Development of novel computational methods suitable for modelling intrinsically disordered proteins

This thesis is submitted to the School of Natural and Environmental Sciences, Newcastle University for the degree of Doctor of Philosophy

João Victor de Souza Cunha

March 2020

Abstract

Proteins without a stable tertiary structure are known as intrinsically disordered or metamorphic. These proteins denoted as IDPs – or protein domains denoted as IDRs – exert crucial roles in cellular signalling, growth and molecular recognition events. Due to their high plasticity, IDPs and IDRs are very challenging for experimental and computational structural studies. To enable these, all-atom molecular dynamics (MD) simulations are used, as they provide insight into structure and dynamics at the atomistic level of detail. However, the current generalist physical models (protein force fields and solvent models) used in MD simulations are unable to generate satisfactory ensembles for IDPs/IDRs when compared to existing experimental data.

This work aimed to improve on the state-of-the-art accuracy for simulations of IDPs/IDRs without sacrificing accuracy for folded domains. Herein, the accuracy of several different force fields frequently used for simulations of proteins was compared, in simulations of both ordered and disordered systems. The results showed that each force field has strengths and limitations.

Given the fact that interactions with the solvent are pivotal for accurate simulations of intrinsically disordered proteins, a novel solvation model was developed, denoted as Charge-Augmented 3 Point water model for Intrinsically disordered Proteins (CAIPi3P). CAIPi3P model was generated through systematic scanning of the dipole moment values calculated for the popular TIP3P three-point water model. By increasing the dipole magnitude, the agreement between experimental and calculated small-angle X-ray scattering (SAXS) curves was massively improved for a series of model IDRs.

To further improve the simulations of proteins containing IDRs, a novel method to assemble force field parameters has been developed. Denoted as Hybrid_FF, it merges parameters from different established force-fields, performing well for structured and disordered regions (AMBER99SB-ILDN and AMBER03ws, respectively), parametrising each secondary structure differently. Testing these joint parameters for a series of IDR-containing proteins showed that such an approach improved the accuracy of the sampled configurations for long disordered regions.

Finally, a software to estimate and analyse the transition dynamics of intrinsically disordered regions has been developed in this work. Named structural quantifier of entropy (SQuE), it uses a first-order approximation to the probability distribution to assess the structural entropy for protein transitions barriers. It is expected that tools developed in this study will generate more accurate IDP/IDR ensembles, broadening the range of biologically relevant systems amenable to atomistic molecular dynamics simulations.

Declaration

The research described within this thesis was performed between September 2016 and October 2019 in the Computational Medicinal Chemistry Laboratory, Bedson Building, School of Natural and Environmental Sciences, Newcastle University, Newcastle-upon-Tyne, UK, NE1 7RU.

All research described within this thesis is original in content and does not incorporate any material or ideas previously published or presented by other authors except where due reference is given.

No part of this thesis has been previously submitted for a degree, diploma or any other qualification at any other university.

Indiciunt Ioannes est?

Acknowledgements

Dear reader,

This is the area that I can talk about and thanks to all the amazing people that I met in these years since the start of my journey into PhD. Therefore, I will be a bit less formal in this section. Apologies in advance.

First, I would like to talk about my work colleagues. I talk a lot, and all of them withstood my monologues, laughed with me and helped me with my work, making my workhours an amazing period of my day. Quico, Danlin, Sylvia, Matt and Ayaz were there in my first two years; especially Quico, which was there also for all lunches we had in the falafel palace. *Sentire muy su faltita!*

Also, Mete, Piotr, Weikang and Ruidi, which joined Bronka's lab and I was there in different phases of their studies. Thanks for allowing me to be your lives and for the trust given. And last but certainly not least, I would like to thank Lanyu, that is making every day better since she started her PhD and has been working with me.

Several other people from the university I would like to thank as well: Mark Garner and all our collaborators from Chemistry, FMS and abroad. Jamie Gibson especially, which, along with being a good friend, always impressed me with his collection of hoodies.

To finalize, there are three people that I would like to thank in a special manner.

First, Paola Lanzoni, my beloved fiancée. Which was always there, giving me strength and love in all moments, and enjoying every brighter moment by my side, making every ray of sun a bit shinier and happy. I am glad to have you by my side as a partner and companion, I will be glad to have you by my side for the years to come.

Second, my supervisor and mentor, Dr Agnieszka K. Bronowska aka Bronka. If there was a moment in my life that I felt lucky was the moment that, beyond all odds, I crossed paths with Bronka. Her genius, limitless creativity, excitement for science, humour and patience made these three years the time that I evolved the most not only as a researcher but also as a human being. It's weird for me to talk seriously about Bronka. She makes science look fun, and every day in the office

is another fun day in my life. I don’t know how to thank her enough for all the opportunities for growth and evolution I had because of her, and I how much patience she had with my blabbing. Also, I would like to thank two sources of inspiration that I had through my PhD: Squee and Henia. Both experts in MD and computational chemistry that helped me a lot these years. Miau.

Third, the most important person in the world for me: My mom. Adriana’s teachings, since I was a baby Jojo, modelled my character and built my human consciousness. I am proud to say that I am her son, and I dedicated this work wholeheartedly to her. As I usually say, if one day I become a strong person such as her, I will feel more and more that I’m navigating towards the right path. All the sacrifices she made and the difficulties she endured on my behalf are actions that I will never forget. Obrigado minha mamae. I wish to extend this to my whole family, my dear sister Carol and my stepfather Kevin. Thanks to being there for me.

As a Brazilian that talks too much, I end my acknowledgements.

Thank you for your patience, my dear reader.

Abbreviations

A	A/Ala	Alanine
	Å	Angström
	AMBER	Assisted Model Building with Energetic Refinement
	Aβ	Amyloid β protein
	APP	Amyloid β-Protein Precursor
	AhR	Aryl hydrocarbon Receptor
	AD	Alzheimer's Disease
	AM1	Austin-Model 1
	AM1-BCC	Austin-Model 1 – Bond Charge Corrections
B	BoNT/A	Botulinum Serotype A
C	C/Cys	Cysteine
	CHARMM	Chemistry at HARvard Macromolecular Mechanics
	CryoEM	Cryo-Electron Microscopy
	CV	Collective Variable
	CG	Conjugate Gradient

	CD	Circular Dichroism
	CAIPi3P	Charge Augmented Three-point Water Model
	CNT	Clostridial Neurotoxin
	Cα	Carbon α
D	D/Asp	Aspartic acid
	Da	Dalton
	DBD	DNA-Binding Domain
	DFT	Density Functional Theory
	DNA	Deoxyribonucleic Acid
	DLB	Dementia with Lewy bodies
E	E/Glu	Glutamic acid
	EM	Electron Micrography
F	F/Phe	Phenylalanine
	FF	Force-Field
G	G/Gly	Glycine
	GB	Generalized-Born

	GBVI/WSA	Generalized-Born Volume Integral/Weighted Surface Area
	GH	Growth Hormone
	GPU	Graphics Processing Unit
	GROMACS	GROningen MAchine for Chemical Simulations
	GROMOS	GROningen Machine Simulation Package
H	H/His	Histidine
	HF	Hartree-Fock
	H-K	Hohenberg-Kohn
	HIF-α	Hypoxia-Inducible Factor α
I	I/Ile	Isoleucine
	IC_{50}	Half Maximal Inhibitory Concentration
	IDP	Intrinsically Disordered Protein
	IDR	Intrinsically Disordered Regions
K	K/Lys	Lysine
	kcal	Kilocalories
	K_d	Dissociation constant

L	L/Leu	Leucine
	L16	Loop 16
	LaP	La Protein
	LaRP	La Related Protein
	LaM	La Motif
	LJ	Lennard-Jones
	logP	Partition Coefficient
M	M/Met	Methionine
	MCSS	Multiple Copy Simultaneous Search
	MD	Molecular Dynamics
	MicroED	Microcrystal Electron Diffraction
	MAS	Magic Angle Spinning
	MM-PBSA	Molecular Mechanics - Poisson-Boltzmann Surface Area
	MSA	Multiple System Atrophy
	mRNA	Messenger Ribonucleic Acid
N	N/Asn	Asparagine

	nm	Nanometer
	NMA	Normal Mode Analysis
	NMR	Nuclear Magnetic Resonance
	NPT	Isothermal-Isobaric Ensemble
	ns	Nanosecond
	NM	Normal Modes
	NOE	Nuclear Overhauser Effect
	NVT	Canonical Ensemble
	NDDO	Neglect of Diatomic Differential Overlap
	NCOA1	Nuclear Receptor Coactivator 1
O	OPLS	Optimised Potential for Liquids Simulations
	OPC	Optimised Placed Charge
P	P/Pro	Proline
	PBC	Periodic Boundary Condition
	PC	Principal Component
	PCA	Principal Component Analysis
	PDB	Protein Data Bank

	PB	Poisson-Boltzmann
	PMF	Potential of Mean Force
	PM3	Parametrisation Method 3
	PM6	Parametrisation Method 6
	ps	Picosecond
	PTB1	Polypyrimidine Tract Binding Isoform 1
	PTM	Post-Translational Modification
	PHF	Paired Helical Filaments
	PD	Parkinson’s Disease
	P	Pressure
	PME	Particle Mesh Ewald
	PAS-B	Per-Arnt-Sim B
Q	Q/Gln	Glutamine
	QM	Quantum Mechanics
	QH	Quasi-Harmonic
R	R/Arg	Arginine
	RBD	Ras Binding Domain

	RMSD	Root-mean Square Deviation
	RMSF	Root-mean Square Fluctuation
	RSD	Root Squared Difference
	RRM	RNA Recognition Motif
S	S/Ser	Serine
	SPC	Single Point Charge
	SASA	Solvent Accessible Surface Area
	SAXS	Small Angle X-ray Scattering
	SCF	Self-Consistent Field
	SQuE	Structural Quantifier of Entropy
	SE	Semi-Empirical
	SD	Steepest Descent
	SARA	Smad Anchor for Receptor Activation
	SEC	Size Exclusion Chromatography
T	T/Thr	Threonine
	TCPTP	T-cell Protein Tyrosine Phosphatase
	TGFa	Transforming Growth Factor Alpha

	TIP3P	Transferable Intermolecular Potential with 3 Points
	TIPS	Transferable Intermolecular Potential functions
	TGFbeta	Transforming growth factor-beta
U	US	Umbrella Sampling
	UGDH	UDP-glucose 6-dehydrogenase
V	v/v	volume/volume
	V/Val	Valine
	VdW	Van der Waals
W	W/Trp	Tryptophan
	WHAM	Weighted Histogram Analysis Method
	WT	Wild Type
Y	Y/Tyr	Tyrosine

List of published works related to this work

De Souza, JV; Reznikov, S; Zhu, R; Bronowska, AK; **Druggability assessment of mammalian PAS domains using computational approaches**, Medicinal Chemistry Communications. 10,7,1126-1137.

De Souza, JV; Sabanes , FZ; Bronowska AK. 2019, **Development of Charge-Augmented Three-Point Water Model (CAIPi3P) for Accurate Simulations of Intrinsically Disordered Proteins**, Under revision. Available on: https://chemrxiv.org/articles/Development_of_Charge-Augmented_Three-Point_Water_Model_CAIPi3P_for_Accurate_Simulations_of_Intrinsically_Disordered_Proteins/7706867

De Souza, JV; Bronowska AK. **Long-Range Entropic Effects on Protein Intrinsically disordered regions.** Submitted. Available on: https://chemrxiv.org/articles/Long-Range_Entropic_Effects_on_Protein_Intrinsically_Disordered_Regions/7701278

De Souza, JV; Zariquiey, FS; Bronowska AK. 2019, **Hybrid_FF: a novel method for parametrizing protein disordered regions for MD.** Under preparation.

Zariquiey, FS; De Souza, JV; Estrada-Tejor, R, Bronowska AK**. If you can't win, then join them: Understanding new ways to target STAT3.** ACS Omega. 4,9,13913-13921

Zariquiey, FS; De Souza, JV; Bronowska AK. , **Cosolvent Analysis Toolkit (CAT): a robust hotspot identification platform for cosolvent simulations of proteins to expand the druggable proteome**, Scientific Reports. 9, 19117, 2019.

Published Source codes

De Souza, JV; Bronowska, AK; SQuEE – Structural Quantifier of Entropy, 2019.

De Souza, JV; Bronowska, AK; Hybrid_FF – Force-Field Hybridizer, 2019.

De Souza, JV; Bronowska, AK; MOL22GMX, 2016.

De Souza, JV; Zariquiey, FS ; Bronowska, AK; MOL22GMX, 2016.

List of Figures

List of Tables

Table of Contents

Introduction

Proteins are the molecular tools used by cells to control processes such as proliferation and survival. Usually, these macromolecules require an organized scaffold to be functional. Nonetheless, with the improvements in experimental assays for biological systems, an increasing number of studies started to focus on proteins that can be functional without an organized scaffold. Proteins without a said organizational level are known as intrinsically disordered or metamorphic proteins. These proteins denoted as IDPs exert crucial roles in cellular signalling, growth and molecular recognition events, and are directly related to several neurodegenerative diseases.

IDPs are very challenging for computational studies, such as all-atom molecular dynamics (MD) simulations. MD simulations can provide insight into structure and dynamics at the atomistic level of detail, which may result in accurate molecular models for protein studies and drug-design approaches. However, the current generalist physical models (protein force fields and solvent models) used in MD simulations are unable to generate satisfactory ensembles for IDPs/IDRs when compared to existing experimental data, since they were biased towards available structural data, mainly focused on proteins with a high organizational level.

- The development of a water model with a better accuracy regarding sampling of IDP ensembles. This led to the development to that charge augmented three-point water model (CAIPi3P).
- A new force field framework to improve the sampling and accuracy of the dynamics for intrinsically disordered regions. This set of parameters were made by merging both AMBER03ws and AMBER99SB-ILDN, resulting in Hybrid_FF force field.
- A software known as a structural quantifier of entropy (SQuE) was developed to analyse the changes caused on loops throughout the dynamics.

Thesis Structure

The thesis starts with a concise explanation of protein chemistry and the thermodynamic driving force behind protein folding. This lays the groundwork to explain the existence of IDPs and their intrinsic characteristics. On the following chapter, the theoretical background for molecular dynamics will be given, and a literature review of most modern methods of MD simulations for IDPs will be done. The following 3 chapters will discuss the results in depth.

Chapter 1: Fundaments of protein folding.

This chapter will be focused on the molecular properties of a protein, such as their composition and structure hierarchy. Upon that, the thermodynamics of a protein folding event will be explained, and how it affects its function.

Chapter 2: Disorder in protein biochemistry

Chapter 2 is focused on IDP biochemistry, with their properties and pathologies are discussed in depth. Important examples related to diseases are shown, alongside a discussion on why this class is so challenging for experimental studies and how computational approaches can mitigate this challenge.

Chapter 3: Theoretical Background

The algorithms behind molecular dynamics simulations are explained in detail in the first part of this chapter. The second part discusses how we can generate new parameters and how can we analyse the generated data. The third part is a concise discussion on recent advances for IDPs simulations,

Chapter 4: Development of the CAIPi3P water model

This chapter will be focused on the development of the CAIPi3P model and its results. This model was tested on IDPs known as benchmarks for new MD methods: histatin 5 and R/S peptide. alongside, CAIPi3P has been tested on a longer IDP, recently deposited on PDB.

Chapter 5: Hybrid_FF and intrinsically disordered regions

To improve simulations on disordered regions within structured proteins, A new force field framework was developed. This approach will be explained in this chapter. This set of parameters were made by merging both AMBER03ws and AMBER99SB-ILDN, resulting in Hybrid_FF force field, which was tested using a series of structures obtained by NMR with terminal IDRs.

Chapter 6: Stability and long-range effect of disordered loops

To analyse and quantify the entropic effect caused by IDRs, A software known as a structural quantifier of entropy (SQuE) was developed to analyse the changes caused on loops throughout the dynamics. This approach is explained in this chapter, with the goal was to calculate the upper values of structural entropy changes and was tested in two different cases, with astounding accuracy.

Chapter 7: Conclusions

This Chapter concludes this works, discussing the improvements and drawbacks offered by the methods developed in this work, as well as proposing areas which this work can be improved.

Chapter 1 - Fundamentals of protein folding

1.1 Proteins structure and function

Proteins are tools used by the cell to do most of the necessary process to thrive and survive. These macromolecules are the controllers of cellular processes like DNA replication, RNA transcription, transport, and many more[1]. Their chains have directionality, i.e. two ends (N- and C) one chemically distinct to the other. Each amino acid is joined to the next one, which forms a peptide bond between the two building blocks.

The C-CO-N-C atomic sequence is known as the protein backbone. It is the same for all the amino acids that comprise a peptide. The twenty proteinogenic amino acids that can be encoded by the human genome are known as standard residues. They differ from each other by the side chain (R position in Figure 2). All the 20 residues are shown in Figure 1. They are divided into four categories: non-polar, positively charged, negatively charged and polar. These characteristics, alongside the side chain size, define a substantial amount of properties of the residue. The sequence of residues that assembles the peptide is called the primary sequence of a protein.

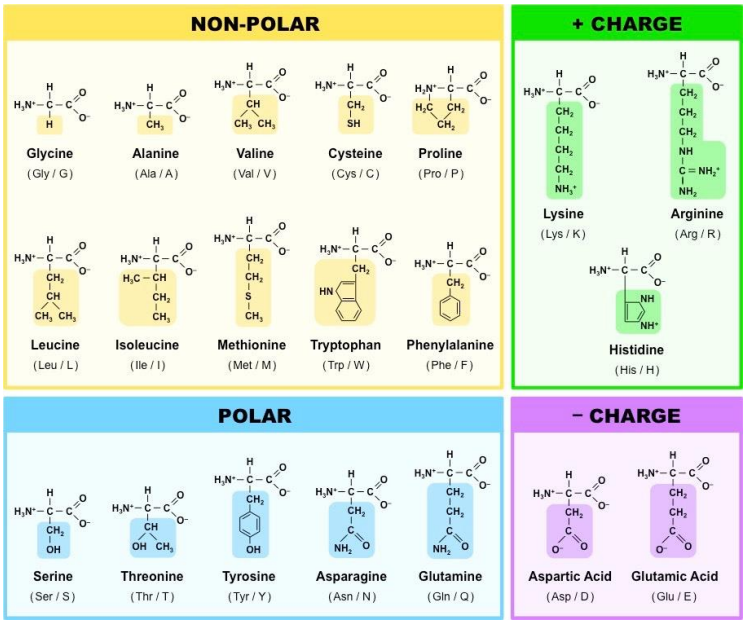

Figure 1: The twenty residues that DNA can encode. Commonly, they are divided into four classes: negatively charged, positively charged, non-polar and polar. Extracted from [2]

The side chain composition defines the biophysical characterisation of the amino acids. Non-polar residues compose most of the protein core since their interaction with water is unfavourable. Some of the residues in this category require extra attention from a modeller perspective, such as glycine and proline. Glycines are the only non-chiral residues, with a sidechain composed only by one hydrogen. Hence, the backbone of the glycine is highly flexible in comparison to the other 19 residues, since there are minimal sidechain steric clashes. However, because of the same reason, glycines are considered disorder promoting residues, often causing substantial fluctuations within protein dynamics[3,4].

Prolines, on the other hand, are overly rigid. The sidechain closing into a pyrrolidine ring with the backbone nitrogen (Figure 1) causes a restriction on possible configurations. Consequently, the region that contains the proline often suffers significant disruption which affects its structure and dynamics[5].

Another residue type that is important to mention is cysteine. The terminal sulphur located at its sidechain confers to it a nucleophilic reactive nature[6,7]. One of the main effects is the formation of disulphide bonds between cysteines. The formation of these bonds restrains the protein conformation, directly affecting its biophysical traits. Also, cysteines are frequent targets of oxidative stress: they

can acquire different protonation states depending on their environment changing their structural configuration[8–10].

Polar residues are mainly located on the protein surface, given the fact they favourably interact with water. Both categories (negatively and positively polar) are essential for modulating interactions, either with ligands or between proteins.

It is essential to say that the polar/charged classification comes from studies of proteins performed in a standard pH of 7. The pH value is crucial when analysing and discussing charged residues since they are prone to be hydrogen donators or receptors.

The charged residues are commonly occurring within enzyme catalytic sites. One example is the catalytic triad located in chymotrypsin[11]. The triad is composed of a serine, a histidine and an aspartic acid, which attacks peptide bonds, breaking proteins in the digestive tract.

The sidechain protonation states of charged residues may change depending on the residue local environment. One of such cases is the protonation on histidine imidazole. Since its one of its pKa is the closest to the human natural pH (pK_a = 6), it is prone to change its protonation state. Because of this, several studies were made on the proteins containing histidines to the protonation effect on its structure[12–14], showing how sensitive proteins are to histidine protonation.

Another vital characteristic of protein residues is the post-translational modification (PTM). Common examples of PTMs are glycosylation and phosphorylation. They are essential to molecular modifications and cellular signalling and happens typically on specific residue sequences[15–18].

1.2 Protein structural hierarchy

All residues follow the same composition for their backbone, as shown in Figure 1. The connection between residues is a peptide bond between the carboxylic acid carbon to the amino group on the adjacent residue. The next atom on the chain is known as the carbon α (Cα).

Each residue has three bonds that belong within the backbone. The N-C peptide bond is non-rotatable, given its partial double bond character[1,5]. Its dihedral value

(known as the ω angles) only accesses two values: 0 degrees (the cis configuration) and 180 degrees (the trans configuration). For protein peptide bonds, most of the residues are found at the trans configuration, given its lower energy in comparison to the cis conformation[1].

The N-Cα and the Cα-C bonds are fully rotatable. These bonds are the key factors on protein dynamics since all the backbone motion are related directly to the dihedral values ψ for the N- Cα bond and φ for the Cα-C bond. Figure 2 shows a representation of both ψ and φ torsional angles located on a protein backbone.

