendly, and easy-to-use manner. Some of the most prominent and popular for simulating biochemical systems are AMBER[52], CHARMM[53], GROMACS[54], DESMOND[55] or NAMD[56], and some of them even have their own force fields, such as AMBER or CHARMM. Other prominent force fields are OPLS[57] or GROMOS[58], and all have several versions to improve the accuracy either in general or in specific molecular systems. Most force fields have a similar functional form (i.e., the potential energy terms that describe the potential energy surface of the system), and the key feature between them relies on the parameterisation of the molecules.

The user can choose the software based on the efficiency of algorithm implementation, availability of advanced sampling or analysis techniques, parallelization on CPUs, GPUs or supercomputers, flexibility to integrate new force fields, graphical user interface and many other considerations. The most common software packages already include many analysis techniques for extracting information about the dynamics and structural properties of molecular systems. However, the use of libraries specialised in molecular modelling, such as MDTraj[59], MDAnalysis[60] or PyTRAJ[61], is a good resource in cases where specific analyses are not found, modifications to them are required or several programs need to be linked. There are also molecular visualisation programs, such as VMD[62] or PyMOL[63], which provide the user with eye-catching qualitative information about the dynamics and the structure of the molecule. Therefore, it is recommended that users explore the various tools and software packages that are available prior to running a simulation, as this can facilitate the study of a desired molecular system in a more efficient and easier manner.

2.3. Coarse-Grained Modelling

Up to this point, we have assumed that molecular modelling is performed at an atomistic resolution, i.e., all atoms of the molecular system (or the solute) are explicitly defined in the simulation box. This high-resolution description of a biochemical system for a medium or large protein generally requires a substantial computational effort to reach long time scales. Regardless of the increase in computational power with GPUs or supercomputers or the approaches within atomistic resolution to enlarge the time step, the cost of capturing biological processes at micro- or millisecond time scales, such as protein aggregation, protein folding, cryptic binding sites, etc, can be prohibitive. For these reasons, a reduction in the atomistic representation of molecular systems is sometimes used in molecular modelling to reduce the complexity and degrees of freedom of the system and thus to increase the simulation efficiency. These medium-to-low resolution molecular descriptions are often referred to as coarse-grained (CG) models, and are based on the assembly of neighbouring atoms within a molecule into *average* particles.

The foundations for the development of multiscale models of complex chemical systems were laid in the 1970s by Michael Levitt, Ariel Warshel and Martin Karplus. They introduced the simplification of the biomolecular complexes in order to perform simulations on long time scales. In fact, they were awarded the 2013 Nobel Prize in Chemistry for this contribution to multiscale simulations and the study of biomolecular

complexes[64]. Since then, many multiscale and coarse-grained models have been developed, progressing from simple to more sophisticated and detailed models, and have become increasingly popular over the years for soft matter research in physics, chemistry and biochemistry. Within protein simulation, CG simulations have proven to be effective in the prediction of structural properties and protein folding mechanisms, and have also shown potential for gaining insight into the protein-protein interactions or the behaviour of molecular membranes.

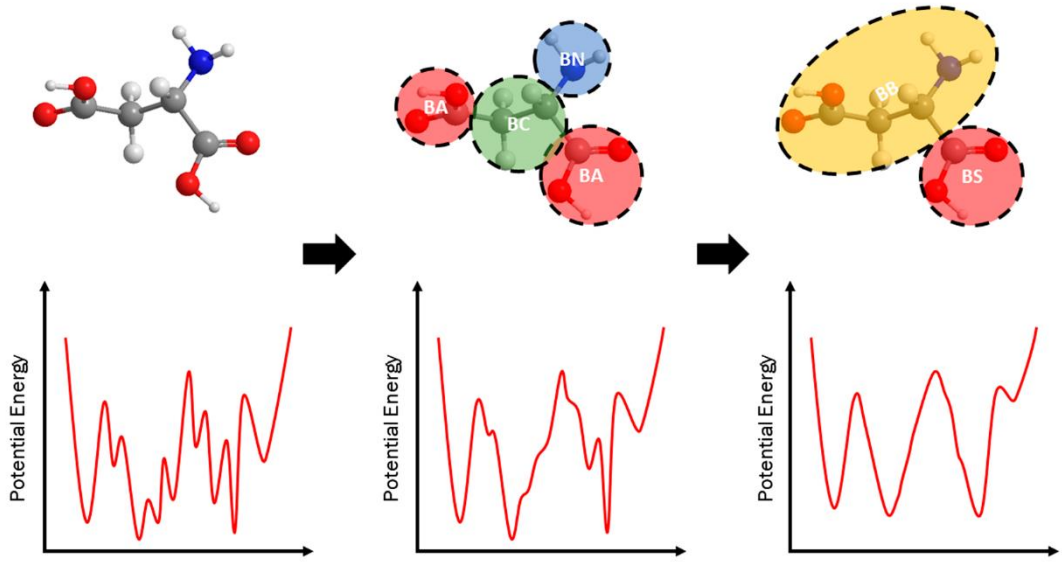


Figure 4. Illustration of the structure and potential energy of aspartic acid with all-atom (left), 4-bead (middle) and 2-bead (right) coarse-grained models.

In the context of biomolecules, CG models reduce the level of representation of the amino acid chain to one, two or more *united* atoms (i.e., the representation that includes several atoms in the same particle) or pseudo-atoms. Sometimes these united atoms are also referred to as beads belonging to a necklace or chain. Depending on the CG model, the level of representation varies, e.g., we can find models with a single united atom per amino acid or several beads to represent the functional groups of an amino acid. Nevertheless, for an adequate modelling we must not forget that these simplified representations must reflect the particularities of the proteins, such as the peptide bond binding in the trans position, the orientations of the side chains, the L-handed conformation of the amino acids, etc. Typically, intermediate models estimate the molecular systems with CG models that assign one or two pseudo-atoms in the backbone and side chain of the amino acids, such as the UNRES[65] and CABS[66] models. There are

CG models with high resolution in the molecular structure, i.e., more pseudo-atoms, with the advantage of providing a better description of the molecular system at the expense of a larger computational time, such as PRIMO[67] or Rosetta[68]. There are also other less common lattice-based models with certain restrictions on geometry and conformations.

Apart from the level of resolution, the design of the force fields in CG models is crucial for the accurate simulation of molecular systems. Three categories can be distinguished according to the approach adopted for their construction: (i) physics-based force fields, which are constructed from all-atom simulations, (ii) knowledge-based force fields, which are derived from the statistics of the structural properties, and (iii) structure-based force fields, which use well-defined protein structures to reproduce the native contacts. The former is usually based on the construction of a potential energy function similar to that described in classical all-atom force fields. However, the terms integrating the potential of the molecular system should be modified due to the pseudo-atom building and require a treatment based on multibody terms representing pseudo-bonds between the united atoms and the non-bonding interactions between the pairs of interacting atoms. There are several strategies for fitting and evaluating the parameters chosen for the CG force fields, generally based on minimising the correlation between the properties of the conformational ensembles generated between the CG and the all-atom molecular systems, such as the radial distribution function, the applied forces or the relative entropy, among others. On the other hand, knowledge-based force fields build the potential energy function through statistical analysis of the conformational properties of the system, such as the relative frequency of atomic contacts, the probability of a correct conformation based on the amino acid sequence and structural features, or the maximisation or minimisation of different criteria (score function, free energy of native states, ...). These methods have demonstrated good efficiency for molecular systems with structural properties similar to those considered in their design, but the transferability of these CG force fields is poor. Finally, the structure-based models are not really popular, since the construction of these force fields through a well-defined initial structure assumes that the relevant interactions of the molecular system are governed by those present in the native structure. In addition to the force fields and models mentioned in the classification, others stand out, such as MARTINI[69,70] (physics-based force field using one to four beads per amino acid), Bereau and Deserno[71] (knowledge-based force field implemented in the

ESPResSO software), OPEP[72] (mixed potential using one to six beads per amino acid), etc[73].

In Chapter 6, CG models are used to assess the ability to sample the conformational space of the histatin-5 peptide using various methods, models, and force fields. Available and most familiar models in our research group include the CG SIRAH model[74,75] implemented in AMBER and GROMACS, and the flexible CG software package ESPResSO[76]. The SIRAH force field is designed according to the knowledge-based approach using structural information available in the Protein Data Bank and canonical structures of DNA and proteins with α-helices and β-sheets to fit the parameters of a classical potential energy function from MD simulations (including familiar concepts of electrostatic interactions, van der Waals, equilibrium angles and distances, force constants, etc.). In addition, SIRAH has developed its own explicit water model, WAT FOUR[77] (WT4), which represents a cluster of 11 water molecules in a tetrahedral transient structure embodied by four beads. SIRAH has proved to reproduce the secondary structures of short peptides, such as chignolin, (AAQAA)3 and the YSEEEERRRR peptide, even from the unfolded state. On the other hand, the ESPResSO software package offers great flexibility in the design of CG models since particles and interactions can be customised with relative ease. Here, the CG resolution was simplified to 2-bead models connected by spring-like bonds and an explicit solvation environment using the "Sugar" library developed by Blanco, P.M. (https://gitlab.com/blancoapa/sugar_library). A major advantage of this software for the present thesis is the implementation of the pH-inclusive methods, such as the constant pH or reaction ensemble method.

2.4. Intrinsically Disordered Proteins in Molecular Dynamics

As mentioned in the introduction, the intrinsic disorder of IDPs confers upon them properties that are translated into high flexibility and rapid interconversion between conformations. Not surprisingly, the first molecular modelling efforts were influenced by the perceived need to reproduce the well-defined three-dimensional structures typical of globular proteins, as the function-structure paradigm was well established in the scientific community. In fact, many of the force fields focused on proteins and peptides were designed based on this assumption and are still in use today. However, such force fields fail when simulating IDPs or intrinsically disordered regions (IDRs), usually due to the overstabilisation of secondary structures, e.g., the α-helix or β-sheet, so that new parameterisations including the features of IDPs are necessary. Fortunately, in recent

decades, IDP-oriented force fields have been published addressing these issues and reproducing experimentally observed properties, such as the radius of gyration (R_g) from SAXS, atomic distances from FRET or the chemical shifts prediction from NMR. However, there are still limitations in the development of these force fields and, more ambitiously, no universal force field exists for both folded and disordered proteins[78,79].

Several strategies have been pursued to develop novel IDP-inclusive force fields. Most of them are based on training (or retraining) the parameters of the (previous) force fields to improve the accuracy in predicting the conformational behaviour and secondary structure propensities. The training data sets are obtained from experiments or quantum mechanics simulations and are intended to capture the properties of IDPs. On the other hand, the importance of protein-water interactions in the simulation of IDPs has been highlighted, so that modifications to water models have also been proposed. Other approaches have also been implemented, such as the use of polarisable force fields, CG models or even force fields designed using machine learning. In this section we review the most important approaches that tackle the IDP-orientated force fields.

The more common option in IDP-specific force fields is to optimise the dihedral angles, in particular the dihedrals φ and ψ of the protein backbone, in order to minimise the overstabilisation of certain secondary structures typically observed in earlier force fields. Therefore, data sets incorporating the structural information of random coils are added to the training processes to recalibrate the force field parameters. Thus, the parameters defining the dihedral angle potential given in Eq. 2 are refitted. Some examples are FF99SB*[80], CHARMM22*[81], OPLS-AA/M[82] or OPLS3[83], all of which have shown improvements in reproducing of the characteristics of certain IDPs. In addition to updating the training sets, the RSFF1[84] and RSSF2[85] force fields have been designed with amino acid-specific parameters to improve the conformational sampling of IDPs, particularly in the secondary structure propensities.

Another strategy adopted in the dihedral angle fashion is the introduction of a corrective term in the potential energy function of the molecular system, the so-called CMAP method[86]. This was first incorporated into the CHARMM22/CMAP[86,87] (or CHARMM27) force field, adding a corrective energy surface that depends on the φ and ψ dihedral distribution of the protein backbone. In general, the two-dimensional distribution generated by the φ and ψ dihedrals of each residue is divided into several

bins, such that the dihedral free energy of a bin i is based on its population during the simulation, ΔG_i^{SIM}.

$$\Delta G_i^{SIM} = RTln\left(\frac{N_i}{N_T}\right) \quad \textbf{Eq. 2.19}$$

where N_i is the number of times that φ and ψ dihedrals are counted within bin i, and N_T is the total number of dihedral combinations during the simulation. The population energy of the residue is compared with a reference energy value derived from experimental databases, ΔG_i^{DB}. If the dihedral free energy of the simulation deviates from the experimental data, an energy correction is applied in the potential energy function, U_i^{CMAP}, which depends on the difference between experimental and simulation free energies.

$$U_i^{CMAP} = \Delta G_i^{DB} - \Delta G_i^{SIM} \quad \textbf{Eq. 2.20}$$

As the number of bins is discretised to a low-dimensional matrix, bicubic interpolation or nearest neighbour methods are typically used to generate a continuous correction energy potential in the system. The CMAP method has finally been implemented in the most popular force fields, such as the AMBER ff14IDPSFF[88], CHARMM36IDPSFF[89] or OPLSIDPSFF[90] force fields, in which Chen and co-workers incorporated the CMAP correction on the 20 standard amino acids, the RSFF2C force field[91], in which the dihedrals of the side chains are also included, and the ESFF1 force field[92], in which the energy correction depends on the sequence environment of the amino acids.

Another line of improvement for simulation of IDPs is the refinement of the protein-water interactions. In addition to the electrostatic interactions that are normally dominant in proteins, it has been found that short-range non-polar interactions are also essential for the compaction and conformation of IDPs. This is usually quantified by the radius of gyration or the end-to-end distance, which can be measured experimentally using SAXS or FRET techniques. Therefore, developing water models that are consistent with experimental observations has become a challenge in IDP simulations. Water models designed based on this consideration typically modify the Lennard-Jones potential parameters of the hydrogen or oxygen atoms of the water molecule. For example, the TIP4P-D model (or the modified version a99SB-disp) increases the dispersion interactions of the water model, resulting in an improved R_g of some IDPs at the cost of

occasionally breaking α-helices and overestimating the R_g, and the CHARMM36m water model adjusts the Lennard-Jonnes well depth parameter of the hydrogen atom,ε_H. The balance of protein-water van der Waals interactions needs to be handled carefully. If overestimated, it will lead to loss of secondary structure and increased solvent exposure, and on the contrary, if underestimated, protein collapse and compaction will occur.

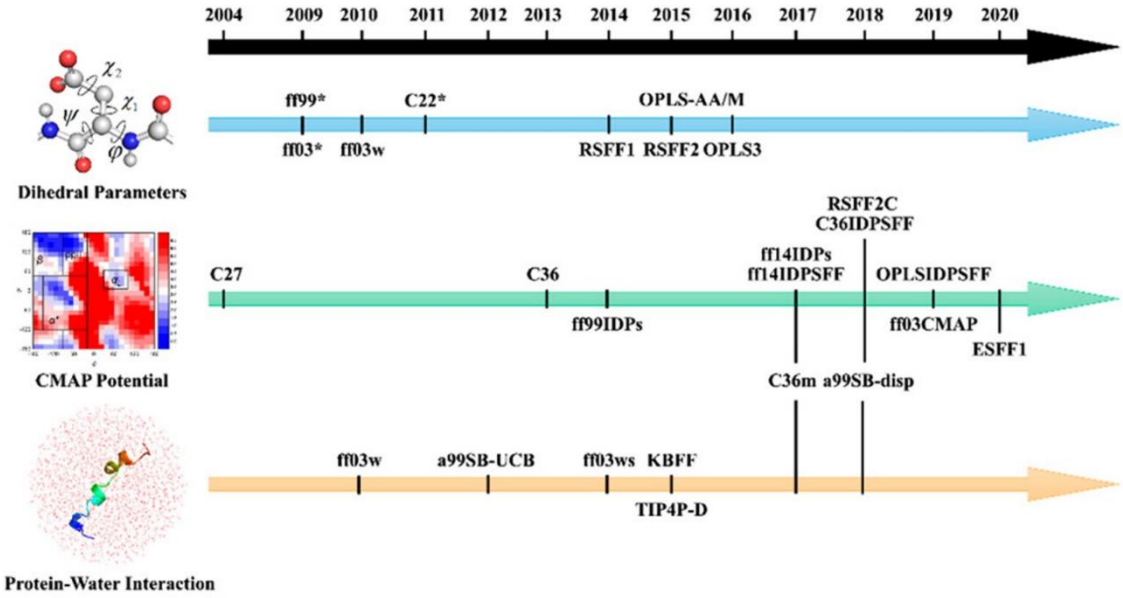


Figure 5. Schematic representation of the dihedral, CMAP and protein-water interaction approaches for improving the simulation of IDPs and the respective force fields and water models designed according to each strategy from Mu *et al.*[79].

Other approaches to improve the parameterisation of simulations include the use of machine learning techniques to calibrate force fields with experimental data, such as SAXS and SANS intensities in the ForceBalance-SAS force field[93]. However, similar to the parameterisation of dihedrals, the selected training set must be carefully considered as it will determine the strengths and weaknesses of the force field. On the other hand, CG IDP-specific force fields have also shown merits in performing IDP simulations, such as the AWSEM-IDP force field[94] in the aggregation process or the OPEP force field in amyloid-β fibril formation. Indeed, they are an interesting choice, particularly in terms of computational efficiency, to observe biological events on long time scales. Finally, dynamic partial charges are increasingly emerging as a promising approach beyond the reparameterisation strategies presented to date. As mentioned in the introduction, IDPs usually contain a high fraction of ionisable or polarisable amino acids, which means that electrostatic interactions and an accurate description of the partial charges according to

the electrochemical environment are critical for the study of IDPs. Therefore, modern polarisable force fields have attempted to address this issue through fluctuating charge models (in OPLS-AA and CHARMM), Drude oscillator models (in CHARMM), induced dipole and Gaussian models for electrostatic interactions (in AMBER), or detailed multipole expansion and complex potentials (in AMOEBA[95]). Although promising, these force fields present computational challenges, and finding suitable solutions to this problem would make them into attractive models for the simulation of IDPs.

2.5. Constant pH Molecular Dynamics

A correct definition of the partial charges of the atoms of a macromolecule is determinant for a realistic simulation of the systems. In biomolecules, the protonation states of ionisable amino acids are critical for the charge distribution of proteins, which in turn can have a profound impact on conformations and consequently on the biological functions. A clear example is proton-coupled conformational dynamics, where the protonation (or deprotonation) of one or two amino acids causes major structural changes in the configuration of a biomolecule. These amino acid protonation states play a central role in other phenomena such as ligand-protein or protein-protein binding processes, mechanisms of membrane channels or ion tunnels, and so on. At present, most simulations usually fix protonation states during MD simulations. Even if some considerations are taken into account during the preparation of the molecular system, e.g., considering the pKa (usually the reference pKa from the individual amino acid) with respect to the *simulated* pH or using protonation state prediction tools such as PROPKA[96] or H++[97], these are insufficient. On the one hand, the reference pKa of the individual amino acids is not reliable for residues within proteins because the pKa can vary significantly depending on the electrochemical environment. On the other hand, pKa estimation tools are poor because they typically use a single structure or a small set of conformations, which is not adequate to accurately predict pKa. Above all, simulations with fixed protonation states do not contemplate the possibility of changing the protonation state during the trajectory, thus neglecting the interplay and interactions between amino acids with dynamic protonation states.

Over the last few decades, several techniques that include the effect of pH and dynamic protonation states of amino acids have been developed, collectively known as constant pH Molecular Dynamics (CpHMD) method[98–103]. The CpHMD techniques explore both the conformational and protonation state space of molecular systems within

the semi-grand canonical ensemble, $(N\Delta\mu_i VT)$. In contrast to the canonical ensemble, the number of protons can fluctuate during the MD simulation, enabling the protonation and deprotonation of amino acids by exchanging these particles with a bath of non-interacting protons in solution, which keeps the chemical potential constant.

In essence, there are two major approaches differing in the treatment of protonation states during the simulation. On the one hand, there is the continuous CpHMD method, first introduced by Brooks and co-workers[101], which relies on λ-dynamics[104] to include the protonation coordinate, λ, of a fictitious mass in the potential energy function and to propagate it during the integration of the molecular system. The protonation coordinate fluctuates between 0 and 1 during the simulation, defining the protonated and deprotonated states at the endpoints, respectively, which in turn modulates the non-bonding potential energy by linear interpolation of the partial charges and van der Waals interactions between the protonated and deprotonated states. Some techniques include a cut-off at 0.2 and 0.8 to accept the protonation state change, although intermediate values of λ lead to physically meaningless transient states that should be discarded in the analysis. Among the most prominent published techniques of continuous CpHMD, there are the implicit[103,105], hybrid[106] and explicit[107,108] solvent methods implemented in CHARMM and AMBER, or the multi-site λ dynamics approach[107,109], CpHMD$^{MS\lambda D}$, implemented in CHARMM and GROMACS. An outstanding advantage of this method is the rapid convergence of the protonation states and pKa, which is even faster when used in conjunction with the enhanced-sampling replica exchange MD (REMD) techniques.

On the other hand, the second approach provides an explicit, physically meaningful description of the protonation states of the titratable residues in the so-called CpHMD method with discrete protonation states (or discrete CpHMD for short)[98–100,102,110,111]. This method is based on propagating the trajectory of the molecular system during the course of a MD simulation, occasionally stopping the conformational sampling to propose new protonation states according to the electrochemical environment of each titratable amino acid. The protonation state change attempt is controlled by the stochastic Monte Carlo criterion, for which a transition free energy between the current protonation state and the proposed protonation state is calculated in Eq. 21 to determine whether the criterion is accepted or not.

$$\Delta G = k_b T(pH - pk_{a,ref}) ln10 + \Delta G_{elec} - \Delta G_{elec,ref} \qquad \textbf{Eq. 2.21}$$

where k_b is the Boltzmann constant, T is the temperature, pH is the pH of the solvent, $pk_{a,ref}$ is the pKa of the reference compound, ΔG_{elec} is the transition free energy associated with the electrostatic interactions between the proposed and the current protonation states, and $\Delta G_{elec,ref}$ is the electrostatic transition free energy between the two states, current and proposed, of the reference compound.

Normally the reference compound is the ionisable amino acid within a dipeptide (i.e., the amino acid with capping groups). The electrostatic transition free energy is obtained as the difference between the electrostatic potentials of the respective partial charge distributions in the titratable amino acid according to the proposed and the current protonation states. Thus, if the protonation state change attempt is accepted, the titratable amino acid is updated with the proposed protonation state. Then, whether the protonation state change is accepted or rejected, the trajectory continues to propagate until it is stopped again, and a new protonation state is proposed. This protocol is repeated until both the conformational and the protonation state space of the molecular system have been sampled.

This method was first proposed by Baptista and co-workers[98,99] using the Poisson-Boltzmann continuum electrostatics as implicit solvent for both conformational and protonation state sampling. Later versions of the stochastic method focused on improving the description or efficiency of the solvent model, such that the generalized-Born method was introduced as a faster alternative for the treatment of the implicit solvent[102,110], or by including explicit water molecules only in the conformational sampling[111,112], which were implemented in AMBER or GROMACS. However, a purely explicit solvent description in the CpHMD method is too costly. It would require proposing the new protonation state and relaxing the water molecules around it for each titratable amino acid before accepting or rejecting the proposal. Otherwise, if the orientation of the water molecules with respect to the current protonation states is maintained, the new protonated states will have very high energy barriers, around ~100 kcal/mol, and therefore the stochastic criterion will almost never accept them. For this reason, implicit solvation is used to sample the protonation states, as it avoids the explicit water problem at the cost of sacrificing accuracy in the modelling water and ions, which can be important in some protonation

state changes. In the next section we will describe in more detail the implementation of CpHMD with implicit and hybrid solvation in the AMBER package.

Both constant pH methods, with either continuous or discrete protonation states, face several challenges. First, the lack of force fields or program architectures tailored for constant pH simulations is probably responsible for inaccuracies in protein conformational sampling[113–116]. Second, the treatment of the net charge fluctuations during the simulation is not trivial. Some techniques have offered ingenious solutions, such as charge compensation through the introduction of co-ions or titratable waters[117–119] or the addition of a background plasma to neutralise the net charge of the system. Third, the computational cost of pKa and protonation state convergence coupled with conformational sampling. This challenge is particularly relevant for hybrid solvent discrete CpHMD simulations, as they require solvent relaxation after accepting a protonation state change. In response, GPU implementations offer an increase in the computational efficiency, and conformational and protonation state sampling has been improved by enhanced-sampling techniques such as pH-based REMD[106,110,111,120] or the reduction of the atomistic resolution with CG models[121,122]. For further information, the reader is referred to the recent review on the current state of constant pH methods by Martins De Oliveira et al.[123].

2.5.1. Constant pH Molecular Dynamics with Discrete Protonation States implemented in AMBER

Most of the constant pH simulations in this thesis are performed by the constant pH Molecular Dynamics method with discrete protonation states implemented in AMBER using the implicit GB solvent[102] or the hybrid explicit/GB version[111]. Therefore, we will dedicate this section to describe the requirements, steps and protocol followed by this implementation for a proper comprehension.

First, the molecular system requires specific residues and the definition of the protonation states and the corresponding partial charge distributions for an accurate representation of the titratable amino acids during the simulation. In addition, the reference pKa described by Bashford et al. and Kyte in Table 2 and the reference electrostatic energies calculated through the dipeptide of the titratable amino acids are required to estimate the transition free energies during the application of the Metropolis Monte Carlo criterion. These parameters are provided by AMBER for several GB models

Residue	$pk_{a,ref}$
Asp	4.0
Glu	4.4
His (δ-state)	6.5
His (ε-state)	7.1
Tyr	9.6
Lys	10.4

Table 2. Titratable residues and pKa values described by Bashford et al.[124] and Kyte[125] used in the CpHMD implementation in the AMBER software.

at ionic strengths of 0.1M and, if other conditions are required, can be calculated by an internal tool. After preparation of the titratable residues and the selection of the GB model, the CpHMD simulation runs the trajectory until a user-defined number of MD steps, τ_{MD}, is reached. The simulation is then paused, and the protocol for protonation state change attempt is executed. There are some subtleties in this step depending on the solvation method used:

i) If the entire simulation is performed with implicit solvation, a single titratable amino acid is randomly selected for the protonation state change attempt. The Monte Carlo criterion is applied and, whether it is accepted or rejected, then MD propagation follows.

ii) If the hybrid solvation method is used, the protonation state change attempt is performed on all titratable amino acids in a random order. If at least one protonation change attempt is accepted, the solvent relaxation is carried out and, subsequently, the MD propagation is continued.

We recall that the protonation state change attempt is based on the Metropolis Monte Carlo criterion using to the transition free energy, Eq. 21, which depends on the electrostatic transition energy estimated by the GB model in this AMBER implementation. Thus, the CpHMD simulation proceeds in an iterative protocol of: (1) propagation of the MD trajectory, (2) proposal and calculation of the transition free energy between the proposed and current protonation states, (3) application of the Monte Carlo criterion and (4) updating the protonation states if necessary. If the hybrid solvation model is used, two additional steps are performed: (i) solvent stripping before step 2, and

(4) solvent relaxation after step 4. This protocol is shown the schematic diagram in Figure 6.

A few considerations should be taken into account in the solvation models. First, the effective global period of the protonation state change attempts must be considered in order to correctly sample the protonation state space of each titratable amino acid. While in explicit solvation it is directly defined by the user with τ_{MD}, in implicit solvation this period depends on the number of titratable amino acids and the τ_{MD}. On the other hand, while the implicit model adjusts the solvent instantaneously after accepting a change in protonation state, explicit solvation requires the relaxation of the water molecules around the solute. For this purpose, the solute is frozen in the simulation box and the water molecules perform a user-defined number of MD steps, τ_{rlx}. This τ_{rlx} must be large enough to produce a new distribution of water molecules around the new protonation states that is uncorrelated with the distribution prior to the change. The implementation of the hybrid solvent method shows that 4 fs of water relaxation is suitable, but due to the high computational cost, a minimum relaxation of 200 ps is suggested since it does not show changes in pKa predictions and has minimal impact on the solvent distributions. Finally, it is worth noting that when two titratable amino acids are close enough, i.e., the titrating hydrogens of the neighbouring residues are within 2 Å of each other, there is a 25% chance that multisite titration will occur. This means that when the protonation state change of one of the amino acids is accepted, the neighbouring amino acid is also changed, so that proton transfer from close titratable sites involved in hydrogen bonding can be captured.

In view of the literature about the constant pH method, the hybrid solvent CpHMD method is generally recommended. Indeed, an implicit solvation model can lead to inaccurate modelling of the molecular system, as pointed out by Machuqueiro and Baptista[126]. In addition, the implementation of the enhanced-sampling REMD technique in discrete CpHMD enables improved conformational and protonation state sampling of molecular systems, thus facilitating a faster and more efficient convergence in the simulations.

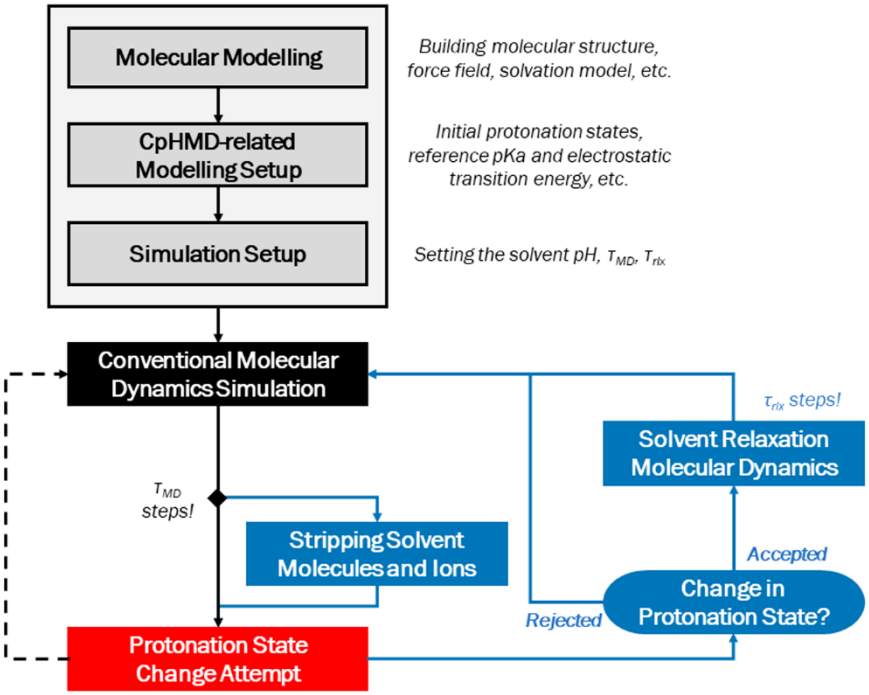


Figure 6. Workflow of the CpHMD method with implicit solvation model (dashed black lines) and explicit water molecules (solid blue lines) implemented in AMBER.

2.5.2. pH-based Replica Exchange Molecular Dynamics

The pH-based replica exchange Molecular Dynamics (pH-REMD) method[106,110,120] is an expanded ensemble technique that allows sampling of the conformational and protonation state space as well as the thermodynamic state space –in this case, the chemical potential of the protons, i.e., the solvent pH– of a molecular system during the simulation. In brief, *N* replicas of the molecular system are ordered according to the solvent pH and, after a certain number of CpHMD steps, an exchange of replicas is attempted. If accepted, the replicas are swapped between the thermodynamic states, i.e., solvent pH, and the CpHMD simulation is then continued. For the replica exchange attempt, the Metropolis Monte Carlo criterion is applied, in this case, using a transition probability, $P_{i\to j}$, defined as:

$$P_{i\to j} = min\{1, exp[ln10(N_i - N_j)(pH_i - pH_j)\} \qquad \textbf{Eq. 2.22}$$

where N_i and pH_i are the number of titrating protons and the solvent pH of replica i, respectively.

Following the protocol shown in Figure 7, the conformational space is sampled continuously, and the thermodynamic state space and protonation states are sampled discretely. The implementation of pH-REMD in AMBER[111] includes some minor refinements to improve performance, such as recommending an even number of replicas and executing the exchange attempts periodically after a user-defined number of MD steps. In addition, the replica exchange is performed between nearest neighbours according to solvent pH, which are alternated during the simulation, so that a replica cannot be exchanged with the same neighbour in succession.

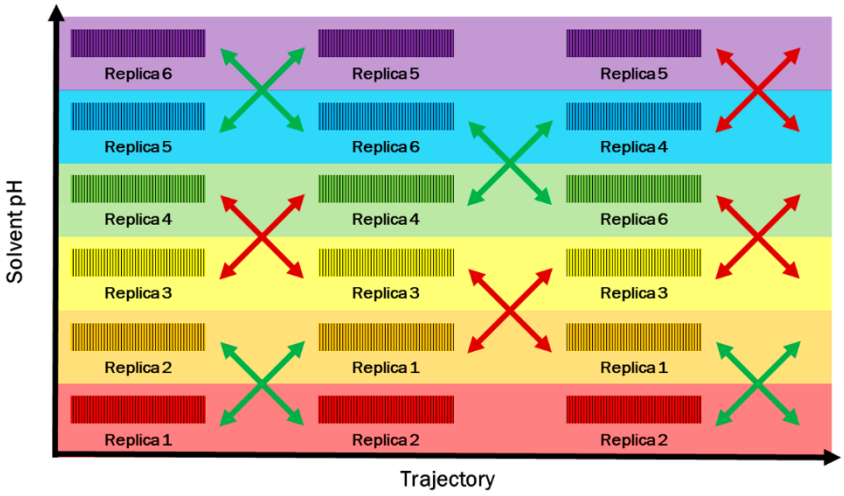


Figure 7. Schematic representation of the pH-REMD protocol. The vertical black lines are the protonation state change attempts and the crossed lines are the accepted (green) and rejected (red) replica exchange attempts. The background reflects the solvent pH conditions from the colour scale of the pH indicator paper.

2.6. Simulation Analysis Techniques

The following section summarises the techniques used to analyse the structural properties and conformational sampling of the trajectories. On the one hand, Ramachandran maps and NMR chemical shift prediction allow a local inspection of the amino acid conformations within macromolecules. On the other hand, the calculation of R_g and SAXS intensities provides a global picture of protein compaction, and the Principal Component Analysis (PCA) facilitates the identification of the conformational space sampled during the simulation. The combination of local and global analysis, combined with the estimation of the secondary structure propensities and comparison with experimental SAXS and NMR observables, has served both to evaluate the ability

of the methods and force fields to capture the conformational space and to provide insight into the preferred conformations of the molecular systems addressed in this thesis.

2.6.1. Ramachandran Maps

Ramachandran maps are the representation of the combination of φ and ψ dihedral angles of the backbone amino acids of a protein in a 2-dimensional map. These φ and ψ angles are defined by specific backbone amino acid atoms, such that φ is the torsion angle between the $N_{(i-1)}$, $C_{(i)}$, $C_{\alpha(i)}$ and $N_{(i)}$ atoms and ψ is the torsion angle between the $C_{(i)}$, $C_{\alpha(i)}$, $N_{(i)}$ and $C_{(i+1)}$ atoms as shown in Figure 8a. From the combination of both angles, it is possible to define the geometry of adjacent amino acids and thus provide conformational and structural information about the amino acids within proteins. This 2D dihedral angle map was originally developed by G. N. Ramachandran et al. in 1963[127], and the regions of the map shown in Figure 8b were identified from hard-sphere calculations in 1968[128].

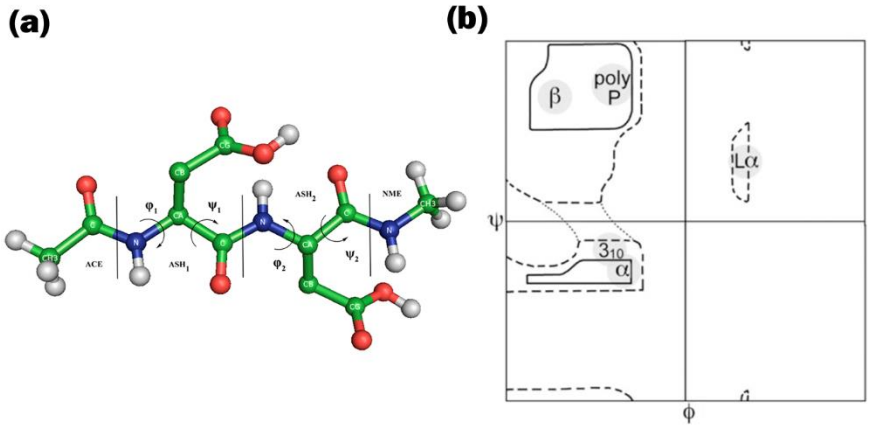


Figure 8. (a) Representation of the backbone dihedral angles φ and ψ of a capped tripeptide. (b) Original Ramachandran map with the relevant conformational regions according to the sphere models (hard-sphere in solid lines, reduced-sphere in dashed lines or relaxed-tau in dotted lines)[128].

As a result of the expansion of high-resolution three-dimensional structure characterisation and its storage in databases, it has been possible to identify the most common regions in the majority of amino acids. These regions are α for α-helix, Lα for left-handed helix, β for β-sheet and ppII for polyproline. Glycine and proline have different patterns on the Ramachandran map. The former is more extended due to the absence of the C_β atom and therefore the dihedral space is less restricted. The latter is

much more compact because the amino acid forms a 5-membered-ring involving the C_α and N atoms of the backbone, which causes a restriction of the space. In addition, the amino acids preceding the proline in the peptide chain are also restricted. There are other schemes for identifying the regions of the Ramachandran map, such as the ABEGO system[129]. Furthermore, as seen in the map, much of the dihedral distribution space is empty due to the intramolecular steric hindrances.

The secondary structure propensities, which are merely the result of the repetition of specific conformations determined by the pattern of φ and ψ dihedral angles in the backbone, can be distinguished in the Ramachandran maps, making them extremely useful for the study of the accessible regions and the energetically favoured conformations based on the identification of dihedral angles, both in *in silico* studies and in structure validation processes.

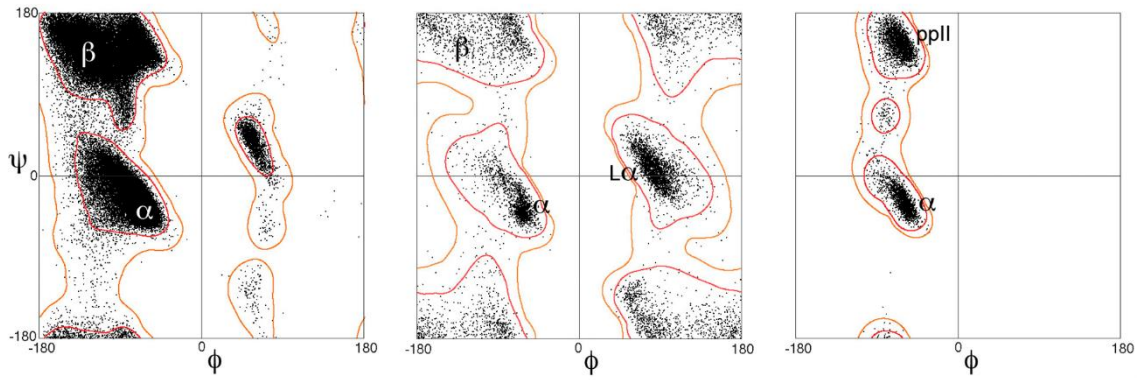


Figure 9. Ramachandra map of (left) most amino acids, (middle) glycine and (right) proline from the work of Lovell et al.[130].

2.6.2. Radius of Gyration

The radius of gyration is a measure of the distribution of atoms with respect to the centre of mass of a molecule. In structural biophysics, it is an important parameter related to the size or overall compactness of a protein, providing useful information about the possible conformation of the systems. In particular, the radius of gyration is the root-mean-square of the distance of N atoms of a macromolecule from its centre of mass, and is calculated as follows:

$$R_g = \sqrt{\left(\frac{\sum_i \|\boldsymbol{r_i}\|^2 m_i}{\sum_i m_i}\right)} \qquad \textbf{Eq. 2.23}$$

where $\boldsymbol{r_i}$ is the distance of atom i from the centre of mass and m_i is the mass of atom i.

The dispersion of values and variance is a good resource for estimating the rigidity or flexibility of a protein, which is of great advantage for the inherent flexibility of IDPs. In addition, R_g can be measured experimentally using the Guinier plot from intensities observed in SAXS. Therefore, conformational ensembles generated during trajectories can be compared with experimental data, giving confidence to simulations performed with a given force field, water model or sampling method. Furthermore, R_g is also indicative of conformational changes in proteins, either due to folding/unfolding events, protein-ligand binding, protein-protein interactions, or others.

2.6.3. Dictionary of Protein Secondary Structure

The Dictionary of Protein Secondary Structure method[131], or DSSP, whose abbreviation originates from the implementation of the "Define Secondary Structure of Proteins" algorithm in the Pascal program, is based on the identification of intramolecular hydrogen bond patterns to assign secondary structure propensities to the amino acids of a protein. To identify these hydrogen bonds of a protein structure with atomic resolution, the partial charges of the carbon and oxygen atoms of the carbonyl group (C=O) and the nitrogen and hydrogen atoms of the amide group (N-H) are first assigned. Next, using a strictly electrostatic model, the electrostatic energy of the atoms involved in the H-bond, ΔG_{elec}^{Hbond}, is computed.

$$\Delta G_{elec}^{Hbond} = q_1 q_2 \cdot \left(\frac{1}{\boldsymbol{r_{ON}}} + \frac{1}{\boldsymbol{r_{CH}}} - \frac{1}{\boldsymbol{r_{OH}}} - \frac{1}{\boldsymbol{r_{CN}}}\right) \cdot f \qquad \textbf{Eq. 2.24}$$

where q_1 and q_2 are -0.42e and +0.20e, respectively, e is the unit electron charge, $\boldsymbol{r_{AB}}$ is the interatomic distance between the atoms *A* and *B* in angstroms, and f is a dimensional factor of 332 kcal/mol.

For this model, an electrostatic energy cut-off of ΔG_{elec}^{Hbond} < -0.5 kcal/mol is set to define a hydrogen bond, although a good binding energy is around 3kcal/mol. Once the hydrogen bond patterns within a molecular structure have been defined, the secondary structure is assigned to each amino acid using the DSSP classification labels.

- The symbols G, H and I encode the 3_{10}-helix, α-helix, and π-helix secondary structures, forming helices with a repeating sequence of hydrogen bonds every three, four, and five residues, respectively.
- The symbols E and B code for two types of β-sheets. The letter E is assigned to β-bulges or extended conformations of anti- or parallel β-sheets. The letter

B corresponds to isolated β-bridges when a single-pair β-sheet bond conformation is formed. If a β-sheet is not long enough, the amino acids are labelled with the letter B.

- The symbol T indicates a hydrogen bond turn. If a 3_{10}-helix, α-helix, or π-helix is not long enough, it is given the letter T.
- The symbol S corresponds to the bends calculated from the angle between the vectors $\boldsymbol{r}_{C_{\alpha,i}C_{\alpha,i+2}}$ and $\boldsymbol{r}_{C_{\alpha,i-2}C_{\alpha,i}}$, with $C_{\alpha,i}$ as the C_α atom at position i of the amino acid chain, so that if an angle is greater than 70º, the amino acids are labelled S. This is the only class that is not based on the electrostatic model, Eq. 24.
- The symbol C (or blank) is used for the amino acids that cannot be classified in the previous labels.

2.6.4. Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method widely applied in data-intensive problems in order to capture as much information as possible in the smallest number of independent variables, also called Principal Components (PCs). The use of this technique implies a loss of precision, at the discretion of the user, in exchange for the simplification, visualisation and analysis of the data. The method usually consists of the following steps:

1. In order to ensure an equal contribution of the initial variables in the PCA, it is crucial to perform a standardisation of these variables to avoid bias due to the sensitivity of the method.
2. Normally, a covariance matrix of *N* dimensions is constructed, where *N* is the number of variables in our data set. The covariance between the variables is introduced to reduce the correlation and to eliminate redundant information contained in them. The resultant matrix is symmetric.

$$\left(\begin{pmatrix} covar(1,1) & \cdots & covar(1,N) \\ \vdots & \ddots & \vdots \\ covar(N,1) & \cdots & covar(N,N) \end{pmatrix}\right) \qquad \textbf{Eq. 2.25}$$

3. By diagonalising the symmetric matrix, we generate the eigenvectors (or PCs) and eigenvalues of the covariance of the data set. This involves calculating new orthogonal vectors, i.e., a set of variables constructed from the linear combination

of the original variables. The advantage of these new variables is that they contain, in descending order, the largest (remaining) variance of the original data. This variance is captured in the eigenvalues and therefore the eigenvalue of the first PC is the largest one, while the eigenvalue of the second PC is the second largest one, and so on up to the *N* principal components generated.

4. Once the *N* principal components have been computed, the user must choose which PCs to use for dimensionality reduction. Typically, the first 2-3 PCs are selected as they can graphically represent the greatest amount of information through two- or three-dimensional plots. A transformation matrix is then created from the selected eigenvectors.
5. Finally, the original data is projected using the transformation matrix into the new space created by the PCs. The points in the PCA map capture more information than the data in the original variables. In fact, the percentage of covariance data included in each PC can be calculated by $eigenvalue_i / \sum_i eigenvalue_i$.

This dimensionality reduction is very useful in protein simulations because there are $3N - 6$ coordinates, where N is the number of atoms, and a large amount of data is generated in each trajectory frame. When PCA is applied to the covariance of the atomic positions of molecular systems, the method is also referred to as essential dynamics[132,133]. Despite the advantages of PCA, some effort is usually required to further reduce the amount of data, such as selecting the position of the C_α atoms or superimposing the conformations to a reference structure to remove translations and rotations of the macromolecule. In this way, the variance within the first PCs can be more easily captured and, consequently, the protein dynamics and motions can be better represented in the PCA plot. In addition, several papers use PCA to generate a population-based energy surface within the selected PCs, so that the conformational preferences, i.e., energy minima, sampled during the simulation can be identified. PCA thus facilitates the interpretation of the simulations to identify the relevant biological events of the proteins.

2.6.5. Small-Angle X-ray Scattering

Small-angle X-ray scattering (SAXS) is an experimental technique based on the measurement of elastic scattering at small angles to determine density differences at the nanoscale. It can provide insight into the dimensionality, size and shape of materials (characterisation of nanopores or distances of ordered systems) and biomolecules

(folded/unfolded state, oligomer formation, state transitions, etc.), depending on the angle range. There are SAXS techniques that can be performed in aqueous conditions, thus offering structural data of proteins in solution, which is extremely valuable to capture the conformational transitions of proteins and, in particular, in the IDPs. In addition, given the limited range of experimental techniques available to characterise IDPs due to the challenges of X-ray diffraction and NMR spectroscopy, SAXS becomes a powerful technique for validating the simulation models and conformational ensembles against experimental observables.

SAXS experiments provide the scattering intensity, $I(\boldsymbol{q})$, with respect to the scattering vector, $\boldsymbol{q}$, which is the gradual variation of the scattering angle 2θ. The scattering intensity is composed of the shape factor, $F(\boldsymbol{q})$, and the structure factor, $S(\boldsymbol{q})$, which give details of the shape and the interaction between the particles of the sample, respectively, such that $I(\boldsymbol{q}) = F(\boldsymbol{q})S(\boldsymbol{q})$. From the plot of $I(\boldsymbol{q})$ and $\boldsymbol{q}$, the radius of gyration of the macromolecule can be calculated. In addition, the pairwise distance distribution function, $p(\boldsymbol{r})$, can also be extracted from the scattering curve, thereby obtaining a measure of the interatomic distances, the shape and the degree of compactness of a macromolecule. Furthermore, the plot of $(\boldsymbol{q}Rg)I(\boldsymbol{q})$ vs $\boldsymbol{q}Rg$, called the Kratky plot, allows the identification of globular conformations of a protein, such that a maximum is found in $\boldsymbol{q}R_g = \sqrt{3}$ regardless of the size of the protein. Therefore, these representations show that this scattering technique is very useful for the global characterisation of macromolecules in solution.

Since this technique is highly applicable to proteins, especially for IDPs, several SAXS intensity profile prediction software have been developed to interpret and validate the conformational ensembles generated from atomistic simulations in contrast to the experimental data. The main feature among the software is the description of the solvation model, distinguishing between implicit solvation (CRYSOL[134], PLUMED[135], FoXS[136]) and explicit water models (WAXSiS[137], Capriqorn[138], 3D-RISM[139]). Other improvements have been implemented, including the fitting of experimental and theoretical intensity profiles, the addition of flexibility to the macromolecule, the correct representation of the conformational state ensembles, etc. For more information, we recommend to the reader this paper which evaluates a variety of SAXS intensity prediction software[140].

2.6.6. Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy in proteins is based on the measurement of the absorption of radio frequency signals generated by a magnetic field in the atomic nuclei of a sample to determine the chemical shifts (CS). In other words, the CS are the resonance frequencies of the atomic nuclei experienced when subjected to a magnetic field, which depend on the local chemical environment of each nucleus within the protein. When NMR is applied at nanometre scale for atom detection, from which a map of atomic bonds, distances between atomic nuclei, and even the dynamics of the proteins can be obtained. Solution NMR spectroscopy can determine CS during protein conformational transitions. It is of particular interest for IDPs which are highly mobile and therefore many conformational transitions are expected within the vast conformational space of these biomolecules. This, together with the difficulty of characterising IDPs by of X-ray crystallography due to their flexibility, makes NMR spectroscopy even more valuable in providing local structural data on IDPs. Indeed, several IDPs have been determined by NMR spectroscopy, but there are some drawbacks related to the concentrations required for characterisation that still need to be addressed. On the other hand, solid-state NMR spectroscopy allows the characterisation of these IDPs when they are in non-soluble states, such as membranes or protein aggregates, which can provide important insights into these complex structures.

Therefore, CS can be extremely useful for the characterisation of protein structures and, as far as *in silico* simulations are concerned, they can also be used as reference values to assess the accuracy of sampling methods, force fields or water models. In fact, many studies have relied solely on the protein backbone CS to validate the conformational ensembles obtained from simulations. For this purpose, several methods have been developed to predict CS from atomic coordinates, which have been improved in recent years. Currently, the methods developed for the CS prediction are based on: (i) sequence homology, (ii) empirical equations derived from classical physics and experimental data, (iii) quantum chemistry, (iv) structure-chemical shifts relation tables, or (v) a combination of the above methods, called hybrid methods. In this thesis, we will use the SPARTA+ program[141], a hybrid method built from an artificial neural network algorithm based on semi-classical equations (e.g., dihedral angles, interactions and backbone flexibility, etc.) and triplet sequence homology assignment from a large database of 580 proteins.

2.7. Bibliography

1. Kettering, F., Shutts, L. W. & Andrews, D. H. A Representation of the Dynamic Properties of Molecules by Mechanical Models. *Phys Rev* **36**, 531–543 (1930).
2. Hill, T. L. On Steric Effects. *J Chem Phys* **14**, 465 (1946).
3. Dostrovsky, I., Hughes, E. D. & Ingold, C. K. 50. Mechanism of substitution at a saturated carbon atom. Part XXXII. The role of steric hindrance. (Section G) magnitude of steric effects, range of occurrence of steric and polar effects, and place of the wagner rearrangement in nucleophilic substitution and elimination. *J Chem Soc* **0**, 173–194 (1946).
4. Alder, B. J. & Wainwright, T. E. Phase Transition for a Hard Sphere System. *J Chem Phys* **27**, 1208–1209 (1957).
5. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
6. Duan, Y. & Kollman, P. A. Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science* **282**, 740–744 (1998).
7. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys Rev* **159**, 98 (1967).
8. Hockney, R. W. & Eastwood, J. W. *Computer Simulation Using Particles* (CRC Press, 1988).
9. Swope, W. C., Andersen, H. C. & Berens, P. H. 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**, 637 (1982).
10. Beeman, D. Some multistep methods for use in molecular dynamics calculations. *J Comput Phys* **20**, 130–139 (1976).
11. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).
12. Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J Comput Phys* **52**, 24–34 (1983).
13. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for molecular simulations. *J Comput Chem* **18**, 1463–1472 (1997).
14. Tuckermar, M., Berne, B. J. & Martyna, G. J. Reversible multiple time scale molecular dynamics. *J Chem Phys* **97**, 1990–2001 (1992).
15. Brooks, C. L., Pettitt, B. M. & Karplus, M. Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids. *J Chem Phys* **83**, 5897–5908 (1985).
16. Tironi, I. G., Sperb, R., Smith, P. E. & van Gunsteren, W. F. A generalized reaction field method for molecular dynamics simulations. *J Chem Phys* **102**, 5451–5459 (1995).
17. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089–10092 (1993).
18. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann Phys* **369**, 253–287 (1921).

19. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–3690 (1984).

20. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 014101 (2007).

21. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J Chem Phys* **81**, 511–519 (1984).

22. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys* **52**, 255–268 (1984).

23. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* **31**, 1695–1697 (1985).

24. Schneider, T. & Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys Rev B* **17**, 1302 (1978).

25. Brown, D. & Clarke, J. H. R. A loose-coupling, constant-pressure, molecular dynamics algorithm for use in the modelling of polymer materials. *Comput Phys Commun* **62**, 360–369 (1991).

26. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **52**, 7182–7190 (1981).

27. Nosé, S. & Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol Phys* **50**, 1055–1076 (1983).

28. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J Chem Phys* **103**, 4613–4621 (1995).

29. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *J Chem Phys* **101**, 4177–4189 (1994).

30. Gilson, M. K., Sharp, K. A. & Honig, B. H. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J Comput Chem* **9**, 327–335 (1988).

31. Honig, B. & Nicholls, A. Classical Electrostatics in Biology and Chemistry. *Science* **268**, 1144–1149 (1995).

32. Clark Still, W., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **112**, 6127–6129 (1990).

33. Schaefer, M. & Karplus, M. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J Chem Phys* **100**, 1578–1599 (1996).

34. Bashford, D. & Case, D. A. Generalized Born Models of Macromolecular Solvation Effects. *Annu Rev Phys Chem* **51**, 129–152 (2000).

35. Sitkoff, D., Sharp, K. A. & Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J Phys Chem* **98**, 1978–1988 (1994).

36. Weiser, J., Shenkin, P. S. & Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J Comput Chem* **20**, 217–230 (1999).

37. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1998).

38. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. Interaction Models for Water in Relation to Protein Hydration. *Intermolecular Forces. The Jerusalem Symposia on Quantum Chemistry and Biochemistry* **14**, 331–342 (1981)

39. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys* **112**, 8910–8922 (2000).

40. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J Phys Chem* **91**, 6269–6271 (1987).

41. Price, D. J. & Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *J Chem Phys* **121**, 10096–10103 (2004).

42. Horn, H. W. *et al.* Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys* **120**, 9665–9678 (2004).

43. Wu, Y., Tepper, H. L. & Voth, G. A. Flexible simple point-charge water model with improved liquid-state properties. *J Chem Phys* **124**, 024503 (2006).

44. Schmitt, U. W. & Voth, G. A. The computer simulation of proton transport in water. *J Chem Phys* **111**, 9361–9381 (1999).

45. Lamoureux, G., Harder, E., Vorobyov, I. v., Roux, B. & MacKerell, A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem Phys Lett* **418**, 245–249 (2006).

46. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* **119**, 5113–5123 (2015).

47. Khalak, Y., Baumeier, B. & Karttunen, M. Improved general-purpose five-point model for water: TIP5P/2018. *J Chem Phys* **149**, 224507 (2018).

48. Abascal, J. L. F. & Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys* **123**, 234505 (2005).

49. Izadi, S., Anandakrishnan, R. & Onufriev, A. v. Building water models: A different approach. *J Phys Chem Lett* **5**, 3863–3871 (2014).

50. Izadi, S. & Onufriev, A. Accuracy limit of rigid 3-point water models. *J Chem Phys* **145**, 074501 (2016).

51. Kadaoluwa Pathirannahalage, S. P. *et al.* Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations. *J Chem Inf Model* **61**, 4521–4536 (2021).

52. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).

53. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).

54. van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).

55. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. in *SC'06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing* 84 (ACM Press, 2006).

56. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781–1802 (2005).

57. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**, 11225–11236 (1996).

58. Oostenbrink, C., Villa, A., Mark, A. E. & van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem* **25**, 1656–1676 (2004).

59. Mcgibbon, R. T. *et al.* Computational Tools MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528–1532 (2015).

60. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

61. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

62. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

63. Schrödinger, L. & DeLano, W. PyMOL. (2020).

64. Levitt, M. Birth and Future of Multiscale Modeling for Macromolecular Systems (Nobel Lecture). *Angew Chem* **53**, 10006–10018 (2014).

65. Liwo, A. *et al.* A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions. *J Mol Model* **20**, 1–15 (2014).

66. Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* **51**, 350–370 (2004).

67. Gopal, S. M., Mukherjee, S., Cheng, Y. M. & Feig, M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins* **78**, 1266–1281 (2010).

68. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein Structure Prediction Using Rosetta. *Methods Enzymol* **383**, 66–93 (2004).

69. Marrink, S. J., de Vries, A. H. & Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *J Phys Chem B* **108**, 750–760 (2003).

70. de Jong, D. H. *et al.* Improved parameters for the martini coarse-grained protein force field. *J Chem Theory Comput* **9**, 687–697 (2013).

71. Bereau, T. & Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J Chem Phys* **130**, 235106 (2009).

72. Sterpone, F. *et al.* The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chem Soc Rev* **43**, 4871–4893 (2014).

73. Singh, N. & Li, W. Recent advances in coarse-grained models for biomolecules and their applications. *Int J Mol Sci* **20**, 3774 (2019).

74. Darre, L., Rodrigo Machado, M., Febe Brandner, A., Carlos Gonzaíez, H. & Pantano, S. SIRAH: A Structurally Unbiased Coarse-Grained Force Field for Proteins with Aqueous Solvation and Long-Range Electrostatics. *J Chem Theory Comput* **11**, 28 (2015).

75. Machado, M. R. *et al.* The SIRAH 2.0 Force Field: Altius, Fortius, Citius. *J Chem Theory Comput* **15**, 2719–2733 (2019).

76. Weik, F. *et al.* ESPResSo 4.0 – an extensible software package for simulating soft matter systems. *Eur Phys J Spec Top* **227**, 1789–1816 (2019).

77. Darré, L., Machado, M. R., Dans, P. D., Herrera, F. E. & Pantano, S. Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *J Chem Theory Comput* **6**, 3793–3807 (2010).

78. Fatafta, H., Samantray, S., Sayyed-Ahmad, A., Coskuner-Weber, O. & Strodel, B. Molecular simulations of IDPs: From ensemble generation to IDP interactions leading to disorder-to-order transitions. *Prog Mol Biol Transl Sci* **183**, 135–185 (2021).

79. Mu, J., Liu, H., Zhang, J., Luo, R. & Chen, H.-F. Recent Force Field Strategies for Intrinsically Disordered Proteins. *J Chem Inf Model* **61**, 1037–1047 (2021).

80. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* **113**, 9004–9015 (2009).

81. Piana, S., Lindorff-Larsen, K., Shaw, D. E., Research, D E Shaw & York, N. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys J* **100**, L47–L49 (2011).

82. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J Chem Theory Comput* **11**, 3499–3509 (2015).

83. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J Chem Theory Comput* **12**, 281–296 (2016).

84. Jiang, F., Zhou, C. Y. & Wu, Y. D. Residue-specific force field based on the protein coil library. RSFF1: Modification of OPLS-AA/L. *J Phys Chem B* **118**, 6983–6998 (2014).

85. Zhou, C. Y., Jiang, F. & Wu, Y. D. Residue-specific force field based on protein coil library. RSFF2: Modification of AMBER ff99SB. *J Phys Chem B* **119**, 1035–1047 (2015).

86. MacKerell, A. D., Feig, M. & Brooks, C. L. Improved Treatment of the Protein Backbone in Empirical Force Fields. *J Am Chem Soc* **126**, 698–699 (2004).

87. Mackerell, A. D., Feig, M. & Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation. *J Comput Chem* **25**, 1400–1415 (2004).

88. Song, D., Luo, R. & Chen, H. F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **57**, 1166–1178 (2017).

89. Liu, H. *et al.* Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins. *Phys Chem Chem Phys* **21**, 21918–21931 (2019).

90. Yang, S., Liu, H., Zhang, Y., Lu, H. & Chen, H. Residue-Specific Force Field Improving the Sample of Intrinsically Disordered Proteins and Folded Proteins. *J Chem Inf Model* **59**, 4793–4805 (2019).

91. Kang, W., Jiang, F. & Wu, Y. D. Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition. *J Chem Theory Comput* **14**, 4474–4486 (2018).

92. Song, D., Liu, H., Luo, R. & Chen, H.-F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J Chem Inf Model* **60**, 2257–2267 (2020).

93. Demerdash, O. *et al.* Using Small-Angle Scattering Data and Parametric Machine Learning to Optimize Force Field Parameters for Intrinsically Disordered Proteins. *Front Mol Biosci* **6**, 64 (2019).

94. Wu, H., Wolynes, P. G. & Papoian, G. A. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J Phys Chem B* **122**, 11115–11125 (2018).

95. Ren, P. & Ponder, J. W. Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J Comput Chem* **23**, 1497–1506 (2002).

96. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK a Predictions. *J Chem Theory Comput* **7**, 525–537 (2011).

97. Anandakrishnan, R., Aguilar, B. & Onufriev, A. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* **40**, W537–W541 (2012).

98. Nio, A., Baptista, M., Martel, P. J. & Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins* **27**, 523–544 (1997).

99. Baptista, A. M. *et al.* Constant-pH molecular dynamics using stochastic titration. *J Chem Phys* **117**, 4184 (2002).

100. Bürgi, R., Kollman, P. A. & van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins* **47**, 469–480 (2002).

101. Lee, M. S., Salsbury, F. R. & Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738–752 (2004).

102. Mongan, J., Case, D. A. & McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* **25**, 2038–2048 (2004).

103. Khandogin, J. & Brooks, C. L. Constant pH molecular dynamics with proton tautomerism. *Biophys J* **89**, 141–157 (2005).

104. Kong, X. & Brooks, C. L. λ-dynamics: A new approach to free energy calculations. *J Chem Phys* **105**, 2414 (1998).

105. Harris, R. C. & Shen, J. GPU-Accelerated Implementation of Continuous Constant pH Molecular Dynamics in Amber: pKa Predictions with Single-pH Simulations. *J Chem Inf Model* **59**, 4821–4832 (2019).

106. Wallace, J. A. & Shen, J. K. Continuous constant pH molecular dynamics in explicit solvent with pH-based replica exchange. *J Chem Theory Comput* **7**, 2617–2629 (2011).

107. Goh, G. B., Hulbert, B. S., Zhou, H. & Brooks, C. L. Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. *Proteins* **82**, 1319–1331 (2014).

108. Huang, Y., Chen, W., Wallace, J. A. & Shen, J. All-Atom Continuous Constant pH Molecular Dynamics with Particle Mesh Ewald and Titratable Water. *J Chem Theory Comput* **12**, 5411–5421 (2016).

109. Knight, J. L. & Brooks, C. L. Multisite λ Dynamics for Simulated Structure-Activity Relationship Studies. *J Chem Theory Comput* **7**, 2728–2739 (2011).

110. Meng, Y. & Roitberg, A. E. Constant pH replica exchange molecular dynamics in biomolecules using a discrete protonation model. *J Chem Theory Comput* **6**, 1401–1412 (2010).

111. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH replica exchange molecular dynamics in explicit solvent using discrete protonation states: Implementation, testing, and validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).

112. Machuqueiro, M. & Baptista, A. M. Constant-pH Molecular Dynamics with Ionic Strength Effects: Protonation-Conformation Coupling in Decalysine. *J Phys Chem B* **110**, 2627–2933 (2006).

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

114. Paul, T. J., Vilseck, J. Z., Hayes, R. L. & Brooks, C. L. Exploring pH Dependent Host/Guest Binding Affinities. *J Phys Chem B* **124**, 6520–6528 (2020).

115. Privat, C., Madurga, S., Mas, F. & Rubio-Martínez, J. On the Use of the Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides. *Polymers* **13**, 99 (2021).

116. Privat, C., Madurga, S., Mas, F. & Rubio-Martinez, J. Unravelling constant ph molecular dynamics in oligopeptides with explicit solvation model. *Polymers* **13**, 3311 (2021).

117. Chen, W., Wallace, J. A., Yue, Z. & Shen, J. K. Introducing titratable water to all-atom molecular dynamics at constant pH. *Biophys J* **105**, L15–L17 (2013).

118. Huang, Y., Chen, W., Wallace, J. A. & Shen, J. All-Atom Continuous Constant pH Molecular Dynamics With Particle Mesh Ewald and Titratable Water. *J Chem Theory Comput* **12**, 5411–5421 (2016).

119. Donnini, S., Ullmann, R. T., Groenhof, G. & Grubmüller, H. Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer. *J Chem Theory Comput* **12**, 1040–1051 (2016).

120. Swails, J. M. & Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J Chem Theory Comput* **8**, 4393–4404 (2012).

121. Barroso Da Silva, F. L., Sterpone, F. & Derreumaux, P. OPEP6: A New Constant-pH Molecular Dynamics Simulation Scheme with OPEP Coarse-Grained Force Field. *J Chem Theory Comput* **15**, 3875–3888 (2019).

122. Grünewald, F. *et al.* Titratable Martini model for constant pH simulations. *J Chem Phys* **153**, 024118 (2020).

123. Martins De Oliveira, V., Liu, R. & Shen, J. Constant pH molecular dynamics simulations: Current status and recent applications. *Curr Opin Struct Biol* **77**, 102498 (2022).

124. Bashford, D., Case, D. A., Dalvit, C., Tennant, L. & Wright, P. E. Electrostatic Calculations of Side-Chain pKa Values in Myoglobin and Comparison with NMR Data for Histidines. *Biochemistry* **32**, 8045–8056 (1993).

125. Kyte, J. *Structure in Protein Chemistry*. (Garland Publishing, Inc., 1995).

126. Machuqueiro, M. & Baptista, A. M. Is the prediction of pKa values by constant-pH molecular dynamics being hindered by inherited problems? *Proteins* **79**, 3437–3447 (2011).

127. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–99 (1963).

128. Ramachandran, G. N. & Sasisekharan, V. Conformation of Polypeptides and Proteins. *Adv Protein Chem* **23**, 283–437 (1968).

129. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. Automatic Classification and Analysis of αα-Turn Motifs in Proteins. *J Mol Biol* **255**, 235–253 (1996).

130. Lovell, S. C. *et al.* Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins* **50**, 437–450 (2003).

131. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

132. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **17**, 412–425 (1993).

133. Meyer, T. *et al.* Essential Dynamics: A Tool for Efficient Trajectory Compression and Management. *J Chem Theory Comput* **2**, 251–258 (2006).

134. Franke, D. *et al.* ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* **50**, 1212–1225 (2017).

135. Tribello, G. A. *et al.* PLUMED 2: New feathers for an old bird. *Comput Phys Commun* **185**, 604–613 (2014).

136. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* **44**, W424–W429 (2016).

137. Knight, C. J. & Hub, J. S. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Res* **43**, W225–W230 (2015).

138. Köfinger, J. & Hummer, G. Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys Rev E* **87**, 52712 (2013).

139. Nguyen, H. T., Pabit, S. A., Meisburger, S. P., Pollack, L. & Case, D. A. Accurate small and wide angle x-ray scattering profiles from atomic models of proteins and nucleic acids. *J Chem Phys* **141**, 22D508 (2014).

140. Bernetti, M. & Bussi, G. Comparing state-of-the-art approaches to back-calculate SAXS spectra from atomistic molecular dynamics simulations. *Eur Phys J B* **94**, 180 (2021).

141. Shen, Y. & Bax, A. SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* **48**, 13–22 (2010).

Chapter 3

Exploring the Polyaspartic Acid Conformations with Constant pH Simulations and Prediction of pKa through Complexation Isotherms

The protonation state of ionisable amino acids can play a paramount role in some biochemical systems, especially when protein binding molecules and/or enzymatic mechanisms are involved. Studying the behaviour of these amino acids, from their charge to the conformations they can adopt, can be crucial to understand these processes. Although experimental techniques can provide valuable insights into these biological functions, these methods are beyond the scope of an atomistic description. Molecular Dynamics (MD) simulation methods help to unravel this puzzle through molecular modelling, which has proven to shed light on hidden mechanisms as a fundamental first step in drug development, protein design, or learn about the structure-function relationship of proteins[1]. While the MD method has demonstrated several successes in the field of computational biochemistry, there are still major challenges to overcome related to the simulation time scale and capturing biological events, the development of accurate general force fields, and so on. These include the dynamic change in the protonation state of the ionisable amino acids and the effect of the solvent pH on them. Fortunately, MD simulations at constant pH have become a popular practice in recent years to overcome this important lack of description in the modelling of molecular systems[2–4]. However, the main application so far has been to estimate the effective pKa of those ionisable amino acids buried within protein structures, although first successful papers elucidating the interaction mechanisms of some proteins have also been published[5–8].

The application of these *constant pH* methods goes beyond the prediction of the pKa values. There is a strong coupling between environmental pH and structure, which consequently modulates the function of the pH-responsive proteins. To discuss and

highlight this issue, this chapter makes a first contact with simulations at constant pH to study its ability to modulate the conformational space of biomolecules. Specifically, it is performed by means of the Constant pH Molecular Dynamics (CpHMD) method[9–11]. Within this approach, there are two branches that are distinguished by the criterion of dynamically changing the protonation state. On the one hand, the continuous CpHMD method[12–14] describes the protonation state of each titratable group by including a λ-protonation coordinate in the Hamiltonian. This λ coordinate in turn allows the description of the protonated or deprotonated states when the λ is above or below a certain value (e.g., when $\lambda < 0.2$, the titratable group is protonated and when $\lambda > 0.8$, it loses the proton). Moreover, this approach allows to describe intermediate states between the fully protonated or fully deprotonated states, which are treated as a proportional contribution of the electrostatic interactions of each of the protonation states. On the other hand, the CpHMD method with discrete protonation states[15–17] performs an exploration of the conformational space according to the principles of MD, while the protonation states are explored with a Monte Carlo (MC) and Continuum Electrics algorithm. Every number of steps along the trajectory, the MD simulation stops and attempts a protonation state change using an MC criterion which is subject to the electrostatic interactions of the titratable amino acid environment.

In this case, we have chosen the discrete CpHMD method to evaluate the capacity of exploring the conformational space of polyaspartic acid as a test model. The main motivation of this selection is because of the explicit and meaningful description of the protonation states during the entire trajectory, which allows a more accurate study of the conformational space at the expense of longer simulation times for a representative sampling of the protonation states. Therefore, we first performed an assessment of the CpHMD method implemented in AMBER by simulating the fully protonated and deprotonated polyaspartic acid under implicit and explicit solvation conditions and comparing them with conventional MD (CMD) simulations under identical conditions. After 2 μs length of simulation for each method, protonation state and solvent model (i.e., 16 μs of polyaspartic acid simulation), the analysis of the conformational space shows serious inconsistencies in the implicit solvent simulations. On the other hand, the explicit solvent simulation partially overcomes the reported limitation of the CpHMD method and succeeds in reproducing structural properties such as the radius of gyration (R_g) and the secondary structure propensity fractions (fpSS) observed in CMD simulations.

After reporting the shortcomings during the validation of the method, we discarded the implicit solvation model and proceeded with pH Replica Exchange Molecular Dynamics (pH-REMD) simulations of polyaspartic acid using only the explicit solvation model[18–20]. The REMD method belongs to the wide range of enhanced sampling techniques (e.g., metadynamics, temperature annealing, accelerated MD, etc.) that accelerate the sampling, in this case of the protonation state space, to reduce the computational cost. In fact, the REMD approach can be applied to several properties of the system, such as temperature, Hamiltonian, pH, redox potential or even perform multidimensional REMD combining these properties, thus enabling simulation over extensive ranges of the property(s) of interest. In brief, the pH-REMD approach consists of running a series of parallel constant pH simulations at several pH values within a pH range. As the simulation progresses, the replicas are swapped with the neighbouring pH values by an exchange probability, which is defined in Eq. 2.22. This method has proven to estimate accurate pk_a values of several amino acids with experimentally determined pk_a[21]. However, the potential of this method the conformational sampling of pH-responsive biomolecules has not been thoroughly investigated. Therefore, in this chapter we also perform a deep conformational exploration of the polyaspartic acid peptide by applying pH-REMD to 16 replicas in the pH range = [1.0, 8.5]. This provides a first insight into the effect of solvent pH in modulating the conformation. In parallel, we exploit the ability of the constant pH in predicting pk_a to discuss the use of Hill/Langmuir-Freundlich (Hill/LF) and Frumkin isotherms in estimating the pk_a of the ionisable amino acids and the peptide.

3.1. Materials and Methods

3.1.1. Polyaspartic Acid Oligopeptide

A linear chain of 10 aspartic acid (Asp or D) amino acids capped by the acetyl (ACE) and methylamine (NME) groups at the N-terminal and C-terminal positions (ACE-Asp_{10}-NME) was constructed using the Leap module of AMBER18[22]. The peptide was parameterised using the ff14SB force field[23]. For the evaluation of the CpHMD method, three peptides were built for comparison: (i) the fully deprotonated peptide (defined by the ASP residue), (ii) the fully protonated peptide (ASH residue) and (iii) the *pH-responsive protonation state* peptide (defined by the AS4 residue, specifically designed for the CpHMD simulations). For convenience, hereinafter, each system will be referred to as (i) ASP_{10}, (ii) ASH_{10} and (iii) $AS4_{10}^{pH=X}$, where X is the solvent pH value in the

CpHMD simulation. To compare the fully deprotonated and protonated peptides with the pH-responsive peptide, solvent pH values of 10.0 and 1.0 were chosen for the CpHMD simulations, respectively. Each of the three systems was prepared under implicit and explicit solvation. Implicit solvation was introduced with the generalized-Born model of Onufriev et al.[24] (gb = 2). For those systems with explicit solvation, the simulation box was defined with dimensions of 77.5x77.5x77.5 $Å^3$ and then filled with TIP3P[25] water molecules and neutralised with Na^+ and Cl^- counterions. Finally, the reference energies of the protonation states were defined using the cpinutil module of AMBER18. The radii of the carboxyl group atoms in the side chain of the AS4 residues were modified in the CpHMD simulations with explicit water molecules as recommended[26].

3.1.2. Simulation Setup

Each simulation box was minimised using the Steepest Descent (SD) method[27] with three levels of restraints: (i) in all atoms of the peptide, (ii) only the backbone atoms and (iii) no restraints. Atom positions were restrained with a force constant of 5 $kcal \cdot mol^{-1} Å^{-1}$. Up to 5,000 steps of SD were performed for each restraint stage, allowing a stepwise relaxation of the peptide with the surrounding solvent.

Before the production runs, the simulation box of each peptide was heated with a linear temperature increase of $1K \cdot ps^{-1}$ for 300 ps and equilibrated in the isobaric-isothermal ensemble (NPT) for 200 ps to adjust the density. A restraint of 5 $kcal \cdot mol^{-1} Å^{-1}$ was applied in the C_α atoms of the peptide. Finally, four 500 ns-length production runs of each peptide system were carried out in the canonical ensemble (NVT). To extend the conformational sampling during the simulations, each of the production runs was started with random velocities satisfying the Maxwell-Boltzmann distribution. The temperature was controlled using a Langevin thermostat[28] (ntt = 3) with a collision frequency of 3 ps^{-1}. The long-range electrostatic interactions were calculated using the Particle Ewald summation method[29]. The SHAKE algorithm was applied to constrain the hydrogen-involving bonds[30]. A cut-off of 10.0 Å was defined for the intermolecular interactions in the explicit solvent simulations.

To include the effect of pH in the simulations, the $AS4_{10}$ peptide was subjected to the discrete CpHMD method using the protocols according to the solvation models: the implicit solvation based on the Generalized Born model (icnstph = 1) and the explicit solvation (icnstph = 2). In the first model, the protonation state change attempt was

performed on a random titratable residue every 10 fs, resulting in an effective global protonation state change attempt of 100 fs (i.e., on average, each amino acid is subjected to one protonation state change attempt every 100 fs). In the second model, the protonation state change attempt was performed every 200 fs on all titratable amino acids, followed by 200 fs of solvent relaxation. The protonation state change attempt in the implicit solvation was performed using the Onufriev et al. model (gb = 2) with a salt concentration of 0.1 M.

	Implicit Solvation	**Explicit Solvation**	**Method**
ASH_{10}	4 x 500 ns	4 x 500 ns	cMD
ASP_{10}	4 x 500 ns	4 x 500 ns	cMD
$AS4_{10}^{pH=1}$	4 x 500 ns	4 x 500 ns	CpHMD
$AS4_{10}^{pH=10}$	4 x 500 ns	4 x 500 ns	CpHMD
$AS4_{10}^{pH=1\text{-}8.5,\ \Delta pH=0.5}$		16 x 200 ns	pH-REMD

Table 1. Details of the simulations performed according to the solvation model, the MD method, and the pH values.

In parallel, we have also performed pH-REMD simulations with explicit water molecules. A total of 16 replicas of 200 ns length were carried out with a pH range from 1 to 8.5 and a pH interval of 0.5 between each replica. The exchange attempt between pHs occurred every 200 fs. All other details specified above for the CMD and CpHMD simulations were maintained.