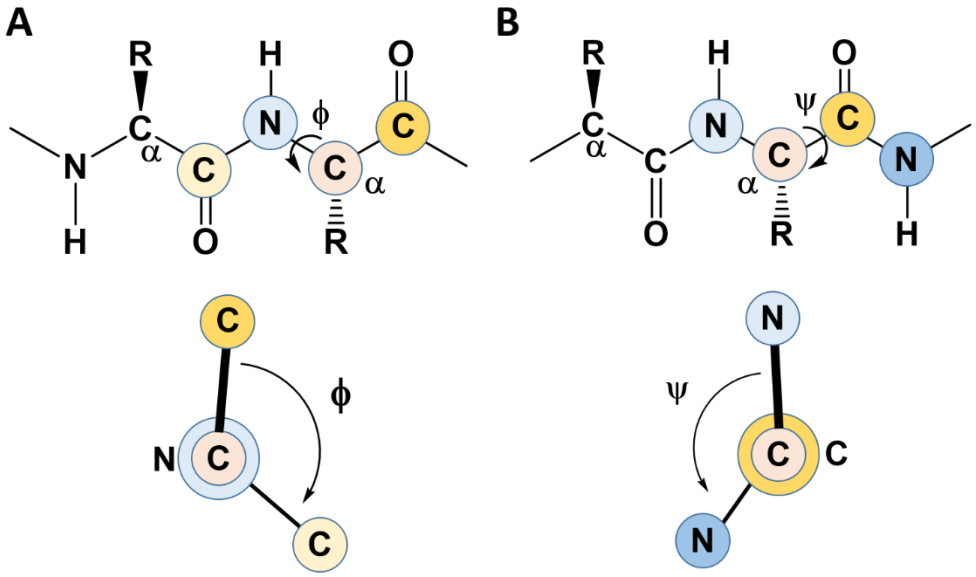

Figure 2: The backbone structure and their respective dihedrals. Most of the dynamics of a protein happen on its dihedral torsions. These torsions are crucial for stabilisation and folding, being unique for each residue. Extracted from [19].

The possible accessed values for the ψ/φ depend on the residue sidechain composition. The steric clash caused by the existence of large sidechains causes the significant difference for distribution of said angles. The distributions of ψ/φ angles are known as the Ramachandran plot[1]. They are essential for the analysis and studies of the next level of protein scaffold: secondary structure.

The protein backbone is prone to form hydrogen bonds between the amide hydrogen and the oxygen from the carboxamide at different residues. These interactions assemble the commonly found secondary structures.

Secondary structures can be found in different scaffolds, being the most common the helices and sheets. There are several types of helices, such as the π-helix, 3-10 helix and α-helix, as the latter being by far the most common[5]. The helical scaffold is frequent because it allows systematic intra-chain hydrogen bonding (as shown in Figure 3), resulting in a more favourable conformation than the unfolded string.

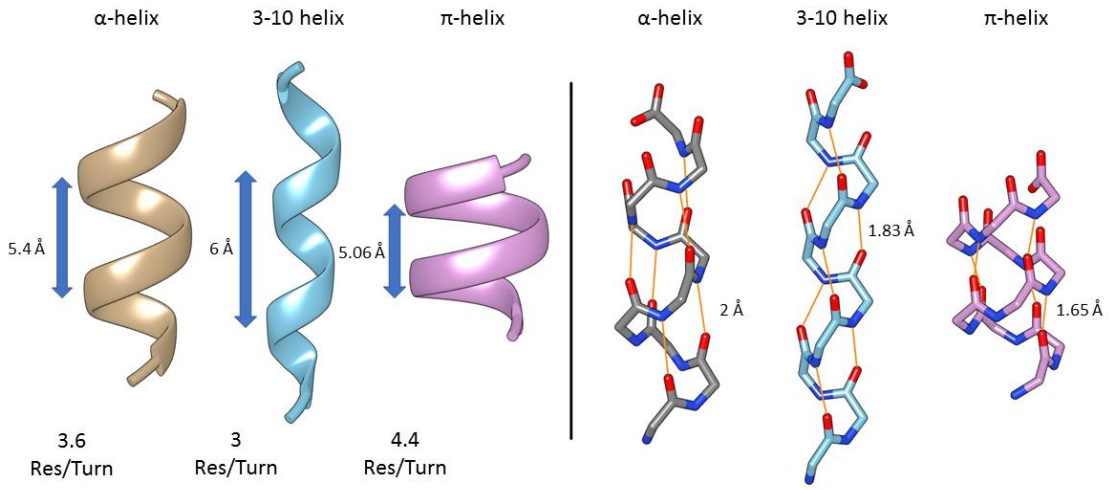

Figure 3: Types of helical structures found on proteins. From these three, the α-helix is by far the most common, given its favourable hydrogen bonds and weaker steric clashes. 3-10 helices and π -helices are less common and typically are crucial for specific protein functions. Made by the author.

α-helices represents the majority of the helical components in the human proteome[4,5,20]. They have a translational periodicity of 5.4 Å with a residue per turn ratio of 3.6. These characteristics result in one hydrogen bond every four residues with an average distance of 2Å, with a ψ and φ values of -57° and -47° respectively. In comparison, the 3-10 helix has -49°, and -26° and the π-helix has -47° and -70° for their ψ and φ values[3], resulting in a tighter and a broader helix respectively.

The ψ/φ in α-helices results in structures containing the lowest degree of steric clashes of all three[1]. However, 3-10 helices occur reasonably frequently (~10%), albeit in shorter sequences (usually no more than four residues). The shorter hydrogen bonding interactions stabilise this scaffold, but the backbone torsion strain makes them unfeasible for long helices. They were first found in essential proteins such as haemoglobins[21] and myoglobins[22].

The second most common secondary structure is the β-sheet. Built by connecting β-strands, they were first proposed by Willian Ashbury in 1930[1,4,5]. The h-bonds forms between each strand, resulting in a pleated sheet. The isomerism contained within the residues forces the sheets to be twisted, increasing the torsional stress on the strand extreme. This stress limits the length of the β-sheets[23,24].

The strand-strand complex comes in two different scaffolds: antiparallel β-sheet and parallel β-sheet. As shown in Figure 4, the backbone h-bonds created between strands are different between the antiparallel and parallel scaffolds: in parallel sheets, the h-bonds requires an angle to form, increasing the distance between HN-O. However, on antiparallel sheets, the h-bonds are closer and parallel between themselves, resulting in a more stable configuration.

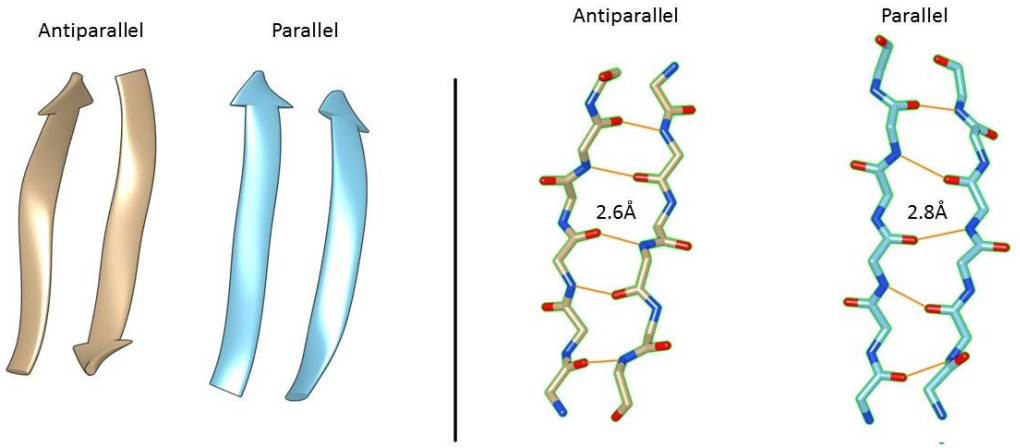

Figure 4: β-sheets configurations. Antiparallel strands result in more favourable inter-strand interactions, being more frequent and more stable. Made by the author.

β-sheets arrangements can be found within several different structural and biological functions [5,25,26]. Usually, bulky side-chain amino acids (i.e. tyrosine and tryptophan) and branched sidechain residues (isoleucine, threonine) are found in the middle of β-sheets. It is also common to have prolines on end-caps of β-strands, to avoid edge-edge interactions[27].

Because of the stable intra-strand organisation, they are directly related to the formation of β-fibrils and protein aggregation exclusion bodies. This fibrillation and aggregation propensity will be discussed in-depth in chapter 2.

Connecting the strands is a structure known as β-turn. This structure is a short sequence (between 2-4 residues) that frequently contains glycines and prolines. Venkatachalam[28] showed in 1968 that are several possible configurations of β-turns to access. They all differ on the average ψ/φ values of their comprising residues, but always have a hydrogen-bond between the first and the last residues. These hydrogen bonds are crucial for maintaining the integrity of the turn and the strands that it connects, and in case of disruption, often affects the protein stability and may result in protein fibrillation[29].

Finally, the residues with no determined secondary structure are often called unstructured loops. Usually comprised of more than five residues, these regions are challenging for both experimental and computational modelling methods. These unstructured yet functional sequences are called intrinsically disordered regions (IDRs). Proteins which are only comprised of unstructured residues are named intrinsically disordered proteins (IDPs). Examples of IDRs and IDPs will be described in Chapter 2.

The combination of helices, loops and sheets form the tertiary structure of a protein. A stable tertiary structure is often its functional form, known as native structure, which contains the requirements for its functionalities, such as catalytic centre and substrate binding sites.

Mathematically, the possibility space for a protein to fold is an astronomically immense value. For example, a protein with 100 residues contains 99 peptide bonds and 198 ψ/φ torsions. If we assume that ψ/φ can assume three stable values, the total number of structural possibilities is 3^196. Hence, if this protein reaches its native state by sequentially sampling all these configurations, it would take longer than the age of existence of the Universe. This is known as the Levinthal paradox[30–33] since small peptides fold within the microsecond to millisecond scale. Therefore, a driving force must exist to guide the folding to its right conformation.

1.3 Water and protein folding

For a protein to achieve a native conformation, it needs to go through several structural changes. The driving force behind these changes is a thermodynamic

quantity known as the system free energy. The two most common free energies a system may have is the Helmholtz free energy and the Gibbs free energy[34]. The Helmholtz free energy is a thermodynamic potential that measures the useful work obtainable from a closed thermodynamic system at a constant temperature and volume (isothermal, isochoric). This makes the Helmholtz free energy useful for systems held at constant volume.

The Gibbs free energy (or free enthalpy) is most commonly used as a measure of thermodynamic potential (particularly in biochemistry) when it is convenient for applications that occur at constant pressure. The Gibbs free energy is represented by Eq.1:

$$G = H - TS = E + PV - TS \quad (1)$$

Where G is the Gibbs free energy, H is the enthalpy, S is the entropy, T is temperature, E is the system internal energy, P is the pressure of the system and V is the volume of the system.

Moreover, the difference between two different thermodynamics i and j states is (Eq. 2):

$$\Delta G = G_j - G_i = (H_j - H_i) - T(S_j - S_i) = \Delta H - T\Delta S \quad (2)$$

The variation in Gibbs free energy between states represents how spontaneous a thermodynamic event is. When a transformation is exergonic (ΔG <0), there is an increase in the entropy of the universe; therefore, it is spontaneous[1,35–37]. However, for a system to be spontaneous, a balance between the enthalpic changes and the internal entropic changes is required. The spontaneity of an event can be tracked using Table 1.

Table 1: Free energy and spontaneity

Enthalpy	**Entropy**	**Spontaneity**	**Free energy**
$\Delta \boldsymbol{H}$ **> 0**	$\Delta S > 0$	If temperature is high	Depends on T
$\Delta \boldsymbol{H}$ **> 0**	$\Delta S < 0$	Not spontaneous	$\Delta G > 0$
$\Delta \boldsymbol{H}$ **< 0**	$\Delta S > 0$	If temperature is low	Depends on T
$\Delta \boldsymbol{H}$ **< 0**	$\Delta S < 0$	Spontaneous	$\Delta G < 0$

Based on this, Jose Onuchic coined a concept known as protein folding energy funnel[38]. As shown in Figure 5, for a protein to achieve its native conformation, the unfolded protein navigates through a free energy surface, moving by several different local minima. Each of these partially folds the protein, reducing both their internal energy and structural entropy[39,40].

From the fully unfolded conformation, which has high entropy and high energy, the molecule starts to fold locally. This transition configuration is known as molten globule: it starts when the free energy decreases and the local interactions assemble in secondary structures. From a molten globular configuration, the system continues navigating the funnel, increasing the favourable intra-protein interactions and water-solvation interactions to achieve its functional form.

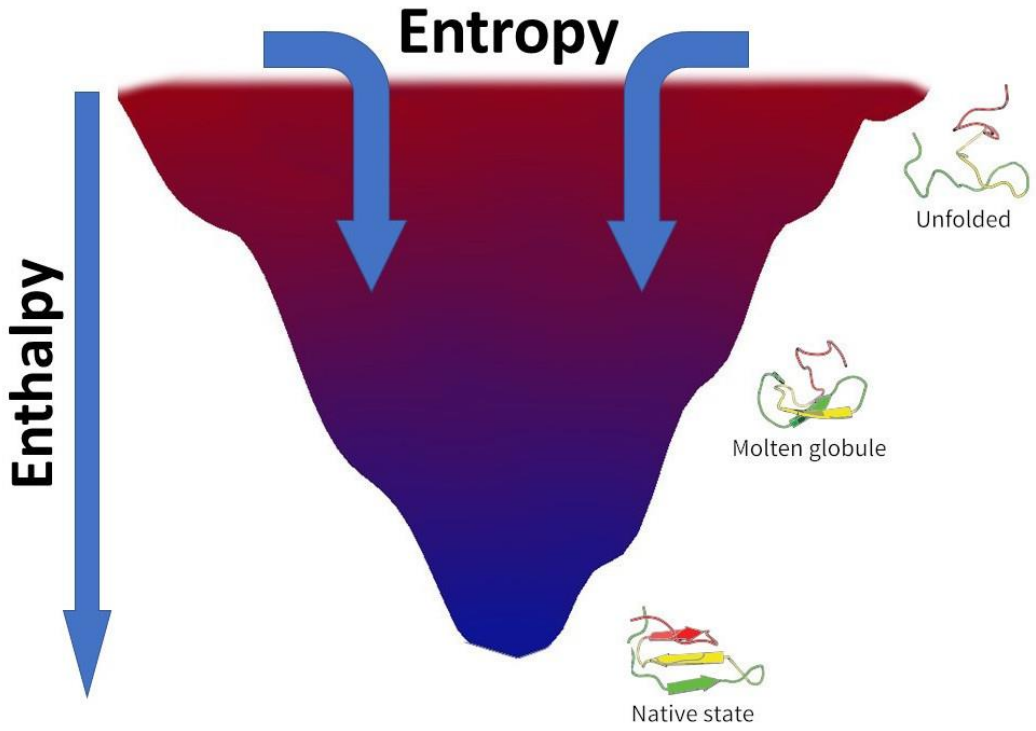

Figure 5: The protein folding funnel. The process of folding decreases both the enthalpy of and entropy. For a stable native structure, the free energy of the system needs to be reduced as well. These three events are only possible by increasing the entropy of the surroundings, while decreasing the entropy of the protein, resulting in a negative free energy change. Made by the author.

However, the decrease in structural entropy should be unfavourable, and at first, should hinder the folding process. Nonetheless, the folding process is pushed forward because there is an increase of entropy in the other component of the system: the solvent.

The water environment plays a crucial role in the process of folding, given the diversity of the 20 natural amino acids. When a hydrophobic particle (i.e. apolar residue) is submerged in water, the solvent molecules organise themselves around the particle, creating a web of hydrogen bonds that surrounds it. Because of this coordinated scaffold, there is a decrease in the entropy in the system (Figure 5). Upon interaction between these hydrophobic constituents, the solvation net is disrupted, resulting in a system entropy increase. Figure 6 shows a representation of the increase of entropy when hydrophobic regions interact

with each other. Usually, this process occurs throughout the structure, resulting in the hydrophobic core of the protein.

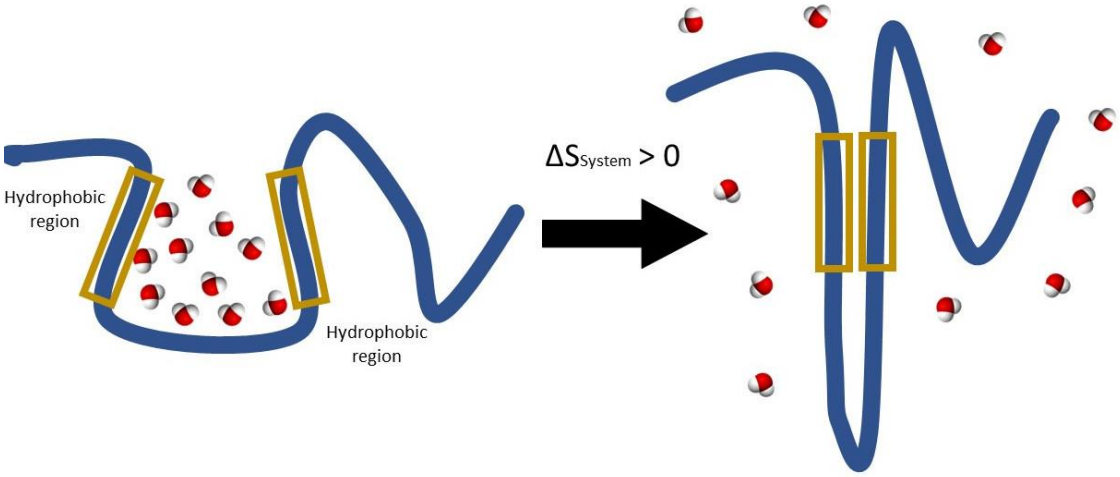

Figure 6: Water organisation around hydrophobic regions. The water molecules organise themselves around the nonpolar areas, reducing the entropy of the system. When the hydrophobic patches interaction free the structured waters, increasing the entropy and decreasing the protein enthalpy, resulting in a negative free energy change. Made by the author.

Most of the known proteins assemble around a hydrophobic core. The core formation is a crucial factor to stabilise protein folding and often dictates its native structure. Consequently, the existence of hydrophobic regions within the primary sequence is fundamental to the formation of a globular state.

Since the folding landscape is uneven, protein folding is not a trivial event. Macromolecules may face barriers in the free energy landscape that can create deep local minima. These barriers may trap them in conformations far away from their native one. One system that safeguards molecules from misfolding is the usage of chaperones, such as heat shock protein 90 (Hsp90). These proteins are folding catalysts, which guide the protein folding by increasing their free energy to overcome energetic barriers[41,42].

This folding funnel concept is qualitatively useful to understand the folding of well-behaved globular molecules. However, it is estimated that more than 20% of the human genome is composed of proteins without a stable tertiary structure[43–45]. Commonly, their primary sequences lack hydrophobic regions, and therefore, they lack a folded core. This generates a unique effect on the energy landscape, creating an effect known as a reverse energy landscape [46–49] (Figure 7). This

effect on the folding landscape emerges from the favourable interaction with the aqueous environment, preventing the formation of a hydrophobic core. Another effect that arises from the lack of a core is the existence of multiple functional states. Several IDP energy minima may have similar free energy; therefore, copies of the same molecule may coexist in heterogeneous conformations[44].

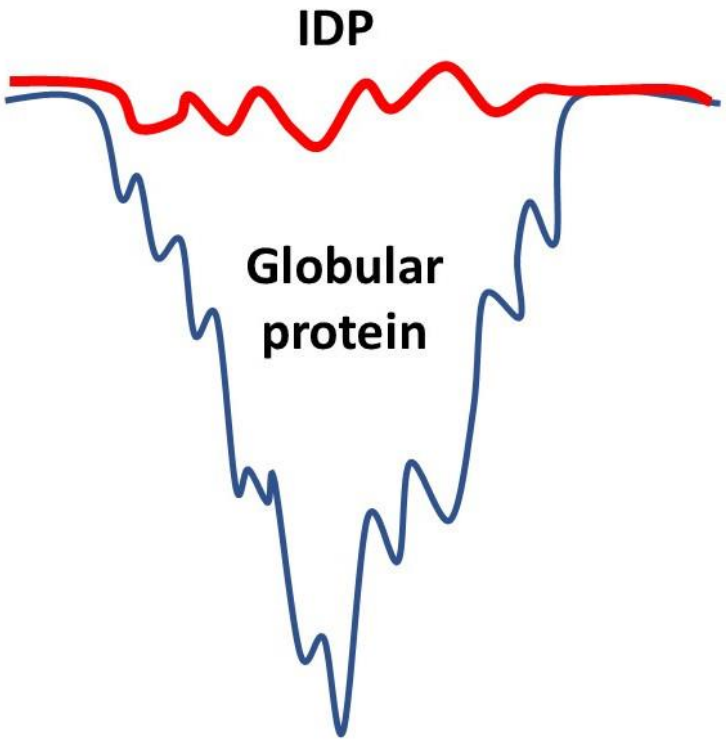

Figure 7: The energy landscape of an IDP. This class of proteins navigates in a specific energy landscape, with several minima with similar magnitudes, resulting in several possible conformations for the IDP to access.

This metamorphism of IDPs is linked to several human diseases. Similarly, to IDPs, IDRs are also affected by this modification in the folding landscape. Both IDPs and IDRs and their related diseases will be discussed in-depth in the next chapter.

Chapter 2 – Disorder in protein biochemistry

2.1 Biological importance of protein disorder

In the early 1990s, the abundance of studies on well-defined structured proteins kept the importance of protein disorder at bay[50]. With the emergence of new biophysical techniques and bioinformatic studies on the complete genome, the scientific community started observing an increased interest in disorder sequences within the human proteome[51].

These unstructured proteins are highly relevant to cellular fitness and homeostasis[51–53]. Known as intrinsically disordered proteins (IDPs), these polypeptides are characterised by the lack of bulky hydrophobic residues and low sequence complexity. As explained in Chapter 1, the lack of such residues prevents the formation of a hydrophobic core, resulting in a protein structure that fluctuates within an ensemble of structures with diverse internal organisation levels. Most of the proteins that compose the eukaryotic proteome contains both structured and disordered regions, especially within loops between structure domains; these loops are named intrinsically disordered regions (IDRs). In this work, IDP will be used as a general term for extensively unstructured proteins, despite the recent discussions in IDPs classification and nomenclature[45].

Frequently, IDPs function as key hubs within protein interaction networks[53,54]. They regulate crucial pathways, such as transcription, translation, cellular signalling, cell cycle and proteostasis. The feature that makes IDPs part of several cellular processes is their structural plasticity, which allows them to acquire a diverse range of functions, partners and environments effects[55,56].

Regarding IDPs primary sequences, there are certain characteristics that are useful for cellular fitness and survival. First, within their sequences, there are small recognition elements that fold upon binding to a partner. This attribute is vital for the assembly of the eukaryotic macromolecular machinery such as the ribosome, performing a function like chaperones[57]. Second, since IDPs structures may fluctuate between partially molten globules and fully unstructured strands, they play an essential role in putting together microtubules and coordinate transmembrane pore formation[58]. Third, when unstructured sequences are

located within globular functional motifs, they act as “entropic” linkers, which help proteins to fold locally with lower entropy[59]. Fourth, they can be used as scavenger peptides, which bind to small ligands to either neutralise or to transport them to different cellular compartments. Finally, they can act as display sites for specific regions of post-translational modifications, acting as signalling hubs in cellular cascades[60]. Figure 8 shows a representation of this classification tree.

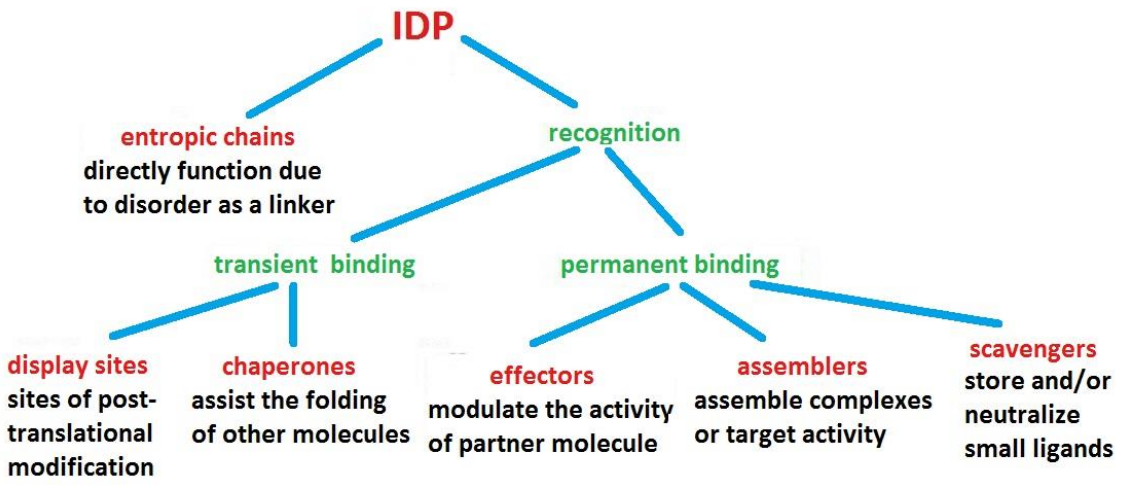

Figure 8: Classification of IDPs. Modified from [61]

The mechanistic procedure of IDP-partner binding is based on the high specificity for its partner, albeit, with a low affinity. Using this characteristic, IDPs can interact with a specific partner activating a cellular network, then undergo structural modifications and unbinding, releasing the partner molecule in the process. Several examples that show IDPs undergoing structural conformational changes after partner binding can be cited, like the neurotoxin botulinum serotype A (BoNT/A) binding to synaptosome-associated protein 25(SNAP25)[62]. The BoNT/A is a clostridial neurotoxin (CNT), which is a causative agent of botulism. It binds to essential proteins called SNAREs, which cleavage is catalysed by the CNTs, impairing neuronal functions.

Another case of IDP structural plasticity being crucial for cellular functions is the Smad proteins binding to the Smad anchor for receptor activation (SARA)[63]. Smad mediates the signalling of the transforming growth factor-beta (TGFbeta) from the transmembrane kinases to the nucleus. SARA recruits Smad2 to the TGFbeta for phosphorylation, allowing it to be transported. The structural

conversion of the SARA from an unstructured loop to an α-helix is central for the interaction specificity with the Smad2 β-sheet[63]. Both examples are shown in Figure 9.

Nonetheless, the unstructured-structure transition is not uniquely necessary for IDPs interactions. Sigalov and coworkers showed for T cell receptor ζ subunit, an IDR from the T cell receptor[64], dimerises without going through structural conformational change.

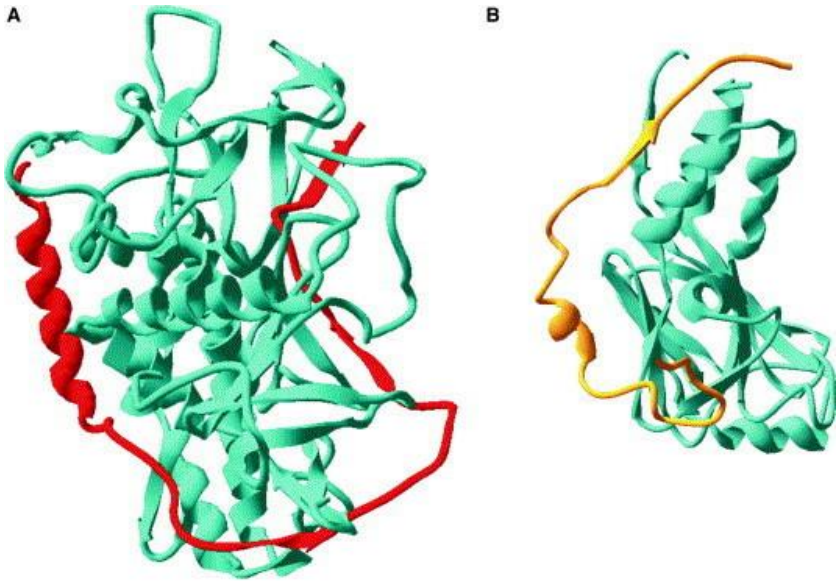

Figure 9: IDPs bounded to their partners, the PDB code is in parenthesis(A) SNAP-25 bound to BoNT/A (1XTG); (B) SARA SBD domain bound to Smad2 MH2 domain (1DEV). Modified from [61]

To attain this degree of conformational plasticity, IDPs typically have primary sequences with a high content of disorder inducing residues (A, R, G, Q, S, P, E and K)[65]. These residues may have a small side chain (A and G), allowing a higher rotational degree of freedom, often containing polar sidechains (Q, S). Lysines, arginines and glutamic acids (K, R, E) have long polar flexible side chains; hence, restraining these residues buried in a structured core is costly for both enthalpy and entropy. Prolines are special cases regarding how they induce disorder, since the pyrrolidine ring incorporated within proline residues causes restraints for the local folding.

In this context, IDPs have a high net charge and low mean hydrophobicity. These attributes mean that water-polypeptide interaction is crucial for a functional conformation, and changes in the environmental composition (i.e. pH, free

radicals, and ionic concentration) may influence the protein stability, misfolding and aggregation propensity[66], resulting in a plethora of human diseases.

2.2 Human diseases related to protein disorder

Protein misfolding causes a range of human diseases. These pathologies arise from the failure of a protein to attain its native structure, losing its function. Alongside with that, proteins may aggregate (or assembling in a fibril conformations), accumulating in the cytosol or toxic inclusion bodies, such as Lewy bodies, within cells[67,68].

Ageing has a direct effect on the *in vivo* protein aggregation propensity[69,70]. Since the cellular machinery becomes more prone to fail with ageing, impaired interactions with endogenous factors (proteins, cellular matrices or small molecules) may cause protein dysregulation. Also, acquired point mutations and failures on PTMs partners directly affect the degradation and aggregation propensity.

Amyloid fibrils are linked to the largest group of misfolding diseases. These fibrils are stable, highly ordered filamentous protein aggregates which originate from the conversion of specific proteins[71]. Subsequentially, these insoluble amyloid structures accumulate within different cells types, causing a series of impediments for cell cycle and function[70]. A representation of the formation cascade for aggregation in amyloid-like fibrils can be seen in Figure 10.

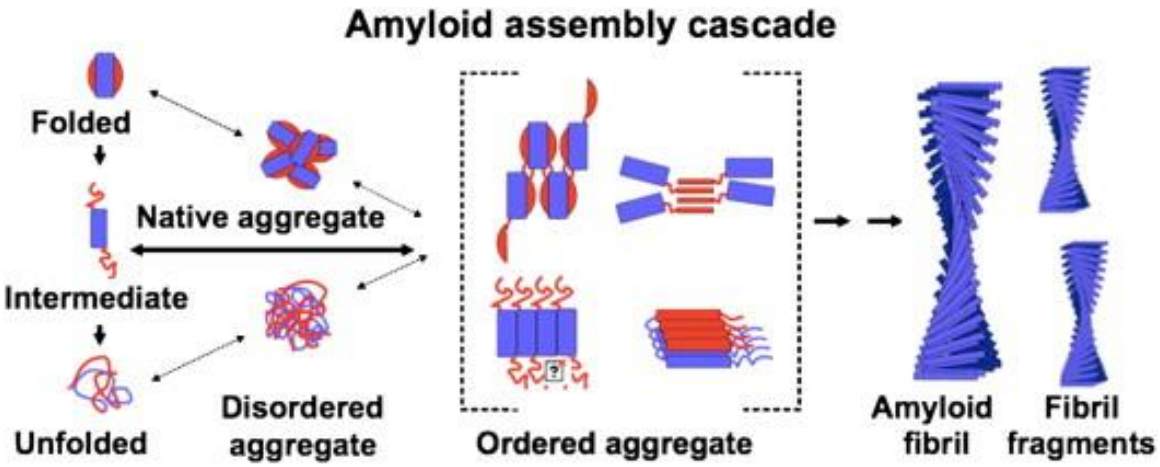

Figure 10: Aggregation cascade: Destabilised by some external effect, the native protein structure may assembly in aggregates, resulting in amyloid fibrillation. Extracted from [72]

Fibrils with amyloid characteristic have a high β-sheet content. Morphologically, they show a core β-sheet structure with a continuous sheet with β-strands running perpendicular to the long fibril axis (Figure 10). Even though the different fibrils have similar morphology, the peptide that constitutes a single subunit may have a acquire different secondary conformations in its native state (β-sheet, α-helix or unstructured).

IDPs are commonly found as proteins that form fibrils. Since the energy barrier required to unfold and refold a protein containing a well-defined hydrophobic core is high, stable globular proteins typically do not start pathological fibrils. On the other hand, IDPs may readily fold in a partially molten globule, which often is the fibril starting point of growth. Nonetheless, IDPs and IDRs in their native conformation perform important cellular tasks, and the trigger for these molecules to form fibrils from misregulation caused by mutations or misrecognition of their binding partner[73].

Several case studies can be discussed as examples of pathogenic IDPs and their mechanisms. One compelling case is the relation between tumour suppressor p53 and the E3 ubiquitin ligase Mouse double minute 2 homolog (Mdm2)[74].

The p53 protein is a transcription factor that target genes involved with cell cycle regulation, i.e. apoptosis. Hence, there is a clear and direct relationship between p53 loss of function and cancer [75]. The Mdm2 binds to the transcription activation domain [76], which has a series of intrinsically disordered regions, blocking the interaction with its genes in three ways. First, it hinders the interactions with other transcription factors sterically, blocking assembly between p53 and partners. Second, Mdm2 acts as a ubiquitin ligase, targeting p53 for degradation. Third, Mdm2 contains a nuclear export signal; hence the complex shown in Figure 11 is transported from the nucleus, preventing its function.

The p53 is central to an extensive cellular information network, regulating signalling cascades critical for cell lifecycle. When Mdm2 disrupts this, the cell goes into a malignant transformation. Cancers that have p53 disruption in their core are found in colon, lung, oesophagus, breast, liver, brain, reticuloendothelial and hemopoietic tissues[75].

The p53 is composed of 3 domains: An N-terminal transactivation domain, a DNA-binding domain (DBD) and a C-terminal regulatory domain[77]. The analysis of the DBD domain shows that both of its terminal regions are intrinsically disordered. These areas are responsible for mediating 70% of p53 interactions with DNA and protein partners [78].

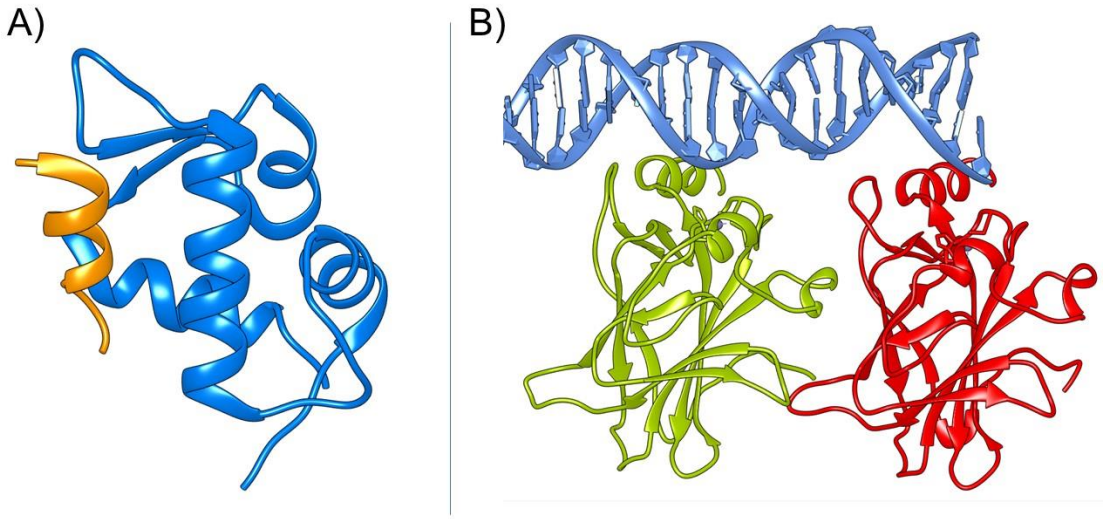

Figure 11: The structure of p53 domains: A) in yellow the part of the IDR located in the transactivation region of the p53, in blue the MDM2 protein (PDB 4HFZ). B) the DNA binding domain of the p53 (PDB 1TSR). Made by the author.

Another condition that IDPs are implicated related is age-dependent Alzheimer disease[79,80]. It often characterised as the accumulation of extracellular amyloid deposits, senile plaques and intracellular fibrillary tangles.

The amyloid β protein (Aβ) is found in high concentration within senile cellular plaques, which are a hallmark of neurodegeneration. Aβ is a 40 – 42 residue peptide produced by endoproteolytic cleavage of the amyloid β-protein precursor (APP). In its native structure, the Aβ peptide does not show any structured regions[80]. However, upon forming a molten globule-like structure, it increases its propensity of fibrillation, hence, partially folded Aβ are indicatives of an early-stage plaque formation.

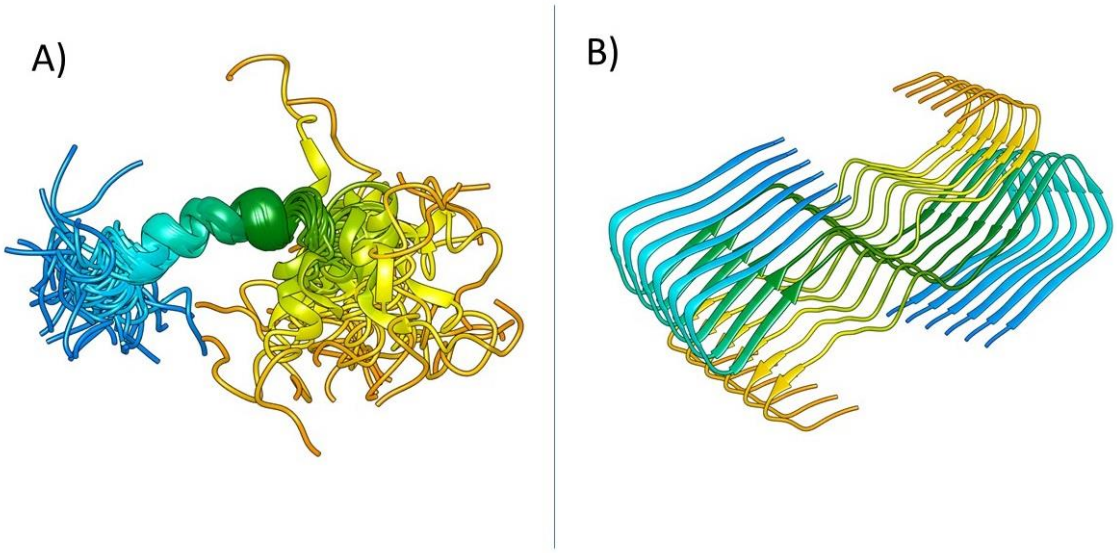

Figure 12: A) helical conformation of the 1-42 Aβ protein in solution (PDB 1IYT)[81]. B) amyloid fibril structure of the Aβ protein (PDB 5KK3)[82]. Made by the author.

Fibrillar tangles are typically found as paired helical filaments (PHF). PHFs assemble around the protein tau, which includes a family of isoforms that associate with microtubules. Interest in tau increased after the discovery made by Delacourte and colleagues[83] of its aggregation in neuronal cells in the progress of AD and several other neurogenerative diseases. The transition from a functional tau to a pathological configuration has been linked with several structural modifications such as mutations and high concentrations of PTMs, mainly phosphorylation[84]. Like Aβ, tau proteins are unstructured in their native conformation; however, they partially fold into a molten globule before assembling in fibrils.

Another known metamorphic IDP related to neurological diseases is α-synuclein. Aggregation of these proteins causes a series of pathological conditions known as synucleinopathies. Clinically, they are characterised by a chronic and progressive decline in motor, cognitive, behavioural, and autonomic functions. Nonetheless, the decline of these functions has a significant overlap with similar neurological conditions caused by different proteins; therefore, the diagnosis is difficult[85].

Common synucleopathies include Parkinson's disease (PD), dementia with Lewy bodies (DLB), Alzheimer's disease (AD), Down's syndrome and multiple system atrophy (MSA). Synuclein inclusion bodies can be found in neurons, deposit in

perikarya, axons or glia. Morphologically diverse, several different inclusion bodies are containing α-synuclein (i.e. Lewy body, Lewy neurites and glial cytoplasmic inclusions). This characteristic makes the treatment even more difficult since the diagnosis and treatment need to be specific to its morphology. Since α-synuclein is an IDP, it shows low structure content in physiological conditions, being a slightly more compact than expected from a full random-coil structure [86]. Using solution NMR, Morar *et al.* showed that the sequence between residue 6 to 37 adopts a helical conformation [86]. This structure correlates with the finding of Ulmer and colleagues, who resolved the structured of the α-synuclein bound to micelles, which shows a highly helical content[87]. This structure and the fibril structure of the α-synuclein can be seen in Figure 13

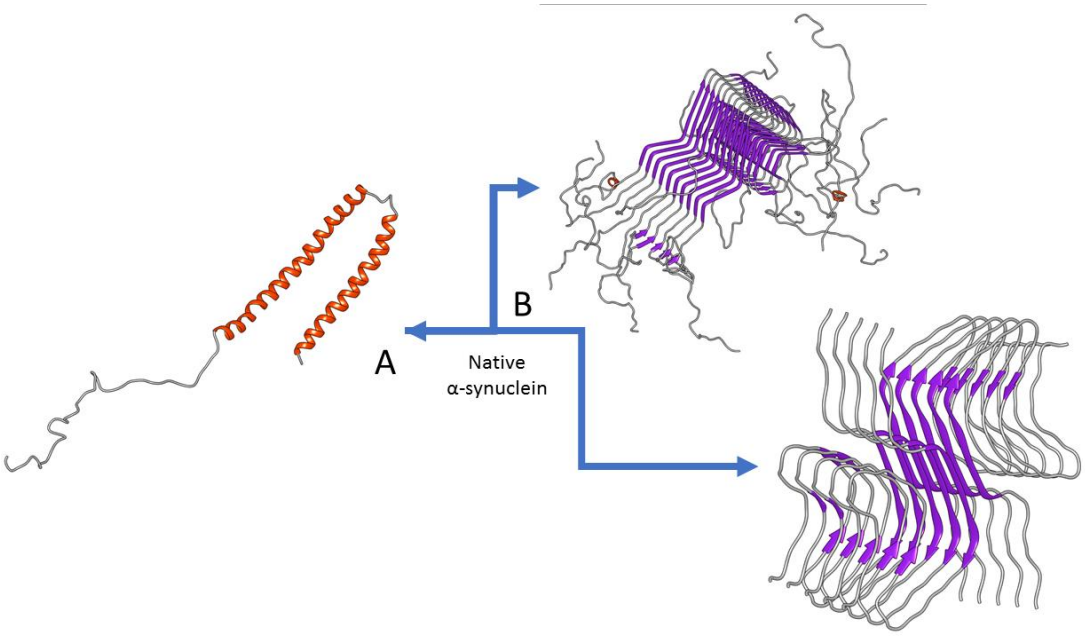

Figure 13: Representation of the α-synuclein metamorphism. As an example, the native α-synuclein may fold into a high helical content structure when bound to micelles (PDB 1XQ8)[87], and different fibril conformations (top structure PDB 2N0A, bottom fibril PDB 6FLT)[88,89].

The α-synuclein is prone to aggregate. Because of its relationship with a series of neurological disabilities, it has been extensively studied[90]. Hence, it became a model system for studies of protein metamorphisms caused by environmental changes. The conformational behaviour of α-synuclein is directly affected by the pH, resulting in fibril like structures on acidic environments[70]. Also, α-synuclein has been shown to form several structurally diverse aggregation complexes, ranging from oligomers (spherical or doughnut) to fully amorphous complexes[70].

Because of the structural diversity of these proteins, Uversky coined the concept protein-chameleon[91]. A protein with such characteristics follows the IDP landscape described earlier, so predicting and assessing the presumably toxic conformation is a very challenging task.