3.1.3. Energetic and Conformational Analysis

The structural analysis of the simulations was performed using the CPPTRAJ module of AmberTools18[31]. The root-mean-square deviation (RMSD) was calculated using the backbone atoms (C, C_α, N, O) of the amino acids. The radius of gyration (R_g) was calculated using the C_α atoms of the peptide. R_g histograms were calculated using a Gaussian kernel density estimator. Secondary structure propensity fractions (fpSS) were estimated using the DSSP method[32]. Energy contributions were extracted directly from the simulation. The conformations of the trajectories were clustered into 15 groups using the bottom-up hierarchical agglomerative algorithm. The RMSD values of the C_α atoms of the aspartic acids were used as a metric for the clustering. The centroid conformations were represented graphically using the Visual Molecular Dynamics (VMD) program[33]. All plots were generated using the Gnuplot utility[34]. Principal Component Analysis

(PCA) technique was performed on the C_α atoms to reduce the dimensionality of the conformational sampling. The pH properties were calculated using the CPHSTATS module of AmberTools18. Data processing was performed by in-house programs in Python 3.6. The deprotonated fractions were fitted to the Hill/LF and Frumkin isotherms using the equations described in Section 3.2.

3.2. Results and Discussion

3.2.1. Assessment of the Constant pH Molecular Dynamics Simulations

To validate the CpHMD method, we forced the pH-sensitive peptides of the CpHMD method to be fully protonated or deprotonated by applying strong acidic and basic conditions. Therefore, we performed the CpHMD simulations at pH 1 and 10, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$, and compared each with the analogues of the CMD method, ASH_{10} and ASP_{10}, respectively. To understand the implications of each method, we have analysed the conformational sampling by PCA and 2D-RMSD of the representative conformations of the clusters, their structural properties such as R_g and fpSS, and finally the time evolution of the potential, electrostatic and van der Waals (vdW) energies.

First, we focused on the set of conformations obtained from the trajectory of each simulation using PCA. After reducing the dimensionality of the peptide coordinates to the Principal Components (PCs), we constructed an energy map from the populations of the first two PCs for each system in Figure 1. For implicit solvation, the conformational sampling of the CMD simulations, ASH_{10} and ASP_{10}, are below PC1 = 0, although the latter is shifted towards negative values and the sampled space shows minor differences. ASH_{10} is predominantly in the region of PC1 = [-10, -3] and PC2 = [-10, 5], which is broader with respect to ASP_{10} which samples the region of PC1 = [-12, -5] and PC2 = [-5, 7]. On the other hand, the CpHMD simulations, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$, are both in the range of PC1 = [0, 10] and PC2 = [-10, 5]. In fact, both simulations share a minimum at approximately [8, 2]. In addition, the peptide reaches low energy conformations distributed in the sampled space when the system is under acidic conditions. However, the CpHMD simulations show smaller preferred regions compared to the CMD simulations, indicating that the pH-responsive simulations in the implicit solvent apparently restricted the accessible conformational space.

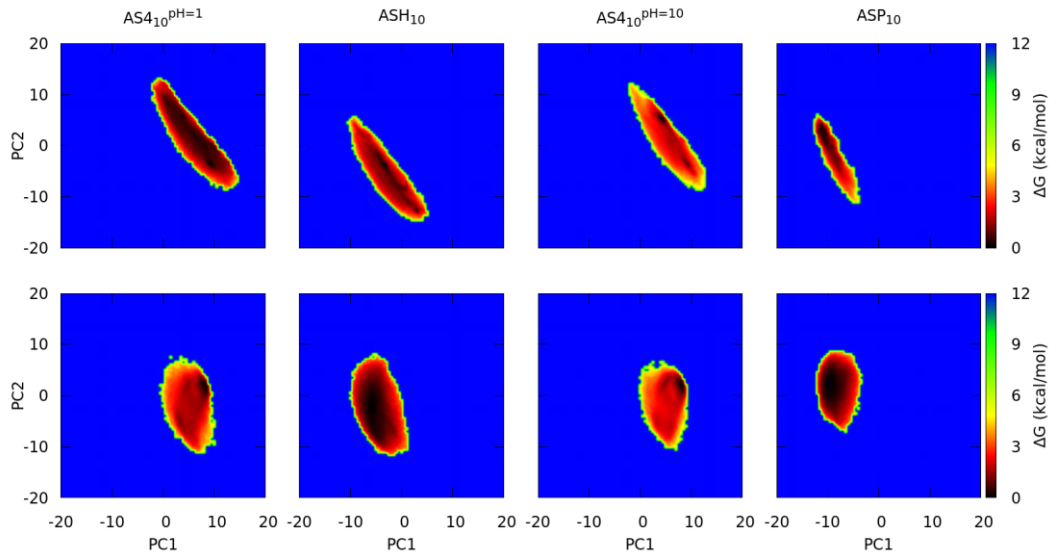


Figure 1. Energy maps based on Principal Component Analysis (PCA) of the fully protonated (ASH_{10} and $AS4_{10}^{pH=1}$) and fully deprotonated (ASP_{10} and $AS4_{10}^{pH=10}$) polyaspartic acid peptides for explicit (top) and implicit (bottom) solvation models.

The explicit solvent simulations show a similar behaviour respect to the implicit solvent analysis. On the one hand, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ sample a similar conformational region in the energy map, approximately at PC1 = [-2, 15] and PC2 = [-8, 12]. When the peptide is under acidic conditions, $AS4_{10}^{pH=1}$, we observe a homogeneous sampling within the region. However, the peptide at basic conditions, $AS4_{10}^{pH=10}$, shows a reduced set of low energy conformations with two minima located at (4, 6) and (10, -4). On the other hand, the CMD simulations, ASH_{10} and ASP_{10}, present regions that are close within the energy map and differ from the regions sampled by the CpHMD method. ASH_{10} samples low energy conformations scattered within a wider region at PC1 = [-10, 6], PC2 = [-15, 5]. In contrast, ASP_{10} exhibits a more compact region at PC1 = [-12, -4], PC2 = [-11, 6] with a low energy conformation set at PC2 ≥ -5. Interestingly, in the protonated state, the preferred conformations lie just below PC1 ≈ 0, meaning that the conformations between the two protonation states should be expected to be distinct for a significant fraction of the conformational population.

The results of the PCA analysis would suggest that the conformational sampling is, surprisingly, *method-dependent* rather than protonation state dependent. Polyaspartic acid simulations in the CpHMD method show a clear tendency to sample similar regions, but with a restraining effect on the accessibility of some conformational regions when is under basic conditions. This restraining effect is also observed in the CMD simulations

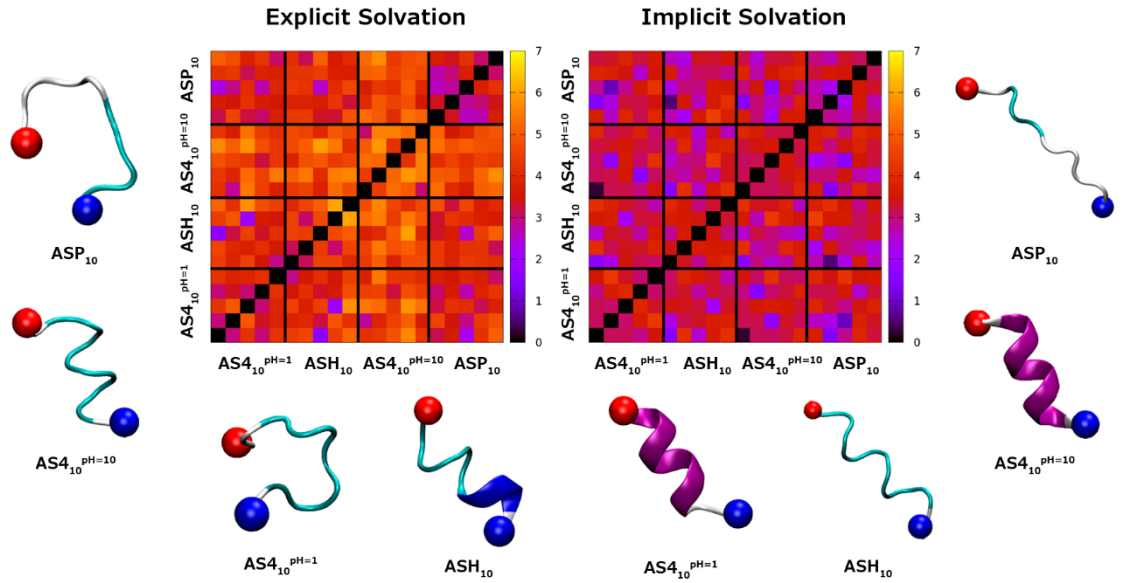


Figure 2. 2D-RMSD of the fifth most populated clusters of the fully protonated (ASH_{10} and $AS4_{10}^{pH=1}$) and fully deprotonated (ASP_{10} and $AS4_{10}^{pH=10}$) polyaspartic acid peptides. The representative conformation of the most populated cluster is plotted for each simulation system.

in which each peptide samples its own conformational space. However, we must stress that the covariance of the coordinate space collected in these PCs may not be sufficient to accurately interpret the conformational sampling of the trajectories. In fact, the amount of covariance data collected in the first two PCs is around ~50%. Therefore, we need to address the comparison between simulation methods with additional analyses on the conformational sampling, such as clustering or the estimation of structural properties.

Thus, the conformations of the simulations of each system were grouped into fifteen clusters by applying the hierarchical agglomerative clustering method. The centroid conformation (i.e., the conformation with the lowest RMSD with respect to the other conformations within a cluster) of the five most populated clusters was extracted and the RMSD between them was calculated, yielding the 2D-RMSD maps in Figure 2. This map is a useful indicator of the structural similarity between the clusters, whose populations are given in Table S1. In addition, the centroid conformation of the most populated cluster was also plotted.

The 2D-RMSD map of the implicit solvation simulations shows a greater similarity between the centroids with values in the range of 3-2 Å (red or purple boxes) and even some values close to ~0 Å (dark purple or black boxes). On the other hand, the

2D-RMSD map of the explicit solvent simulations shows a greater structural heterogeneity between the centroids of the clusters. Most of the values are in the range of 3-6 Å (orange boxes), with some exceptions such as the values at ~2 Å (purple boxes), which are infrequently. The simulations with explicit water molecules can capture structurally more distant conformations in conformational space, while the simulations with implicit solvent retain similar conformations in time. Indeed, the centroids of cluster 0 in ASH_{10} and $AS4_{10}^{pH=10}$ at explicit solvation present an α-helix-forming conformation, while $AS4_{10}^{pH=1}$ and ASP_{10} show a disordered structure. Interestingly, the peptides with opposite protonation state have similar structures in the main cluster centroids, although the populations are different (Table S1). The population distributions of the clusters in the protonated state agree with small deviations. In addition, the ASP_{10} centroids show large RMSD values (~6 Å) with respect to the other simulation conditions, indicating that structurally more distinct conformations are expected. On the other hand, when using the implicit solvation model, the peptide reaches helical conformations in the ASH_{10}, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$. In contrast, ASP_{10} shows a more spatially extended structure. It should be emphasised that the 2D-RMSD map indicates that the centroid conformations of $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ in the main cluster are very similar in structure. These clusters represent more than 80% of the conformational sampling of these peptides, which means that most of the trajectories of the CpHMD simulations have a high structural similarity. On the contrary, the population of clusters does not agree between the analogous systems, and, in addition, they are close between the simulations performed with the same method.

Except for $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ in the implicit solvation, the 2D-RMSD map does not allow to distinguish whether the peptides are conformationally similar to each other when they are in the same protonation state. Low RMSD values are observed in the 2D-RMSD map for systems with different protonation (e.g., in explicit solvation, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ show two centroids with ~2 Å values) and with similar protonation (e.g., $AS4_{10}^{pH=1}$ and ASH_{10} also show two centroids with ~2 Å values). Therefore, we decided to further explore the implications of using each simulation method in the polyaspartic acid decapeptide and analysed some structural properties in Figure 3, such as R_g and fpSS.

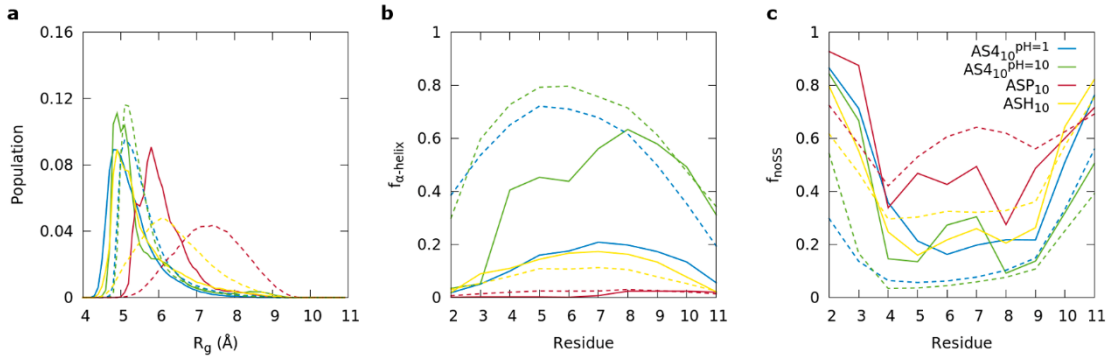


Figure 3. (a) Radius of gyration, (b) fraction of α-helix conformation and (c) fraction of absence of secondary structure of the fully protonated (ASH_{10} and $AS4_{10}^{pH=1}$) and fully deprotonated (ASP_{10} and $AS4_{10}^{pH=10}$) polyaspartic acid peptides in implicit (dashed lines) and explicit (solid lines) solvation model.

Starting from R_g, the implicit solvent simulations show a clear behaviour: ASH_{10} and ASP_{10} have different R_g distributions with respect to the other peptides, including those in the explicit solvent. These distributions have broad R_g values ranging from 4.5 to 9.5 Å. In contrast, the CpHMD simulations, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$, show narrow and large peaks around ~5 Å. In the explicit solvent simulations, the protonated peptides, ASH_{10} and $AS4_{10}^{pH=1}$, exhibit similar R_g distributions with a single peak at ~4.7 Å. On the other hand, the deprotonated peptides, ASP_{10} and $AS4_{10}^{pH=10}$, show separated distributions with their maximum peaks at ~6 and ~4.7 Å, respectively, suggesting that they do not have similar structural compactness. In fact, an increase in R_g of deprotonated peptides is expected because the negative charges of side chain carboxyl groups repel the neighbouring like-charged groups, thus elongating the peptide chain in order to reduce the energy penalty of having negatively charged groups in close proximity[35]. Surprisingly, $AS4_{10}^{pH=10}$ does not show this behaviour and instead the two peaks are observed at R_g values close to the protonated state.

Next, we focused on the α-helix formation and the absence of secondary structure (non-SS), which are the main fractions observed in the centroids in Figure 2. In the α-helix plot, Figure 3b, the α-helix fractions per residue already show a method-dependent behaviour decoupled from the protonation state in the implicit simulation methods. $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ have a high α-helix content (~70%), whereas ASH_{10} and ASP_{10}, barely reach a 20% of the content. On the other hand, the protonated peptides in explicit solvent, ASH_{10} and $AS4_{10}^{pH=1}$, correspond correctly in the α-helix fraction with values

below the ~20%. For the deprotonated ones, ASP_{10} and $AS4_{10}^{pH=10}$, the α-helix fractions exhibit opposite behaviour. The former barely contains any α-helix fraction, while the latter shows larger fractions (>40%), even showing an α-helix content close to the implicit solvation observation (~70%) for the second half of the peptide chain. As an indicator of the degree of disorder in the conformations, we have calculated the fraction of conformations whose structure is not classified by the DSSP method, hence termed non-SS in Figure 3c. The non-SS propensity shows a similar behaviour with respect to the other structural properties (R_g and α-helix fraction). The CpHMD simulations using the implicit solvation model have a low content, implying the expectation of highly structured conformations for these peptides. In the CMD simulations, ASH_{10} and ASP_{10} show non-SS fractions of around ~33% and ~55% respectively. ASP_{10} has the highest content of disordered conformations of all the simulations presented in this section. In contrast, the explicit solvent simulations agree between the protonated peptides, ASH_{10} and $AS4_{10}^{pH=1}$, which present a conformational sampling with a low percentage of disorder in the structure (~25%). The deprotonated peptides, ASP_{10} and $AS4_{10}^{pH=10}$, disagree in the non-SS fractions with values around ~40% and below ~20% respectively.

The other SS propensities are shown in Appendix A, Figure A1. For 3_{10}-helix and turn formation, the method-dependent behaviour is not observed for the implicit solvation model. However, the 3_{10}-helix fraction is not consistent between the analogous systems, while the turn structure has a similar content for all systems. The bend formation again underlines the inconsistency observed in the previous SS propensities: a clear deviation is found depending on the method. On the other hand, the explicit solvent simulations show that the protonated state cannot reproduce the 3_{10}-helix content, and, for the first time, the deprotonated peptides have similar fractions. ASH_{10} stands out with a ~20% formation in residues 3-6 of the peptide chain, while $AS4_{10}^{pH=1}$ barely contains a ≥10% fraction. In the turn structure, significant fractions below ~40% are observed in the protonated state with a good agreement between the analogue simulations. The deprotonated peptides show a lower turn content for both ASP_{10} and $AS4_{10}^{pH=10}$, which exhibit a different distribution of the content within the peptide chain. This turn-content-dependent behaviour between the protonated and deprotonated states is explained by Milorey B. et al. paper[36], in which work was reported that the protonated aspartic acid in the GDG trimer has a high propensity to form turn-like conformations, whereas the deprotonated state loses the ability to form turn structures due to the side chain charges[37].

We assume this explanation can be extended in the case of the decapeptide. Finally, the bend formation is consistent with the observations highlighted in this section: the ASP_{10} content shows a significant deviation with respect to $AS4_{10}^{pH=10}$, with a difference of up to ~40% for some residues, while the peptide in the protonated state exhibits similar contents of around ~20%.

Several observations can be made from the analysis of the conformational space and structural properties. First, there is a deviation of the conformational sampling according to the simulation method when the implicit solvation model is used, thus demonstrating a limitation of the fully protonated or fully deprotonated peptides to reproduce the properties of the analogous simulations in CMD. The results reported in this chapter justify this statement, especially in the 2D-RMSD map, in which we can observe a great structural similarity (RMSD ≈ 0 Å) between the main centroids of the CpHMD simulations with a population ≥ 80%. Inevitably, these highly populated clusters lead to structural properties with similar tendencies and therefore do not agree with the analogous simulations of the CMD method. Fortunately, the explicit solvation model escapes from this deficiency, but leads to a second observation. On the one hand, the simulations of the peptide in the protonated state agree on the structural properties (R_g and SS fractions), whereas the energy map does not. As we have already mentioned, it is likely that the PCs do not sufficiently capture the covariance of the coordinates in the first two PCs to accurately represent the conformational space, so focusing on the reproduction of the conformational properties allows us to guarantee a certain reliability of the CpHMD method. On the other hand, it is worth noting that the simulations in the deprotonated state again show inconsistencies with the results reported so far, especially in the structural properties. In the latter case, we have focused on the role of the solvent, specifically the sodium counterions (Na^+). For this purpose, we have analysed the radial distribution function (RDF) of the oxygen atoms of the carboxyl group of the side chain in Figure A2. Indeed, there is a significant deviation in the distribution of the counterions around the oxygens of each system. In the peptide simulation of the CMD method, an intense peak at ~2.5 Å and a second one of lower intensity at ~4.5 Å are observed. Conversely, the CpHMD simulation shows only a very low intensity peak at 2.5 Å, indicating that the Na^+ counterions do not interact with the carboxyl groups at short distances. Such evidence could explain the discrepancies between the CMD and CpHMD simulations in the deprotonated state.

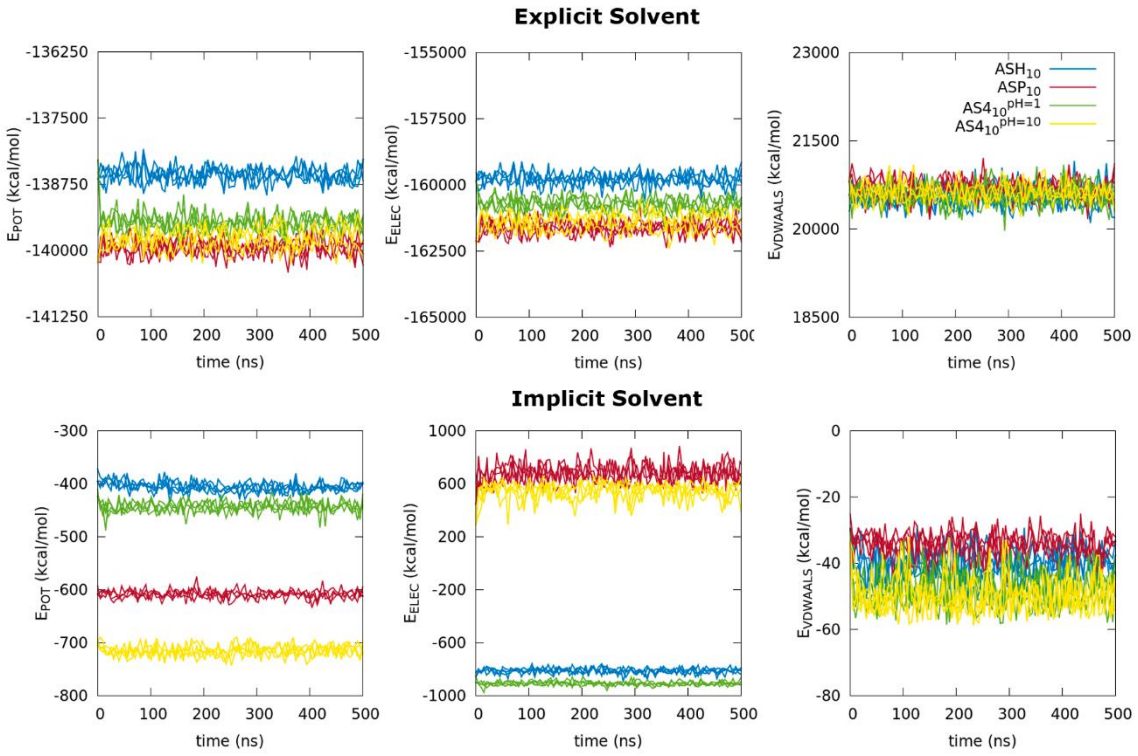


Figure 4. Potential, electrostatic and van der Waals energy contributions of the fully protonated (ASH_{10} and $AS4_{10}^{pH=1}$) and fully deprotonated (ASP_{10} and $AS4_{10}^{pH=10}$) polyaspartic acid peptides.

Given the difficulty in understanding the reported weaknesses of the CpHMD method, we have investigated the energy contributions of the peptide simulations. The potential, electrostatic and vdW energy contributions for the explicit (top) and implicit (bottom) solvent simulations are plotted in Figure 4. In the implicit solvent, the potential energy has a gap of ~100 kcal/mol between the deprotonated systems, ASP_{10} and $AS4_{10}^{pH=10}$. In the protonated systems, ASH_{10} and $AS4_{10}^{pH=1}$, this gap is reduced to ~40 kcal/mol, but is still remarkable. This potential energy can be divided into internal (bond, angular, dihedral, etc), electrostatic and vdW contributions. Focusing on the electrostatic interactions, it is noted that the electrostatic energies of the peptides with the same protonation state also display energy gaps. ASH_{10} and $AS4_{10}^{pH=1}$ have an energy gap around ~100 kcal/mol, while ASP_{10} and $AS4_{10}^{pH=10}$ present a small but fluctuating energy gap between the two simulations. Curiously, this energy gap is more pronounced in the protonated state. On the other hand, the vdW contributions show minimal differences with respect to the electrostatic term, so it does not seem to cause this energy divergence. In the explicit solvent simulations, the potential energy also shows these energy gaps between analogous systems. On this occasion, the protonated peptides, ASH_{10} and

$AS4_{10}^{pH=1}$, exhibit a difference of ~1250 kcal/mol, while the deprotonated ones, ASP_{10} and $AS4_{10}^{pH=10}$, reduce it by half, ~500 kcal/mol. The increase in the energy gaps could be due to the interaction of the polyaspartic acid decapeptide with the counterions and water molecules. In fact, the electrostatic energy contribution shows a similar behaviour between ASH_{10} and $AS4_{10}^{pH=1}$, but it does not entirely explain the energy gaps observed in the total potential energy. The vdW contributions are consistent between the two protonation states.

The energetic contributions show that the structural divergence according to the simulation method is mainly due to the electrostatic energy contribution. On the one hand, the simulations with implicit solvation present smaller energy gaps, although they show larger structural differences in the conformational sampling of the analogous systems in the protonated state. Interestingly, the peptides in the protonated state minimise the energy gap in the potential contribution, although the electrostatics shows similar energy gaps for both protonation states. Therefore, other energetic contributions such as internal energy come into play in the energetic divergence, which needs to be studied in more detail in the future. On the other hand, simulations with explicit solvent increase the energy gaps between the analogous peptides, especially in the protonated state. It should be stressed that this increase is probably due to the interaction of the peptide with the surrounding solvent molecules. Similar to the implicit solvation, the divergence in the potential energy is not fully explained by the electrostatic contribution in the explicit solvent simulations. Despite the failure to reproduce the potential energy, the protonated peptides, ASH_{10} and $AS4_{10}^{pH=1}$, show a surprisingly good reproduction of the structural properties (R_g and fpSS), consistent with the experimental evidence mentioned above. On the other hand, and contrary to $AS4_{10}^{pH=10}$, we believe that ASP_{10} follows the properties of an ionised polyanion, since larger R_g distributions and low turn-content for the polyaspartic acid in CMD simulations as described in the literature are observed. A deeper examination focusing on the structural properties and energy contributions in the CpHMD method is therefore necessary and will be discussed in Chapters 3 and 4 of this thesis.

3.2.2. pH-Responsive Conformations at pH Conditions around the Intrinsic pKa

3.2.2.1. Progressive Shift of Conformational Properties with Solvent pH

Next, we will study the effect of solvent pH at values around the intrinsic pk_a of the aspartic acid amino acids. To this end, we have used the pH-REMD method, which involves running a series of parallel replicas within a range of pH values separated by a ΔpH. These parallel trajectories are swapped during the simulation by applying an exchange probability criterion. Therefore, 16 replicas of the polyaspartic acid decapeptide were simulated in explicit solvation conditions within a pH range = [1, 8.5] and ΔpH = 0.5 during 200 ns per replica. We then performed a conformational analysis and estimated the apparent pk_D and intrinsic pk_a by fitting the deprotonation fractions to the Langmuir-Freundlich and the Frumkin isotherms.

We followed the PCA protocol to construct the energy maps for each solvent pH condition, as shown in Figure 5. All of them are approximately in the range of PC1 = [-10, 10] and PC2 = [-10, 10], thus suggesting that the conformational space of the peptide is not broad or, more certainly, that the conformational sampling of each replica is in the same region. The energy map of $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=2}$ sampled homogeneously through this region (broad dark region, ΔG ≈ 0 kcal/mol). Interestingly, a progressive convergence towards specific minima within the conformational space is observed as we increase the solvent pH. It finally culminates in two minima at mildly basic conditions, pH 8, located approximately at (0, -5) and (-1, 0). Therefore, increasing the solvent pH promotes these conformational regions and makes them more populated in the PCA maps. Indeed, the 2D-RMSD map in Figure A3 also shows this restraining effect on the representative conformations of the three most populated clusters per replica. At pH > 3, the RMSD values decrease significantly, indicating that the representative conformations have structural similarity with RMSD < 1 Å, whereas these low-RMSD values are not prevalent at pH ≤ 3. However, it cannot be ignored that the RMSD values increase again at pH 8, indicating that there are structural differences between the conformations at mild basic pH compared to the simulations at mild acidic pH.

To interpret the promotion of specific conformations with increasing pH, we have focused on the structural properties such as radius of gyration, α-helix fraction and number of hydrogen bonds shown in Figure 6. The R_g distributions in Figure 6a show the progressive increase of the peak ~5.2 Å at pH = 1 with a maximum at pH = 6, reaching a

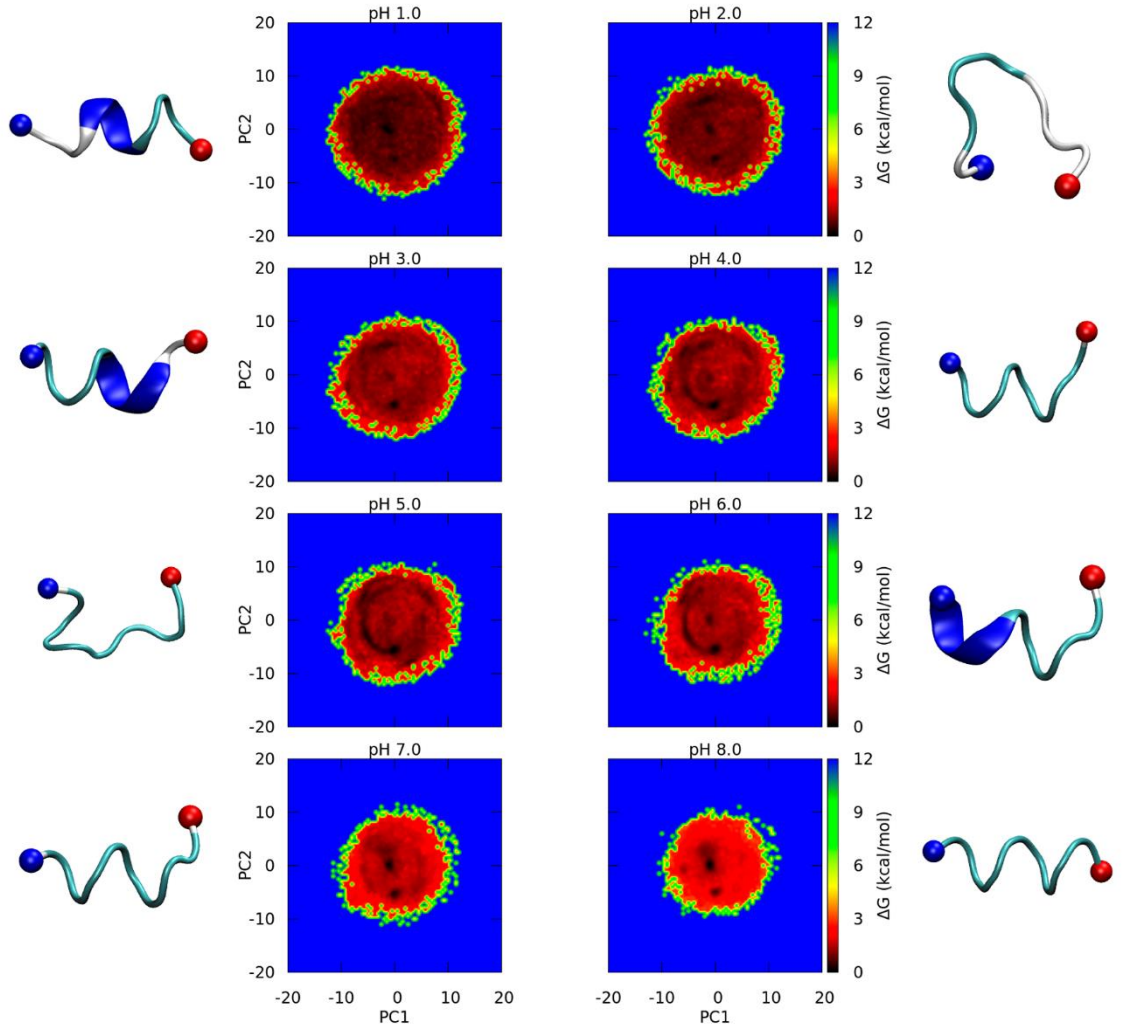


Figure 5. Energy maps based on Principal Component Analysis (PCA) of pH-REMD trajectories from pH 1.0 to 8.5 with ΔpH = 0.5. The representative conformation of the most populated cluster of the trajectory at each pH condition is also shown in the figure.

population of 20% in this simulation. It then decreases and shifts to a second peak with Rg ~4.8 Å at mildly basic conditions, pH 8. In fact, the number of conformations with R_g > 6 Å gradually increases when 3 > pH > 6. Therefore, the peptide can reach extended structures under strongly acidic conditions and is compacted when the pH becomes neutral or mildly basic. This behaviour could be explained by the shift in the H-bond distributions observed in Figure 6b. Interestingly, the peptide forms more H-bonds in mildly acidic conditions, as can be seen in the distributions with peak ~5 at pH = 5-6. The distribution of R_g and H-bonds suggests that at these pH conditions there is an interplay of polar interactions that would compact the peptide structure. On the other hand, there is

an interruption in H-bond formation at pH 7-8, reaching a distribution with a peak at ~2 and a significant reduction in the number of H-bonds.

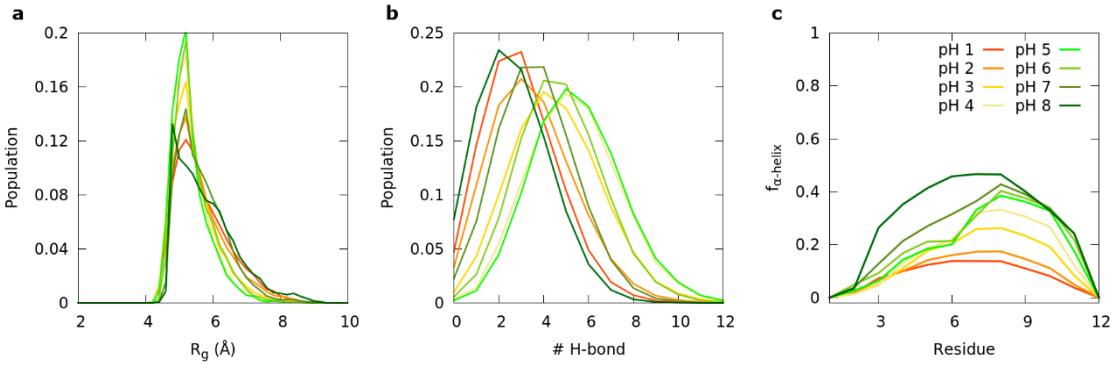


Figure 6. (a) Radius of gyration, (b) number of H-bonds and (c) α-helix fraction of the polyaspartic acid peptide in the explicit solvation pH-REMD simulation.

All this discussion is supported by the evolution of the α-helix fraction in Figure 6b. $AS4_{10}^{pH=1}$ has a low α-helix content, which gradually increases as the pH is changed to mildly basic conditions. Indeed, the formation of the α-helix causes the peptide to gain structure and thus become more compact reaching populations of ~40%, in particular from residues 6 to 11. $AS4_{10}^{pH=8}$ has the highest α-helix content despite having a low amount of H-bonds. This radical change in the conformational properties between simulations at acidic pH conditions, from 3.0 to 6.0, and neutral or slightly basic simulations, pH 7 or pH 8, suggests that the predominant conformations formed at these solvent pH conditions are significantly different. In fact, the population of the main clusters collected in Table S2 in conjunction with the 2D-RMSD plot (Figure A3) makes these conformational changes evident. Interestingly, the main conformation at mild basic pH, with ~55% population, is structurally similar to the main cluster at acidic conditions (pH = 2), which could explain the similar R_g distributions despite the different α-helix fractions.

3.2.2.2. Determination of pKa by Complexation Isotherms

One of the most remarkable points demonstrated by the CpHMD method is the ability to predict the intrinsic acid dissociation constant pk_a of the titratable amino acids. Several papers reported so far have proven that experimental pk_a can be obtained with CpHMD within an acceptable error[6,20,21,38]. In these works, the so-called effective or apparent pk_a is usually obtained by collecting the deprotonation fraction, f_d, at several pH values and fitting it to the titration curve defined by the Hill equation[39], which is

widely used in enzymatic catalysis. Fortunately, f_d can be easily obtained over a wide range of pHs using the pH-REMD method. However, there has been little discussion of the use of the Hill equation to predict pk_a in simulations at constant pH. Here, we would like to explore this topic in more depth and make a number of clarifications about the Hill equation, which is basically the well-known Langmuir-Freundlich (LF) isotherm in heterogeneous catalysis for negative cooperativity ($n < 1$), and the Frumkin isotherm in the context of pk_a predictions.

First, the proton must be identified as a ligand of the polyaspartic acid peptide, which is a macromolecule or receptor with i binding sites provided by each aspartic acid amino acids. When the proton (H^+) and the binding site *i* in the deprotonated state ($-S_i^-$) complex, the functional group in the protonated state ($-S_iH$) is formed. Therefore, this reaction is defined with an equilibrium binding constant, $K_{c,i}$, which depends on the concentration (or activity) of each component of the reaction as described in Eq. 1.

$$-S_i^- + H^+ \rightleftarrows -S_iH \qquad K_{c,i} = \frac{[-S_iH]}{[-S_i^-] \cdot a_{H^+}} \qquad \textbf{Eq. 3.1}$$

In terms of a binding process study, the deprotonation fraction of each aspartic acid, $f_{d,i}$, becomes the degree of dissociation, α_i, or conversely the coverage, θ_i, defined as $\theta_i = 1 - \alpha_i$ or $\theta_i = [occupied\ i - sites]/[total\ i - sites]$. From θ_i and $K_{c,i}$, we can estimate the apparent pk_a, pk_{app}, generalising the Handerson-Hasselbalch (HH) equation, Eq. 2. In an ideal case, where the amino acids do not interact with the neighbouring residues, the Henderson-Hasselbalch (HH) behaviour is fulfilled and the $pk_{app,i}$ becomes a pH-independent constant. On the contrary, for a non-ideal case, the $pk_{app,i}$ depends on the solvent pH and then becomes a distribution of pk_a[40–42].

$$pk_{app,i} \equiv logK_{c,i} = pH + \begin{cases} \log\left(\frac{\theta_i}{1-\theta_i}\right) \\ log\left(\frac{1-\alpha_i}{\alpha_i}\right) \end{cases} \qquad \textbf{Eq. 3.2}$$

Under non-interaction ideal conditions, it is possible to estimate the *intrinsic* acid dissociation constant of the amino acid i within a protein, the intrinsic $pk_{a,i}$. The intrinsic $pk_{a,i}$ is defined as the pk associated with the free energy when the ionisation state changes from neutral to charged and no other (coulombic) electrostatic interactions of the

chemical environment are involved in the process. In addition, the intrinsic $pk_{a,i}$ estimation must be averaged over the entire conformational ensemble of the macromolecule. Then, the $pk_{a,i}$ shift of an amino acid within a protein with respect to the $pk_{a,i}$ of a free amino acid is attributed to the desolvation penalty (or Born effect) and the background interaction energy, which is due to the interactions with the permanent dipoles of the protein. These contributions are critical when calculating the intrinsic $pk_{a,i}$ of an amino acid by free energy calculations[43,44].

The constant pH simulations enable the prediction of $pk_{app,i}$ without free energy calculations, allowing us to construct the $pk_{app,i}$ distribution by fitting $f_{d,i}$ or α_i to a complexation isotherm. Each isotherm has its own particularities and therefore some notes are necessary. Under ideal conditions, the Langmuir isotherm, derived from the adsorption model of the same name, equivalent to the Monod equation, commonly used for the growth of microorganisms and non-cooperativity binding of enzymes, can model the binding process of a ligand-receptor complex. Therefore, the Langmuir and Monod are the same particular solution of the Hill/Langmuir-Freundlich isotherm, called the Hill/LF isotherm from now, for the ideal case. The Langmuir isotherm, expressed in Eq. 3, gives the ideal behaviour of each of the amino acids in our peptide, which should be identical between the several binding sites, regardless of their position in the peptide chain. The binding constant $K_{c,i}$ then satisfies the condition that it does not depend on pH or proton activity, so that $K_{c,i} = k_i \neq f(pH)$ and hence $pk_{a,i} = logk_i$.

$$\theta_i = \frac{k_i a_{H^+}}{1 + k_i a_{H^+}} \qquad \textbf{Eq. 3.3}$$

The Hill/LF isotherm allows the interpretation of the binding process of the ligand-receptor complex in a non-ideal scenario. In biochemistry, the Hill equation identifies positive (n > 1) or negative (n < 1) cooperativity in enzymatic reactions according to the Hill coefficient, *n*. However, this negative cooperativity can also be explained by the heterogeneity of the binding sites according to the LF isotherm, which is probably the best interpretation in this study. The Hill/LF isotherm is expressed in Eq. 4a in terms of the average binding constant $k_{m,i}$, but it can also be defined in terms of the average acid dissociation constant, $k_{am,i}$, with $k_{am,i} = \left(k_{D,i}\right)^{n_i}$. $k_{D,i}$ is the dissociation constant at the half occupation which is satisfied when $\alpha_i = \theta_i = 0.5$ and n_i is an empiric parameter to fit the isotherm to the $f_{d,i}$ values. From the coverage in Eq. 4a, we can obtain

the Hill/LF equation in Eq. 4b to estimate the $pk_{app,i}$ distribution and n_i for each amino acid. Next, we rearranged the Hill equation in Eq. 4c so that it depends on $f_{d,i}$ or α_i as considered in the constant pH-framework to estimate the $pk_{D,i}$, and pk_{app} of the amino acids by fitting the simulated $f_{d,i}$'s. However, it must be emphasised that this $pk_{D,i}$ only corresponds to the pk_{app} at pH conditions where $\alpha_i = \theta_i = 0.5$ is satisfied. At other pH conditions it is necessary to calculate the $pk_{app,i}$ and for this reason the pk_{app} distribution plays a key role in determining the real protonation state fractions when the polyaspartic acid is far from the $pk_{D,i}$ conditions.

$$\theta_i = \frac{\left(k_{m,i}a_{H^+}\right)^{n_i}}{1+\left(k_{m,i}a_{H^+}\right)^{n_i}} = \frac{a_{H^+}^{n_i}}{k_{am,i}+a_{H^+}^{n_i}} = \frac{a_{H^+}^{n_i}}{\left(k_{D,i}\right)^{n_i}+a_{H^+}^{n_i}} = \frac{1}{\left(\frac{k_{D,i}}{a_{H^+}^{n_i}}\right)^{n_i}+1} \qquad \textbf{Eq. 3.4a}$$

$$pk_{app,i} = pH + \frac{1}{n_i}\log\frac{1-\alpha_i}{\alpha_i} \quad \xrightarrow{\alpha_i=\theta_i=0.5} \quad pk_{app,i} = pk_{D,i} = pH \qquad \textbf{Eq. 3.4b}$$

$$f_{d,i} = \alpha_i = \frac{1}{1+10^{n_i(pk_{D,i}-pH)}} \qquad \textbf{Eq. 3.4c}$$

One of the most convenient features of fitting these isotherms is their ability to estimate the intrinsic $pk_{a,i}$ of the amino acids. By definition, we must calculate the $pk_{a,i}$ when the aspartic acids are in the neutral state, which in this case is the limit when $\alpha \to 0$ (fully protonated peptide). Unfortunately, the Hill/LF isotherm is not able to reproduce the binding properties at low dissociation levels[40]. In fact, the conversion of $k_{am,i} = \left(k_{D,i}\right)^{n_i}$ leads to lower values due to the deviation of the isotherm at these conditions. The source of this deviation is that the Hill/LF isotherm does not distinguish between interactions and heterogeneity since $k_{am,i}$ represents the average acid dissociation constant, assuming a heterogeneity-dispersion of the $k_{a,i}$ values provided by n_i. Thus, everything is captured within this empirical constant, which ultimately translates into a poor ability to model the binding processes at low coverage conditions, close to the neutral net charge, i.e., when the intrinsic $pk_{a,i}$ can be determined.

In view of the limitations of the empirical Hill/LF isotherm, we propose the Frumkin isotherm as an alternative isotherm capable of predicting the intrinsic $pk_{a,i}$ of amino acids, defined in Eq. 5. In this isotherm, k_i is the binding dissociation constant and δ_i is a physically meaningful parameter derived solely from a mean-field model of

interactions between the binding sites. The complexation isotherm can be used to the binding constant directly, from which we can estimate the p$k_{app,i}$ when $\alpha_i \to 0$ and decoupled from the effects of the interactions.

$$\theta_i = \frac{k_i c_{H^+} e^{2\beta\delta_i(1-\theta_i)}}{1 + k_i c_{H^+} ei^{2\beta\delta_i(1-\theta_i)}} \qquad \textbf{Eq. 3.5}$$

Both complexation isotherms enable the prediction of the $logk_{c,i}$ or $pk_{app,i}$ distribution from the $f_{d,i}$ values. To quantitatively compare the isotherms, we can compute the average pk_i, the $pk_{m,i}^{dist}$, and the standard deviation, σ_i^{dist}, from the distributions and thus assess whether the predicted distributions are comparable. The definitions of $pk_{m,i}^{dist}$ and the σ_i^{dist} are given in Eq. 6 and Eq. 7 for the Hill/LF and Frumkin isotherms[40], respectively.

$$pk_{m,i}^{dist} = pk_{D,i};\ \sigma_i^{dist} = \frac{\pi}{(ln10)}\sqrt{\frac{1-{n_i}^2}{3{n_i}^2}} \qquad \textbf{Eq. 3.6}$$

$$pk_{m,i}^{dist} = pk_{a,i} + \frac{\beta\delta_i}{(ln10)}\ ;\ \sigma_i^{dist} = \frac{1}{(ln10)}\sqrt{\frac{(\beta\delta_i)^2}{3} + 2\beta\delta_i} \qquad \textbf{Eq. 3.7}$$

The constant pH approach in AMBER sets a reference $pk_{a,i}$ of 4.0 for the aspartic acid by default. After the previous explanation, we proceeded to build the titration curves using the f_d values from the pH-REMD simulation and the Hill/LF and Frumkin isotherms in Figure 7. All constants and parameters extracted from the fitting of f_d to the isotherms are summarised in Table 2. For clarity, depending on whether the Hill/LF or Frumkin isotherms are used, $pk_{D,i}$ or the intrinsic $pk_{a,i}$ are obtained directly from the fitting, respectively. The reported Hill/LF isotherms show that all the amino acids of the peptide tend to shift their effective $pk_{D,i}$ to higher pH values, even reaching an effective $pk_{D,i}$ value of the reference pk_D + 1.4 at residue 11. More importantly, the Hill coefficients are far from the ideal case (n = 1), confirming that the titration of these amino acids follows a non-ideal behaviour. On the other hand, the intrinsic $pk_{a,i}$ of the amino

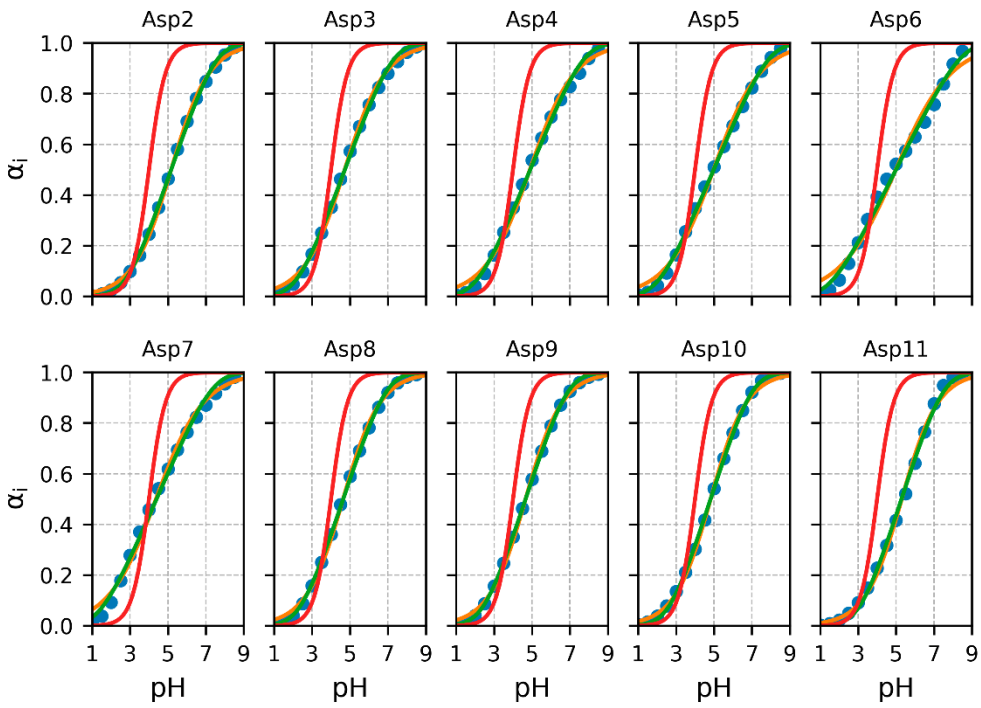


Figure 7. Titration curves estimated from the deprotonated state fractions obtained from the pH-REMD simulation (blue circles). The red lines are the Langmuir isotherm for an ideal case estimated using the reference $pk_{a,i}$ defined by AMBER. Orange and green lines are the Hill/LF and Frumkin isotherms, respectively.

acids obtained from the Frumkin isotherms are significantly lower than the reference value. We attribute this to the fact that the intrinsic $pk_{a,i}$ is obtained in the limit of non-interaction between amino acids, which cannot be fully achieved in the pH-REMD simulations due to the electrostatic interactions of the partial charges present in the all-atom model. Therefore, from both isotherms, we can confirm that (i) the amino acids are easily protonated ($pk_{D,i}$ > reference $pk_{a,i}$) to reduce the electrostatic repulsion exerted by the negative charges and (ii) the free energy to reach the neutral or fully protonated state decreases with respect to the free amino acid (intrinsic $pk_{a,i}$ < reference $pk_{a,i}$). Another interesting observation is the values of $pk_{D,i}$ and intrinsic $pk_{a,i}$ according to the position of the amino acid in the chain (not to be confused with the $pk_{app,i}$ distributions). While the distribution of $pk_{D,i}$ follows a rather unclear trend, the intrinsic $pk_{a,i}$ is conditioned by the position of the amino acid in the peptide. The intrinsic $pk_{a,i}$ decreases when the amino acid is placed in a more central position within the chain, which is also confirmed by the δ_i-interaction parameter. This parameter increases for the central amino acids,

suggesting that there are stronger interactions, and we therefore assume that the aspartic acids favour earlier neutralisation.

In fact, these strong interactions imply that the $pk_{app,i}$ is not constant and depends on the solvent pH, as we can observe in Figure 8. In these plots, the $pk_{app,i}$ varies significantly with pH, with variations of $\sigma_{i,LF}$= 1.5-2.5 and $\sigma_{i,F}$ = 1.3-2.1 with respect to the $pk_{m,i}$ of the distribution (i.e., $pk_{app,i}$ when α_i = θ_i = 0.5 if the protonation state sampling is extensive). When the simulated f_d's are compared with the isotherms, the Frumkin isotherm is more accurate and thus gives a $\sigma_{i,F}$ that is more reliable for the simulated values. Therefore, we would like to highlight the importance of understanding that the $pk_{app,i}$ in a non-ideal scenario can range widely between $\pm\sigma_i$, especially for these approaches that use $pk_{D,i}$ to determine an initial and fixed protonation state in MD simulations, as this can be tricky.

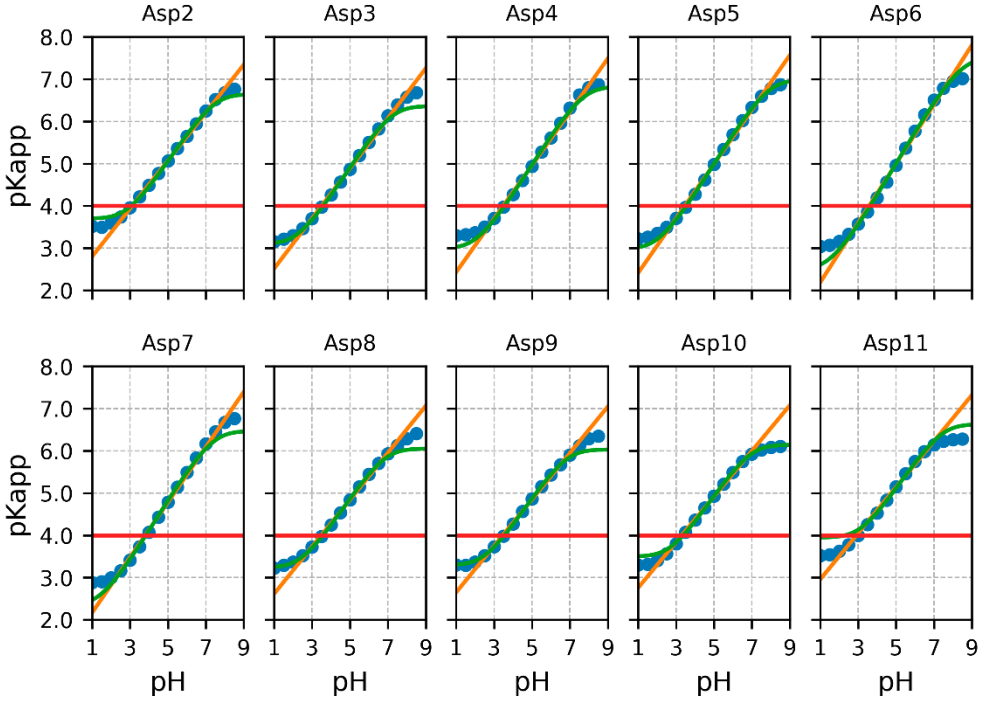


Figure 8. $pk_{app,i}$ distributions estimated from the deprotonated state fractions obtained from pH-REMD simulations (blue circles). The red lines are the Langmuir isotherm for an ideal case estimated using the reference $pk_{a,i}$ defined by AMBER. The orange and green lines are the Hill/LF and Frumkin isotherms, respectively.

	$pk_{D,LF} \pm \sigma_{LF}$	$n_{LF} \pm \sigma_{LF}$	$pk_{a,F} \pm \sigma_F$	$\delta_F \pm \sigma_F$	$pk_{m,LF}^{dist}$	σ_{LF}^{dist}	$pk_{m,F}^{dist}$	σ_F^{dist}
PEPTIDE	4.87 ± 0.03	0.38 ± 0.01	3.21 ± 0.02	3.79 ± 0.04	4.87	1.93	4.86	1.53
ASP 2	5.17 ± 0.02	0.43 ± 0.01	3.70 ± 0.04	3.39 ± 0.08	5.17	1.64	5.17	1.42
ASP 3	4.72 ± 0.03	0.41 ± 0.01	3.10 ± 0.05	3.76 ± 0.11	4.72	1.77	4.73	1.52
ASP 4	4.89 ± 0.05	0.37 ± 0.01	2.99 ± 0.07	4.41 ± 0.07	4.89	1.99	4.91	1.70
ASP 5	4.98 ± 0.05	0.36 ± 0.01	2.98 ± 0.04	4.61 ± 0.09	4.98	2.07	4.98	1.75
ASP 6	4.99 ± 0.09	0.30 ± 0.02	2.49 ± 0.11	5.79 ± 0.24	4.99	2.53	5.00	2.07
ASP 7	4.37 ± 0.06	0.35 ± 0.01	2.33 ± 0.10	4.77 ± 0.20	4.37	2.14	4.40	1.80
ASP 8	4.64 ± 0.03	0.44 ± 0.01	3.24 ± 0.03	3.24 ± 0.06	4.64	1.59	4.64	1.37
ASP 9	4.66 ± 0.02	0.45 ± 0.01	3.29 ± 0.02	3.16 ± 0.04	4.66	1.56	4.66	1.35
ASP 10	4.82 ± 0.02	0.46 ± 0.01	3.50 ± 0.03	3.05 ± 0.06	4.82	1.52	4.82	1.32
ASP 11	5.30 ± 0.04	0.45 ± 0.02	3.95 ± 0.07	3.09 ± 0.14	5.30	1.54	5.29	1.33

Table 2. pk and parameters of the Langmuir-Freundlich (LF) and Frumkin (F) isotherms. pk_D is pk_{app} at $\alpha = \theta = 0.5$ and n is the Hill coefficient in the LF isotherm. The intrinsic pk_a and δ parameter are given for the Frumkin isotherm. The pk_m^{dist} and σ^{dist} are the average and standard deviation of the pk_{app} distribution of the LF or F isotherms of the fitting.

Up to this point we have studied the proton-amino acid binding process of an amino acid i in the peptide, but the above approaches can also be applied to the proton equilibrium of the whole peptide in Eq. 8. The binding and dissociation constants, logK_c and pK_{app}, can also be estimated from the complexation isotherms. For this, we use the same formalism described above, but this time our receptor becomes the polyaspartic acid peptide and the deprotonated fraction is calculated from the average of the degree of dissociation of the amino acids at each solvent pH, where $\alpha = \frac{\sum_i \alpha_i}{N}$ and N is the number of titratable side chains. Therefore, the dependence of α and pK_{app} of the peptide on the solvent pH can be estimated using Eq. 9 and the Hill/LF and Frumkin isotherms.

$$-S^- + H^+ \rightleftarrows -SH \qquad K_c = \frac{[-SH]}{[-S^-] \cdot a_{H^+}} \qquad \textbf{Eq. 3.8}$$

$$pK_{app} \equiv logK_c = pH + \begin{cases} \log\left(\frac{\theta}{1-\theta}\right) \\ log\left(\frac{1-\alpha}{\alpha}\right) \end{cases} \qquad \textbf{Eq. 3.9}$$

In this case, K_c is the average binding affinity of the proton (or ligand) at a given pH. That is, the average of the microstates of binding of the proton to the different binding sites of the polyaspartic acid. Thus, the electrostatic interactions and the availability of the binding sites depend on the pH of the solvent, regardless of the binding site to which the proton binds within the chain. The ideal conditions for the peptide are therefore only satisfied when the binding sites are identical, i.e., there is no heterogeneity in the chain, and there are no interactions with neighbouring amino acids. Under these conditions, it is satisfied that the $logK_c = pK_{app} \neq f(pH)$, resulting in $pK_{app} = logK_c = logK$ where K is the intrinsic binding constant of all binding sites ($K = k_i$). However, the polyaspartic acid cannot be treated under ideal conditions because the strong electrostatic interactions of the amino acids depend on their position in the chain and cause the deviation from the ideal case. Therefore, the titration curves of the amino acids of the polyaspartic acid peptide are described by heterogeneous isotherms and the interpretation of the peptide protonation must be explained by the average of the individual isotherms of each aspartic acid within the chain.

After fitting α in Figure 9, the Hill/LF isotherm predicts a pK_D = 4.87 ± 0.03 and n = 0.38 ± 0.01, while the Frumkin isotherm gives us an intrinsic pK_a = 3.21 ± 0.02 and the interaction parameter δ = 3.79 ± 0.04. Again, the Frumkin isotherm fits better than

the Hill/LF model at the ends of the simulated f_d's. In this case, the LF/Hill isotherm shows negative cooperativity (n < 1) for the peptide and the δ parameter of the Frumkin isotherm suggests strong electrostatic interactions, thus confirming that the solvent pH plays a key role in the pK_{app} distribution of the peptide. Indeed, the intrinsic pK_a is far from the reference pk_a of the individual amino acids ($pk_{a,i}$ = 4.0) and, even under the conditions of the fully protonated peptide (pH < 2), the intramolecular interactions and chain heterogeneity still determine the affinity of the proton to bind to the peptide. Therefore, the presence of other negatively charged amino acids makes the polyaspartic acid become a stronger acid because it tends to donate the initial protons more easily (intrinsic pK_a < reference pk_a) compared to the free amino acid due to the intramolecular electrostatic interactions.

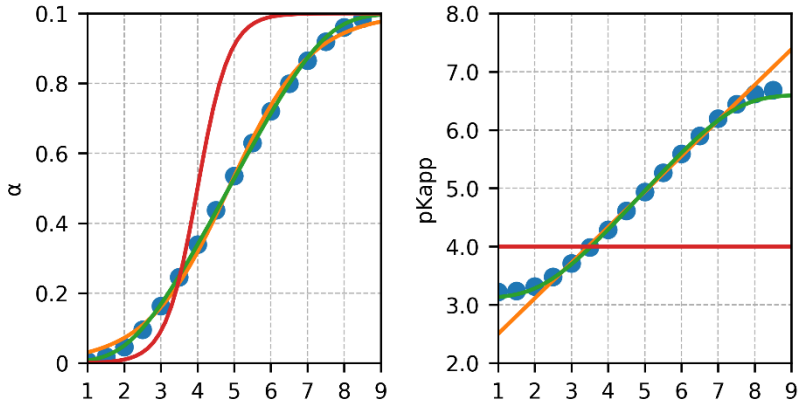


Figure 9. Titration curves (left) and pK_{app} distribution (right) of the polyaspartic acid. The red lines are the Langmuir isotherm for an ideal case estimated by reference $pk_{a,i}$ defined by AMBER. The orange and green lines are the Hill/LF and Frumkin isotherms, respectively, after fitting the data extracted from the simulation shown in blue circles.

Through the case of the polyaspartic acid, we have highlighted the advantages of using the Hill/LF and Frumkin isotherms to fit the deprotonation fractions of the pH-REMD simulation for both individual amino acids and the peptide. For the individual amino acids, the Hill/LF isotherm allows us to correctly predict the $pk_{app,i}$ curves except at the limits of α_i, and it directly provides us with the $pk_{D,i}$, i.e., the $pk_{app,i}$ when the amino acid is half occupied, which is a common value used to estimate the protonation state of amino acids in CMD simulations. The behaviour of $pk_{D,i}$ as a function of the position of the residues in the chain appears to be more complex and is difficult to

correlate with other properties of the peptide. On the other hand, the Frumkin isotherm is outstanding in its ability to predict the titration curve and the apparent $pk_{a,i}$ distribution even at low α_i. The intrinsic $pk_{a,i}$ values indicate that the residues located in the middle of the peptide sequence are lower compared to the residues at the ends of the chain. Furthermore, the titration curve of the peptide is also estimated from the isotherms, using the protonation fractions of each amino acid as average values that the protons experience when bound to the peptide. The observations made for each isotherm are also repeated in the case of the peptide, and the pK_{app}, pK_D and intrinsic pK_a are predicted. Furthermore, the average, $pk_{m,i}^{dist}$, and standard deviations, σ_i^{dist}, of the distributions of $pk_{app,i}$ and pK_{app} are in agreement between the Hill/LF and Frumkin isotherms, providing confidence in the estimations made in this work.

3.3. Conclusions

pH-REMD simulations can be a powerful technique to extend the conformational and protonation state sampling over a wide range of solvent pHs. In this work, we observed in the PCA energy maps that polyaspartic acid decapeptide shifts from homogeneous conformational sampling at strong acidic pH conditions to the promotion of a set of conformations as the solvent pH becomes mildly basic. The structural properties suggest that this behaviour is given by an increase in α-helix formation, which we also identify in the CpHMD simulation of $AS4_{10}$ at pH = 10 with explicit solvent (Section 3.1). $AS4_{10}^{pH=10}$ has a large α-helix content above 60%, whereas $AS4_{10}$ at mildly basic conditions exhibits an α-helix content below 45%. Therefore, the conformational space has apparently not converged even at pH conditions higher than the $pk_{D,i}$ + 2, highlighting the importance of considering the effect of the solvent pH in the pH-responsive biomolecules even when the pH conditions are apparently far from the effective $pk_{D,i}$. In addition to the structural analysis, this chapter also emphasises the benefits of using pH-REMD to predict the effective $pk_{D,i}$ and the intrinsic $pk_{a,i}$ of the titratable amino acids using the Hill/LF and Frumkin isotherms. The isotherms also predicted the pH-dependent $pk_{app,i}$ distribution for each amino acid, which provides more information for accurate protonation state prediction than the effective $pk_{D,i}$, i.e., the solvent pH at which the titratable amino acid is in the 50/50% ionised and neutral states, respectively. The intrinsic $pk_{a,i}$ values suggest that the carboxyl group of the aspartic acids releases the proton more readily as the amino acids are closer to the centre

of the peptide chain. Furthermore, the overall pK_D and pK_a of the peptide has also been estimated from the isotherms and average values of the individual amino acids.

On the other hand, we found some concerns in the evaluation of the CpHMD method. First, the implicit solvent simulations are not able to reproduce the conformational space of the CMD simulations. A clear conformational bias is observed depending on the simulation method, indicating a strong limitation of the CpHMD method in this solvent condition. The explicit solvation simulations are spared from this shortcoming since the peptide in the protonated state agrees in the structural properties when both CMD and CpHMD methods are compared. In the deprotonated state, the conformational and energetic observations are again in disagreement, apparently motivated by a different spatial distribution of the Na^+ counterions during the simulation. The results reported so far raise serious concerns about the accuracy of the CpHMD method implemented in AMBER18 with respect to the conformational description of the ionisable amino acids. This conflicts with the demonstrated ability of the CpHMD to predict the effective $pk_{D,i}$ or the intrinsic $pk_{a,i}$ of buried amino acids in proteins, since protonation states are tightly coupled to the conformational sampling (and vice versa). If the method shows inaccuracies in the conformations obtained from the simulations, large errors in the predictions of the effective $pk_{D,i}$ and intrinsic $pk_{a,i}$ should be expected. Moreover, this shortcoming becomes critical if pH-dependent conformational ensembles or ligand-protein mechanisms (and so many other biological events) are intended to be captured by the discrete constant pH method implemented in AMBER. Therefore, there is an urgent need to address this issue in the context of the CpHMD simulations in order to improve the ability to accurately sample both conformational and protonation fractions and, ultimately, to better predict pk_a.

3.4. Bibliography

1. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol Syst Des Eng* **2**, 9–33 (2017).

2. Harris, R. C., Tsai, C.-C., Ellis, C. R. & Shen, J. Proton-Coupled Conformational Allostery Modulates the Inhibitor Selectivity for β-Secretase. *J Phys Chem Lett* **8**, 4832–4837 (2017).

3. Radak, B. K. *et al*. Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *J Chem Theory Comput* **13**, 5933–5944 (2017).

4. Huang, Y., Yue, Z., Tsai, C.-C., Henderson, J. A. & Shen, J. Predicting Catalytic Proton Donors and Nucleophiles in Enzymes: How Adding Dynamics Helps Elucidate the Structure–Function Relationships. *J Phys Chem Lett* **9**, 1179–1184 (2018).

5. Machuqueiro, M. & Baptista, A. M. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins* **72**, 289–298 (2008).

6. di Russo, N. v., Estrin, D. A., Martí, M. A. & Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput Biol* **8**, e1002761 (2012).

7. Mcdougal, O. M., Granum, D. M., Swartz, M., Rohleder, C. & Maupin, C. M. pK a Determination of Histidine Residues in α-Conotoxin MII Peptides by 1 H NMR and Constant pH Molecular Dynamics Simulation. *J Phys Chem B* **117**, 2653–2661 (2013).

8. Ma, S., Henderson, J. A. & Shen, J. Exploring the pH-Dependent Structure-Dynamics-Fuction Relationship of Human Renin. *J Chem Inf Model* **61**, 16 (2021).

9. Beroza, P., Fredkin, D. R., Okamura, M. Y. & Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. *Proc Nat Acad Sci USA* **88**, 5804–5808 (1991).

10. Mertz, J. E. & Pettitt, B. M. Molecular Dynamics at a Constant pH. *Int J High Perform Comput Appl* **8**, 47–53 (1994).

11. Baptista, A. M., Martel, P. J. & Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins* **27**, 523–544 (1997).

12. Lee, M. S., Salsbury, F. R. & Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738–752 (2004).

13. Donnini, S., Tegeler, F., Groenhof, G. & Grubmüller, H. Constant pH Molecular Dynamics in Explicit Solvent with λ-Dynamics. *J Chem Theory Comput* **7**, 1962–1978 (2011).

14. Goh, G. B., Hulbert, B. S., Zhou, H., Brooks III, C. L. & Brooks, C. L. Constant pH Molecular Dynamics of Proteins in Explicit Solvent with Proton Tautomerism. *Proteins* **82**, 1319–1331 (2014).

15. Baptista, A. M., Teixeira, V. H. & Soares, C. M. Constant-pH molecular dynamics using stochastic titration. *J Chem Phys* **117**, 4184 (2002).

16. Bürgi, R., Kollman, P. A. & van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins* **47**, 469–480 (2002).

17. Mongan, J., Case, D. A. & McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* **25**, 2038–2048 (2004).

18. Meng, Y. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J Chem Theory Comput* **6**, 1401–1412 (2010).

19. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).

20. Hofer, F., Kraml, J., Kahler, U., Kamenik, A. S. & Liedl, K. R. Catalytic Site p K_a Values of Aspartic, Cysteine, and Serine Proteases: Constant pH MD Simulations. *J Chem Inf Model* **60**, 3030–3042 (2020).

21. Swails, J. M. & Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J Chem Theory Comput* **8**, 4393–4404 (2012).

22. Case, D. A. *et al.* Amber 2018. *University of California, San Francisco* (2018).

23. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

24. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383–394 (2004).

25. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

26. Yeager, A. v, Swails, J. M. & Miller, B. R. Improved Accuracy for Constant pH-REMD Simulations through Modification of Carboxylate Effective Radii. *J Chem Theory Comput* **13**, 4624–4635 (2017).

27. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicist*. (1999).

28. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 2017).

29. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089 (1998).

30. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

31. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

32. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

33. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

34. Williams, T. *et al.* Gnuplot 4.6: an interactive plotting program. *Software Manual* 238 (2012).

35. Landsgesell, J. *et al.* Simulations of ionization equilibria in weak polyelectrolyte solutions and gels. *Soft Matter* **15**, 1155–1185 (2019).

36. Milorey, B., Schwalbe, H., O'Neill, N. & Schweitzer-Stenner, R. Repeating Aspartic Acid Residues Prefer Turn-like Conformations in the Unfolded State: Implications for Early Protein Folding. *J Phys Chem B* **125**, 11392–11407 (2021).

37. Rybka, K. *et al.* Disorder and order in unfolded and disordered peptides and proteins: A view derived from tripeptide conformational analysis. II. Tripeptides with short side chains populating asx and β-type like turn conformations. *Proteins* **81**, 968–983 (2013).

38. Wallace, J. A. *et al.* Toward accurate prediction of pKa values for internal protein residues: The importance of conformational relaxation and desolvation energy. *Proteins* **79**, 3364–3373 (2011).

39. Weiss, J. N. The Hill equation revisited: uses and misuses. *FASEB J* **11**, 835–841 (1997).

40. Lluís Garcés, J. *et al.* Complexation isotherms in metal speciation studies at trace concentration levels. Voltammetric techniques in environmental samples. *Phys Chem Chem Phys* **4**, 3764–3773 (2002).

41. Garcés, J. L., Koper, G. J. M. & Borkovec, M. Ionization Equilibria and Conformational Transitions in Polyprotic Molecules and Polyelectrolytes. *J Phys Chem B* **110**, 10937–10950 (2006).

42. Companys, E. *et al.* Electrostatic and specific binding to macromolecular ligands. *Colloids Surf A Physicochem Eng Asp* **306**, 2–13 (2007).

43. Casasnovas, R., Ortega-Castro, J., Frau, J., Donoso, J. & Muñoz, F. Theoretical pKa calculations with continuum model solvents, alternative protocols to thermodynamic cycles. *Int J Quantum Chem* **114**, 1350–1363 (2014).

44. Dutra, F. R., Silva, C. de S. & Custodio, R. On the Accuracy of the Direct Method to Calculate p K_a from Electronic Structure Calculations. *J Phys Chem A* **125**, 65–73 (2021).

This chapter has been published in Privat, C., Madurga, S., Mas, F. Rubio-Martínez, J. On the Use of Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides. *Polymers* **13**, 99 (2021)

Chapter 4

On the Use of the Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides

Highlighting the importance of solvent pH in charge-structure coupling for protein conformational space sampling during MD simulations is one of the main objectives of this thesis. Indeed, due to the wide range of pH values that oscillate in the human body, from 4.5 in lysosomes to 8.0 in mitochondria[1], the ability of pH to modulate some biomolecules is central for an in-depth study of the biological functions of pH-responsive proteins. Under this premise, the available approaches that include the effect of pH in conventional MD (CMD) simulations, such as the use of pKa prediction tools (PROPKA[2] or H++[3]), have been insufficient until recently. In response to the need for more accurate modelling of the ionisable biomolecules depending on the environmental pH, a collection of MD methods *at constant pH* has emerged over the last decades under the name of Constant pH Molecular Dynamics (CpHMD) techniques[4–7]. These methods introduce the dynamic change of the protonation states of the titratable amino acids (and other biomolecules if necessary) during the course of simulations by setting the semi-grand canonical ensemble ($N\Delta\mu VT$). The most successful approaches are mainly (i) the methods with discrete protonation states sampled by a stochastic criterion, the so-called discrete CpHMD[5,6,8–11], and (ii) the methods that describe the protonation states by introducing protonation coordinates into the potential energy function, also known as continuous CpHMD methods[7,12–15]. Further details of each technique and the various methods developed can be found in theoretical background (Section 2.5).

Both continuous and discrete CpHMD methods have shown promising results in the prediction of amino acid pKa and conformational sampling of proteins, as well as the role of pH in ligand-protein mechanistic studies[16–21]. However, some shortcomings have

also been pointed out, such as the lack of accuracy in the description of some physical properties or the trapping of the molecular systems in local minima. Some papers have reported the strengths and weaknesses of these methods[22–27], while some reviews have summarised their development in the past[28,29] and present[30] in order to better comprehend the current limitations of constant pH simulations. Fortunately, the CpHMD methods have been refined over the years through modified force fields, water models or algorithm improvements, but also by adopting enhanced-sampling approaches to overcome the drawbacks of lack of convergence and sampling[10,11,14,31–35]. Several methods of CpHMD have been implemented in popular simulation packages, such as CHARMM[36], GROMACS[37] or AMBER[38], and the application of these techniques is gaining more and more followers nowadays. In particular, both continuous[35,39–42] and discrete[8,11] CpHMD methods and the enhanced-sampling pH-based replica exchange Molecular Dynamics (pH-REMD) method[18] have been implemented in the AMBER package, which is a well-known simulation package for the simulation of biomolecules.

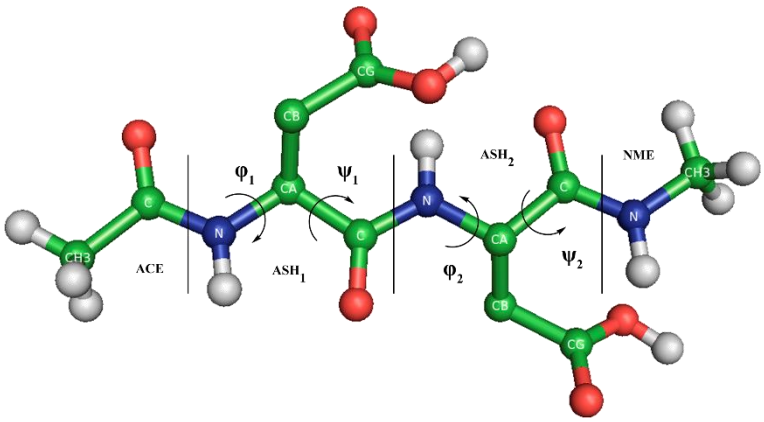


Figure 1. Capped Asp_2 tripeptide in the syn-O2 protonated state. The amino acids, capping groups and dihedral angles φ and ψ are indicated. The θ angle is formed by the CG_1, CA_1, CA_2, and CG_2 atoms. Non-polar hydrogens of the amino acids are hidden. Subscripts refer to monomers 1 and 2.

In line with the simulations performed in Chapter 3, the discrete constant pH method is used in this work because the advantages of the explicit description of the protonated states during the conformational sampling of the molecular systems. However, due to the shortcomings concerning the poor reproducibility of the conformational sampling and the structural properties in CpHMD simulations compared with CMD method presented in the previous chapter, we now present a fundamental revision of this method from the basics in order to clarify the reported limitations. As the inclusion of

CpHMD methods in protein studies is becoming increasingly common, this chapter examines the accuracy of the CpHMD method with discrete protonation states implemented in AMBER on simple biomolecules from a conformational perspective. For this purpose, and after finding that implicit solvation models show larger deviations in the CpHMD simulations, capped di(amino acid) tripeptides with six titratable amino acids were constructed to perform simulations on the microsecond scale using a Generalised Born model for implicit solvation in both CMD and CpHMD methods. The titratable tripeptides were simulated under strong acidic or basic pH conditions to ensure a fully protonated or deprotonated state, whereas the CMD simulations were carried out with a fixed protonation state. In this manner, the tripeptide simulations can be compared at solvent pH conditions that result in a similar protonation state, regardless of the method. Ramachandran maps and energy contributions from the tripeptide trajectories were therefore analysed to find the source of the deficiencies observed in the previous chapter. Thus, in this chapter we discuss some of the successes and weaknesses of the CpHMD with discrete protonation states in an implicit solvation model implemented in the AMBER18 version.