2.3 Experimental and computational determination of IDP characteristics

Several techniques can experimentally determine the molecular characteristic of IDPs, yet, all of them have their drawbacks. Typically, the assessment of a protein atomic configuration is done by X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM). Several IDP characteristics can also be determined using biophysical assays such as circular dichroism (CD), small-angle X-ray scattering (SAXS), size-exclusion chromatography (SEC), mass spectrometry techniques, and many more.

For X-ray crystallography, a significant shortcoming is the necessity of a stable crystal. Crystals require a unique internal repeating structure to diffract, and therefore, IDP-based crystal would lose the information on its possible accessible ensemble. Alongside that, the IDP crystal growth would be almost impossible given the lower organisation level of IDPs. Hence, for IDPs, X-ray crystallography is mostly used to study structured protein – IDP complexes[92,93].

Solution NMR studies of IDPs usually will give more information regarding their molecular organisation. NMR nuclear Overhauser effect (NOE) can give an internal distance list between labelled atoms, and consequently, a modelled dynamical ensemble. Also, chemical exchange NMR studies have been used to study residue-solvent accessibility for IDPs, helping to elucidate how dynamics within the microsecond time scale work[94]. IDPs represents a challenging problem for NMR techniques: first, the requirement of labelling the sample may affect the protein production, and second, NMR samples need to be highly soluble, as precipitation into aggregates directly affects the NMR assignment. Another pivotal NMR-based technique is the solid-state NMR (SSNMR)[95,96]. SSNMR is especially useful for very large and poorly soluble proteins, such as amyloid fibrils. SSNMR uses the magic-angle spinning (MAS) to acquire the inter-atomic

distance list. While useful, it may require a protein to be fixed to a membrane, reducing its applicability.

Cryo-EM for IDPs faces similar challenges as X-ray crystallography. The sample preparation requires the protein to be restrained within an amorphous ice sheet. Also, the requirement for several electron micrographs of different aspects of the same conformation creates a barrier for the application of Cryo-EM for IDPs. However, this technique has been successfully used to assess polymorphisms of the amyloid fibrils[97].

SAXS uses the scattering profile of a protein in solution. One of the most exciting information that SAXS can give is the internal radial distribution, which can be used to model the external molecular surface. Alongside with that, SAXS can be used to obtain an accurate measure of the protein hydrodynamic radius; hence it is commonly applied to the studies of IDPs. How the data is acquired and processed is explained in the next section.

Another biophysical technique commonly used for IDPs is the circular dichroism (CD). In CD, the light polarisation properties of proteins are used to obtain a semiquantitative percentage of the secondary structure of a sample. Since the protein is in solution, it became a useful method to understand pH and temperature effects on disordered systems.

Amidst all experimental techniques, several computational tools can be used to predict disorder and asses its dynamical states. First, there are several curated online databases of protein sequences and biophysical information for IDPs, such as Disprot[98], D2P2[99], MobiDB[100], IDEAL[101] and pE-DB[102]. The latter contains a significant amount of structural information as well. Alongside with databases, online prediction tools for disordered regions were developed. These methods are based either on the mentioned databases or in physicochemical properties of the primary sequence. The most known webserver predictors are shown in Table 2

Table 2: List of prediction servers for IDPs and IDRs.

Server name	Method	Prediction	Advantages	Disadvantages
DisEMBL[103]	Machine Learning	Disordered loops and high mobile regions.	Fast, online server	Not clear outputs, restricted training dataset
Globplot[104]	Empirical analysis	Regions with the propensity for globularity (probability of secondary/random coil formation)	Fast, online server, Prediction of IDP-Structural effect.	Empirical approach results in a high flexibility but does has a lower than machine learning methods.
Pondr[105]	Machine Learning	Random coils and molten globule regions	Fast, online server, Prediction of IDP-Structural effect, analysis of structural dynamics.	Restricted training data sets
FoldIndex[106]	Empirical analysis	Regions with high net charge and low hydrophobicity	Structural analysis, physical-chemical properties calculated	Inaccurate regarding complex motifs.
IUPred[107]	Energy empirical analysis	Regions that lack 3D structures	Clear output. Motif definition, domain recognition.	The binary classification does not allow the analysis of possible molten-globule structures.

None of these web servers can give atomic resolution and information regarding intrinsic molecular dynamics. From this point of view, atomistic molecular dynamics (MD) simulations offer an advantage over other methods. Using MD techniques, the scientific community can acquire information about its protein conformational landscape and dynamics.

2.4 Experimental techniques for structural studies

Acquiring atomistic insight of molecules is invaluable for protein science. Although many experimental assays give us information with a molecular resolution, there are no techniques with zero caveats for IDPs.

Three main experimental techniques for structure assignment applicable to IDPs are X-ray crystallography, nuclear magnetic resonance and transmission cryo-electron microscopy. Another essential method to assess structural states with dynamical data is small-angle X-ray scattering (SAXS).

2.4.1 X-ray crystallography

X-ray crystallography counts for 90% of the structures deposited in the Protein Databank[5]. The cornerstone for this technique is the Huygens-Fresnel[5] principle (Figure 14), which states and predicts how waves behave when diffracting.

The diffraction phenomena occur when light passes through any slit. As stated by the Huygens-Fresnel principle, the incident wavefront should be treated as a new punctual wave source. These emerging waves interact with each other creating patterns which are unique to their wavelength and the grid structural characteristics. The outcome is a series of constructive and destructive wave interferences.

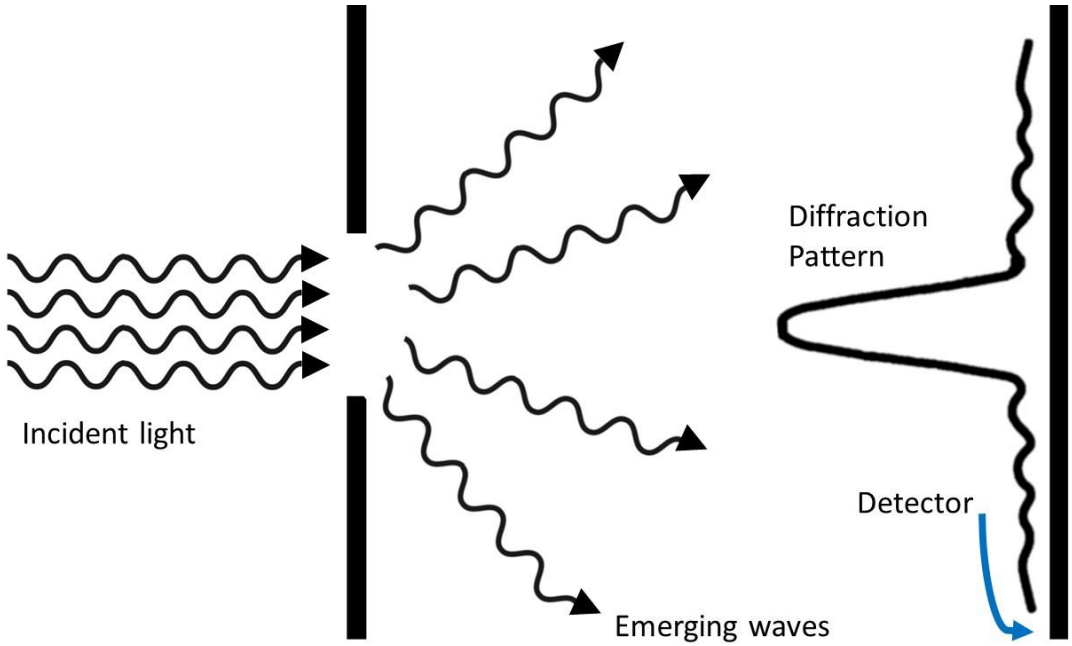

Figure 14: The concept of slit diffraction is based on the Huygens-Fresnel principle. After interfering with the slit, the light emerges as a new set of waves which interacts with the other emerging waves, generating regions of constructive interference

Given the uniqueness of the diffraction pattern, it is possible to discover the internal organization of the diffraction grid by studying the constructive-destructive interferences obtained in a detector. Biological X-ray crystallography uses a crystal of a biomolecule as the grid. The electronic density of this biomolecule composes the grid, which can be calculated from the diffraction pattern, as shown in Figure 15. The incident light has a wavelength of 0.1 nanometres, so it results in a molecular model at an atomic scale. This technique, although highly successful in determining protein structures, has two known caveats: first, it requires a stable crystal and second, its results in a snapshot of a crystal component, and may not represent a full picture of the native protein.

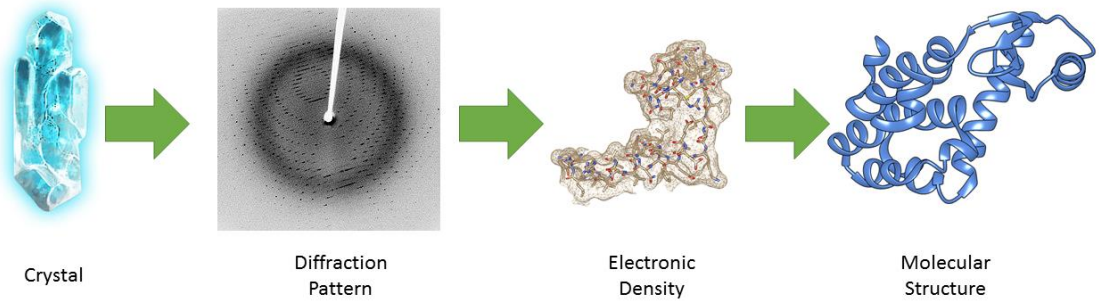

Figure 15: Workflow for solving a protein structure through x-ray crystallography: After the crystallisation, the protein is placed in an x-ray beam, resulting in the diffraction pattern. From the diffraction pattern, one can calculate through mathematical methods, the electronic density that caused it. Using the obtained density, it is possible to fit the protein residues, given the internal restrains, into the density map.

The crystallisation is usually a highly challenging aspect. Typically, it is required for the protein to have a stable tertiary structure and a well-defined core. The need for a crystal with sufficient size and stability significantly limits the number of proteins which can be studied by this technique.

Because of the X-ray diffraction requires a single crystal to get a successful diffraction pattern, X-ray crystallography, it is unable to describe the full dynamics of the protein. This static snapshot addresses important questions about the backbone structure, the side-chain interactions, and potential small molecule binding sites. Nevertheless, it is unable to clarify several characteristics related to the dynamics of the macromolecule in water, like conformational changes caused by thermal fluctuations[5].

2.4.2 Nuclear magnetic resonance

The second technique for structure assignment is the solution nuclear magnetic resonance (NMR). Solution NMR relies on the nuclear spin, which is an intrinsic characteristic of the particles. The nuclear spin interacts with external magnetic fields and responds to it in different manners. This response is directly related to the microenvironment of the respective residue. Depending on intra- and intermolecular interactions, the decay frequency of that atom differs from the other atoms in different molecular environments.

Experimentally, the solution NMR experiment goes as follows: first, a sample is created using labelled isotopes. For proteins, this requires a bacterial minimal culture medium, which contains only compounds labelled with isotopes. After expression, the sample is inserted in a powerful parallel magnetic field, which aligns the atomic spins. In sequence, the parallel magnetic field is turned off, and each atom decays with a different frequency. The decay is captured by detectors that surround the sample, obtaining the resultant magnetic field caused by the spins. By applying a Fourier transform in the resulting data, one can obtain the frequencies that are specific per atoms.

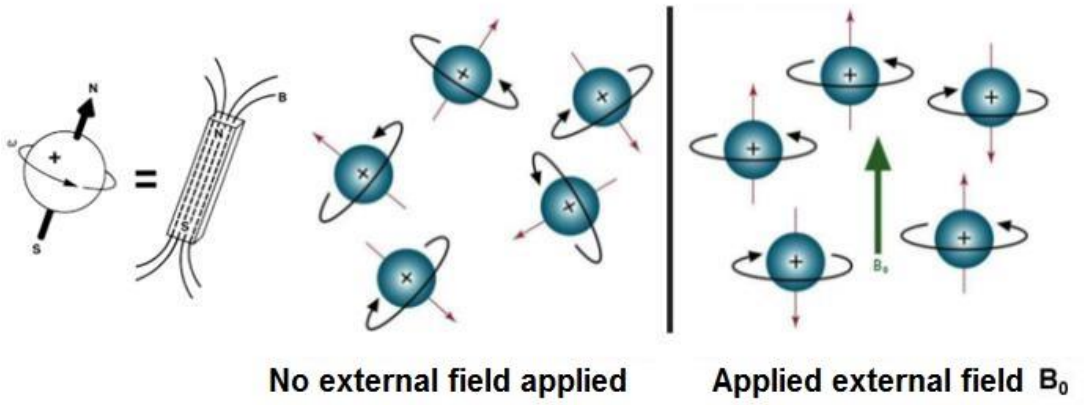

Figure 16: External magnetic field effect on atomic spins. When a strong magnetic field is applied to the sample, a part of the particles has their spins aligned to the externally applied field. When this external field is turned off, the atomic spin relaxation field is acquired by detectors, resulting in the resonance frequencies. Extracted from [108].

Because of the spin-spin coupling, atoms change the magnetic resonance of its neighbours. One important effect caused by atomic relaxation is the Nuclear Overhauser Effect (NOE)[109]. With the NOE, the interatomic distances can be calculated.

The structure of the protein can be discovered with the obtained list of distances. Given the fact that the list is finite and often incomplete, the structure solving the problem via NMR becomes a multi-solution problem. Hence, it allows NMR to calculate an ensemble of possible arrangements which gives insight on the dynamics of the system, resulting in a representative ensemble of the molecule dynamics in water.

One of the disadvantages of solution NMR is the upper threshold on molecular weight. Since it is based on the spin rotational decay, larger molecules directly affect the decay velocity, resulting in NMR spectra with a low resolution. Another disadvantage is the difficulties tied with the sample acquirement given the need for isotopes, producing NMR proteins can be time and money consuming.

2.4.3 Cryo-electron microscopy

The first application of electron microscopy (EM) to biological molecular modelling took place in 1968, where electron micrographs were used to reconstruct the T4 phage tail[110]. In 50 years, the technique improved significantly, achieving results comparable to single-crystal X-ray structure determination [111].

The basic principle of the EM is based on the scattering properties of electrons. This property allows EM to obtain a coherent image of the molecule; however, it is needed to prepare the particles in a cryopreservative environment, employing an ultra-fast freezing method. Doing so causes the water molecules to freeze in an amorphous configuration. Hence, the scattering profile of a coherent electron beam generates a series of magnified images of the sample.

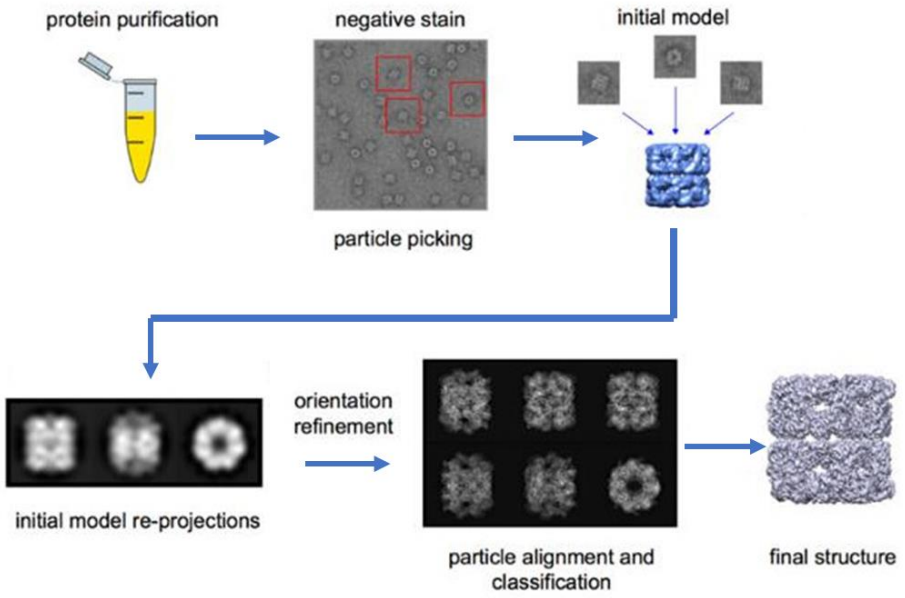

Figure 17: CryoEM structure acquiring scheme: The flash freeze of a purified protein sample undergoes the electron microscopy procedure, while pictures of different protein orientations are taken. Afterwards, the images go through roto translation alignments, resulting in the first model. The structure can be found after orientation refinement and new steps of classification. Extracted from [108].

After the images are acquired, an alignment algorithm is applied to overlay the single-molecule pictures. This organised overlay generates an initial model, which is refined by new sets of images with different orientations. Since it does not need crystals or isotopes, the generation of functional samples for cryo-EM is often more straightforward than the methods previously mentioned. Another advantage in comparison to X-ray crystallography is that the final structure should be a better representative of the native structure.

The most significant disadvantage of cryo-EM is the threshold of molecular weight. Since it has a reasonable level of noise, the method is limited to bigger molecules. This threshold, however, has continuously been lowered by equipment modernisation and better software, nowadays, some cryo-EM models obtain almost atomic resolution for stable and organised models.

2.4.4 Small-angle X-ray scattering

During late 1930, Guinier *et. al*.[112] devised an experiment using the scattering profiles of metal pertained within the small-angle region in the X-ray diffraction pattern to acquire structural data. Based on this work, methods were devised to understand and study colloidal materials. By obtaining the diffraction pattern difference between the pure solvent and the colloid, the small-angle profile can be calculated, resulting in the internal radial density. This method can help understanding protein dynamics in an aqueous solvent since SAXS does not require a crystal.

The internal radial density of the sample results in a model of the molecular envelope of the system. Using this, SAXS is a standard tool to study disordered and partially disordered systems. However, the SAXS model is not a structure, since it is an outcome from a low-resolution analysis of the average envelope. Mathematically, the definition of the final atomic envelope model given by SAXS is a multi-solution problem. Hence, the final model is the most probable one but an incomplete representation of the molecular states.

As said, the main drawback of this technique is the lack of atomic resolution, in comparison to the information obtained via X-ray crystallography, NMR or high-

resolution cryo-EM. Regardless, it gives crucial dynamical data on disordered molecules, which can be challenging for the aforementioned techniques.

The results obtained from these techniques directly affect molecular mechanics methods. On this topic, Chapter 3 explains in detail the MD algorithm and its methods, alongside the shortcomings in applications to IDPs and IDRs.

Chapter 3 – Theoretical background

Several computational methods have been developed in the second half of the twentieth century to study how molecules behave at atomic scales. Some of these were focused on biological systems, like proteins and DNA. In this chapter, the discussion will be focused on molecular mechanics and all-atom molecular dynamics simulations: the principles, key approximations, strengths and limitations.

3.1 Biomolecular simulations

The molecular dynamics (MD) algorithm is a computational method to study the time-dependent changes (e.g. motions, conformational changes) in molecular systems. This method started to be developed in the late 1950s[113] for applications in theoretical physics, but it quickly spread to different areas of science ranging from material structures to biomolecular protein studies.

To be applied within the biomolecular context, the scientific community devised several methodologies through the years. Numerous studies use MD calculations for a plethora of applications, like the study of conformational states of interest in calculating and assessing molecular interactions between different molecules[113,114].

Given the high complexity in describing the molecular dynamics through time, no direct analytical solution can be calculated. Hence, numerical approaches are used to calculate how molecular systems evolve through time. Because MD algorithms are stochastic, the sampled system might not represent all possible conformations of the system. This is known as the MD sampling problem.

Regardless of this disadvantage, MD simulations are used to predict macroscopic thermodynamic properties. This ability to predict observable biophysical characteristics comes from the fact that the MD simulations generate a time-driven ensemble that follows the ergodic hypothesis[35,113–115]. This means that the sampled molecular states are representatives of the macro-ensemble that it belongs to, and its average values should represent the average observables.

3.1.1 Molecular dynamics theoretical framework

The theoretical basis of the MD simulations is that each atom is a hard-sphere. This representation is based on the Born-Oppenheimer approximation, which will be described in detail in section 4.2. On these spherical atoms, a Hamiltonian is built to describe the potential energy field applied to it. With the energetics described, the force exerted on these atoms is calculated via Eq. 3:

$$\vec{F_i} = -\vec{\nabla} U_i(t, \vec{r_1}, \vec{r_2}.\vec{r_3}, \dots, \vec{r_N}) \tag{3}$$

Where F_i is the resulting force in atom i, U_i is the resulting potential energy in atom i, which is a function of all N atoms in the system, and $\vec{\nabla}$ is the gradient operator (Eq. 4):

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}\hat{x} + \frac{\partial}{\partial y}\hat{y} + \frac{\partial}{\partial z}\hat{z}\right) \tag{4}$$

Using Newton's second law, the acceleration in atom i can be calculated as Eq. 5:

$$\vec{a_i} = \frac{\vec{F_i}}{m_i} \tag{5}$$

where m_i is the mass of the atom i.

With the acceleration given by Eq. 5, the velocities and the new positions can be calculated by Eq. 6 and Eq. 7:

$$\vec{v_i}(\vec{r_i}t + \Delta t) = \int_t^{t+\Delta t} \vec{a_i}\, dt \tag{6}$$

$$\vec{r_i}(\vec{r_i}, t + \Delta t) = \int_t^{t+\Delta t} \vec{v_i} dt \tag{7}$$

Where $\vec{v_i}$ and $\vec{r_i}$ is the velocity and position of atom i in the time $(t + \Delta t)$ respectively.

With this approach, the algorithm can calculate the evolution in time of the atomic positions. This procedure can be repeated for any number of times, as described in Figure 18.