4.1. Materials and Methods

4.1.1. Capped Tripeptides

Tripeptides (ACE-X-X-NME, hereinafter X_2) capped at the extremes by the acetyl (ACE) and N-methyl (NME) groups were constructed for the protonated, deprotonated and titratable residues of X, where X = {lysine, tyrosine, cysteine, histidine, glutamic acid, aspartic acid} amino acids (Figure 1). The residues LYN, CYM, HID, HIE, GLU, and ASP were used for the deprotonated forms and the residues LYS, TYR, CYS, HIP, GLH and ASH for the protonated forms in the CMD simulations. The titratable peptides in the CpHMD method were built using the residues LYS, TYR, CYS, HIP, GL4 and AS4 (using the AMBER convention). ff14SB force field[43] and constph.lib (only in CpHMD) were loaded into the LEaP module of AMBER18. The CPIN file was then generated for the titratable systems, specifying the initial protonation state according to the solvent pH and the Generalised Born (GB) model of Onufriev et al.[44] (igb = 2). The lysine, tyrosine and cysteine amino acids have two possible protonation states: the deprotonated and protonated forms. Histidine has up to three protonation states, which

Residue	CMD	CpHMD			Prot. State.	Intrinsic pKa
		pH 1	pH 12	pH 14		
ASP	✓				D	4.0
ASH	✓				P	
AS4		✓	✓		T	
GLU	✓				D	4.4
GLH	✓				P	
GL4		✓	✓		T	
HIE	✓				D	7.1 (ε) 6.5 (δ)
HID	✓				D	
HIP	✓	✓	✓		P/T	
CYM	✓				D	8.5
CYS	✓	✓	✓		P/T	
TYR	✓	✓			P/T	9.6
LYN	✓				D	10.4
LYS	✓	✓		✓	P/T	

Table 1. Simulations performed for each residue type and method. Protonation state is defined as deprotonated (D), protonated (P) or titratable (T) form. Some residues can be used to generate both protonated CMD simulations and titratable peptides in the CpHMD method. The intrinsic pKa values of the side chains are used according to Mongan et al.[8] in the AMBER implementation.

are classified into the deprotonated and protonated forms: the doubly protonated HIP state for the protonated form, and the ε (HIE) and δ (HID) states for the deprotonated (or neutral) histidine. HIE and HID are defined by the position of the hydrogen on the N-epsilon and N-delta nitrogen, respectively, in the neutral form. The δ-state was chosen as the initial protonation state for the CpHMD simulations of histidine. Finally, the glutamic acid and aspartic acid can be found in the deprotonated form or up to four states in the protonated form. These protonated states depend on the position and orientation of the hydrogen (*syn* or *anti*) when one of the oxygen atoms of the carboxyl group (O1 or O2) is protonated. The four protonatable sites of the side chain of the residue AS4 in Appendix B, Figure B1. State 1 (syn-O2 protonation) was chosen as the initial protonation state in the CpHMD simulations at acidic solvent pH conditions, which is the default protonated

state in the CMD method. Counterions were implicitly considered in the solvation model with an ionic strength of 0.1 M.

4.2.2. Simulation Setup

Each system was minimised according to a three-stage protocol with different restraints: (i) on all atoms, (ii) on the backbone atoms, and (iii) on the free system. 5000 steps (maximum) of the steepest descent method[45] were performed per stage. Restraints were introduced with force constants of 5.0 $kcal \cdot mol^{-1} \cdot A^{-2}$. In the titratable systems, the implicit CpHMD method[8] (icnstph = 1) was turned on to define the protonation state of the amino acids, but without changing the protonation states (ntcnstph > 5.000).

After the minimisation step, a heating simulation was performed by linearly increasing the temperature (1 $K \cdot ps^{-1}$) of the capped tripeptide up to 300 K. The system was then equilibrated by keeping the tripeptide at 300 K for 200 ps in the isobaric-isothermal ensemble (NPT). To increase the conformational exploration[46], four replicas were generated for each system using the final coordinates of the equilibration step but resetting the initial velocities. 1 μs per replica were performed with the implicit solvent method, using the Generalised Born model of Onufriev et al. (igb = 2) and an ionic strength of 0.1 M. The SHAKE algorithm[47] constrained the bond lengths. A Langevin thermostat[48] with a collision frequency of 3 ps^{-1} was chosen for the thermal bath and no periodic boundary conditions (PBCs) were required. For the titratable simulations, an implicit CpHMD method was used with a frequency of protonation state change attempt of 0.01 ps^{-1} (ntcnstph = 5). Strong pH conditions were set to ensure a dominant protonation state during CpHMD simulations, with pH values of 12.0 and 1.0 chosen for the deprotonated and protonated forms, respectively. The only exception was the capped Lys_2 tripeptide, which required a higher basicity in the solvent (pH 14.0). Table 1 summarises the residue type, simulation method and solvent pH of the production runs.

4.1.2. Energy and Conformational Analysis

The energies, coordinates and output files were updated every 2, 10 and 20 ps, respectively. The energy terms and normalised histograms of each term were calculated using the CPPTRAJ module[49]. The dihedral angles (φ, ψ and an angle related to the orientation of the side chains with respect to the C_α atoms, hereinafter referred to as *θ angle*) were also obtained with CPPTRAJ. An in-house tool transformed the dihedral

angles generated during the simulation into Gibbs free energies using Eq. 1, thus facilitating the construction of the potential energy surface in the Ramachandran maps[50].

$$\Delta G = -k_B T ln(N_i/N_{max}) \qquad \textbf{Eq. 4.1}$$

where k_B is the Boltzmann constant, T is the temperature, and N_{max} and N_i are the maximum population and the population of a cell i after applying a grid to the distribution of the dihedral angles φ and ψ with a spacing of 1°. The Ramachandran map was divided into nine conformational regions (C_5, P_{II}, α_D, β_2, C^7_{axial}, α_L, α', α_R and C^7_{eq}) according to the Rubio-Martinez et al.[51] in Figure B2 and the global populations in each conformational region were calculated. Each amino acid of the tripeptides was analysed separately, resulting in two sets of conformational data corresponding to the N-terminal amino acid (set 1) and the C-terminal amino acid (set 2). The minima of the Ramachandran maps were located using a larger grid spacing (2°) to reduce the apparition of false minima. All plots were generated using GNUPLOT (version 4.6)[52].

4.2. Results and Discussion

4.2.1. Gibbs Free Energies in Ramachandran Space

The conformational sampling of each system was analysed by means of the Ramachandran map. Since the capped tripeptides have two amino acids with their backbone dihedral angles φ and ψ (Figure 1), the pair φ/ψ dihedrals dihedral pair of each monomer (the N-terminal and C-terminal amino acid) was represented. The reported results of the simulations start with the basic pKa amino acids, continue with the specific case of the histidine and end with those with a carboxyl group in the side chain.

4.2.1.1. Basic pKa Amino Acids

In this group we include those titratable amino acids with an intrinsic pKa greater than 7.0. The conformational sampling of this set of capped tripeptides is represented in the Ramachandran maps for each simulation condition (CMD at the top and CpHMD at the bottom). The LYS systems are shown in Figure 2, and TYR and CYS are found in Figure B3 and Figure B4. The deprotonated form of tyrosine is not available in the AMBER libraries for the CMD method, so only the simulations of the TYR system in the protonated form were performed. However, the partial charges of the deprotonated tyrosine can be calculated as it has been proven to play an important role in the conformation of some proteins[53].

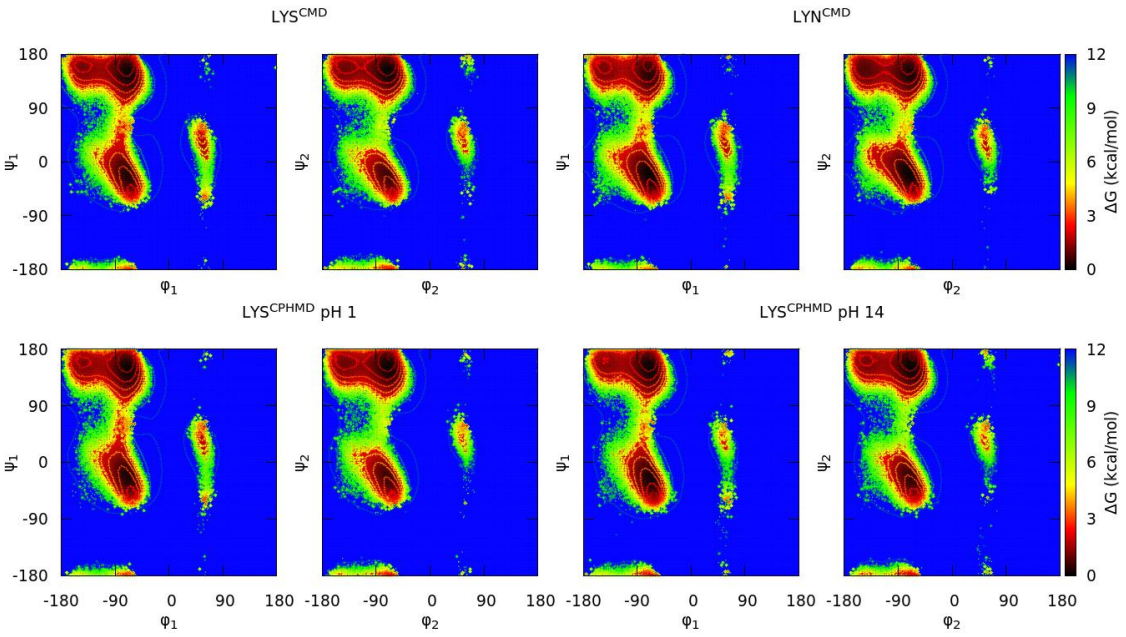


Figure 2. Gibbs free energies in Ramachandran space of the capped Lys_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. The protonated forms are on the left (CMD; top—CpHMD; bottom) and the deprotonated forms are on the right (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

Comparison of the two simulation methods in the Ramachandran plots shows that the LYS protonated forms (LYS^{CMD} and LYS^{CpHMD} at pH 1) are in agreement. Instead, the deprotonated simulations (LYN^{CMD} and LYS^{CpHMD} at pH 14) exhibit smooth differences in the depth of the minima. For further clarification, the conformational profile of the capped tripeptides was studied by dividing the Ramachandran map into nine regions according to Rubio-Martinez et al., which are associated with a specific conformation (C_5, P_{II}, α_D, β_2, C_7^{axial}, α_L, α', α_R, and C_7^{eq}). By calculating the population of each region, the conformational propensities of each amino acid were estimated. The population ratios allow a quantitative analysis of the conformational sampling of the simulation methods by identifying the most stable regions. The populations of these regions for each monomer are shown in Figure 3. In general, the P^{II} and α^R conformations predominate over all others. The protonated form of the LYS systems shows close population ratios between the counterparts (LYS^{CMD} and LYS^{CpHMD} at pH 1). The deprotonated CMD simulation (LYN^{CMD}) has a different population profile with respect

to the other systems, showing a behaviour far from the CpHMD analogue (LYS^{CpHMD} at pH 14). In contrast, LYS^{CpHMD} at pH 14 has similar conformational populations with respect to LYS^{CMD} and LYS^{CpHMD} at pH 1.

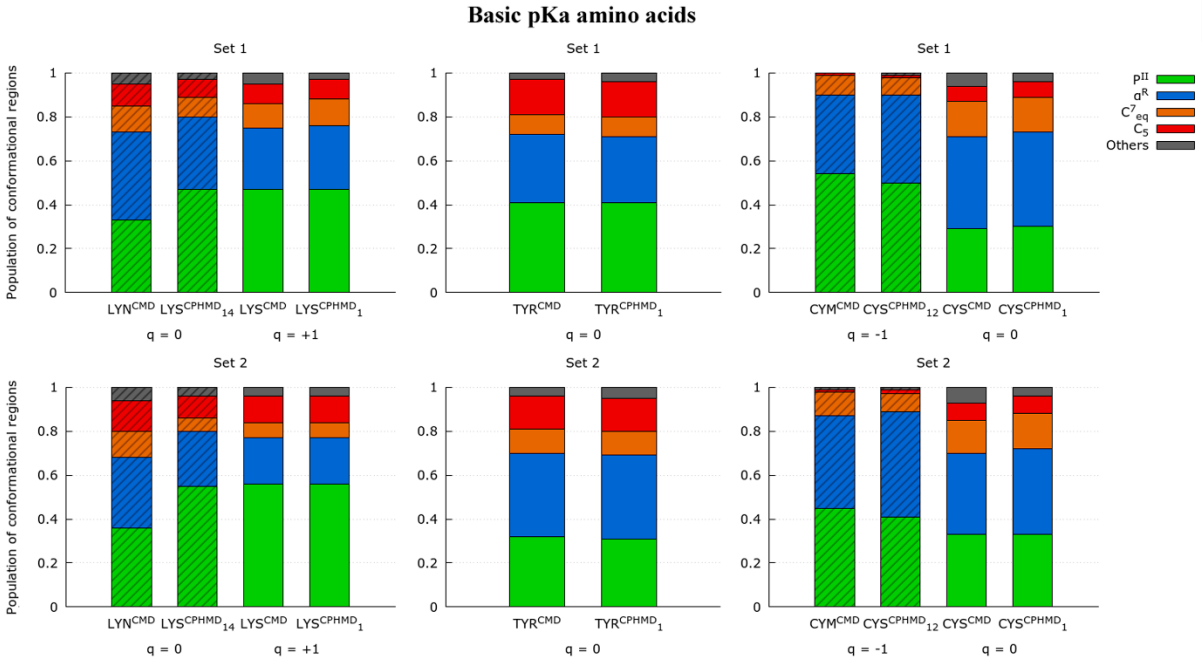


Figure 3. Ratio of the four most populated conformational regions (P^{II}, α^{R}, C^{7}_{eq}, and C_5 in green, blue, orange, and red, respectively) in the Ramachandran map of the amino acids LYS, TYR, and CYS. The labels indicate the residue, the simulation method (in the superscript), and the solvent pH (in the subscript, only for the CpHMD simulations). The subtitles indicate the set of dihedrals corresponding to monomer 1 (N-terminal) or monomer 2 (C-terminal amino acid). The net charge of the amino acids is indicated below the systems (q = −1, 0 or +1). The box style (striped or solid) indicates those peptides in the same protonation state, regardless of the method used. The classification 'others' (grey) includes the conformational regions β_2, α', α_D, α_L, and C_7^{axial}.

A good accordance for the protonated systems of the TYR amino acid (TYR^{CMD} and TYR^{CpHMD} at pH 1) is observed in Figure B3. Except for barely noticeable differences in the populations of minor conformational regions (C_7^{axial} and α_L), the Ramachandran maps and the population ratios shown in Figure 3 are in good agreement. For the CYS systems shown in Figure B4, the conformational profiles show a similar trend to the TYR systems. Therefore, a high consistency between CMD and CpHMD counterparts is also observed in the Ramachandran maps and population ratios for the protonated (CYS^{CMD} and CYS^{CpHMD} at pH 1) and deprotonated forms (CYM^{CMD} and CYS^{CpHMD} at pH 12).

These observations proved that the CpHMD method was generally consistent in the conformational sampling of these amino acids, except for the deprotonated LYS form. A first weakness is thus identified since the Ramachandran maps of the deprotonated LYS^{CpHMD} system were unable to reproduce the conformational profile of the well-established CMD method.

4.2.1.2. Histidine

This amino acid has pKa values of 6.5 and 7.1 for the δ and ε states, respectively. Depending on the position of the hydrogen in the neutral form, histidine can be found in the δ (N-delta atom) or the ε (N-epsilon atom) state. Thus, two protonation states coexist when the imidazole ring of the side chain becomes neutral, modulating the conformational sampling of the peptide depending on the position of the hydrogen during the simulation.

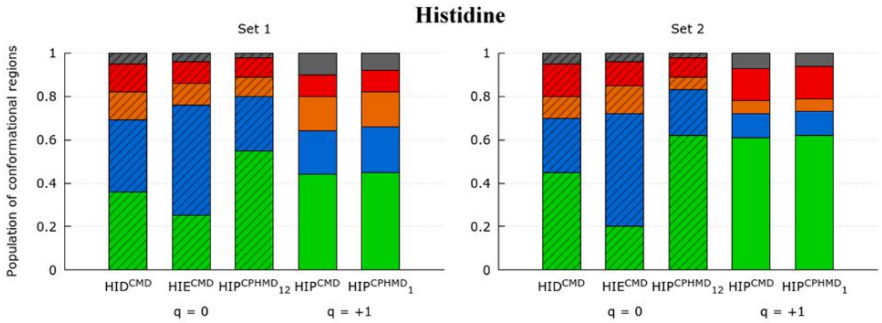


Figure 4. Ratio of the fourth most populated conformational regions (P^{II}, α^{R}, C^{7}_{eq}, and C_5 in green, blue, orange, and red, respectively) in the Ramachandran map of the HIS amino acid. The labels indicate the residue, the simulation method (in the superscript) and the solvent pH (in the subscript, only for the CpHMD simulations). The subtitles indicate the set of dihedrals corresponding to monomer 1 (N-terminal amino acid) or monomer 2 (C-terminal amino acid). The net charge of the amino acids is indicated below the systems (q = −1, 0 or +1). The box style (striped or solid) indicates those peptides in the same protonation state, regardless of the method used. The classification 'others' (grey) includes the conformational regions β_2, α', α_D, α_L, and C_7^{axial}.

The Ramachandran maps of the histidine simulations in Figure B5 illustrate the dihedral distribution obtained from the conformational sampling. The protonated peptides (HIP^{CMD} and HIP^{CpHMD} at pH 1) show similar conformational profiles in the Ramachandran maps. The population ratios confirm this observation: the HIP^{CMD} and HIP^{CpHMD} at pH 1 simulations show close population ratios in Figure 4. In contrast, the

deprotonated simulations (HIE^{CMD}, HID^{CMD} and HIP^{CpHMD} at pH 12) exhibit deviations in the depth of the minima of the Ramachandran plots. In addition, the population ratios of the CMD simulations (HIE^{CMD} and HID^{CMD}) are not in agreement with the HIP^{CpHMD} at pH 12 system. In this case, the HIP^{CpHMD} simulation at basic pH conditions has population ratios closer to the protonated form rather than to its CMD analogue. The population ratios of HIE^{CMD} and HID^{CMD} are far from being similar, suggesting that the position of the hydrogen in the N-epsilon and N-delta atom plays an important role in the conformational sampling of the deprotonated forms.

While the protonated forms are in good conformational agreement, the deprotonated forms of histidine indicate that the CpHMD method at basic pH conditions is unable to reproduce the conformational sampling of the CMD counterparts. As HIP^{CpHMD} coexists between the δ and ε protonation state in the neutral form at pH 12, one might expect a population profile resulting from the combination of the profiles of both states. Instead, the P^{II} conformation of the CpHMD systems at basic pH conditions behaves similarly to the protonated simulations, which is a fact that is also observed for the LYS systems.

4.2.1.3. Acidic Amino Acids

Glutamic acid and aspartic acid are two amino acids characterised by the four protonatable sites in the carboxyl group. Although both residues are similar, except for additional methyl group in the glutamic acid side chain which results in a shift in the pKa, the Ramachandran maps and population ratios do not behave similarly.

On the one hand, the conformational sampling of the GLU systems illustrated in the Ramachandran maps (Figure B6) follows the trend of the results observed for the LYS and HIS systems. The Ramachandran plots and population ratios of the protonated simulations (GLH^{CMD} and $GL4^{CpHMD}$ at pH 1) are in a good agreement in Figure B6 and Figure 5. However, this is not the case for the deprotonated systems (GLU^{CMD} and $GL4^{CpHMD}$ at pH 12), whose population ratios deviate significantly from each other. In fact, it is shown that $GL4^{CpHMD}$ at pH 12 has a similar population profile with respect to the GLH^{CMD} and $GL4^{CpHMD}$ at pH 1. This fact is no longer surprising, since it also occurs in previous systems (LYS and HIP).

On the other hand, the ASP peptides stand out since the protonated simulations (ASH^{CMD} and $AS4^{CpHMD}$ at pH 1) show a slight disagreement in the minima of the

Ramachandran maps, but not as pronounced as those observed in the deprotonated ones (Figure 6). Nevertheless, the population ratios in Figure 5 confirm that this disagreement is due to smooth differences in the population of each conformation (including C_7^{eq} and C_5). The deprotonated systems of the ASP amino acid (ASP^{CMD} and $AS4^{CpHMD}$ at pH 12) show a greater dissimilarity in the Ramachandran maps and population ratios.

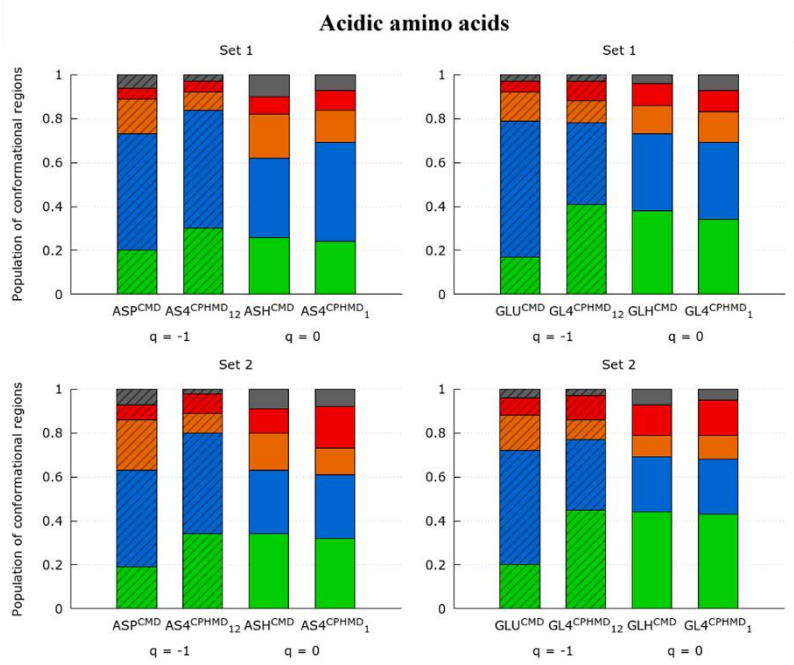


Figure 5. Ratio of the four most populated conformational regions (P^{II}, α^R, C^7_{eq}, and C_5 in green, blue, orange, and red, respectively) in the Ramachandran map of the ASP and GLU amino acids. The labels indicate the residue, the simulation method (in the superscript) and the solvent pH (in the subscript, only for the CpHMD simulations). The subtitles indicate the set of dihedrals corresponding to monomer 1 (N-terminal) or monomer 2 (C-terminal amino acid). The net charge of the amino acids is indicated below the systems (q = −1, 0 or +1). The box style (striped or solid) indicates those peptides in the same protonation state, regardless of the method used. The classification 'others' (gray) includes the conformational regions β_2, α', α_D, α_L, and C_7^{axial}.

Apart from the differences in the deprotonated forms, which are also observed in the previous amino acid sets, another factor apparently interferes by causing small changes in the conformational sampling of the protonated forms. These deviations could arise from the multiple protonatable position of the hydrogen when the side chains are protonated. In addition, the conformational sampling of ASP is probably more sensitive to the position of this proton given that the carboxyl groups of the successive aspartic

acids are closer compared to the GLU systems, which have an additional methyl group in the side chain.

The Ramachandran plots demonstrated the consistency of the CpHMD method in reproducing the conformational sampling of the protonated forms of the basic pKa, histidine and acidic amino acids. However, some shortcomings were noted for the deprotonated forms of all systems (except for CYS).

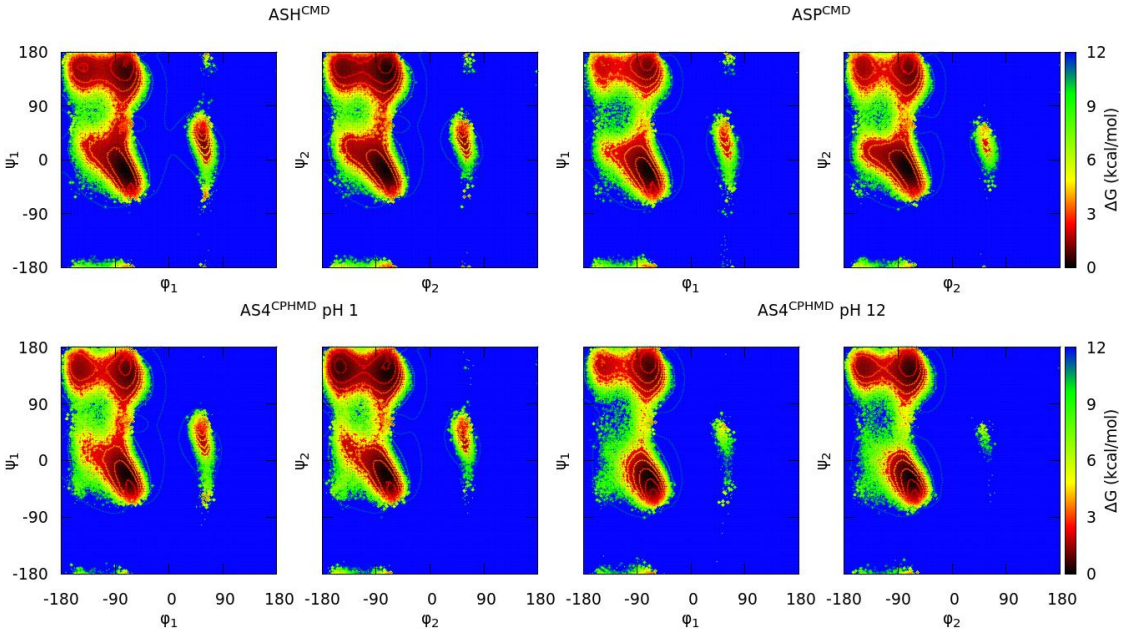


Figure 6. Gibbs free energies in the Ramachandran space of the capped Asp_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (only for the CpHMD simulations). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. The protonated forms are on the left (CMD; top—CpHMD; bottom) and the deprotonated forms are on the right (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

The main reason of this inconsistency in the deprotonated forms is the mismatch in the partial charges when comparing the CMD and CpHMD counterparts. Table B1 lists the partial charges of the individual acid atoms. Indeed, AMBER manifested that CpHMD residues always use the partial charges of the protonated form, called *reference* residue, in the backbone atoms and only change the partial charges of the side chain atoms when the residue reaches another protonation state[8]. It is therefore not surprising that the

electrostatic interactions are not fully reproducible when using the CpHMD method. We hope that this limitation can be overcome in future updates of the method.

On the other hand, the anchoring of the hydrogens in the CMD simulations with respect to the *dynamic* protons in the CpHMD method is another reason for the observed deviations. The CpHMD method has a hydrogen atom in all the protonatable sites during the simulation and activates them (by changing the partial charges of the side chain) according to the protonation state. Under this consideration, there are two scenarios: (i) the histidine and (ii) the acidic amino acids. For the histidine, the protonated form of the CpHMD method has the two hydrogen atoms activated as the reference residue (HIP) of the CMD method, so there is no difference between them. Therefore, the conformational sampling of the protonated simulations should be, and is, very similar. However, the deprotonated forms of histidine have different protonation state sampling. The HIP^{CpHMD} simulation at pH 12 coexist in the δ and ε states over time, whereas the CMD method fixes one protonation state state (HIE or HID) during the simulation. Apart from the failure to reproduce the electrostatics due to the partial charges, the deprotonated forms of the histidine are not entirely comparable due to the change in position of the activated hydrogen during the CpHMD simulations. The change in position of the hydrogen atom in the CpHMD simulations then leads to different conformational sampling compared to the CMD simulations, in which the hydrogen is fixed at the N-delta or N-epsilon atom positions.

The acidic amino acids present a similar problem, but this time in the protonated forms. These residues have four protonatable sites (the *anti* or *syn* position in each oxygen of the carboxyl group), which implies a greater allocation of the hydrogen atom when the side chain is protonated in comparison with the CMD method, in which the hydrogen is bound in the syn-O2 position. In fact, the populations of the protonation states during the CpHMD simulation were 96% and 4% (on average) for the *syn* and *anti* positions, respectively, with these percentages equally distributed between the two oxygen atoms. In the CMD simulations, the hydrogen is bonded to the O2 oxygen atom. Therefore, the change in position of the hydrogen is only achieved by rotating the bonds of the carboxyl group, which is more expensive compared to the CpHMD method. The CMD and CpHMD at pH 1 simulations are then not fully comparable due to the different sampling of the protonation states. However, the multiple protonatable positions of the CpHMD simulations are far from causing significant deviations in the conformational sampling of

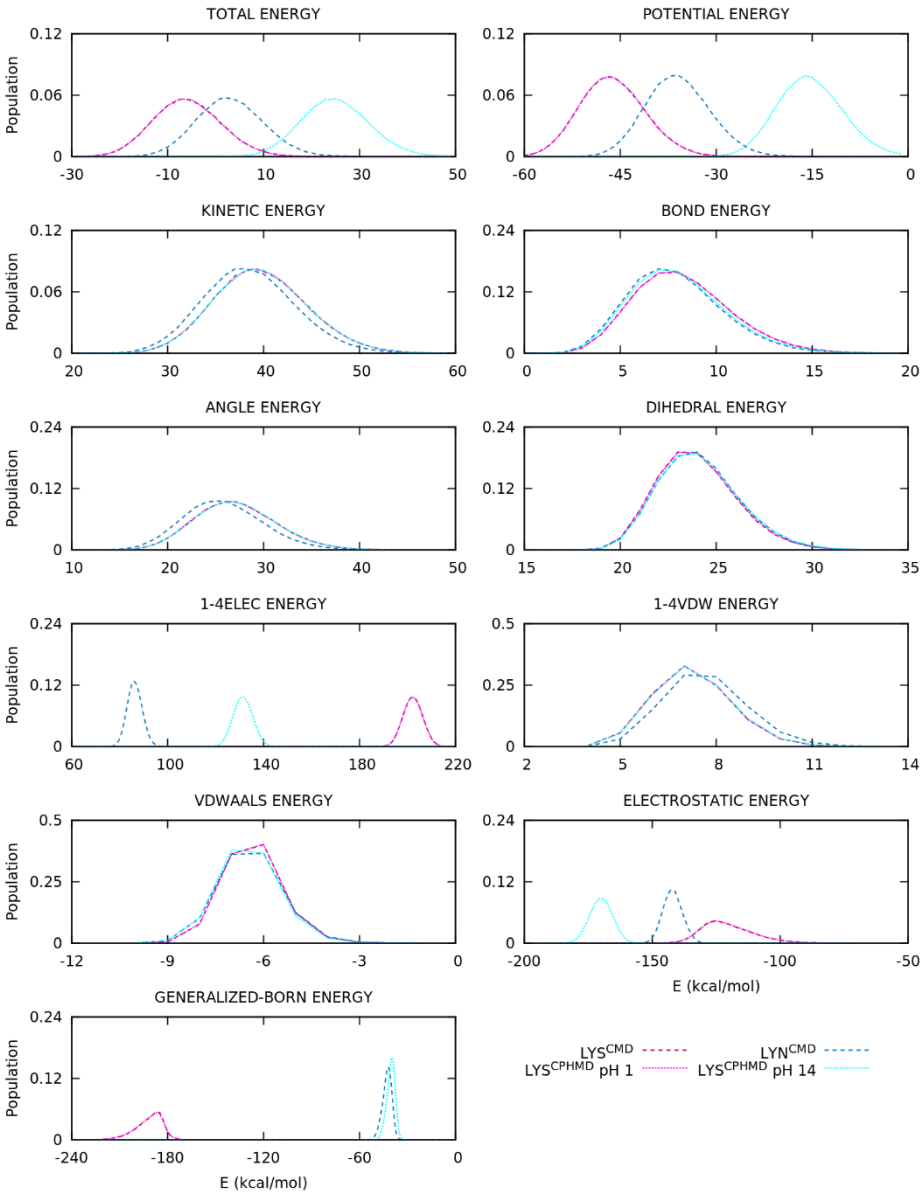


Figure 7. Energy distributions of the capped Lys_2 tripeptide. Global, inner, van der Waals, and electrostatic terms are shown. Dotted and dashed lines are CpHMD and CMD simulation methods, respectively.

the acidic amino acids as observed in the Ramachandran maps and population ratios reported above.

4.2.2. Energy Contributions

The energy terms of the AMBER's force field provide further information for the interpretation of the conformational sampling divergence. Therefore, normalised distributions of the contribution energies (total, kinetic, and potential, and each term of

the potential energy) were computed using the energy values from the simulation and plotted using GNUPLOT. The partition of the electrostatic energy into backbone and side chain contributions was also performed using CPPTRAJ to clarify the effects of the mismatch between the partial charges. This section focuses primarily on the electrostatic contribution, but other energy terms are also illustrated and some internal energies are highlighted during the analysis.

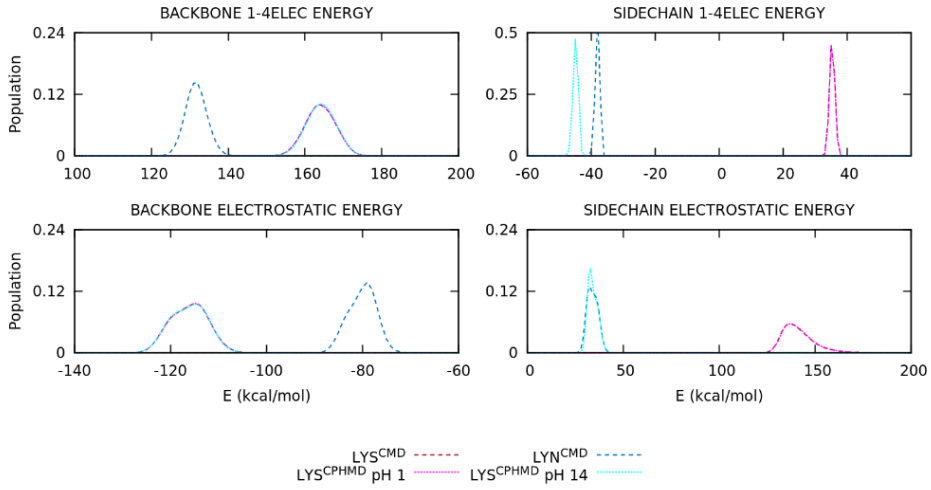


Figure 8. Energy distribution of the 1–4 and long-range electrostatics of the backbone and side chain atoms of the capped Lys_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

The energy distributions of the basic pKa amino acids are shown in Figure 7, Figure B7 and Figure B8 for the LYS, TYR, and CYS systems, respectively. In the LYS system, the overlapping of the protonated simulations (LYS^{CMD} and LYS^{CpHMD} at pH 1) is observed in all energy terms of Figure 7. On the contrary, the deprotonated LYS systems (LYN^{CMD} and LYS^{CpHMD} at pH 14) show a significant shift in the 1–4 electrostatic interactions, as well as in the long-range electrostatics, which has a distinct shape in the distribution. To understand the effect of the partial charges restriction in the implementation of the CpHMD method in AMBER, the electrostatic terms of all simulations were decomposed into backbone and side chain atoms. The separation of the electrostatics in the LYS systems reveals that the contribution of the protonated systems (LYS^{CMD} and LYS^{CpHMD} at pH 1) perfectly overlaps in both backbone and side chain atoms of the amino acid (Figure 8). However, a deviation is observed in both backbone electrostatic terms and the 1–4 electrostatics of the side chain distributions of the

deprotonated simulations (LYN^{CMD} and LYS^{CpHMD} at pH 14). This deviation in side chain electrostatics may be related to the partial charge of the C_β atom (Table B2).

In the TYR system, only the energy distributions of the protonated simulations (TYR^{CMD} and TYR^{CpHMD} at pH 1) are available in Figure B7. Both distributions overlap perfectly, as does the decomposition of the electrostatics in the Figure B9. These results are consistent with those observed in the Ramachandran maps. On the other hand, the energy distributions of the CYS systems also show a good overlap in the protonated simulations (CYS^{CMD} and CYS^{CpHMD} at pH 1) in Figure B8. However, the deprotonated systems (CYM^{CMD} and CYS^{CpHMD} at pH 12) display mild shifts in the total, potential, dihedral and 1–4 electrostatic energies, and different shapes in the 1–4 and long-range electrostatics. The decomposition of the electrostatics in the deprotonated simulations evidences a modest shift in the distributions of the electrostatics in both side chain and backbone atoms (Figure B10). The backbone electrostatics of CYS^{CpHMD} at pH 12 suggest that the deprotonated form modulates the conformational sampling in such a manner that the distribution shape ultimately becomes similar to that of CYM^{CMD}. Furthermore, the conformational sampling of the deprotonated CYS systems (CYM^{CMD} and CYS^{CpHMD} at pH 12) in Figure B4 and Figure 3 are in surprising agreement although some energy terms differ.

The protonated simulations of the HIS amino acid (HIP^{CMD} and HIP^{CpHMD} at pH 1) show a large overlap of the energy distributions in Figure B11. However, the deprotonated forms (HIE^{CMD}, HID^{CMD} and HIP^{CpHMD} at pH 12) exhibit dissimilarities in several energy terms (i.e., total energy, potential energy, electrostatics and internal energies). The distribution of the CpHMD simulations does not reproduce the δ or ε state of the neutral HIS as observed in the plots. This fact was expected given the coexistence of the two protonation states in the CpHMD simulations. Instead, the electrostatic energy of HIP^{CpHMD} at pH 12 shows two peaks representing these states, but far from the energy range shown in the deprotonated CMD forms. To unravel this behaviour, the decomposition of the electrostatics is illustrated in Figure B12. The distributions of the protonated simulations (HIP^{CMD} and HIP^{CpHMD} at pH 1) follow the trend of the global electrostatics. On the contrary, the deprotonated simulations (HIE^{CMD}, HID^{CMD} and HIP^{CpHMD} at pH 12) show distinct distributions in all contributions. The backbone electrostatic energies show that the distributions of HIP^{CMD} and HIP^{CpHMD} at pH 1 and 12 overlap, while the HIE^{CMD} and HID^{CMD} systems have their singular distributions. The

side chain contributions are more coherent as the distribution of HIP^{CpHMD} at pH 12 is closer to the deprotonated simulations (HIE^{CMD} and HID^{CMD}) rather than to the protonated ones (HIP^{CMD} and HIP^{CpHMD} at pH 1). Focusing on the deprotonated CpHMD system, this behaviour in the backbone atoms is explained by the incorrect assignment of partial charges. The deviation of the electrostatic energy in the side chain atoms is due to the sum of two factors: (i) the partial charges of the side chain atoms vary with time due to the alternation between the δ and ε neutral states during the CpHMD simulation, which then modulates the conformational sampling, and (ii) the distributions of the electrostatic decompositions for the HIP^{CpHMD} at pH 12 are calculated using fixed partial charges of the HID or the HIE residues, ignoring the actual protonation state of the residues during the CpHMD simulation. Then, these distributions of HIP^{CpHMD} at pH 12 systems should be considered as rough approximations.

The ASP and GLU amino acids introduce the multiple protonatable sites into the CpHMD simulations. The energy distributions are illustrated in Figure 9 and Figure B13, respectively. For the ASP amino acid, in contrast to the previous amino acid sets, the energy distributions of the protonated systems (ASH^{CMD} and $AS4^{CpHMD}$ at pH 1) do not overlap due to the electrostatics (1–4EE, long-range EE, and, for the first time, Generalised Born contributions) as well as the angular and dihedral energies. Some deviations with respect to the CMD counterpart are expected because of the multiple protonation states over time. The deprotonated systems (ASP^{CMD} and $AS4^{CpHMD}$ at pH 12) show similar total and potential energies, but the same behaviour is observed in the electrostatic, angular, and dihedral contributions. In fact, the distribution shift is more pronounced for the electrostatic interactions. The angular and dihedral terms of the $AS4^{CpHMD}$ systems at acidic and basic pH conditions overlap strongly between them, except for their analogues (ASH^{CMD} and ASP^{CMD}). The electrostatic decomposition into backbone and side chain atoms in Figure 10 proves that the latter contribution causes the divergence in the electrostatics for the protonated simulations. This fact is probably related to the change in protonation states (and partial charges) during the simulation. The two peaks shown in the side chain electrostatics in the $AS4^{CpHMD}$ at pH 1 correspond to the *syn-O1* and *syn-O2* protonation states in their most stable conformation. For the deprotonated simulations, a mismatch in the distributions in both the side chain and backbone contributions is observed. This can be readily explained by the different partial

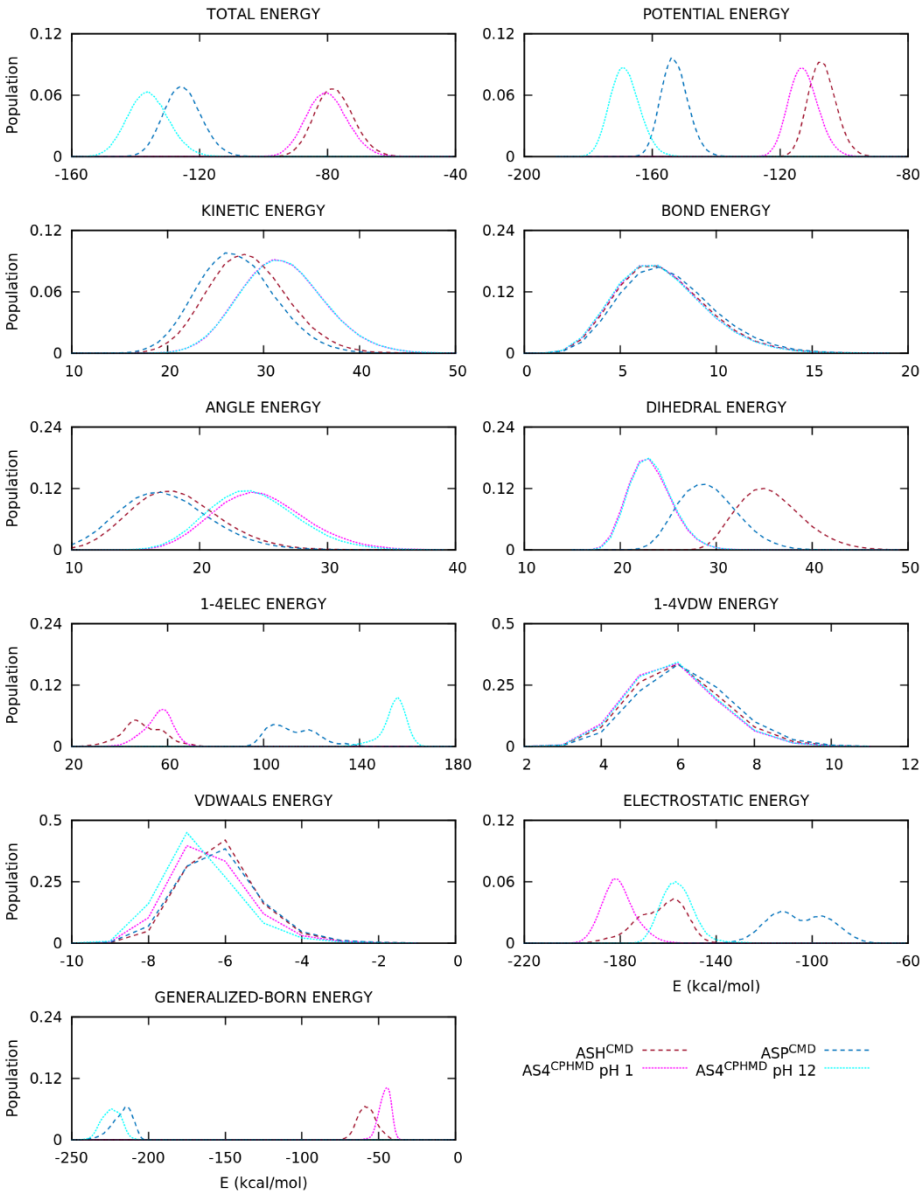


Figure 9. Energy distributions of the capped Asp_2 tripeptide. Global, inner, van der Waals, and electrostatic terms are illustrated. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

charges of the backbone atoms, while the shift in the side chain is probably caused by the partial charge of the C_β atom.

The energy distributions of the GLU systems show similar results to those observed for the ASP amino acid. The protonated simulations (GLH^{CMD} and $GL4^{CpHMD}$ at pH 1) show variations in the kinetic and potential energies, specifically in the angular, dihedral, and electrostatic terms (Figure B13). However, the deviations in the electrostatic energy are smaller than those in the ASP systems since the distributions agree in the

energy range, but the contours do not fit. On the contrary, the distributions of the deprotonated systems (GLU^{CMD} and $GL4^{CpHMD}$ at pH 12) have a larger shift for the total, potential, and 1–4 electrostatic terms, and a similar energy range for the long-range electrostatics. The decomposition of the electrostatic terms (Figure B14) shows that the backbone atoms reproduce the electrostatic interactions in the protonated systems (GLH^{CMD} and $GL4^{CpHMD}$ at pH 1). The electrostatic potential of the side chain atoms is inconsistent, which we assume is the result of the multiple protonation state. The distributions of the deprotonated systems (GLU^{CMD} and $GL4^{CpHMD}$ at pH 12) evidence deviations in the backbone and side chain contributions for both electrostatic terms. The contours in the side chain electrostatics suggest different protonation state sampling in the CpHMD simulations compared to the CMD counterparts.

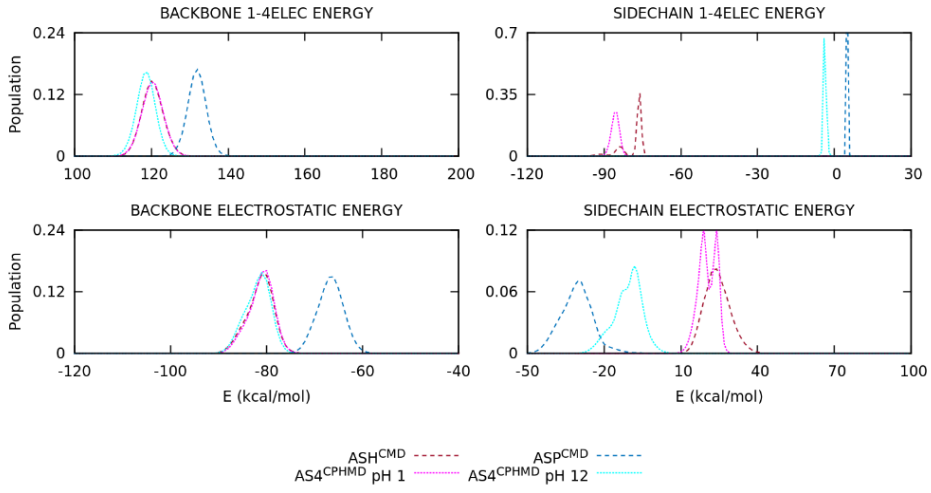


Figure 10. Energy distribution of the 1–4 and long-range electrostatics of the backbone and side chain atoms of the capped Asp_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

In the analysis of the energy contributions some energy-related deficiencies are identified. An accurate description of the electrostatics is crucial to ensure the reproducibility of the simulation and thus obtain a satisfactory conformational sampling. The energy decomposition helped to clarify several points. On the one hand, the backbone electrostatic energy shows that the protonated simulations are in agreement while the deprotonated ones do not match. As discussed in the previous section, the CpHMD method fixes the partial charges of the backbone atoms to the reference residue, i.e., the protonated state, regardless of the protonation state of the amino acid. This approach gives an inaccurate description of the electrostatic interactions when the residue is deprotonated

and therefore the deprotonated CpHMD simulations cannot reproduce the electrostatic distributions of the CMD counterpart. In fact, the backbone electrostatics of the deprotonated CpHMD simulations usually overlaps with the distributions of the protonated systems, although smooth shifts can be observed as a result of the different conformational sampling. In particular, the CYS systems might be controversial as they have a correct *global* electrostatic distribution, but when the backbone and side chain contributions are considered separately, the CMD and CpHMD counterparts clearly do not have similar distributions. On the other hand, the side chain electrostatic energies generally show deviations in the deprotonated simulations of all amino acids and the protonated simulations of HIS and acidic amino acids. These deviations are caused by two factors: (i) the modified partial charge of the C_β atom to ensure a net charge change of ±1.0, which affected the distributions of the deprotonated forms, and (ii) those amino acids with multiple protonatable sites in the CpHMD method that are not comparable to the CMD counterparts since the partial charges of the side chain atoms of the CpHMD residues vary during the simulation accordingly to the different protonation states. This is observed for the deprotonated form of HIS and the protonated form of the acidic amino acids, including in the Generalised Born electrostatics of the latter.

The energy distributions also suggest that the angular and dihedral energies are not properly described in these multiple protonatable amino acids. It seems plausible that the divergence in these two terms is not due to the partial charges and could instead be caused by (i) the activation and deactivation of the hydrogen during the protonation change and/or (ii) how the CpHMD-specific residues and these ghost hydrogen atoms are introduced into the residues.

4.2.3. Side Chain Orientation and Atom Distances

Finally, the dihedral angles φ and ψ and the characteristic dihedral, which is constructed by the backbone C_α atoms and a selected side chain atom of each amino acid, were used to define a new representation of the conformational space. This dihedral, called *angle θ*, is more suitable for providing insight on the orientation of the side chains with respect to the backbone chain. This *side chain-orientation* space is then divided into four sets: the φ_i/θ and θ/ψ_i, where i is the N-terminal (monomer 1) or C-terminal (monomer 2) amino acid. Figure 11 illustrates the θ dihedral angle and Table B3 gives the selected atoms for the θ angle for each amino acid. The map of the capped His_2

tripeptide is illustrated in Figure 12 and Figure 13, and the other maps are shown in Appendix B (Figures B15–B19). The distribution of interatomic distance between the selected atoms is shown in Figure B21.

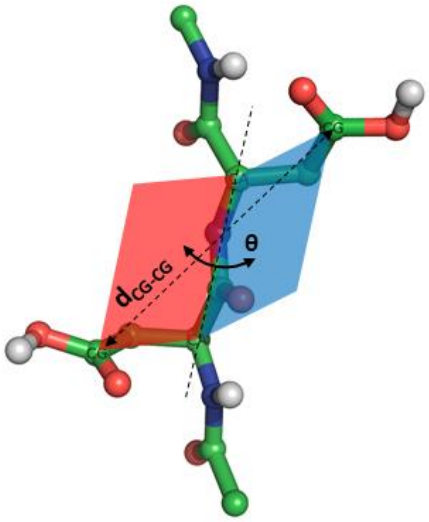


Figure 11. Dihedral angle (θ) constructed using the C_α atoms (CA) and the selected atoms in the side chain. In this case the carboxylic carbon atoms (CG) are selected. Table B3 gives the atom selection for each amino acid.

In general, the deprotonated and protonated simulations are consistent with the results of the Ramachandran maps. The protonated systems of all amino acids, except GLU and ASP, show a good agreement of the conformational sampling as well as the distances of the specific atoms. In contrast, the GLU and ASP systems exhibit mild deviations in both conformational sampling and atomic distances. For all the amino acids, the conformational sampling of the deprotonated forms diverges between the CMD and CpHMD counterparts, being of minor relevance for CYS and LYS and more significant for HIS, GLU and ASP.

This subsection is consistent with the reported results of the Ramachandran maps and energy distributions. However, the definition of this new angle and the construction of these maps (in the φ/θ and θ/ψ space) provide new information about HIP^{CpHMD} at pH 12. The atomic distances and the plots are more similar to the HID^{CMD} system rather than to HIE^{CMD}, which seems plausible since the side chain electrostatics of HIP^{CpHMD} at pH 12 are closer to the HID^{CMD}. Indeed, this conclusion is in line with the population of the δ state during the CpHMD simulation (77% and 81% for monomers 1 and 2, respectively)

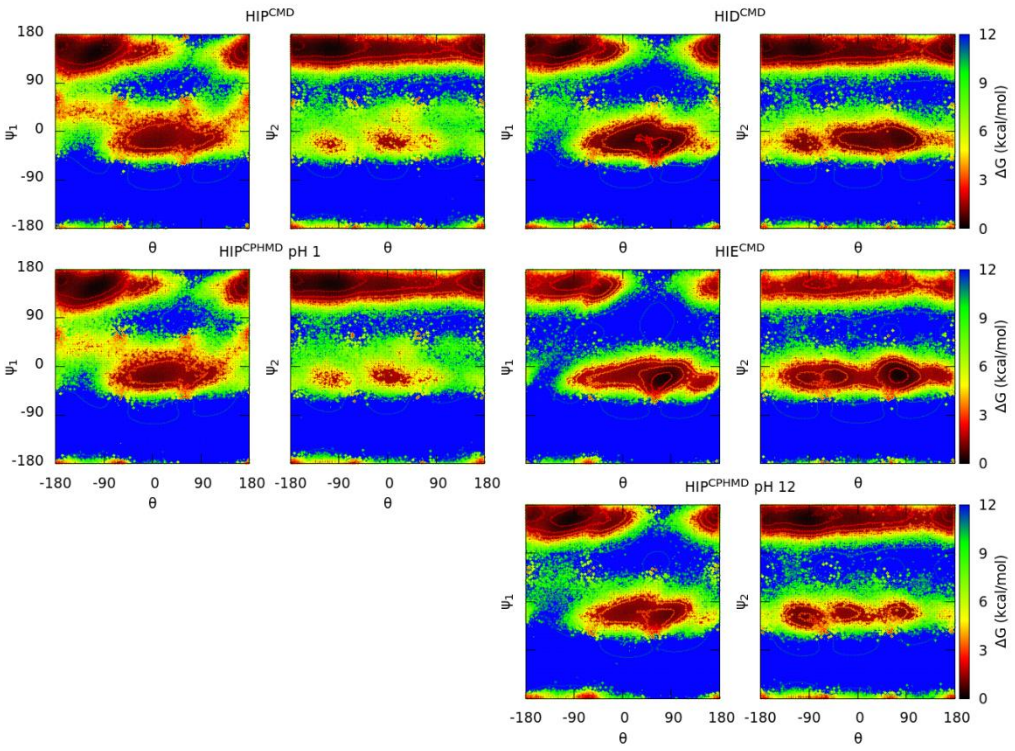


Figure 12. Gibbs free energies in the *side chain-orientation* space of the capped His_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). The dihedral angles ψ and θ are used in this plot for each monomer (ψ_1 from the N-terminal amino acid; ψ_2 from the C-terminal amino acid). The protonated forms are on the left and the deprotonated forms are on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

in contrast to the ε state (23% and 19%). On the other hand, residues GLU and ASP show different behaviour. In these systems, the dihedral plots show that the CMD and CpHMD counterparts (e.g., in the case of GLU, the GLH^{CMD} and $GL4^{CpHMD}$ at pH 1 systems for the protonated form, and the GLU^{CMD} and $GL4^{CPHMD}$ at pH 12 systems for the deprotonated one) have a similar conformational sampling, although closer atomic distances are shown when using the same simulation method. Even though the deviation in atomic distance is small, it may be due to a failure to correctly describe the angle and dihedral energies.

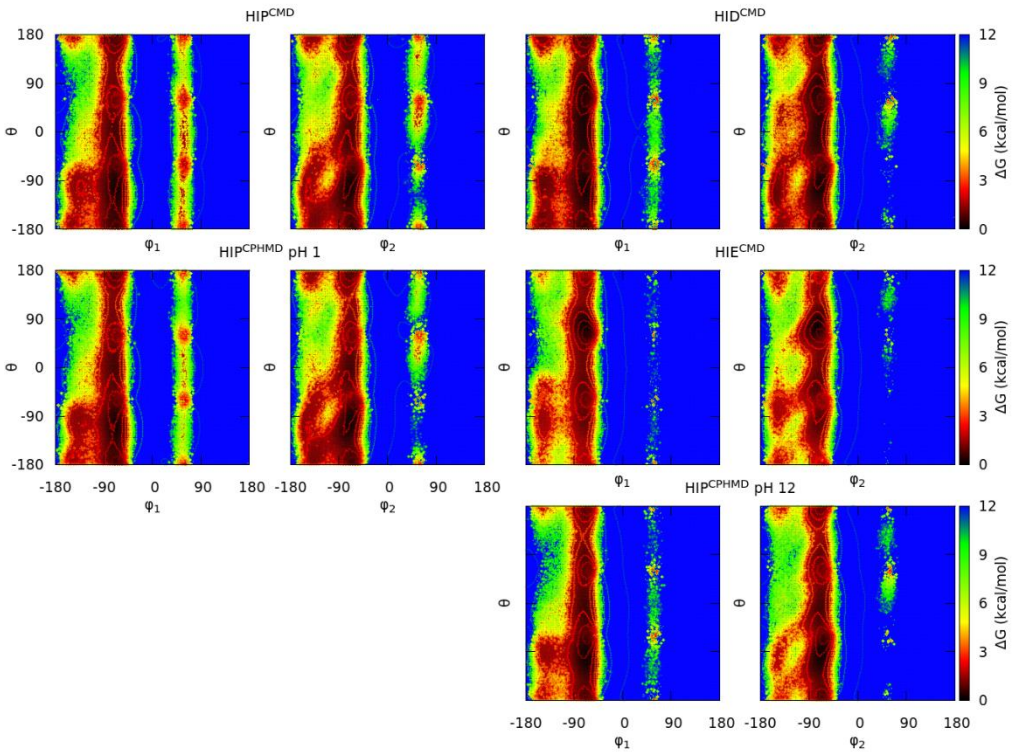


Figure 13. Gibbs free energies in the *side chain-orientation* space of the capped His_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). The dihedral angles φ and θ are used in this plot for each monomer (φ_1 from the N-terminal amino acid; φ_2 from the C-terminal amino acid). The protonated forms are on the left and the deprotonated forms are on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

4.3. Conclusions

Ramachandran maps and energy distributions have shown that the CpHMD method can reproduce the conformational sampling of the protonated forms of the tripeptides simulated with the CMD method. For the deprotonated forms, the different assignment of partial charges of the backbone atoms in the AMBER implementation leads to inaccuracies in the conformational profiles and energy distributions with respect to the CMD simulations. The electrostatic distributions show good agreement for the protonated forms, while the deprotonated ones exhibit significant deviations. The decomposition of the energy into backbone and side chain contributions reveals that the backbone electrostatics of the protonated form, that is, the reference state, in the protonated CMD simulations and both protonated and deprotonated CpHMD simulations have similar distributions. Instead, the deprotonated CMD systems have their own distribution

according to the assigned partial charges. The mismatch in energy between the deprotonated forms and the overlap of the energy distribution of the deprotonated CpHMD systems with the distributions of the protonated forms is due to the fixed partial charges of the backbone atoms in the CpHMD simulations. In addition, minor deviations in the side chain electrostatic energies are observed in the deprotonated forms due to the modified partial charge of the C_β atom. The acidic amino acids also do not overlap perfectly in the side chain electrostatics due to the multiple protonatable sites in the CpHMD simulations, thus showing an energy distribution with two peaks corresponding to the protonation states in the *syn* position of each oxygen atom. Furthermore, these *multi-site protonatable amino acids*, which also include the deprotonated HIS^{CpHMD}, show deviations in the angular and dihedral energies. Due to the different sampling of protonation states in the CMD and CpHMD methods, the Ramachandran maps and the energy distributions of these residues are not strictly comparable. Thus, the change in protonation states might be considered an advantage for sampling the conformational space rather than an inaccurate description of the amino acids.

The CpHMD method represents an improvement in the simulation of the biomolecules. The *dynamic* protonation states provided by the CpHMD methods allow the protonation state sampling according to the chemical environment (and therefore a greater conformational sampling) during the course of the simulations. For amino acids that have more than one protonation state in the protonated form, the fast mobility of the hydrogen atoms may provide a better description rather than CMD simulations. However, the Ramachandran maps reveal a shortcoming in the conformational sampling of the deprotonated CpHMD simulations due to the fixed partial charges of the backbone atoms. Therefore, we recommend using the CpHMD method in the AMBER implementation with caution, since the effects of incorporating inaccurate partial charges in the backbone atoms are unknown, and comparing structural protein descriptors (R_g, chemical shifts, FRET measurements…) with experimental data whenever possible.

4.4. Bibliography

1. Asokan, A. & Cho, M. J. Exploitation of Intracellular pH Gradients in the Cellular Delivery of Macromolecules. *J Pharm Sci* **91**, 903–913 (2002).

2. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK a Predictions. *J Chem Theory Comput* **7**, 525–537 (2011).

3. Anandakrishnan, R., Aguilar, B. & Onufriev, A. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* **40**, W537–W541 (2012).

4. Nio, A., Baptista, M., Martel, P. J. & Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins* **27**, 523–544 (1997).

5. Baptista, A. M. *et al.* Constant-pH molecular dynamics using stochastic titration. *J Chem Phys* **117**, 4184 (2002).

6. Bürgi, R., Kollman, P. A. & van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins* **47**, 469–480 (2002).

7. Lee, M. S., Salsbury, F. R. & Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738–752 (2004).

8. Mongan, J., Case, D. A. & McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* **25**, 2038–2048 (2004).

9. Meng, Y. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model. *J Chem Theory Comput* **6**, 1401–1412 (2010).

10. Itoh, S. G., Damjanović, A. & Brooks, B. R. pH replica-exchange method based on discrete protonation states. *Proteins* **79**, 3420–3436 (2011).

11. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).

12. Khandogin, J. & Brooks, C. L. Constant pH molecular dynamics with proton tautomerism. *Biophys J* **89**, 141–157 (2005).

13. Donnini, S., Tegeler, F., Groenhof, G. & Grubmüller, H. Constant pH Molecular Dynamics in Explicit Solvent with λ-Dynamics. *J Chem Theory Comput* **7**, 1962–1978 (2011).

14. Wallace, J. A. & Shen, J. K. Continuous constant pH molecular dynamics in explicit solvent with pH-based replica exchange. *J Chem Theory Comput* **7**, 2617–2629 (2011).

15. Goh, G. B., Hulbert, B. S., Zhou, H., Brooks III, C. L. & Brooks, C. L. Constant pH Molecular Dynamics of Proteins in Explicit Solvent with Proton Tautomerism. *Proteins* **82**, 1319–1331 (2014).

16. Długosz, M., Antosiewicz, J. M. & Robertson, A. D. Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain. *Phys Rev E* **69**, 021915 (2004).

17. Machuqueiro, M. & Baptista, A. M. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins* **72**, 289–298 (2008).

18. Swails, J. M. & Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J Chem Theory Comput* **8**, 4393–4404 (2012).

19. Russo, D. V. di, Estrin, N. v, Martí, D. A. & Roitberg, M. A. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput Biol* **8**, 1002761 (2012).

20. Mcdougal, O. M., Granum, D. M., Swartz, M., Rohleder, C. & Maupin, C. M. pKa Determination of Histidine Residues in α-Conotoxin MII Peptides by 1 H NMR and Constant pH Molecular Dynamics Simulation. *J Phys Chem B* **117**, 2653–2661 (2013).

21. Sarkar, A., Lal Gupta, P. & Roitberg, A. E. pH-Dependent Conformational Changes Due to Ionizable Residues in a Hydrophobic Protein Interior: The Study of L25K and L125K Variants of SNase. *J Phys Chem B* **123**, 5742–5754 (2019).

22. Khandogin, J., Chen, J. & Brooks III, C. L. Exploring atomistic details of pH-dependent peptide folding. *Proc Natl Acad Sci USA* **103**, 18546–18550 (2006).

23. Williams, S. L., Blachly, P. G. & Mccammon, J. A. Measuring the successes and deficiencies of constant pH molecular dynamics: A blind prediction study. *Proteins* **79**, 3381–3388 (2011).

24. Machuqueiro, M. & Baptista, A. M. Is the prediction of pKa values by constant-pH molecular dynamics being hindered by inherited problems? *Proteins* **79**, 3437–3447 (2011).

25. Wallace, J. A. *et al.* Toward accurate prediction of pKa values for internal protein residues: The importance of conformational relaxation and desolvation energy. *Proteins* **79**, 3364–3373 (2011).

26. Sun, Z., Wang, X. & Song, J. Extensive Assessment of Various Computational Methods for Aspartate's pKa Shift. *J Chem Inf Model* **57**, 1621–1639 (2017).

27. Buslaev, P. *et al.* Best practices in constant pH MD simulations: accuracy and sampling. *J Chem Theory Comput* **18**, 6134–6147 (2022).

28. Mongan, J. & Case, D. A. Biomolecular simulations at constant pH. *Curr Opin Struct Biol* **15**, 157–163 (2005).

29. Chen, W., Morrow, B. H., Shi, C. & Shen, J. K. Recent development and application of constant pH molecular dynamics. *Mol Simul* **40**, 830–838 (2014).

30. Martins De Oliveira, V., Liu, R. & Shen, J. Constant pH molecular dynamics simulations: Current status and recent applications. *Curr Opin Struct Biol* **77**, 102498 (2022).

31. Williams, S. L., de Oliveira, C. A. F. & Mccammon, J. A. Coupling Constant pH Molecular Dynamics with Accelerated Molecular Dynamics. *J Chem Theory Comput* **6**, 560–568 (2010).

32. Mao, A. H. *et al.* Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci USA* **107**, 8183–8188 (2010).

33. Chen, W., Wallace, J. A., Yue, Z. & Shen, J. K. Introducing Titratable Water to All-Atom Molecular Dynamics at Constant pH. *Biophys J* **105**, L15–L17 (2013).

34. Yeager, A. v, Swails, J. M. & Miller, B. R. Improved Accuracy for Constant pH-REMD Simulations through Modification of Carboxylate Effective Radii. *J Chem Theory Comput* **13**, 4624–4635 (2017).

35. Harris, R. C. & Shen, J. GPU-Accelerated Implementation of Continuous Constant pH Molecular Dynamics in Amber: pKa Predictions with Single-pH Simulations. *J Chem Inf Model* **59**, 4821–4832 (2019).

36. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).

37. van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).

38. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).

39. Wallace, J. A. & Shen, J. K. Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange. *J Chem Theory Comput* **7**, 2617–2629 (2011).

40. Wallace, J. A. & Shen, J. K. Charge-leveling and proper treatment of long-range electrostatics in all-atom molecular dynamics at constant pH. *J Chem Phys* **137**, 184105 (2012).

41. Huang, Y., Chen, W., Wallace, J. A. & Shen, J. All-Atom Continuous Constant pH Molecular Dynamics With Particle Mesh Ewald and Titratable Water. *J Chem Theory Comput* **12**, 5411–5421 (2016).

42. Huang, Y., Harris, R. C. & Shen, J. Generalized Born Based Continuous Constant pH Molecular Dynamics in Amber: Implementation, Benchmarking and Analysis. *J Chem Inf Model* **58**, 1372–1383 (2018).

43. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput* **11**, 3696–3713 (2015).

44. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383–394 (2004).

45. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicist*. (1999).

46. Perez, J. J., Santos Tomas, M. & Rubio-Martinez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J Chem Inf Model* **56**, 1950–1962 (2016).

47. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

48. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 2017).

49. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

50. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–99 (1963).

51. Rubio-Martinez, J., Tomas, M. S. & Perez, J. J. Effect of the solvent on the conformational behavior of the alanine dipeptide deduced from MD simulations. *J Mol Graph Model* **78**, 118–128 (2017).

52. Williams, T. *et al.* Gnuplot 4.6: an interactive plotting program. *Software Manual* 238 (2012).

53. Pietropaolo, A. *et al.* Unveiling the Role of Histidine and Tyrosine Residues on the Conformation of the Avian Prion Hexarepeat Domain. *J Phys Chem B* **112**, 5182–5188 (2008).

This chapter has been published in Privat, C., Madurga, S., Mas, F., Rubio-Martinez, J. Unravelling Constant pH Molecular Dynamics in Oligopeptides with Explicit Solvation Model. *Polymers* **13**, 3311 (2021)

Chapter 5

Unravelling Constant pH Molecular Dynamics in Oligopeptides with Explicit Solvation Model

The advantages of including the dynamic change of the protonation state depending on the chemical environment, and thus the charge-structure coupling, in the simulations are more than obvious for the study of proteins. To achieve this consideration in Molecular Dynamics (MD) simulations, this thesis has already introduced several times the Constant pH Molecular Dynamics (CpHMD) technique[1–5] developed over the last two decades, and more specifically we have focused on the approach with discrete protonation states[4,6–8]. By explicitly describing the protonation states, the discrete CpHMD method allows us to obtain a realistic atomistic representation of the conformational and protonation space of molecular models. The emergence of the simulations at constant pH and their recent popularisation through the implementation in software packages such as AMBER[9], CHARMM[10] or GROMACS[11] has not yet permitted an in-depth study of the potential of these approaches in the conformational space of proteins. While they have demonstrated a great ability to predict pKa or even reproduce mechanisms and conformational configurations of some proteins[12–21], our study on polyaspartic acid in Chapter 3 reveals some shortcomings in terms of conformational sampling. To explore these observations further, a detailed evaluation of the discrete CpHMD method with implicit solvation based on Generalised Born was performed. After carrying out CpHMD simulations of capped tripeptides and comparing them with conventional MD (CMD) simulations[22], some drawbacks regarding the conformational sampling of the titratable residues were observed. We concluded that the rough approximation made to the fixed partial charges of the backbone atoms in the titratable residues led to significant deviations in the conformational sampling of the deprotonated forms of the titratable amino acids. However, given the extensive literature demonstrating the potential of this method in certain applications and the advantages of protonation sampling over

conformations, we believe that these limitations can eventually be overcome in order to improve the accuracy of the constant pH simulations.

In this chapter, therefore, we proceed with the evaluation of the discrete CpHMD implemented in AMBER to shed light on the extent of the limitations mentioned above. Here, the hybrid solvation method[8], i.e., implicit solvation for the protonation state change attempt and explicit solvation for the conformational sampling, was used on the capped tripeptides in simulation boxes with TIP3P water molecules[23]. Unfortunately, the analysis of the conformational space and energy distributions indicates that the inconsistencies persist in the deprotonated state during the CpHMD simulations, regardless of the solvation method. To pursue our goal, we examined whether the position of the aspartic acids within a non-polar oligopeptide are also influenced by these shortcomings. We therefore performed extensive 8-microsecond simulations with two oligopeptides containing two aspartic acids in different positions: (i) separated and terminal and (ii) adjacent and central. Using both structural properties and energy maps, we prove that it is possible to minimise the deviations when the titratable amino acids are sufficiently distant, thus providing a better understanding of the limitations of the CpHMD method implemented in AMBER for large biomolecule studies.

5.1. Materials and Methods

In light of the objectives of this work, we prepared two sets of peptides to provide more insight into the performance of the CpHMD method. The first set of simulations includes the capped tripeptide to investigate the limitations reported in the previous chapter[22], but this time including explicit water molecules in the simulations. The second set includes oligopeptides with two aspartic acids placed at different positions in the sequence to study the effect of the distance between titratable amino acids when using the CpHMD method.

5.1.1. Capped Tripeptides

Six tripeptides consisting of two consecutive amino acids with acetyl (ACE) and N-methyl (NME) capping groups at the extremes of the sequence (ACE-X-X-NME) were constructed using the LEaP module of the AMBER suitcase. The titratable amino acids available in the CpHMD method of AMBER18 version[24] were Asp (D), Glu (E), His (H), Cys (C), Tyr (Y), and Lys (K). The residues and pH conditions used in the tripeptide simulations are listed in Table 1. Asp and Glu have specific titratable residues (AS4 or

GL4) due to the multiple positions of the proton in the protonated form, as shown in Figure 1. The other amino acids use the residue in the protonated form (HIP, CYS, TYR, and LYS) as the titratable residue in the CpHMD method. Since histidine has two protonation states in the neutral form, the delta (δ) and epsilon (ε) protonation states, the corresponding tripeptides in these states were prepared using the HID and HIE residues in the CMD simulations, respectively. Tyrosine was not simulated in the deprotonated form, as it is not parameterised in this protonation state in the CMD method. Finally, a simulation box with a minimum distance of 14.0 Å from any atom of the tripeptides was constructed and filled with TIP3P water molecules. If necessary, counterions were added until the net charge of the simulation box was neutralised. Any solvent molecule within 1.0 Å of the solute was removed to avoid overlapping between molecules.

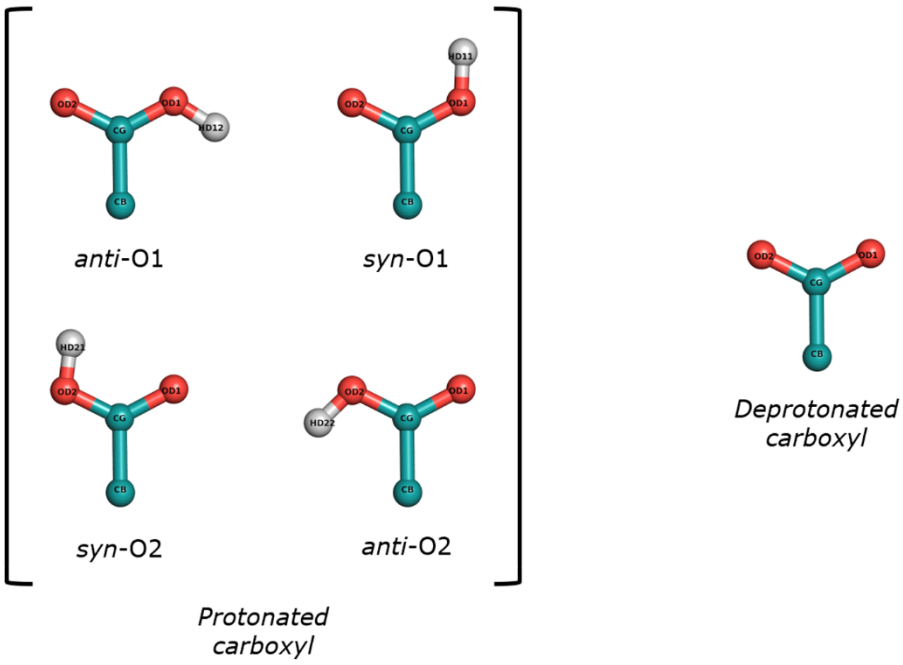


Figure 1. Protonation states of the carboxyl group in the side chains of the residues AS4 or GL4. There are four protonated states depending on the position (*syn* or *anti*) of the hydrogen atom with respect to the charged oxygen.

5.1.2. Oligopeptides

The second set of simulations was two oligopeptides consisting of a linear chain of eight alanine interrupted by two aspartic acids in different positions: (1) adjacent and central (ACE-A-A-A-A-D-D-A-A-A-A-NME or $A_4D_2A_4$) or (2) separated and terminal (ACE-D-A-A-A-A-A-A-A-D-NME or DA_8D). The ACE and NME capping groups

were added at the extremes of the peptides as we indicated in the sequences. These oligopeptide systems were defined as cubic boxes of 77.5 Å per axis and filled with TIP3P water molecules. If necessary, counterions were added until the net charge of the simulation box was neutralised. Any solvent molecule within 1.0 Å of the solute was discarded to avoid overlapping between solute and solvent molecules.

5.1.3. Preparation of the Input Peptide Structures

The ff14SB force field[25] and the constph.lib library (for CpHMD simulations only) were loaded into the LEaP module to parameterise the capped tripeptides and the oligopeptides. The *cpinutil.py* script then prepared the protonation states of the titratable residues using the Generalised Born model of Onufriev et al.[26] (igb = 2) for an ionic strength of 0.1 M. The residues AS4 and GL4 were defined in the *syn*-O2 protonated state at acidic pH conditions, and the residue HIP started as the neutral δ-state protonation state at basic pH conditions. The script also modified the intrinsic radii of the carboxylate oxygens in the topology file of those peptides containing residues AS4 and GL4[27].

5.1.4. All-Atom Conventional and Constant pH Molecular Dynamics Simulations

All the peptide systems were minimised using the steepest descent method[28] in three levels of restriction. Restrictions with a force constant of 5 kcal·mol^{-1}·Å^2 were applied in (1) all peptide atoms, (2) backbone atoms only and (3) no restrictions, during 5000 steps at each restriction level. In the CpHMD simulations, we did not turn on the protonation state change attempt during minimisation. Next, the systems were heated from 0 to 300 K with a linear increase of 1 K·ps^{-1} in the canonical ensemble (NVT) and then equilibrated for 200 ps in the isobaric-isothermal ensemble (NPT). Using the last coordinates after equilibration, four replicas with random initial velocities following a Maxwell–Boltzmann distribution were generated, and production runs of 500 ns (4 replicas × 500 ns = 2 μs per simulation) were performed in the canonical ensemble in order to increase the conformational sampling[29]. A Langevin thermostat[30] was set up with a collision frequency of 3 ps^{-1}. Periodic boundary conditions and the SHAKE algorithm were employed in the simulations. In the hybrid solvent CpHMD method of the AMBER implementation, the frequency of the protonation state change attempt was set to 0.2 ps^{-1} and water molecules were relaxed 0.2 ps after a successful attempt. Fully protonated or deprotonated states of the titratable amino acids were ensured by applying strong acidic (pH = 1) and basic (pH = 12) pH conditions in the CpHMD simulations. The titratable

LYS residue required an increase in the pH value at basic conditions (pH = 14.0). In the oligopeptide simulations, we set the solvent pH to 10.0 for a fully deprotonated state since aspartic acid has a low intrinsic pKa value. The simulations performed in this study are summarised in Table 1. All the MD calculations were carried out using the GPU version of the PMEMD software.

5.1.5. Energetic and Conformational Analysis

In all simulations the conformational configurations of the trajectory were collected every 10 ps. The energy contributions were later recalculated with a cut-off of 10.0 Å and using the trajectories after stripping the solvent molecules. The electrostatic energies were computed using Particle Mesh Ewald with a long-range correction for periodicity. In addition, the electrostatic potential was calculated dividing the capped tripeptides into the backbone atoms, including the capping groups, and the side chain atoms. In the CpHMD simulations, we also obtained the protonation fractions and the populations of each protonation state of the titratable amino acids using the *cphstats* program available in AMBER to confirm that the CpHMD simulations were performed in fully protonated or deprotonated states.

The radial distribution functions (RDFs) and the dihedral angles φ and ψ of the tripeptides were calculated using the CPPTRAJ module[31]. RDFs were computed using the distance of the water molecules around specific atoms of the side chains of each amino acid. An in-house script calculated the Ramachandran energy maps by transforming the dihedral data into Gibbs free energy as given in Eq. 1.

$$\Delta G = -k_b T ln(n_i/n_{max}) \qquad \textbf{Eq. 5.1}$$

where k_b is the Boltzmann constant, T is the temperature, and n_{max} and n_i are the maximum population and the population of a cell i in a grid of dihedral angles with a spacing of 1°. We classified the regions of the Ramachandran maps according to scheme of Rubio-Martinez et al. [32] as shown in Appendix C, Figure C1.