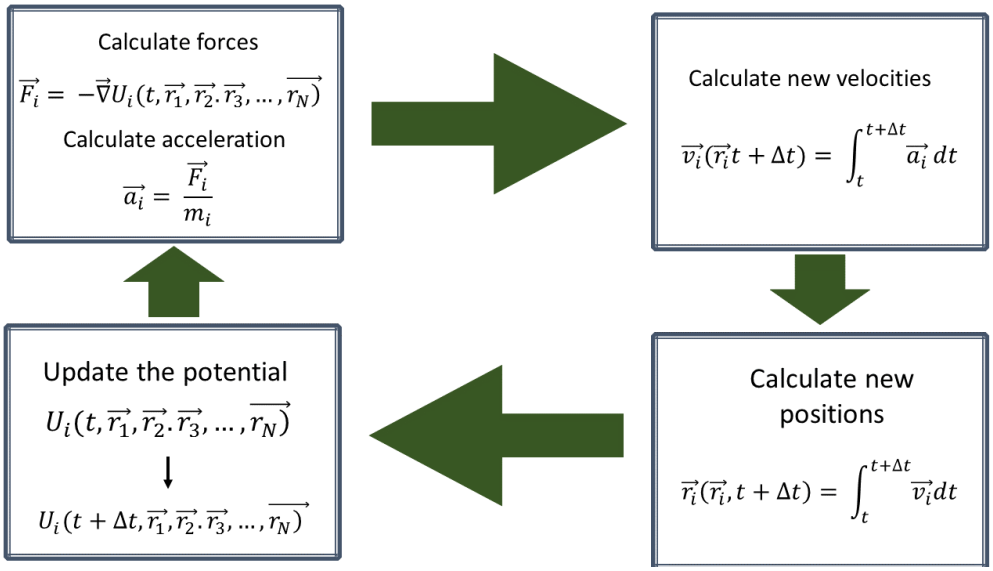

Figure 18: Molecular dynamics framework of integration cycles

These integration cycles are usually calculated by using a leapfrog integration method[113,114]. This method defines the velocities and the positions as time-dependent Taylor series, which can be readily integrated to obtain its related primitive function.

Usually, the timestep used in simulations is 2 fs. This value is used because it's higher than the period of vibration for bonds between non-hydrogen atoms (C, N S, O). Because it is higher than the vibration period of bonds containing hydrogen, these bonds are restrained throughout the simulation, usually using the LINCS distance restraint algorithm[113].

The cornerstone of the integration cycles is the potential energy function. The set of descriptors and the functions required to define a potential is called a force-field.

3.1.2 Potential energy description and force fields

The first thing to be assigned to a system before a simulation is its atomic descriptions. A molecular energy potential is required to describe how the inter

and intramolecular interactions work. These energy descriptors are called force fields (FF).

AMBER FF is one of the most widely used empirical force fields in biomolecular simulations community, and the one used in this work. The potential energy of the system can be described by decomposing the potential in two layers: internal and non-bonded potentials (Eq. 7).

$$E_{total} = E_{Internal} + E_{non-bonded} \tag{7}$$

For the intra-molecular energies, the AMBER model describes the system in a series of harmonic oscillators for bond, angle and dihedral vibrations (Eq. 8 and Eq. 9):

$$E_{Internal} = E_{bonds} + E_{angles} + E_{dihedrals} \tag{8}$$

$$E_{Internal} = \sum_{bonds} k_b(r_0 - r_i)^2 + \sum_{angles} k_a(\phi_0 - \phi_i)^2 + \sum_{dihedrals} \frac{1}{2} V_n[1 + cos(n\theta + \gamma)] \tag{9}$$

Where k_b is the force constant for the bonds, with equilibrium length r_0, k_a is the angles force constant for the angles, with equilibrium angles of φ_0. V_n is the force constant for the torsions, with phase value as γ, and n is the torsion multiplicity.

The bond and angle terms are harmonic oscillators. Both use a quadratic potential to emulate the energy well that describes the motion of a specific set of atoms. This approximation rises certain drawbacks for MD simulations. One of the most crucial disadvantages is the impossibility to simulate the breaking and creation of covalent bonds, given the fact that harmonic quadratic potential never goes to zero.

To better describe the torsional values, the dihedral term is a Fourier series for the same quartet of atoms. In other words, the same torsional set of atoms (i, j, k, l) may contain a series of terms to describe its torsions more accurately. The

torsional potential is described as a series because of the different atomic orbital hybridisations for the same atom type. Also, some special torsions, such as the amine bond in the protein backbone, require an out-of-plane dihedral potential to maintain the configuration of non-rotatable bonds. A detailed description of the generation of new parameters on bonded potentials is explained in section 4.2.

The second layer for the potential energy is the non-bonded interaction potential $E_{non-bonded}$. These terms are often described as electrostatic Coulomb potential and a 12-6 Lennard-Jones potential (Eq. 10):

$$E_{non-bonded} = \sum_i \sum_j \frac{C q_i q_j}{\epsilon r_{ij}} + \sum_i \sum_j 4\varepsilon\left(\frac{r_m}{r_{ij}}^{12} - \frac{2r_m^6}{r_{ij}^6}\right) \quad (10)$$

where C is a unit constant, q_i is the respective charge of the atom i, ϵ is the electrostatic permissivity of the medium, r_{ij} is the distance between atom i and j, ε is the Lennard-Jones well depth and r_m are the Van der Waals constants for the respective atoms. The Coulombic term describes charge-charge interactions. It is critical in describing hydrogen bond formation, polar-polar interactions, and solvent-protein interactions in polar environments. The Lennard-Jones (LJ) potential is an empirical description of how electronic clouds interact. The first r^{12} term describes Pauli repulsion due to orbital overlap, and the r^6 represents short-term attraction due to orbital dispersion forces[113]. The Lennard-Jones potential is shown in Figure 19.

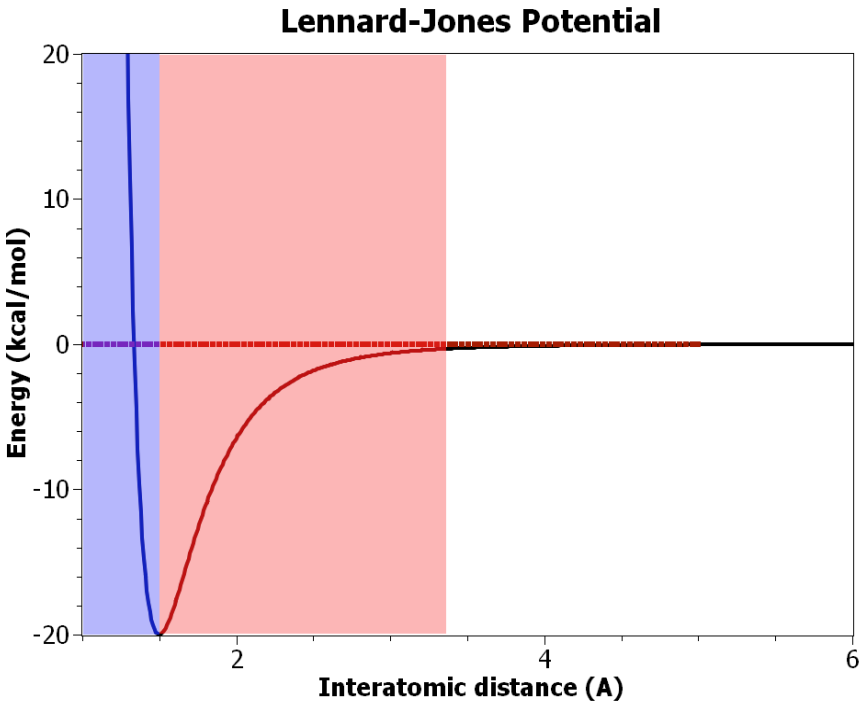

Figure 19: The Lennard-Jones potential: in blue, the repulsive region and in red the attractive part.

One necessary approximation that is taken regarding non-bonded interaction is the exclusion list for bonded chains. Mainly, the energetics between bonded atoms are primarily at the quantum level. Therefore, the MD algorithm does not calculate the non-bonded interaction for connected atoms. This approximation is crucial to reduce the amount of computational power required to simulate the system of interest.

The number of atomic pairs for interactions grows exponentially with the number of atoms, as shown in Equation 10. For that, a system decomposition in layers needs to be applied to reduce computational requirements. The most common approach is to use the Particle Mesh Ewald[116] algorithm.

Particle Mesh Ewald consists of an interpolation grid method in the reciprocal space. The interaction potential is divided into two layers (Eq. 11):

$$U_{Total} = U_{short-range} + U_{long-range} \quad (11)$$

In the short-range layer, the summation of the energy is made directly in the real space, calculating the pair-wise interactions one by one, as shown in Eq. 12.

$$U_{short-range} = \sum_i \sum_j U_{Elec}(r_{i,j}) + U_{LJ}(r_{i,j}) \quad (12)$$

Using a pair-wise calculation on the whole system would be very demanding computationally. Therefore, the long-range term using PME will be (Eq. 13):

$$U_{long-range} = \int \widetilde{U}_{Long-range} * |\tilde{\rho}|^2 dk \quad (13)$$

Where $\widetilde{U}_{long-range}$ and $\tilde{\rho}$ are the Fourier transform of the potential and charge density, respectively. Both components of the interaction potential converge quickly in their own space, resulting in greater accuracy and a reduction of computational requirement. The cut-off distance that defined the boundary between PME description and pairwise energetic description is system dependent. This cut off radius is usually, for simulation in equilibrium, on the nm scale.

The values used in this work (1 nm) comes from the fact that the simulation box is created with this distance from the edge of the protein, hence, the surface interaction should be calculated in the short-range pairwise method[113].

Given the periodicity implicit in the Fourier space, this method requires the use of periodic boundary conditions (PBC)[113].

The application of PBC in molecular dynamics simulations not only solves the periodicity required by PME but also to avoids the appearance of finite-size effects[113,114]. The PBC is exemplified in Figure 20, which emulates the existence

of virtual boxes, where a copy of the particle that enters the box wall exits on the other side. As such, the number of particles in the box remains constant.

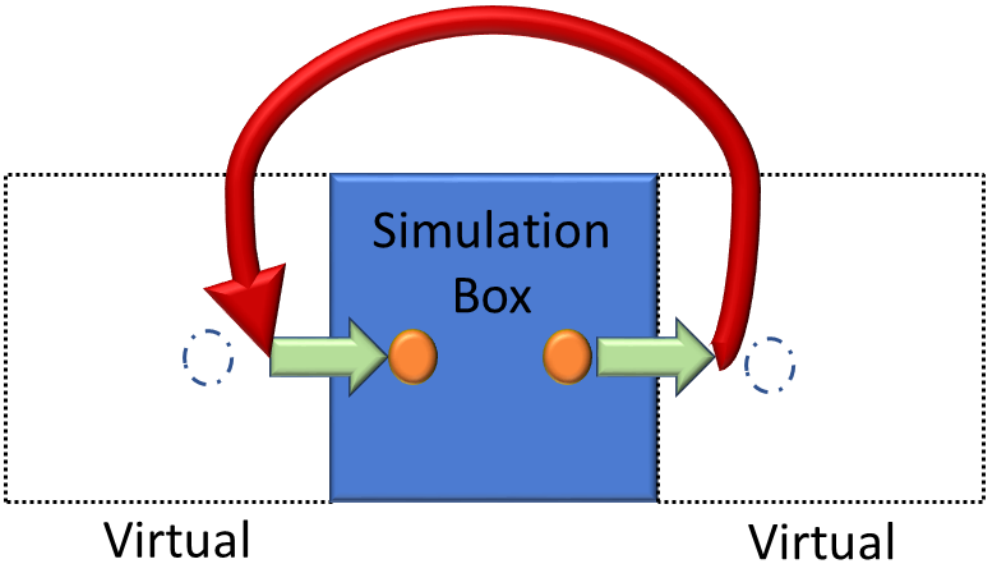

Figure 20: Periodic boundary condition (PBC) representations, within a 3D box, the escaping water molecules are placed on the other side of its simulation box, creating in a virtually infinite system. Made by the author.

In all the work done in this thesis, a cubic box was used centred around the protein, but there are several different unitary box types, *i.e.* triclinic and dodecahedral boxes.

3.1.3 Simulation environment – solvent and ions

After defining the box and its periodic boundaries, the solvation box and its components need to be defined. In MD, there are two methods to describe how the molecular system will be solvated: implicit or explicit solvent.

Given the limited computational power of early MD studies, implicit solvent methods were broadly used. These are based on adding new terms to the system Hamiltonian, resulting in the total energy of the system as (Eq. 14):

$$E_{Total} = E_{Bonded} + E_{Non-bonded} + E_{Solvation} \quad (14)$$

The most popular method to use as an implicit solvent in a molecular dynamics simulation is to solve the Poisson-Boltzmann equation[117]. The computational requirement to solve the Poisson-Boltzmann equation is significant, given its high amount of integral calculations required to define the electrostatic density and the resulting potential.

A known and popular approximation is the linearised Poisson-Boltzmann, also known as Generalized Born (GB). This approximation uses an extra parameter known as Born radii, which are specific for each atom, to approximate the PB integrals to a linear Coulombic-like interaction potential. However, this requires an accurate estimation of Born radii. Several works recently have been published using implicit solvent[118–120], either GB or PB. Regardless, the number of works published with implicit solvent has been decreasing with the rise of more powerful computers, development of consumer-range GPUs and increased popularity of explicit representation of the solvent.

Explicit solvent uses solvent models to simulate a more realistic and complete environment. Given the fact that any molecule can be used as the solvent, a range of applications have been devised, from solvent-solvent to protein-solvent interactions. For the simulations of biomolecules, a solvent model that adequately represents a water environment is crucial to model the system accurately. The most used water model is the three-site transferable intermolecular potential or TIP3P water model. Devised by Jorgensen *et al*.[121] in 1983, it consists of a three atoms molecular scaffold of an H_2O molecule, using experimentally structural parameters.

Several other water models have been developed[122], such as TIP4P[123] and OPC[124]. A more in-depth discussion on water models and their applicability will be carried out in Section 4.4.

The biological environment needs to be reflected in the simulation for a more accurate picture of protein simulations. Therefore, ions are added to the simulation box to neutralise the simulation box and to emulate the free ions in a physiological concentration.

The addition of ions can vary significantly between applications. Study of ions in MD simulations shows that the different concentration and ionic composition does affect assessed states throughout the sampling run[113].

Usually, the addition of ions is made by calculating the overall charge distribution throughout the box. Upon inspection of the resulting electrostatic potential, a particular solvent component of the system, i.e. a specific water molecule, is swapped by a cation or an anion in areas which will minimise the electrostatic potential. With the ions added to the system, neutralising the simulation box, and – most commonly – mimicking physiological concentration of ions, the setup is completed.

3.1.4 Structural energy minimisation procedures

To assure the structural stability of the system, energy minimisation procedures are needed. Despite the method used to acquire the protein structure, (experimental procedures or computational modelling), the arrangement might have slight atomic overlays, which may result in an unstable molecular dynamic.

If the total energy is too high, the resulting force vector will have a momentaneous high intensity. When such an event occurs, the integration cycle crashes, given the fact that the sudden motion created by this force may cause the system box to explode.

The energy minimisation of the system prior to any MD simulations is one approach to reduce the probability of a system crash during the equilibration cycle. The two most commonly used methods of optimisation to a minimum in molecular mechanics are the steepest descent algorithm (also known as gradient descent), and the conjugate gradient algorithm. As discussed in Section 2.2, the configuration of the system may get trapped in a local energy well, so a more robust method of minimisation might be needed to sample closer to the global minimum.

The steepest descent (SD)[125] algorithm is a method based on the derivative of the variable to be optimised, in this case, the potential energy. As such, for every minimisation cycle, the 3N dimension position vector r_{n+1} can be calculated by Eq. 15:

$$r_{n+1} = r_n + \frac{F_n}{MAX(F_n)} h_n \quad (15)$$

Where r_n is the starting position, F_n is the force applied in that atom, $MAX(F_n)$ IS the maximum force applied in any atom and h_n is the atomic displacement for that cycle. Hence, this calculation goes over all atoms in the system, and the convergence criterion is either a predefined number of cycles or an upper threshold of the system highest force.

This method is simple and often quick, but given the fact it requires only orthogonal gradients to be calculated, it is prone to get trapped within local energetic wells [125]. This effect may be attenuated by tuning the parameter h (the maximum allowed displacement per cycle), with incremental decreases in a series of cycles to improve the final configuration.

The second approach, conjugate gradients (CG), requires a more complex mathematical approach[126]. Since SD uses a serial orthogonal approach, it may not be as efficient, since it will approach the solution in a zig-zag pattern[125,126]. Conjugate gradients use the fact that the gradient ∇ of the function f can be described as (Eq. 16):

$$\nabla f(x) = Qx - b \quad (16)$$

Where Q is the Hessian matrix and b is the configurational vector where f has its critical point. In the case of MD application, f is the potential energy U and $\nabla f(x)$ is the force vector. The criterion of minimisation needs to be that the resulting force is a null vector, therefore (Eq. 17):

$$Qx = b \quad (17)$$

Suppose a solution x* existing for Equation 17. This solution can be described as a linear combination of an orthogonal basis d, defined by vectors d_i, so (Eq. 18):

$$x* = \sum_{1}^{n} \alpha_i d_i \tag{18}$$

At the minimum (Eq. 19):

$$Qx* = b = \sum_{1}^{n} \alpha_i Q d_i \tag{19}$$

Therefore, using the configurations of the system, the basis *d* can be defined.

This method, since it scans the minima more efficiently, is more computationally demanding and slower than steepest descent. Regardless, it still is a derivative-based method, since it depends on the Hessian matrix for calculation. Even yielding more accurate results than steepest descent on finding the closest minimum point, it may still get trapped into deep local minima.

From the biomolecular standpoint, the configuration that the protein starts is often close to the global minimum, since it should be a representative of its native state. Regardless, problems with experimental data when resolving structures may result in erroneous side chains configurations, which minimisation procedures resolve. With the system energetically minimised, the thermodynamic variables of the system in question need to be defined.

3.1.5 Thermodynamic macro variables

As stated before, given the ergodic status of MD, accurately defining the characteristics of the single cell is crucial to predicting macromolecular system properties. Hence, the first variable that needs to be resolved is the temperature of the system.

Theoretically, the minimised static model acquired after the minimisation procedure has a temperature of 0 K since there is no dynamical atomic motion assigned to it. To increase the temperature, one needs to assign a specific condition of the experiment.

These conditions are macro variables the ensemble will have as constants. For a temperature increase of the system (equilibration runs), the NVT (or canonical) ensemble is used.

The NVT ensemble is a statistical ensemble that represents the probability of accessible states in a predefined configuration. In this case, the variables set as constants are the number of particles (N), the volume of the system (V) and the temperature (T). As such, the probabilities assigned to each microstate of the system are (Eq. 20):

$$\rho = \frac{e^{\frac{-E}{kT}}}{Z} \tag{20}$$

Where *k* the Boltzmann constant, T is temperature, E is the state energy, and *Z* is the partition function defined in Eq. 21:

$$Z = \sum_{1}^{n} e^{\frac{-En}{kT}} \tag{21}$$

Since the probabilities in this ensemble do not depend on any other variable, i.e. pressure, NVT ensemble can be used for heating of the system in the equilibration phase.

To be assured that the heating will not change nor affect its starting structural conformation, position restraints are applied. Often, these restraints are utilised via the addition of a harmonic potential on selected protein atoms, i.e. protein backbone[113].

Afterwards, a distribution of velocities is applied to the atoms to reach a macrostate temperature T in a restrained configuration that follows Eq. 22:

$$\sum_{i=1}^{N} m_i |V_i|^2 /_2 = \frac{kT}{2}(3N - Nc) \tag{22}$$

Where m_i is the mass and $|V_i|$ is the average velocity of atom i, N is the total number of particles, and N_c is the number of restrained components.

This is a Boltzmann-Maxwell distribution of velocities which reaches the desired temperature[127]. A thermostat algorithm is applied to the system to update the temperature throughout the integration timesteps. The simplest method to change the temperature is by rescaling the velocity for every new step to the temperature T, so eq. 22 turns into Eq. 23:

$$\sum_{i=1}^{N} m_i |V_i|^2 /_2 \rightarrow \sum_{i=1}^{N} m_i \gamma |V_i|^2 /_2 \quad (23)$$

Where γ is (Eq. 24):

$$\gamma = \sqrt{T /_{T_i}} \quad (24)$$

and T_i is the temperature of step i.

Since the temperature rescales directly with the velocity, these methods do not allow thermal fluctuations through the system. Based on velocity scaling, the Berendsen thermostat was devised. The idea follows that a weak coupling of the system updates the average temperature to a temperature bath. Because of the weak coupling, the temperature does not scale directly with the velocity [114,127]. Therefore the γ for a Berendsen thermostat is (Eq. 25):

$$\gamma = \sqrt{1 + \frac{\Delta t}{\tau} (T /_{T_i} - 1)} \quad (25)$$

where τ is a coupling term called “rise time” This term controls how strongly the system ‘feels’ the temperature bath. The problem of scaling methods is the fact

that they do not allow stochastic variations on the velocity since they scale it directly. Other thermostats address this problem, such as Nosé-Hoover thermostat[114]. Often these thermostats require more computational time but can simulate a proper canonical ensemble. A way of reaching a middle ground is to use a method called velocity rescaling[128]. This method, implemented in GROMACS, add a Wiener stochastic function to the γ term. Therefore, the velocity scaling becomes randomised, sampling a full canonical ensemble. This thermostat was the one used throughout this work, since its faster and generates the proper required ensemble. To simulate proper experimental procedures, molecular simulations are calculated in room temperature, with average temperature values varying within the 298-300K. The rise time is usually on the picosecond scale, which was assigned to this work the value of 0.1 ps. This means that the temperature should be updated every 50 steps of simulations, which allows the atoms to disperse the temperature evenly between their neighbour.

3.1.6 Pressure equilibration and barostats

After the stage of thermal equilibration, the volume re-configuration needs to be set. The reason is the box setup does not change in the NVT ensemble, so it may not be the most accurate volume for the box of interest. This arises the need for equilibration to set up the remaining macro thermodynamic variables, such as pressure. Another reason this next step is required is that, especially for a biomolecule, the experiments aimed to be modelled are done at the constant pressure environment. Hence, in the second stage of equilibration, we modify the variables defined as constants, switching from the NVT to the NPT ensemble (N – number of particles, P – pressure and T – temperature, also known as an isothermal-isobaric ensemble).

The system needs to be coupled to a pressure control (barostat) as well as a temperature bath (thermostat). The two most popular ways are the weak Berendsen coupling[125] and the Parrinello-Rahman barostat[129]. The Berendsen coupling barostat functions similarly to its thermostat since it scales the volume of the box though time to attain a predefined pressure. This barostat belongs to

a class called isotropic scaling since it does not change the overall shape of the box, just the size equally in all dimensions.

The Parrinello-Rahman, on the other hand, does an anisotropic scaling and is calculated by the scaling differently atomic coordinates throughout the system. Again, this comparison between two barostats follows the same parallels for the thermostats: the Berendsen barostat, albeit simple, does not allow local fluctuations to sample a correct NPT ensemble, which Parrinello-Rahman does.

Given the requirement to generate proper NPT ensemble, the Parrinelo-Rahman barostat is used and regulated to sampled states with a 1 bar pressure with a 2 ps rescaling time. The rescaling time is the period the system requires to update the box characteristics to achieve the reference pressure, which requires a reasonable time to equilibrate and reorganize the particles within the box. Hence, it is usually 10 times slower than the rise time for the thermostat.

3.1.7 Production simulations and analysis

With the thermodynamic macro variables defined and system equilibrated, the dynamic ensemble can be calculated. As described before, the integration cycle drives the calculations. After the equilibration steps, the atomic harmonic restrains are disabled, and the protein can be properly sampled and simulated. The parameters and the methods are usually the same as the pressure equilibrations, in an NPT configuration, to be able to sample experimental properties of the system in question.

Typically, to ensure that the system has been properly equilibrated, a series of metrics are used. The most common one is the root-mean-square deviation (RMSD) in the function of time and the root-mean-square fluctuation (RMSF)[113].

3.2 Force fields and parameter development

A crucial problem in molecular mechanics is the sampling of molecular ensembles[113]. MD simulations are restrained to a finite amount of time and may thus prevent the complete sample of the respective configurational space. This

inability also occurs given the fact that the energetic potential functions may be inaccurate and may bias the conformations that the system can acquire.

In this section, it will be discussed the inner workings of the most used conventional force fields.

3.2.1 Development of force-fields: parameters and functions

There are several ways of decomposing the energetic description of a molecular system, given the fact that the classical Hamiltonians have the additive property[113]. The most common way to describe the molecular structure is to define a series of characteristics such as:

- Partial atomic charge
- Atomic radius and mass
- Hybridisation and bond configuration
- Structural energy for angles
- Torsional energetic landscape

The partial atomic charge, albeit a useful concept, is not a quantum mechanical molecular characteristic. Assigning partial charges is a tool to classically mimic the electronic distribution in a molecule caused by the atomic electronegativities.

There are several ways to calculate the partial atomic charges for the complex molecules.

Quantum mechanical (QM) methods calculate most of these atomic charge descriptors. For a better explanation on QM calculations for biomolecular simulations, they are usually divided into different layers of complexity: Hartree-Fock (HF), density functional theory (DFT), and semi-empirical methods (SE).

Hartree-Fock method (HF) focuses on solving the Schrödinger equation for the molecular wave function. In 1927, D.R. Hartree developed a procedure which he called Self-Consistent Field (SCF)[130], to approximate the molecular wave function by assigning a linear combination of atomic orbitals for the time-independent Schrödinger equation. Two approximations that allowed the solution of the wavefunction problem is the Born-Oppenheimer and the lack of electron motion correlation.

The Born-Oppenheimer approximation dictates that the nucleus motion is not affected by the electronic movement, given the fact that the nucleus is 1000 times heavier than the electron. Therefore, the electron-nucleus motion is decorrelated. Hence, the solution of the problem can be narrowed down to electronic motion only. Along with that, HF decorrelates the motion between electrons, since the cross-motion integrals are analytically unsolvable[131].

These approximations allow us to calculate the electronic distribution of a molecule, within a range of accuracy, but the decorrelation approximation brings some setbacks: Given the fact that electronic interactions are neglected, there is an undervaluation of the emission energetics, and orbital distribution are often misplaced[131]. However, HF has a reasonably accurate charge fitting for organic molecules, since they are usually calculated in their electronic ground state and omission of the electron correlation effects is not significantly detrimental.

Density functional theory (DFT) is hailed as a highly accurate method for QM calculations. It relies on the concept that the electronic properties of a many-body system could be uniquely dependent on its electron density. With the theoretical footing based on two Honenberg-Kohn (H-K) theorems, it has been frequently used for computational *ab initio* molecular studies. The first H-K theorem says that an electron density yields a unique set of ground-state properties. Hence, the ground-state calculations can be done using electron density functions which reduces the 3N dimensional problem to a continuous distribution. To couple with that, the second H-K theorem says that the ground state can be acquired by minimising the energy functional assigned to the molecule[131].

Another way to calculate partial atomic charges is through semi-empirical approaches, which often hail a high level of accuracy for biomolecular applications (i.e. organic fragments, small molecules of biological interest).

Semi-empirical calculations use approximations to emulate molecular orbitals. John Pople and coworkers introduced these methods during the 1970s[132]. Their work introduced methods such as NDDO (neglect of diatomic differential overlap). The goal of these methods was to fit reduced orbitals to *ab initio* calculations, minimising the computational power requirements. Nowadays, these methods are often replaced by the second generation semi-empirical QM calculations.

Based mainly on the NDDO approach, several improved methods were created, such as AM1[133], PM3[134] and PM6[134]. The second-generation semi-empirical quantum mechanical calculation methods are more accurate, given the improvement on computational power, which allows a higher number of parameters to be used on said models.

On a usage point of view for organic molecules, the Austin model 1 (AM1) is one of the most used, since it is frequently used for calculations of charges using AM1 with bond charge correction (AM1-BCC), developed by Jakalian *et al*. [135]. AM1 is based on Pople's work using NDDO by parametrising the repulsion between close atoms within the electronic structure. The BCC (bond correction charges) is a set of additive terms to correct the AM1 population charges. Given the fact the AM1-BCC was trained on a set of organic molecules, it describes the electronic behaviour of charges accurately. However, it is not well suited for calculation of charges of metal-organic complexes.

Usage of QM methods allows parametrisation of the valence structure using structural optimisation in the ground state. Since throughout an MD simulation, the hybridisation configuration does not change, the right structure must be built. Optimisation algorithms can use several different QM levels of theory to assess the right scaffold of a molecule and to calculate equilibrium bond lengths and angles values.

On classical mechanics, the bonds are modelled using a harmonic oscillator potential function, which requires a parameter that represents the strength of the said bond. This parameter is called the bond spring constant (k). There are several methods to obtain bond k, which does not rely on *ab initio* QM calculations. One of such ways is to use the acquire using solution nuclear magnetic resonance[113] (NMR) experiments.

Similar to bonds, the angle energetic functional can be modelled using a harmonic potential function as well, assigning a force spring constant k for the angle energetics, and a starting angle of equilibrium for the respective configuration. The angles force constants can also be calculated similarly to the bonds, either via QM or experimental means[136].

As explained in Section 2.4, the overall structural states accessed in a protein structure are mainly caused by the accessible Φ/Ψ backbone torsional dihedrals per specific residue. As such, the molecular torsions should be modelled in more details, requiring a sophisticated function that can describe its periodicity, as shown in Figure 21. Each of the Fourier terms carries a force constant V_n, which the strength of that function regarding the other terms.

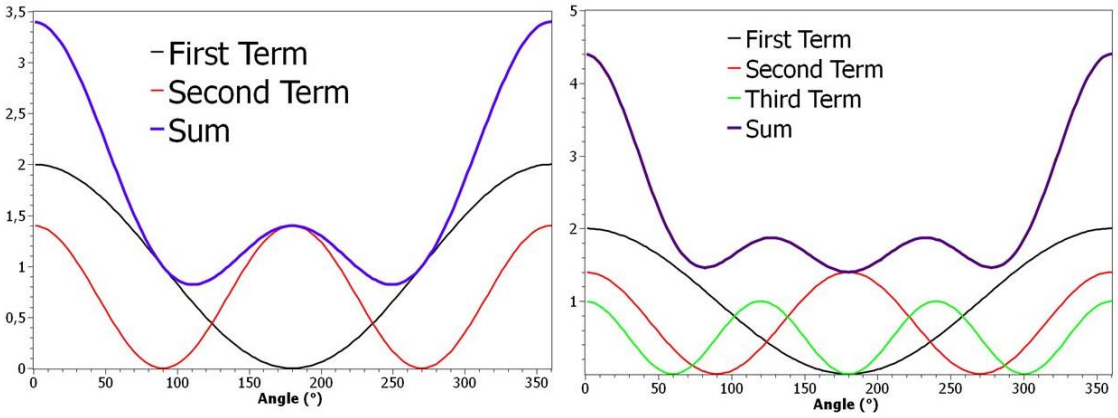

Figure 21: Example of successive addition on the Fourier series for a description of dihedral potential. Made by the author.

Fourier series is a mathematical method able to describe different torsional configurations by using different periodicities. A known way to calculate these terms and to parametrise torsions are to scan the torsional landscape using an accurate QM theory level and from the obtained surface, calculate the vibrational frequencies via Fourier transform.

Another function that is crucial for the molecular description is the out-of-plane (improper) dihedral. Improper dihedrals describe atoms that are required to stay within a specific plane (i.e. aromatic rings). This is also vital for protein backbones, given the fact that the peptide bond between residues is planar.

Based on these functions, there are a series of different force fields available to the scientific community nowadays. The most popular ones include AMBER (Assisted Model Building with Energetic Refinement)[137], OPLS (Optimised Potential for Liquids Simulations)[138], GROMOS (Groningen Molecular Simulation)[139], and CHARMM (Chemistry at HARvard Macromolecular Mechanics)[140].

The difference between these forcefields are on how the parameters were acquired and to which experimental data they were fitted. An example that will be discussed in Sections 4.5 and 6.1 is the differences between AMBER03ws and AMBER99SB-ILDN.

Given the fact that benchmarking of new force fields is biased towards the experimental data available to use to evaluate the results, the applicability of different force fields varies. The result is a series of force fields for various applications. Since the most accessible data for biomolecules are in the condensed state (i.e. for proteins, near-native folded conformations) most of the force fields are focused on replicating folded protein results. For a time, this results in a lack of force fields that can replicate data for different conformational states accurately of biomolecules like fibril-like structures, molten globule-like conformations, or unfolded configurations [141].

Alongside with the right protein parametrisation, the accuracy of the results will also depend on the capacity of the environment to interact with the solute in a satisfactory way. Hence, how the solvent is described is pivotal for a proper simulation and to be sure of how reliable the generated data is[115].

3.3 Methods for analysis of molecule dynamics

3.3.1 Root-mean-square deviation

Root-mean-square deviation (RMSD) metric is used to analyse the conformational changes of a molecular group through time in comparison to a reference configuration. RMSD is defined as (Eq. 26):

$$RMSD\ (t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i(t) - \overline{x_i})^2 + (y(t) - \overline{y_i})^2 + (z_i(t) - \overline{z_i})^2} \quad (26)$$

Where N is the number of atoms, and $\overline{x_i}$, $\overline{y_i}$ and $\overline{z_i}$ are the reference positions for the atom i.

3.3.2 Root-mean-square fluctuation

Root-mean-square fluctuation (RMSF) is used to analyse the average atomic fluctuation in comparison to a reference configuration. It is defined as (Eq. 27):

$$RMSF_{atom} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i - \overline{x})^2 + (y_i - \overline{y})^2 + (z_i - \overline{z})^2} \quad (27)$$

Where T is the total number of frames of the trajectory, and $\overline{x_i}$, $\overline{y_i}$ and $\overline{z_i}$ are the reference positions for the atom i. RMSFs are often calculated for a group of atoms (i.e. residues) for the analysis of local spatial fluctuation.

3.3.3 Principal component analysis

To extract the essential dynamics of a molecule in a MD simulation, principal component analysis (PCA) can be used. By calculating the 3Nx3N matrix of average covariance C, where N is the number of atoms in the molecule (Eq. 28):

$$C = \langle (X_i - \overline{X_i})(X_j - \overline{X_j}) \rangle \quad (28)$$

Upon diagonalisation of matrix C by using an orthonormal transformation R the diagonal matrix of eigenvalues can be obtained (Eq. 29):

$$R^T C R = diag(\lambda_1, \lambda_2 \ldots \lambda_{3N}) \quad (29)$$

Where the columns in R is the eigenvectors and λ_i are the eigenvalues of C, respectively. These eigenvectors are known as the principal components (PC) of the trajectory, and they represent directions which the molecule had the largest correlated motion. The trajectories projection into the PCs can be obtained via (Eq. 30):

$$p(t) = R^T(r(t) - \bar{r}) \quad (30)$$

Where p is a position matrix as a function of time projected into the new basis R.

3.4 Interaction evaluation methods

Aside from defining the energetics of molecular dynamics, these energy descriptors are used in different applications, such as molecular docking and interaction energetics for MD.

3.4.1 Molecular docking

Docking methods are used to understand and predict molecular interaction between partners. As a primary application, the interaction between a protein receptor and a drug-like organic molecule is calculated using a preselected energy scoring function. Commonly, the sum of the intermolecular electrostatic component and its Lennard-Jones component between the components is used as the scoring function (Eq. 31):

$$E_{interaction} = \sum_{i-Rec} \sum_{j-Lig} \frac{C q_i q_j}{\epsilon r_{ij}} \sum_{i-Rec} \sum_{j-Lig} 4\varepsilon \left(\frac{r_m}{r_{ij}}^{12} - \frac{2r_m^6}{r_{ij}}^{6}\right) \quad (31)$$

However, in Equation 31, the indices i and j are comprised only by the inter-molecular atoms, resulting in the vacuum interaction energy between two components. Since this calculation can be done fast, they are used for high-throughput virtual screening of databases of drug-like ligands[142].

The issue with these calculations is the conformational space that both interacting components can achieve. Flexible ligands with several rotatable bonds may have a substantial amount of conformations which needs to be tested against a considerable number of receptor configurations. Usually, these calculations consider the receptor rigid to reduce the amount of computational power required. For the ligands, which typically is a small organic molecule, the algorithm generates a set of rigid conformers, emulating ligand flexibility. Figure *22* exemplifies this procedure.

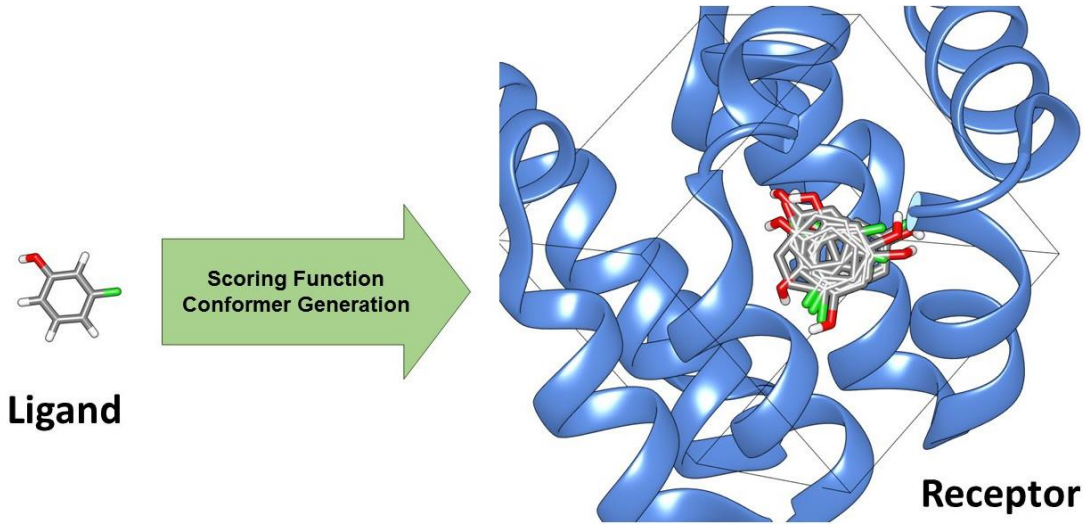

Figure 22: Molecular docking of molecules into protein receptors – the description of the ligand uses classical force-fields, and a set of conformers is generated prior docking, and the best-evaluated conformers are selected. Made by the author.

The AMBER scoring function is used by several docking software, like Autodock[143] and UCSF Dock 6[144]. Several other docking programs have been developed, including open-source (Vina) and commercial such as MOE and SeeSAR. SeeSAR uses a knowledge-based scoring function named HYDE[145], which uses a regression model scoring function to calculate interaction and desolvation energy in a fast and reliable way.

Alongside with the rigid body approximations, molecular docking has another drawback: it is a single point calculation. This means that it does not use a dynamical ensemble, but a single conformation to a single calculation. Albeit fast, molecular docking loses the information on dynamics and it is not considered a free energy of binding. One of the simplest methods to calculate the free energy of binding is MMPBSA[113,146].

3.4.2 Molecular Mechanics Poisson-Boltzmann surface area (MMPBSA) interaction energy

In MMPBSA framework, two components need to be calculated for a proper assessment of the free energy values related to a thermodynamics event: the enthalpy change ΔH and the entropy term ΔS (Eq. 32). MMPBSA is a standard

method to calculate interaction free energetics in complexes. It decomposes the ΔG as:

$$\Delta G = \Delta H - T\Delta S = \Delta E_{interaction} + \Delta E_{desolv-polar} + \Delta E_{desolv-nonpolar} + T\Delta S_{normal\ modes} \quad (32)$$

Where $\Delta E_{interaction}$ is the inter-molecular vacuum interaction, $\Delta E_{desolv-polar}$ is the polar desolvation energy, $\Delta E_{desolv-nonpolar}$ is the nonpolar desolvation energy and $\Delta S_{normal\ modes}$ entropy calculated via normal modes, and T is the temperature of the system.

To solve the free energy calculation, it uses only the simulation of the complexed state. Hence, it is known as an end-point free energy method. Similarly, to the molecular docking, MMPBSA calculates the interaction energy between two bound molecular entities; however, it does so for all frames of a trajectory. The procedure to extract free energy is exemplified in Figure *23*.

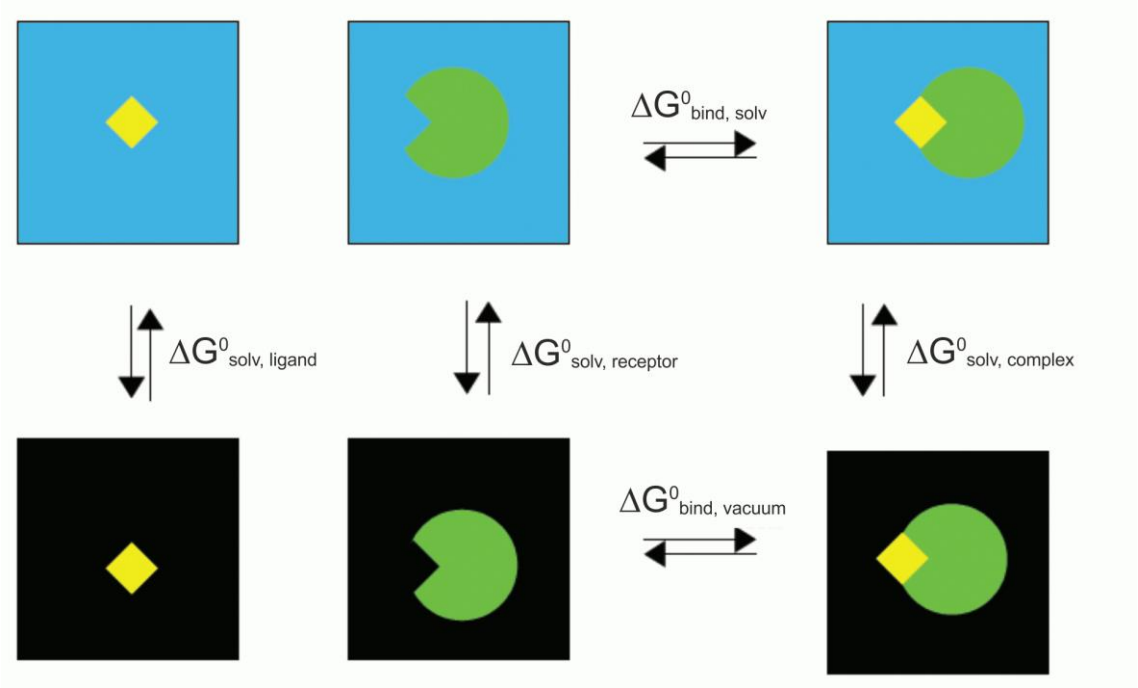

Figure 23: The thermodynamic cycle used in MMPBSA. Since the binding event is often a reversible process, the thermodynamic process can be described in a purely empirical way: the complex Is moved from a solvated environment to vacuum. The energetic cost of this displacement is the receptor-ligand complex desolvation energy. From the complex vacuum state,

the components of the system are dissembled, obtaining the interaction energy between system constituents. Finally, the separated pieces are moved back into a solvated box, resulting in the solvation energy per component. The sum of all these terms is the overall binding free energy in solution. Extracted from [147]

To calculate the polar desolvation energy, MMPBSA used the Poisson-Boltzmann equation (Eq. 33):

$$\nabla \cdot \varepsilon(\vec{r})\nabla\phi(\vec{r}) = -4\pi\rho(\vec{r}) \quad (33)$$

where $\varepsilon(\vec{r})$ is a predefined dielectric distribution function for the solvated molecular system, $\phi(\vec{r})$is the potential distribution function, and $\rho(\vec{r})$is the fixed atomic charge density. This equation is either solved by finite-difference[126] solution or by the Generalised-Born linear approximation[148]. By solving it, the $\Delta\mathrm{Gpolar}$ can be extracted from the electrostatic density and its related energy potential.