The conformational properties of the oligopeptides were analysed by the radius of gyration (R_g) and secondary structure fractions (fpSS) using the CPPTRAJ module. R_g was calculated using the C_α atoms of the peptides. fpSS was estimated by the DSSP method using all backbone atoms. All trajectories were superimposed onto the linear

CAPPED TRIPEPTIDES

Residue	CMD	CpHMD			PS	Intrinsic pKa
		pH 1	pH 12	pH 14		
ASP	✓				D^-	4.0
ASH	✓				P^n	
AS4		✓	✓		T	
GLU	✓				D^-	4.4
GLH	✓				P^n	
GL4		✓	✓		T	
HIE	✓				D^n	6.6
HID	✓				D^n	
HIP	✓	✓	✓		P^+/T	
CYM	✓				D-	8.5
CYS	✓	✓	✓		$P^{n/}$T	
TYR	✓	✓			P^n/T	9.6
LYN	✓				D^n	10.4
LYS	✓	✓		✓	$P^{+/}$T	
			DA_8D			
		pH 1	**pH 10**			
ASP	✓				D^-	4.0
ASH	✓				P^n	
AS4		✓	✓		T	
			$A_4D_2A_4$			
		pH 1	**pH 10**			
ASP	✓				D^-	4.0
ASH	✓				P^n	
AS4		✓	✓		T	

Table 1. Summary of the simulations indicating the peptide, the residue type, the simulation method, the protonation state (PS), and the intrinsic pKa of the amino acids. The PS labels indicate protonated (P), deprotonated (D) or titratable (T) residues, while the superscripts refer to the positive (+), neutral (n) or negative (-) charge of the side chains. The PS of the titratable residues depends on the solvent pH conditions (1, 12, or 14).

conformation before applying the Principal Component Analysis (PCA). PCA was applied by using the covariance of the C_α atom positions to build the transformation matrix. Subsequently, the conformational configurations were projected in a space defined by the first two Principal Components (PCs) to calculate the Gibbs free energy, as indicated in Eq. 1., with a grid spacing of 0.2°. Finally, the trajectories were clustered with the hierarchical agglomerative (bottom-up) approach using the root-mean-square displacement (RMSD) of the C_α atom positions as the distance metric. The conformational configurations were divided into 15 clusters with a sieve of 20 frames. The RMSD values between all the representative conformations of each cluster (2D-RMSD) were then calculated. All plots were generated with GNUPLOT[33].

5.2. Results and Discussion

5.2.1. Capped Tripeptides in Explicit Water Molecules

First, the capped tripeptides were simulated in explicit water molecules using CMD and CpHMD methods. The Ramachandran maps of the capped tripeptides were constructed by representing the backbone dihedral angles φ and ψ of each of the two monomers (the N- or C-terminal amino acid) of the tripeptide. We divided these maps into nine regions defined in Figure C1 according to the predominant conformation. The populations of each conformational region were calculated to provide a conformational profile of each peptide. In addition, the distributions of energy contributions were plotted, and the electrostatics was recalculated by removing the water molecules. In this study, we have focused on the latter contributions, which are fundamental in the change of the protonation states of the titratable amino acids. Finally, the effect of the electrostatic interactions on the solvent molecules was analysed by means of the RDFs of the water molecules around the tripeptides.

The capped Asp tripeptide is mainly discussed in this section to assess the strengths and weaknesses of the acidic amino acids in the CpHMD method when the explicit solvation model was introduced in the simulations. In the previous chapter, we carried out a similar study of the capped tripeptides, but using the implicit solvent model, and finally reported inconsistencies in the approach due to the assignment of the partial charges of the backbone atoms, among other possible artifacts. In this chapter we intend to clarify whether the reported CpHMD limitations persist with explicit solvent. We will also discuss the results observed for the other tripeptides reported in the Appendix C.

5.2.1.1. Conformational Sampling Inconsistencies in Deprotonated Forms of Amino Acids with Multiple Protonation States

The combinations of the dihedral angles ψ and φ of each amino acid were represented in the Ramachandran maps to obtain a profile of the secondary structure of each amino acid within the capped tripeptides. We then defined a grid on these maps to obtain the population fraction of each bin and thus calculate the Gibbs free energies. In addition, we measured the population ratios of the nine conformational regions as described in Materials and Methods.

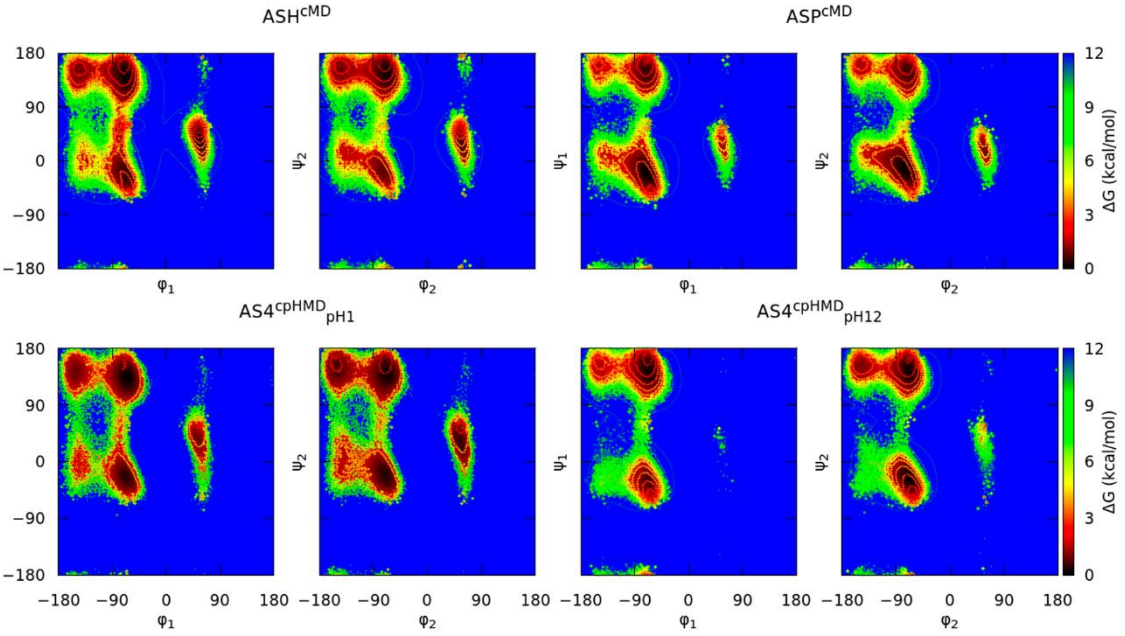


Figure 2. Ramachandran maps of the capped Asp_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation condition has two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

The Ramachandran maps of the capped Asp_2 tripeptide are illustrated in Figure 2. In the protonated form of the Asp_2 tripeptide, the conformational distributions of the CMD (ASH^{CMD}) and CpHMD ($AS4^{CpHMD}_{pH1}$) simulations do not fully satisfy the minima of the main populated regions (P_{II}, α_R, C_7^{eq}, and C_5), nor do the contours of the α_R region. We also observed this behaviour in the energy maps of the deprotonated form (ASP^{CMD} and $AS4^{CPHMD}_{pH12}$), in which again the minima and the contours of the α_R region do not agree between the methods. To quantitatively compare the simulation methods, we have

plotted the population of the main conformational regions in Figure 3. Here, the protonated simulations (ASH^{CMD} and $AS4^{CpHMD}{}_{pH\ 1}$) show small deviations (about ~10% maximum) in the populations of the regions, which can be accepted within a tolerance due to differences in the protonation state sampling between the methods. Thus, despite the deviation at the minima, the conformational populations of the protonated aspartic acid tripeptides are generally in agreement. However, when the Asp_2 tripeptide is deprotonated (ASP^{CMD} and $AS4^{CpHMD}{}_{pH12}$), the systems exhibit strong deviations in the conformational regions. A low population ratio of the α_L conformation confirms that this region is not sampled in the deprotonated CpHMD system.

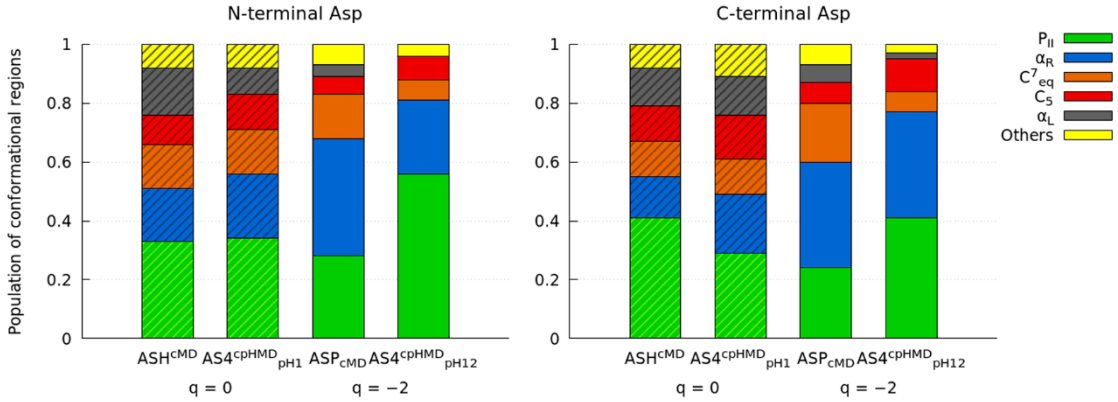


Figure 3. Populations of the conformational regions (P_{II}, α_R, $C^7{}_{eq}$, $C_{5,}$ and α_L) in the Ramachandran maps of each amino acid of the capped Asp_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q). The striped and solid box represent the protonated and deprotonated states, respectively.

In this chapter we also performed the simulations for each of the titratable amino acids available in AMBER. On the one hand, there are the hydrophilic amino acids Glu (acidic) and His (basic). The former is structurally similar to the Asp amino acid, but with an additional methyl group in the side chain and a slight shift in the intrinsic pKa. Indeed, the Ramachandran maps of the protonated (GLH^{CMD} and $GL4^{CpHMD}{}_{pH1}$) and deprotonated forms (GLU^{CMD} and $GL4^{CpHMD}{}_{pH12}$) show a similar behaviour as the Asp_2 tripeptide in Figure C2. The conformational populations confirm this behaviour in Figure C3, in which the populations of the conformational regions in the Ramachandran maps are clearly different in the deprotonated form. The case of histidine is more complex because the neutral form of the imidazole ring in the side chain can be defined as N-delta nitrogen (δ) or N-epsilon nitrogen (ε) depending on the position of the hydrogen. When found in the

protonated form, histidine is doubly protonated (and positively charged) in the imidazole ring. The Ramachandran maps of the capped His_2 tripeptide show a good agreement in the protonated form (HIP^{CMD} and $HIP^{CpHMD}{}_{pH\ 1}$) as observed in Figure C4. The population ratios of the regions show this tendency in Figure C5. However, the neutral δ and ε states of the His_2 tripeptide are remarkably different in CpHMD ($HIP^{CpHMD}{}_{pH\ 12}$) when compared to CMD (HID^{CMD} and HIE^{CMD}). $HIP^{CpHMD}{}_{pH12}$ shows a singular conformational distribution in the Ramachandran maps and population ratios, more similar to the protonated form rather than to the neutral HID^{CMD} or HIE^{CMD} tripeptides.

On the other hand, the hydrophilic basic Lys, the hydrophobic aromatic Tyr and the hydrophilic polar Cys amino acids constitute a set of titratable residues with intrinsic pKa values > 7.0 (i.e., 10.4, 9.6, and 8.5, respectively). The protonated form of lysine (LYS^{CMD} and $LYS^{CpHMD}{}_{pH1}$), tyrosine (TYR^{CMD} and $TYR^{CpHMD}{}_{pH1}$) or cysteine (CYS^{CMD} and $CYS^{CpHMD}{}_{pH1}$) in the capped tripeptides show closer conformational sampling in the Ramachandran maps (Figures C6–C8) and the populations of the conformational regions (Figures C9–C11) when the simulation methods are compared. The conformational sampling of the cysteine in the deprotonated form (CYM^{CMD} and $CYS^{CpHMD}{}_{pH12}$) is also consistent between the two methods. However, the deprotonated form of lysine (LYN^{CMD} and $LYS^{CpHMD}{}_{pH14}$) shows mild but not significant deviations in the conformational profile in the Ramachandran map and population ratios. We remind that tyrosine in the deprotonated form was not evaluated in this work due to the lack of parameterisation in the ff14SB force field, although the partial charges of the side chain atoms in the deprotonated state are available in the CpHMD libraries.

The conformational distributions of the capped tripeptides show that the conformational samplings of the deprotonated forms generally do not agree when the simulation methods are compared, except for those amino acids with pKa > 7, for which the deviations are small or acceptable within a tolerance. The protonated forms of the amino acids in the tripeptides agree in the conformational samplings, although those with multiple protonation states (Asp, Glu, and His) have mild shifts in the populations of the conformational regions. Furthermore, the inclusion of TIP3P water molecules generally leads to an increase in the P_{II} population, except in a few specific cases, but still shows the deviations between simulation methods that were reported in the previous chapter. In that report, we attributed the observed inconsistencies in the conformational sampling mainly to the crude approach of the partial charges in the backbone and C_β atoms, among

other minor reasons. Thus, despite the inclusion of explicit water molecules, the deviations of the deprotonated forms are not corrected when comparing the simulation methods.

5.2.1.2. Energy Contributions Reveal Deficiencies in Reproducing Electrostatic Interactions

Each energy term involved in the simulations was calculated using the CPPTRAJ module and then we compared the normalised distributions of each simulation method. For the electrostatic interactions we calculated the 1–4 and long-range interactions (i) with TIP3P water molecules and (ii) ignoring the solvent. Although this chapter only illustrates the electrostatic energies after stripping off the solvent molecules, both cases are considered in the discussion of the following section. This decision was made because the large proportion of solvent-solvent interactions caused a masking effect and thus hindered the consequences of fixing the backbone partial charges of the titratable amino acids. To examine the effect of the electrostatic interactions of the solute on the solvent, the RDFs of the water molecules around each amino acid were calculated.

As can be seen in Figure C12, the energy distributions of the Asp tripeptide show deviations in the electrostatic, dihedral and angular contributions in both the protonated (ASH^{CMD} and $AS4^{CpHMD}{}_{pH1}$) and deprotonated forms (ASP^{CMD} and $AS4^{CpHMD}{}_{pH12}$). The protonated form has electrostatic distributions in a close energy range but with distinct contours. In contrast, the deprotonated form exhibits distributions in distant energy ranges between the simulation methods. The CpHMD simulations share the same angular and dihedral distributions regardless of the solvent pH conditions and, in addition, are not consistent with the CMD analogues. To unravel the consequences of the incorrect partial charges in the CpHMD simulations, the electrostatic distributions of the backbone and side chain atoms were computed separately in Figure C13. In the protonated form (ASH^{CMD} and $AS4^{CpHMD}{}_{pH1}$), the backbone electrostatic energies agree in both the 1–4 and long-range terms. However, the electrostatic energy of the side chain atoms does not match the distributions. The deprotonated form (ASP^{CMD} and $AS4^{CpHMD}{}_{pH12}$) shows mild shifts but similar contours in both the backbone and side chain electrostatic energy distributions. Nevertheless, the backbone electrostatic distribution of the tripeptide at basic pH conditions ($AS4^{CpHMD}{}_{pH12}$) is closer to the protonated ones (ASH^{CMD} and $AS4^{CpHMD}{}_{pH1}$) rather than to the ASP^{CMD} system.

Therefore, the energetic contributions of the simulations with explicit solvent molecules behave in a similar fashion to those with an implicit solvation model. Apart from some deviations in the electrostatic energies, which can be expected since the explicit water molecules are a more accurate solvent model, the reported inconsistencies were also demonstrated in the previous chapter. There, we explained that the failure to reproduce the electrostatics in the deprotonated form are primarily due to the partial charges approach in the CpHMD method, where the partial charges of the backbone atoms of the AS4 residue are fixed at the values of the protonated (or reference) residue during the simulation. In addition, the partial charge of the C_β atom is also adjusted to ensure a change in net charge of ±1.0 when the protonation state of a titratable amino acid is changed. For this reason, the electrostatic interactions of the deprotonated form do not match in the backbone, which instead shows closeness to the distributions of the protonated form. Other factors are probably involved in these inconsistencies, such as the definition of dummy hydrogen atoms as ghost atoms or the different protonation state sampling. Note that the protonated form in the CpHMD method starts in the syn-O2 protonation state, but rapidly changes to other protonated states over time after accepting a protonation state change attempt. At the end of the simulations, we calculated the populations on each protonation state and found that $AS4^{CpHMD}{}_{pH1}$ is mainly populated a 47.2% and 45.8% in the syn-O1 and syn-O2 protonated states, respectively. The protonation state sampling in the CMD simulations is slower because the protonation state change is achieved by rotating the bonds and angles of the carboxyl groups. In our previous work, we suggested that this faster protonation state sampling is probably the reason for the deviations observed in the conformational profiles. It should then be investigated whether the protonation state sampling represents a consistent improvement in the conformational sampling of these peptides.

The radial distribution functions of the TIP3P water molecules around the capped tripeptides were calculated to understand the effect of the partial charges in the solute-solvent electrostatic interactions. The RDFs of the Asp_2 tripeptide in both protonated and deprotonated forms are in good agreement in Figure 4. The former (ASH^{CMD} and $AS4^{CpHMD}{}_{pH1}$) shows smooth changes in the contours of the distributions. On the other hand, the deprotonated form (ASP^{CMD} and $AS4^{CpHMD}{}_{pH12}$) shows only a slight shift in the distributions. In any case, these deviations are not significant and can be accepted within

a tolerance. Furthermore, the N-terminal or C-terminal position of the Asp amino acid in the tripeptide sequence does not affect the RDFs.

For the other titratable amino acids, the Glu tripeptide shows a similar behaviour to the Asp tripeptide in the energy distributions as observed in Figures C14 and C15. The electrostatic interactions and the dihedral and angular energies do not agree between the methods for both the protonated (GLH^{CMD} and $GL4^{CpHMD}{}_{pH1}$) and deprotonated (GLU^{CMD} and $GL4^{CpHMD}{}_{pH12}$) forms. The case of the His tripeptide is more challenging since the δ and ε neutral states were fixed in the CMD simulations, whereas the CpHMD method allowed the exchange between both protonation states. Therefore, HID^{CMD} and HIE^{CMD} are not strictly comparable with the $HIP^{CpHMD}{}_{pH12}$ simulation. All energy contributions of the protonated form (HIP^{CMD} and $HIP^{CpHMD}{}_{pH1}$) are in agreement in Figure C16. The deprotonated form (HID^{CMD}, $HIE^{CMD,}$ and $HIP^{CpHMD}{}_{pH12}$), on the other hand, does not agree when the simulation methods are compared, but they show total energy distributions in a close range. Only the electrostatic interactions exhibit notable shifts between the CpHMD and CMD simulations. Note that the energy distributions of $HIP^{CpHMD}{}_{pH12}$ were calculated by fixing the partial charges of the side chain atoms in one of the two protonation states, which is a very rough approximation. Thus, the energy distributions of HID^{CMD} and HIE^{CMD} could be considered as the energy boundaries within which the distribution of $HIP^{CpHMD}{}_{pH12}$ should fall. The electrostatic energies were also split into the backbone and side chain atoms in Figure C17. The backbone electrostatic contribution of $HIP^{CpHMD}{}_{pH12}$ overlaps with the protonated form as observed for other peptides. The side chain electrostatic distribution shows mild shifts with respect to the HID^{CMD} and HIE^{CMD} simulations, suggesting that the source of the deviation is mainly due to the failure to reproduce the backbone electrostatics.

Despite the deficiencies in the electrostatics, the RDFs of each protonation form of the capped Glu and His tripeptides show good overlapping in Figure 4. The Glu tripeptide has mild shifts in the protonated (GLH^{CMD} and $GL4^{CpHMD}{}_{pH1}$) and deprotonated (GLU^{CMD} and $GL4^{CpHMD}{}_{pH12}$) forms. For the His tripeptides, the RDFs of the simulations in the protonated form (HIP^{CMD} and $HIP^{CpHMD}{}_{pH1}$) overlap perfectly, and the deprotonated $HIP^{CpHMD}{}_{pH12}$ also overlaps with HIE^{CMD} and HID^{CMD}.

Finally, the energy contributions of the protonated form of Lys (LYS^{CMD} and $LYS^{CpHMD}{}_{pH1}$), Tyr (TYR^{CMD} and $TYR^{CpHMD}{}_{pH1}$) and Cys (CYS^{CMD} and $CYS^{CpHMD}{}_{pH1}$) tripeptides are in agreement, as can be observed in Figures C18–C20. However, the

deprotonated form of Lys (LYN^{CMD} and LYS^{CpHMD}_{pH14}) shows shifts in the distributions of the electrostatic energy and therefore in the total energy. The deprotonated Cys tripeptide (CYM^{CMD} and CYS^{CpHMD}_{pH12}) also fails in the overlapping of the electrostatic and dihedral energies. For the protonated forms of the tripeptides with pKa > 7.0, the division of the electrostatic energy into backbone and side chain atoms overlaps perfectly in all electrostatic contributions (Figures C21-C23). Nevertheless, the deprotonated forms show mild shifts in the side chain electrostatics, and the backbone electrostatic energy of the CpHMD simulations overlaps with the energy distributions of the protonated forms. The RDFs show good agreement in both protonated and deprotonated forms (Figure 4), suggesting that the approach in the partial charges of the backbone atoms does not have a significant effect on the distribution of water molecules around the tripeptides.

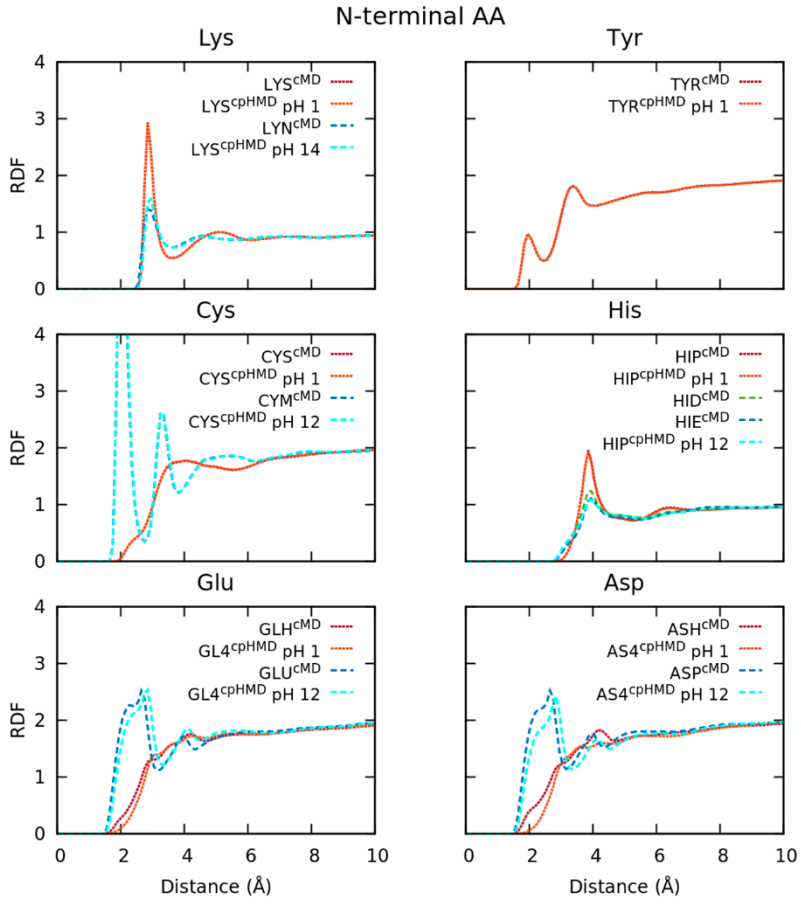


Figure 4. Radial distribution functions (RDFs) of the water molecules around each amino acid of the capped tripeptides. Only the N-terminal amino acid of each tripeptide structure is shown in this plot. The simulations in the protonated form are represented with dotted lines and the deprotonated ones with dashed lines.

Because of the explicit description of the solvent molecules in the simulations, smooth changes in the contours or the energy range of the distributions are observed when these distributions are compared with those distributions of the simulations with the implicit solvent. Nevertheless, all the simulations show similar behaviour regardless of the solvation model. It should be noted that the His amino acid could not be fully compared because the rough approximation to calculate the electrostatic energy and therefore the distributions in the CpHMD simulation at pH 12 should be averaged over the population of each state (δ or ε). In this case, the ε-state is more populated in the explicit solvation model (30% and 22% for N- and C-terminal amino acids, respectively) than in the simulations with the implicit solvent (23% and 19%). However, the δ-state still predominates at strong basic pH conditions (70% and 78%), which is consistent with the evidence observed in the Ramachandran maps and conformational populations.

5.2.2. Titratable Aspartic Acids in Adjacent and Terminal Positions in Oligopeptides

After evaluating the CpHMD method, we constructed two oligopeptides with eight Ala and two Asp amino acids in (1) separated and terminal (DA_8D peptide) and (2) adjacent and central ($A_4D_2A_4$ peptide) positions as test models. These simulations were designed to assess whether the failure of electrostatic interactions or other reported shortcomings of the CpHMD method persist in these oligopeptides and whether the distance between titratable amino acids can minimise the shortcomings.

Therefore, 8 μs length simulations were performed for each of these oligopeptides, $A_4D_2A_4$ and DA_8D, in the protonated and deprotonated forms of the Asp amino acid using the CMD and CpHMD methods. We then analysed the conformational sampling of these peptides by clustering the trajectories and building energy maps in the PCA space. Other properties related to the conformational sampling were calculated, such as R_g and secondary structure propensities. Finally, the distributions of the energy contributions were also calculated to finish the study of the extent of the implications of using the CpHMD method in these peptides.

5.2.2.1. The Position of the Titratable Amino Acids Modulates the Conformational Sampling

First, the conformations of the trajectories were used to construct the covariance matrix within the PCA approach in order to project the conformational sampling in the PC1 and PC2 space. The Gibbs free energies were then calculated by generating a grid in this new space and calculating the populations of each bin. From the eigenvalues of the

PCs, it was estimated that ~50% of the conformational sampling data was collected in these energy maps.

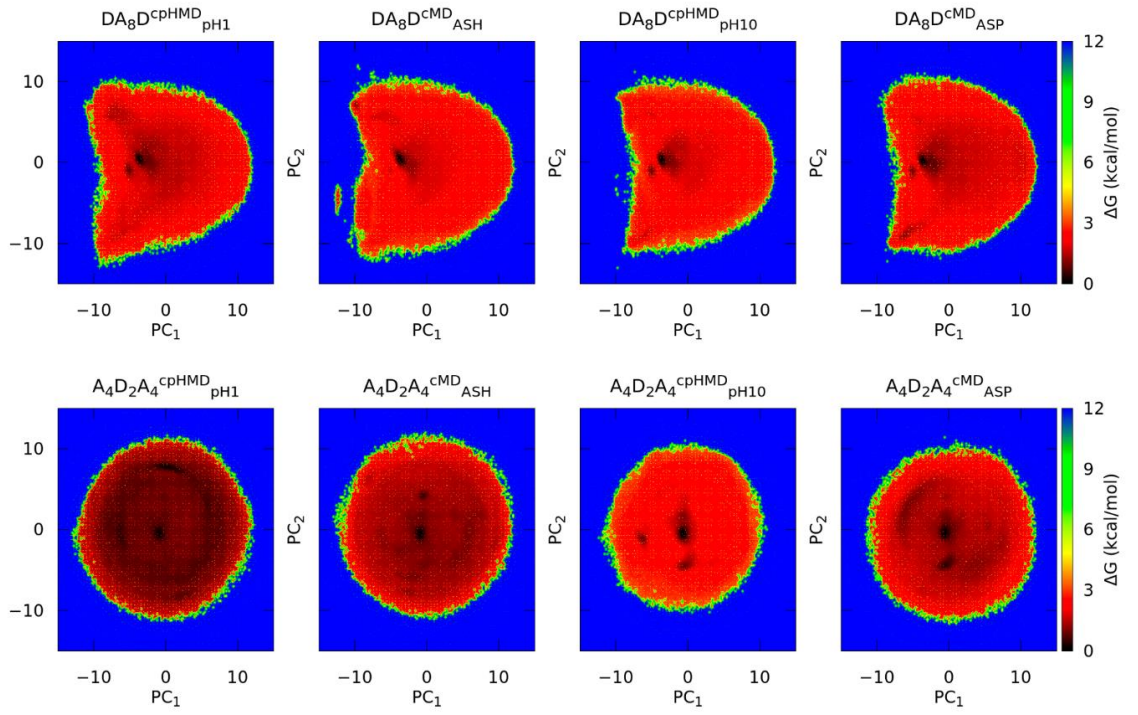


Figure 5. Gibbs free energies in the PC1 and PC2 space of the oligopeptides. The four plots at the top correspond to the DA_8D peptide and the four at the bottom correspond to the $A_4D_2A_4$ peptide. The subtitles indicate the peptide system, the simulation method (in superscript) and the residue label or the solvent pH (in subscript).

The energy maps in PCA space of the DA_8D and $A_4D_2A_4$ peptides are illustrated in Figure 5. The oligopeptide with terminal titratable amino acids (DA_8D) shows a similar conformational sampling regardless of the protonation state or even the simulation method. The location of the minima or populated regions is apparently more difficult to reproduce in the energy maps. Some subtleties are appreciated in the maps, e.g., the $DA_8D^{CpHMD}{}_{pH1}$ system is more distributed in the space since a wide dark area is observed or a new minimum appears in $DA_8D^{CMD}{}_{ASH}$. To quantitatively compare the conformational sampling of each system, the trajectories were clustered and the populations of the main clusters are shown in Figure 6. The clusters were ordered by population ratio, which does not necessarily mean that the cluster labels represent identical or close regions in the conformational sampling. Both protonated and deprotonated forms show good agreement in the populations of each cluster when the simulation methods are compared. In fact, the populations are similar between protonated

and deprotonated oligopeptides. The 2D-RMSD map of the representative conformation of the most populated clusters was calculated in Figures C24 and C25 to measure the structural similarity. In the protonated form, the superimposition of the representative conformations of the clusters C0 and C1 has a low RMSD value, indicating that $DA_8D^{CMD}{}_{ASH}$ and $DA_8D^{CpHMD}{}_{pH1}$ exhibit a close conformational sampling for, at least, ~40% of the trajectory. Other RMSD values show a good fitting between low populated clusters. However, the 2D-RMSD map of the deprotonated form indicates lower but still good fitting values. The two most populated clusters, C0 and C1, are present in both simulation methods but the order of their population changes. Other exchanges between low populated clusters are observed in the 2D-RMSD of the protonated and deprotonated forms, or even some representative conformations that apparently do not fit any other cluster.

To analyse the convergence of the simulations, the distribution of the first three PCs at 2, 4, and 8 μs were computed for the protonated and deprotonated forms of DA_8D in Figure C26. In general, the distributions do not change significantly over the reported times, but the observed peaks do, suggesting that more simulation time may be required for the stabilisation of the PC distributions. Given that the peaks at 4 and 8 μs showed small but still significant shifts in some systems, we extended the DA_8D simulations to 10 μs to ensure convergence. From 8 to 10 μs there were no notable variations in the distributions. We therefore concluded that simulation lengths of 8 μs were sufficient to sample the conformational space of the oligopeptides extensively.

On the other hand, the $A_4D_2A_4$ peptide shows remarkable differences in the energy maps in Figure 5. The protonated form ($A_4D_2A_4^{CMD}{}_{ASH}$ and $A_4D_2A_4^{CpHMD}{}_{pH1}$) is widely distributed in the conformational space. The conformational populations of the clusters confirm this observation as the ratios of the most populated clusters are very high. The 2D-RMSD values are not encouraging since the representative conformation of the most populated cluster, C0, of $A_4D_2A_4^{CpHMD}{}_{pH1}$ does not match any of the most populated clusters of $A_4D_2A_4^{CMD}{}_{ASH}$ or the cluster C1 of $A_4D_2A_4^{CpHMD}{}_{pH1}$. Since all clusters have closer populations and good RMSD values are observed in the 2D-RMSD plot and in other clusters, we assume that the conformational sampling is not very different between the methods. The simulations of the oligopeptide in the deprotonated form ($A_4D_2A_4^{CMD}{}_{ASP}$ and $A_4D_2A_4^{CpHMD}{}_{pH10}$) show a more restricted conformational sampling in Figure 5, especially for the $A_4D_2A_4^{CpHMD}{}_{pH10}$, which clearly exhibits three minima in

the map. Indeed, this system shows ~63% of the population contained within the three most populated clusters, which stands out with a difference of ~15% population when compared to the three most populated clusters of the peptide in the CMD method, $A_4D_2A_4{}^{CMD}{}_{ASP}$. However, the RMSD values indicate a good fit between the representative conformations of the clusters of both CMD and CpHMD methods in Figures C27 and C28, suggesting that there were small structural changes.

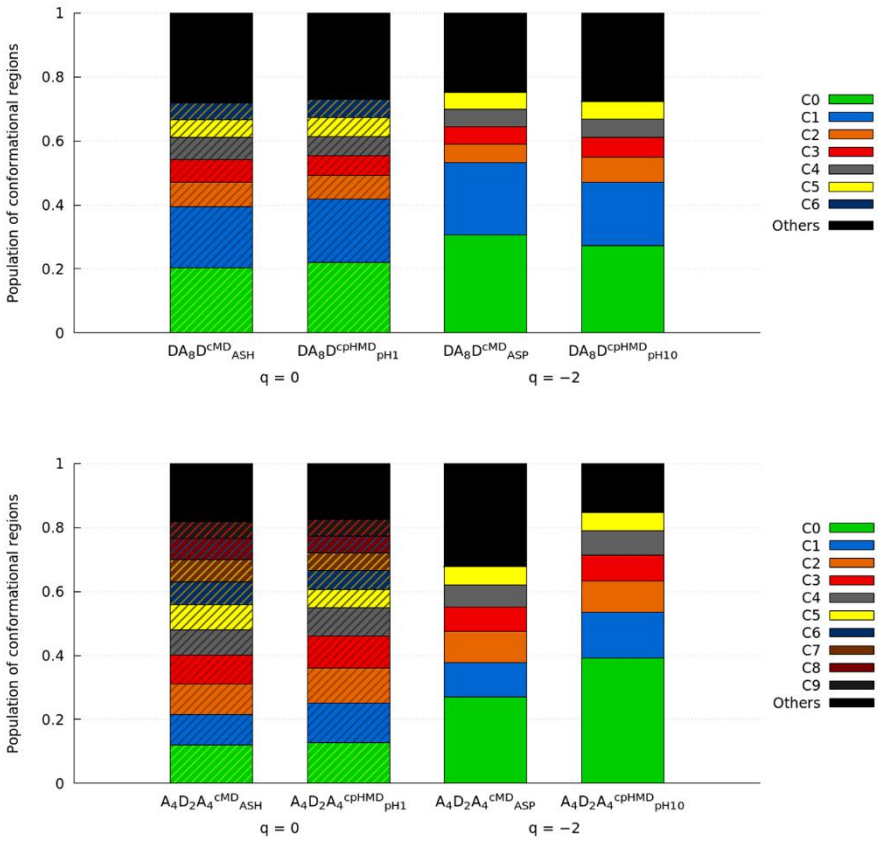


Figure 6. Representation of the clusters with a population ratio >5% (DA_8D and $A_4D_2A_4$ at the top and the bottom, respectively) in all simulation methods. The labels on the x-axis indicate the system, the simulation method (in superscript) and the solvent pH (in subscript, CpHMD simulations only). The total charge of the tripeptide is given below the systems (q = −2, 0). The box style (striped or solid) represents these systems in the same protonation state, regardless of the simulation method.

Thus, similar population ratios between clusters and tolerable agreement in the RMSD of representative conformations (but in exchange order) are found in the protonated form, while the deprotonated form shows better RMSD values between the simulation methods but more shifts in the population fractions. In order to explain the

behaviour observed in the energy maps and the clusters, we calculated some conformational properties to check if these tendencies are also present in these structural indicators.

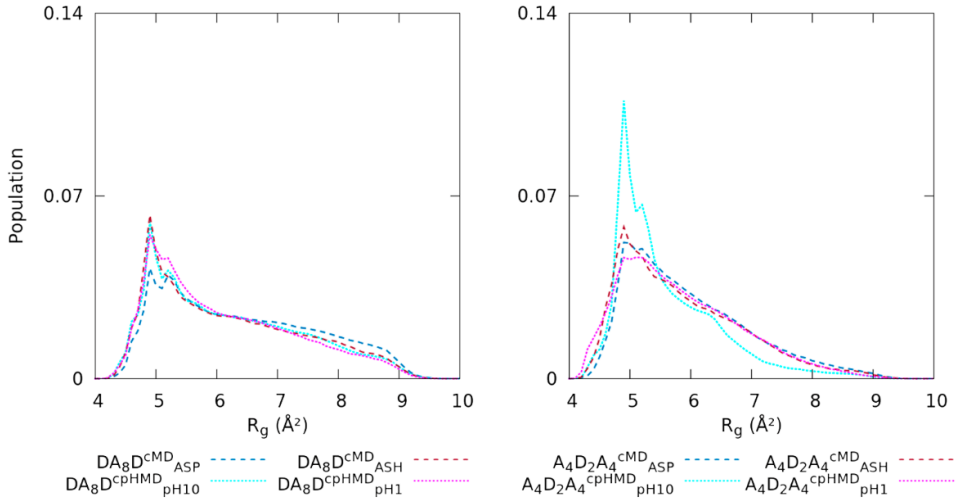


Figure 7. Comparison of the normalised distributions of R_g of the DA_8D (left) and $A_4D_2A_4$ (right) peptides. The simulation methods are represented in dashed (CMD) and dotted (CpHMD) lines. Cyan and blue colours indicate the deprotonated form and magenta and red colours indicate the protonated from.

5.2.2.2. Terminal Titratable Residues Accurately Describe Conformational Properties

The radius of gyration of the peptides was calculated in Figure 7 to measure the dispersion of atoms around the centre of mass as an indicator of structural compactness. The R_g distributions of the DA_8D peptide are in good agreement for both protonated and deprotonated forms, except for the two peaks located at ~5 Å. Nevertheless, the protonated form ($DA_8D^{CMD}{}_{ASH}$ and $DA_8D^{CpHMD}{}_{pH1}$) is fairly similar on the first peak, while the deprotonated form ($DA_8D^{CMD}{}_{ASP}$ and $DA_8D^{CpHMD}{}_{pH10}$) only overlaps on the second peak. Thus, the DA_8D peptide is consistent between the simulation methods, but smooth deviations in the peaks are observed. On the other hand, the $A_4D_2A_4$ peptide disagrees remarkably on the deprotonated form ($A_4D_2A_4{}^{CMD}{}_{ASP}$ and $A_4D_2A_4{}^{CpHMD}{}_{pH10}$). The tail of the R_g distribution of $A_4D_2A_4{}^{CpHMD}{}_{pH10}$ decays faster and the first peak is larger than in the CMD simulation, suggesting more compacted conformations compared to $A_4D_2A_4{}^{CMD}{}_{ASP}$. The protonated form ($A_4D_2A_4{}^{CMD}{}_{ASH}$ and $A_4D_2A_4{}^{CpHMD}{}_{pH1}$) has similar distributions but with a mild shift in the first peak. The deviation in the deprotonated form is consistent with the conformational sampling analysed in the energy maps and clusters.

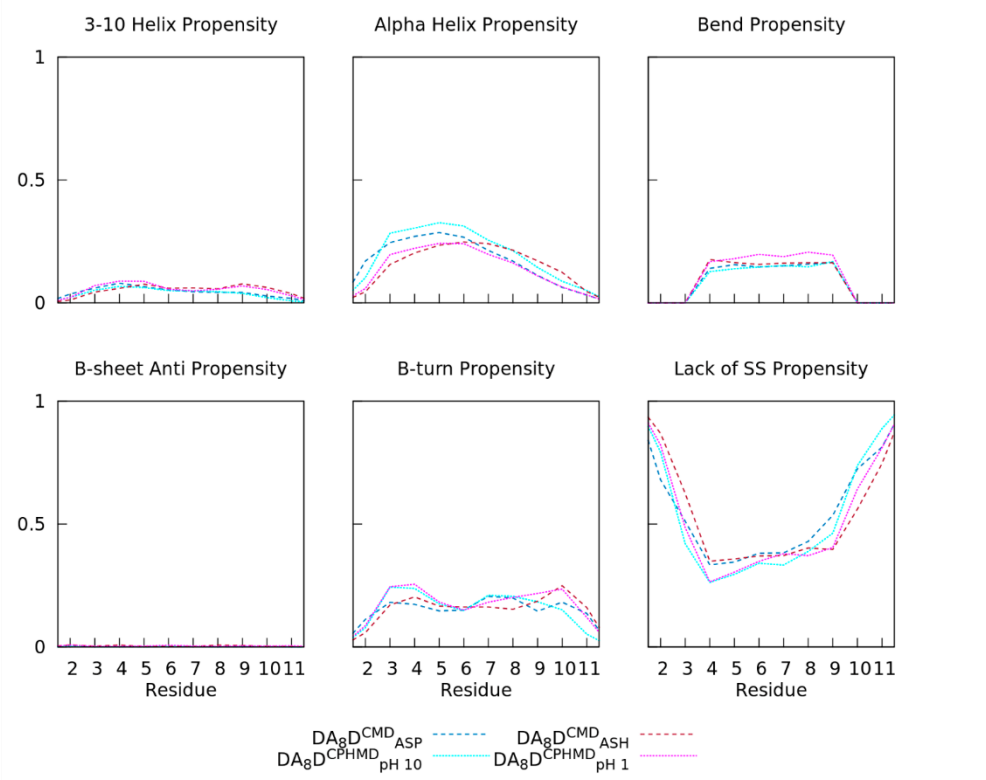


Figure 8. Secondary structure propensity fractions (fpSS) of each amino acid in the DA_8D oligopeptide using the DSSP algorithm. β-sheets and π-helices are omitted due to lack of content. The dashed and dotted lines indicate the CMD and CpHMD simulation methods.

The secondary structure propensity fractions (fpSS) of the peptides were calculated by using the DSSP method. In Figure 8, the DA_8D peptide shows good agreement between the two protonation forms within a tolerance of ~5%. The bend and 3_{10} helix structures overlap strongly when the simulation methods are compared, and the other SS propensity fractions (α-helix, bend, or turn) show mild but not significant deviations. The propensity to form a random coil is higher (~35%) than helices (~25%) or other secondary structures. On the other hand, the $A_4D_2A_4$ peptide is more diverse with respect to the fpSS in Figure 9. Neither the protonated nor the deprotonated forms overlap in the CMD and CpHMD simulations, even for those SS (turn and 3_{10} helix) with low fractions. In general, deviations of up to 20% are observed in the fpSS plots, except for the deprotonated form ($A_4D_2A_4^{CMD}{}_{ASP}$ and $A_4D_2A_4^{CpHMD}{}_{pH10}$) which stands out in the α-helix conformation. The high propensity for α-helix formation in the $A_4D_2A_4^{CpHMD}{}_{pH10}$ is consistent with the high compactness found in the R_g distribution. Thus, the conformational properties of the peptides with adjacent titratable Asp amino acids show greater deviations in the fpSS and R_g, apparently depending on the simulation method.

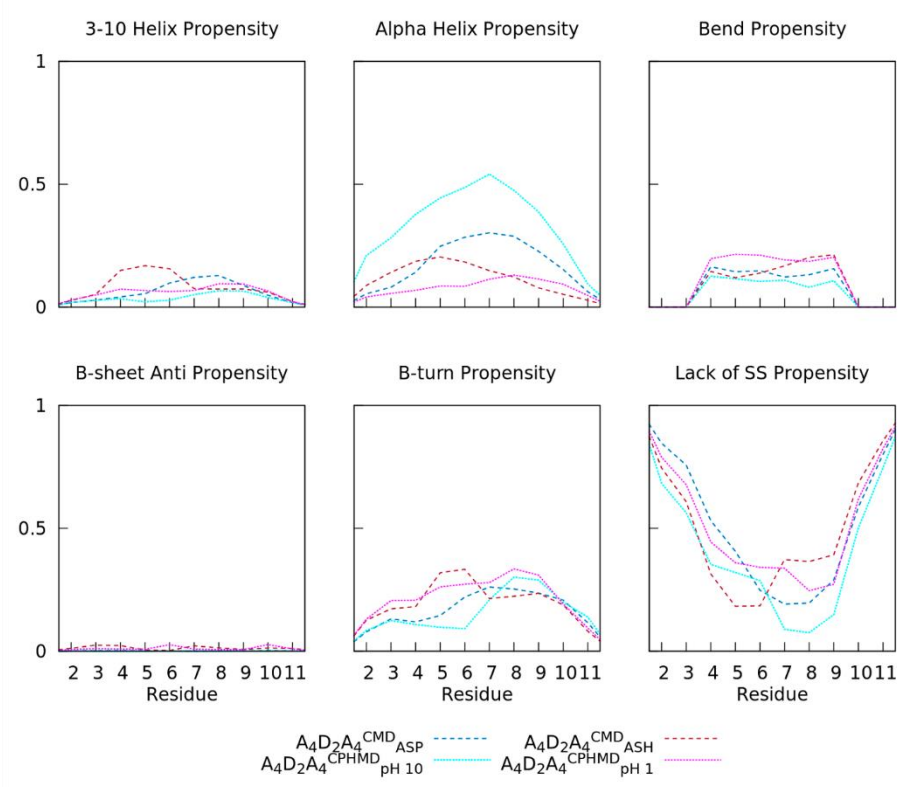


Figure 9. Secondary structure propensity fractions (fpSS) of each amino acid in the $A_4D_2A_4$ oligopeptide using the DSSP algorithm. β-sheets and π-helices are omitted due to lack of content. The dashed and dotted lines indicate the CMD and CpHMD simulation methods.

5.2.2.3. Electrostatic and Dihedral Energy Description Causes Deviations in Conformational Sampling and Structural Properties

To identify the source of the deviations in conformational sampling and structural properties, we now focus on the energy contributions of the simulations. We recalculated the intra- and intermolecular energies after removing the solvent molecules. The energy distributions of the peptides with the solvent molecules were also considered in the discussion of this section.

The total energy of the DA_8D peptide shows mild shifts in the distributions of the CMD and CpHMD simulations due to the mismatch of the angular, dihedral and electrostatic (1–4 and long-range) contributions in Figure 10 and Figure 11. On the one hand, the electrostatic energy distributions of the protonated ($DA_8D^{CMD}{}_{ASH}$ and $DA_8D^{CpHMD}{}_{pH1}$) and deprotonated ($DA_8D^{CMD}{}_{ASP}$ and $DA_8D^{CpHMD}{}_{pH10}$) forms are in a close energy range but have different contours in Figure 10. The CpHMD simulations have a

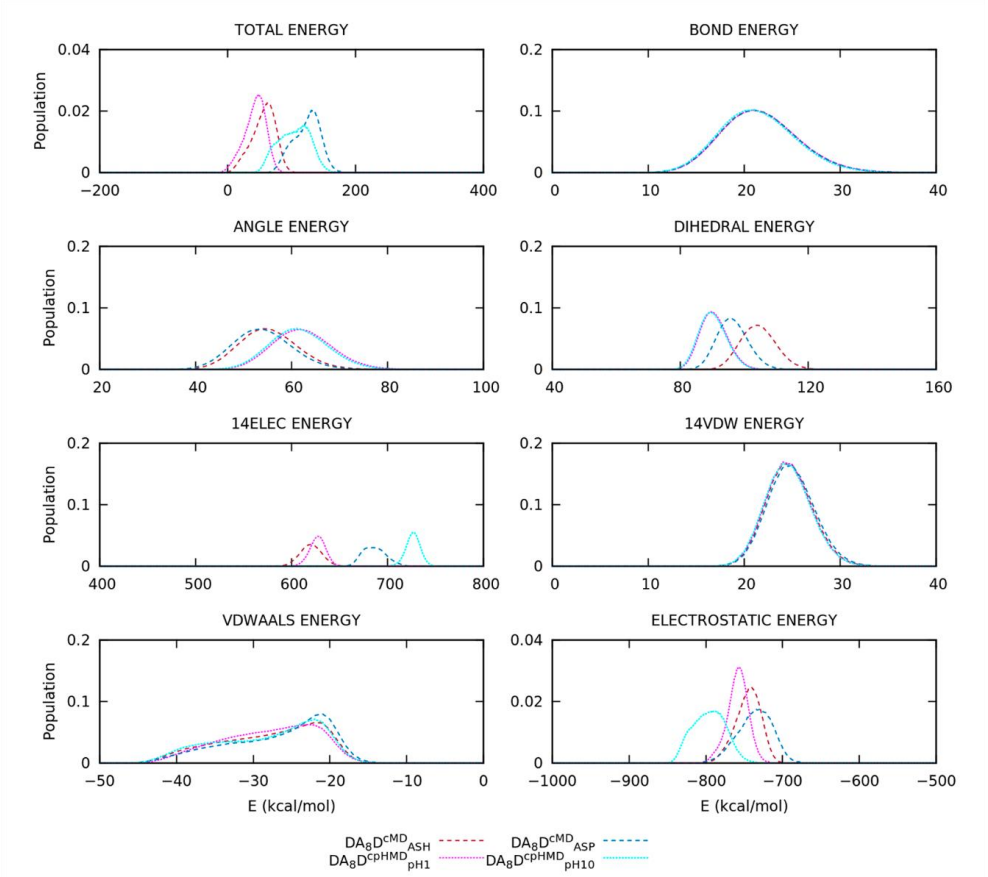


Figure 10. Normalised distributions of total, intra- and intermolecular energies of the DA_8D peptide. The CMD and CpHMD simulation methods are shown as dashed and dotted lines, respectively. The water molecules were not included in the calculations.

large peak on electrostatics, whereas the CMD simulations show broad distributions. On the other hand, the distributions of the angular and dihedral energies overlap in the CpHMD simulations, regardless of the solvent pH, which is an observation repeated in previous systems, i.e., the tripeptides in the explicit and implicit solvation model. Finally, the simulation methods show smooth deviations in the energy distributions of the DA_8D peptide, especially in the deprotonated form, consistent with the behaviour observed in the structural properties (R_g and fpSS). However, the deviations caused by the CpHMD residues are not sufficient to significantly modify the conformational sampling of the DA_8D peptide. In other words, when the CpHMD residues are sufficiently separated in the peptide chain, the reported shortcomings of the CpHMD are minimised.

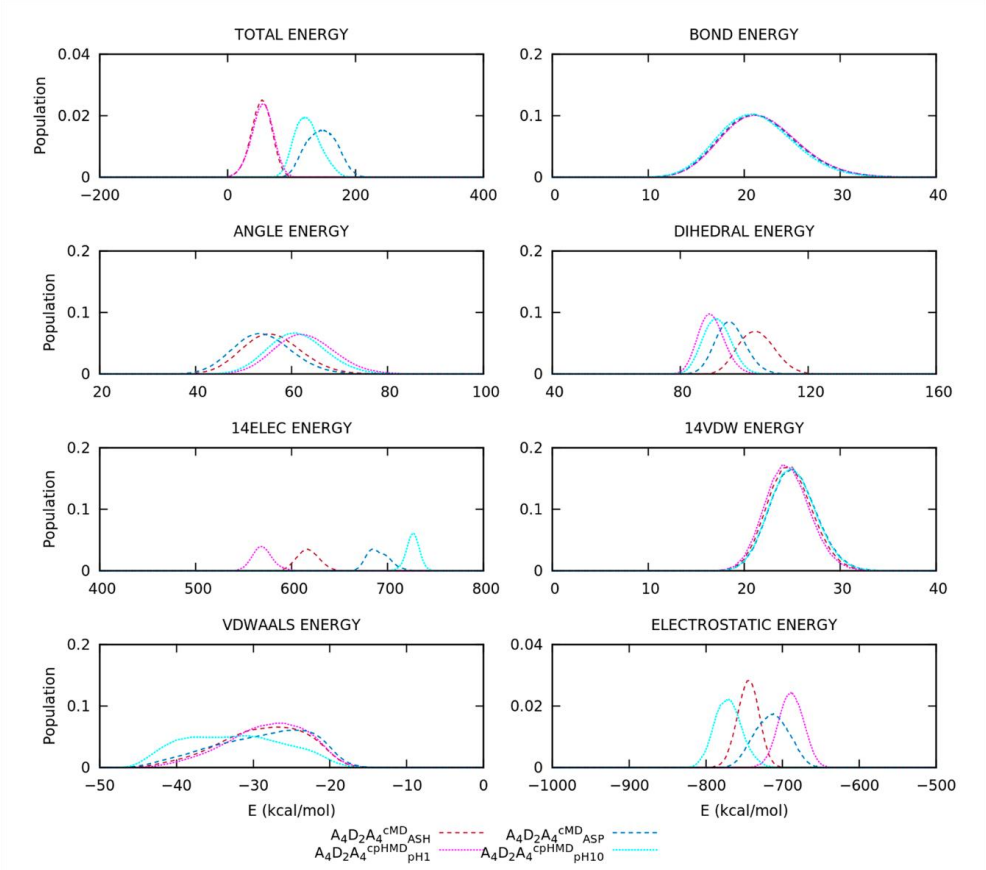


Figure 11. Normalised distributions of total, intra- and intermolecular energies of the $A_4D_2A_4$ peptide. The CMD and CpHMD simulation methods are shown as dashed or dotted lines, respectively. The water molecules were not included in the calculations.

The $A_4D_2A_4$ peptide, on the other hand, shows a good overlapping of the total energy distributions for the protonated form ($A_4D_2A_4^{CMD}{}_{ASH}$ and $A_4D_2A_4^{CpHMD}{}_{pH1}$) in Figure 11. In this case, the electrostatic interactions are poorly reproduced between the methods, with energy distributions in far ranges. The titratable amino acids of the $A_4D_2A_4$ peptide are closer, and therefore the interactions involving partial charges not correctly assigned in the backbone atoms play a more relevant role in the electrostatic interactions. This can be observed in the deviations of the electrostatic energy distributions of both the protonated and deprotonated ($A_4D_2A_4^{CMD}{}_{ASP}$ and $A_4D_2A_4^{CpHMD}{}_{pH10}$) simulations. Furthermore, the deprotonated form shows mild deviations on the total energy, including the angular, dihedral and even van der Waals energies. The dihedral and angular energies are not accurately reproduced and the distribution of the van der Waals interactions of $A_4D_2A_4^{CpHMD}{}_{pH10}$ is significantly different from the other simulations.

5.3. Conclusions

We have extended the previous work on the titratable residues of the CpHMD method in Chapter 3 and 4 by introducing explicit solvent molecules in the simulations of the capped tripeptides. The Ramachandran maps and energy distributions show similar behaviour to the results reported in the simulations with the implicit solvation model. Thus, the inconsistencies in the conformational and energy analyses of the deprotonated form of the CpHMD simulations are still related to the rough approximation adopted in the assignment of the partial charges of the backbone atoms, especially in those amino acids with multiple protonation states. In these *multisite protonatable* amino acids, minor shifts in the electrostatics and the conformational populations of the protonated form are observed. We assume that these small deviations are the result of the distinct protonation state sampling between the methods. In fact, the protonation state sampling of the CpHMD could be considered an advantage for the simulations with Asp, Glu, or His amino acids. In parallel, other minor artefacts could be involved in the above inconsistencies, such as the adjustment of the partial charge of the C_β atom or the ghost atoms during the simulation. It should be noted that the dielectric constant of water was underestimated by the TIP3P water model, and simulations with other explicit water models may yield different electrostatic profiles. However, since the discrepancy between the simulations is primarily due to the assignment of partial charges of the backbone atoms, it is expected that the shortcomings remain independent of the water model.

After examining the strengths and weaknesses of the CpHMD method when using explicit solvent in the simulations, we investigated the effect of the position of the titratable amino acids in a non-polar chain. On the one hand, the DA_8D oligopeptide shows no remarkable deviations in the energy maps, clustering and conformational properties of both protonated and deprotonated forms, suggesting that biomolecules with spatially separated titratable amino acids can reproduce the conformational sampling of the CMD simulations. On the other hand, the simulations of $A_4D_2A_4$ in the protonated form show good agreement in the conformational and energy analyses when comparing the CMD and CpHMD methods. In contrast, the deprotonated form exhibits important deviations in the measured properties (R_g, fpSS), the energy maps, and the clustering, indicating that the incorrect partial charges of the deprotonated state significantly affect the electrostatic interactions and thus modulate the conformational sampling. In order to make further progress in the identification of the CpHMD deficiencies, it would be

desirable to study other properties related to the electrostatic environment of the titratable amino acids, such as the presence of polar or charged amino acids, the addition of the ionic strength, etc. However, this third chapter concludes with the evaluation of the simulations at constant pH with discrete protonation states to pursue the objective of exploring the effects of the solvent pH on intrinsically disordered proteins, but leaves the door open to look for strategies to minimise the reported limitations in the deprotonated form of titratable amino acids in the CpHMD method. Hopefully, a more accurate description of the electrostatic interactions can be achieved in the simulations at constant pH in the near future.

5.4. Bibliography

1. Beroza, P., Fredkin, D. R., Okamura, M. Y. & Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. *Proc Nat Acad Sci USA* **88**, 5804–5808 (1991).
2. Mertz, J. E. & Pettitt, B. M. Molecular Dynamics at a Constant pH. *Int J High Perform Comput Appl* **8**, 47–53 (1994).
3. Baptista, A. M., Martel, P. J. & Petersen, S. B. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins* **27**, 523–544 (1997).
4. Baptista, A. M. *et al.* Constant-pH molecular dynamics using stochastic titration. *J Chem Phys* **117**, 4184 (2002).
5. Lee, M. S., Salsbury, F. R. & Brooks, C. L. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins* **56**, 738–752 (2004).
6. Bürgi, R., Kollman, P. A. & van Gunsteren, W. F. Simulating proteins at constant pH: An approach combining molecular dynamics and Monte Carlo simulation. *Proteins* **47**, 469–480 (2002).
7. Mongan, J., Case, D. A. & McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* **25**, 2038–2048 (2004).
8. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).
9. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).
10. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).
11. van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).
12. Mongan, J. & Case, D. A. Biomolecular simulations at constant pH. *Curr Opin Struct Biol* **15**, 157–163 (2005).

13. Williams, S. L., Blachly, P. G. & Mccammon, J. A. Measuring the successes and deficiencies of constant pH molecular dynamics: A blind prediction study. *Proteins* **79**, 3381–3388 (2011).

14. Machuqueiro, M. & Baptista, A. M. Is the prediction of pKa values by constant-pH molecular dynamics being hindered by inherited problems? *Proteins* **79**, 3437–3447 (2011).

15. Wallace, J. A. *et al.* Toward accurate prediction of pKa values for internal protein residues: The importance of conformational relaxation and desolvation energy. *Proteins* **79**, 3364–3373 (2011).

16. Swails, J. M. & Roitberg, A. E. Enhancing Conformation and Protonation State Sampling of Hen Egg White Lysozyme Using pH Replica Exchange Molecular Dynamics. *J Chem Theory Comput* **8**, 4393–4404 (2012).

17. di Russo, N. v., Estrin, D. A., Martí, M. A. & Roitberg, A. E. pH-Dependent Conformational Changes in Proteins and Their Effect on Experimental pKas: The Case of Nitrophorin 4. *PLoS Comput Biol* **8**, e1002761 (2012).

18. Ellis, C. R., Tsai, C.-C., Hou, X. & Shen, J. Constant pH Molecular Dynamics Reveals pH-Modulated Binding of Two Small-Molecule BACE1 Inhibitors. *J Phys Chem Lett* **7**, 944–949 (2016).

19. Sarkar, A., Lal Gupta, P. & Roitberg, A. E. pH-Dependent Conformational Changes Due to Ionizable Residues in a Hydrophobic Protein Interior: The Study of L25K and L125K Variants of SNase. *J Phys Chem B* **123**, 5742–5754 (2019).

20. Dobrev, P. *et al.* Probing the Accuracy of Explicit Solvent Constant pH Molecular Dynamics Simulations for Peptides. *J Chem Theory Comput* **16**, 2561–2569 (2020).

21. Buslaev, P. *et al.* Best Practices in Constant pH MD Simulations: Accuracy and Sampling. *J Chem Theory Comput* **18**, 6134–6147 (2022).

22. Privat, C., Madurga, S., Mas, F. & Rubio-Martínez, J. On the Use of the Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides. *Polymers* **13**, 99 (2021).

23. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926 (1998).

24. Case, D. A. *et al.* Amber 2018. *University of California, San Francisco* (2018).

25. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput* **11**, 3696–3713 (2015).

26. Onufriev, A., Bashford, D. & Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **55**, 383–394 (2004).

27. Yeager, A. v, Swails, J. M. & Miller, B. R. Improved Accuracy for Constant pH-REMD Simulations through Modification of Carboxylate Effective Radii. *J Chem Theory Comput* **13**, 4624–4635 (2017).

28. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicist*. (1999).

29. Perez, J. J., Santos Tomas, M. & Rubio-Martinez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J Chem Inf Model* **56**, 1950–1962 (2016).

30. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 2017).

31. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

32. Rubio-Martinez, J., Tomas, M. S. & Perez, J. J. Effect of the solvent on the conformational behavior of the alanine dipeptide deduced from MD simulations. *J Mol Graph Model* **78**, 118–128 (2017).

33. Williams, T. *et al.* Gnuplot 4.6: an interactive plotting program. *Software Manual* 238 (2012).

Chapter 6

Extensive Conformational Sampling of the Intrinsically Disordered Protein Histatin-5 Using All-Atom and Coarse-Grained Force Fields and Constant pH Molecular Dynamics Simulations

Intrinsically disordered regions (IDRs) or proteins (IDPs) challenge the structure-function paradigm established for more than 100 years until the end of the 20th century. The scientific community supported that the proteins exert their biological function through a well-defined three-dimensional structure, but IDPs broke the established structure-function scheme[1,2] by possessing a high degree of flexibility that allows them to adopt diverse conformational ensembles over time through disorder-to-order transitions. This dynamic conformational nature, in turn, makes them promiscuous and versatile proteins that can play their biological role with high specificity in binding processes, thus becoming key actors in various cellular processes (transcription and translation regulation, protein phosphorylation, self-assembly regulation, cellular signalling...) of eukaryotes[3]. Capturing the structures of IDPs as a first step in the study of the mechanisms of action of biological functions may then be promising for therapeutic applications, but unfortunately the characterisation of IDPs by experimental techniques is extremely difficult because they only determine the average observable of the several conformational ensembles that IDPs coexist over time. Fortunately, approaches such as Molecular Dynamics (MD) or Monte Carlo (MC) methods become very relevant in this context since they can model IDPs to generate conformational ensembles[4,5]. However, to map the entire conformational space of IDPs, a high computational effort is required, especially for MD.

Among the shortcomings of the MD method, an accurate parameterisation in the modelling of IDPs must be considered in order to accurately describe the conformational

space. In recent decades, novel force fields and water models have been proposed to overcome this challenge, since most of the available force fields were originally designed to mimic the properties of well-defined globular proteins. The force fields ff14IDPSFF[6], A99SB-disp[7], ff19SB[8], CHARMM36*[9], etc., have been successful in the simulating IDPs[10–12]. On the other hand, the role of the solvent in simulations of IDPs is increasingly recognised since protein-water interactions are essential for the ordering and disordering of proteins. The water models TIP4P-D[13], OPC[14] and the A99SB-disp[7] in combination with specific force fields have also shown potential among the many models available[8,15]. It is therefore expected that the use of IDP-specific force fields and water models will become more common in simulation studies of IDPs in the coming years.

In this chapter, in addition to evaluating the performance of some force fields and water models in IDPs, we also include the effect of solvent pH in the simulations. This property is essential in the simulation of pH-responsive proteins since the charge of the amino acids is often a determining factor in the structure of these proteins. IDPs generally have a high percentage of polar or ionisable amino acids in the sequence, so the pH of the environment plays a critical role in conformational sampling. Therefore, we employed the constant pH Molecular Dynamics (CpHMD) method[16] discussed in the previous chapters[17,18] to evaluate the effect of charge regulation over time on the conformational sampling of the simulations and vice versa. To this end, a reference IDP model, the human salivary peptide histatin-5[19], was selected as the target biomolecule for this study. This 24-amino acid disordered peptide is an antifungal agent found in the saliva. The sequence of histatin-5 is rich in histidine (~30%) and other ionisable amino acids such as aspartic acid, lysine, tyrosine or arginine. Many all-atom and coarse-grained (CG) MD or MC simulations of this peptide with various force fields and water models are collected in the literature, usually compared with SAXS, CD and NMR experiments[10,20–28]. In this case, we have evaluated resolution models (all-atom and CG), force fields (ff14SB[29], ff14IDPSFF, SIRAH[30] and Sugar), water models (TIP3P[31], TIP4P-D and WT4[32]), and simulation method (conventional or constant pH) by performing one of the most extensive conformational sampling reported in the literature (~8 million conformations and ~110 μs in total). The trajectory of each simulation was grouped into conformational clusters and analysed using SAXS intensity profiles or NMR chemical shifts. Among the simulations performed, the coarse-grained SIRAH/WT4 simulation with multiple seeds and the all-atom ff14IDPSFF/TIP4P-D simulation stand out in terms of reproducibility of

the experimental data. Furthermore, it is clear that the ability to modulate the charge during the simulations reduces the structure fraction in the all-atom simulations, while those simulations that favour extended conformations cause the deprotonation of the histidine amino acids. Thus, this chapter highlights the importance of using appropriate molecular models and methods to simulate IDPs to obtain an accurate conformational and protonation state sampling.

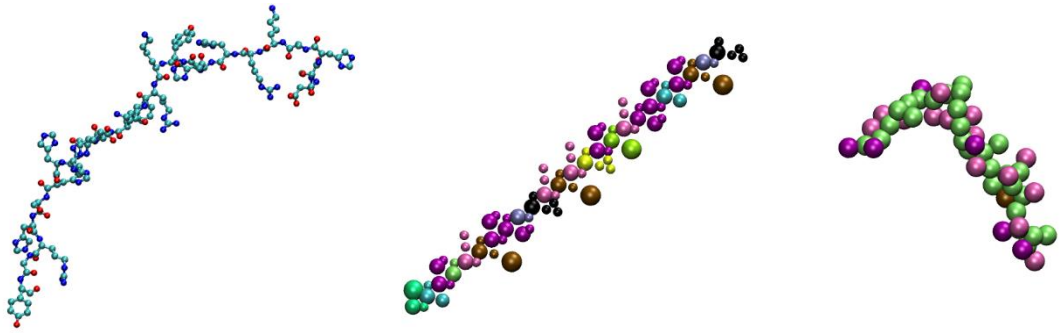


Figure 1. Histatin-5 models with different particle resolutions: all-atom (left), coarse-grained with SIRAH force field (middle) and coarse-grained with ESPResSO/Sugar library (right).

6.1. Materials and Methods

6.1.1. All-Atom Molecular Dynamics

The human salivary peptide histatin-5 was modelled as a linear chain consisting of the amino acids DSHAKRHHGYKRKFHEKHHSHRGY using the Leap module of AMBER18[33]. The peptide was parameterised with force fields ff14SB[29] or ff14IDPSFF[6]. For the conventional MD (CMD) simulations, the protonation states of the seven histidine amino acids at pH 7.0 were predicted using the PropKa tool[34], finally assigning the ε-state (HIE residue) to all of them. On the other hand, the CpHMD simulations required the assignment of the doubly protonated amino acid (HIP residue) to allow the protonation state change during the simulation. The CPHSTATS library was used to correctly define the protonation states. Each peptide was oriented according to its moments of inertia inside a box of dimensions 100x100x100 Å. The simulation box was filled with the TIP3P[31] or TIP4P-D[13] water molecules, and the net charge of the system was neutralised with Cl^- counterions. The energies and partial charges of the protonation states of the HIP residues were assigned with the CPINUTIL module.

The preparation of the peptide consisted of a three-step minimisation protocol to reduce internal stresses and relax the system. Using the steepest descent (SD) method[35], a first minimisation step was performed by restraining all peptide atoms to allow the solvent to adapt around the solute. In a second step, only the backbone atoms were restrained during the minimisation. In the final step, the simulation box was minimised with no restraints on the system. A maximum of 5000 SD steps were performed in each stage and a force constant of 5 kcal·mol^{-1} was applied to the restrained atoms. After relaxation, the simulation boxes were gradually heated with a linear increase of 1 K·ps^{-1} for 300 ps to a final temperature of 300K in the canonical ensemble (NVT). The volume of the simulation box was then adjusted by applying a pressure of 1.0 atm on the isobaric-isothermal ensemble (NPT) until a density close to 1.0 mol·A^{-3} was reached.

Simulation	Resolution	Method	FF/Water Model	Production runs
SBW3	All-atom	CMD	ff14SB/TIP3P	4 x 5 μs
SBW3pH	All-atom	CpHMD	ff14SB/TIP3P	4 x 5 μs
IDPW3pH	All-atom	CpHMD	ff14IDPSFF/TIP3P	4 x 5 μs
IDPW3pH^{R}	All-atom	CpHMD	ff14IDPSFF/TIP3P	32 x 100 ns
IDPW4DpH	All-atom	CpHMD	ff14IDPSFF/TIP4P-D	4 x 2 μs
SRH	CG	CMD	SIRAH	4 x 10 μs
SRH^{R}	CG	CMD	SIRAH	32 x 100 ns
SGR	CG	CMD	Sugar	4 x 5·10^{6} steps
SGRpH	CG	CpHMD	Sugar	4 x 5·10^{6} steps

Table 1. Details of the histatin-5 simulations. The resolution level, the simulation method, the force field and water model and the simulation time are given in the table.

Once the preparation of the simulation boxes was complete, 4 replicas of each simulation box were generated to allow further sampling of the conformational space as suggested by J. Rubio et al[36]. Each replica was assigned different initial velocities according to Maxwell-Boltzmann distribution and production runs of 5 μs length were performed. The temperature of the simulations was controlled using the Langevin thermostat[37] with a collision frequency of 3 ps^{-1}. The long-range electrostatic interactions were calculated with the Particle mesh Ewald method[38]. The hydrogen-involving bonds were constrained with the SHAKE algorithm[39]. The intermolecular interactions were considered within a cut-off of 10.0 Å. In the CpHMD simulations, a solvent pH of 7.0 was fixed for all the simulations[16]. The protonation state change attempts were performed every 200 fs, and after accepting a protonation state change, the water molecules were

relaxed for another 200 fs. During the protonation state change attempt, an ionic strength of 0.1 M was assigned for the electrostatic energy calculation.

To further sample the conformational space, 32 conformations were chosen from the ESPResSO CG simulations (SGR) after clustering as initial structures to perform 100 ns length simulations with the all-atom ff14IDPSFF force field for a *multi-seed* simulation (IDPW3pH^{R}). Full details of the simulations in terms of simulation method, the level of representation, the force field and water model and the initial configurations are given in Table 1.

6.1.2. Coarse-Grained Molecular Dynamics

The histatin-5 simulations with a CG representation were performed using the SIRAH[30] and ESPResSO[40] software packages. For the former, the linear chain constructed in the all-atom model was transformed into a CG model using the *cgconv* tool[41]. The CG beads were then parameterised with the SIRAH force field and WT4 water molecules[32] were added up to 20 Å to solvate the system. Cl^- counterions were added to the CG representation (ClW) to neutralise the net charge of the system. In this case, only a two-step minimization was performed. First, the simulation box was relaxed by applying restraints only to the GN and GO beads with a force constant of 2.4 kcal·mol^{-1} and then a second minimisation was carried out without restraints. For both minimisations 5000 SD steps were performed. Next, the simulation box was heated to 300K for 500 ps in the canonical ensemble, and then equilibrated to a pressure of 1.0 atm for 25 ns in the isobaric-isothermal ensemble. Finally, 4 replicas of the simulation box with different initial velocities were generated to perform 10-µs production runs in the isobaric-isothermal ensemble (NPT), with a total simulation time of 40 µs. The Langevin thermostat controlled the temperature of the simulations with a collision frequency of 50 ps^{-1}, and long-range electrostatic interactions were computed using the Particle Ewald mesh method. As the constant pH method is not available in SIRAH, the effect of pH was not included in the simulations using this software package. In addition, the sampling capacity of SIRAH was also tested with a multiple seed simulation of 32 production runs of 100 ns, using different initial structures obtained from the ESPResSO simulations (SGR), to examine whether the conformational sampling could be improved.

On the other hand, the Sugar library developed by Blanco P.M. (https://gitlab.com/blancoapa/sugar_library) was employed to prepare the simulation box in the ESPResSO software. A 2-bead linear model based on the histatin-5 sequence was built inside a 66.8x66.8x66.8 nm^3 box. The simulation box was then filled with ions until the charge of the system was neutralised and an ionic strength of 0.1M was achieved. The Lennard-Jones interactions were defined with the WCA potential and the electrostatic interactions with the P3M potential. After setting up the simulation box, it was freely minimised for 10.000 steps using the SD method. A temperature of 300 K was set in the Langevin dynamics, which was chosen to perform the simulation with the Velocity Verlet integrator and the Langevin thermostat. Four production runs of 5,000,000 steps were carried out and the configurations and properties of the histatin-5 peptide were extracted every 500 steps. In addition, a further 4 replicas were simulated using the constant pH method available in ESPResSO. A reference pKa of 6.8 for the histidine amino acids were defined according to Hass M. and Mulder F.A.A.[42]. CpHMD simulations were performed at neutral pH conditions (pH = 7).

6.1.3. Conformational Space and Structural Properties

The dimensionality of the conformational space sampled from each of the simulations was reduced by the Principal Component Analysis (PCA) method using the CPPTRAJ module[43]. The transformation matrix was obtained by diagonalising the covariance matrix of the C_α atoms of all the simulations. The simulations were then projected into the PCA space, and the Gibbs free energies of the populations contained in the PC1 and PC2, i.e., the PCs containing more structural information, were calculated with a grid of $\delta(PC) = 1.0$. On the other hand, the conformations of the trajectories were grouped into 15 clusters using the hierarchical agglomerative clustering method and the covariance of the position of the C_α atoms as a metric. The goodness of clustering was also calculated using the average distance from the centroid, the Davis-Bouldin Index (DBI), the pseudo-F statistics (psF) and the SSR/SST. From the representative conformations (or centroids) of the clusters with a population > 10%, the SAXS intensity profiles were estimated using the Fast X-ray Scattering (FoXS) server of Sali Lab[44]. The theoretical SAXS profiles were compared with experimental scattering data of histatin-5 at 1.26mg/l in 20mM Tris, 150 mM NaCl, pH 7.0 and at 298 K (SASDHH8)[45]. Separately, the chemical shifts (CS) of all conformations were calculated with the SPARTA+ software[46] integrated in the library MDTraj[47]. A linear regression was then performed

between the predicted (or simulated) HA atom CS and the experimental CS from Raj P.A. et al.[48]. The slope, intercept and r-value of the linear regression were extracted to assess the fit between the experimental and predicted CS of the HA atoms.

The radius of gyration (R_g) of the conformations was calculated excluding hydrogen atoms in the CPPTRAJ module. The R_g distributions were generated and normalised for comparison between the simulations. The secondary structure propensity fractions were estimated with the DSSP method[49]. The α-helix, 3_{10} helix and π-helix structures were grouped into the "helix" class. The parallel and anti-parallel β-sheet, β-bulge and isolated β-bridge structures were grouped into the "β-sheet" class. Protonation states were calculated by the *cphstats* module in the all-atom simulations and manually in the CG simulations. All plots were generated using GNUPLOT v4.6[50].

6.2. Results and Discussion

In this chapter we have focused on the implications of using various simulation setups, either at the level of structural resolution, force fields or water models, simulation method or sampling strategy, on one of the most widely used IDP models: the histatin-5 peptide. In terms of the molecular representation, we can distinguish between the simulations performed at all-atom resolution and those where the atoms or amino acids are simplified into beads in the CG model. The all-atom simulations were performed using two force fields, the popular ff14SB and the IDP-specific ff14IDPSFF, and two water models, the common 3-point water model TIP3P and the 4-point model with corrections for protein-water dispersion interactions TIP4P-D. In addition, given the relevance of the solvent pH to the definition of the protonation states and hence conformations, most simulations were carried out using the CpHMD method. To extend the sampling of ff14IDPSFF, we also performed a multi-seed simulation using 32 initial configurations extracted from the ESPResSO CG simulation. This CG simulation initially showed a good R_g distribution with respect to the experimental R_g determined by SAXS but was eventually surpassed by other simulations.

On the other hand, simulations with CG resolution are mainly divided according to the bead model and parametrisation. By means of the ESPResSO software and the Sugar library, histatin-5 was modelled based on a 2-bead model and simulated with the CMD and CpHMD methods. In contrast, the SIRAH software does not have the constant

pH method, and therefore a CMD simulation was launched from an *N*-bead model of histatin-5, in which each amino acid has a different number of beads according to the atomic structure. Visual inspection of the SIRAH simulation revealed that the flexibility of the peptide from the initial linear structure was low, resulting in structures with poor conformational diversity. To address this issue, and similar to the IDPW3pH^R simulation, we performed a second simulation using 32 initial configurations extracted from the clustering of the ESPResSO GC simulation to expand the conformational sampling.