Moreover, to calculate the apolar desolvation energy, a linear empirical formula is used to emulate the energy cost do solvate-desolvate apolar regions. The most common is the accessible surface area calculation (Eq. 34)

$$\Delta G_{non-polar} = \gamma * SASA + b \quad (34)$$

SASA is the accessible surface area for a respective atom. The surface tension γ and the correction term b are usually set to be constant for all solute molecules; for example, these are 0.00542 kcal/mol-Å^2 and 0.92 kcal/mol, respectively.

To obtain the entropy term $\Delta S_{normal\ modes}$, an approximative method is used. One of the most used methods is the normal modes (NM) approximation. This method is based on calculating the vibration dynamics of the molecule. Each vibration is then decomposed as a harmonic one- or two-dimension oscillator. With this simplification, the entropy can is calculated using the Boltzmann weighted entropy equation (Eq. 35):

$$S = -k_b \int \tilde{\rho}(p,q) \ln\left(\tilde{\rho}(p,q)\right) dqdp \tag{35}$$

where S is entropy, k_b is the Boltzmann factor, $\tilde{\rho}(p,q)$ is the normalized probability of the decomposed system within the momentum and configurational phase space. NM uses a local decomposition for entropy calculations, which may result in high fluctuations, slow convergence and inaccurate results.

Another method used is the quasi-harmonic (QH) approximation. QH uses the concept of essential dynamics of sampled trajectory. The ensemble essential dynamics are calculated by diagonalizing the interatomic correlation matrix, and each of the obtained macro-dynamics is approximated as a harmonic potential.

From this, the procedure is similar to the earlier method. A representation of each method related to the phase space is shown in Figure *24*.

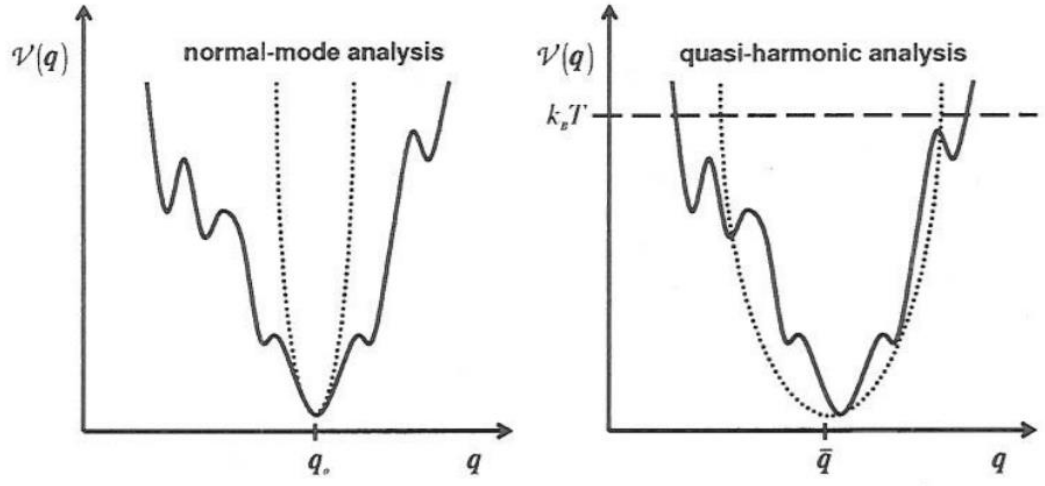

Figure 24: Differences between approximations used by normal-mode (NM) and quasi-harmonic (QH) methods: NM probes local vibrations within the configurational space, approximating each of them as a harmonic oscillator. QH probes the overall extent of the phase space, calculating over states that can be accessed for the temperature of the simulation.

MMPBSA has shown success in several different cases[119,149,150]. However, the approximative nature of it results in a relatively inaccurate method for binding free energies.

Another class of methods for calculation binding free energies is by using enhanced sampling techniques[113,125,151]. Examples of enhanced sampling

techniques are free energy perturbation, metadynamics, adaptative biasing force and umbrella sampling. They rely on sets of simulations to scan a predetermined potential energy surface. Umbrella sampling was used in this work to study ligand-receptor binding energy.

3.4.3 Umbrella sampling

Umbrella sampling belongs to a class of enhanced sampling method known as bias driven techniques[152]. Umbrella sampling is defined by generating a biasing potential which is a function of a single reaction coordinate, resulting in a system Hamiltonian as (Eq. 36):

$$E_{System} = H_{FF} + \omega(\xi) \tag{36}$$

where H_{FF} is the energy of the system and $\omega(\gamma)$ is the biasing potential as a function of the γ, the reaction coordinate related to the predefined collective variable (CV).

With a sampled reaction coordinate, a series of simulations are carried for each sampled window (Figure *25*). Hence, the average force exerted per umbrella window can be calculated via Eq. 37:

$$\langle -\nabla(\mathrm{E(r)}) \rangle = \frac{\int \exp[-\beta E(r)] \, -\nabla(\mathrm{E(r)}) d^N r}{\int \exp[-\beta E(r)] \, d^N r} \tag{37}$$

By integrating this average force throughout the reaction coordinate ξ, the PMF landscape can be calculated by integrating over the mean forces of the pathway (Eq. 38):

$$A(\xi) = \int_{\xi\mathrm{initial}}^{\xi\mathrm{final}} \langle -\nabla(\mathrm{E(r)}) \rangle d\,\xi \tag{38}$$

The energy landscape $A(\xi)$ holds the binding energy information related to this process. Figure *25* shows a representation of the sampled windows and a related final PMF curve.

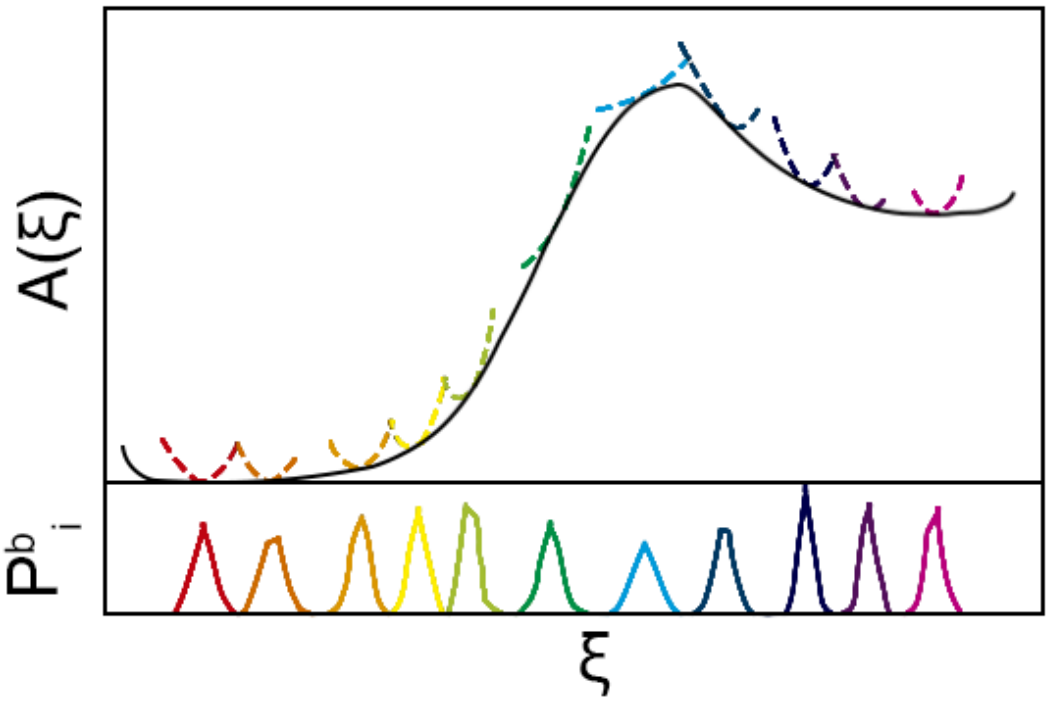

Figure 25: PMF representation and its related windows.

Regardless of the method, they all require accurate force fields and solvation models to yield meaningful results.

3.5 Water models compatible with most commonly used force fields

There are several ways to model the effect of the solvent in biomolecular simulation, but they can be assigned to two different classes, implicit and explicit solvation, as explained in Section 4.2.

Explicit solvation is the usage of molecular water models that explicitly fill the simulation box. These water models can be divided into three main categories: empirical, polarizable and *ab initio* models. These three classes and their respective advantages and disadvantages can be seen in the table

Table 3 Description of the water models reviewed by Ouyang and Bettens[122].

Class	Application	Advantages	Disadvantages	Examples
Empirical	Large systems, such as proteins and materials for studies on atomic scales.	Fast, simple to parametrise, high transferability between systems	Inaccurate water-water interactions and bulk water parameters. No changes for electrostatic parameters.	TIP3P, TIP4P, TIP4P/2005, OPC, TIP5P
Polarizable	Smaller molecules, such as ligands and smaller proteins for biological systems.	Accurate electronic distribution and its dynamical changes in different environments	Requires a higher computational load, which constraining the maximum size.	ASP, SAPT, AMOEBA
Ab-Initio	Water-water interactions and full analysis of electronic distribution of water	High accurate dynamics regarding water properties, useful for studies of quantum properties of water.	Requires an extreme amount of computational power in comparison to the other 2 models, which does not scale for large biological systems.	CC-pol, MB-Pol

Empirical models are based on water structural and physicochemical characteristics, some of the most known families are the SPC (Single Point Charge)[153] and the TIPs (Transferable Intermolecular Potential functions)[154]; which would later be modified to TIPnP series (nP standing for n-Point).

One of the main characteristics of this class is the fact that the partial charges are static. Therefore, there is no momentaneous atomic dipole change. Hence, the partial charge is represented by a fixed value, resulting in an inaccuracy of the electronic effects which emerges from the water interactions.

On the other hand, using this class of waters results in a massive reduction in computational use for each simulation; therefore, it became the most popular class to be used in biomolecular mechanics applications.

Obtaining water parameters requires structural information such as the H-O-H angle, the bond distances and their atomic specifications. The SPC family uses an angle of 109.5° and an H-O length of 1Å. These are the theoretical ideal values for a single water molecule in vacuum. From the SPC basic model, several different models were generated with varying degrees of success, such as the SPC/E[155] and SPC/Fw[156].

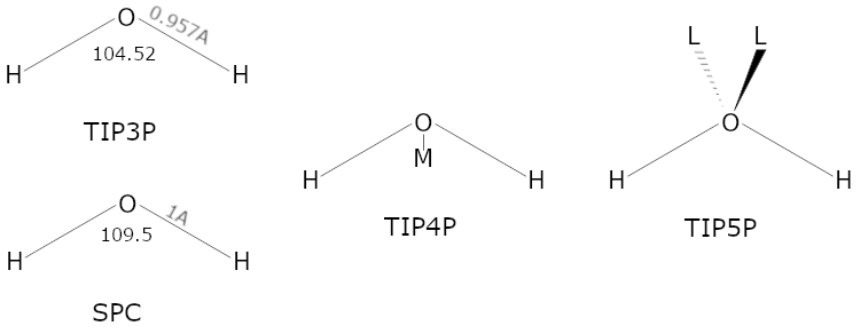

Figure 26: Representations of 4 scaffolds of water models – TIP3P, TIP4P, TIP5P and SPC. The M is the TIP4P dummy atoms, and L represents the 2 extra points in the TIP5P model.

The TIPnP family relies on dummy atoms to represent electronic displacements on the water molecule. Similar to SPC, the TIP3P[121] models rely on three atoms to describe the water structure. However, the structural parameters are different: TIP3P uses the average H-O-H angle given by NMR studies of 104.52° with a shorter H-O distance of 0.957Å in comparison to the SPC scaffold. Because of this agreement, TIP3P gained significant popularity among the computational chemistry community.

Because of this popularity, TIP3P has been modified throughout the years, not only in the internal parameters (i.e. TIP3P/Fw[156]) but also adding dummy atoms to represent electronic configurations better, resulting in the TIP4P[123] and TIP5P[157] models.

The TIP4P water model adds a virtual dummy atom (*i.e.* a massless point charge, or an M-site) within the H-O-H plane (Figure 26). This displaces the negative charge position, placing it closer to the hydrogens, reducing the dipole moment. From this, a series of models were developed, namely, the TIP4P/2005[158], which has modifications of the LJ parameters to improve water interactions, TIP4P-Ew[159], which applies another set of modifications to improve on the bulk water parameters in comparison to TIP4P/2005.

Continuing in the TIPnP family, Mahoney and Jorgensen developed the TIP5P. This water model uses M-sites to mimic the lone pairs surrounding the oxygen

atom in the water (Figure 26). This generated a series of water models such as TIP5P-Ew[160] and TIP5P-2018[161].

It is essential to discuss the creation of the OPC (optimised placed charge) water model[124], since the idea behind the parametrization or the OPC charges influenced the creation of CAIPi3P. This water model was created on modifying the water charge positions to fit the bulk water experimental data better.

Polarisable models rely on semi-empirical approaches to add the induction/polarisation effect. Often, these are incorporated explicitly via dipole-dipole polarizability, using perturbation theory to treat intermolecular forces. Since this class uses a more refined theory level, it requires more computation power, setting an upper threshold on the system size. Notable water models included in this class is the ASP family[162], the SAPT family[163], and the AMOEBA family[164].

Nowadays, one of the new classes of water models is the ones based on *ab initio* data. Using large datasets of energetic values calculated by QM methods, these highly accurate methods require large amounts of computational power, not being used often for biomolecular simulations[122].

As a criterion of quality for the model, bulk water parameters are often calculated (i.e. density, dipole moment, dielectric constant, the heat of vaporisation, first peak radial density distance between oxygen atoms, diffusion coefficient, isobaric heat capacity, thermal expansion coefficient, isothermal compressibility). For the cited empirical water models, OPC has the best agreement with most of the most calculated water parameters[124]. Regardless, TIP3P is still one of the most used water models, even showing significant deviation between experimental bulk properties and calculated ones.

This deviation is a reflection on the static classical structure of the water model, which fails to represent the changes in the electronic clouds caused by the environment. The TIP4P class fails as well to calculate bulk water parameters accurately, nonetheless, shows a significant improvement on the TIP3P and SPC bulk water values.

Within this paradigm, it is known that the water heavily influences the obtained ensembles in MD simulations, since it modulates the overall conformations and the final energetic landscape. This is particularly important for the IDP class of proteins, given their unique energetic landscape.

3.6 Recent advances in molecular modelling of IDPs and IDRs

Given the challenges that intrinsically disordered proteins bring into structural biology and biochemistry, computational methods can be crucial, helping decipher their molecular configuration at the atomistic level of detail and function. However, MD weaknesses get more problematic when applied to proteins with less organised structures.

First, the simulation timescale required to emulate large conformational changes in large IDPs is computationally unfeasible. Therefore, methods that enhance sampling or change the resolution of the system (i.e. coarse grain[113]) are crucial to understanding the IDPs dynamics within a longer timescale[165].

Regardless of the relative success of IDP molecular studies using MD, using different, often simplified methods[166], can help to overcome the sampling problem, such as Gō models[167], metadynamics[168] and Monte Carlo ensemble generation[169]. Within this scope, the accuracy and reliability of the generated ensembles will be as good as the quality of the mathematical modelling tools used to define the quality of molecular states.

The force field accuracy is the second challenge faced by molecular simulations. Their resulting energy landscape is usually inaccurate for simulations of protein domains lacking well defined secondary structures. Developments have been made on how the system energy is calculated to overcome the biasing problem: energetic functions to evaluate generated ensembles and new parameters to improve the accuracy of said established force fields.

Modelling novel energetic functions are very challenging. Typically, these new functions rely on QM approximations or empirical modifications on interactions functions. Examples of this are the soft-core interaction potential[170], and well established but not often used functions like flat-bottom interaction[113] and Morse potential[113]. When applied to IDPs, one of the most successful approaches is the

grid-based CMAP correction[171]. When applied to CHARMM and AMBER force fields, it yielded exciting results. CMAP is based on applying correction terms to backbone φ-ψ torsion calculations. Using this approach, Ye and colleagues generated the AMBERFF99idp force-field, which applies CMAP corrections to specific residues[171]. Later, they extended the concept to all residues, creating the AMBERFF14idp[172]. For the case studies used on their work, the results showed reasonable improvements, especially for proteins bearing intrinsically disordered regions.

On the other hand, improvement of parameters is often based on the most used functions (discussed in Section 4.2). These modifications affect sets of atomic and molecular descriptors to improve a specific characteristic of the system.

The goal of the scientific community regarding molecular mechanical force fields is to generate a generalist force field which could accurately sample all kinds of proteins; folded and disordered alike. Generalist force fields developed to date fail to grasp the dynamics of overly flexible regions. Approaches to solving this problem came for different types of force fields. For AMBER, the AMBER03w[173] and AMBER03ws[174] force fields are often used to simulate stretched conformations of flexible proteins. Mainly, their modifications rely on changing LJ parameters, modifying how water-protein interaction happens. Best and colleagues achieved a consistent improvement on simulations for IDPs by multiplying the well depth parameter ε between the water oxygen the protein atoms by a constant factor[174]. The results are promising for full IDPs, but for structured proteins bearing intrinsically disordered regions (IDRs), the secondary structure within the structured domains often falls apart, showing inaccuracies for highly structured proteins. To overcome this issue, Hybrid-FF force field framework was created in this work, which assumes different structural parameters between regions containing secondary structures, and regions purely comprised by unstructured loops. How it was developed and the benefits it generates will be discussed in Chapter 5.

CHARMM family has its series of force fields tailored for IDPs. One of the cases with the most successful history is the CHARMM36m[175]. Both AMBER03ws and CHARMM36m and force fields were able to generate correct results for IDPs in comparison to experimental data.

Alongside the accuracy of the force-field parameters, how the solute-solvent interactions are parametrized profoundly affect the final ensemble. Therefore, several specific water models were created to be used alongside these force fields.

The TIP4P water scaffold is often the basis for generating new models for biomolecular MD simulations. Being highly compatible with the scale parameters from AMBER03ws, the TIP4P/2005 is frequently used for IDPs simulations. The usage of the AMBER03ws+TIP4P/2005 scales the interaction within the order of 1.1 times the usual water-protein dispersion interaction, resulting in a better agreement for the chain dimension of the ACTR[174] and used for benchmarking in several studies[176].

This model was the basis for the TIP4P-D water model, which restrains the water O ε, resulting in better electrostatic interaction between protein-water[177]. Nonetheless, Henriques *et. al.* showed that TIP4P-D did not yield significant improvement for the simulation of histatin5 in comparison with the TIP4P/2005[186]. A drawback of the TIP4P model is the addition of a dummy atom, which increases the computational system requirement. Based on the modifications of the force fields and the TIP4P models and their specific modifications, CAIPi3P model was developed in this study, which does not require a dummy atom and has increased electrostatic interaction with the protein.

In the next chapters, the CAIPi3P model and Hybrid_FF force field will be discussed, and their result compared to a series of water models and protein force fields.

Chapter 4 - Development of the CAIPi3P water model

As outlined in Chapter 2, intrinsically disordered proteins (IDPs) exert important roles in cellular signalling, growth and molecular recognition events. Due to their high plasticity and a lack of fixed tertiary structure, IDPs are very challenging for experimental structural studies. Hence, all-atom molecular dynamics (MD) simulations are widely employed to provide detailed atomic insight in IDPs dynamics governing its functional mechanisms. However, the current generalist force fields and solvent models are unable to generate satisfactory ensembles for IDPs when compared to existing experimental data.

This chapter describes the development of the **C**harge-**A**ugmented **3** **P**oint water model for **I**ntrinsically disordered **P**roteins (CAIPi3P). CAIPi3P has been generated by performing a systematic scan of partial atomic charges assigned to the popular molecular scaffold of the three-point TIP3P water model. The results showed that explicit solvent MD simulations employing CAIPi3P solvation improved the SAXS scattering profiles considerably for three different IDPs. Not surprisingly, this improvement was further enhanced by using CAIPi3P water in combination with the protein force field parametrised for IDPs. The work presented in this chapter also demonstrated the applicability of CAIPi3P to proteins containing globular (structured) as well as intrinsically disordered regions/domains, which will be discussed further in Chapter 6. The results highlight the crucial importance of solvent effects for generating molecular ensembles of IDPs to reproduce the experimental data available. Hence, it is concluded that the newly developed CAIPi3P solvation model is a valuable tool for molecular simulations of intrinsically disordered proteins and assessing their molecular dynamics.

4.1 MD simulations, solvation effects and IDPs

Atomistic molecular dynamics (MD) simulations can reliably assess dynamic properties in equilibrium structures of molecular systems of interest, given an ergodic sampling and an accurate force field. The force field parameters are calibrated to reproduce properties measured by experiments or simulations. Considering the immense complexity of macromolecular systems, and sensitivity

of weak (hydrogen-bonding and dispersion) non-covalent interactions in the liquid phase, contributing to intra-solute, solute-solvent and solvent-solvent interactions, even minor inaccuracies in models and their parameters can adversely impact the results of atomistic molecular simulations, especially of challenging systems such as intrinsically disordered proteins (IDPs). IDPs are elusive to experimental studies; thus, atomistic simulations are a crucial tool to provide detailed insight into their complex structure, dynamics, and function. Unfortunately, computational studies of IDPs are often found to disagree with experimental data. Free energy landscape of IDPs is inverted compared to the structured proteins[49], which makes computational studies focusing on IDPs very challenging. Discrepancies between theory and experiments are commonly attributed to either force field biases[178,179] or insufficient sampling. This motivated the development of molecular force fields designed to handle IDPs[180–182] and to apply enhanced sampling techniques[183–187] or restraints derived from experimental data (e.g. solution NMR) in simulations of IDPs[186,188,189]. The outcomes of those efforts were successful to various extents. However, IDPs simulations still require parameter improvements[186,190–193].

IDPs have disordered structures in aqueous solution, and while either dehydrated or interacting with lipid membranes, they exhibit increased amounts of ordered secondary structures[141]. This clearly shows that IDPs are highly sensitive to solvation effects[194,195] and suggests that focusing on improvement of the water models used in the simulations may offer a more accurate yet computationally feasible framework for reliable simulations of this class of proteins.