6.2.1. TIP4P-D Water Model and Multi-Seed SIRAH Simulations Agree with SAXS and NMR Experimental Data

One of the most widely used experimental properties to study the model protein histatin-5 is the intensity profile obtained by SAXS. From this profile and various approximations, the R_g of the histatin-5 structure can be predicted. In this work, we used the experimental intensity profile of the histatin-5 at pH 7, 150 ml NaCl and temperature 298K, which conditions are reproduced in the simulations. From the conformational sampling of histatin-5, the global R_g of each simulation and the $R_g^{cluster}$ of the four most populated clusters of each simulation were calculated. The clustering of conformational states is necessary to capture the structure-related conformations of histatin-5 and their abundance during the simulation. The population percentage of the clusters and the average distance of all configurations with respect to the centroid of each cluster are given in Appendix D, Table D1. Although we will focus on to the clusters in the conformational sampling analysis in Section 3.2, the representative conformations of the four most populated clusters were used to predict the SAXS intensity profiles. The R_g and the intensity profile of the centroids were then estimated. The correlation between the experiment and theoretic intensity profiles was evaluated using the χ^2 fitting function, for which a good fit can be assumed if $\chi^2 < 3.0$. The theoretical SAXS intensity profiles with the SASDHH8 experimental scattering are illustrated in Figure 1. The theoretical R_g values, the χ^2 for all simulations and the experimental R_g from the SASDHH8 data are given in Table 2.

MODEL	R_g	$R_g^{cluster}$	SAXS		R_g^{exp}	NMR		
			R_g	χ		n	y_0	r
SBW3	9.0 ± 1.2	8.5 ± 0.4	8.5	9.2		1.2 ± 0.1	-0.4	0.95
		9.0 ± 0.7	8.4	10.1				
		8.6 ± 0.5	8.4	8.1				
		10.6 ± 0.8	11.0	**2.0**				
SBW3pH	9.7 ± 1.5	8.7 ± 0.6	8.3	11.0		1.3 ± 0.1	-0.8	0.95
		8.9 ± 0.5	8.7	8.8				
		9.9 ± 0.7	9.6	4.2				
		9.6 ± 0.7	9.9	5.5				
IDPW3pH	10.8 ± 2.0	10.6 ± 1.1	10.8	3.8		**1.0 ± 0.2**	0.2	0.83
		9.3 ± 0.5	9.1	7.9				
		10.4 ± 0.7	10.5	**2.6**				
		10.4 ± 1.0	9.8	5.0				
IDPW3pH^R	11.5 ± 2.3	10.3 ± 0.8	10.3	3.1		1.3 ± 0.1	-1.2	0.93
		9.4 ± 0.8	8.8	10.0				
		9.7 ± 0.8	9.3	5.7				
		14.0 ± 1.1	14.2	**1.5**				
IDPW4DpH	14.7 ± 2.7	17.8 ± 1.2	18.9	4.8	**13.7 ± 0.1**	**0.9 ± 0.2**	0.4	0.79
		15.7 ± 1.1	16.1	**2.6**				
		11.9 ± 1.0	11.8	**2.9**				
		10.5 ± 0.8	10.4	3.4				
SRH	11.1 ± 0.5	10.9 ± 0.3	10.8	3.0		0.6 ± 0.2	1.8	0.61
		10.6 ± 0.3	10.7	3.3				
		11.5 ± 0.3	11.5	**2.5**				
		10.8 ± 0.3	10.8	**2.6**				
SRH^R	12.1 ± 1.6	13.2 ± 0.7	13.5	**1.3**		**1.1 ± 0.2**	-0.6	0.75
		11.9 ± 0.7	12.1	**1.1**				
		13.3 ± 0.5	13.0	**1.1**				
		10.2 ± 0.3	10.1	3.8				
SGR	*13.2 ± 1.6*	14.9 ± 5.1	11.7	**2.2**		0.4 ± 0.1	2.9	0.70
		11.8 ± 4.6	8.1	11.8				
	14.2 ± 7.0	8.1 ± 3.8	7.1	12.2				
		20.4 ± 5.5	15.0	3.8				
SGRpH	*12.9 ± 1.6*	9.5 ± 4.4	8.1	9.6		0.4 ± 0.1	2.9	0.69
		15.7 ± 5.1	11.6	**2.6**				
	14.2 ± 7.0	13.7 ± 4.8	10.1	**2.4**				
		21.8 ± 5.2	16.5	**2.8**				

Table 2. Radius of gyration, SAXS and NMR properties of the simulations. R_g was calculated for the entire simulation and for each of the most populated clusters. R_g and χ^2 of the centroid conformation of the clusters in SAXS column were predicted with the FoXS server using the SASDHH8 data[45]. The slope, intercept and r-value of the linear regression of the simulated and experimental[48] HA CS are given in the table.

The reported R_g values range from 9.0 to 14.8 Å depending on the simulation setup, with some values quite distant from the experimentally determined R_g of 13.7 ±

0.07 Å. Among the simulations that fail to reproduce the experimental R_g are SBW3, SBW3pH, IDPW3pH, IDPW3pHR and SRH. Except for the latter, these systems are all-atom simulations whose force fields have already been shown to be unsuccessful in reproducing the experimental R_g in shorter time lengths[45]. Although such an extensive exploration of the conformational space has not yet been carried out, the results suggest that increasing the time length is not sufficient to capture extended conformations of the peptide. Adding the effect of the solvent pH using the constant pH method and thus enabling the charge regulation of histidine is also insufficient. Indeed, CpHMD simulations with the ff14SB and ff14IDPSFF force fields have not been reported in the literature. Despite not overcoming the force field related limitations to reproduce the experimental R_g, we do observe a shift towards larger R_g values when the protonation state sampling is included in the simulations. Actually, R_g increases from 9.0 ± 1.2 Å in SBW3 to 9.7 ± 1.5 Å in SBW3pH. In parallel, Sullivan et al.[25] reported histatin-5 simulations using the ff14IDPSFF/TIP3P with an R_g of 7.48 and 9.87 Å for the two most populated clusters after a 1 μs-length simulation, whereas in this work we observe average R_g values of 10.6 ± 1.1 Å and 9.3 ± 0.5 Å, and a global R_g of 10.8 ± 2.0 Å. Therefore, the constant pH method apparently leads to conformations with larger R_g closer to the experimentally determined one, thus confirming that the dynamic protonation states influence in the conformational sampling as expected due to the charge-structure coupling. The R_g distributions for these simulations can be observed in Figure 2, showing that SBW3, SBW3pH and IDPW3pH have a maximum population peak found in an R_g range far from the experiment.

For none of these simulations are the theoretical SAXS intensity profiles in agreement with the experimental scattering. In fact, the χ^2 values are greater than 3.0, indicating a poor fit. Only cluster C4 of SBW3 (6% of the population) and cluster C3 of IDPW3pH (12%) show a $\chi^2 < 3.0$, suggesting that extended conformations with intensity profiles similar to the SAXS experiment can be sampled but not representatively. Similarly, the IDPW3pH^R and SRH simulations show a few clusters with good χ^2 values, despite not reproducing the experimental R_g. The IDPW3pH^R simulation has an average R_g of 11.5 ± 2.3 Å and the distribution is much more populated at $R_g > 14.0$ Å compared to the previous simulations. Apparently, the multi-seed conformational sampling yields more extended conformations in agreement with the SAXS experiment, since the third cluster (11% population) has a χ-value < 2.0 and an R_g value very close to the

experimental one (14.0 ± 1.1 Å). Therefore, the ff14IPDSFF can capture extended conformations more frequently, but only by applying an *enhanced-sampling* strategy. On the other hand, the SRH simulation has an average R_g of 11.1 Å with a low standard deviation of 0.5 Å, indicating a poor structural diversity. Qualitative analysis of the trajectory revealed that histatin-5 in the SIRAH model is not very flexible when starting from a linear structure. Nevertheless, it is remarkable for its ability to obtain conformations with a theoretical intensity profile close to the experimental one given the χ^2-values < 3 observed in the clusters C3 and C4.

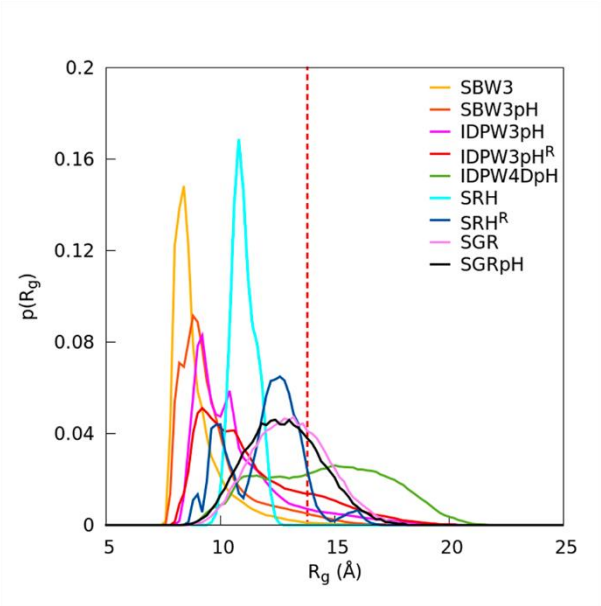


Figure 2. R_g distributions of the simulations performed with histatin-5. The dashed line represents the experimentally determined R_g by SAXS.

On the other hand, the IDPW4DpH, SRH^R and SGRpH simulations show a global R_g and theoretical SAXS intensity profiles in agreement with the experimental data. For atomistic simulations, the TIP4P-D water model, characterised by the correction of the protein-water dispersion forces, has already demonstrated its potential to reproduce experimentally the histatin-5 radius of gyration with the ff14IDPSFF[25] and A99SB-ILDN[10] force fields. In these works, the R_g has values of 13.5 and 13.2 Å, respectively, whereas here an R_g above the experimental value is observed with a large deviation, 14.8 ± 2.7 Å, is observed. In fact, the R_g distribution is very broad in a range of values between 8 and 21 Å, demonstrating the wide conformation spectrum that can be sampled by introducing the TIP4P-D water model. Due to the constant pH and extensive

conformational sampling in this work, it is unclear whether the water model together with the protonation state sampling results in an overly extended conformational sampling, whether these conformations are simply the product of the extension of the conformational sampling from the previous work, or a combination of both factors. These overly extended conformations are also reflected in the SAXS intensity profiles, where the main cluster (25%) with an R_g of 18.1 Å has a poor χ^2 value, while the other two clusters (21% and 11%) have R_g values not so far from the experiment and χ^2 values < 3.0. The simulation that really stands out in reproducing the SAXS intensity profiles is the SRH^R simulation. The three most populated clusters (~40% of the trajectory) show χ^2-values very close to 1.0, demonstrating an excellent fit. Furthermore, both the C1 and C3 clusters have an average R_g and standard deviation within the experiment range. Conversely, the global R_g of the SRH^R simulation does not agree with the experiment, although the distribution is close to the experimentally determined R_g.

Finally, we examine the SGR and SGRpH simulations. Both simulations show an average R_g of 13.2 and 12.9 Å with a standard deviation of 1.6 Å when analysing the 2-bead model. The results are extremely good with respect to the experimental R_g and, in fact, the distributions for both simulations are the most uniform and close to the experiment. However, the conformations adopted by the 2-bead model are not reproducible at all-atom resolution because many peptide bonds adopt a cis-orientation. In the pursuit of reconstructing realistic conformations, we imposed trans peptide bonds in the back-mapping. We then calculated the average R_g of these conformations and obtained 14.2 ± 7.0 Å for both simulations. The reported R_g from back-mapping indicates that all-atom conformations are indeed different from the 2-bead model. Despite this rough approximation, some clusters can produce theoretical intensity profiles with good agreement with the experiment. SGR has a χ^2-value of 2.2 for its first cluster (21%), while SGRpH has three clusters (~40%) with χ^2-values < 3.0. Clearly, the conformational ensembles generated by ESPResSO simulations would be greatly improved with a tool capable of successfully back-mapping the 2-bead model to the all-atom resolution integrated in the Sugar library.

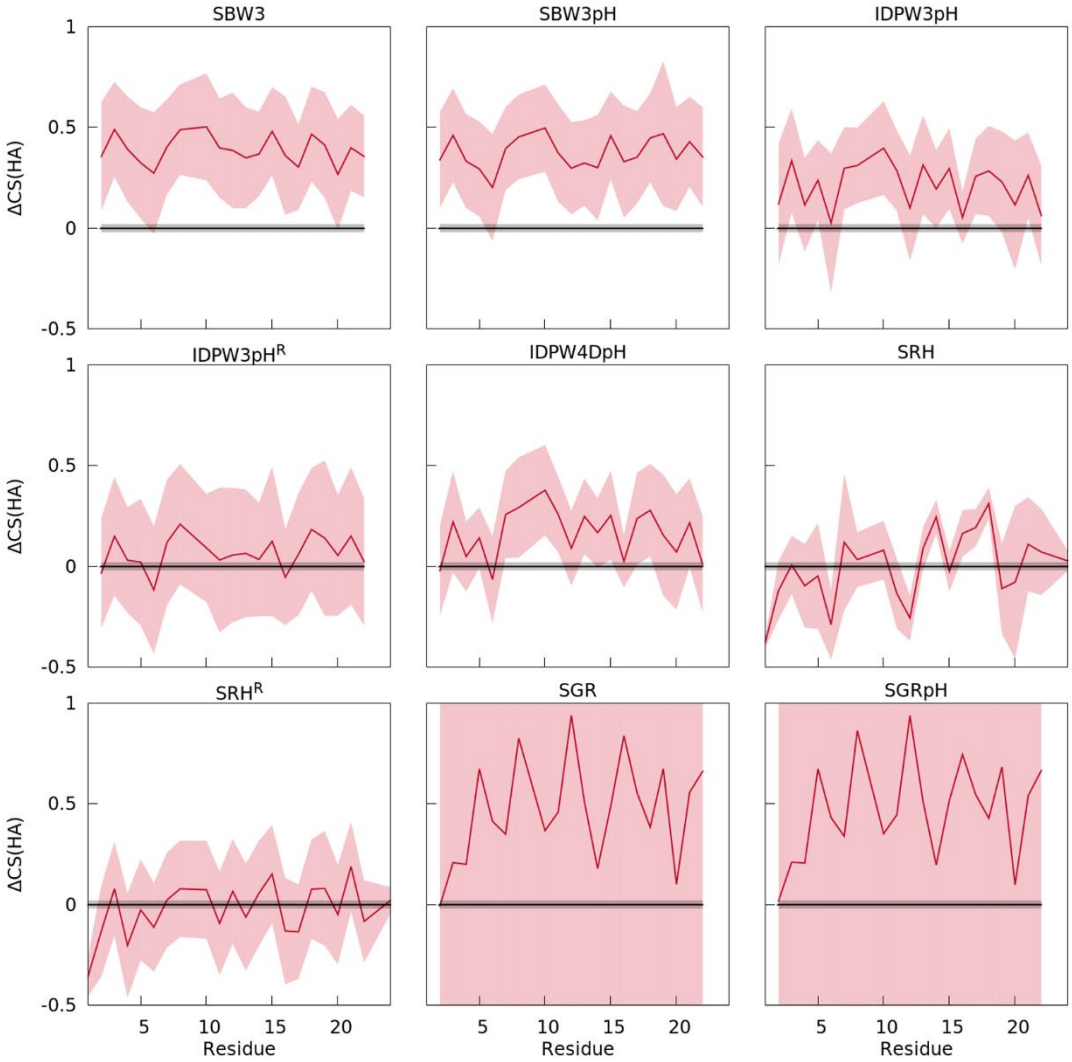


Figure 3. Deviations between the average predicted CS from simulations (in red) with respect to the experiment (in black) determined by Raj P.A. et al.[48]. The red and grey shades represent the associated error of the calculated chemical shifts in the simulations and experiment, respectively.

To further compare the simulations with respect to the available histatin-5 experiments, we have taken the HA atoms CS determined by NMR at pH 3.8, H2O/D2O solvent and 30ºC conditions from the work of Raj et al.[48]. In this case, the solvent pH in the simulations does not match the experiment, so deviations would be expected. Nevertheless, the IDPW3pH, IDPW4DpH and SRH^R simulations stand out in the linear regression between predicted and experimental CS, showing the best slope and r-values in Table 2. The slope and r-value provide information about the fit and correlation of the predicted and experimental data, respectively. In all three simulations we find a slope of

~1.0 within the standard deviation, while the two all-atom simulations show better r-values. In addition, the deviation between the simulated and experimental average CS and the associated errors for each HA atom of the amino acids are illustrated in Figure 3. It can be observed that the deviation of ΔCS(HA) is less than 0.5, which indicates a good result, and furthermore the error is lower than most simulations. In fact, SRH^R stands out with a very low ΔCS(HA) compared to the other simulations. On the other hand, the SBW3, SBW3pH and $IDPW3pH^R$ simulations also exhibit a good fit, and the r-values indicate that the experimental and predicted CS have a good correlation. In this case, the ΔCS(HA) values of SBW3 and SBW3pH are observed further apart, while, surprisingly, $IDPW3pH^R$ has a low ΔCS(HA) but a higher deviation compared to other simulations so far. In contrast, the SRH, SGR and SGRpH simulations dramatically fail in the linear regression of the CS(HA). Although the deviation of ΔCS(HA) in SRH is small, and the error even smaller, the linear fit shows that it is unable to reproduce the experimental CS. On the other hand, the conformations generated from the 2-bead model show a large deviation in ΔCS(HA) and an incredibly large error. These observations can be understood if we consider (i) the rigidity of the SRH simulation and (ii) the inaccurate back-mapping of the GC to all-atom resolution in the SGR and SGRpH simulations.

Therefore, the comparison with the available experimental data strongly indicates that the IDPW4DpH and SRH^R simulations are superior in reproducing the R_g, SAXS intensity profiles and NMR chemical shifts. Similarly, it suggests that the force field ff14IDPSFF holds potential to obtain conformations detected by SAXS but requires ingenuity in sampling strategies to representatively capture the conformational space. Finally, it points to the importance of a tool to accurately perform the back-mapping of the ESPResSO simulations given that the R_g distributions with the 2-beads model are promising but unable to convert these CG conformations into all-atom configurations that can reproduce the experimental data.

6.2.2. Disordered Structures Are Essential to Reproduce Experimental Observables

After validating the IDPW4DpH and SRH^R simulations against the SAXS and NMR observables, we next studied the conformational sampling of the simulations through the energy maps based on the PCA space. For this type of analysis, there is always the dimensionality problem of the $3N$ variables, which are the three Cartesian coordinates multiplied by N number of atoms of the peptide model. To handle the large number of

data, the PCA method has become a popular solution to reduce the complexity of this problem, which consists in efficiently transforming the coordinate data into the PCs by diagonalisation of the coordinate covariance matrix. From the eigenvectors and this transformation matrix, we can project the conformations into the PCA space and then calculate the Gibbs free energies within a set of PCs through a grid-based population analysis.

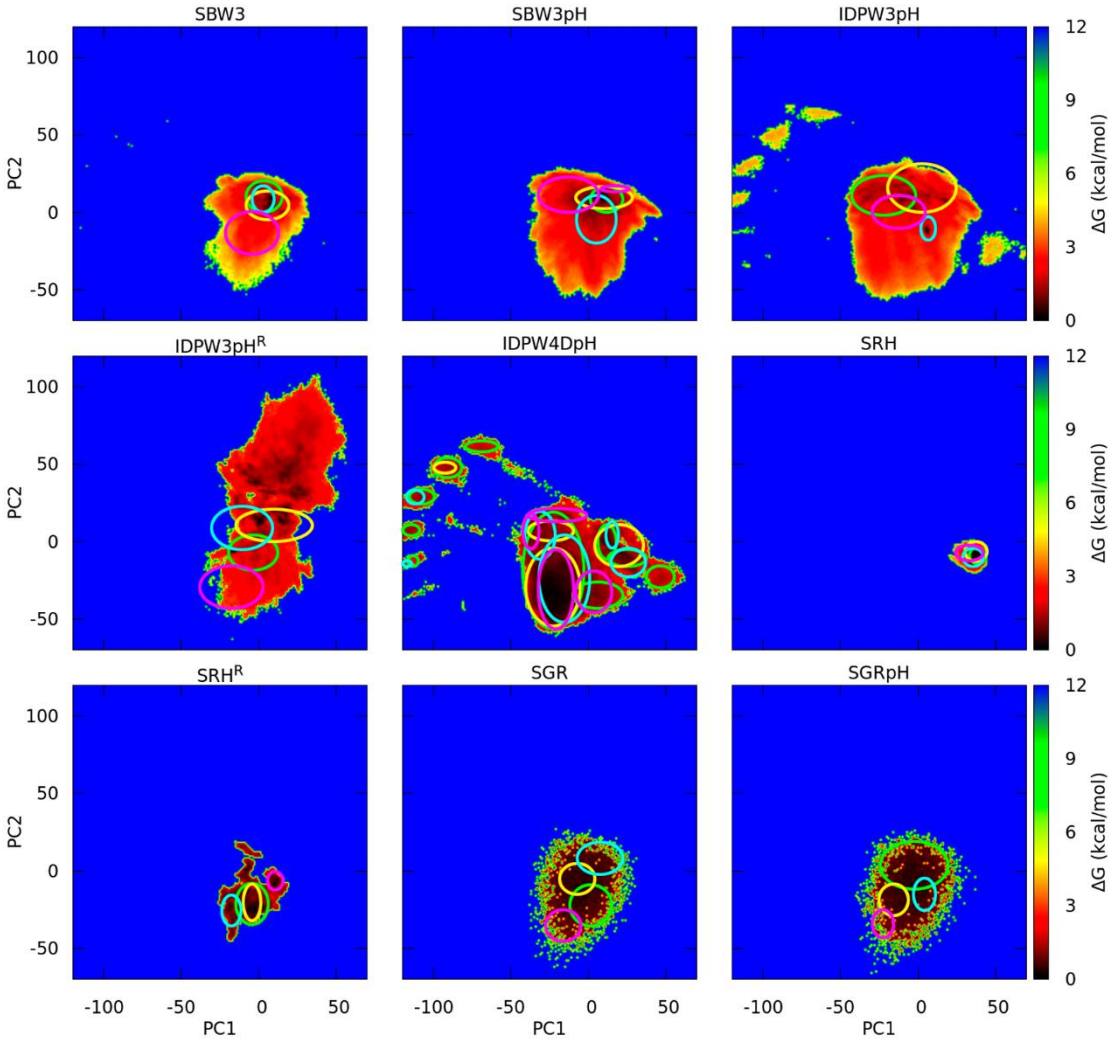


Figure 4. Energy maps in the PCA space of the conformational sampling of the histatin-5 simulations. The population Gibbs free energy is shown in black for the highest population regions and in blue for the regions not sampled. The conformational sampling of clusters C1, C2, C3 and C4 are indicated with green, yellow, cyan, and purple circles.

In this case, to represent and study the conformational space of the histatin-5, we used the first two PCs, which contain 45% of the covariance of the positions of the

configurations generated during the trajectory, to build the energy maps. The PCA energy maps of the simulations and the region sampled by each cluster within this space are illustrated in Figure 4. A quick look reveals significant dissimilarities between the simulations depending on the resolution of the model, all-atom or CG, and even between the bead models within the CG resolution. The all-atom simulations sample the conformational space more widely and locate the conformational minima more accurately. In contrast, the CG simulations perform an apparently homogeneous exploration within a narrower conformational space. Several observations are confirmed within the CG simulations. First, the SRH simulation is rigid and hardly adopts different conformations during the trajectory, which is reflected in the reduced space sampled in the energy map. On the other hand, the SGR and SGRpH simulations show a similar conformational space without any noticeable change when the pH-dependent protonation state sampling is introduced in the simulation. Nevertheless, there is a distinct clustering of the conformations, which makes the CpHMD simulation able to locate centroids with a better fit to the SAXS intensity profiles. Finally, the SRH^R simulation, the most promising simulation according to the SAXS and NMR experiments, also shows a restricted conformational space. There is more heterogeneity within the conformational sampling, highlighting in particular the location of clusters in the region of PC1 = [-25, 0] and PC2 = [-40, -15], which ones have good results in the SAXS intensity profiles. Therefore, in order to correctly reproduce the experimental observations, this region sampled by these clusters seems to be essential.

Among the all-atom simulations, we can distinguish between (i) the simulations that fail to reproduce any of the experimental observables, such as SBW3 and SBW3pH, (ii) the simulations that reproduce some SAXS or NMR observables moderately, such as IDPW3pH and IDPW3pHR, and (iii) finally the simulation that agrees with the experimental observables, IDPW4DpH. Based on this categorisation, the SBW3 and SBW3pH simulations exhibit the narrowest conformational sampling with respect to the other all-atom simulations. Furthermore, the SBW3 simulation is extremely restricted in the central region, around the point (0, 10). Fortunately, SBW3pH can explore a conformational space and locate more widespread minima within the sampled region. However, most of the clusters in both simulations are located around the (0, 10) point, which could probably mean that this set of conformations is not able to reproduce the experimental data. On the other hand, the IDPW3pHR simulation stands out for the large

conformational sampling due to the multi-seed simulation strategy. The low energy regions are distributed throughout the space, although the clusters are mainly located around the (0,0) point. This suggests that, despite the extended conformational sampling, the simulation prefers conformations located in the regions that we have previously detected in SBW3 and SBW3pH. Finally, the energy maps of IDPW3pH and IDPW4DpH show a similar conformational space, but the distribution of the minima within it is completely different. The first simulation samples small and discrete regions close to the point (0, 0), thus indicating a set of preferred conformations that do not fit the experimental measurements, as we have discussed in the previous section. In contrast, the IDPW4DpH simulation breaks with these conformational preferences and samples the conformational space more broadly and homogeneously within the defined space. The clusters are scattered throughout the energy map, particularly cluster C1, which includes more separated regions within the map. This is probably the reason behind the inability of the cluster C1 to capture the SAXS intensity profile. Given the complexity of the cluster distributions within the energy maps, more effort would be needed to understand the relationship between the conformational sampling and the experimental data. On the other hand, the long simulation times apparently allow for the exploration of regions significantly distant from the conformational sampling found in the other simulations, although both IDPW3pH and IDPW4DpH have different water models.

To understand the impact of the simulation setups on the conformational sampling, the secondary structures of the trajectories were also analysed using the DSSP method. Figure 5 illustrates the fractions of the secondary structure propensities (fpSS) of the histatin-5 conformations adopting β-sheets (anti- and parallel), helices (α, 3_{10} and π), turns and bends. The non-represented fraction in the plot corresponds to the absence of structure in the peptide, i.e., random coil structures. The SBW3 and SBW3pH simulations show higher structure compared to the other all-atom simulations. In particular, SBW3 achieves conformations with a higher helix fraction, a phenomenon already reported in the literature for the ff14SB force field[8,51]. In contrast, the SBW3pH simulation combines helix and β-sheet fractions depending on the fragment of the peptide and a reduced turn fraction in comparison with SBW3. This would indicate that the influence of the charge regulation affects the fpSS. In addition, the fpSS of the IDPW3pHR simulation also indicates a homogeneous appearance of β-sheet, as occurs in SWB3pH, although with a reduced formation of loops (turns and bends) and therefore an

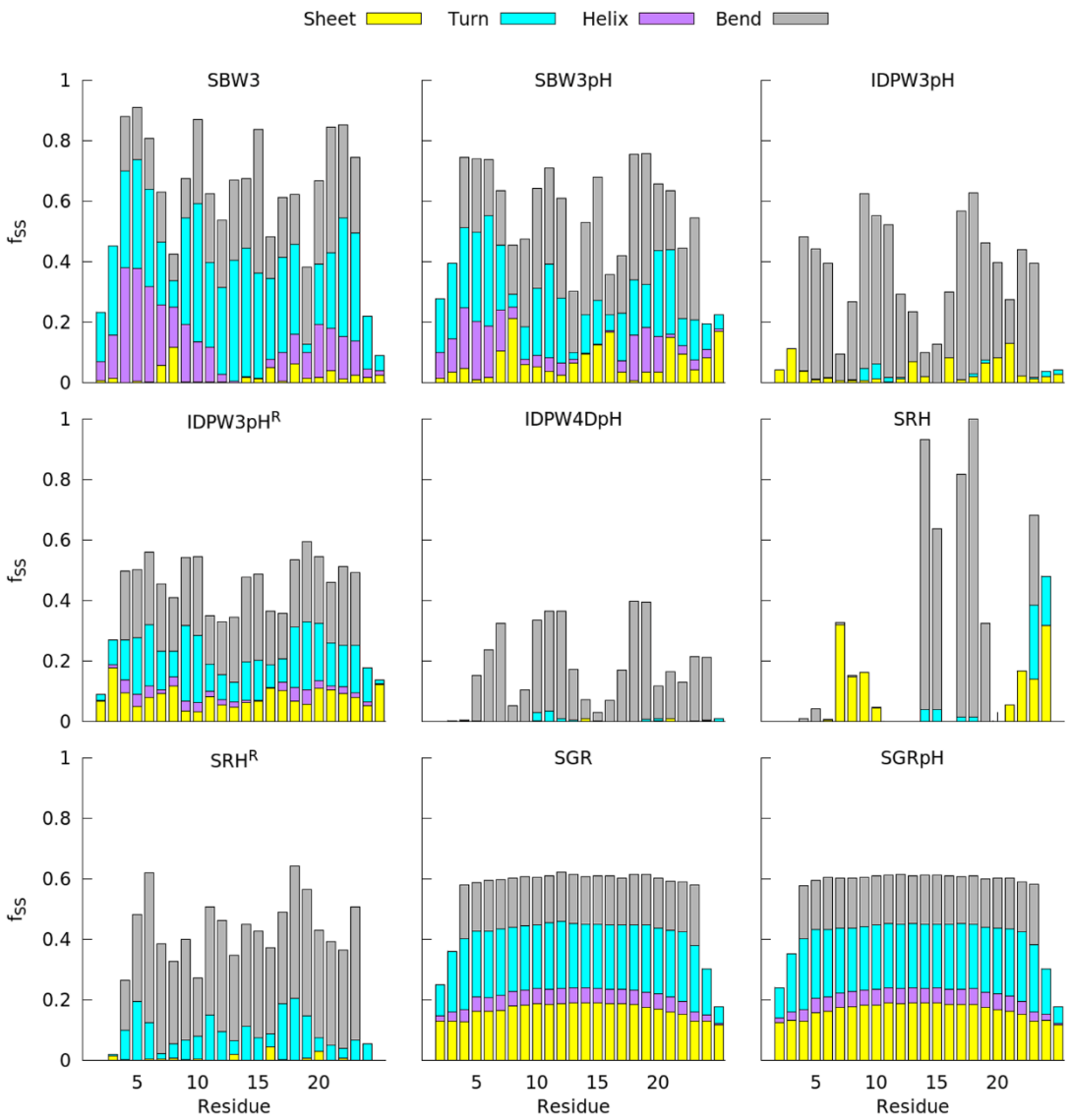


Figure 5. Secondary structure propensity fractions from the histatin-5 simulations. Anti- and parallel β-sheet, β-bulges and isolated β-strands are grouped in the "sheet" class, and α-, 3_{10}- and π-helices in the "helix" class.

increased appearance of random coil conformations. When examining the fpSS of the IDPW3pH simulation, it can be seen that prolonging of the simulation reduces the β-sheet formation and mainly promotes bends. Therefore, the extension of the simulation time of the histatin-5 peptide with the ff14IDPSFF force field favours the exploration of disordered conformations. Furthermore, when including the TIP4P-D water model in the simulation, which promotes the formation of more extended structures as we have observed in the R_g distributions, we can observe in the IDPW4DpH simulation that all fpSS are reduced, remaining only a small fraction of bends. These secondary structure propensities are also observed in the CHARMM36IDPSFF and CHARMM36m force

fields with the TIP3P modified water model, in which histatin-5 is practically a random coil according to the DSSP analysis[28]. In fact, both A99SB-disp with TIP4P-D-type and A99SB-ILDN with TIP4P-D water model show some helix or β-sheet content, pointing out that the choice of both force field and water model is essential for an accurate conformational sampling of IDPs such as the results reported in this work.

On the other hand, the SGR and SGRpH simulations show practically identical fractions with high β-sheet, turn and bend content, which must be the main reason behind the failure when comparing the experimental observables with the predictions extracted from the conformations after back-mapping. Furthermore, this reinforces that the conformational sampling is very similar and independent of the solvent pH when using ESPResSO. With respect to the SRH simulation, the fpSS indicate the highest disorder in the conformations with a punctual β-sheet formation in some specific amino acids. This could be explained by the low flexibility of SIRAH in the histatin-5 simulation. Finally, the SRH^R simulation shows a predominance of high loop formation (turns and bends, especially the latter). This turn/bend fraction is superior when compared to the IDPW4DpH simulation, which also has excellent predictions of the SAXS and NMR observables. Considering that the three SRH^R clusters shows a better R_g and SAXS intensity profile compared to IDPW4DpH, we could certainly attribute this inability of clusters C1 and C4 of IDPW4DpH to the formation of too extended and disordered structures. Therefore, the ff14IDPSFF/TIP4P-D combination would reproduce the SAXS intensity profile and the conformational space of histatin-5 more accurately with all-atom simulations if it did not promote the formation of disordered structures so much.

6.2.3. Conformational Sampling Determines the Protonation Fraction of Histidines

A particular interest of this thesis is the coupling of conformational sampling and the peptide protonation states. Studying the effects of conformation on protonation fractions, or vice versa, is difficult because both properties are interdependent. In this chapter we will only discuss the seven histidines of the histatin-5, ignoring the other titratable amino acids since the protonation state sampling was performed on the histidine amino acids. The protonation states of other ionisable amino acids were fixed given that the intrinsic pKa values are expected to be far from the solvent pH and therefore fully deprotonated states for the aspartic and glutamic acids and fully protonated states for

lysine or tyrosine can be assumed. Figure 6 shows the average protonation fraction of the histidines and the standard deviation from the four replicates of each simulation.

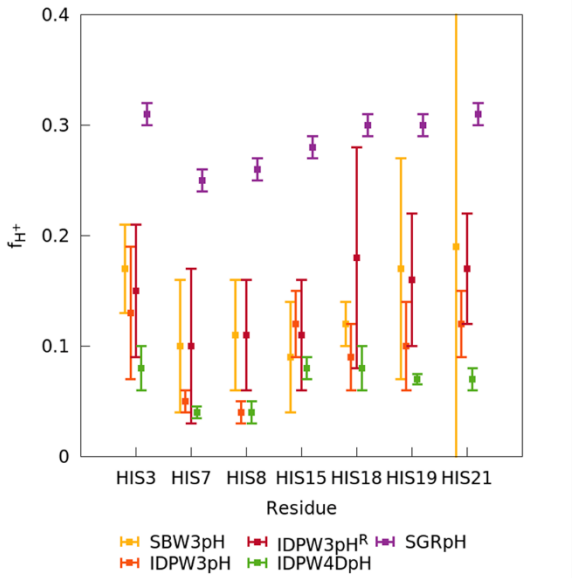


Figure 6. Protonation fractions of the histidine amino acids from the histatin-5 trajectories simulated with the CpHMD method. The standard deviation is calculated from the four replicas performed for each simulation.

The pKa of the imidazole ring of the histidine is 6.0, therefore the deprotonated state is expected at pH conditions above the intrinsic pKa. All simulations show a protonation fraction less than 50%, indicating that indeed the histidines are in a deprotonated state, despite the electrochemical environment provided by the histatin-5 peptide. In contrast, depending on the simulation setup, different protonation fractions can be observed in the imidazole ring. The CG simulation of SGRpH shows a higher protonation compared to the other simulations and has a protonation fraction of ~30%. The other simulations show a much lower protonation fraction of 20-5%. Starting with SBW3pH and IDPW3pHR, both simulations have similar protonation fractions for the histidine amino acids except for His-18. Furthermore, SBW3pH and IDPW3pHR exhibit the largest deviation in the protonation fraction among the replicas, especially at the end of the peptide chain. On the other hand, IDPW4DpH stands out for a high deprotonation, with all histidines having a protonation fraction below 10% and a low standard deviation. IDPW4DpH shows the largest disorder in the structure and now the highest deprotonation in the histidines, suggesting that the disordered structure and the low protonation fractions

may be related. In fact, the protonation fractions of the IDPW3pH simulation lie between the SBW3pH/IDPW3pHR and IDPW3pH protonation fractions. For most of the histidines the protonation fractions are more similar to the SBW3pH/IDPW3pHR although HIS-7, HIS-8 or HIS-18 are closer to the IDPW4DpH protonation fractions. We remind that the conformational sampling of IDPW3pH is also similar to IDPW4DpH, although the secondary structure propensity fractions are higher structure in the former.

Therefore, there is a clear relationship between the structure and the charge of the histatin-5. The structured conformations present in SBW3pH, IDPW3pHR or SGRpH tend to protonate the histidines, probably due to a higher (or stronger) number of intramolecular interactions between the amino acids of the peptide as the conformations are more compact and have a lower R_g. On the contrary, extended conformations with higher R_g make the histidines more solvent-exposed and with fewer intramolecular interactions, apparently favouring the deprotonation. This highlights the relevance of using an accurate force field not only for conformational sampling, but also for protonate state sampling when running simulations at constant pH. As we observed, the protonation fraction varies significantly with the force field, even when histatin-5 is simulated at pH conditions of minimum capacitance, i.e., with a low charge regulation response, either by donating or accepting protons according to the electrostatic interactions of the environment. Therefore, at pH conditions of maximum capacitance, as reported by Blanco P.M. et al.[26] or Skepö et al.[21], the coupling of conformations and charge regulation will be stronger. Not to mention the effect of the ionic charge in the solvent, which may play a key role in the conformation-charge dependence of histatin-5 and would be of particular interest to study in conjunction with the IDP-specific force fields developed in recent years.

6.3. Conclusions

In this chapter we have carried out an extensive conformational sampling of the model IDP histatin-5 using several all-atom and CG simulations to evaluate the force fields (ff14SB, ff14IPDSFF, SIRAH, Sugar), the water models (TIP3P, TIP4P-D), the simulation method (CMD, CpHMD) and the sampling strategy. The all-atom simulations have outperformed previous work in the literature in terms of simulation time length, thus providing a reliable conformational study of the histatin-5. To assess the conformational ensembles, we have relied on the experimental R_g, SAXS intensity profiles and NMR

chemical shifts. Among the all-atom simulations, the combination of the ff14IDPSFF force field and the TIP4P-D water model stands out with extended conformations, a broad R_g distribution and good reproducibility of the experimental observables. This conformational sampling behaviour of histatin-5 derives from the TIP4P-D water model, which enhances protein-solvent dispersion interactions as previously demonstrated in the literature[13,23]. In addition, we must stress that ff14IDPSFF also contributes to the random coil formation during the trajectory at long simulation times. On the other hand, the SRH^R simulation with CG resolution succeeds in reproducing the experimental NMR and SAXS observables with high accuracy, but a multi-seed sampling strategy was required to improve the conformational sampling. The other simulations have not been as successful as those mentioned in the analysis. However, the 2-bead model CG in the ESPResSO/Sugar simulation may be considered as good R_g distributions were reported. An adequate back-mapping of the CG model to all-atom resolution would be promising for this software package. Finally, the pH-dependent protonation state sampling with the CpHMD method yields an increase of disordered conformations in histatin-5.

Therefore, simulations at constant pH are recommended because the charge regulation according to the electrochemical environment confers more flexibility to histatin-5 in all-atom simulations, which would be critical for an accurate conformational sampling. However, this structure-charge coupling works in both directions, and to properly define the protonation state fractions, an accurate parameterisation of the force fields is also required. As discussed above, compact structures favour further protonation of the peptide, whereas, on the contrary, histidines are more deprotonated when extended conformations are reached. Therefore, it is necessary to carefully choose the force fields, water models and sampling strategies in order to obtain an accurate conformational sampling and reproduce the experimental observables, but also to sample the protonation fractions of the ionisable amino acids.

6.4. Bibliography

1. Chouard, T. Structural biology: Breaking the protein rules. *Nature* **471**, 151–153 (2011).
2. Ferreon, C., Chris, A., Ferreon, M., Trivedi, R. & Nagarajaram, H. A. Intrinsically Disordered Proteins: An Overview. *Int J Mol Sci* **23**, 14050 (2022).
3. Jakob, U., Kriwacki, R. & Uversky, V. N. Conditionally and Transiently Disordered Proteins: Awakening Cryptic Disorder To Regulate Protein Function. *Chem Rev* **114**, 6779–6805 (2014).

4. Chong, S.-H., Chatterjee, P. & Ham, S. Computer Simulations of Intrinsically Disordered Proteins. *Annu Rev Phys Chem* **9**, 4 (2017).

5. Ciemny, M., Badaczewska-Dawid, A., Pikuzinska, M., Kolinski, A. & Kmiecik, S. Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields. *Int J Mol Sci* **20**, 606 (2019).

6. Song, D., Luo, R. & Chen, H.-F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **57**, 1166–1178 (2017).

7. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Nat Acad Sci USA* **115**, E4758–E4766 (2018).

8. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J Chem Theory Comput* **16**, 528–552 (2020).

9. Huang, J. *et al.* CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Methods* **14**, 71–73 (2017).

10. Henriques, J., Cragnell, C. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput* **11**, 3420–3431 (2015).

11. Gong, X., Zhang, Y. & Chen, J. Advanced Sampling Methods for Multiscale Simulation of Disordered Proteins and Dynamic Interactions. *Biomolecules* **11**, 1416 (2021).

12. Mu, J., Liu, H., Zhang, J., Luo, R. & Chen, H.-F. Recent Force Field Strategies for Intrinsically Disordered Proteins. *J Chem Inf Model* **61**, 1037–1047 (2021).

13. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J Phys Chem B* **119**, 5113–5123 (2015).

14. Izadi, S., Anandakrishnan, R. & Onufriev, A. v. Building Water Models: A Different Approach. *J Phys Chem Lett* **5**, 3863–3871 (2014).

15. Kadaoluwa Pathirannahalage, S. P. *et al.* Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations. *J Chem Inf Model* **61**, 4521–4536 (2021).

16. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).

17. Privat, C., Madurga, S., Mas, F. & Rubio-Martínez, J. On the Use of the Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides. *Polymers* **13**, 99 (2021).

18. Privat, C., Madurga, S., Mas, F. & Rubio-Martinez, J. Unravelling Constant pH Molecular Dynamics in Oligopeptides with Explicit Solvation Model. *Polymers* **13**, 3311 (2021).

19. Pollock, J. J., Denepitiya, L., MacKay, B. J. & Iacono, V. J. Fungistatic and fungicidal activity of human parotid salivary histidine-rich polypeptides on Candida albicans. *Infect Immun* **44**, 702–707 (1984).

20. Iovino, M., Falconi, M., Marcellini, A. & Desideri, A. Molecular dynamics simulation of the antimicrobial salivary peptide histatin-5 in water and in trifluoroethanol: a microscopic description of the water destructuring effect. *J Pept Res* **58**, 45–55 (2001).

21. Kurut, A. A., Henriques, J., Forsman, J., Skepö, M. & Lund, M. Role of histidine for charge regulation of unstructured peptides at interfaces and in bulk. *Proteins* **82**, 657–667 (2014).

22. Cragnell, C., Durand, D., Cabane, B. & Skepö, M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins* **84**, 777–791 (2016).

23. Henriques, J. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the TIP4P-D Water Model and the Representativeness of Protein Disorder Models. *J Chem Theory Comput* **12**, 3407–3415 (2016).

24. Jephthah, S., Staby, L., Kragelund, B. B. & Skepö, M. Temperature Dependence of Intrinsically Disordered Proteins in Simulations: What are We Missing? *J Chem Theory Comput* **15**, 2672–2683 (2019).

25. Sullivan, S. S. & Weinzierl, R. O. J. Optimization of Molecular Dynamics Simulations of c-MYC1-88—An Intrinsically Disordered System. *Life* **10**, 109 (2020).

26. Blanco, P. M., Madurga, S., Garcé, J. L., Mas, F. & Dias, R. S. Influence of macromolecular crowding on the charge regulation of intrinsically disordered proteins. *Soft Matter* **17**, 655 (2021).

27. Shrestha, U. R., Smith, J. C. & Petridis, L. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun Biol* **4**, 243 (2021).

28. Jephthah, S., Pesce, F., Lindorff-Larsen, K. & Skepö, M. Force Field Effects in Simulations of Flexible Peptides with Varying Polyproline II Propensity. *J Chem Theory Comput* **17**, 6634–6646 (2021).

29. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

30. Machado, M. R. *et al.* The SIRAH 2.0 Force Field: Altius, Fortius, Citius. *J Chem Theory Comput* **15**, 2719–2733 (2019).

31. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

32. Darré, L., Machado, M. R., Dans, P. D., Herrera, F. E. & Pantano, S. Another Coarse Grain Model for Aqueous Solvation: WAT FOUR? *J Chem Theory Comput* **6**, 3793–3807 (2010).

33. Case, D. A. *et al.* Amber 2018. *University of California, San Francisco* (2018).

34. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pK a Predictions. *J Chem Theory Comput* **7**, 525–537 (2011).

35. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicist*. (1999).

36. Perez, J. J., Santos Tomas, M. & Rubio-Martinez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J Chem Inf Model* **56**, 1950–1962 (2016).

37. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 2017).

38. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089 (1998).

39. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

40. Limbach, H. J., Arnold, A., Mann, B. A. & Holm, C. ESPResSo-an extensible simulation package for research on soft matter systems. *Comput Phys Commun* **174**, 704–727 (2006).

41. Machado, M. R. & Pantano, S. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **32**, 1568–1570 (2016).

42. Hass, M. A. S. & Mulder, F. A. A. Contemporary NMR Studies of Protein Electrostatics. *Annu Rev Biophys* **44**, 53–75 (2015).

43. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

44. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* **44**, (2016).

45. Sagar, A., Jeffries, C. M., Petoukhov, M. v., Svergun, D. I. & Bernadó, P. Comment on the Optimal Parameters to Derive Intrinsically Disordered Protein Conformational Ensembles from Small-Angle X-ray Scattering Data Using the Ensemble Optimization Method. *J Chem Theory Comput* **17**, 2014–2021 (2021).

46. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* **48**, 13–22 (2010).

47. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528–1532 (2015).

48. Raj, P. A., Marcus, E., Sukumaran, D. K. & Disease, P. Structure of Human Salivary Histatin 5 in Aqueous and Nonaqueous Solutions. *Biopolymers* **45**, 51–67 (1998).

49. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

50. Williams, T. *et al.* Gnuplot 4.6: an interactive plotting program. *Software Manual* 238 (2012).

51. Duong, V. T., Zihao, C., Thapa, M. T. & Luo, R. Computational Studies of Intrinsically Disordered Proteins. *J Phys Chem B* **122**, 10455–10469 (2018).

This chapter has been published in Privat, C., Madurga, S., Mas, F., Rubio-Martinez, J. Molecular Dynamics Simulations of an α-Synuclein NAC Domain Fragment with a ff14IDPSFF IDP-specific Force Field Suggest β-Sheet Intermediate States for Fibrillation. *Physical Chemistry Chemical Physics* **24**, 18841–18853 (2022)

Chapter 7

Molecular Dynamics Simulations of α-Synuclein NAC Domain Fragment with ff14IDPSFF IDP-specific Force Field Suggest β-Sheet Intermediate States for Fibrillation

Parkinson's disease (PD) is the second most common neurodegenerative disorder in the world's population, particularly in the people over 60 years of age.[1,2] The development of PD is mainly attributed to the aggregation of misfolded α-synuclein (αS) protein in Lewis bodies, which ultimately leads to the loss of dopaminergic neurons.[3–5] To date, no definitive cure for this disease has been found. The initial stages of the mechanism of formation of these fibrils from the αS monomers are still unknown, which in turn hinders the design of drugs to treat PD. The micelle-bound αS monomers[6,7] and the fibril oligomers[8] have been characterised, but the structure of the protein in free solution or the intermediate conformations of the fibrillation process have not been reported. The difficulty in identifying the intermediates of fibrillation is due to the transient nature of αS as an intrinsically disordered protein (IDP). IDPs are distinguished for their structural disorder and their ability to rapidly interconvert between conformational states over time. Some experimental techniques, such as nuclear magnetic resonance (NMR), small-angle X-ray scattering (SAXS) or far-UV dichroism[8–13], can capture structural properties of IDPs, but only average observations of the conformational ensemble of the protein are obtained. At this point, computational studies come into play as a resource capable of providing insights into IDPs through atomistic simulation methods. However, αS fibrillation occurs on a time scale that is computationally inaccessible to traditional simulation methods. Some studies have attempted to overcome this limitation by ingenious approaches ranging from enhanced-sampling techniques[14–17], coarse-grained models[18,19], simulations of specific fragments of αS[18,20] or guiding the simulations an experimental data bias.[21,22] Although these efforts have led to very

promising results in understanding the αS fibril mechanism, further research is still needed.

αS is a 140-aa presynaptic protein found mainly in nervous tissue, and its function is still poorly understood[23]. It has been associated with several biological processes, such as synaptic vesicle recycling, regulation of DNA repair or involvement in neuronal apoptosis[24]. The primary structure of αS is divided into (i) the N-terminal domain (1-60 aa), (ii) the non-amyloid-β component (NAC) domain (61-95 aa) and (iii) the C-terminal domain (96-140 aa). The first domain consists of 7 imperfect repeats of 11 amino acids which confer an amphipathic character and an overall positive charge. These repeats contain abundant KTKEGV segments, which have a propensity to adopt α-helix conformations and allow αS to bind to membranes[6,25]. The NAC domain acts as a hydrophobic core for fibrillation[26]. The C-terminal domain is highly charged and mobile due to the abundance of acidic amino acids in its chain. Some studies suggest that the C-terminal tail is responsible for inhibiting fibril formation by burying the NAC domain, thus preventing interactions between the monomers prior to oligomerisation[17,18]. Interestingly, while the membrane-bound monomeric αS structures have a high α-helix content in the N-terminal and NAC domains, the fibril conformation is precisely characterised by the agglomeration with β-sheets. Apparently, the β-sheet structure may be critical in the early stages of the fibrillation process, although a study suggests that an intermediate α-strand/sheet intermediate may be necessary for the fibrillation mechanism[27].

Most atomistic simulation studies to date use classical force fields that generally ignore the transient nature of IDPs and lead to conformational biases. Historically, these classical force fields have been parameterised to accurately reproduce well-defined, experimentally determined three-dimensional structures. Fortunately, in recent years, some force fields have been developed to include disorder structure in their parameterisation. In fact, previous assessments of standard and modern force fields have demonstrated a high sensitivity in obtaining IDP conformational ensembles[28,29]. Approaches to incorporate the intrinsic disorder of IDPs can range from adjusting the dihedral parameters (CHARMM22*[30], RSFF2[31] or OPLS-AA/M[32] force fields), adding a grid-based energy correction term to the φ/ψ dihedral energy surface called CMAP method (CHARMM36[33], ff14IDPSFF[34], ESFF1[35]) or refining the protein-water interactions (a99SB-disp[36], ff03ws[37], CHARMM36m[38]). Among these new force fields,

ff14IDPSFF has been developed as a promising force field capable of correcting the dihedral distributions of all 20 amino acids from the popular ff14SB[39] force field by adding the CMAP energy term. To provide further confidence, this IDP-specific force field has been shown to improve the description of chemical shifts in the α-synuclein protein.[34]

In this chapter, we performed simulations of a fragment of the αS protein using the ff14SB and ff14IDPSFF force fields, which we refer to as E-αSNAC^{SB} and E-αSNAC^{IDP} respectively, to examine the generated conformational ensembles and gain insight into the intermediate states of fibrillation. This αS fragment is defined by the 35-97 amino acids that constitute the fibril core of the Greek-key topology adopted by the αS fibril according to Tuttle et al.[8], and includes the NAC domain and a segment of the N-terminal domain. In addition, the conformational space of this fragment was previously explored by 18 μs T-REMD simulation with the CHARMM27* force field[20]. The work of Jain et al. concluded that there is an α-helix content at room temperature, while β-sheets are formed at high temperatures. Here, we have carried out 6 μs-length simulations for each force field and observed a significant bias in the secondary structure propensities after analysing the conformational ensembles and contact maps. The ff14SB simulations preserve the α-helices of the micelle-bound αS structure, while the conformational sampling of ff14IDPSFF stands out with random coil and β-sheet structures, the latter present in the αS fibrils. Furthermore, we have verified that ff14IDPSFF shows a closer agreement with four sets of chemical shifts obtained from NMR in solution by determining the linear equations and the Pearson correlation coefficients. The IDP-specific force field is able to adopt intermediate states with β-sheet conformations that are not normally found in conventional force fields with 6 μs-length simulations, which demonstrates the potential of this force field to study the conformational space of the αS and its role in the fibrillation mechanism.

7.1. Materials and Methods

7.1.1. Human α-Synuclein Protein Structure

The Protein Data Bank database (https://www.rcsb.org/) contains many αS structures available to the scientific community. For our study we selected the structure characterised by solution NMR with PDB-code 1XQ8 (human-micelle bound αS).[6] The 1XQ8 structure was cleaved, retaining the core amino acids of the protein (35-97) that

comprise the NAC domain and a segment of the N-terminal domain. We have renamed this selection of amino acids as the *extended α-synuclein NAC domain* (E-αSNAC) fragment.

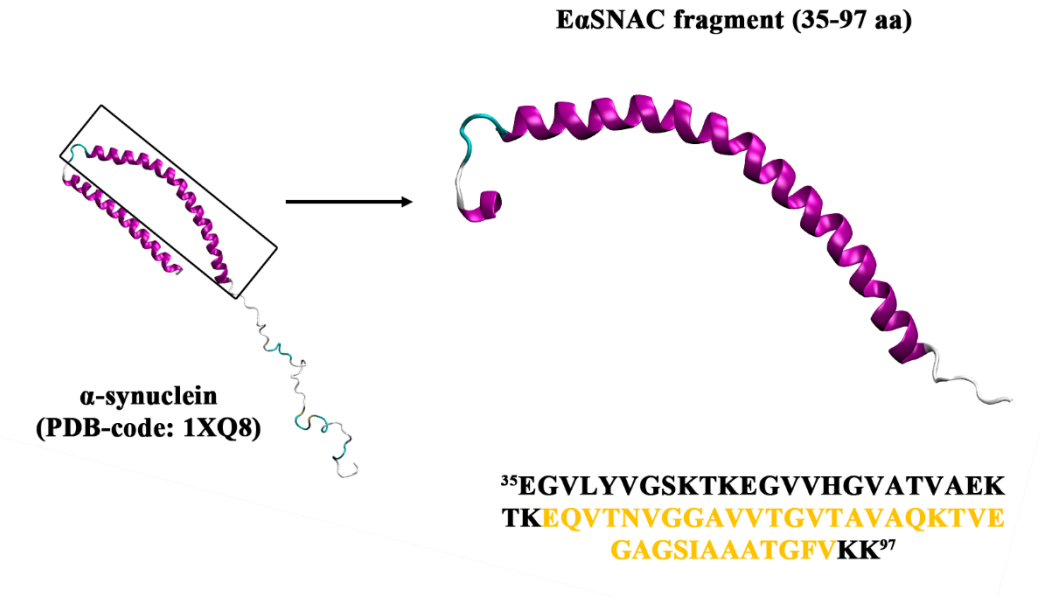


Figure 1. NMR structure of the human micelle-bound α-synuclein (PDB-code: 1XQ8) on the left and structure and amino acid sequence of the E-αSNAC fragment on the right. The NAC domain amino acids (residues 61-95) are coloured in orange.

7.1.2. Structure Preparation and Simulation Setup

E-αSNAC was oriented according to its principal moments of inertia (I_x, I_y, I_z) using an internal script of the research group. Next, the LEaP module of AMBER18[40] was used to parameterise the model with the ff14SB[39] force field and to define a box of dimensions 140x140x140 $Å^3$ as the simulation system. The simulation box was filled with TIP3P water molecules[41], with a space of 1.0 Å between any amino acid of the protein and the water molecules. The net charge of the system was neutralised by adding Na^+ or Cl^- counterions. The hydrogen mass was distributed among the amino acid atoms using the ParmEd module to increase the time step from 2 fs to 4 fs[42]. We then built a second model system with a new topology that included the parameterisation of the IDPs-specific force field ff14IDPSFF[34].

A three-phase protocol was then applied to minimise the molecular models. This protocol consisted of using the steepest descent (SD) method[43] with three levels of restraints on the protein to relax the internal tensions of the system after the addition of

the solvent molecules. In the first minimisation stage, 5000 SD steps were performed restricting the entire protein to relax only the surrounding water molecules. In the second, the side chains of the amino acids and the solvent molecules were slowly relaxed during 5000 SD steps, applying restraints only to the backbone atoms of the protein. Finally, the entire system was freely minimised for a further 5000 SD steps. All the above restraints were defined by a force constant of 5 kcal·mol^{-1}·Å $^{-2}$.

The simulation box was heated with a linear increase of 1 K·ps^{-1} in the canonical ensemble (NVT) for 300 ps until a final temperature of 300K was reached. A second equilibration step was then performed for 300 ps in the isobaric-isothermal ensemble (NPT) to adjust the volume of the system box to a pressure of 1.0 atm. During the heating and equilibration steps, the protein backbone atoms were restrained with a force constant of 5 kcal·mol^{-1}·Å^{-2}. To increase the conformational sampling, three replicas of the system were generated[44] and random initial velocities were assigned following a Maxwell-Boltzmann distribution. Each replica was integrated for a time length of 2 μs. Trajectory coordinates were recorded every 20 fs and the output data every 40 fs. The SHAKE algorithm[45] constrained the hydrogen-involving bonds during the simulation and the temperature was maintained at 300K with the Langevin thermostat[46] and a collision frequency of 3.0 ps^{-1}. A 9.0 Å cut-off and periodic boundary conditions (PBCs) were applied.

7.1.3. Conformational Analysis

The simulations were analysed by calculating the structural properties and comparing the NMR experiments with the conformational ensembles after estimating the chemical shifts. The visualisation of the trajectories and the illustration of the conformations were performed with VMD software[47]. Plots were generated using Gnuplot (version 4.6)[48].

Conformational properties were calculated with the CPPTRAJ module[49] of AMBER18. The root-mean-square deviation (RMSD) of the protein backbone (C_α, C, N, O atoms) was calculated with respect to the reference structure, which corresponds to the 1XQ8 structure determined by NMR in solution. We also calculated the root-mean-squared fluctuation (RMSF) and the radius of gyration (R_g) of the C_α atoms to provide insight into the flexibility and compactness of the conformations obtained from the simulation. The secondary structure propensity fractions (fpSS) of the protein were

estimated using the DSSP method.[50] They were categorised into five classes: β-strand (isolated β-bridges, β-bulges, and extended strands), helices (3_{10} helix, α-helix, and π-helix), coil (no secondary structure assignment), turn (isolated hydrogen-bonded turn) and bend. To provide further insight into the secondary structures, intramolecular contacts were calculated using CPPTRAJ, for which we accepted as a contact any atom (except hydrogen) with a distance of less than 8.0 Å to another amino acid atom. The contacts that are defined in the initial structure are referred to as native contacts, while the new contacts that appear during the simulation are referred to as non-native contacts.

For clustering, the RMSD of the C_α atoms was first calculated as a distance metric using a 5-frame sieve. Then, starting with each conformation as an individual cluster, the clusters were merged according to the average distance between the members of the clusters until all conformations were grouped into 15 clusters. The RMSD of the centroid conformations of the five most populated clusters was calculated and plotted on the 2D-RMSD plot. The conformational space was represented by Principal Component Analysis (PCA) technique. The covariance of the distance between C_α atoms was used as metric. The conformations of the trajectories were projected onto the first two PCs (PC1 and PC2), which represent 33% of the covariance. Next, an in-house script estimated the Gibbs free energy of the populations in PC1 and PC2 according to Eq. 1, where n_i and n_{max} are the population in bin i and the bin of maximum occupation, respectively to build the energy maps in the PCA space.

$$\Delta G = -k_b T ln(n_i/n_{max}) \qquad \textbf{Eq. 1}$$

The chemical shifts of the ^{15}N, ^{13}C, $^{13}C_\alpha$ and $^{13}C_\beta$ atoms of one-fifth of the conformations were estimated using SPARTA+[51]. The averages of the chemical shifts of each atom were calculated and the linear regression and Pearson correlation coefficients between the simulated and experimental chemical shifts were then obtained using Gnuplot and SciPy[52], respectively. The NMR measured chemical shifts were obtained from the Biological Magnetic Resonance Data Bank (BMRB) database (https://bmrb.io/) with the following IDs: 18857[53], 19337[54], 25527[55] and 6968[56].

7.2. Results

6 μs simulations of the extended αS NAC (E-αSNAC) domain fragment for the ff14SB (E-αSNAC^{SB}) and ff14IDPSFF (E-αSNAC^{IDP}) force fields were performed with three replicas of 2 μs-length and random initial velocities. A total simulation time of 12 μs was run in this study, regardless of the parametrisation.

7.2.1. Mobility and Compactness of Trajectories

First, the conformational properties (RMSD, RMSF and radius of gyration) of the ff14SB and ff14IDPSFF simulations were calculated. After superimposing the conformations from the trajectories on the NMR-characterised 1XQ8 structure, the RMSD of the protein backbone (C_α, C, N, O atoms) was calculated for each simulation and illustrated in Figure 2A. The fluctuations of the amino acids during the simulation are illustrated in Figure 2B.

Distinct RMSD and RMSF are observed depending on the replica, highlighting the advantage of performing multiple production runs of the system to explore the conformational space extensively. The time-dependent RMSD in Figure 2A shows that the ff14SB simulation has values below 21 Å. Interestingly, replica 1 yields conformations with RMSD values around 15 Å and the distribution shows conformations predominantly in two RMSD ranges (~17 and ~21 Å). Furthermore, replica 2 shows a large peak in the distribution and a stable time-dependent RMSD, suggesting that the protein conformation does not change significantly during the trajectory. On the other hand, the time-dependent RMSD of ff14IDPSFF apparently fluctuates more compared to the ff14SB simulation. In fact, replica 2 stands out with values above 21 Å, even reaching conformations with RMSD around 25 Å. The peaks of the distributions of ff14IDPSFF are around 20 and 24 Å. Furthermore, the fluctuations of the residues according to each force field simulation are illustrated in Figure 2B, in which it is observed that the RMSF of ff14SB exhibits different behaviour between replicas at residues 35-47 and 65-75, while the rest of the protein is similar. In contrast, ff14IDPSFF shows a distinct RMSF in almost the entire E-αSNAC fragment, even exhibiting Δ(RMSF) of ~10 Å at residues 65-72. The RMSD of ff14IDPSFF shows larger distances with respect to the ff14SB counterpart, indicating a conformational sampling more distant from the α-helix-rich initial structure, while the RMSF shows less pronounced divergence between the ff14SB

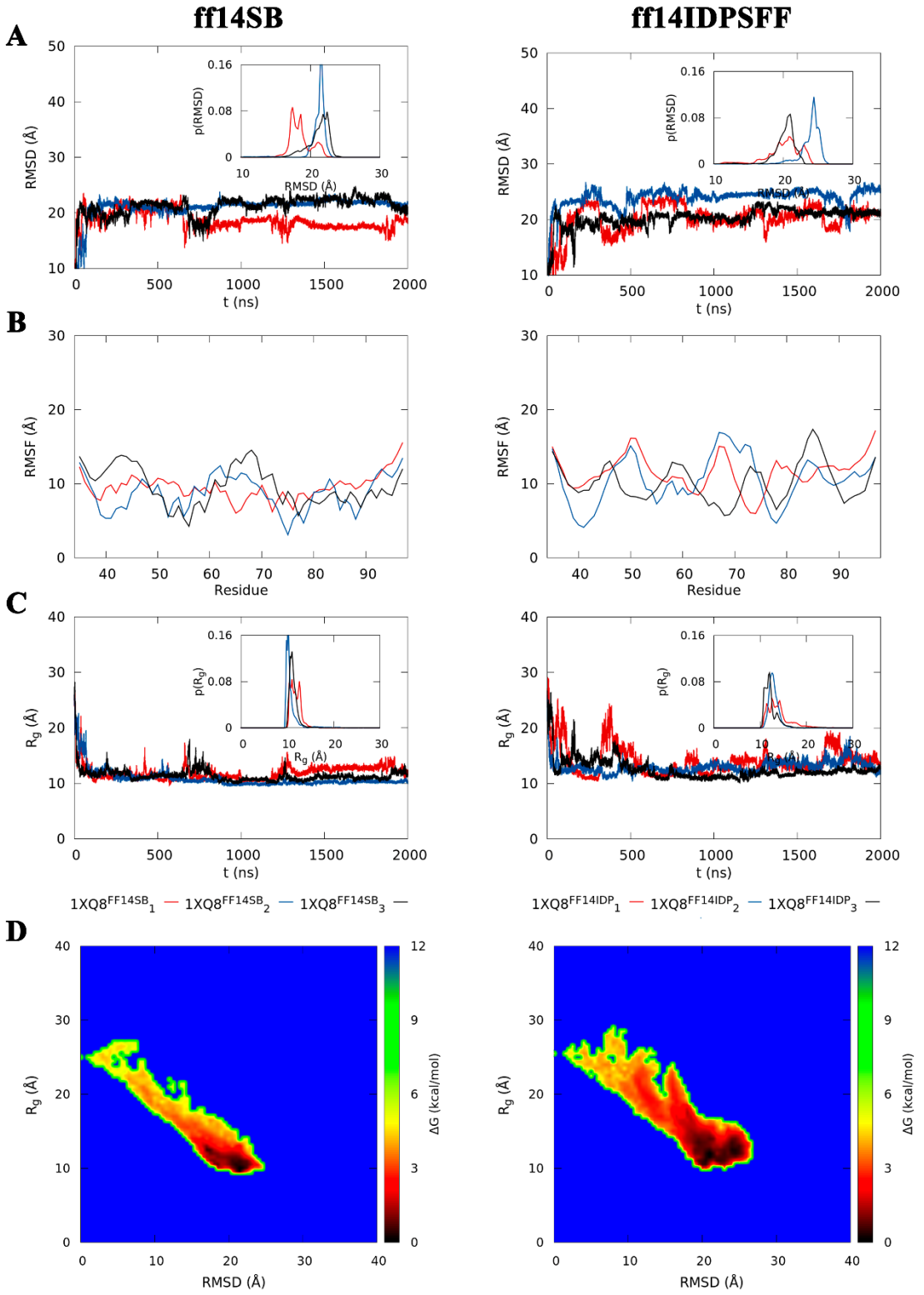


Figure 2. Conformational properties of the ff14SB (left) and ff14IDPSFF (right) simulations of E-αSNAC. (A) RMSD, (B) RMSF and (C) R_g of each replica are shown in red, blue and black lines, respectively. The distributions of RMSD (p(RMSD)) and radius of gyration (p(R_g)) are also illustrated within each plot. (D) Energy maps of the RMSD and radius of gyration expressed in kcal·mol^{-1}.

replicas. This may indicate that the IDP-specific force field samples regions in conformational space where amino acid mobility varies significantly.

To examine the compactness of the E-αSNAC fragment, the time-dependent radius of gyration and the distributions are illustrated in Figure 2C, and the energy maps constructed in the RMSD/R_g space are shown in Figure 2D. The average R_g values are (11.4 ± 1.5) Å for E-αSNAC^{SB} and (13.2 ± 2.1) Å for E-αSNAC^{IDP}. The ff14SB simulation shows a stable R_g, consistent with the narrow distributions at ~11 Å. In contrast, the R_g of ff14IDPSFF exhibits more significant fluctuations than the ff14SB force field along with broad distributions located at ~13Å. Although the R_g averages of the force fields are close, the conformations of ff14SB appear modestly more compact compared to ff14IDPSFF.

On the other hand, the energy maps in Figure 2D show that ff14SB has a thin, highly populated region depicted in black in the plot. The ff14IDPSFF simulation enlarges this black region and samples conformations with higher RMSD and R_g values, which is consistent with the reported observations and suggests that the IDP-specific force field can capture distant conformations from the initial α-helix-rich structure.

7.2.2. Secondary Structure Propensities and Contact Maps

The DSSP method of Kabsch and Sander[50] was applied to all conformations to calculate the secondary structure propensity factors (fpSS) in Figure 3A. For convenience, the α-helix, π-helix, and 3_{10}-helix were grouped into the helix class and the parallel and antiparallel β-sheets, β-bulges, and isolated β-strands into β-strand. The other SS propensities mantained their conventional names (turn, bend, and coil). The fpSS in E-αSNAC^{SB} indicate a predominant helix content (39%) that coexists with coil (25%), turn (20%) and bend (14%) conformations. In the case of ff14IDPSFF, the fpSS of E-αSNAC^{IDP} show a large random coil content (44%) together with β-strand (26%), bend (16%) and turn (12%) conformations.

The DSSP analysis reveals that the parameterisation of the E-αSNAC fragment leads to significant changes in the adopted conformations and therefore in the fpSS. ff14SB preserves the helices found in the NAC and N-terminal domains of the 1XQR structure for a considerable simulation time, especially at residues 40-50, 50-70 and 85-90. Nevertheless, bends, turns and random coil conformations are formed during the

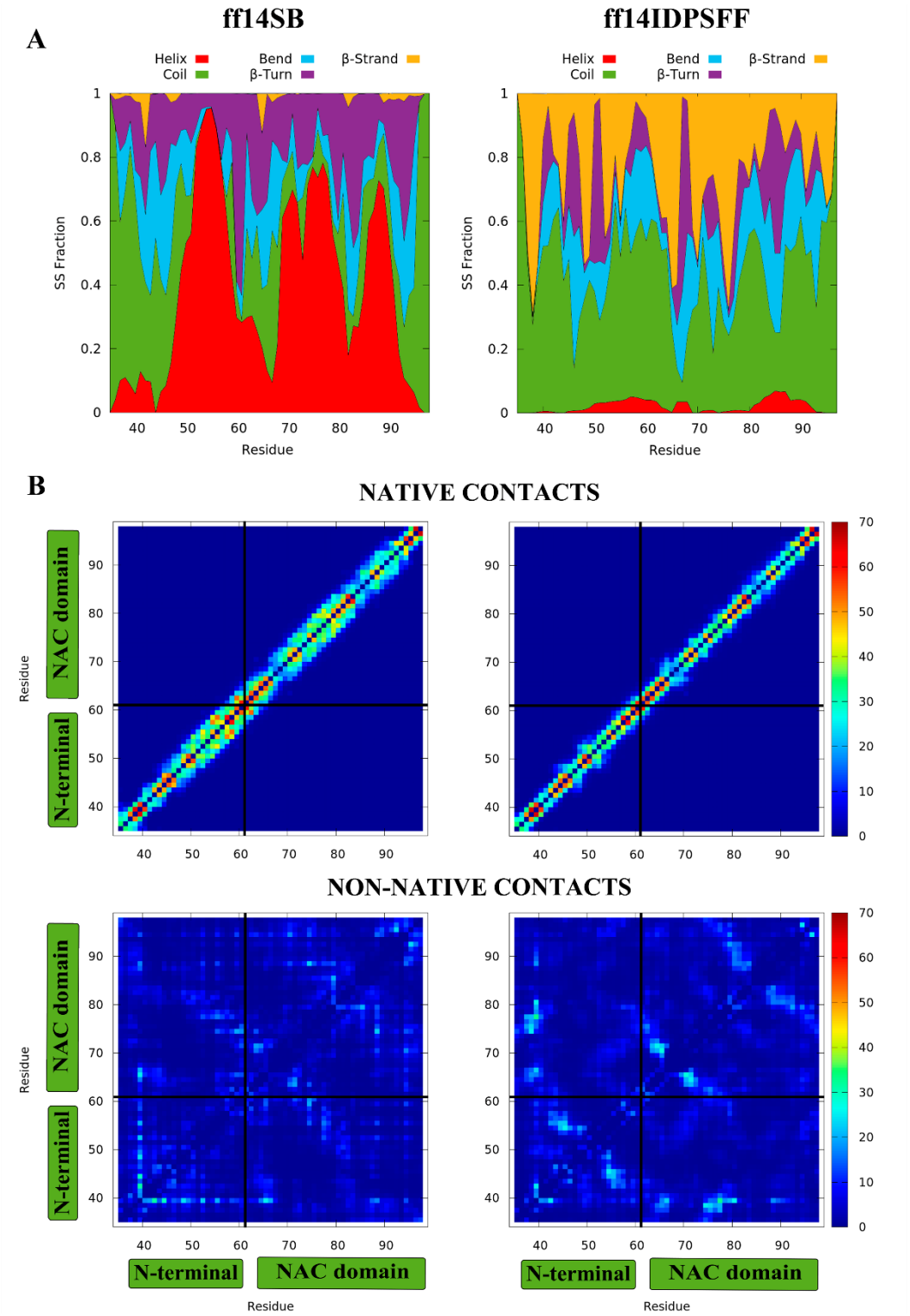


Figure 3. (A) fpSS of E-αSNAC. The helix class (red) includes α-helix, π-helix and 3_{10}-helix, and the β-strand class (orange) includes β-sheets, β-bulges and isolated β-strands. Turns, bends and random coils are coloured in purple, cyan and green, respectively. (B) Native (middle) and non-native (bottom) contact maps of E-αSNAC. The N-terminal and NAC domains are separated in the plot by black lines.

simulation and break with the high helicity of the native structure. Surprisingly, ff14IDPSFF shows an almost complete absence of helical conformations, which are instead replaced by random coils and other conformations in lower abundance such as β-strands, bends and turns. The large random coil content together with the breaking of the native helices could explain the greater flexibility indicated by RMSD and R_g.

In addition, the residue contacts were tracked using an 8 Å distance cut-off. For each force field, two contact maps are illustrated in Figure 3B. Native contacts are those distances between atoms below the cut-off identified in the 1XQ8 structure and conserved during the conformational sampling. Non-native contacts are those contacts produced in the simulation that do not appear in the reference NMR structure. The contacts of each individual atom pair in a residue are normalised between 0 and 1.0 by the total number of conformations and summed to the respective residue. Values greater than 1.0 are expected because the normalisation is only performed between pairs of contacts, meaning that each residue has, on average, the indicated number of contacts with all possible atom pairs that can form in the protein within the cut-off.

The ff14SB simulation shows strong contacts on the diagonal of the native contact map in Figure 3B, which are associated with the formation of α-helices. Residues with a high population of contacts on the map diagonal also have a large helix content in the DSSP map. In the non-native contact maps of ff14SB, the fragment exhibits spurious antidiagonal contacts identified as antiparallel β-sheets. Furthermore, the region 35-50 is rich in contacts, apparently promoting random coil conformations as suggested by the DSSP map. On the other hand, the native contacts of ff14IDPSFF are less abundant in the contact map, a fact consistent with the lack of helix content in the DSSP analysis. The diagonal and antidiagonal contacts of the non-native contact map are remarkable for the E-αSNAC^{IDP} fragment. Several amino acids show antidiagonal contacts with each other (residues 43-58, 58-75, 78-95). In addition, the contact map also shows parallel contacts between the amino acids 37-43 and 75-81 and, less frequently, between the amino acids 53-59 and 69-75. In both force fields, Tyr39, the unique tyrosine in the N-terminal domain, makes strong contacts with few residues. In the ff14SB simulation, Tyr39 interacts particularly with the amino acids S42, T44, V48, V52, V55, A65, and V66. This number of contacts is reduced in the ff14IDPSFF simulation, where Tyr39 interacts with S42, V66, V77, and L80 but allows Leu38 to make contacts with residues V74-V77. Many regions show overlapped contacts between several amino acids, which suggest

dynamic interconversion of distinct β-strand conformations. Thus, β-sheet formation is apparently not restricted to specific amino acids, but instead occurs in distinct and even shared regions of the E-αSNAC fragment.

7.2.3. Conformational Sampling in Principal Component Analysis

Principal Component Analysis (PCA) was performed on all conformations obtained from the simulations with the ff14SB and ff14IDPSFF force fields. Each trajectory was projected into first two eigenvalues (PC1 and PC2) and then transformed into Gibbs free energies as described in Methods. The PCA energy maps and the distributions of PC1 and PC2 are illustrated in Figure 4A. The conformations of the ff14SB and ff14IDPSFF simulations were clustered into 15 clusters using the hierarchical agglomerative clustering method. The centroids of the three most-populated clusters are shown in Figure 4B. The 2D-RMSD of the centroids of the five most-populated clusters is represented in Appendix E, Figure E1.

The PCA energy map of ff14SB distinguishes five low energy regions in the conformational sampling. The distributions of the PC1 and PC2 show several peaks, suggesting that certain conformations are preferred and, conversely, the access to some regions within the conformational space is apparently restricted. After clustering, the centroid conformations of the two most populated clusters have a high helix content in the structure, while the centroid of the third most populated cluster has a disordered structure. In contrast, E-αSNAC^{IDP} shows a broad conformational sampling as well as a more dispersed population within it, which is consistent with the wide PC1 and PC2 distributions. ff14IDPSFF samples a region defined by PC1 = [-100, -50], PC2 = [-50, 100] that does not appear in the ff14SB simulation. In fact, the centroid of the most populated cluster in the E-αSNAC^{IDP} is located in this region and exhibits a high β-strand content. On the contrary, the region delimited by PC1 = [25,50], PC2 = [0,50] is not sampled in ff14IDPSFF, where the centroids of the most populated clusters of ff14SB are located. The exploration of certain regions in the PCA space is then favoured or restricted depending on the force field used.

The centroid conformations of the three most populated clusters, representing >75% of the population, are illustrated in Figure 4B. The population of each cluster is given below. 84% of the conformations in the ff14SB simulations are found within the first three clusters. In contrast, the three most-populated clusters account for 76% of the

conformations in ff14IDPSFF, suggesting that the conformational sampling is more dispersed in the IDP-specific force field. Furthermore, the centroid conformations also show distinct secondary structure propensities in the simulations. E-αSNAC^{SB} preserves the helices of the native structure, with the exception of the cluster 3 centroid. Given the importance of the NAC domain for the fibril formation, this region (residues 61-95) is

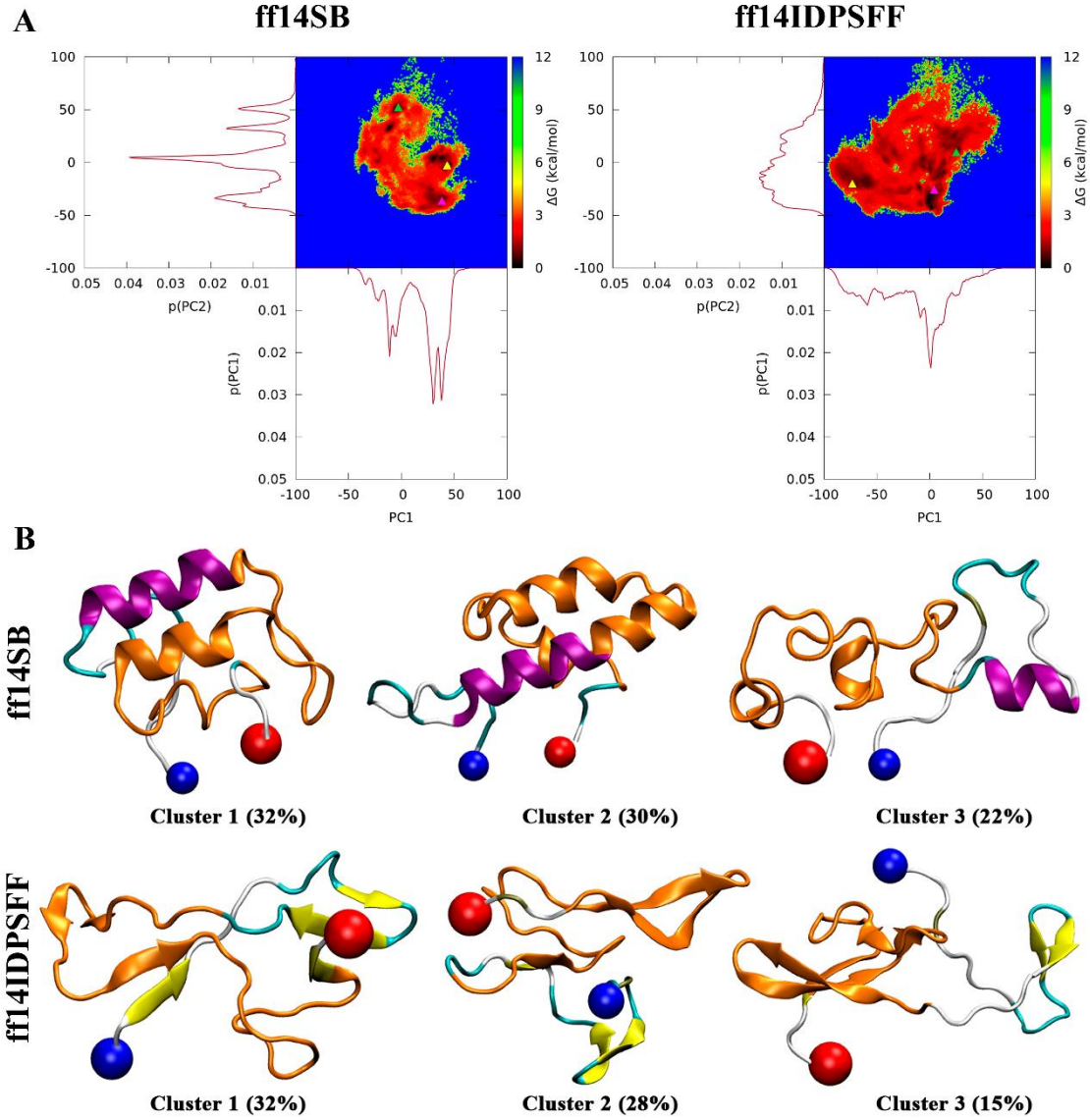


Figure 4. (A) Gibbs free energy maps in PCA space of the extended αS NAC domain fragment using the ff14SB (left) and ff14IDPSFF (right) force fields. The centroid conformations of clusters 1, 2 and 3 are marked with yellow, pink, and green triangles, respectively. The probability distributions of the PC1 and PC2 are also represented. (B) Centroid conformations of the three most populated clusters of the ff14SB (top) and ff14IDPSFF (bottom) simulations. The NAC domain (residues 61-95) is coloured in orange.

coloured in orange. Centroids 1 and 2 exhibit α-helix conformations, especially at amino acids 70-80 and 86-90. The 60-70 amino acids become loops or turns to allow the change of direction of the protein structure and in turn enable the interaction of the α-helix of the cleaved N-terminal and the NAC domain in order to form α-hairpins. Indeed, the cluster 2 centroid captures part of the helix formation (~30%) in the N-terminal domain that interacts with the rest of the E-αSNAC fragment.

The centroids of the ff14IDPSFF simulation do not show any helix formation in the structure and instead adopt β-strands, and even β-hairpins, formed within the NAC domain itself or together with amino acids of the N-terminal domain. The centroid of cluster 1 has β-strands (residues 1-4, 14-20, 40-42, 58-61) formed between the NAC and the N-terminal domain, and a short β-hairpin (residues 29-35) in the NAC domain itself. The centroids of clusters 2 and 3 contain larger β-hairpins involving several amino acids of the NAC domain. These β-strands, whether forming β-hairpins or not, appear in different amino acid regions as indicated by the DSSP method and the contact maps.

We also examined the goodness of clustering using the distance-to-centroid, Davies-Bouldin index (DBI), pseudo-F statistic (psF) and SSR/SST magnitudes listed in Table E1. The distance-to-centroid calculates the mean distance of the conformations captured within a cluster with respect to the centroid, the DBI measures whether the separation of the clusters and the classification of the conformations is correct, the pseudo-F statistics aims to capture the tightness of the clusters and the SSR/SST indicates the percentage of variance captured in the clustering. The magnitudes shown in the table indicate that the DBI and SSR/SST values are similar between the clustering of the ff14SB and ff14IDPSFF simulations, which means that the conformations obtained from the simulation are captured and classified with the same degree of goodness. On the other hand, the distance-to-centroid and the psF magnitudes show lower and higher values respectively in the ff14SB simulations, indicating a higher similarity between conformations within the narrower clusters. Conversely, although the clustering of the ff14IDPSFF is ranked with the same degree of goodness, the conformations within a cluster show greater heterogeneity and the clusters are broader in conformational space.

7.2.4. Simulated and Experimental NMR Chemical Shifts

Finally, the chemical shifts ($\delta^{15}N$, $\delta^{13}C$, $\delta^{13}C_\alpha$, $\delta^{13}C_\beta$) of the protein backbone were predicted using the SPARTA+ program[51]. The results of the ff14SB and ff14IDPSFF

simulations were compared with measured NMR data of full-length αS obtained from four data sets available in the Biological Magnetic Resonance Data Bank (BRMB). The linear regression of each set of measured and predicted chemical shifts for each atom and force field are illustrated in Figures E2-E9. The relationships between the predicted and experimental chemical shifts were analysed with the Pearson correlation coefficient, r-value, in Figure 5.

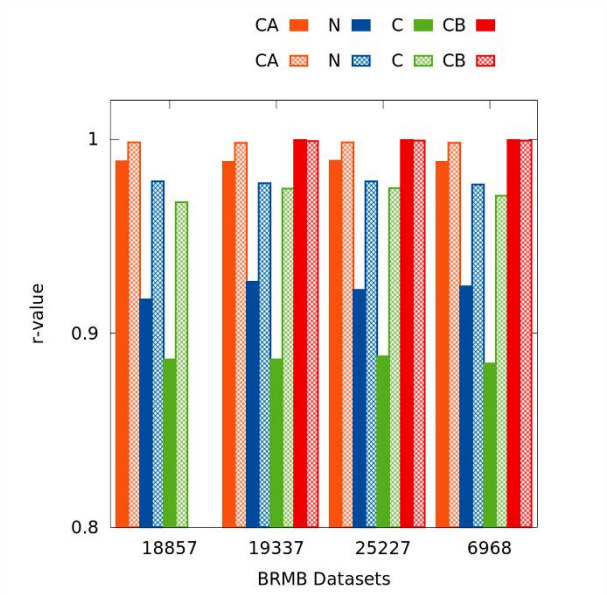


Figure 5. Pearson correlation coefficients between experimental and simulated chemical shifts using four BRMB data sets indicated on the x-axis (BRMB ID: 18857, 19337, 25227 and 6968) of the ff14SB (solid) and ff14IDPSFF (pattern) simulations. Orange, blue, green, and red boxes are C^{α}, N, C and C^{β} atoms, respectively. The C^{β} chemical shifts were not available in the BRMB 18857 data set.

The linear regression of the C_{α} and C_{β} atoms shows a high reproducibility of the chemical shifts in both force fields. Only the slope of the linear equation for the C_{α} atom shows a small deviation in contrast to the other data sets, which fit accurately. On the other hand, the linear regression of the chemical shifts of the N atom exhibits a larger slope and intercept for the ff14IDPSFF force field with values around 0.95 and 4.5-7.0, respectively. In contrast, the predicted chemical shifts of the N atom in ff14SB deviate significantly from the experimental ones, thus fitting its linear equation with slopes around 0.78. Finally, the chemical shifts of the C atoms show less reproducibility compared to other chemical shifts, independent of the force field. However, they are not negligible, with slopes of ~0.88 and ~0.85 for ff14SB and ff14IDPSFF, respectively. The

Pearson correlation coefficient is often used to assess whether two variables, in this case the simulated and measured NMR, are correlated. The r-values obtained from the simulations are generally well predicted, being >0.88, but show differences between the N and C atoms depending on the force field. Of note are the C_α and C_β atoms, which have r-values very close to 1.00 for all simulations regardless of the force field. The r-values of the chemical shifts of the N and C atoms are also remarkable with values around 0.97 in the ff14IDPSFF simulation. On the other hand, the r-values of these atoms in ff14SB are around 0.91 and 0.88, respectively, being lower with respect to the ff14IDPSFF force field. Then, the comparison between the measured NMR and the predicted chemical shifts reveals a better correlation for the IDP-specific force field, suggesting that the conformational ensembles obtained from the ff14IDPSFF simulations are more suitable for exploring the conformational properties of αS.

7.3. Discussion

In this chapter, the ff14SB and ff14IDPSFF simulations of the extended NAC domain fragment of αS show different conformational sampling, highlighting the preservation of the α-helix in the former while the IDP-specific force field simulations coexist between random coils and β-strands. The exploration of the PCA space reveals specific regions depending on the parameterisation of the protein. Apparently, the most populated clusters of the simulations are located in regions *exclusive* of each force field. Although RMSD and R_g have close average values, the ff14IDPSFF simulation shows a broader distribution and vaguely larger values. The analysis of the conformational properties indicates that the ff14SB force field achieves more restricted and compacted conformations, while the ff14IDPSFF force field explores more flexible conformations distant from the reference structure. Furthermore, the clustering indicators point out that the conformations obtained from ff14IDPSFF are clustered more widely and with greater structural difference, such an observation expected due to the structural diversity provided by the IDP-specific force field features.

According to the work of Yu et al.[18], which collected experimental data of full-length αS under different measurements conditions, the α-helix, β-sheet and turn contents are within 10%-48%, 0%-20% and 17%-41%, respectively. These results are in agreement with the ff14SB simulations, which have a ~40% of α-helix content and a negligible β-strand propensity (<1%). On the other hand, Chen et al.[17] reported the secondary structure content from the far UV CD spectrum of monomeric αS in solution:

6% α-helix, 34% β-strand, 18% turns, and 42% other conformations (including random coil, bends and other helices). These propensities are in good agreement with E-αSNAC^{IDP}. Despite the reproducibility of experimental data with simulations of E-αSNAC, it should be noted that the distribution of secondary structure in αS is heterogeneous. The E-αSNAC fragment removes the disordered C-terminal domain (residues 95-140) and part of the N-terminal domain (residues 1-37), where an extensive α-helix conformation is found in the structure of the membrane-bound αS. Neglecting these regions may lead to variations in the secondary structure content when compared to the full-length protein experiments. Furthermore, it must be considered that the secondary structure ratios determined by the experimental techniques are conditioned by the interactions that several monomers of the protein can establish in solution. In our simulations we use a single αS fragment, in which intermolecular interactions are absent and may lead to deviations from the experimental data.

Fortunately, computational studies focusing on fragments of αS have been performed [18,20,57] motivated by the prohibitive cost of performing long-time simulations of the full protein. Jain et al.[20] performed 18μs T-REMD simulations of the E-αSNAC fragment with CHARMM27* force field in which a higher content of α-helix (~62%) was determined at the expense of a lower content of random coil, turn and bend (~16%, ~11%, ~8%, respectively). These results are consistent with the ff14SB simulations despite the more distributed secondary structure content. Nevertheless, the contact map reported in the work of Jain et al. indicates that residues 70-85 have a 20-40% probability of forming antidiagonal contacts, which are related to β-strand conformations. The DSSP map of the ff14IDPSFF simulation shows this type of contacts and the β-sheet content is within the probability. On the other hand, Yu et al.[18] used a hybrid-resolution model to perform long-time simulations of αS and a short fragment (36-55 amino acids) parameterised with the CHARMM27 force field and CMAP correction. They found β-hairpin formation in the 36-55 amino acid fragment. They suggest that strong interactions between the C-terminal and this β-hairpin region reduce the access to β-strand formation and increase the required folding time. The E-αSNAC structure includes this β-hairpin region, which are N-terminal residues close to the NAC domain. The ff14IDPSFF simulation is in agreement with the results of Yu and co-workers as β-hairpin conformations are found in the centroids of clusters 2 and 3.

Furthermore, Chen et al. found β-strands in the NAC domain of some cluster centroids after performing trFRET-guided DMD simulations of full-length αS[17]. More importantly, this observation is supported by previous studies pointing out that the hydrophobic central amino acids of αS form β-strands as the first step of the oligomerisation and fibril formation.[18,58–60] On the other hand, a recent study by Balupuri et al.[27] suggested that the aggregation of αS could be achieved by an intermediate with an α-strand/sheet conformation found in the critical NAC region of amino acids 72-74. Indeed, we analysed the conformations of these three amino acids, but did not detect any α-strand content, irrespective of the force field. Interestingly, this critical NAC region in the ff14IDPSFF simulation shows a high turn content between the adjacent β-sheets, as well as a higher density of contacts in the 63-65 region compared to ff14SB. Similarly, other regions apparently crucial for αS fibrillation also show specific contacts that are only observed in ff14IDPSFF. Residues 74-81, which are part of the critical segment of the NAC domain for fibril formation[26], show significant contacts with amino acids 37-39. For these 74-81 amino acids, ff14SB shows only a weak, sparse, and heterogeneous contact density with residues 48-67. On the other hand, ff14IDPSFF shows another region of high contacts between residues 43-49 and 54-57, in which β-sheet propensity is observed in the DSSP map. In this case, amino acids 43-49 are part of one of the seven imperfect KTKEGV repeats reported in the NAC-domain, probably involved in the association of the protein with membrane lipids although it has also been hypothesised that they may play a role in the tetramerisation of αS[25,61]. However, this fragment is part of the compact hydrophobic β-sheet-rich structure in the fibrils. The map also shows a region of small antidiagonal contacts between amino acids 74-83 and 86-97, which probably adopt the β-hairpin observed in the cluster 2 representative. To ensure the convergence of these β-sheet structures observed in the clusters, we have illustrated the β-sheet propensity of E-αSNAC^{IDP} in Figure E10. The convergence plot shows that the β-sheet content is stable after 6 μs of simulation in the ff14IDPSFF force field. Finally, the review by Meade et al.[62] listed the most important mutations (A30P, E46K, H50Q, G51D, A53T/E) that affect the rate of αS fibrillation. We have examined whether these amino acids form relevant interactions, but unfortunately none of the force field show remarkable contacts.

One of the key points in understanding the results presented in this chapter is the conformational bias that exists in the force fields according to the studies reported so far.

Duong et al.[63] suggested that there are secondary structure preferences in the ff14SB and ff14IDPSFF force fields. In addition, several studies have pointed out that the ff14SB force field overestimates helix formation due to the use of globular protein structures in the parameterisation.[39,63,64] On the other hand, ff14IDPSFF is a relatively new force field that has incorporated the flexibility of IDPs by adding and optimising the CMAP correction terms in the potential energy function of ff14SB. After studying short peptides and the RNA-binding protein HIV-1 Rev, Duong et al.[63] concluded that the ff14IDPSFF force field promotes random coil conformations and disordered secondary structures consistent with experiments. An example of the potential of this force field are the works of Song et al.[34] and Dan et al.[65], in which the simulation of the all-atom microtubule-associated Tau protein with ff14IDPSFF was able to capture β-sheet conformations that were also observed in experiments. In a comparative study between force fields, β-hairpin was found in conformational ensembles of β-amyloid proteins using IDP-specific force fields.[66] A similar trend in β-strand content is observed in E-αSNAC^{IDP} after performing conventional 6 μs-length simulations with ff14IDPSFF. In addition to the conformational bias in the force field, Yu et al.[18] demonstrated with the extensive hybrid-model PACE simulations that C-terminal interactions affect in the β-hairpin formation of the 38-53 region. Therefore, E-αSNAC^{IDP} has two factors that facilitate random coil or β-strand conformations, i.e. (i) the promotion of random coil conformations by the ff14IDPSFF force field and (ii) the absence of the C-terminal domain and its interactions with the NAC domain.

Finally, the correlations between predicted and measured NMR chemical shifts indicate that ff14IDPSFF reproduces more accurately the experiments of αS in solution, especially for ^{13}C and ^{15}N atoms. Indeed, it is consistent with previous works reporting that a promising feature of the ff14IDPSFF force field is the improved prediction of chemical shifts compared to ff14SB.[34,63,65,66] Our results with the IDP-specific force field from the conformational analysis of this and previous studies and the contrast of the NMR chemical shifts point to a very successful exploration of the αS conformational space within affordable MD simulation times lengths.

7.4. Conclusions

The αS protein adopts a wide range of conformations during time evolution due to its intrinsic disorder typical of IDPs. Because of the difficulty in experimentally characterising the conformations of these highly flexible proteins, atomistic simulations

come into play, especially those that have attempted to incorporate the features of IDPs in recent years. In this study, we selected an αS fragment (residues 35-97) to reproduce the conformational space through 6 μs simulations using the classical ff14SB AMBER force field and the IDP-specific ff14IDPSFF force field developed by Song et al. [34]. The results indicate that the classical force field preserves the conformations typically found in the micelle-bound α-synuclein structure, notable for the high presence of α-helices. On the other hand, the ff14IDPSFF force field provides conformational ensembles dominated by structural disorder and low formation of β-strands, which are apparently not easily accessible in conventional force fields. To validate these simulations, we performed the linear regression and reported the Pearson's correlation coefficients between the predicted and the experimentally measured chemical shifts, demonstrating that ff14IDPSFF reproduces NMR data more accurately. Therefore, the results presented in this chapter suggest that ff14IDPSFF is reliable for exploring αS conformations not normally found in well-established force fields, and provide additional evidence to the body of work pointing to β-sheet formation as an intermediate state to fibril formation in the αS protein. An in-depth study of αS with IDP-specific force fields could certainly shed some light on the mechanism of protein fibril formation, and subsequently contribute to a more complete picture of the pathogenesis of neurodegenerative disorders derived from synucleinopathies.

7.5. Bibliography

1. Goedert, M. Alpha-synuclein and neurodegenerative diseases. *Nat Rev Neurosci* **2**, 492–501 (2001).
2. Hirsch, L., Jette, N., Frolkis, A., Steeves, T. & Pringsheim, T. The Incidence of Parkinson's Disease: A Systematic Review and Meta-Analysis. *Neuroepidemiology* **46**, 292–300 (2016).
3. Spillantini, M. G. *et al.* α-Synuclein in Lewy bodies. *Nature* **388**, 839–840 (1997).
4. Dawson, T. M. & Dawson, V. L. Molecular Pathways of Neurodegeneration in Parkinson's Disease. *Science (1979)* **302**, 819–822 (2003).
5. Breydo, L., Wu, J. W. & Uversky, V. N. α-Synuclein misfolding and Parkinson's disease. *Biochim Biophys Acta Mol Basis Dis* **1822**, 261–285 (2012).
6. Ulmer, T. S., Bax, A., Cole, N. B. & Nussbaum, R. L. Structure and Dynamics of Micelle-bound Human α-Synuclein. *J Biol Chem* **280**, 9595–9603 (2005).
7. Rao, J. N., Jao, C. C., Hegde, B. G., Langen, R. & Ulmer, T. S. A Combinatorial NMR and EPR Approach for Evaluating the Structural Ensemble of Partially Folded Proteins. *J Am Chem Soc* **132**, 8657–8668 (2010).
8. Tuttle, M. D. *et al.* Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. *Nat Struc Mol Bio* **23**, 409–415 (2016).

9. Uversky, V. N., Li, J. & Fink, A. L. Evidence for a Partially Folded Intermediate in α-Synuclein Fibril Formation. *J Biol Chem* **276**, 10737–10744 (2001).

10. Matthew M. Dedmon, Kresten Lindorff-Larsen, John Christodoulou, Michele Vendruscolo, and & Dobson*, C. M. Mapping Long-Range Interactions in α-Synuclein using Spin-Label NMR and Ensemble Molecular Dynamics Simulations. *J Am Chem Soc* **127**, 476–477 (2004).

11. Sung, Y.-H. & Eliezer, D. Residual structure, backbone dynamics, and interactions within the synuclein family. *J Mol Biol* **372**, 689–707 (2007).

12. Bertoncini, C. W. *et al.* Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proc Natl Acad Sci USA* **102**, 1430–1435 (2005).

13. Schwalbe, M. *et al.* Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and a-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* **22**, 238–249 (2014).

14. Wu, K.-P., Weinstock, D. S., Narayanan, C., Levy, R. M. & Baum, J. Structural Reorganization of α-Synuclein at Low pH Observed by NMR and REMD Simulations. *J Mol Biol* **391**, 784–796 (2009).

15. Ilie, I. M., Nayar, D., den Otter, W. K., van der Vegt, N. F. A. & Briels, W. J. Intrinsic Conformational Preferences and Interactions in α-Synuclein Fibrils: Insights from Molecular Dynamics Simulations. *J Chem Theory Comput* **14**, 3298–3310 (2018).

16. Zhang, Y. *et al.* High-speed atomic force microscopy reveals structural dynamics of α-synuclein monomers and dimers. *J Chem Phys* **148**, 123322 (2018).

17. Chen, J. *et al.* The structural heterogeneity of α-synuclein is governed by several distinct subpopulations with interconversion times slower than milliseconds. *Structure* **29**, 1048-1064.e6 (2021).

18. Yu, H., Han, W., Ma, W. & Schulten, K. Transient β-hairpin formation in α-synuclein monomer revealed by coarse-grained molecular dynamics simulation. *J Chem Phys* **143**, 243142 (2015).

19. Ramis, R. *et al.* A Coarse-Grained Molecular Dynamics Approach to the Study of the Intrinsically Disordered Protein α-Synuclein. *J Chem Inf Model* **59**, 1458–1471 (2019).

20. Jain, K., Ghribi, O. & Delhommelle, J. Folding Free-Energy Landscape of α-Synuclein (35–97) Via Replica Exchange Molecular Dynamics. *J Chem Inf Model* **61**, 432–443 (2021).

21. Allison, J. R., Rivers, R. C., Christodoulou, J. C., Vendruscolo, M. & Dobson, C. M. A Relationship between the Transient Structure in the Monomeric State and the Aggregation Propensities of α-Synuclein and β-Synuclein. *Biochemistry* **53**, 28 (2014).

22. Ahmed, M. C. *et al.* Refinement of α-Synuclein Ensembles Against SAXS Data: Comparison of Force Fields and Methods. *Front Mol Biosci* **8**, 216 (2021).

23. Maroteaux, L., Campanelli, J. T. & Scheller, R. H. Synuclein: A Neuron-Specific Protein Localized to the Nucleus and Presynaptic Nerve Terminal. *J Neurosci* **8**, 2804–2815 (1988).

24. Bendor, J. T., Logan, T. P. & Edwards, R. H. The Function of α-Synuclein. *Neuron* **79**, 1044–1066 (2013).

25. Bussell, R. & Eliezer, D. A Structural and Functional Role for 11-mer Repeats in α-Synuclein and Other Exchangeable Lipid Binding Proteins. *J Mol Biol* **329**, 763–778 (2003).

26. Giasson, B. I., Murray, I. V. J., Trojanowski, J. Q. & Lee, V. M.-Y. A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of α-Synuclein Is Essential for Filament Assembly. *J Biol Chem* **276**, 2380–2386 (2001).

27. Balupuri, A., Choi, K.-E. & Kang, N. S. Computational insights into the role of α-strand/sheet in aggregation of α-synuclein. *Sci Rep* **9**, 59 (2019).

28. Henriques, J. O., Cragnell, C. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput* **11**, 3420–3431 (2015).

29. Kang, W., Jiang, F. & Wu, Y. How to strike a conformational balance in protein force fields for molecular dynamics simulations? *WIREs Comput Mol Sci* **12**, e1578 (2022).

30. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys J* **100**, L47–L49 (2011).

31. Yang, S., Liu, H., Zhang, Y., Lu, H. & Chen, H. Residue-Specific Force Field Improving the Sample of Intrinsically Disordered Proteins and Folded Proteins. *J Chem Inf Model* **59**, 4793–4805 (2019).

32. Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J Chem Theory Comput* **11**, 3499–3509 (2015).

33. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* **34**, 2135–2145 (2013).

34. Song, D., Luo, R. & Chen, H.-F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **57**, 1166–1178 (2017).

35. Song, D., Liu, H., Luo, R. & Chen, H. F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J Chem Inf Model* **60**, 2257–2267 (2020).

36. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Nat Acad Sci USA* **115**, E4758–E4766 (2018).

37. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J Chem Theory Comput* **10**, 5113–5124 (2014).

38. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73 (2017).

39. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

40. Case, D. A. *et al.* Amber 2018. *University of California, San Francisco* (2018).

41. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

42. Hopkins, C. W., le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **11**, 1864–1874 (2015).

43. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicist*. (1999).

44. Perez, J. J., Santos Tomas, M. & Rubio-Martinez, J. Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory. *J Chem Inf Model* **56**, 1950–1962 (2016).

45. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

46. Allen, M. P. & Tildesley, D. J. *Computer Simulation of Liquids*. (Oxford University Press, 2017).

47. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **14**, 33–38 (1996).

48. Williams, T. *et al.* Gnuplot 4.6: an interactive plotting program. *Software Manual* 238 (2012).

49. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

50. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

51. Shen, Y. & Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* **48**, 13–22 (2010).

52. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

53. Vandova, Gergana; Tamiola, Kamil; Oktaviani, Nur; Mulder, Frans. Backbone 1H, 13C, and 15N chemical shift assignments for alpha-synuclein at different pH and temperature (2013).

54. Kang, L., Janowska, M. K., Moriarty, G. M. & Baum, J. Mechanistic Insight into the Relationship between N-Terminal Acetylation of α-Synuclein and Fibril Formation Rates by NMR and Fluorescence. *PLoS One* **8**, e75018 (2013).

55. Porcari, R. *et al.* The H50Q Mutation Induces a 10-fold Decrease in the Solubility of α-Synuclein*. *J Biol Chem* **290**, 2395–2404 (2015).

56. Bermel, W. *et al.* Protonless NMR Experiments for Sequence-Specific Assignment of Backbone Nuclei in Unfolded Proteins. *J Am Chem Soc* **128**, 3918–3919 (2006).

57. Pujols, J. *et al.* Small molecule inhibits α-synuclein aggregation, disrupts amyloid fibrils, and prevents degeneration of dopaminergic neurons. *Proc Nat Acad Sci USA* **115**, 10481–10486 (2018).

58. Brodie, N. I., Popov, K. I., Petrotchenko, E. v., Dokholyan, N. v. & Borchers, C. H. Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. *PLoS Comput Biol* **15**, e1006859 (2019).

59. Graen, T., Klement, R., Grupi, A., Haas, E. & Grubmüller, H. Transient Secondary and Tertiary Structure Formation Kinetics in the Intrinsically Disordered State of *α* -Synuclein from Atomistic Simulations. *ChemPhysChem* **19**, 2507–2511 (2018).

60. Healey, M. A., Woodside, M. T. & Tuszynski, J. A. Phase transitions and structure analysis in wild-type, A30P, E46K, and A53T mutants of α-synuclein. *Eur Biophys J* **45**, 355–364 (2016).

61. Dettmer, U., Newman, A. J., von Saucken, V. E., Bartels, T. & Selkoe, D. KTKEGV repeat motifs are key mediators of normal α-synuclein tetramerization: Their mutation causes excess monomers and neurotoxicity. *Proc Nat Acad Sci USA* **112**, 9596–9601 (2015).

62. Meade, R. M., Fairlie, D. P. & Mason, J. M. Alpha-synuclein structure and Parkinson's disease – lessons and emerging principles. *Mol Neurodegener* **14**, 29 (2019).

63. Duong, V. T., Zihao, C., Thapa, M. T. & Luo, R. Computational Studies of Intrinsically Disordered Proteins. *J Phys Chem B* **122**, 10455–10469 (2018).

64. Reid, L. M. *et al.* How well does molecular simulation reproduce environment-specific conformations of the intrinsically disordered peptides PLP, TP2 and ONEG? *Chem Sci* **13**, 1957–1971 (2022).

65. Dan, A. & Chen, H.-F. Secondary structures transition of tau protein with intrinsically disordered proteins specific force field. *Chem Biol Drug Des* **93**, 242–253 (2019).

66. Rahman, M. U., Rehman, A. U., Liu, H. & Chen, H.-F. Comparison and Evaluation of Force Fields for Intrinsically Disordered Proteins. *J Chem Inf Model* **60**, 4912–4923 (2020).

Chapter 8

Outlook: The Synuclein Protein Family

The environmental factors within a cell that can affect proteins are many and varied. In fact, in addition to the presence of other proteins or peptides that may interact with our target protein, there are also a number of physicochemical properties provided by the cellular environment that can alter the biological functions of a protein[1]. This is particularly true in the case of IDPs since the intrinsic disorder and disorder-to-order transitions confer them the ability to respond rapidly to the external stimuli, which is of paramount relevance to their regulatory functions. Therefore, the cell composition, such as the presence of inorganic ions, pH, metabolites or even the electrochemical environment resulting from the presence of other molecules or ions, can greatly alter the functionality of the proteins. For example, regarding the synuclein protein family, several studies have shown that the presence of metal ions promotes the onset of neurodegenerative diseases through the protein-metal interactions of many amyloidogenic IDPs such as α-synuclein, β-amyloid or Tau[2]. On the other hand, pH is also crucial in the modulation of the protein conformations through the charge-structure coupling, since the environmental pH affects the protonation state of the ionisable amino acid side chains, as we have already mentioned in this thesis. This property is of great interest in IDPs due to the abundance of ionisable amino acids in their sequence and the wide pH range within the cellular compartments[3]. On the other hand, more complex factors may be also involved within the cell, such as viscosity and molecular crowding, which define the steric hindrance and the excluded volume effect (which is closely related to the effective concentration) in the cytosol and therefore have an influence on the compaction and folding of globular proteins and IDPs.

The synuclein protein family is not exempt from the influence of these environmental factors on cells. Indeed, the metal ion binding capacity of αS has already been mentioned. In addition, other factors, such as salt concentration or pH, have also been investigated. In particular, solvent pH at mildly acidic conditions has been shown to accelerate αS aggregation and to alter the fibril structure with different fibril typologies governed by the kinetic control of the reaction[4–6]. Interestingly, β-synuclein has also been found to

fibrillate at mildly acidic conditions, although in principle it exists as a soluble monomer under physiological conditions. Therefore, the work of Moriarty et al.[7] presents the pH as an on/off-switch for βS fibrillation via the interaction of certain glutamic acids. Furthermore, the study of αS/βS chimeras shows that intramolecular interactions between domains are decisive for stabilising or inhibiting the fibril formation. Therefore, motivated by this work and other in-silico simulations on the ability of pH or protonation states to modify the conformational space of αS, the future research of this thesis will focus on an in-depth study of the conformational space of αS and βS by all-atom CpHMD simulations[8] including novel IDP-specific force fields and water models.

Here we present the first results of the 2 μs-length exploration of αS and βS using the CpHMD method and the ff14IDPSFF force field[9] with TIP3P water molecules at physiological (7.3) and mildly acidic (5.8) pH conditions.

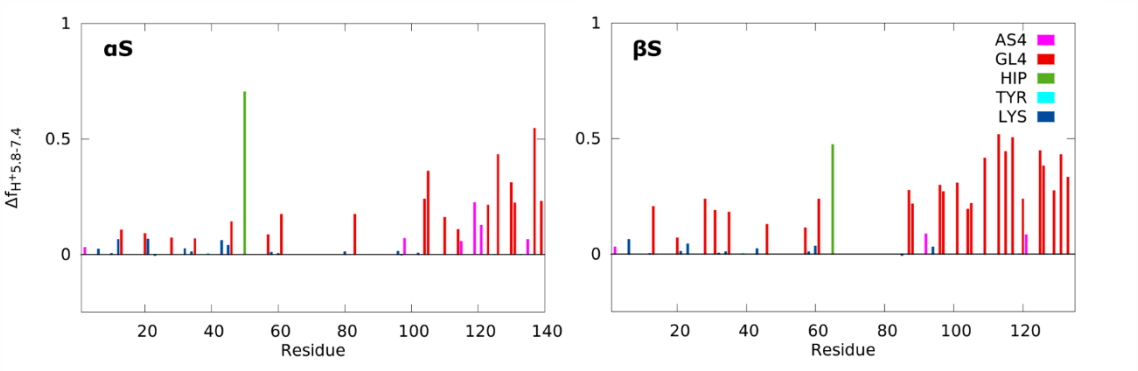


Figure 1. Protonated state fractions shifts of αS (left) and βS (right) proteins from solvent pH 7.4 to 5.8.

The shifts of the protonated state fractions of αS and βS between solvent pH 5.8 and 7.4 are illustrated in Figure 1. In general, βS shows a larger change in the net protein charge ($|\Delta q|$ ~9.0e) compared to αS ($|\Delta q|$ ~5.5e). Furthermore, the charge variation is mainly in the C-terminal domain, which is known for its inhibitory effect in preventing fibril formation. Regarding the different amino acids, on the one hand, the histidine in both proteins shows a high sensitivity to the change of the solvent pH, because the intrinsic pKa of the side chain of the imidazole ring is 6.8 in ideal conditions. However, the shift of the protonated state fraction in αS is more pronounced in comparison to βS. On the other hand, as pointed out in the study by Moriarty and co-workers, glutamic acids are predominant among the amino acids that undergo the changes in protonation state fraction, suggesting that they play a key role in understanding the fibrillation capacity of

βS and the observed topologies and aggregation rate of αS. Furthermore, aspartic acids are not prominent in βS, whereas αS shows some mild shifts in the C-terminal tail. Other amino acids do not appear to be relevant in these simulations at constant pH.

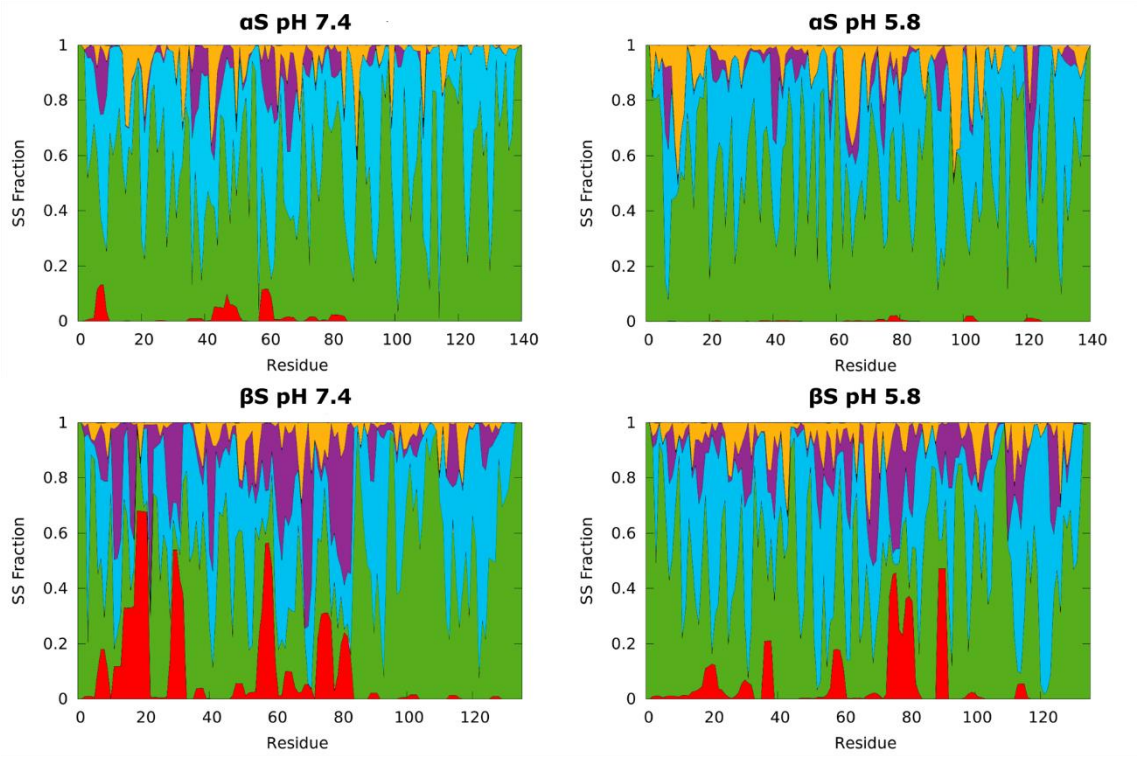


Figure 2. Secondary structure propensity fractions of αS and βS at pH 5.8 and 7.4. α-, 3_{10}- and π-helix are grouped in red, β-sheets, β-bulges and isolated β-strands are grouped in orange, and turns, bends and random coil conformations are coloured in purple, blue and green, respectively.

The secondary structure propensity fractions (fpSS) of the four simulations, αS and βS at pH 5.8 and 7.4, are illustrated in Figure 2. In αS, the α-helix structure found in the initial membrane-bound structure is completely broken. This was observed earlier in Chapter 7 with the extended NAC-domain αS fragment which significantly lost the helix content with the ff14IDPSFF force field. While remnants of the helix content remain at pH 7.4 conditions, this low helix content disappears at pH 5.8 conditions. On the other hand, the random coil and bend conformations are mainly promoted, and β-strand structures appear sporadically with a content generally below 20%. For βS, the formation of helices and turns is more significant, especially in the NAC and N-terminal domains at pH 7.4. In contrast, βS reduces the content of helices and turns at pH 5.8 and, in turn, these structures are more distributed along the protein. It is therefore possible that the helix formation in

these two domains is important for the prevention of fibrillation. On the other hand, β-strand formation also occurs in βS with a low content as in αS. Compared to αS, βS contains a lower content of bends and random coil conformations although they are still predominate in the conformational sampling.

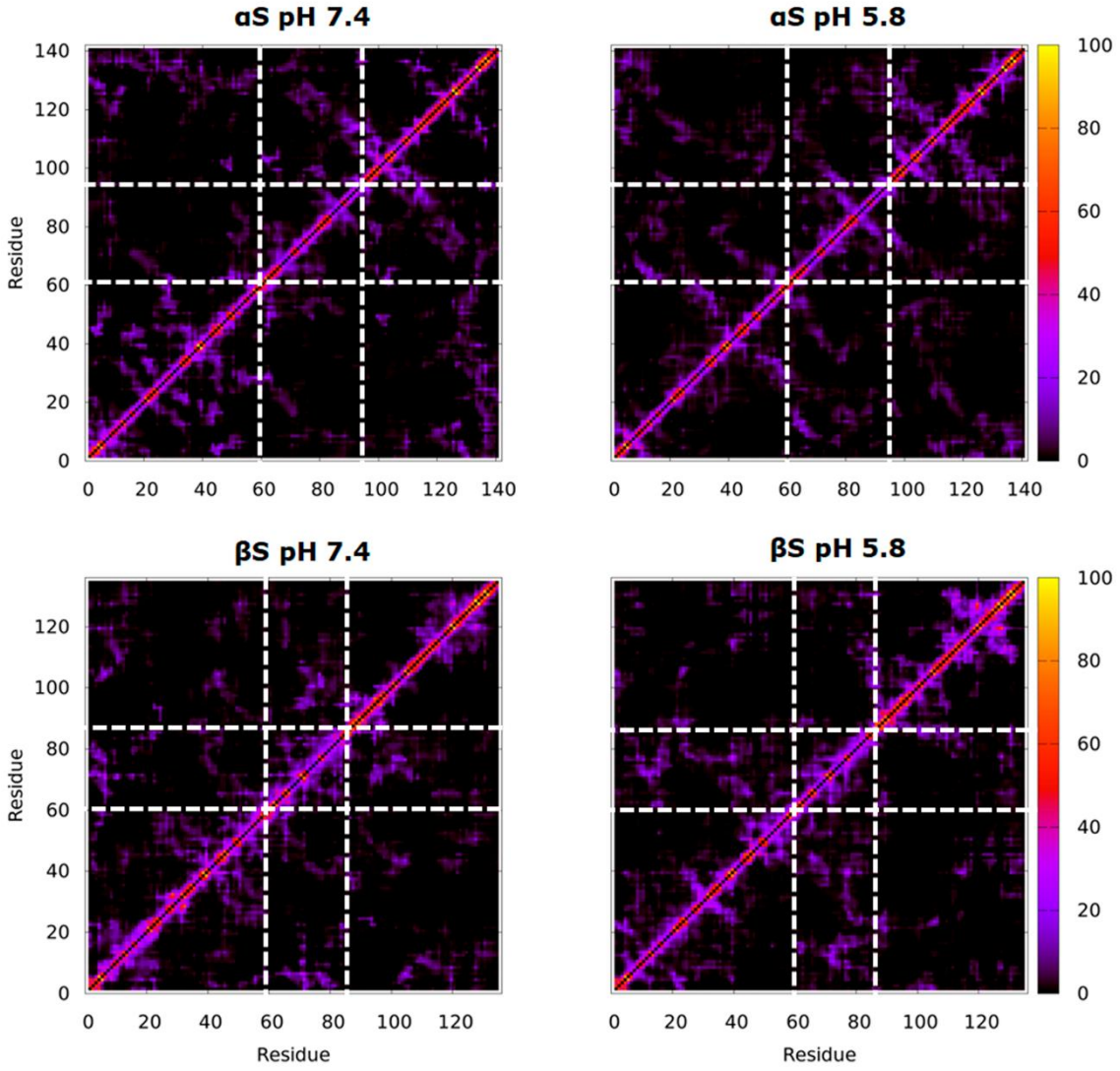


Figure 3. Intramolecular contact maps of αS and βS at pH 5.8 and 7.4. The N-terminal, NAC and C-terminal domains are separated by white dashed lines. The colour indicates the frequency of contacts during the trajectory.

Finally, we present the map of the intramolecular contacts during the trajectories in Figure 3. On the one hand, αS at pH 7.4 shows a higher number of intramolecular contacts within the N-terminal domain, and some minor ones between the C-terminal tail and the N-terminal domain as well as between the last amino acids of the NAC domain and the first residues of the C-terminal domain. In contrast, the intramolecular contacts

disappear significantly at pH 5.8 conditions, with only some antidiagonal contacts typical of β-strand in some sets of amino acids at the end of the N-terminal and NAC domains, as well as in the middle of the C-terminal domain. The disappearance of these contacts could explain the absence of structure and the increased presence of random coil in the fpSS plot, since at pH 5.8 conditions the frequency of the contacts is reduced. On the other hand, at pH 7.4, βS establishes significant contacts within the NAC domain and a higher number of contacts on the diagonal of the N-terminal and NAC domains. Other minor contacts are present between these two domains. In contrast, at pH 5.8 the diagonal contacts are drastically reduced and some antidiagonal contacts appear in the C-terminal domain. There are also sporadic contacts between the NAC and N-terminal domains, which may indicate an interaction between these domains to promote fibrillation after helical breaking present at mildly acidic conditions.

Therefore, the results of this preliminary study suggest changes in the intramolecular interactions and the secondary structure propensities that are dependent on the pH and the protonation state of the glutamic acids present in the αS and βS chains, although aspartic acids and histidine may also be important, especially for the former. On this basis, our perspectives for further research on the synuclein protein family are summarised below:

(i) A study of the titration curves of ionisable amino acids within the αS and βS structures from pH-REMD simulations and the estimation of the pKa by means of complexation isotherms. In this way, we can determine which amino acids are most sensitive to the pH change and reduce the number of amino acids included in the CpHMD simulation.

(ii) Explore the conformational space of the αS and βS proteins using IDP-specific force fields (ff14IDPSFF, a99SB-disp or ff19SB), water models (TIP4P-D, OPC or others depending on the force fields) and the CpHMD method at different pH conditions. In addition, include CG simulations to facilitate further sampling of the conformational space to develop multi-seed sampling strategies.

(iii) Extend the study for γ-synuclein and oligomers of αS and βS, and, if possible, contrast all the simulations with NMR or SAXS experimental techniques to provide reliability in the combination of force field, water models and constant pH method carried out in the simulations.

Overall, we hope to gain a deeper understanding of the mechanism of fibrillation of the synuclein protein family and the solvent pH dependence through these simulations at constant pH. Ultimately, based on the success of these studies, we would like to contribute to the treatment of synucleopathies and, in the future, support this ongoing research with drug design projects to advance the therapies for neurodegenerative diseases.

Bibliography

1. Theillet, F.-X. *et al.* Physicochemical Properties of Cells and Their Effects on Intrinsically Disordered Proteins (IDPs). *Chem Rev* **114**, 6661–6714 (2014).
2. Breydo, L. & Uversky, V. N. Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases. *Metallomics* **3**, 1163 (2011).
3. Casey, J. R., Grinstein, S. & Orlowski, J. Sensors and regulators of intracellular pH. *Nat Rev Mol Cell Biol* **11**, 50–61 (2010).
4. Tycko, R. Physical and structural basis for polymorphism in amyloid fibrils. *Protein Science* **23**, 1528–1539 (2014).
5. McGlinchey, R. P., Jiang, Z. & Lee, J. C. Molecular Origin of pH-Dependent Fibril Formation of a Functional Amyloid. *ChemBioChem* **15**, 1569–1572 (2014).
6. Buell, A. K. *et al.* Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. *Proc Nat Acad Sci USA* **111**, 7671–7676 (2014).
7. Moriarty, G. M. *et al.* A pH-dependent switch promotes β-synuclein fibril formation via glutamate residues. *J Biol Chem* **292**, 16368–16379 (2017).
8. Swails, J. M., York, D. M. & Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J Chem Theory Comput* **10**, 1341–1352 (2014).
9. Song, D., Luo, R. & Chen, H.-F. The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. *J Chem Inf Model* **57**, 1166–1178 (2017).

Chapter 9

Conclusions

The research conducted in this thesis is presented as a modest contribution to a deeper understanding of the pH-dependent charge-structure coupling of proteins, in particular for the so-called intrinsically disordered proteins (IDPs). Through the investigations of Chapter 3 to Chapter 8, which range from the study of the effect of pH on short peptides to the sampling of the conformational space and the protonation states of IDPs, this final chapter summarises and outlines the concluding remarks of this doctoral thesis.

Several issues of paramount importance for incorporating the effect of pH and the intrinsic disorder of IDPs into MD simulations have been addressed in this investigation. First, we have considered the available techniques within the Constant pH Molecular Dynamics (CpHMD) approach, selecting the CpHMD method with discrete protonation states for a physically meaningful representation of the peptides and proteins along the entire trajectory. Through the conformational analysis of the polyaspartic acid peptide in Chapter 3, we identified a few inconsistencies in the CpHMD method, which then led us to investigate in more detail the limitations and capabilities of this approach implemented in AMBER using titratable amino acid tripeptides described in Chapters 4 and 5.

- The implementation of the CpHMD method with discrete protonation states in the AMBER software package induces deviations in the reproduction of the conformational space and structural properties in the deprotonated state of the tripeptides compared to the conventional MD (CMD) simulations. These observations can be explained by the fact that the set of partial charges of the reference residue (i.e., the protonated amino acid) remains fixed throughout the simulation, leading to a misassignment of the partial charges on the backbone atoms of the amino acids in the deprotonated state. Other contributions may also be involved in the observed deviations such as the description of the dihedral angles.
- The CpHMD simulations of the tripeptides with ionisable amino acids with a single protonation site successfully reproduce the conformational profiles of the CMD

simulations in the protonated form. On the other hand, the tripeptides with titratable amino acids possessing multiple protonation states show mild deviations in the conformational space, suggesting that the sampling of the protonation states promotes distinct conformational preferences within the Ramachandran map. Therefore, the inclusion of pH has a potential benefit on the conformational sampling of larger biomolecules derived from proper protonation state sampling.

- By means of the pH-REMD method, we have shown the coupling charge-structure in a wide range of pHs around the pK_D of the polyaspartic acid peptide, proving that the conformations and titratable amino acids are sensitive to the solvent pH, even at conditions apparently far from the effective pK_D. At this point, we have highlighted the importance of estimating the titration curve and the pk_{app} distribution of the individual amino acids, as well as those for the entire peptide, in the pH range of the pH-REMD simulations using the Hill/Langmuir-Freundlich (Hill/LF) and Frumkin complexation isotherms. On the one hand, the Hill/LF isotherm gives the effective $pk_{D,i}$ and pK_D of the titratable amino acids and the peptide, respectively, when the protonation/deprotonation fraction is 50%, but the isotherm does not fit to the simulated values at the limits of the titration curves. On the other hand, the Frumkin isotherm directly provides the intrinsic $pk_{a,i}$ and pK_a for the titratable amino acids and peptide, respectively, and a physically meaningful δ-parameter of the electrostatic interactions of the molecular system. Interestingly, the $pk_{a,i}$ values obtained from the isotherms differ from the $pk_{a,i}$ = 4.0 set by the default in the AMBER program, and a relationship is observed between the position of the individual titratable amino acids within the chain and the $pk_{a,i}$ values. Therefore, it is recommended to study the $pk_{app,i}$ distribution of the amino acids before running simulations at constant pH.
- We have tested the limitations of the CpHMD method implemented in AMBER on alanine decapeptides with two aspartic acids in different positions and conclude that, if the ionisable amino acids are sufficiently distant, the deviations in the conformational space are negligible. This raises the possibility of strategies that minimise the effect of the deprotonated residues in simulations at constant pH. However, the extent of such deviations in more complex systems (i.e., including diverse electrochemical environments with other ionisable amino acids, solvents with different dielectric constants, inorganic ions, ionic strengths, etc.) also requires further study.

On the other hand, we have focused on current force fields in order to incorporate the inherent flexibility of IDPs into MD simulations since both conformational and protonation state sampling need to be properly addressed. Therefore, through an extensive study of the histatin-5 model IDP with different force fields, water models, simulation methods and sampling techniques, in Chapter 6 we have emphasised the importance of a good parameterisation to reproduce the NMR and SAXS experimental observables, and we have included the dynamic change of protonation states in these simulations.

- The all-atom simulation with the ff14IDPSFF force field and the TIP4P-D water model has demonstrated successful performance in reproducing the SAXS radius of gyration and NMR chemical shifts. Both the force field and the water model are known to improve the description of IDPs by incorporating the CMAP correction energy and enhancing the water-protein dispersion forces, respectively.
- Furthermore, when using a coarse-grained resolution in the histatin-5 simulations, a multi-seed sampling strategy is required to obtain good results with the SIRAH force fields and the WAT4 water model. Based on this study, we are currently planning to investigate protocols for novel sampling strategies based on transitions from a CG resolution with a reduced number of beads to all-atom resolutions in a stepwise manner, i.e., using intermediate CG models with a higher level of description.
- The strong charge-structure coupling is demonstrated in the histatin-5 simulations, as the protonation fractions vary significantly depending on the force field, the water model or the resolution of the molecular model used.

Finally, in light of the conclusions of the previous chapters, in Chapter 7 and Chapter 8 we have begun a study of the synuclein protein family, which are involved in several biological functions that are currently still being elucidated. However, the fibrillogenesis of these proteins is associated with certain physiological pathologies such as α-synuclein aggregation in Lewis bodies in Parkinson's disease. Therefore, the ultimate goal of this research is to gain a deeper understanding of the mechanism of fibril formation and the pH dependence, as reported in *in vivo* studies.

- Extensive sampling of the extended NAC domain fragment of α-synuclein reveals significant differences between the ff14SB and ff14IDPSFF force fields. The ff14SB force field, which is well established in biomolecular simulations, maintains the formation of helices in the trajectory. In contrast, the IDP-specific force field differs from its predecessor by a high formation of random coil conformations and a

remarkable β-strand content, which are found in β-sheet structures within the Greek-key topology of α-synuclein fibrils. Thus, the ff14IDPSFF force field suggests that these disordered structures with sporadic transitions to β-strand formation could be the intermediate states of the fibrillation process.

- On the other hand, the first results of simulations at constant pH with the ff14IDPSFF force field on α-synuclein and β-synuclein show certain differences between the two proteins, such as an almost complete helix unfolding in α-synuclein or a greater sensitivity to solvent pH in β-synuclein due to the protonation of the amino acids of the C-terminal domain. Given the pH-dependent fibrillation response of these proteins, an in-depth study including all the factors discussed in the present thesis would be very enriching and hopefully will be performed in the near future.

List of Publications

Articles

- Privat, C., Madurga, S., Mas, M. & Rubio-Martínez, J. On the Use of the Discrete Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides. *Polymers* **13**, 99 (2021).
- Privat, C., Madurga, S., Mas, M. & Rubio-Martínez, J. Unravelling Constant pH Molecular Dynamics in Oligopeptides with Explicit Solvation Model. *Polymers* **13**, 3311 (2021).
- Privat, C., Madurga, S., Mas, M. & Rubio-Martínez, J. Molecular Dynamics Simulations of an α-Synuclein NAC Domain Fragment with a ff14IDPSFF IDP-Specific Force Field Suggest β-Sheet Intermediate States of Fibrillation. *Physical Chemistry Chemical Physics* **24**, 18841–18853 (2022).

Appendix A

Supporting Information to "Exploring the Polyaspartic Acid Conformations with Constant pH Simulations and Prediction of pKa through Complexation Isotherms"

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Explicit Solvent	ASP_{10}	33%	17%	14%	8%	6%
	ASH_{10}	20%	12%	10%	9%	8%
	$AS4_{10}^{pH=1}$	16%	13%	12%	12%	11%
	$AS4_{10}^{pH=10}$	44%	18%	15%	4%	4%
Implicit Solvent	ASP_{10}	32%	14%	13%	6%	5%
	ASH_{10}	26%	17%	8%	8%	8%
	$AS4_{10}^{pH=1}$	80%	4%	3%	3%	2%
	$AS4_{10}^{pH=10}$	88%	3%	2%	2%	1%

Table A1. Population fractions of the five most populated clusters of ASP_{10}, ASH_{10}, $AS4_{10}^{pH=1}$ and $AS4_{10}^{pH=10}$ in implicit and explicit solvation models.

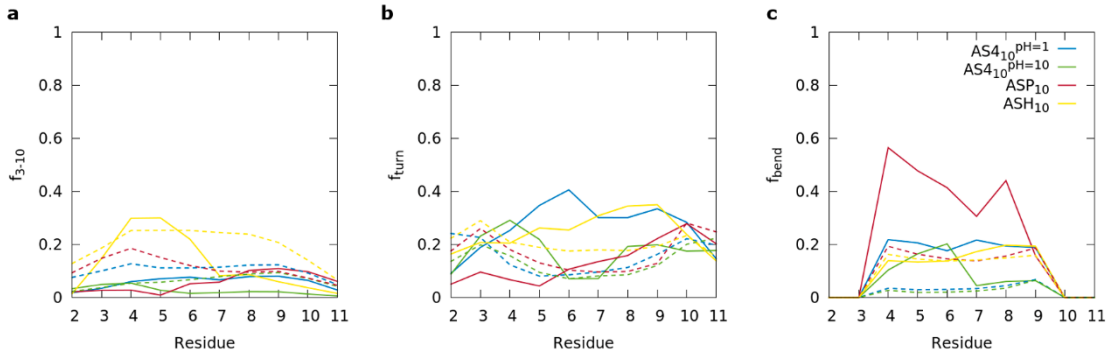


Figure A1. Fractions of (a) 3-10 helix, (b) β-turn and (c) bend of the fully protonated (ASH_{10} and $AS4_{10}^{pH=1}$) and fully deprotonated (ASP_{10} and $AS4_{10}^{pH=10}$) polyaspartic acid peptides in implicit (dashed lines) and explicit (solid lines) solvation models.

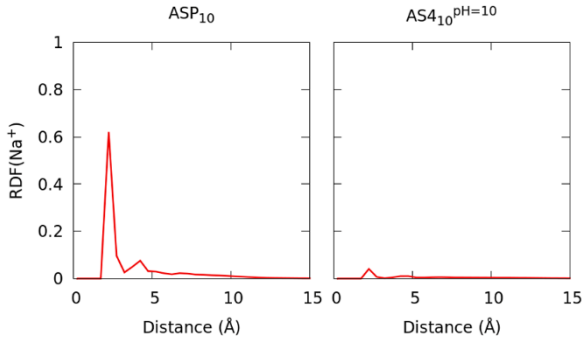


Figure A2. Radial distribution function (RDF) of Na^+ counterions around the oxygen atoms of the carboxyl group of the aspartic acid side chains in the deprotonated state of the CMD (left) and CpHMD (right) simulations with explicit solvation model.

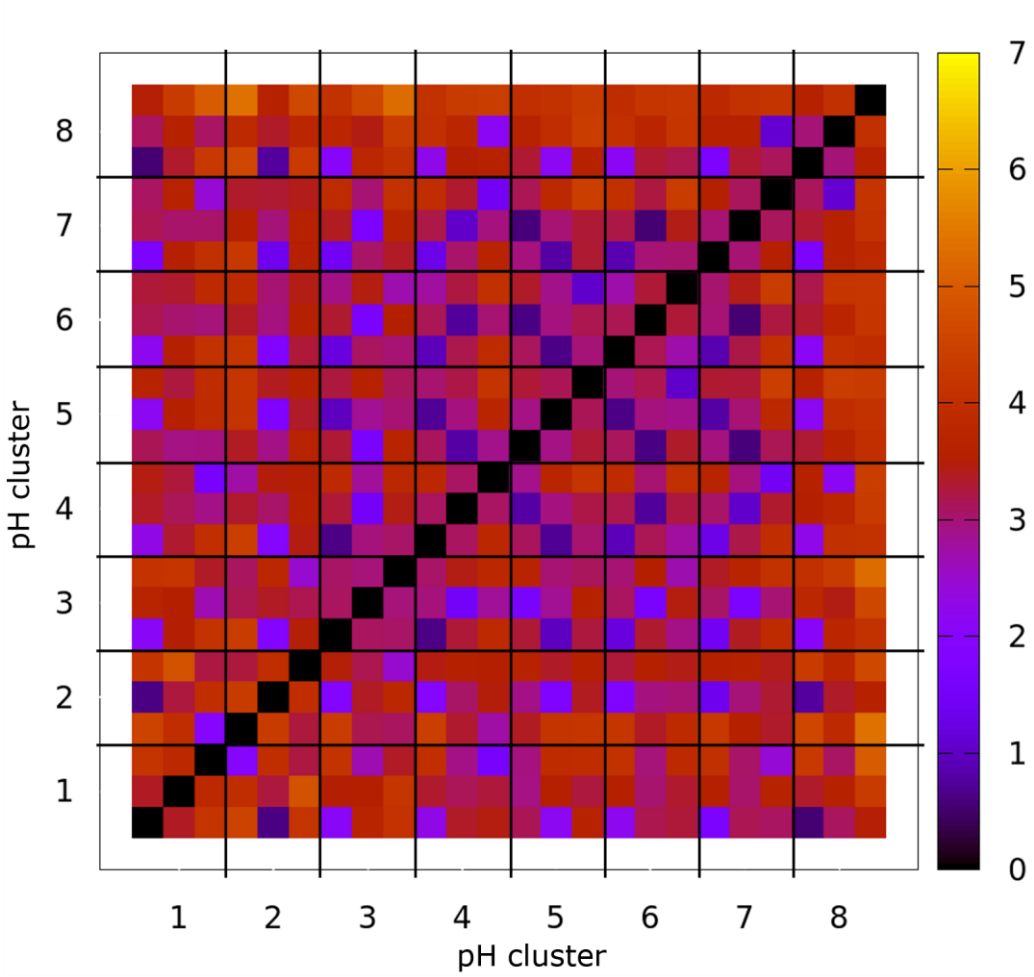


Figure A3. 2D-RMSD of the three most populated clusters of the pH-REMD trajectories from pH = 1.0-8.0 with ΔpH = 0.5. The colour labels indicate the RMSD value (in Å) between the centroid conformations of each cluster.

%	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
$AS4_{10}^{pH=1}$	14	10	10	9	7
$AS4_{10}^{pH=1.5}$	17	13	12	10	9
$AS4_{10}^{pH=2}$	17	12	11	8	7
$AS4_{10}^{pH=2.5}$	22	20	9	9	8
$AS4_{10}^{pH=3}$	27	15	15	8	8
$AS4_{10}^{pH=3.5}$	23	16	15	9	7
$AS4_{10}^{pH=4}$	20	17	14	12	8
$AS4_{10}^{pH=4.5}$	18	18	14	13	11
$AS4_{10}^{pH=5}$	21	20	14	10	8
$AS4_{10}^{pH=5.5}$	21	16	14	9	9
$AS4_{10}^{pH=6}$	29	20	16	11	8
$AS4_{10}^{pH=6.5}$	23	18	18	11	9
$AS4_{10}^{pH=7}$	28	15	12	12	11
$AS4_{10}^{pH=7.5}$	36	18	15	12	5
$AS4_{10}^{pH=8}$	55	8	7	7	6
$AS4_{10}^{pH=8.5}$	55	7	7	6	5

Table A2. Population fractions of the five most populated clusters of $AS4_{10}^{pH=2.5\text{-}6.0,\ \Delta pH=0.5}$ in the pH-REMD simulation.

Appendix B

Supporting Information to "On the Use of Constant pH Molecular Dynamics to Describe the Conformational Space of Peptides"

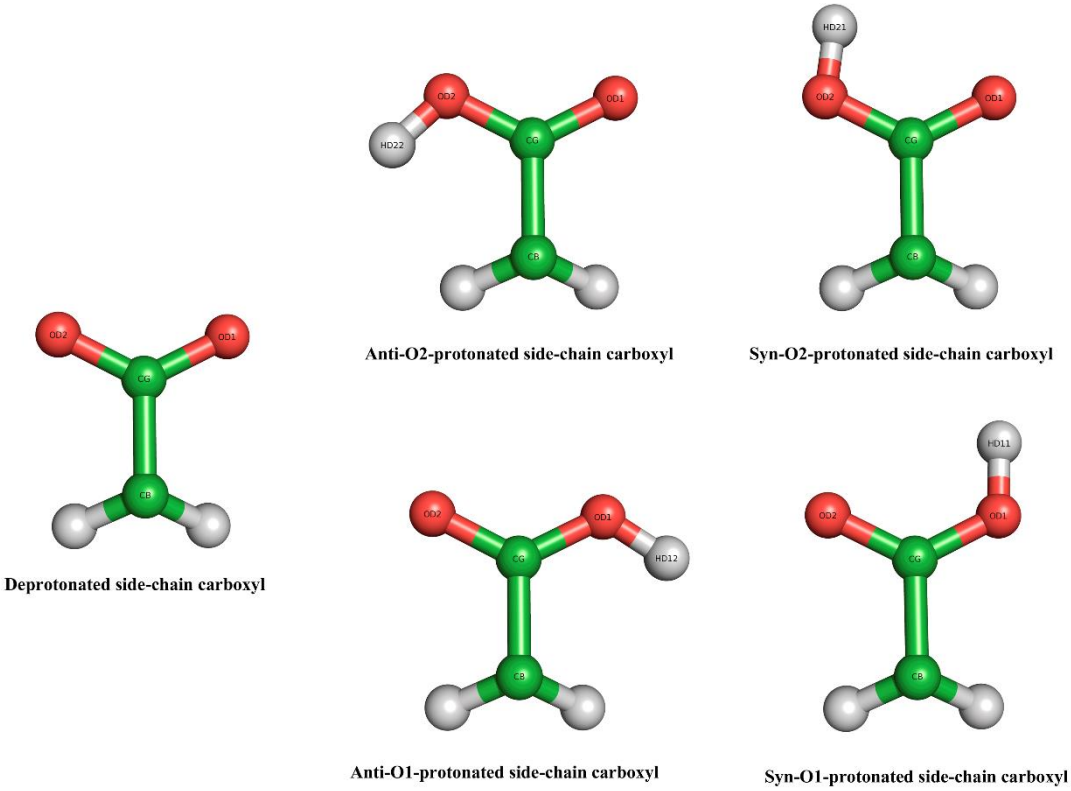


Figure B1. Protonatable sites in the aspartic acid side chain. There are four protonatable sites corresponding to the *anti* or *syn* position with respect to each oxygen of the carboxyl group (OD1 and OD2). The CpHMD method builds a residue with the four hydrogens, and only make one or none are active, depending on the protonation state.

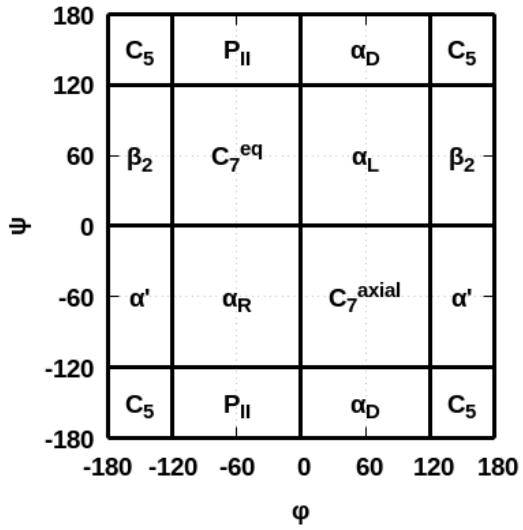


Figure B2. Classification of the nine secondary structure regions (C_5, P_{II}, α_D, β_2, C_7^{eq}, α_L, α', α_R and C_7^{axial}) in the Ramachandran map by J. Rubio-Martinez et al.[1].

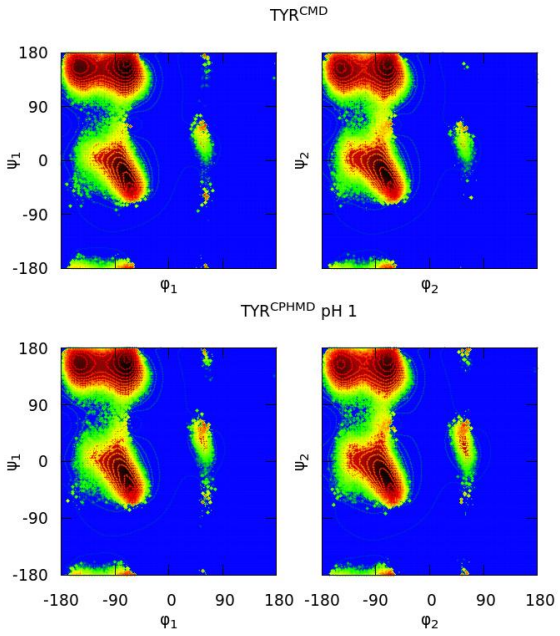


Figure B3. Gibbs free energies in the Ramachandran space of the capped Tyr_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. Only the protonated forms are shown for TYR residue (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

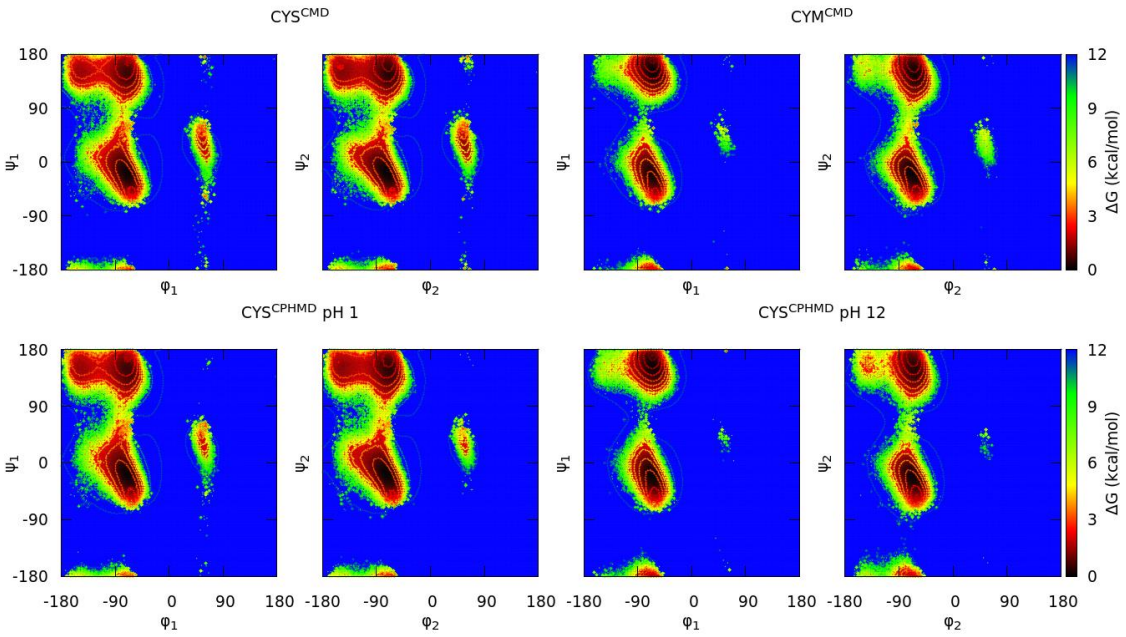


Figure B4. Gibbs free energies in the Ramachandran space of the capped Cys_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. The protonated forms are on the left (CMD; top—CpHMD; bottom) and the deprotonated forms are on the right (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

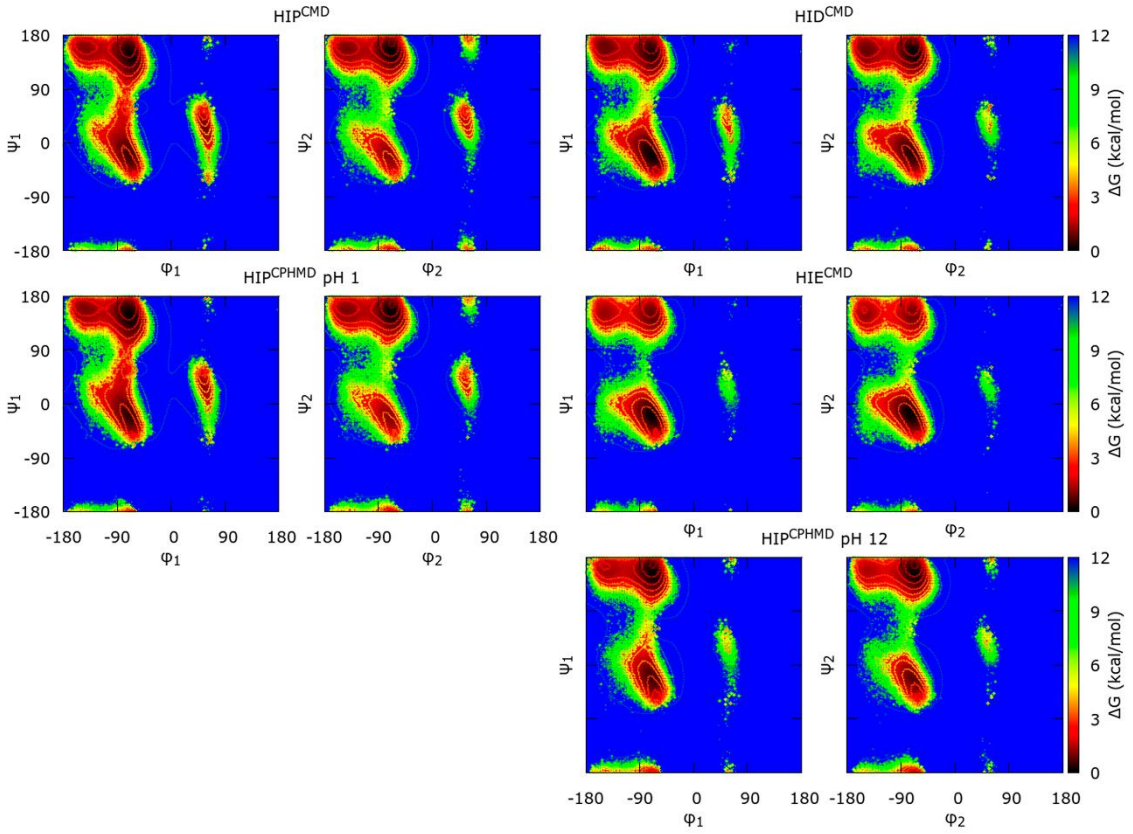


Figure B5. Gibbs free energies in the Ramachandran space of the capped His_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. The protonated forms are on the left (CMD; top—CpHMD; bottom) and the deprotonated forms are on the right (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

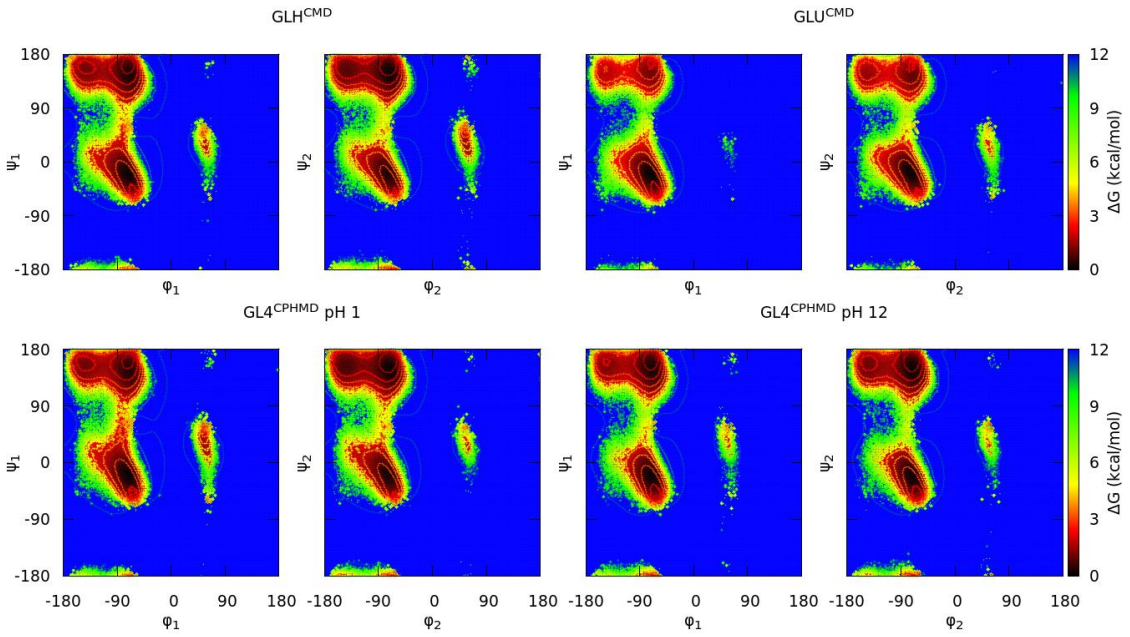


Figure B6. Gibbs free energies in the Ramachandran space of the capped Glu_2 tripeptide. Each subtitle indicates the residue, the simulation method (in the superscript) and the solvent pH (for the CpHMD simulations only). Both sets of dihedrals (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid) are illustrated. The protonated forms are on the left (CMD; top—CpHMD; bottom) and the deprotonated forms are on the right (CMD; top—CpHMD; bottom). The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

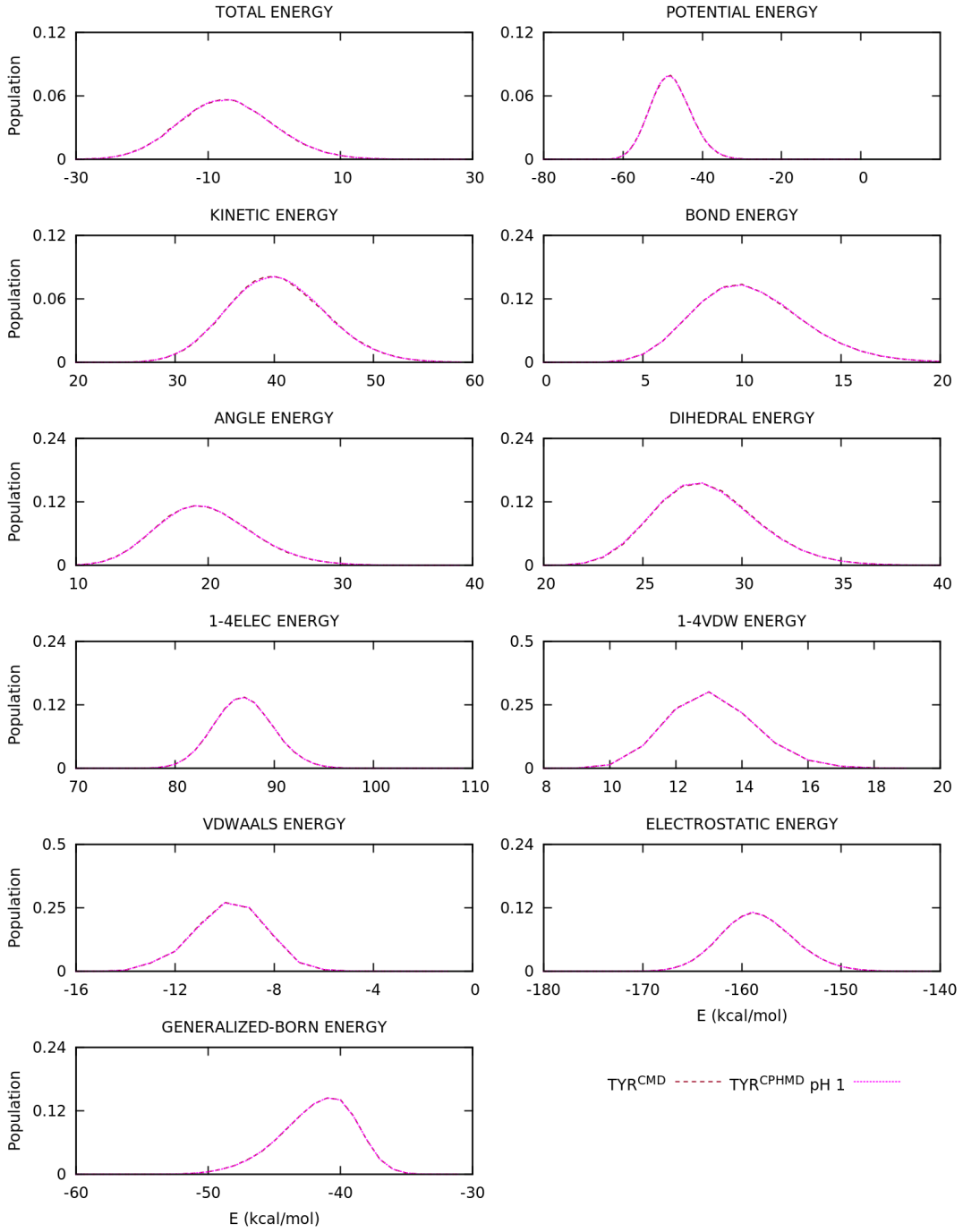


Figure B7. Energy distributions of the capped Tyr_2 tripeptide. Global, inner, van der Waals and electrostatic terms are show. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

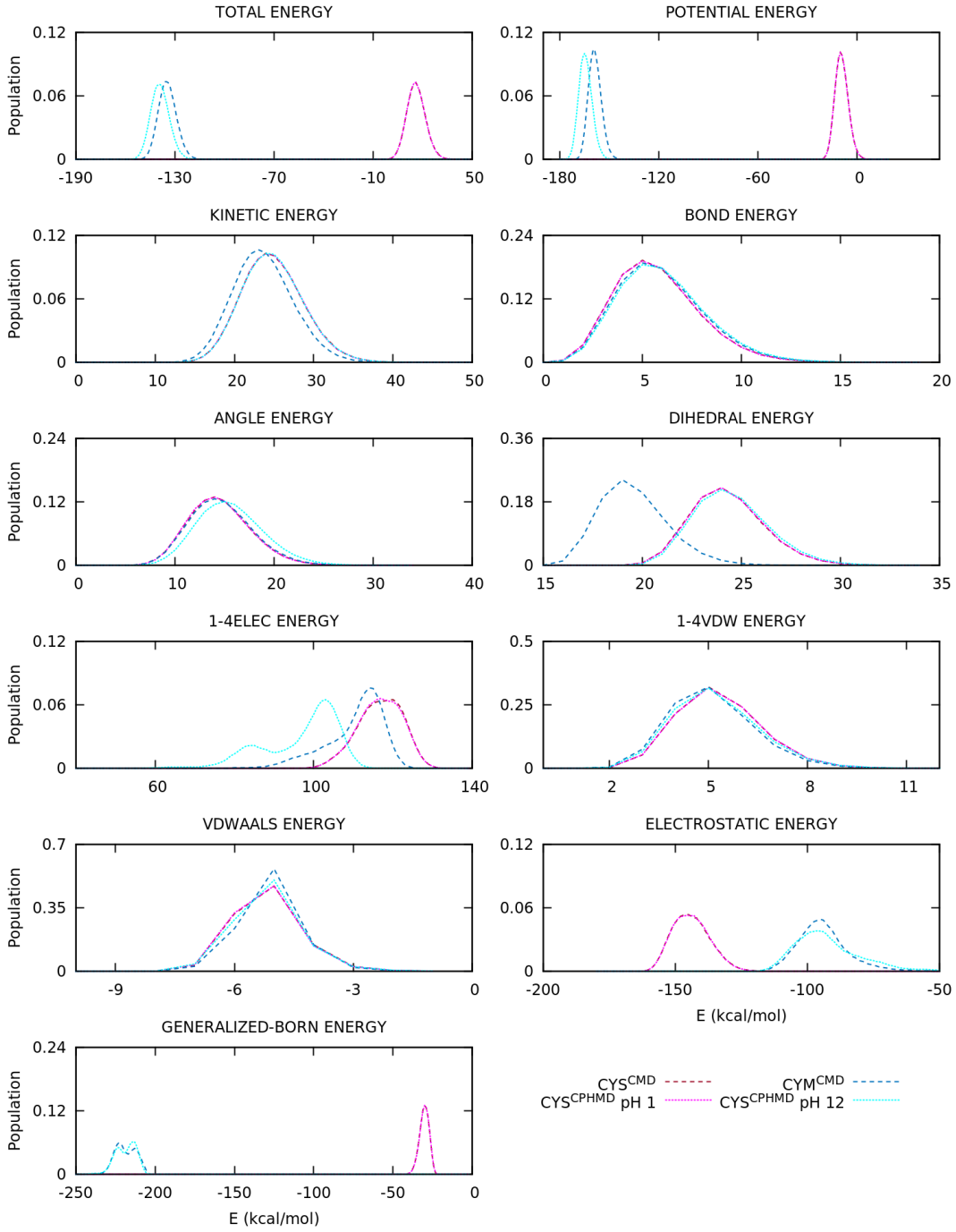


Figure B8. Energy distributions of the capped Cys_2 tripeptide. Global, inner, van der Waals and electrostatic terms are shown. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

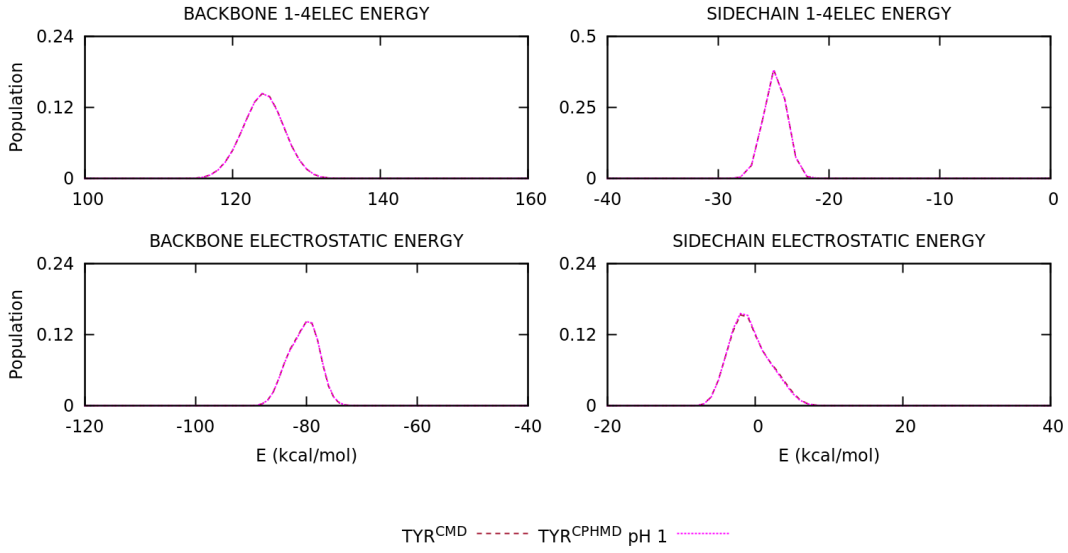


Figure B9. Energy distribution of the 1-4 and long-range electrostatics of the backbone and side chain atoms of the capped Tyr_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

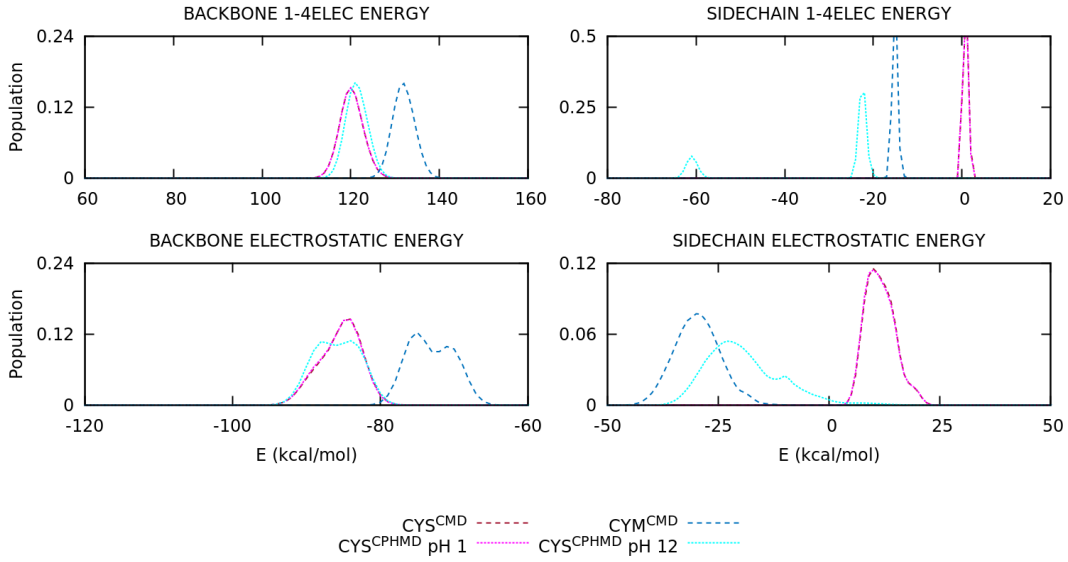


Figure B10. Energy distribution of the 1-4 and long-range electrostatics of the backbone and side chain atoms of the capped Tyr_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

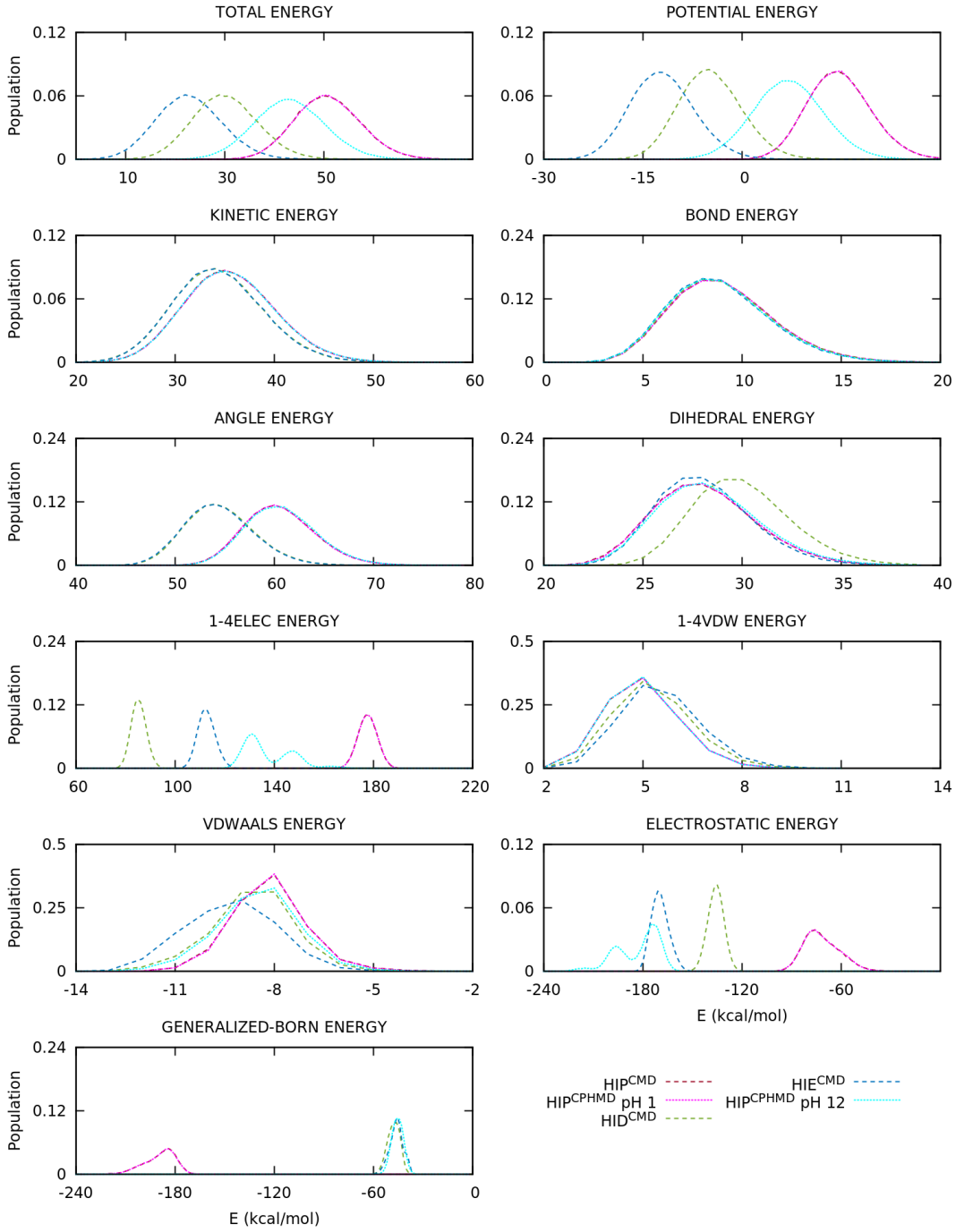


Figure B11. Energy distributions of the capped His_2 tripeptide. Global, inner, van der Waals and electrostatic terms are illustrated. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

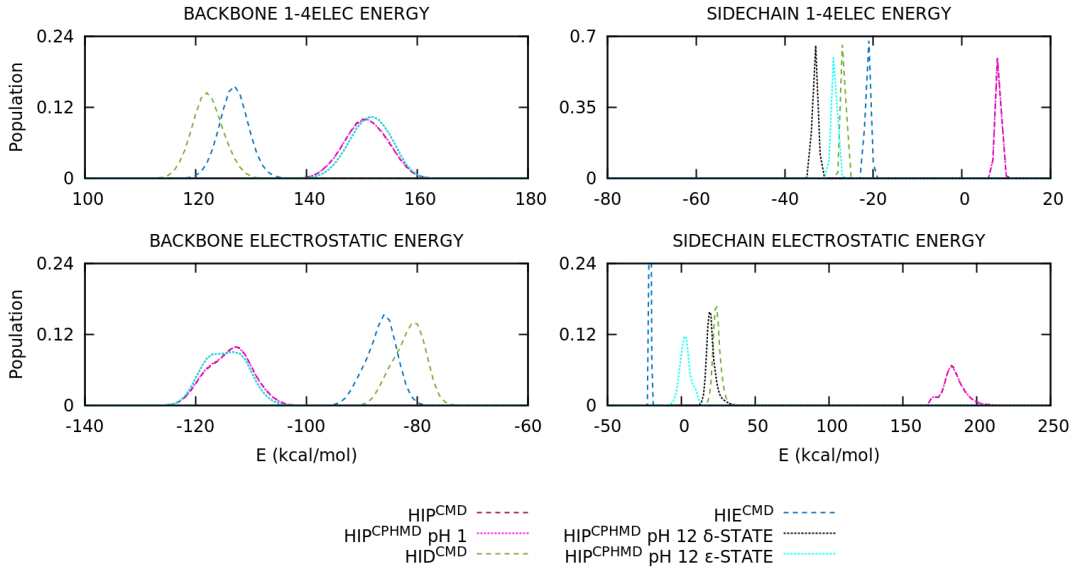


Figure B12. Energy distribution of the 1-4 and long-range electrostatics of the backbone and side chain atoms of the capped His_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively. The labels δ-STATE and ε-STATE refer to the partial charges used to calculate electrostatic energies of the side chain.

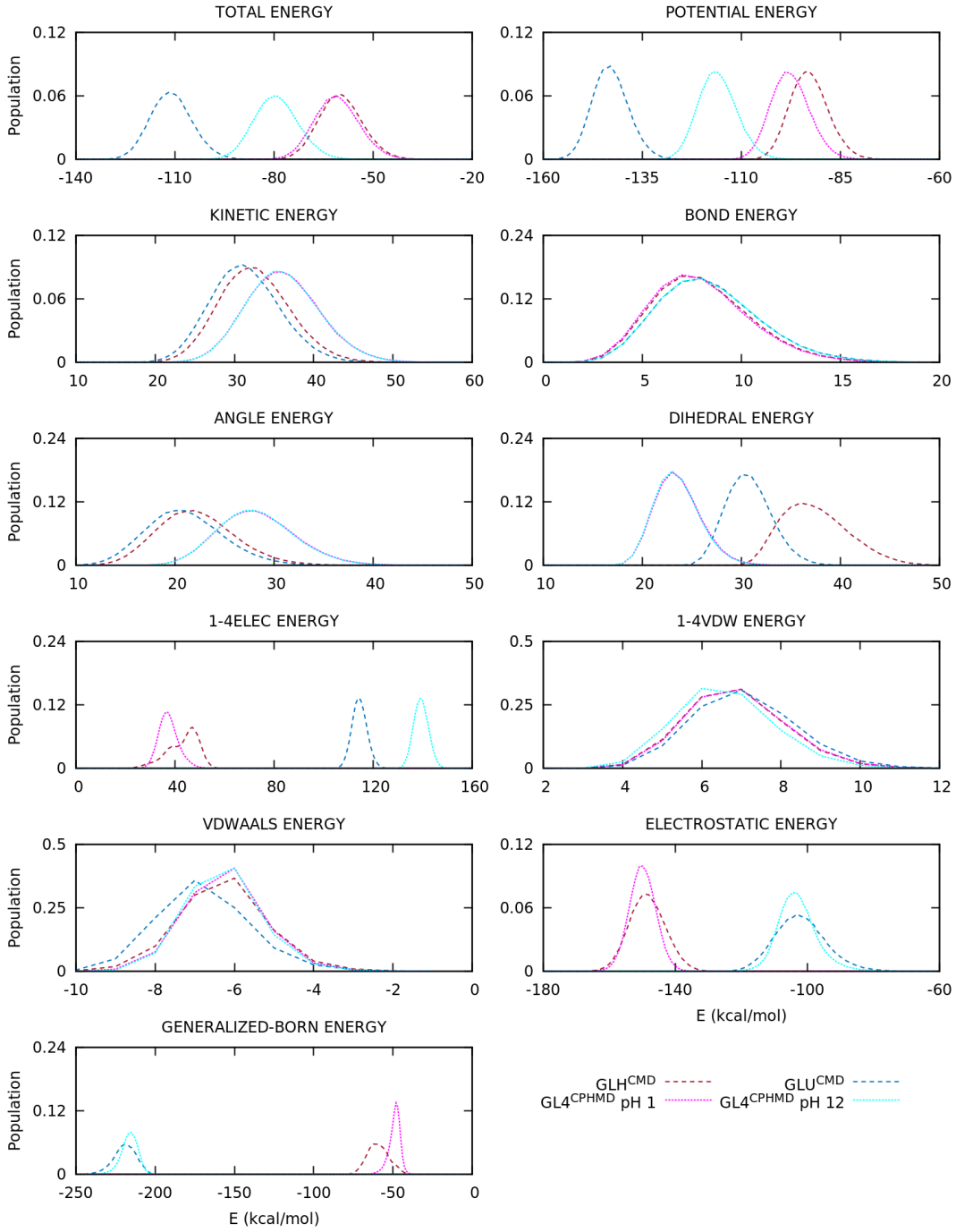


Figure B13. Energy distributions of the capped Glu_2 tripeptide. Global, inner, van der Waals and electrostatic terms are illustrated. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

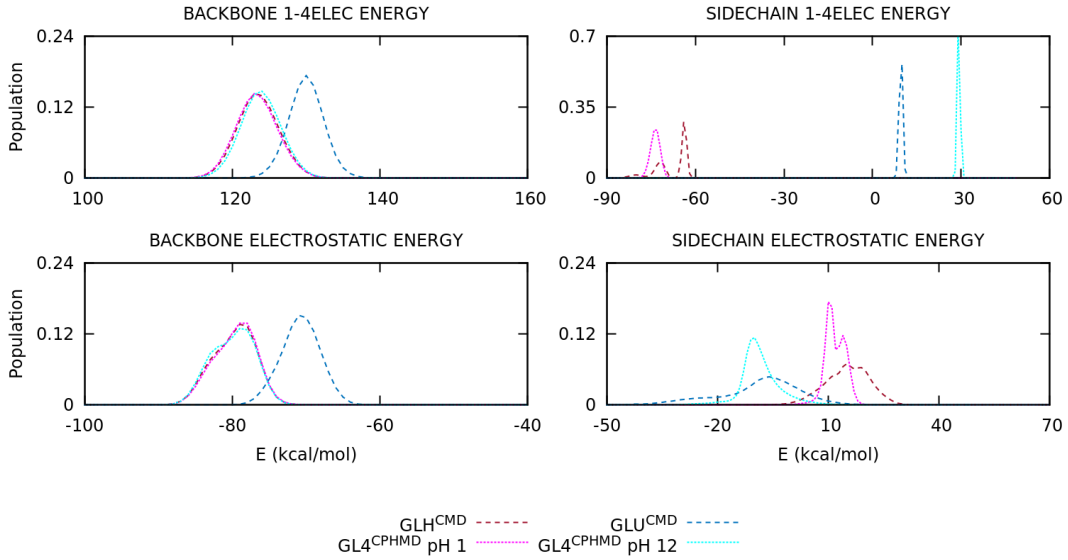


Figure B14. Energy distribution of the 1-4 and long-range electrostatics of the backbone and side chain atoms of the capped Glu_2 tripeptide. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

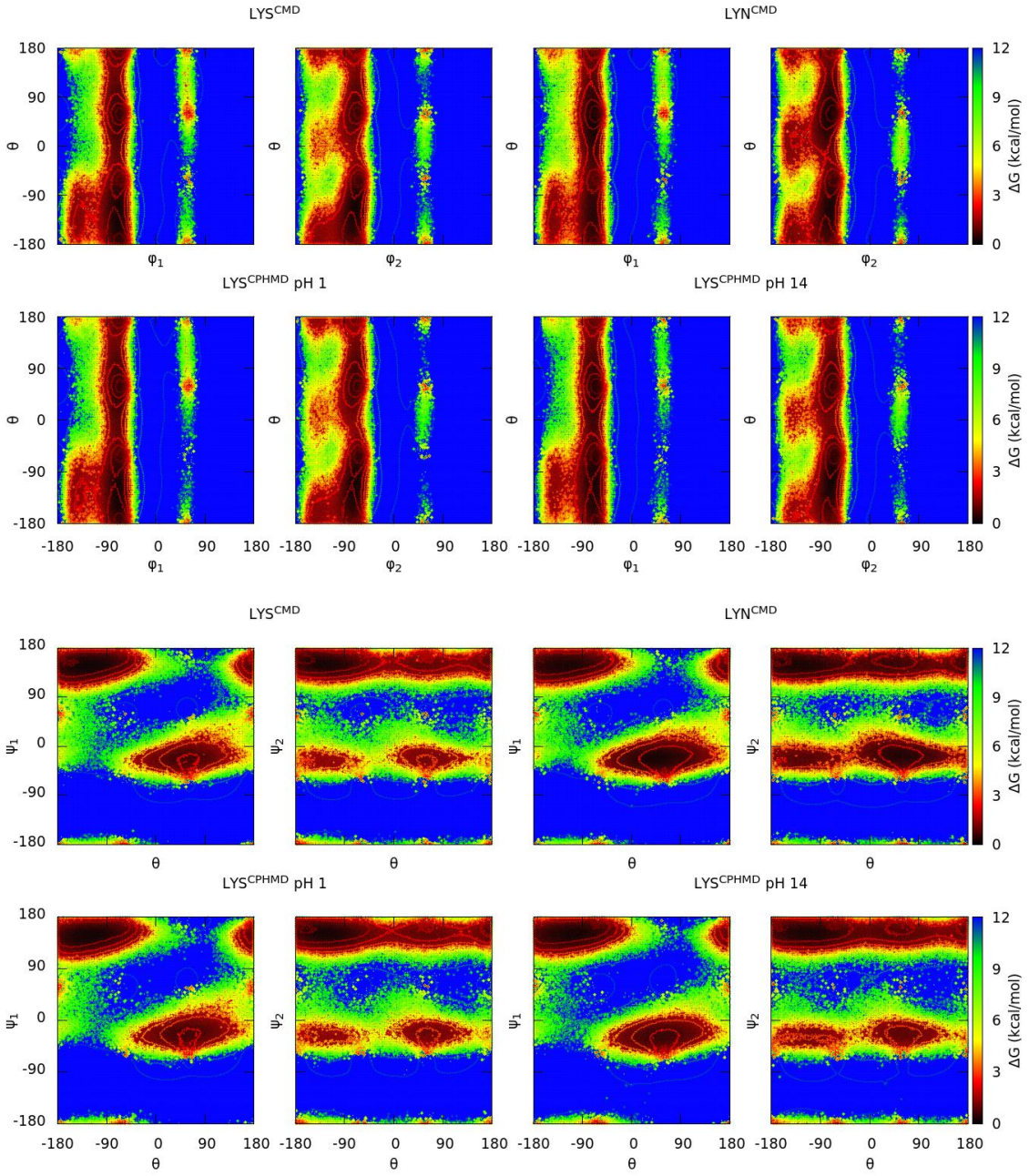


Figure B15. Gibbs free energies in the *sidechain-orientation* space of the capped Lys_2 tripeptide. The labels indicate the residue, the simulation method (in the superscript) and the pH (for the CpHMD simulations only). Four sets of dihedral angles are represented in this plot, using the θ dihedral together with the φ or ψ dihedral of each monomer (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid). The protonated forms are on the left and deprotonated forms on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

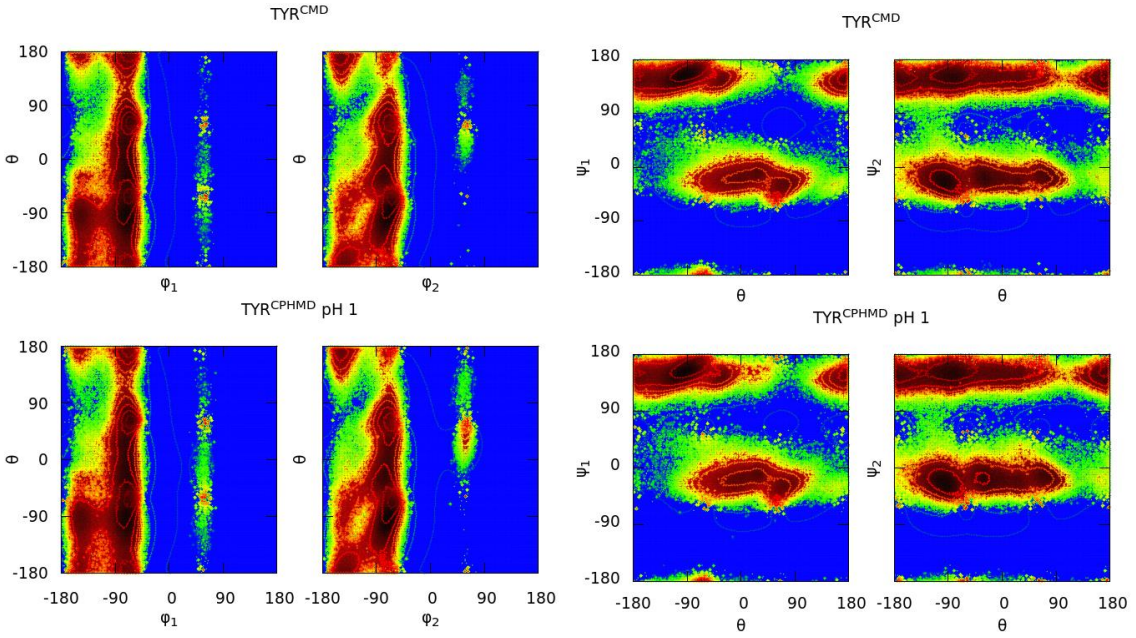


Figure B16. Gibbs free energies in the *sidechain-orientation* space of the capped Tyr_2 tripeptide. The labels indicate the residue, the simulation method (in the superscript) and the pH (for the CpHMD simulations only). Four sets of dihedral angles are represented in this plot, using the θ dihedral together with the φ or ψ dihedral of each monomer (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid). Only the protonated forms are illustrated for TYR residue. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

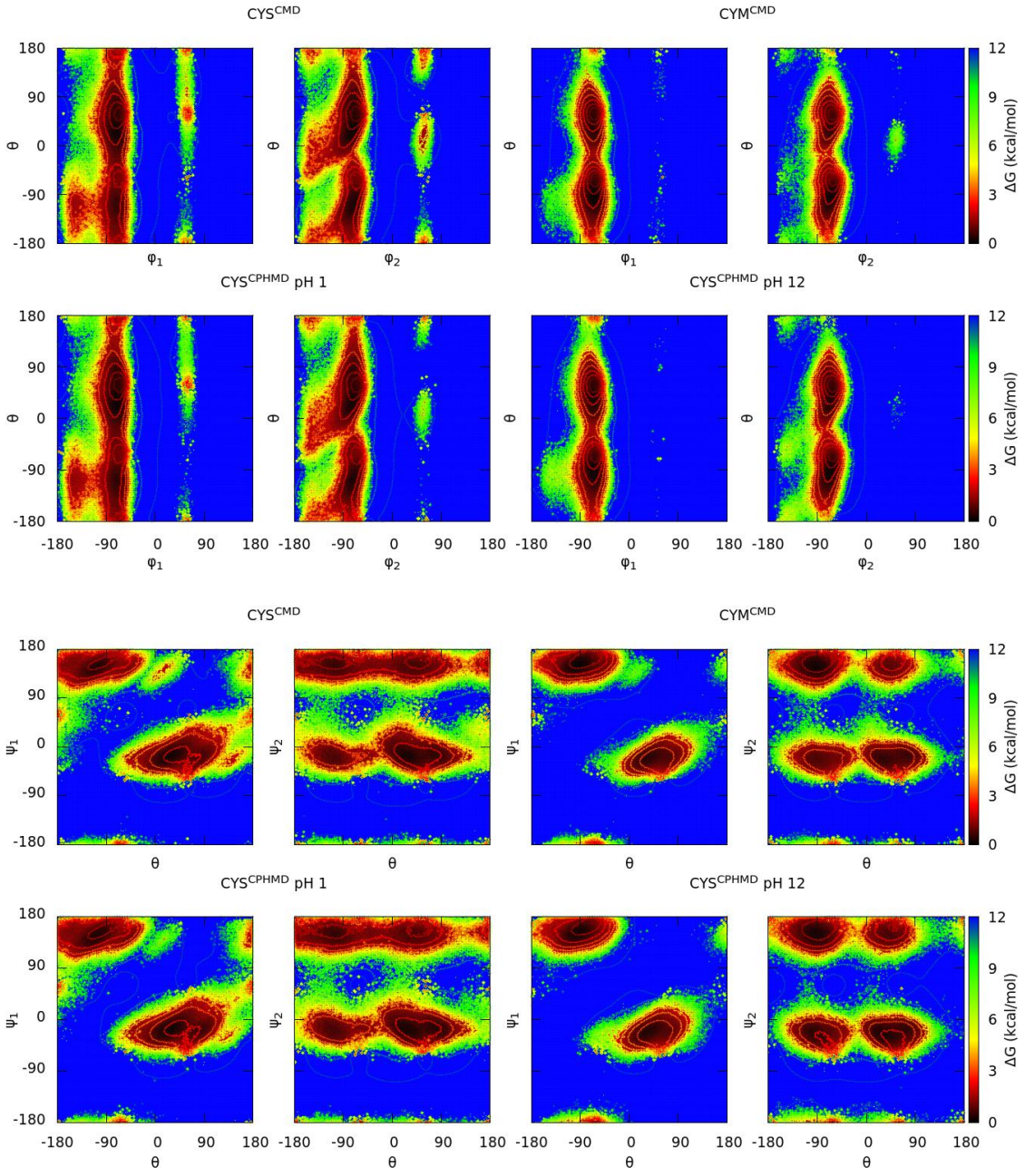


Figure B17. Gibbs free energies in the *sidechain-orientation* space of the capped Cys_2 tripeptide. The labels indicate the residue, the simulation method (in the superscript) and the pH (for the CpHMD simulations only). Four sets of dihedral angles are represented in this plot, using the θ dihedral together with the φ or ψ dihedral of each monomer (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid). The protonated forms are on the left and deprotonated forms on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

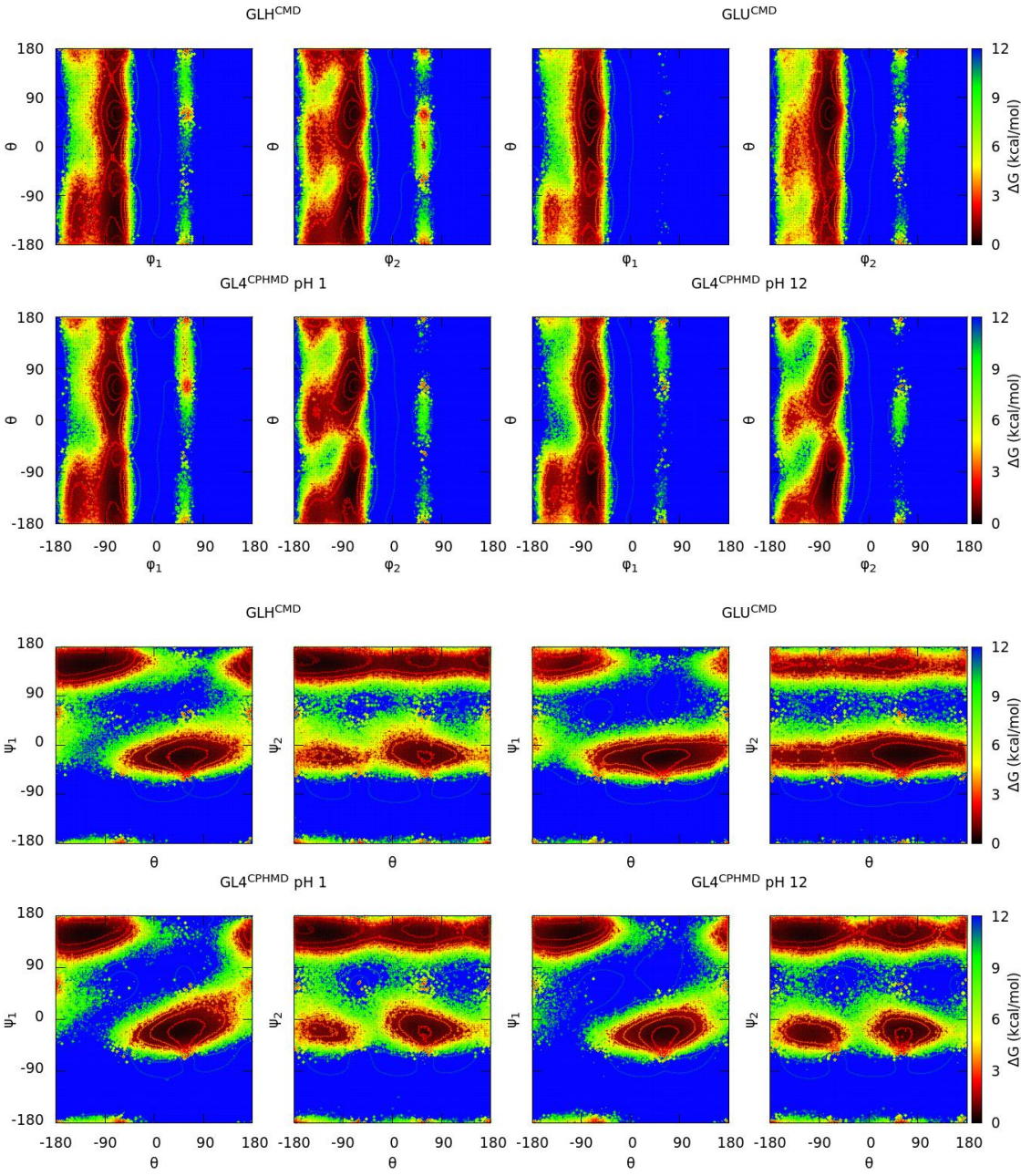


Figure B18 Gibbs free energies in the *sidechain-orientation* space of the capped Glu_2 tripeptide. The labels indicate the residue, the simulation method (in the superscript) and the pH (for the CpHMD simulations only). Four sets of dihedral angles are represented in this plot, using the θ dihedral together with the φ or ψ dihedral of each monomer (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid). The protonated forms are on the left and deprotonated ones on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

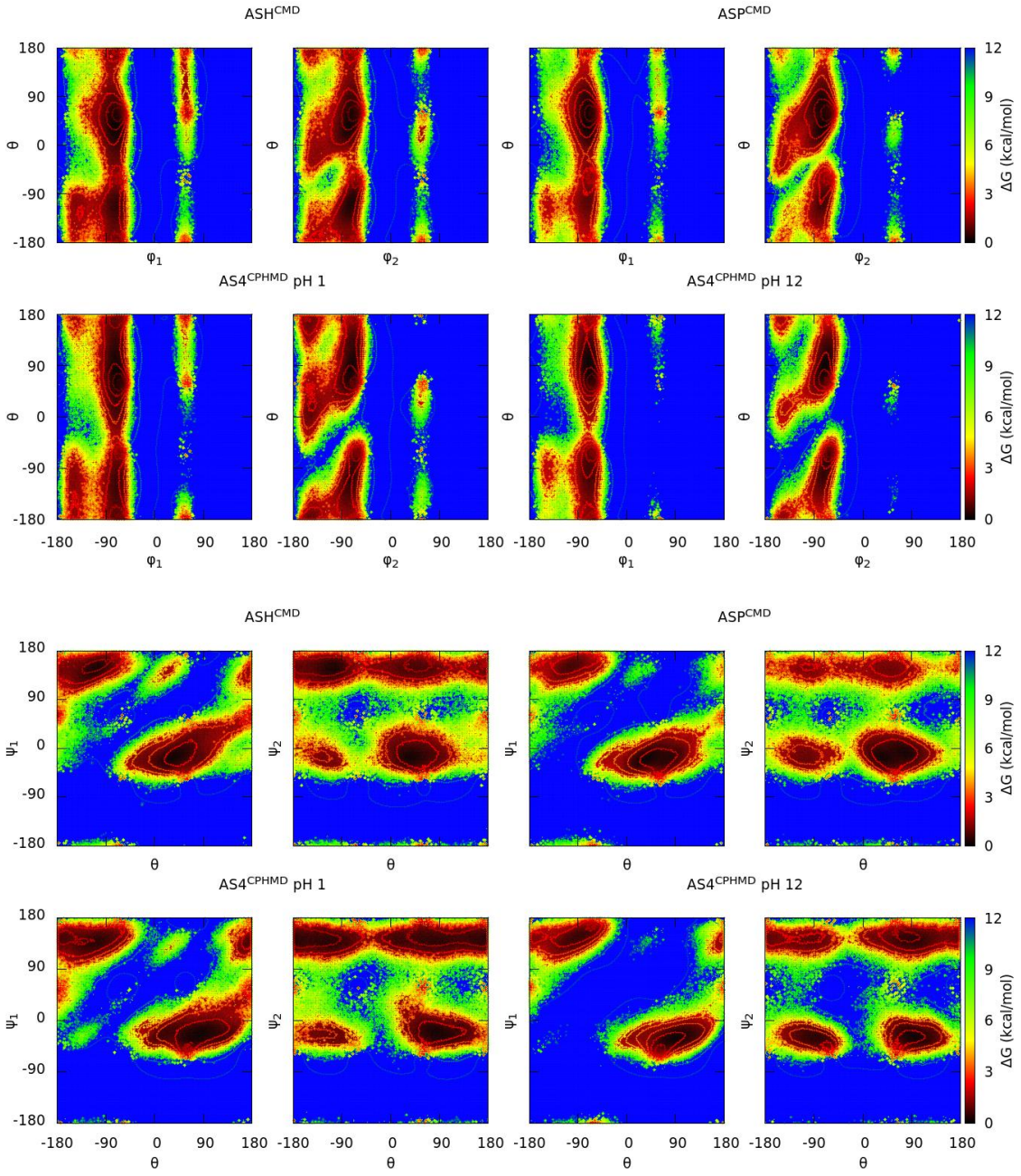


Figure B19. Gibbs free energies in the *sidechain-orientation* space of the capped Asp_2 tripeptide. The labels indicate the residue, the simulation method (in the superscript) and the pH (only for the CpHMD simulations). Four sets of dihedral angles are represented in this plot, using the θ dihedral together with the φ or ψ dihedral of each monomer (φ_1/ψ_1 from the N-terminal amino acid; φ_2/ψ_2 from the C-terminal amino acid). The protonated forms are on the left and deprotonated forms on the right. The solid lines indicate an increase of 0.6 kcal/mol in the energy values.

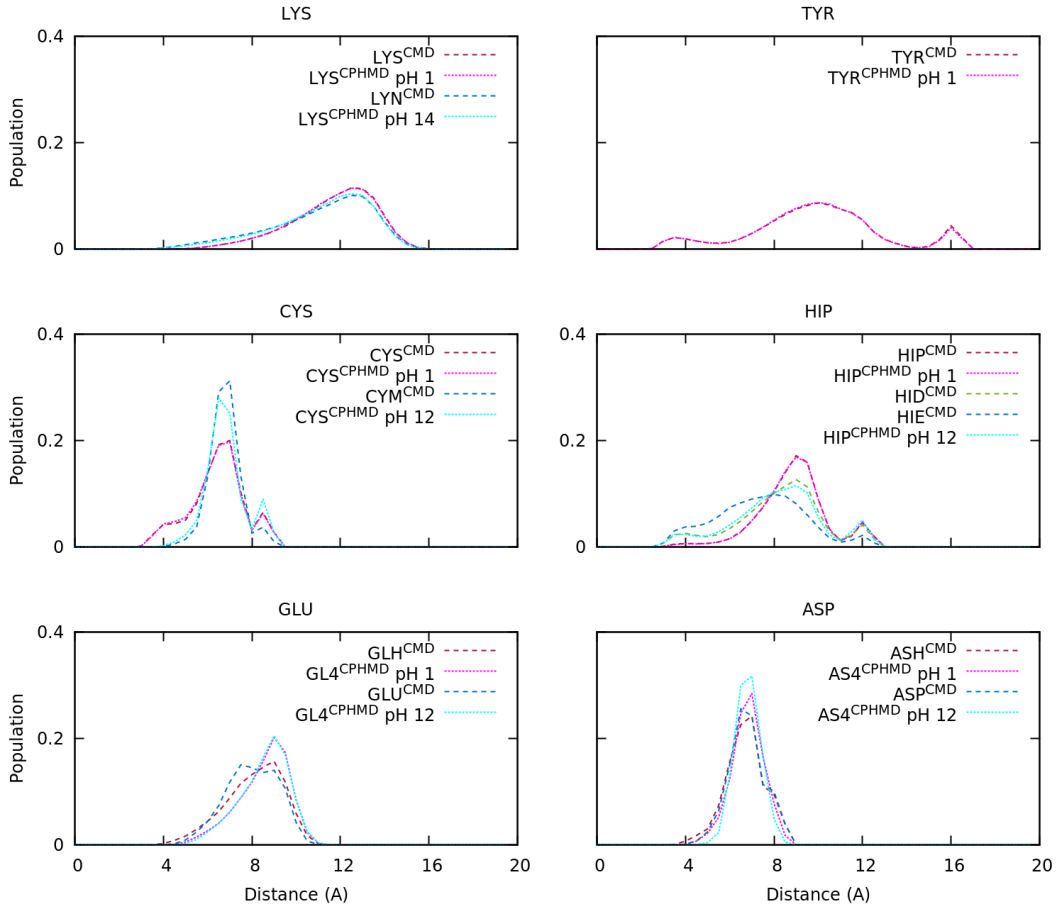


Figure B20. Distribution of the interatomic distance between the atoms of the side chain selected for the construction of the dihedral angle θ (Table B3).

ATOM	GLH	GLU	pH12	P-sO$_2$	P-aO$_2$	P-sO$_1$	P-aO$_1$	ATOM	ASH	ASP	pH12	P-sO$_2$	P-aO$_2$	P-sO$_1$	cP-aO$_1$
N	-0,4157	-0,5163	-0,4157	-0,4157	-0,4157	-0,4157	-0,4157	**N**	-0,4157	-0,5163	-0,4157	-0,4157	-0,4157	-0,4157	-0,4157
H	0,2719	-0,2936	0,2719	0,2719	0,2719	0,2719	0,2719	**H**	0,2719	0,2936	0,2719	0,2719	0,2719	0,2719	0,2719
CA	0,0145	-0,0397	0,0145	0,0145	0,0145	0,0145	0,0145	**CA**	0,0341	-0,0381	0,0341	0,0341	0,0341	0,0341	0,0341
HA	0,0779	0,1105	0,0779	0,0779	0,0779	0,0779	0,0779	**HA**	0,0864	-0,088	0,0864	0,0864	0,0864	0,0864	0,0864
CB	-0,0071	0,056	-0,0398	-0,0071	-0,0071	-0,0071	-0,0071	**CB**	-0,0316	-0,0303	-0,1783	-0,0316	-0,0316	-0,0316	-0,0316
HB2	0,0256	-0,0173	-0,0173	0,0256	0,0256	0,0256	0,0256	**HB2**	0,0488	-0,0122	-0,0122	0,0488	0,0488	0,0488	0,0488
HB3	0,0256	-0,0173	-0,0173	0,0256	0,0256	0,0256	0,0256	**HB3**	0,0488	-0,0122	-0,0122	0,0488	0,0488	0,0488	0,0488
CG	-0,0174	0,0136	0,0136	-0,0174	-0,0174	-0,0174	-0,0174	**CG**	0,6462	0,7994	0,7994	0,6462	0,6462	0,6462	0,6462
HG2	0,0430	-0,0425	-0,0425	0,0430	0,0430	0,0430	0,0430	**OD1**	-0,5554	-0,8014	-0,8014	-0,5554	-0,5554	-0,5554	-0,5554
HG3	0,0430	-0,0425	-0,0425	0,0430	0,0430	0,0430	0,0430	**OD2**	-0,6376	-0,8014	-0,8014	-0,6376	-0,6376	-0,6376	-0,6376
CD	0,6801	0,8054	0,8054	0,6801	0,6801	0,6801	0,6801	**HD21**	0,4747	-	0,0000	0,4747	0,0000	0,0000	0,0000
OE1	-0,5838	-0,8188	-0,8188	-0,5838	-0,5838	-0,5838	-0,5838	**C**	0,5973	0,5366	0,5973	0,5973	0,5973	0,5973	0,5973
OE2	-0,6511	-0,8188	-0,8188	-0,6511	-0,6511	-0,6511	-0,6511	**O**	-0,5679	-0,5819	-0,5679	-0,5679	-0,5679	-0,5679	-0,5679
HE2	0,4641	-	0,0000	0,4641	0,0000	0,0000	0,0000	**HD22**	-	-	0,0000	0,0000	0,4747	0,0000	0,0000
C	0,5973	0,5366	0,5973	0,5973	0,5973	0,5973	0,5973	**HD11**	-	-	0,0000	0,0000	0,0000	0,4747	0,0000
O	-0,5679	-0,5819	-0,5679	-0,5679	-0,5679	-0,5679	-0,5679	**HD12**	-	-	0,0000	0,0000	0,0000	0,0000	0,4747
HE22	-	-	0,0000	0,0000	0,4641	0,0000	0,0000								
HE11	-	-	0,0000	0,0000	0,0000	0,4641	0,0000								
HE12	-	-	0,0000	0,0000	0,0000	0,0000	0,4641								

Table B1. Partial charges of the protonated and deprotonated forms of the Glu and Asp amino acids in the CMD and CpHMD simulations. *pH(X)* and *p-(X)* refer to the partial charges used in the CpHMD method and other labels correspond to the CMD residues. Both Glu and Asp amino acids have four protonated states: the *syn* (P-sO$_X$) and *anti* (P-aO$_X$) positions on the oxygen atoms (O1 or O2) of the carboxyl group.

ATOM	LYN	LYS	pH1	pH14	ATOM	TYR	pH1	ATOM	CYM	CYS	pH1	pH12	ATOM	HID	HIE	HIP	pH1	pH12ε	pH12δ
N	-0,416	-0,3479	-0,3479	-0,3479	N	-0,4157	-0,4157	N	-0,4160	-0,4160	-0,4160	-0,4160	N	-0,416	-0,4160	-0,3480	-0,3480	-0,3480	-0,3480
H	0,272	0,2747	0,2747	0,2747	H	0,2719	0,2719	H	0,2720	0,2720	0,2720	0,2720	H	0,2720	0,2720	0,2750	0,2750	0,2750	0,2750
CA	-0,072	-0,2400	-0,2400	-0,2400	CA	-0,0014	-0,0014	CA	-0,0350	0,0210	0,0210	0,0210	CA	0,0190	-0,0580	-0,1350	-0,1350	-0,1350	-0,1350
HA	0,099	0,1426	0,1426	0,1426	HA	0,0876	0,0876	HA	0,0510	0,1120	0,1120	0,1120	HA	0,0880	0,1360	0,1210	0,1210	0,1210	0,1210
CB	0,048	-0,0094	-0,0094	-0,1096	CB	-0,0152	-0,0152	CB	-0,2410	-0,1230	-0,1230	-0,3590	CB	-0,0460	-0,0070	-0,0410	-0,0410	-0,1110	-0,1012
HB2	0,034	0,0362	0,0362	0,0340	HB2	0,0295	0,0295	HB2	0,1120	0,1110	0,1110	0,1110	HB2	0,0400	0,0370	0,0810	0,0810	0,0402	0,0367
HB3	0,034	0,0362	0,0362	0,0340	HB3	0,0295	0,0295	HB3	0,1122	0,1112	0,1112	0,1112	HB3	0,0402	0,0367	0,0810	0,0810	0,0402	0,0367
CG	0,06612	0,0187	0,0187	0,0661	CG	-0,0011	-0,0011	SG	-0,8844	-0,3119	-0,3119	-0,8844	CG	-0,0266	0,1868	-0,0012	-0,0012	-0,0266	0,1868
HG2	0,01041	0,0103	0,0103	0,0104	CD1	-0,1906	-0,1906	HG	-	0,1933	0,1933	0,0000	ND1	-0,3811	-0,5432	-0,1513	-0,1513	-0,3811	-0,5432
HG3	0,01041	0,0103	0,0103	0,0104	HD1	0,1699	0,1699	C	0,5973	0,5973	0,5973	0,5973	HD1	0,3649	-	0,3866	0,3866	0,3649	0,0000
CD2	-0,0376	-0,0479	-0,0479	-0,0377	CE1	-0,2341	-0,2341	O	-0,5679	-0,5679	-0,5679	-0,5679	CE1	0,2057	0,1635	-0,0170	-0,0170	0,2057	0,1635
HD2	0,01155	0,0621	0,0621	0,0115	HE1	0,1656	0,1656						HE1	0,1392	0,1435	0,2681	0,2681	0,1390	0,1435
HD3	0,01155	0,0621	0,0621	0,0115	CZ	0,3226	0,3226						NE2	-0,5727	-0,2795	-0,1718	-0,1718	-0,5727	-0,2795
CE2	0,32604	-0,0143	-0,0143	0,3260	OH	-0,5579	-0,5579						HE2	-	0,3339	0,3911	0,3911	0,0000	0,3339
HE2	-0,0335	0,1135	0,1135	-0,0336	HH	0,3992	0,3992						CD2	0,1292	-0,2207	-0,1141	-0,1141	0,1292	-0,2207
HE3	-0,0335	0,1135	0,1135	-0,0336	CE2	-0,2341	-0,2341						HD2	0,1147	0,1862	0,2317	0,2317	0,1147	0,8620
NZ	-1,0358	-0,3854	-0,3854	-1,0358	HE2	0,1656	0,1656						C	0,5973	0,5973	0,7341	0,7341	0,7341	0,7341
HZ1	-	0,3400	0,3400	0,0000	CD2	-0,1906	-0,1906						O	-0,5679	-0,5679	-0,5894	-0,5894	-0,5894	-0,5894
HZ2	0,38604	0,3400	0,3400	0,3860	HD2	0,1699	0,1699												
HZ3	0,38604	0,3400	0,3400	0,3860	C	0,5973	0,5973												
C	0,5973	0,7341	0,7341	0,7341	OH	-0,5679	-0,5679												
O	-0,5679	-0,5894	-0,5894	-0,5894															

Table B2. Partial charges of the protonated and deprotonated forms of the Lys, Tyr, Cys and His amino acids in the CMD and CpHMD simulations. *pH(X)* and *p-(X)* refer to the partial charges used in the CpHMD method and other labels correspond to the CMD residues. The histidine has two states in the neutral form: the ε- (pH12-ε) and δ- (pH12-δ) states.

	System	Atomic distance		Dihedral angle θ
LYS	LYS^{CMD}	11.89 ± 1.90	NZ-NZ	NZ-CA-CA-NZ
	$LYS^{CpHMD}{}_{1}$	11.87 ± 1.90		
	LYN^{CMD}	11.30 ± 2.34		
	$LYS^{CpHMD}{}_{14}$	11.39 ± 2.22		
TYR	TYR^{CMD}	10.13 ± 3.02	OH-OH	OH-CA-CA-OH
	$TYR^{CpHMD}{}_{1}$	10.07 ± 2.96		
CYS	CYS^{CMD}	6.71 ± 1.24	SG-SG	SG-CA-CA-SG
	$CYS^{CpHMD}{}_{1}$	6.68 ± 1.23		
	CYM^{CMD}	7.04 ± 0.71		
	$CYS^{CpHMD}{}_{12}$	7.08 ± 0.91		
HIP	HIP^{CMD}	9.10 ± 1.60	CE1-CE1	CE1-CA-CA-CE1
	$HIP^{CpHMD}{}_{1}$	9.12 ± 1.58		
	HIE^{CMD}	7.63 ± 2.02		
	HID^{CMD}	8.49 ± 2.04		
	$HIP^{CpHMD}{}_{12}$	8.42 ± 2.07		
GLU	GLH^{CMD}	8.34 ± 1.31	CD-CD	CD-CA-CA-CD
	$GL4^{CpHMD}{}_{1}$	8.75 ± 1.15		
	GLU^{CMD}	8.23 ± 1.17		
	$GL4^{CpHMD}{}_{12}$	8.79 ± 1.10		
ASP	ASH^{CMD}	6.95 ± 0.89	CG-CG	CG-CA-CACG
	$AS4^{CpHMD}{}_{1}$	7.05 ± 0.74		
	ASP^{CMD}	7.02 ± 0.81		
	$AS4^{CpHMD}{}_{12}$	7.06 ± 0.57		

Table B3. Average and standard deviation of the interatomic distances calculated from the selected atoms at the end of the side chains. The set of atoms used to define the dihedral angle θ in each tripeptide is also given.

Bibliography

1. Rubio-Martinez, J., Tomas, M. S. & Perez, J. J. Effect of the solvent on the conformational behavior of the alanine dipeptide deduced from MD simulations. *J Mol Graph Model* **78**, 118–128 (2017).

Appendix C

Supporting Information to "Unravelling Constant pH Molecular Dynamics in Oligopeptides with Explicit Solvation Model"

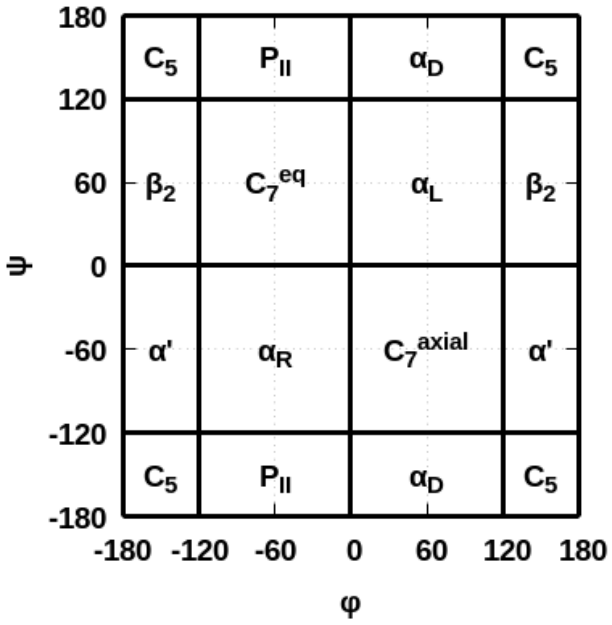


Figure C1. Classification of the nine secondary structure regions (C_5, P_{II}, α_D, β_2, C_7^{eq}, α_L, α', α_R and C_7^{axial}) in the Ramachandran map according to J. Rubio-Martinez et al. [1].

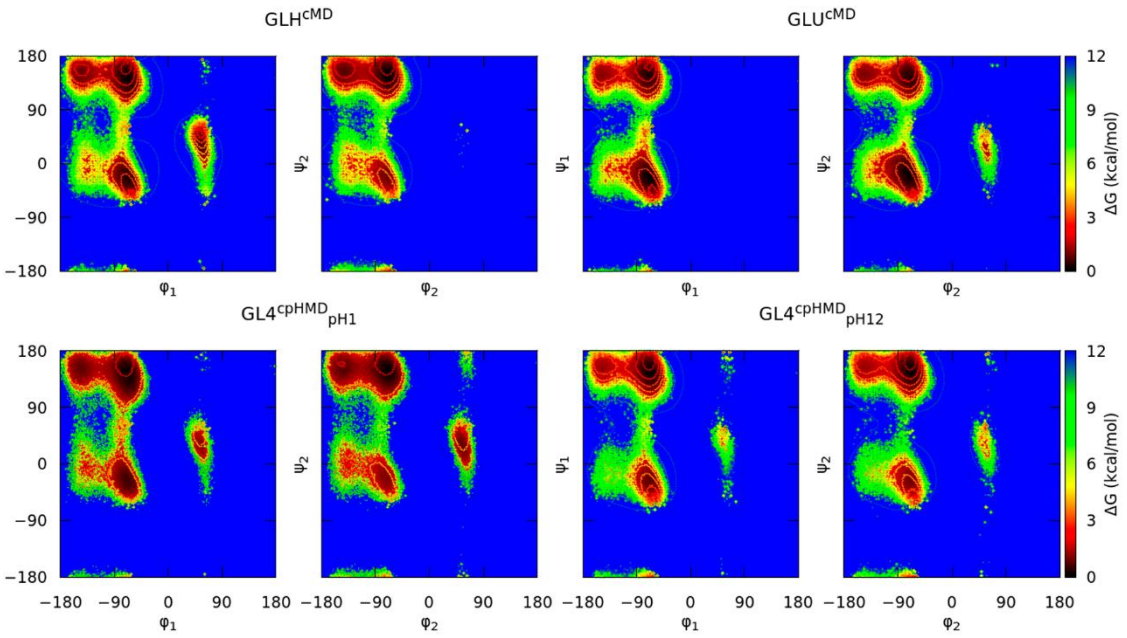


Figure C2. Ramachandran maps of the capped Glu_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation condition has two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

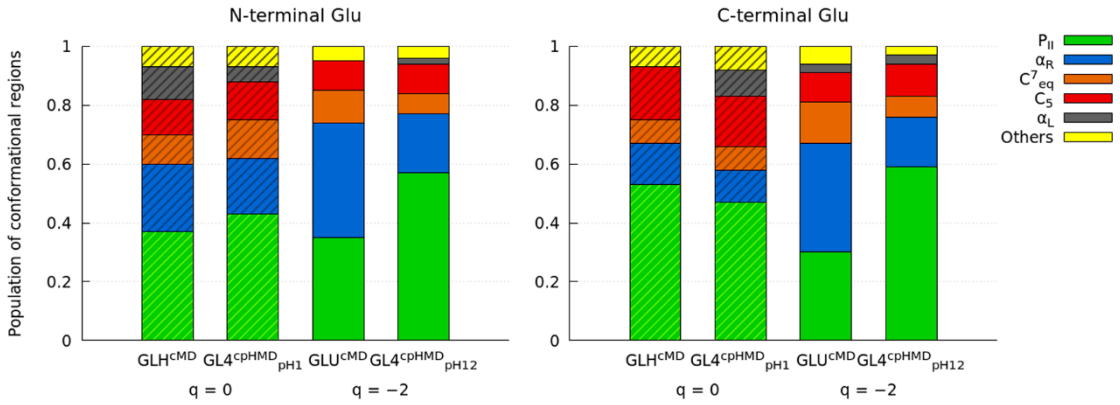


Figure C3. Populations of the conformational regions (P_{II}, α_R, C^7_{eq}, C_5, and α_L) in the Ramachandran maps of each amino acid of the capped Glu_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q). The striped and solid box represent the protonated and deprotonated states, respectively.

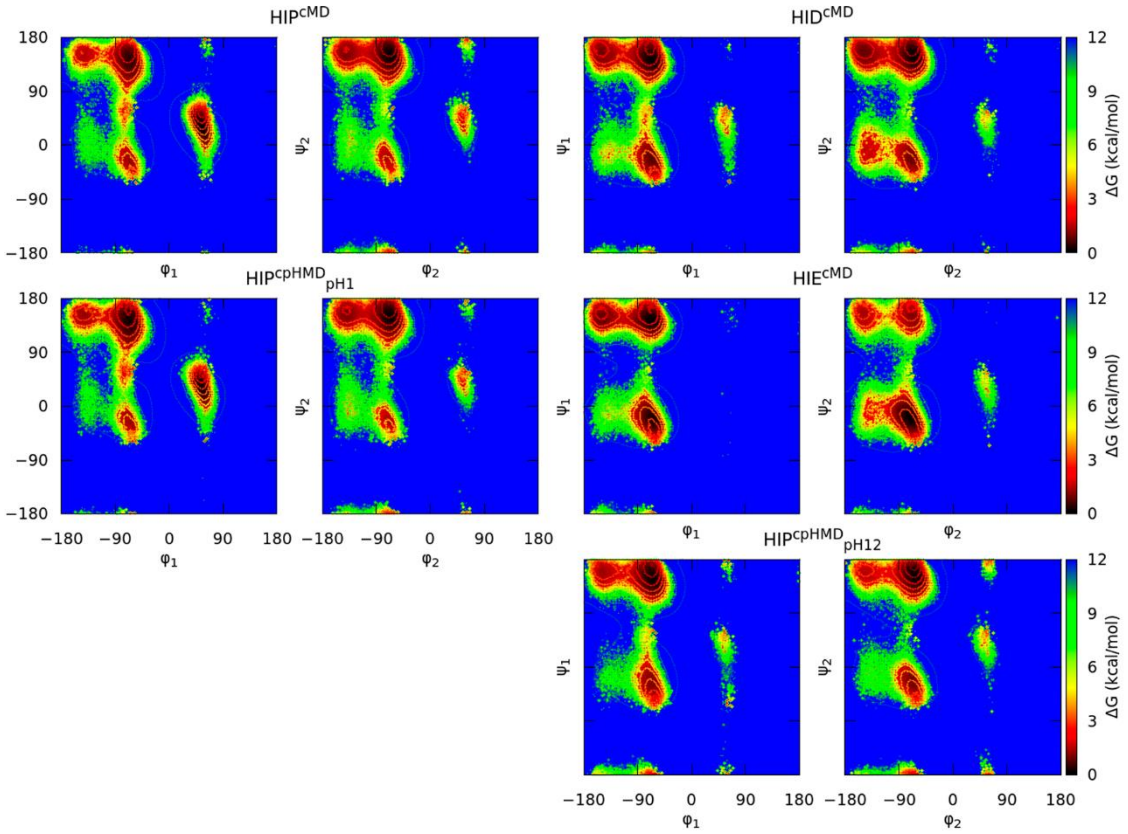


Figure C4. Ramachandran maps of the capped His_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation condition has two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

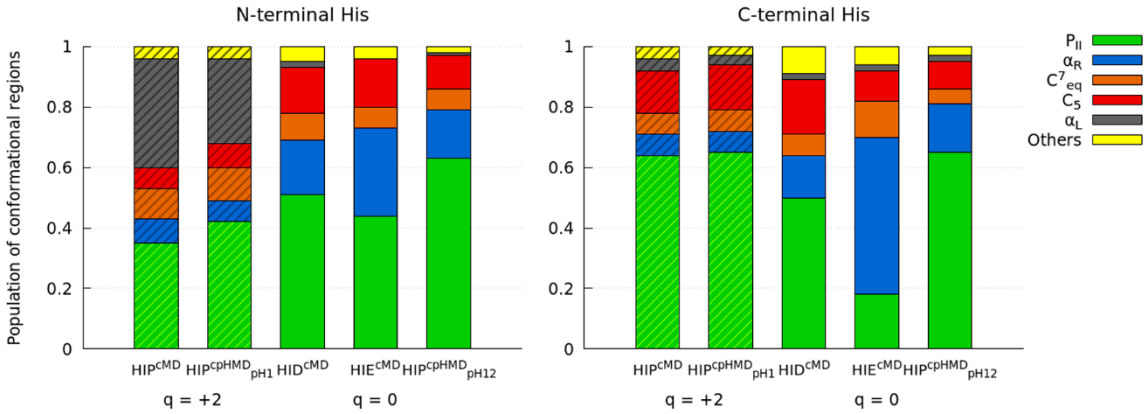


Figure C5. Populations of the conformational regions (P_{II}, α_R, C^7_{eq}, C_5, and α_L) in the Ramachandran maps of each amino acid of the capped His_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q). The striped and solid box represent the protonated and deprotonated states, respectively.

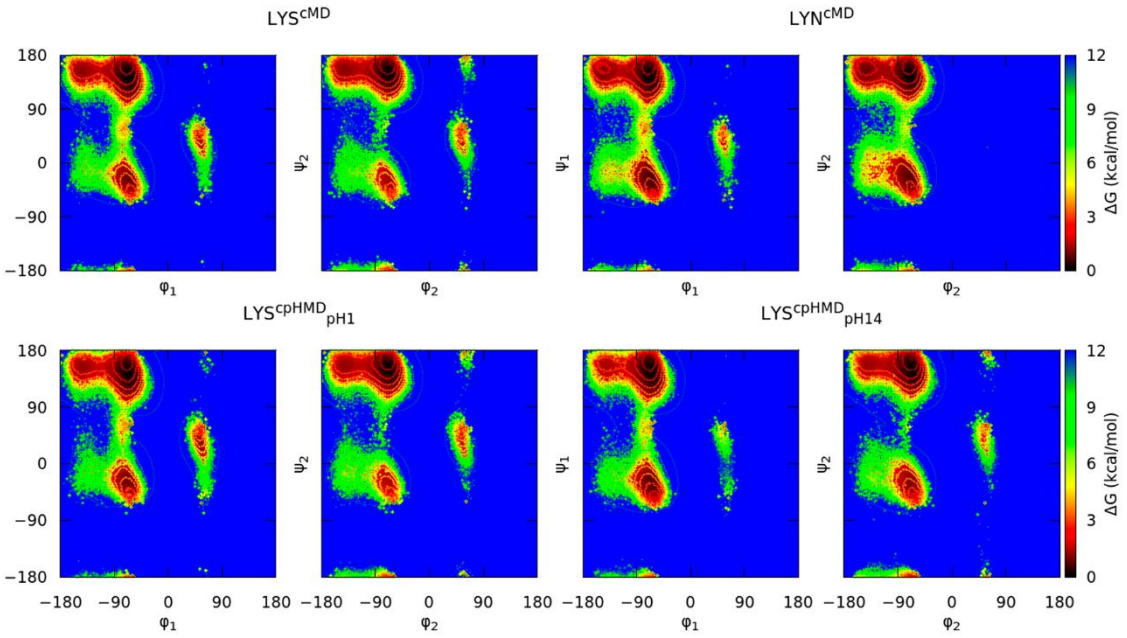


Figure C6. Ramachandran maps of the capped Lys_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation condition has two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

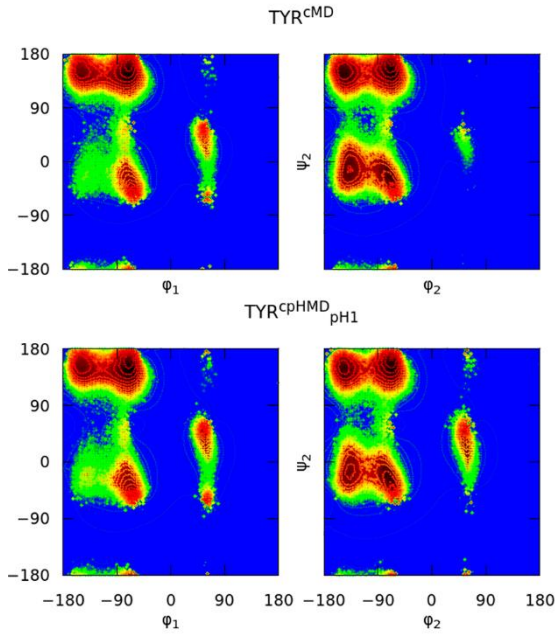


Figure C7. Ramachandran maps of the capped Tyr_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation condition has two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

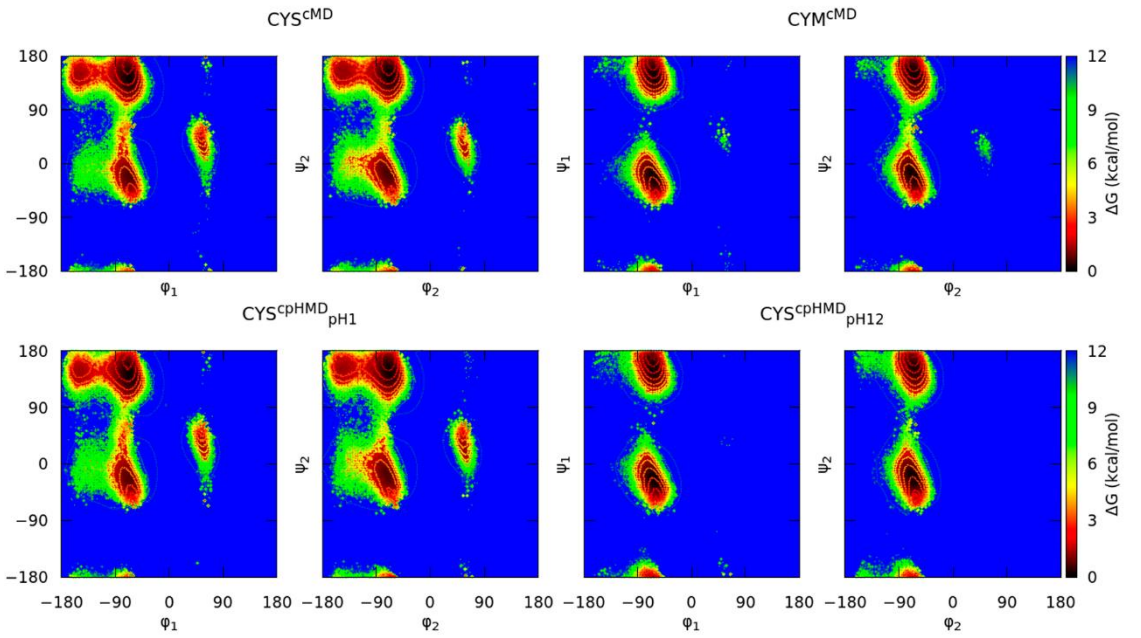


Figure C8. Ramachandran maps of the capped Cys_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. Each simulation conditions have two energy maps corresponding to the set of backbone dihedral angles of the N-terminal (φ_1/ψ_1) or the C-terminal amino acid (φ_2/ψ_2). The solid lines indicate an increase of 0.6 kcal/mol in the energy map.

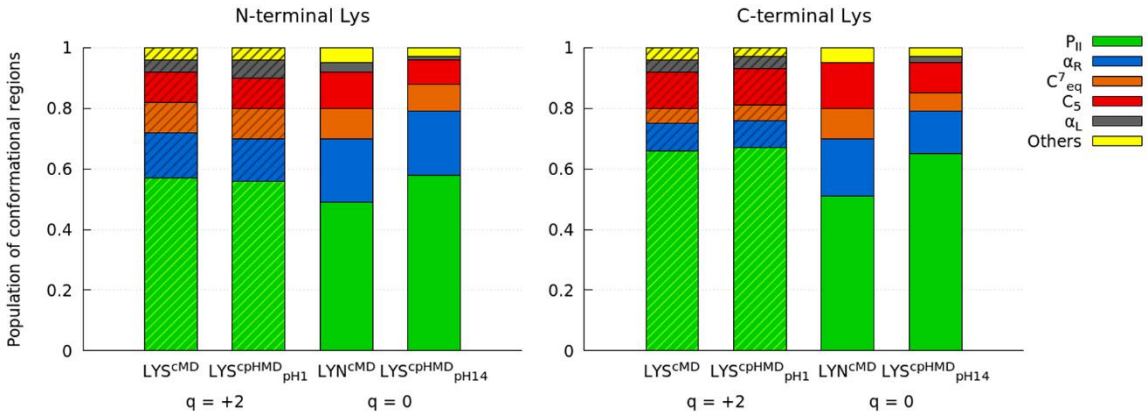


Figure C9. Populations of the conformational regions (P_{II}, α_R, C^7_{eq}, C_5, and α_L) in the Ramachandran maps of each amino acid of the capped Lys_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q). The striped and solid box represent the protonated and deprotonated states, respectively.

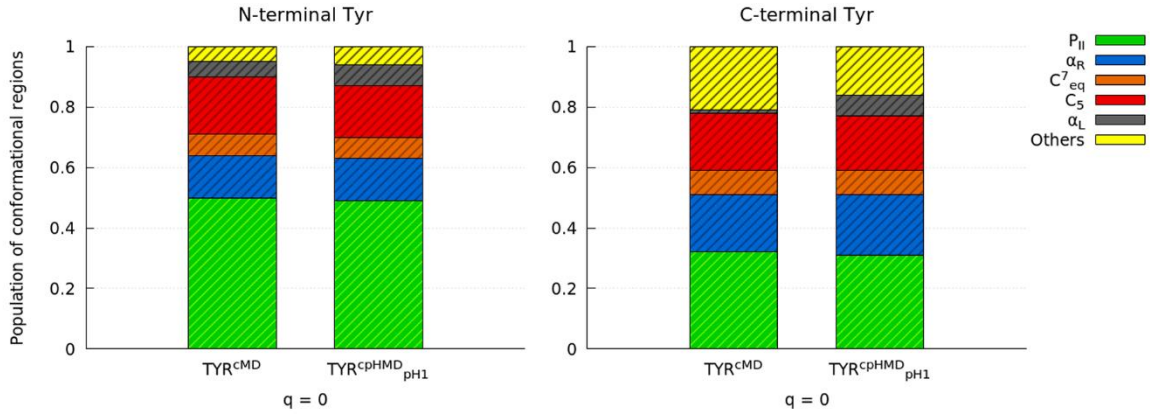


Figure C10. Populations of the conformational regions (P_{II}, α_R, C^7_{eq}, C_5, and α_L) in the Ramachandran maps of each amino acid of the capped Tyr_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q).

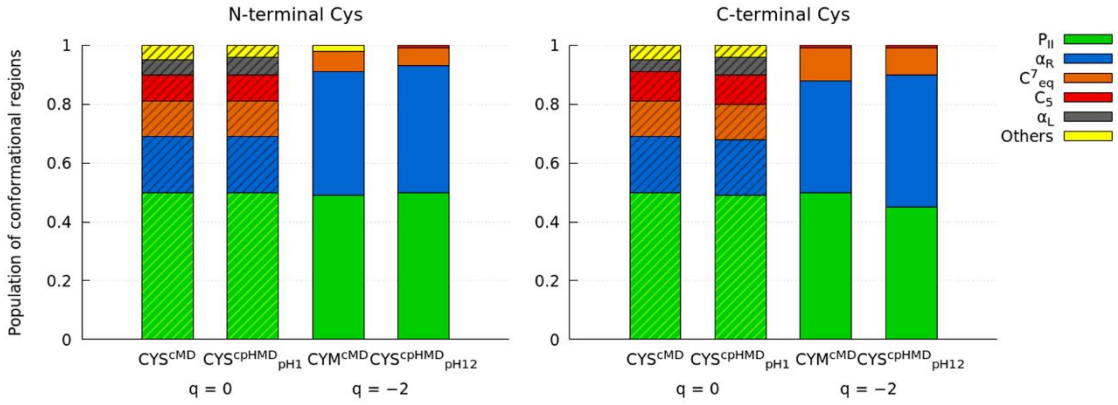


Figure C11. Populations of the conformational regions (P_{II}, α_R, C^7_{eq}, C_5, and α_L) in the Ramachandran maps of each amino acid of the capped Cys_2 tripeptide. The titles indicate the residues with the simulation method and the solvent pH in superscript and subscript, respectively. The net charge of the tripeptide is shown below (q). The striped and solid style represent the protonated and deprotonated states, respectively.

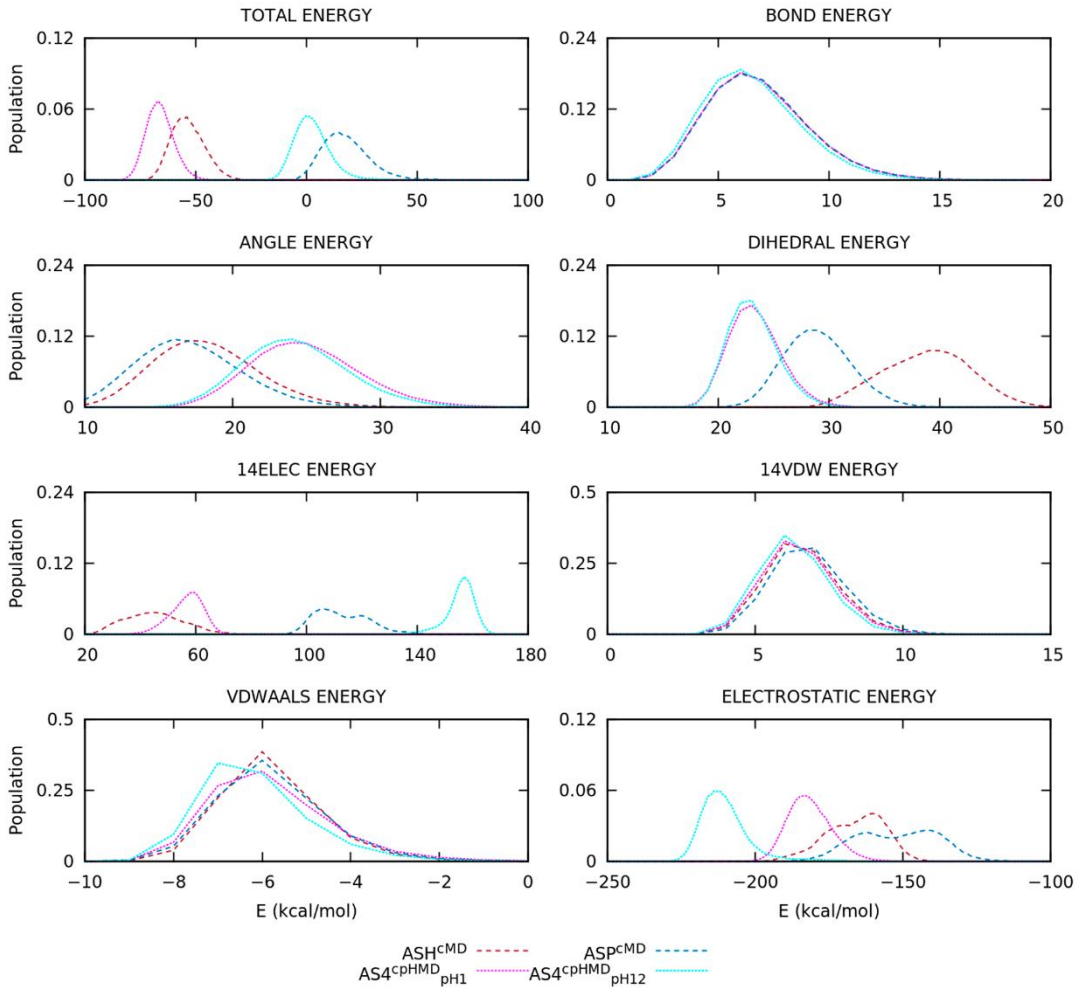


Figure C12. Energy distributions of the capped Asp_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

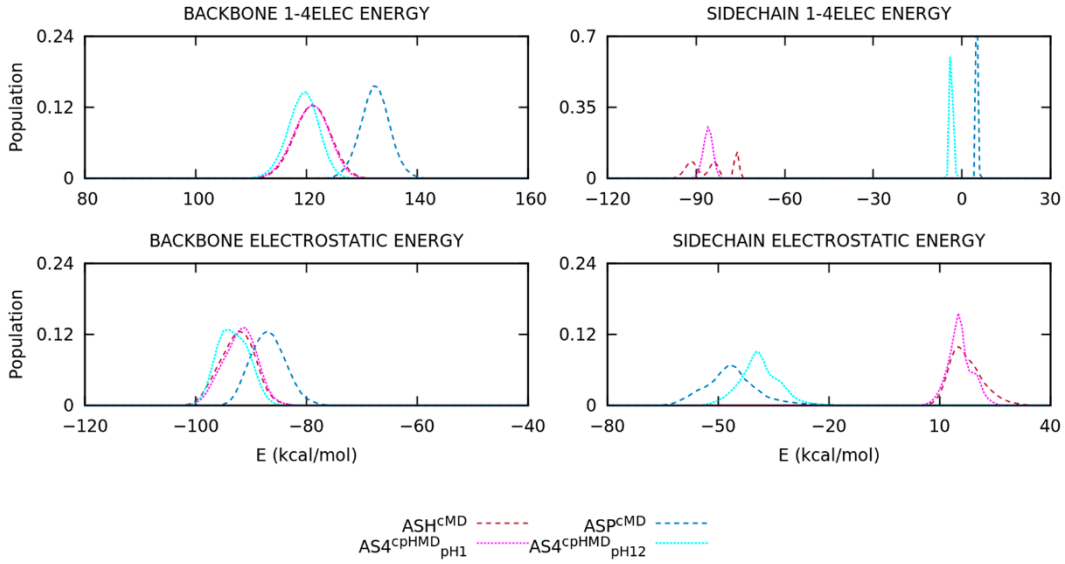


Figure C13. Energy distribution of the 1-4 and long-range electrostatics capped Asp_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

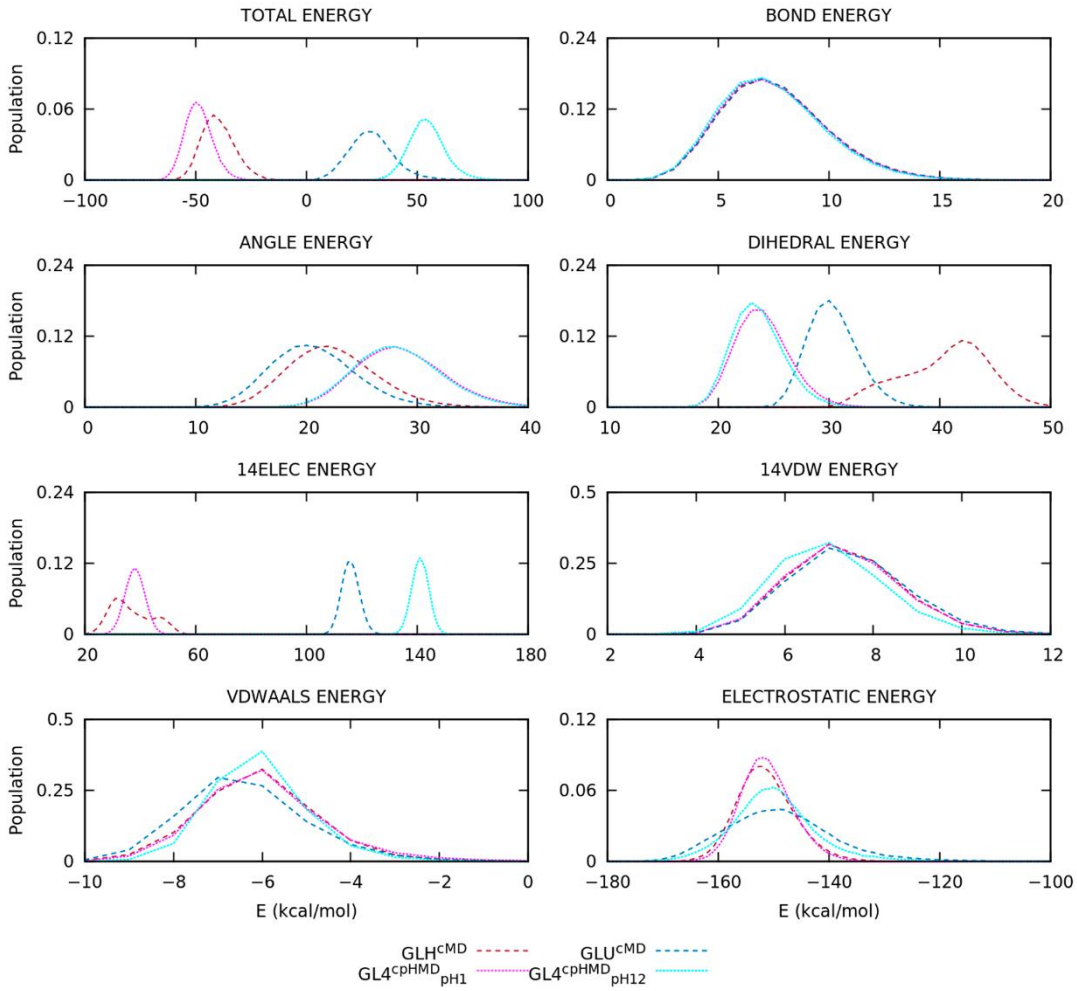


Figure C14. Energy distributions of the capped Glu_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

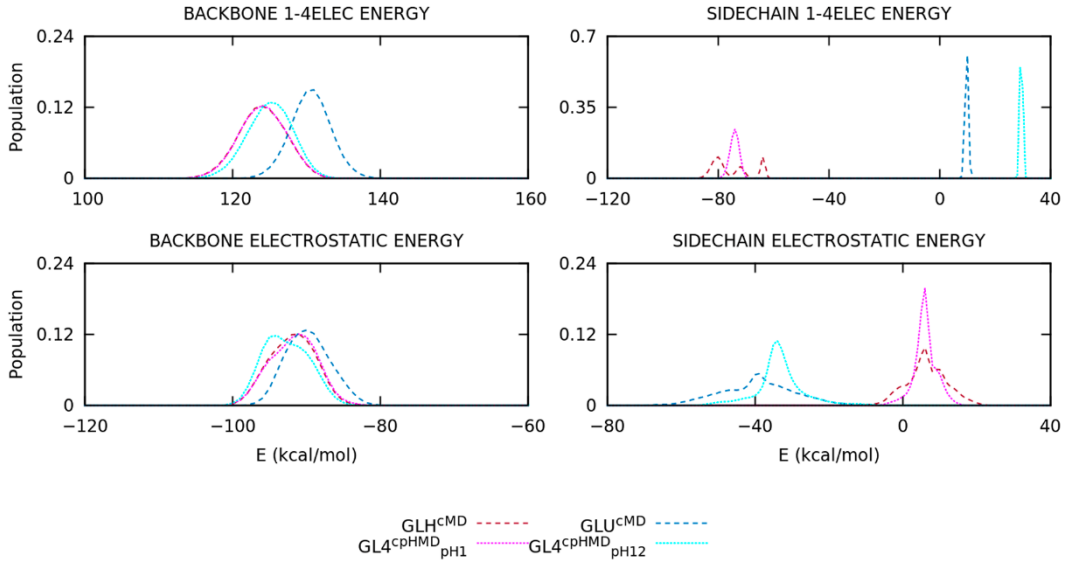


Figure C15. Energy distribution of the 1-4 and long-range electrostatics capped Glu_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

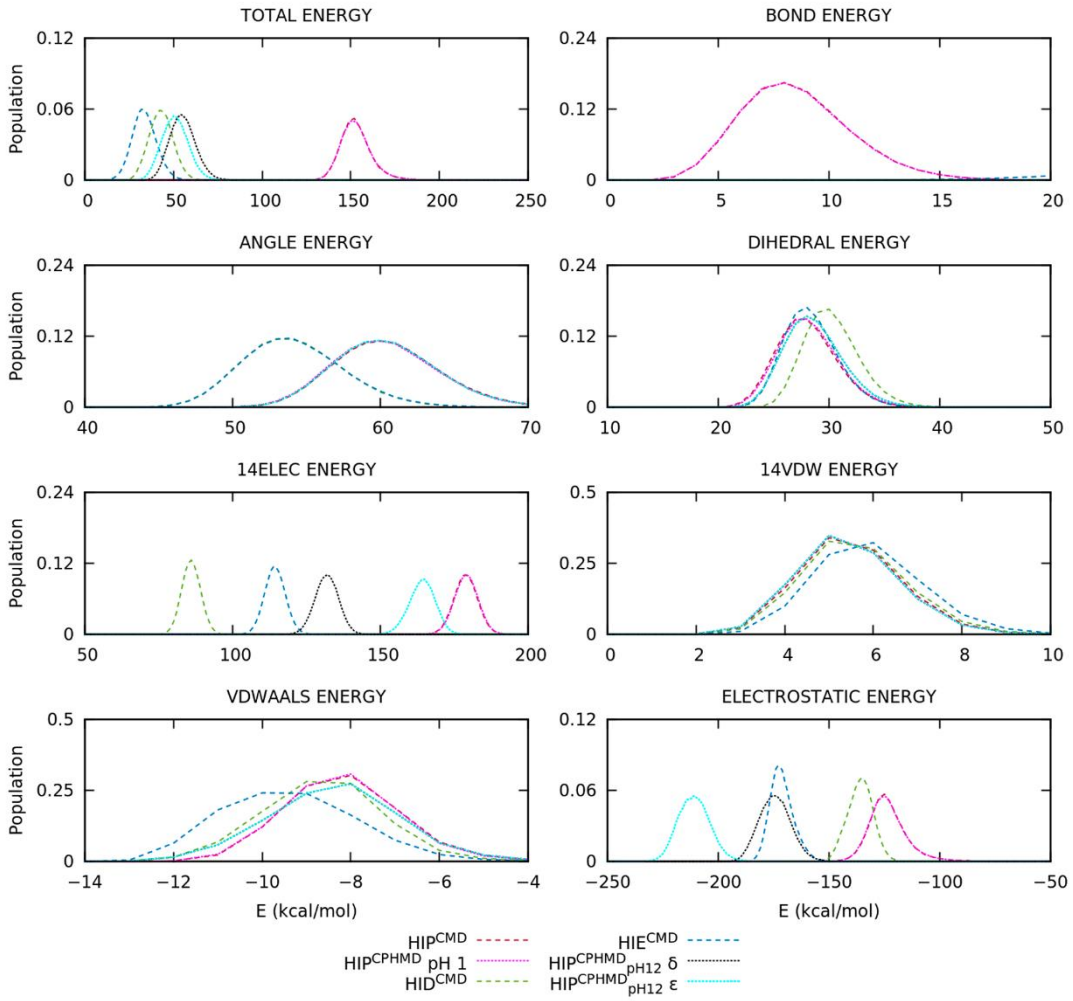


Figure C16. Energy distributions of the capped His_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulations, respectively. $HIP^{CpHMD}{}_{pH12}\ \delta$ and $HIP^{CpHMD}{}_{pH12}\ \varepsilon$ are the energy distributions calculated using partial charges fixed on the δ and ε protonation state.

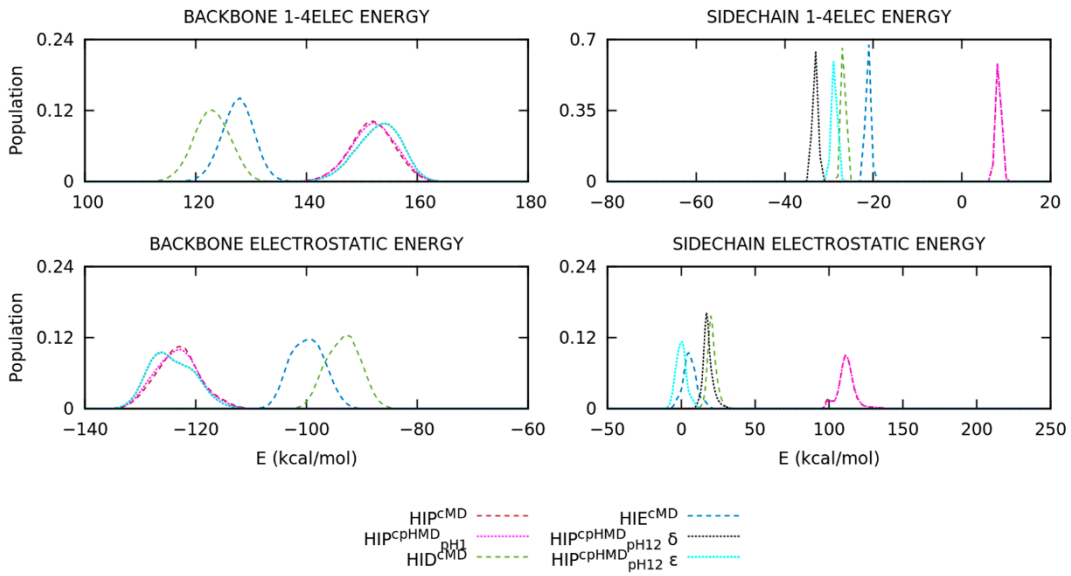


Figure C17. Energy distribution of the 1-4 and long-range electrostatics capped His_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulations, respectively. $HIP^{CpHMD}{}_{pH12}\ \delta$ and $HIP^{cpHMD}{}_{pH12}\ \varepsilon$ are the energy distributions calculated using partial charges fixed on the δ and ε protonation state.

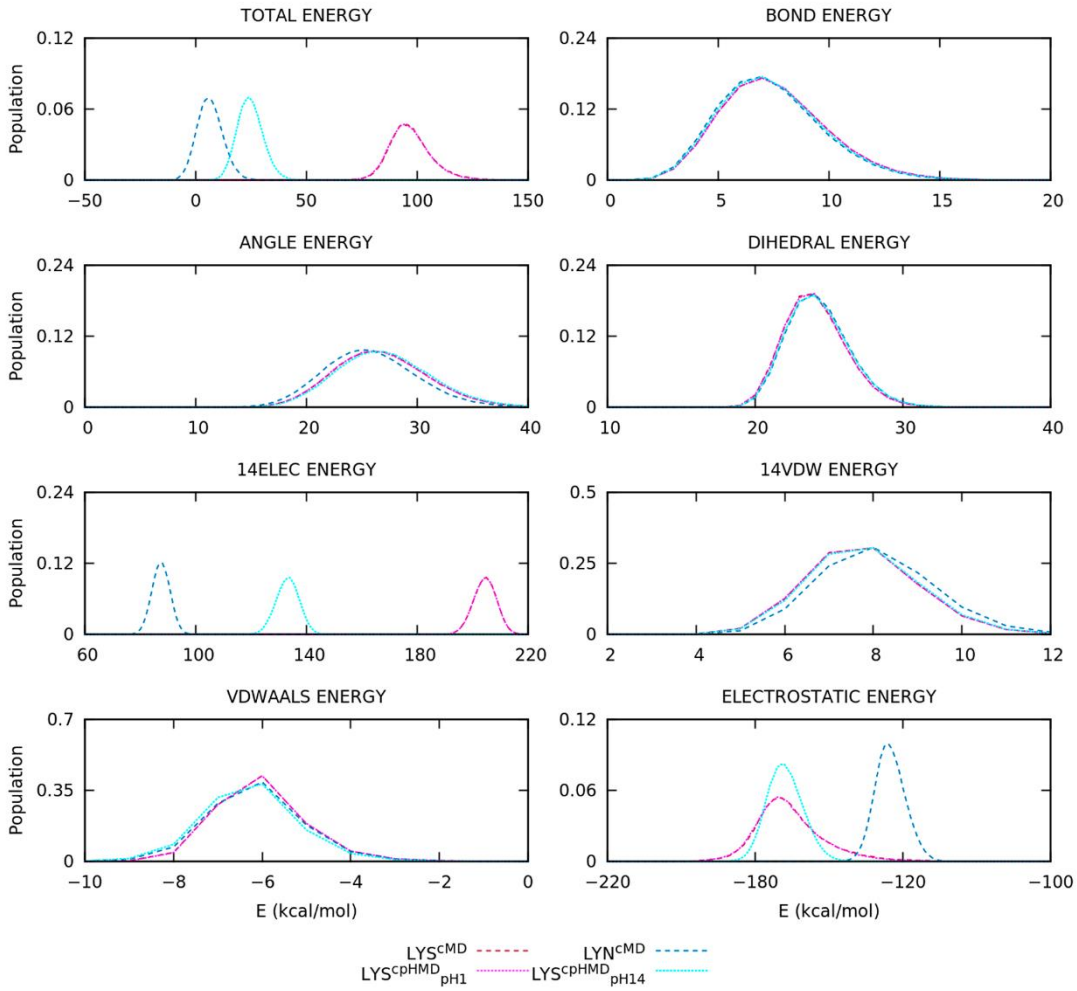


Figure C18. Energy distributions of the capped Lys_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

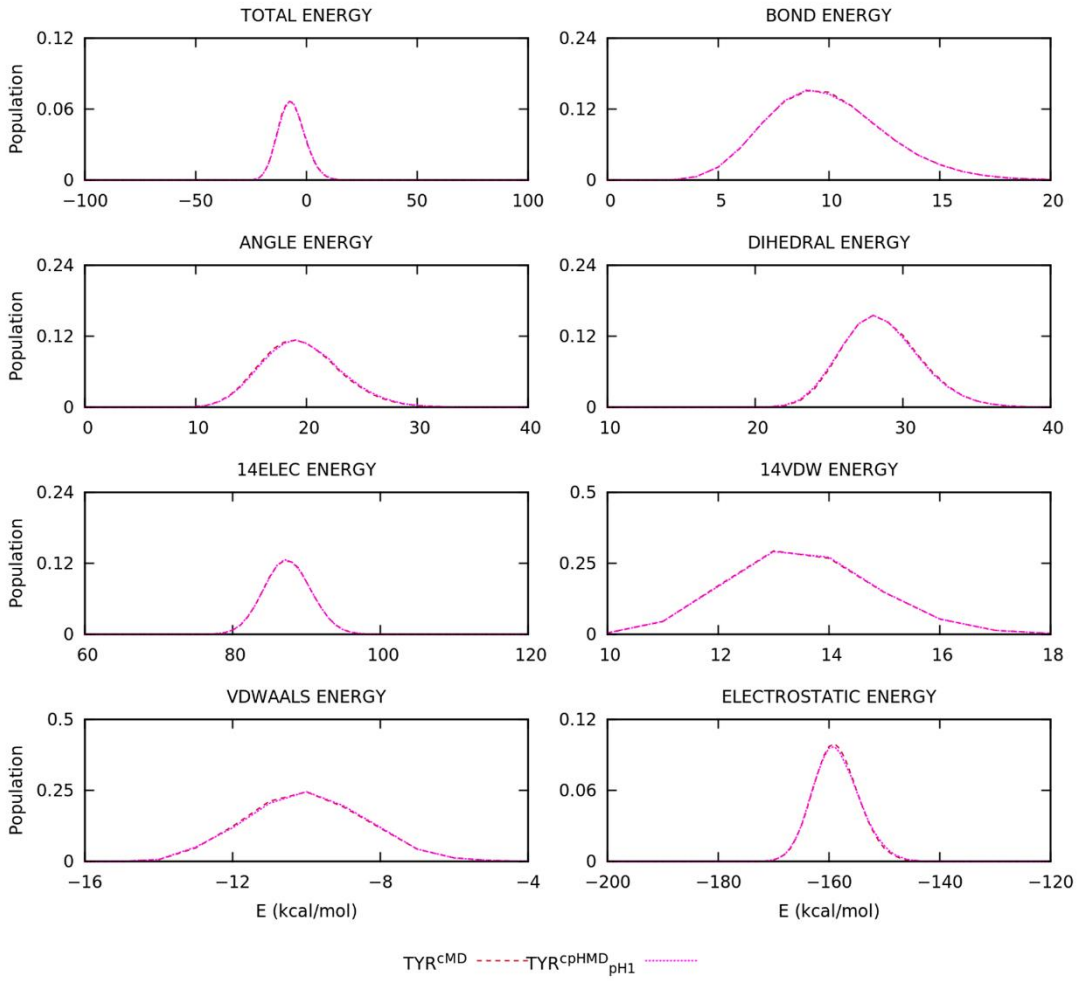


Figure C19. Energy distributions of the capped Tyr_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

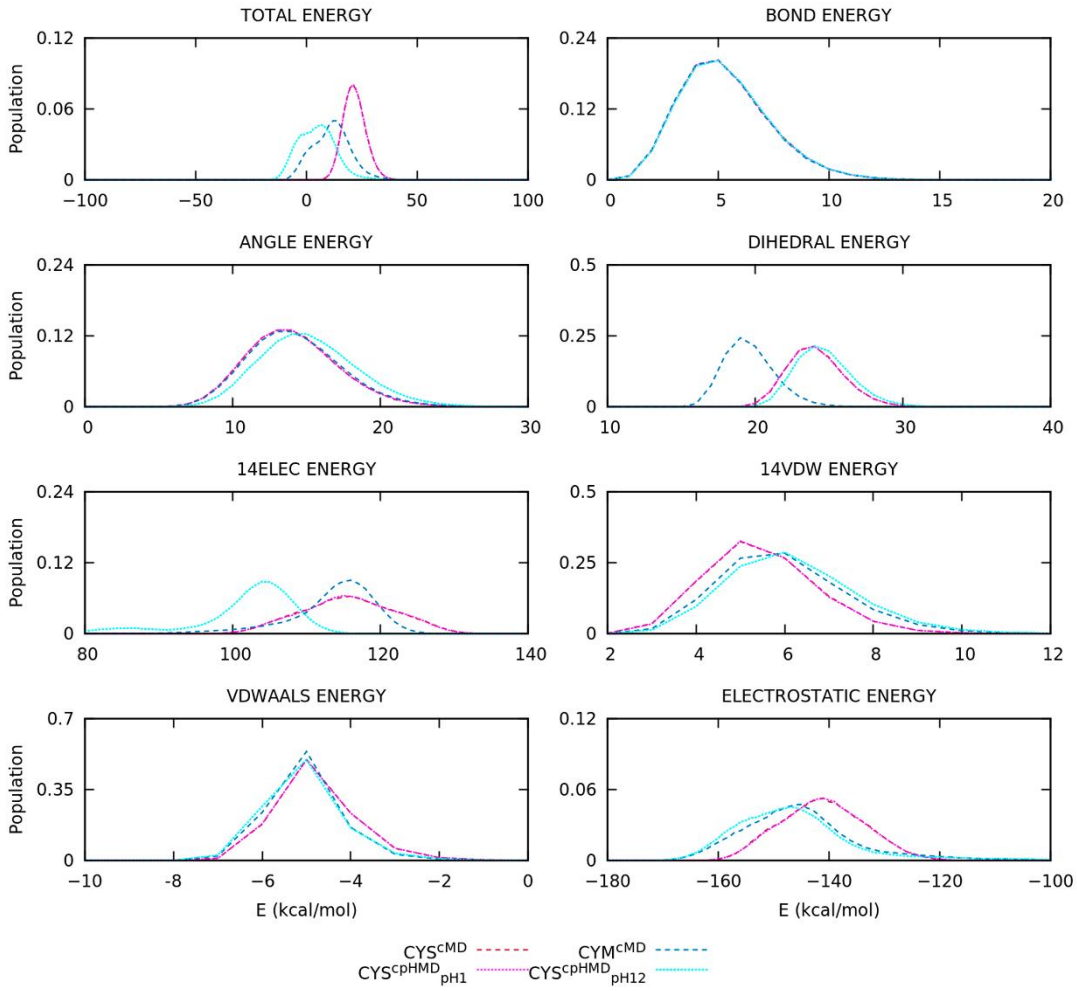


Figure C20. Energy distributions of the capped Cys_2 tripeptide without solvent molecules. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

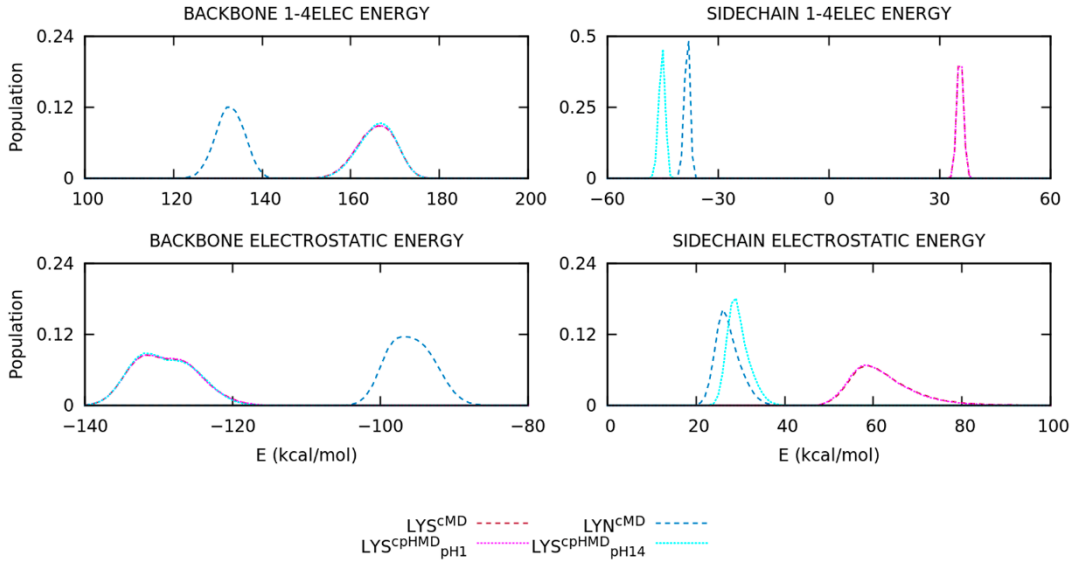


Figure C21. Energy distribution of the 1-4 and long-range electrostatics capped Lys_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

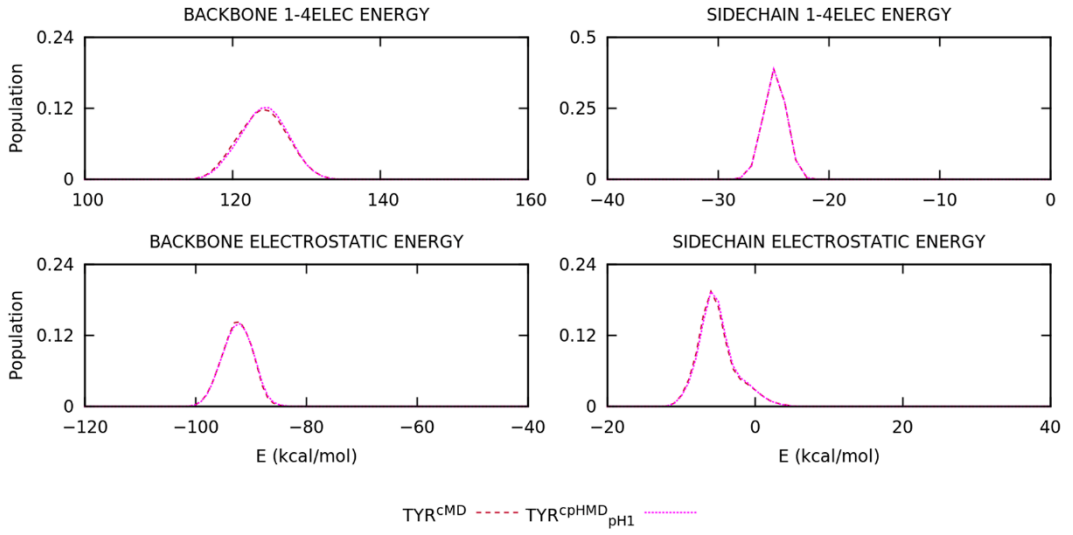


Figure C22. Energy distribution of the 1-4 and long-range electrostatics capped Tyr_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

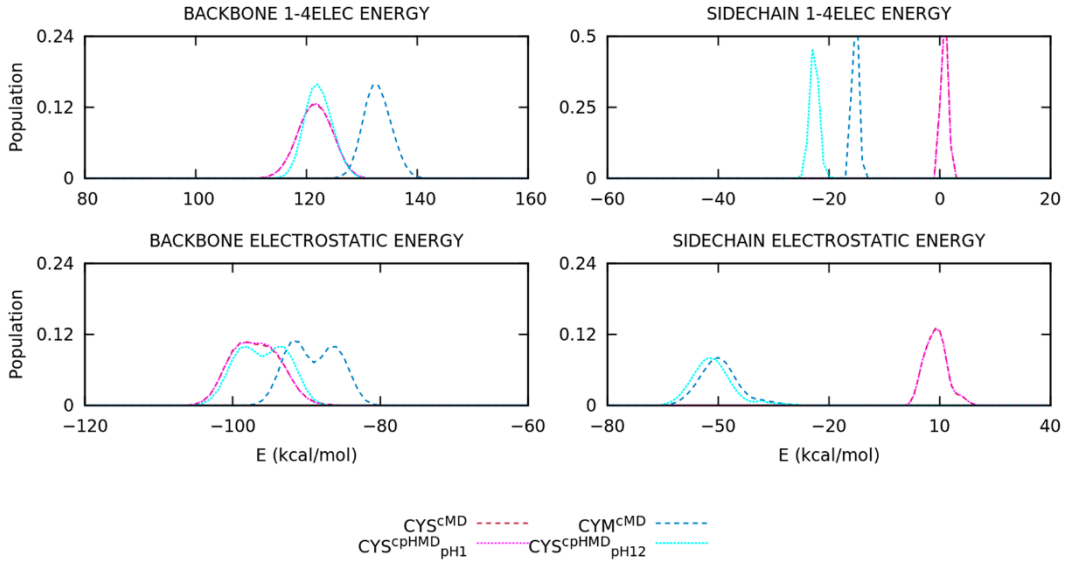


Figure C23. Energy distribution of the 1-4 and long-range electrostatics capped Cys_2 tripeptide divided into backbone and side chain atoms. The dotted and dashed lines are the CpHMD and CMD simulations, respectively.

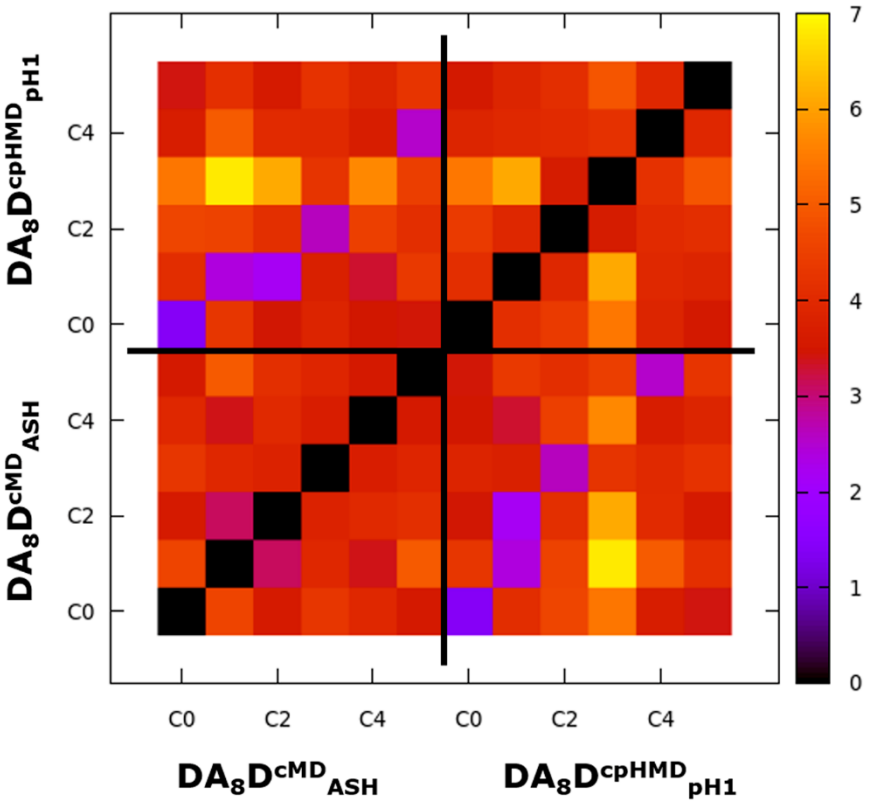


Figure C24. 2D-RMSD map of the first six representative conformations of the DA_8D peptide in the protonated form ($DA_8D^{CMD}{}_{ASH}$ and $DA_8D^{CpHMD}{}_{pH1}$). The RMSD is calculated using the C_α atoms of the peptides.

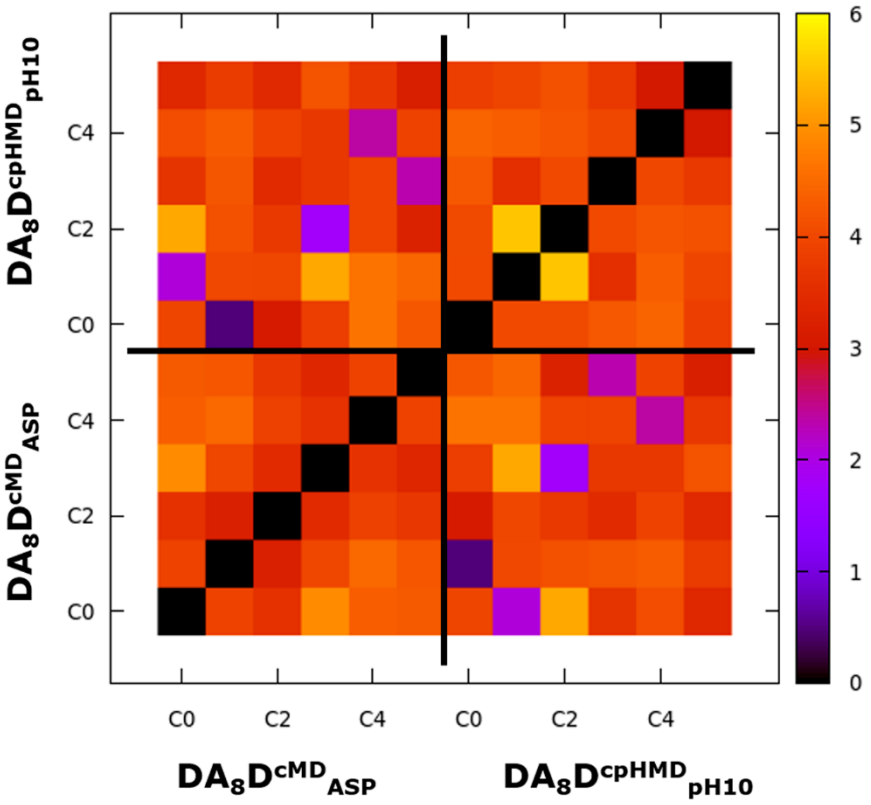


Figure C25. 2D-RMSD map of the first six representative conformations of the DA_8D peptide in the deprotonated form ($DA_8D^{CMD}{}_{ASP}$ and $DA_8D^{CpHMD}{}_{pH10}$). The RMSD is calculated using the C_α atoms of the peptides.

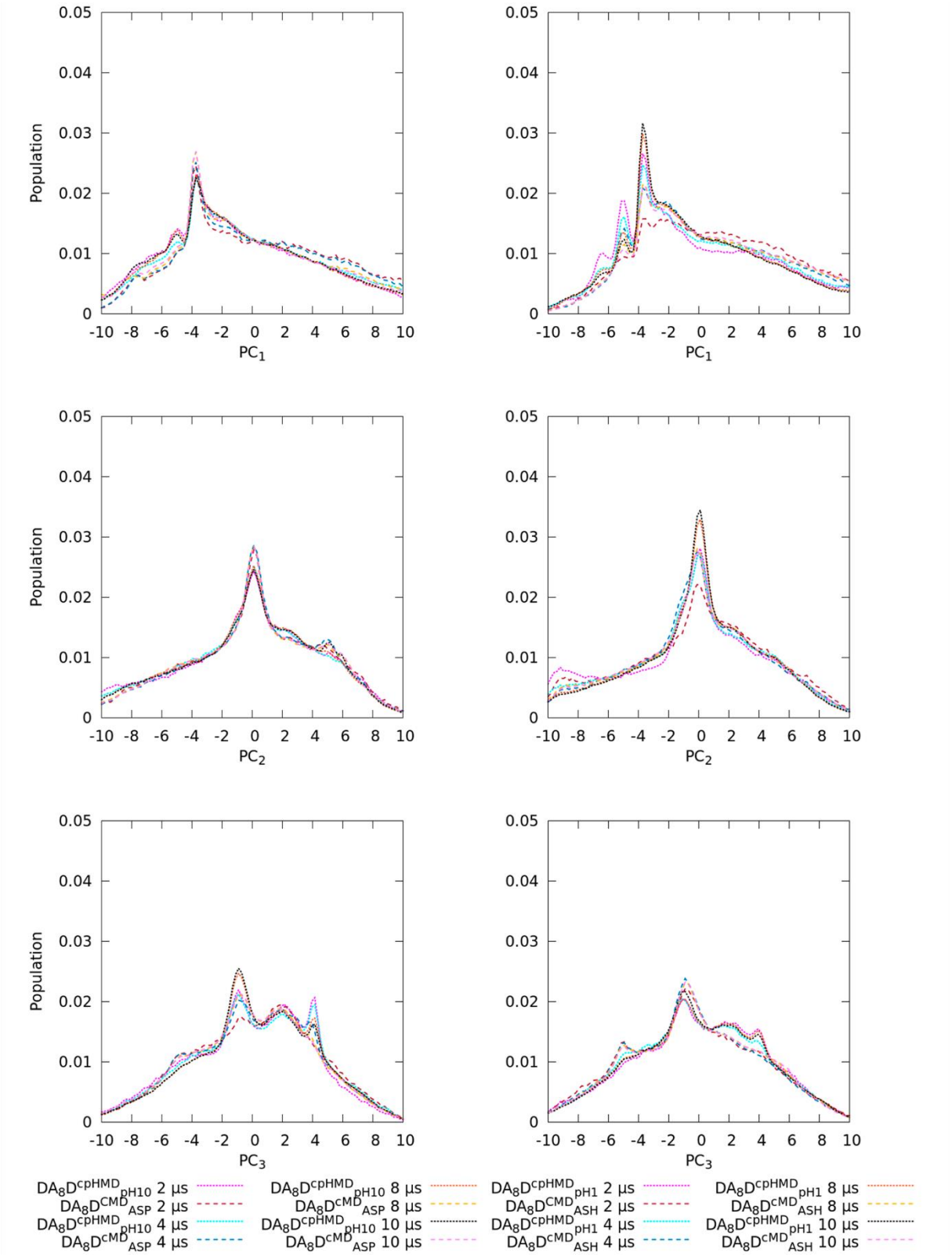


Figure C26. Distribution of the first three PCs at different simulation times (2, 4, 8 and 10 µs) of the DA_8D peptide. The deprotonated and protonated form are on the left and right, respectively. The dotted and dashed lines are the CpHMD and CMD simulation methods, respectively.

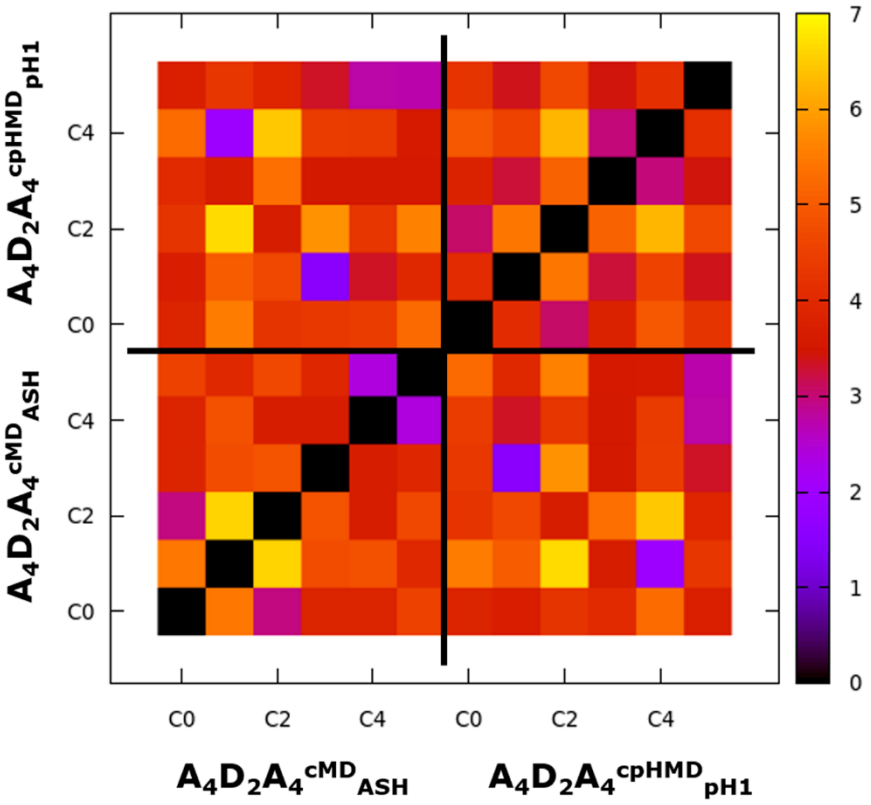


Figure C27. 2D-RMSD map of the first six representative conformations of the $A_4D_2A_4$ peptide in the protonated form ($A_4D_2A_4^{CMD}{}_{ASH}$ and $A_4D_2A_4^{CpHMD}{}_{pH1}$). The RMSD is calculated using the C_α atoms of the peptides.

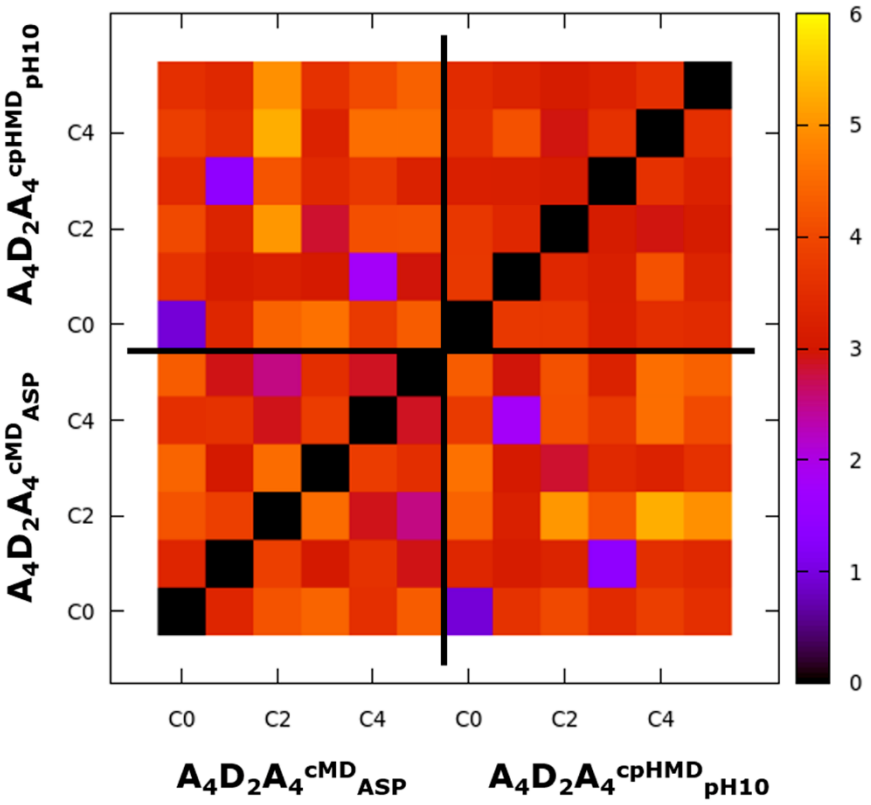


Figure C28. 2D-RMSD map of the first six representative conformations of the $A_4D_2A_4$ peptide in the deprotonated form ($A_4D_2A_4^{CMD}{}_{ASP}$ and $A_4D_2A_4^{CpHMD}{}_{pH10}$). The RMSD is calculated using the C_α atoms of the peptides.

Bibliography

1. Rubio-Martinez, J.; Tomas, M.S.; Perez, J.J. Effect of the solvent on the conformational behavior of the alanine dipeptide deduced from MD simulations. *J. Mol. Graph. Model.* **2017**, *78*, 118–128, doi:10.1016/j.jmgm.2017.10.005.

Appendix D

Supporting Information to "Extensive Conformational Sampling of the Intrinsically Disordered Protein Histatin-5 Using Coarse-Grained and All-Atom Force Fields and Constant pH Molecular Dynamics"

MODEL	C1		C2		C3		C4		DBI	psF	SSR/SST
	%	distC	%	distC	%	distC	%	distC			
SBW3	47	3.8	20	4.1	15	3.7	6	4.3	2.0	100717	0.41
SBW3pH	20	3.4	15	4.0	14	4.2	13	4.5	1.9	144434	0.50
IDPW3pH	34	4.7	24	4.4	12	3.7	5	4.4	1.9	167350	0.54
IDPW3pH^{R}	14	4.5	14	4.4	11	4.5	10	4.6	1.9	52643	0.56
IDPW4DpH	25	4.7	21	4.5	11	4.3	8	4.0	1.7	66039	0.61
SRH	34	1.4	15	1.5	11	1.4	10	1.4	1.7	42779	0.60
SRH^{R}	17	3.5	12	3.4	10	3.2	9	3.0	1.2	9503	0.70
SGR	21	6.1	20	6.0	19	6.0	17	6.1	2.2	1835	0.39
SGRpH	38	6.4	18	6.1	13	5.8	9	5.7	2.1	1717	0.38

Table D1. Populations and distances-to-centroid of the four most populated clusters (C0, C1, C2, and C3). The David-B, pseudo-F and SSR/SST indices of each clustering are also shown in the table.

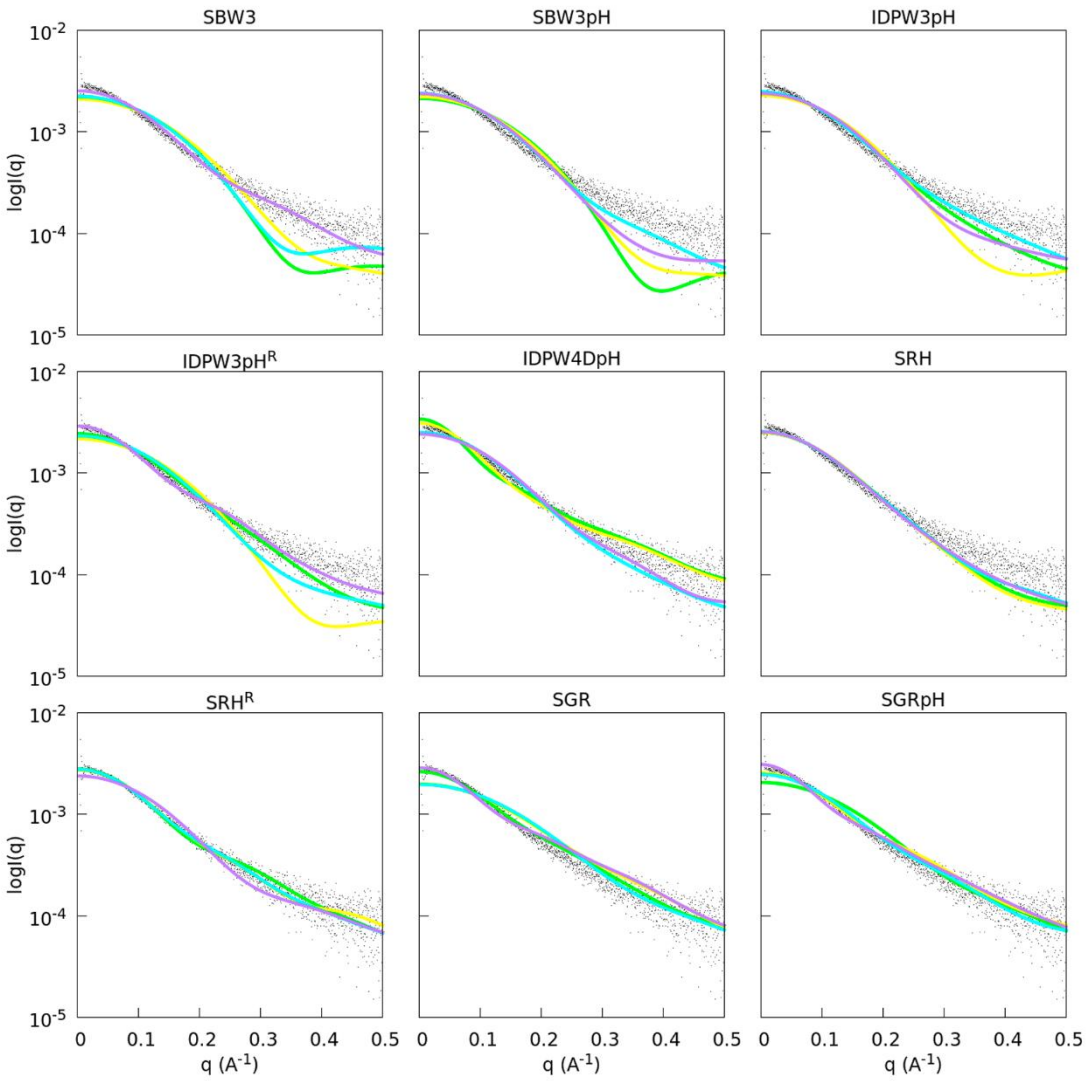


Figure D1. Theoretical SAXS intensity profiles for the three most populated clusters (red, green, and blue lines for clusters C0, C1, and C2, respectively) fitted to the experimental SASDHH8 scattering (black dots).

Appendix E

Supporting Information to "Molecular Dynamics Simulation of α-Synuclein NAC Domain Fragment with ff14IDPSFF IDP-specific Force Field Suggest β-Sheets Intermediate State for Fibrillation"

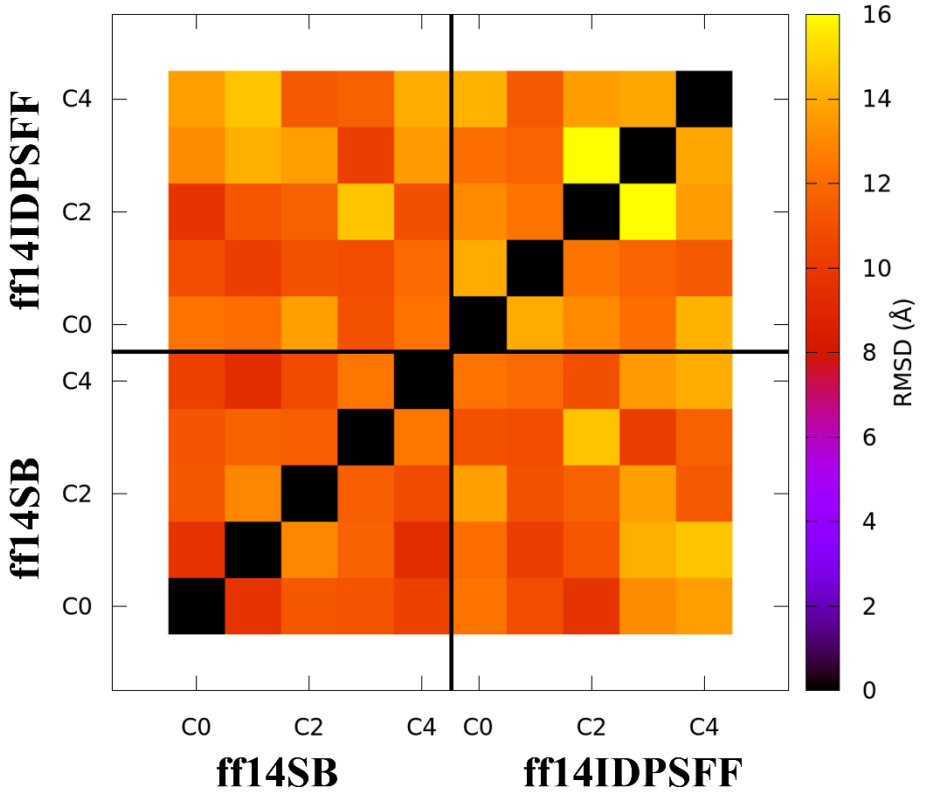


Figure E1. 2D-RMSD map of the representative conformations of the five most populated clusters of EαSNAC using the ff14SB and ff14IDPSFF force fields. The RMSD is calculated with the C_α atoms of the peptides.

	ff14SB	**ff14IDPSFF**
$\mathbf{\Delta d_{c0}^{centroid}}$ **(Å)**	4.2	6.1
$\mathbf{\Delta d_{c1}^{centroid}}$ **(Å)**	4.4	5.7
$\mathbf{\Delta d_{c2}^{centroid}}$ **(Å)**	4.7	6.9
$\mathbf{\Delta d_{c3}^{centroid}}$ **(Å)**	5.2	5.5
$\mathbf{\Delta d_{c4}^{centroid}}$ **(Å)**	4.1	5.6
DBI	1.33	1.37
psF	40985	33522
SSR/SST	0.657	0.610

Table E1. Clustering indicators of the ff14SB and ff14IDPSFF simulations. $\Delta d_{cX}^{centroid}$, DBI, psF and SSR/SST magntiudes are the average distance-to-centroid of cluster X, the Davis-Bouldin index, the pseudo-statistic F and the sum of squares regression/sum of squares total, respectively.

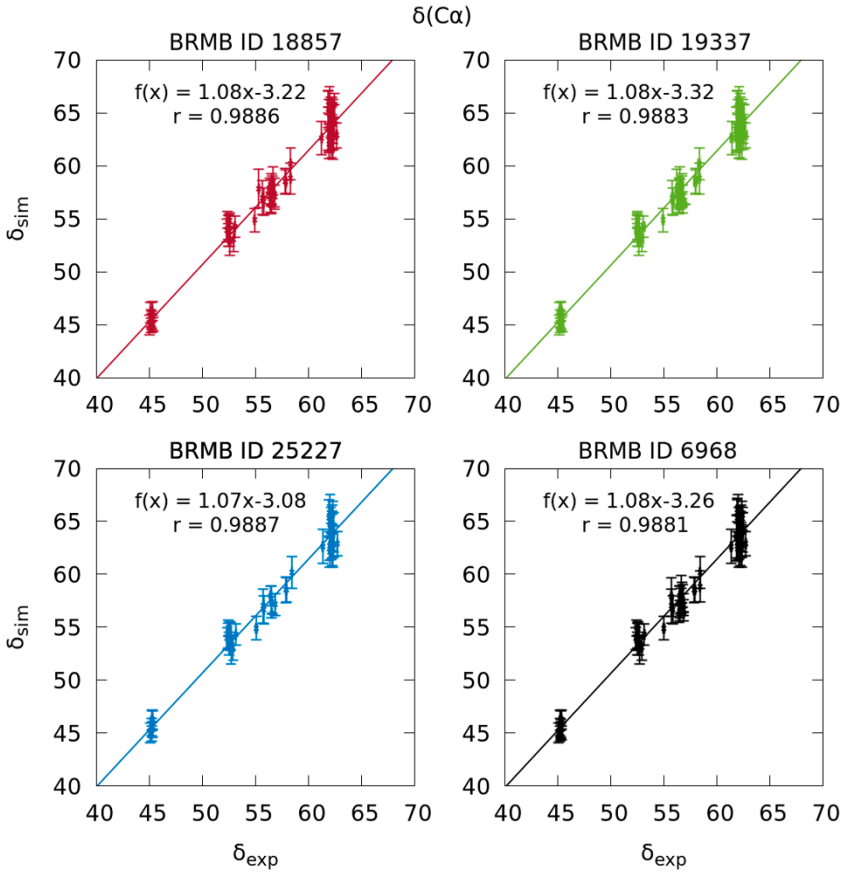


Figure E2. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C_α atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14SB simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

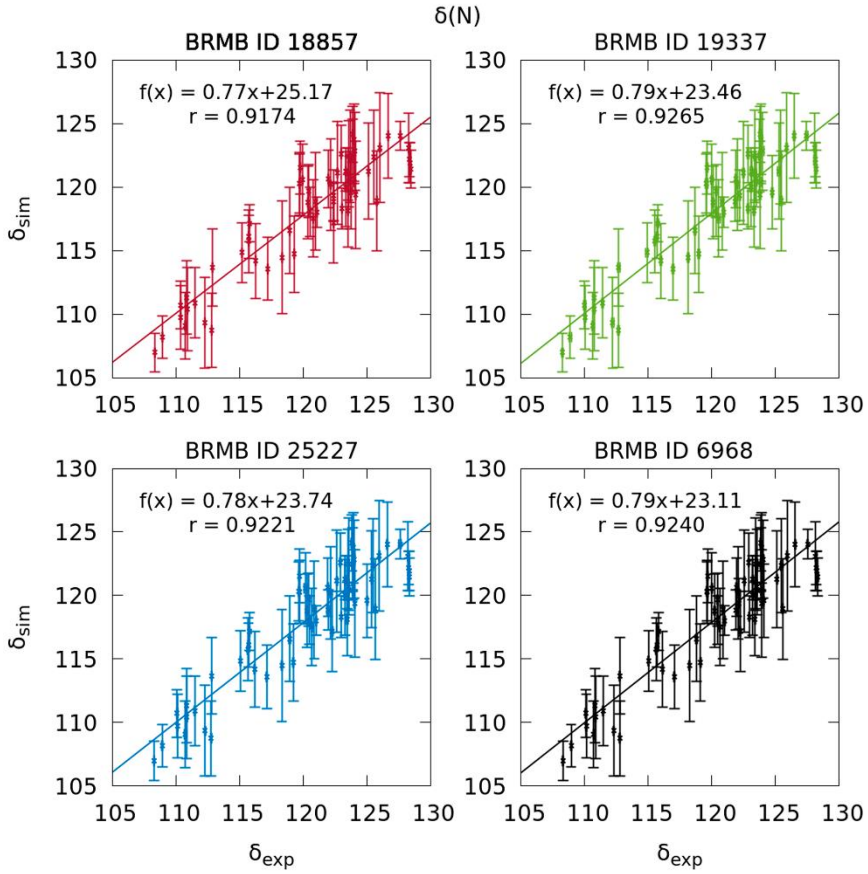


Figure E3. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the N atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14SB simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

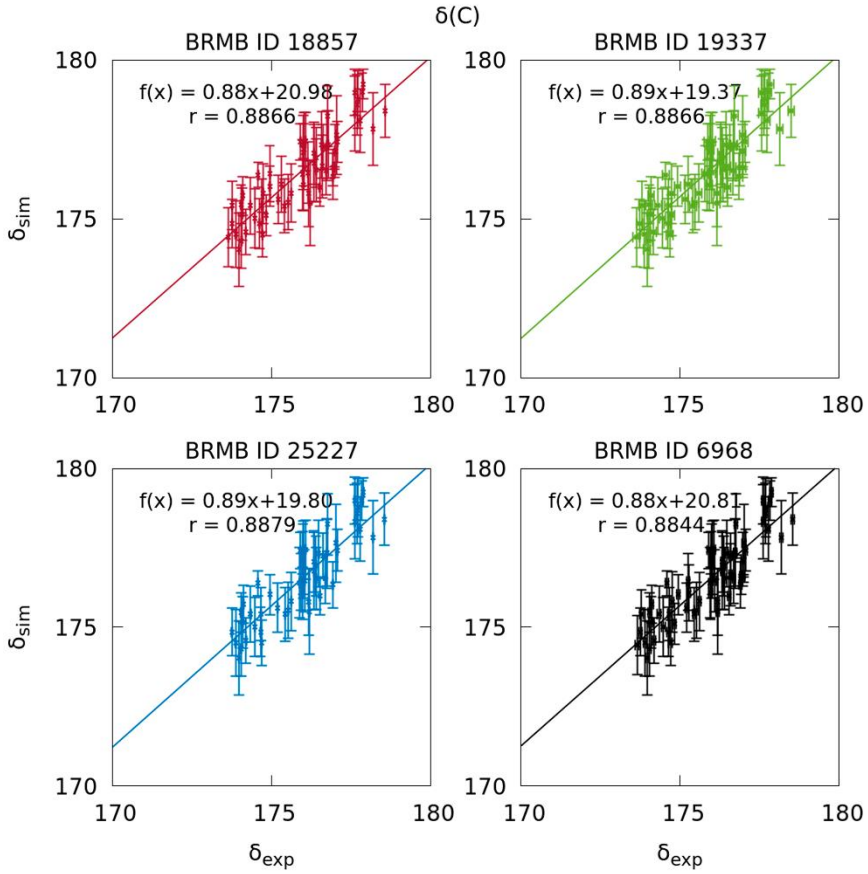


Figure E4. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14SB simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

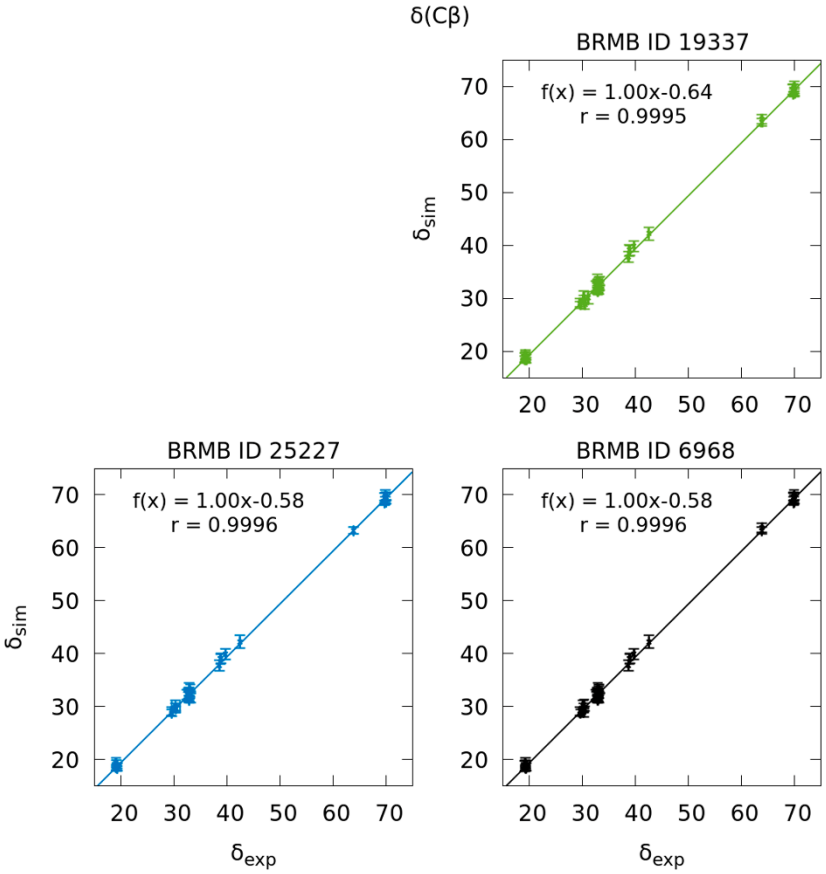


Figure E5. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C_β atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14SB simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

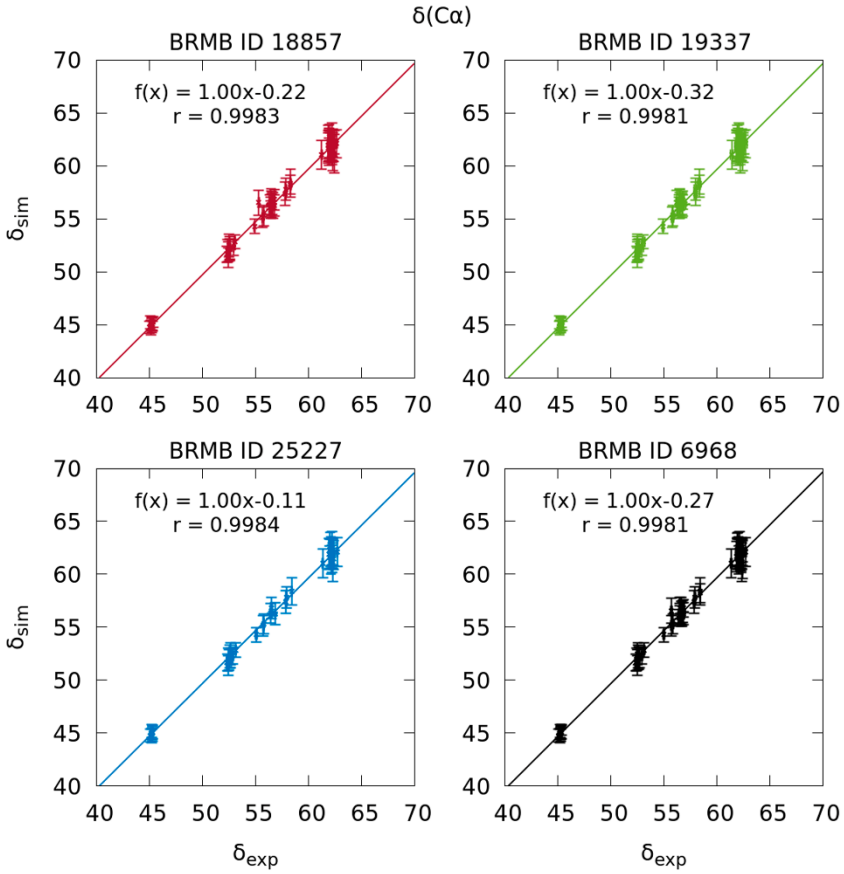


Figure E6. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C_α atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14IDPSFF simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

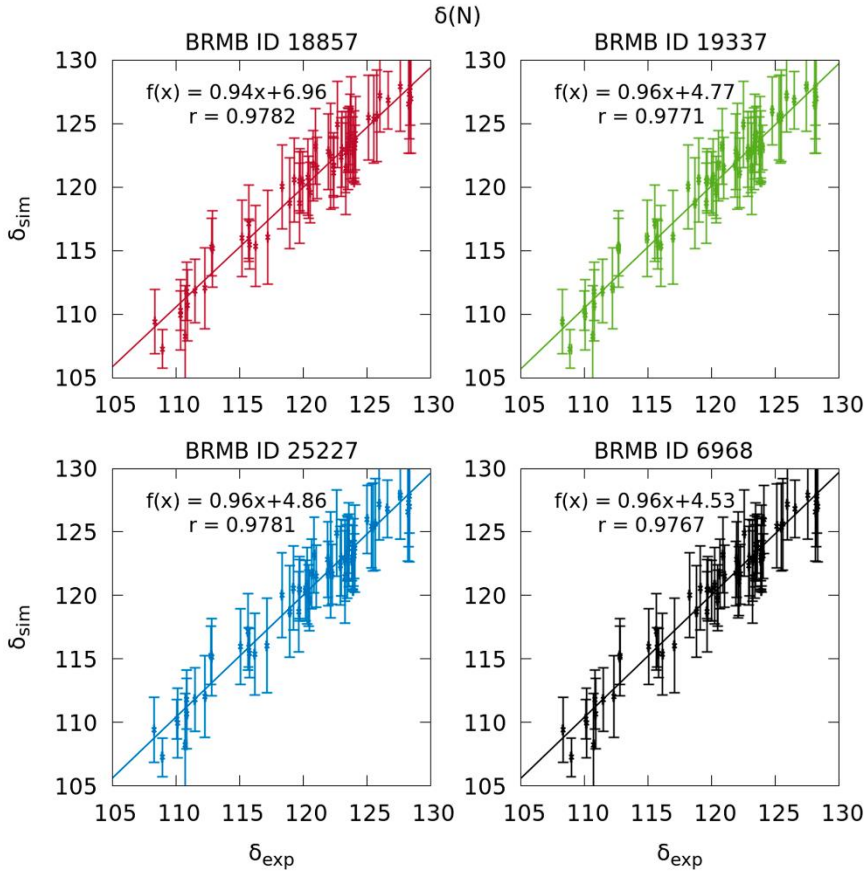


Figure E7. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the N atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14IDPSFF simulation. The deviation of the measured and predicted chemical shifts are represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

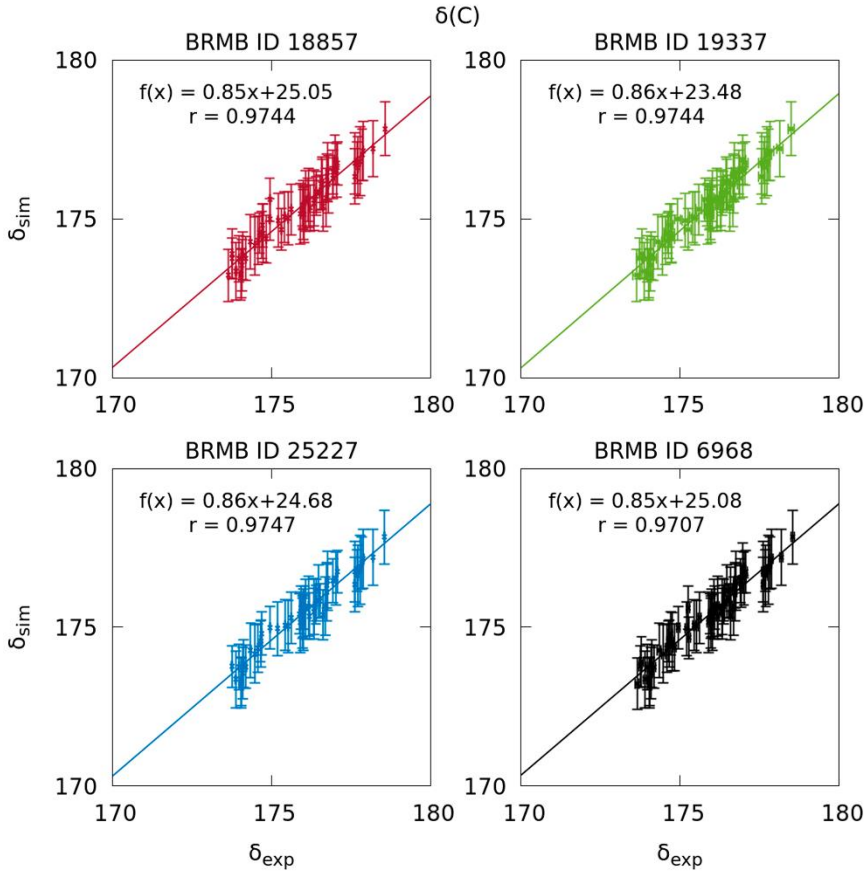


Figure E8. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14IDPSFF simulation. The deviation of the measured and predicted chemical shifts are represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

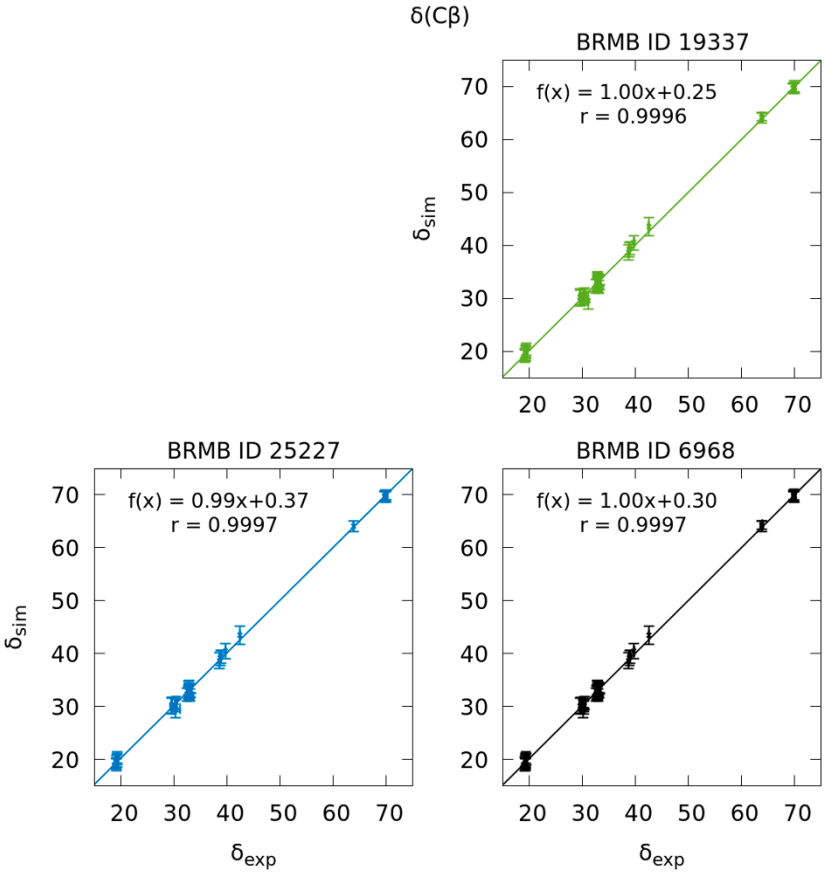


Figure E9. Correlation between the measured (δ_{exp}) and predicted (δ_{sim}) chemical shifts of the C_β atom for the four BRMB data sets (18857, 19337, 25227 and 6968) in the ff14IDPSFF simulation. The deviation of the measured and predicted chemical shifts is represented with vertical and horizontal error bars, respectively. The linear equations obtained by fitting the chemical shift data and the Pearson correlation coefficient (r) are also shown in the plot.

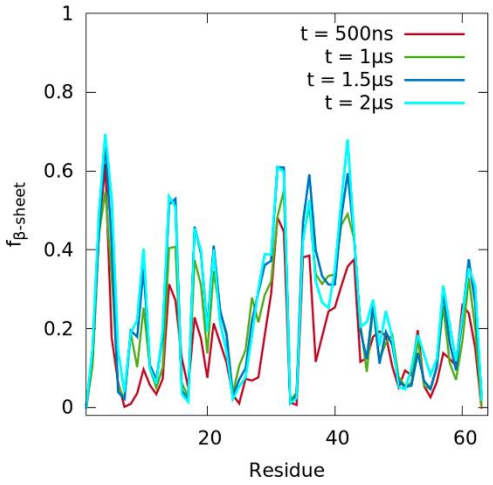


Figure E10. Fraction of β-sheet content at four simulation times (0.5, 1, 1.5 and 2 μs) of EαSNAC in the ff14IDPSFF simulation. The β-sheet content is defined as the sum of anti-parallel and parallel β-sheets, β-bulges and isolated β-strands propensities determined by the DSSP method.