The complexity of the water properties, combined with multiple possible levels of approximation, has led to the proposal of dozens of water models. Simplified classical water models, such as widely popular three-point SPC[153] and TIP3P[121] models, are currently indispensable components of atomistic MD toolkits. Nevertheless, despite several decades of extensive research, these models are still far from perfect. To start, none of them accurately reproduces the critical properties of bulk water[196].

In simulations of IDPs, the best-performing water models have a charge distribution with a large dipole moment, a significant quadrupole moment, and negative charge out of the molecular plane, to give symmetrically ordered

tetrahedral hydration[197]. It has been observed that the dipole moment calculated for TIP3P model is too low, resembling of a dipole moment of an isolated water molecule in vacuum (2.3 D), rather than of a dipole in the liquid bulk state (~3 D). The exact value of the liquid water dipole is still debated; however, in this study relied on the results of the most recent first-principles simulations of liquid state water. To improve the properties of the TIP3P water model, it seemed crucial to adjust the dipole moment by augmenting partial atomic charges of the water molecule (Figure 27). The performance of such an improved model denoted as **C**harge-**A**ugmented **3** **P**oint Water model for **I**ntrinsically disordered **P**roteins (CAIPi3P).

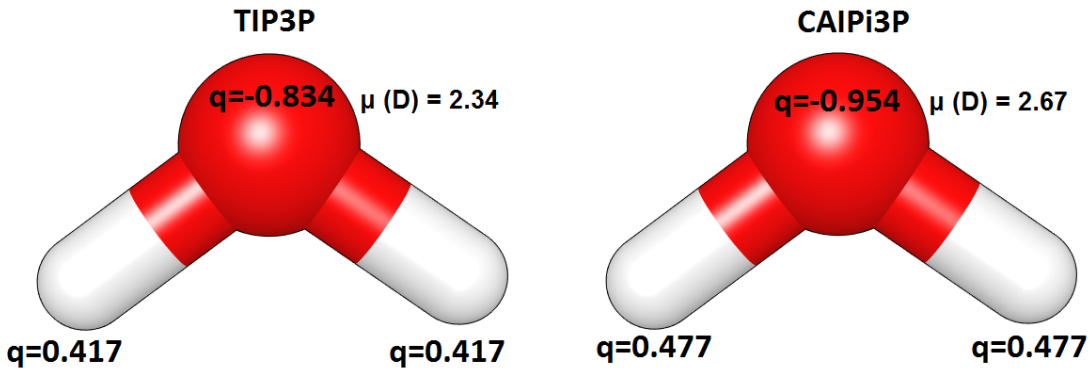

Figure 27: Partial charges for TIP3P and CAIPi3P water models.

CAIPi3P model has been subsequently tested on model IDPs: histatin 5, R/S-peptide, and partially disordered *At2g23090* protein from *A. thaliana*. It has been observed that the dipole moment adjustment dramatically improved the performance of the model, in terms of reproducibility of experimental data for IDPs, without negatively affecting the performance and data reproducibility for the folded regions/domains of partially disordered systems or the 'structured' globular proteins for both lysozyme and ubiquitin, which will be shown in the next section.

4.2 Methodology

To assess the role of the solvation effects in reproducing the experimental parameters of IDPs, and to evaluate the applicability of CAIPi3P model to studies of "mixed" ordered-disordered systems, three models were selected (histatin 5[176] and R/S peptide[175]) and At2g23090, which is partially disordered. To determine

the performance of the model, comparisons were made between CAIPi3P and established water models.

Fully extended conformations of histatin 5 (amino acid sequence: DSHAKRHHGYKRKFHEKHHSHRGY) and R/S-peptide (amino acid sequence: GAMGPSYGRSRSRSRSRSRSRS) were built using UCSF Chimera[198] package since their experimental atomistic structures were not available. The conformational ensemble of *A. thaliana At2g23090 (*PDB code: 1WVK), obtained by solution NMR, was used to calculate the small-angle X-ray scattering (SAXS) distribution and radius of gyration. The lowest-energy conformer was selected as a starting point for all-atom molecular dynamics (MD) simulations.

For all systems investigated, missing hydrogen atoms were added, and several combinations of protein and water parametrisations were chosen, as summarised in Table 4. All simulations were performed using the Gromacs 5.3 suite[199]. The combinations of the force field and water models used are summarised in Table 4. For each combination, a 1 nm cubic box was centred on the structure.

The system was solvated with the necessary number of the water molecules to fill the protein simulation box. Next, sodium and chloride ions were added to the system at a concentration of 0.1 M to neutralise the simulation unit and to mimic the “physiological” salt concentration. The bonds were constrained using the LINCS algorithm[200], setting a 2 fs time step. The electrostatic interactions were calculated using particle-mesh Ewald method[116], with a non-bonded cut-off set at 1 nm. All structures were energy minimised using the steepest descent algorithm for 20,000 steps. The minimisation was stopped when the maximum force fell below 1000 kJ/mol/nm using the Verlet cutoff scheme. This was followed by an equilibration run (NVT ensemble) of 20 ps with a time step of 2 fs and position restraints applied to the backbone, where the system was heated from 0 to 300 K; and another equilibration (NPT ensemble) at the constant temperature (300 K, 20 ps, 2fs step) with backbone position restraints applied, and the constant pressure (1 bar). The temperature was set constant at 300 K by using an alternative Berendsen[128] thermostat (τ = 0.1 ps). The pressure was kept constant at 1 bar by using a Parrinelo-Rahman barostat with isotropic coupling (τ = 2.0 ps) to a pressure bath[201]. Finally, three production replica runs (NPT ensemble) of

100 ns were run for each system, using every force field – solvation model combination.

Ubiquitin (PDB code: 1UBQ) and lysozyme (PDB code: 253L) were selected for comparative runs to assess the effect of CAIPi3P water model on globular proteins with no IDRs. The simulation methodology was the same as the one described for the IDPs, with the exception that only AMBER99SB-ILDN force field was used in combination with either the TIP3P or CAIPi3P solvation model.

Table 4: Systems simulated and their respective force field/solvent combinations

Protein	PDB Code	Force Field	Water model
Histatin5	-	AMBER99SB-ILDN[202]	TIP3P; CAIPi3P
		AMBER03ws[173]	TIP3P; CAIPi3P; TIP4P/2005[158]
R/S Peptide	-	AMBER99SB-ILDN[202]	TIP3P; CAIPi3P
		AMBER03ws[173]	TIP3P; CAIPi3P; TIP4P/2005
At2g23090	1WVK	AMBER99SB-ILDN[202]	TIP3P; CAIPi3P
		AMBER03ws[173]	TIP3P; CAIPi3P; TIP4P/2005

CRYSOL[203,204] software was used to calculate the SAXS scattering patterns, along with the GNOM[203] software to calculate radial density distributions. The root square difference (RSD) between the experimental density curves and the curves extracted from simulations were made using an in-house script. The *gmx gyrate* module from the Gromacs suite was used to calculate the radii of gyration from the obtained trajectories. RMSF and RMSD values were calculated using the Gromacs suite (*gmx rms* and *gmx rmsf*, respectively).

4.3 Results

4.3.1 Parametrisation of CAIPI3P water model

Unlike globular proteins, intrinsically disordered proteins (IDPs) do not have a proper hydrophobic core. As such, long-range electrostatic interactions play an important role in defining of IDR behaviour[194,195]. Therefore, to accurately predict the dynamics of IDRs from the atomistic molecular simulations, the protein-water interactions need to be re-calibrated.

A systematic scan for dipole moment magnitude was made using the molecular scaffold of the popular TIP3P framework[121] as a template. Different dipole moments were tested using histatin 5 as a reference system since it's IDP model that has a uniform charge distribution through its structure. The model that showed the best agreement with experimental SAXS radial distribution was selected for CAIPi3P, and its partial charge values were kept constant throughout the other systems. The atomic charges and the value for the dipole moment are shown in Table 5. CAIPi3P model has been developed to scale solute-solvent interactions through electrostatic interactions. This follows the idea of scaling solute-solvent interactions, which was the basis for the creation of AMBER03ws[174], which has modifications on the LJ parameters for the protein-water interactions. This force field is tailored for simulations of IDPs, has been designed to be fully compatible with the TIP4P/2005 water model.

Table 5: Partial atomic charges and resulting dipole moments for CAIPi3P, TIP3P, and TIP4P/2005 water models

	O charge (e)	*Dummy atom charge (e)*	*H charge (e)*	*Dipole moment (D)*
CAIPi3P	-0.954	-	0.477	2.69
TIP3P	-0.834[205]	-	0.417 [205]	2.36 [205]
TIP4P/2005	-	-1.128 [158]	0.5564[158]	2.30 [158]
Experimental	-		-	2.5-3[205]

4.3.2 MD simulations of a full-length IDP: histatin 5

Histatin 5 belongs to the family well characterised antimicrobial peptides secreted in human salivary (submandibular) glands[176]. It is a highly water-soluble IDP that has been used as a model in computational studies[176,179]. Although there is no experimental structure of histatin 5 available to date, Henriques *et al.*[176] published its SAXS data. In their study, the best agreement between simulations and experimental results has been achieved using AMBER03ws force field with the TIP4P/2005 water model.

We found that although usage of AMBER03ws improved the sampling compared to AMBER99SB-ILDN (Figure 28.A), combining it with CAIPi3P solvation model rather than TIP3P has achieved an improvement regarding the difference between calculated and experimental SAXS curves, resulting in a curve significantly closer to the experimental. The combination of AMBER03ws (protein) and CAIPi3P (solvent) outperformed AMBER03ws+TIP4P/2005 (RSD = 0.0110 for AMBER03ws+TIP4P/2005 in comparison to 0,007 for the AMBER03ws+CAIPi3P), which has been considered the state-of-the-art[176]. Although the AMBER03ws+CAIPi3P combination has shown the closest agreement with the experimental data. The improvement resulting from the application of CAIPi3P was apparent regardless of the protein force field used (Figure 28.A) since both simulations with CAIPi3P (AMBER99SB-ILDN+CAIPi3P and AMBER03ws+CAIPi3P) sample curves closer to the experimental curve. It is very encouraging in the context of the transferability of the CAIPi3P model and its applicability to studies of intrinsically disordered macromolecules.

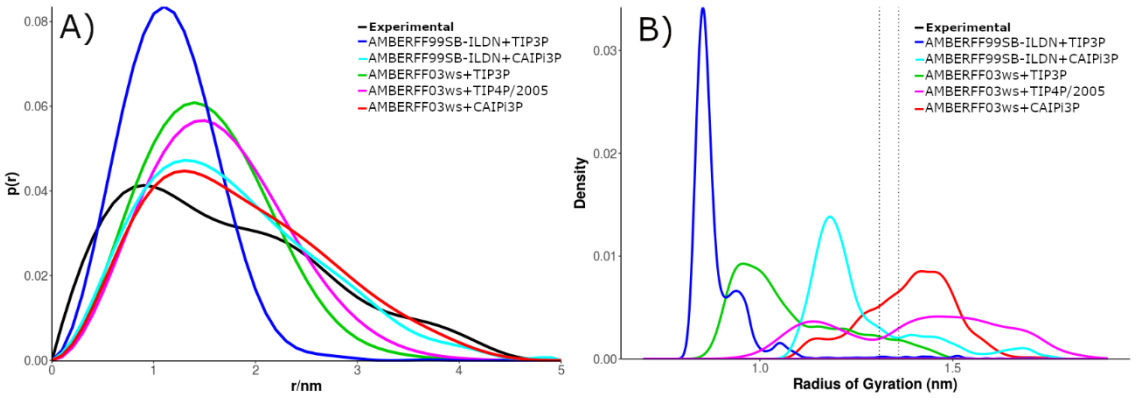

Figure 28: Small-angle X-ray scattering (SAXS) radial distributions and calculated radii of gyration of histatin 5: A) SAXS distributions for five chosen combinations of protein and water parametrisations; B) Distributions of the sampled radius of gyration for the combinations shown in panel A; the radius experimental interval is highlighted using black dashed lines.

As showed in Figure 28 .B, both AMBER03ws and AMBER99SB-ILDN force fields attained reasonable sampling of the experimental radii of gyration, since the radius of gyration distribution for both AMBER03ws+TIP4P/2005 and AMBER03ws+CAiP3P can sample the within the experimental radius of gyration values more than the other combinations (Figure 29.B). The AMBER03ws combined with CAIPi3P had its distribution peak around 1.4 nm, sampling more expanded conformations than any of the combined sets.

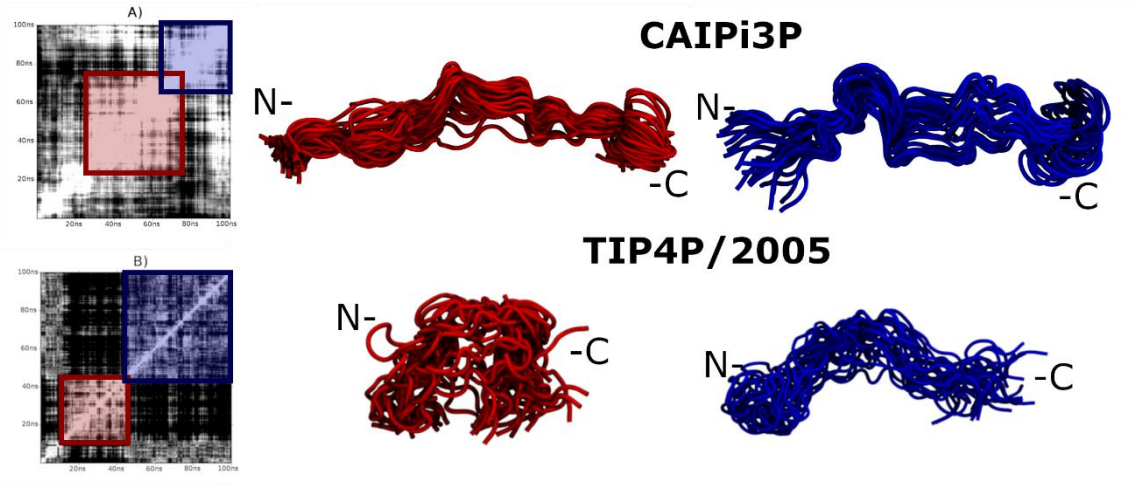

Figure 29: RMSD matrices and their respective clusters obtained by AMBER03ws. A) CAIPi3P matrix B) TIP4P/2005 matrix. The structures and their respective areas in the RMSD matrix are coloured similarly. The matrix similarity threshold were 3 angstroms, hence, lower values have a white colour and higher values have a black colour.

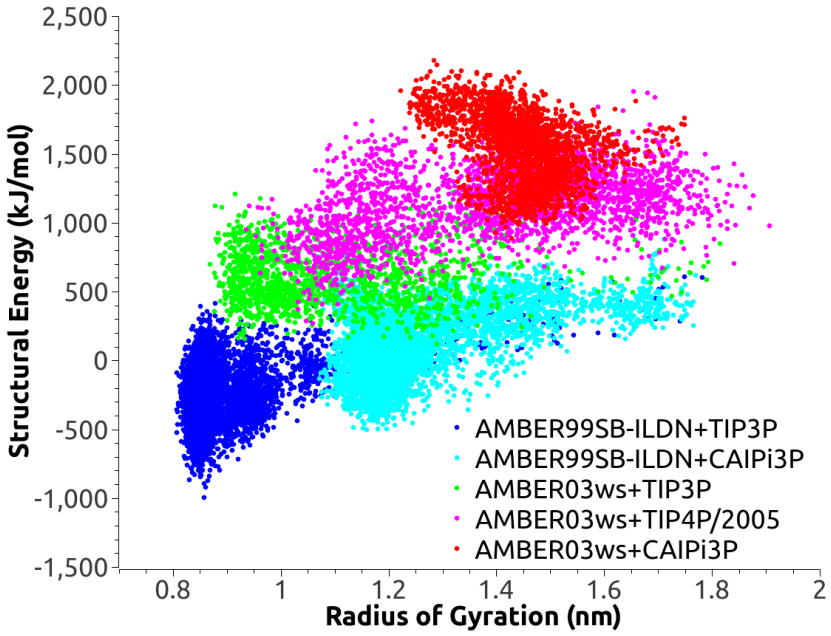

Figure 30: Histatin 5 internal potential structural energy in the function of its radius of gyration.

The solute-solvent and solvent-solvent long-range electrostatic interactions play a significant role in defining of the conformational landscape of IDPs. The solvation model is, therefore, crucial for the sufficient sampling of the IDPs. Figure 29 shows that two clusters obtained by the simulations using CAIPi3P, calculated from the RMSD matrix, are a very similar one to another, as can be seen in their respective intersection area in the RMSD matrix. TIP4P/2005, on the other hand, sampled two sparse conformations, having the system collapsed on itself for nearly half of the simulation time, represented by the TIP4P/2005 blue protein cluster. The most compact (self-collapsed) conformation affected the radial distribution, which resulted in the ensemble with the radial distribution resembling that of a globular protein, which directly affects the calculated sampled, resulting in the gaussian-like distribution for the SAXS density (Figure 29).

The description of solute-solvent interactions increased upon using the CAIPi3P model. The internal potential energy for the protein increased, showing that protein intramolecular interactions should be disrupted, and the solute-solvent interactions should be increased, as can be seen in the energy difference between AMBER03ws and AMBER99SB-ILDN clusters in Figure 31. Although AMBER03ws can increase the radius of gyration by increasing the solute-solvent

contribution, CAIPi3P samples the correct configurations by increasing the Coulombic solute-solvent and solvent-solvent interactions, increasing the structural potential energy, as can be seen in the red cluster in Figure 30, which has both a high value of structural energy and radius of gyration.

4.3.3 The CAIPi3P effect on the sampling of the charged repeats of R/S-peptide

Arginine – serine repeats (R/S repeats) play an essential role in cellular signalling since the phosphorylation of the serine residues is crucial for the regulation of many enzymes and receptors. Because of the accumulation of highly polar arginine and serine residues, intrinsically disordered R/S-peptide is highly polar itself. As such, it presents a challenging IDP to be correctly modelled. Several studies on its dynamics have been performed, using solution NMR and SAXS[175,206].

The calculated SAXS parameters and radii of gyration for the R/S peptide and their comparison to the experimental data available are shown in Figure *31*. The choice of the protein force field played a critical role in reproducing the experimental data, as shown in Figure *31*.A, where the curves for AMBER03ws are shown to be significantly closer to the experimental distribution. Application of both TIP4P and CAIPi3P water models reproduced the experimental SAXS radial distribution, but the usage of AMBER03ws played a more significant role on sampling states similar to the experimental conformations, as shown in the Figure 31.A. Nonetheless, simulations performed using the CAIPi3P model achieved better sampling than TIP4P (achieving a 0.007 RSD in comparison to the 0.008 difference given by TIP4P/2005). Also, CAIPi3P can sample within the experimental range for the radius of gyration, as shown by the calculated density curves (Figure 31 .B), on which the black lines represent possible experimental values and AMBER03ws+CAIPi3P (red curve) sampled several states within this range.

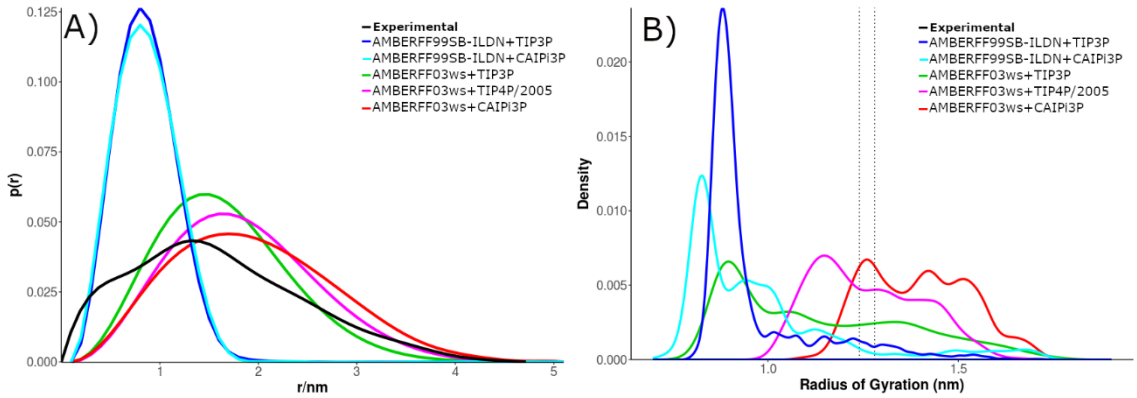

Figure 31: Small-angle X-ray scattering (SAXS) radial distributions and calculated radii of gyration of R/S-peptide: A) SAXS distributions for five chosen combinations of protein and water parametrisations; B) Distributions of sampled radius of gyration for the combinations shown in panel A; the radius experimental interval is highlighted using dashed lines.

Regardless of the solvation model used, the R/S peptide simulated with the AMBER99SB-ILDN force field collapsed on itself after 30ns of simulation, resulting in a very different distribution when compared to the experimental data. (difference between calculated and experimental SAXS curves: CAIPi3P RSD=0.033 and TIP3P RSD=0.034, respectively; Figure *31* .A). This is likely to arise from the differences between force fields, mainly in values of torsion parametrisation for bulky residues.

Employing AMBER03ws force field improved the agreement with the experimental data, regardless of the water model (TIP4P/2005 RSD = 0.008, and CAIPi3P RSD = 0.007). However, the production trajectories obtained with CAIPi3P water show a higher distribution peak in the denoted experimental range of the radius of gyration (1.3 nm; Figure *31* .B)., resulting in a predicted radius of gyration of 1.3 nm.

For the R/S peptide sampling assessment, two clusters of each combination of protein force field/ water molecule were selected for visual inspection. CAIPi3P clusters remained in an open, extended conformation for approximately 85% of the simulation run, as shown in the RMSD matrix in Figure 32, on the white cluster shown on the matrix. Both solvent models enabled interactions between the N-terminal region and the 16 residues R/S repeat region. The modified CAIPi3P model shielded the interactions and avoided a collapsed structure for the first 60 ns of the simulation. In the TIP4P/2005 model, a partial collapse occurred early

in the simulation, and it is highlighted by the TIP4P/2005 red ensemble in Figure 32 (lower panel).

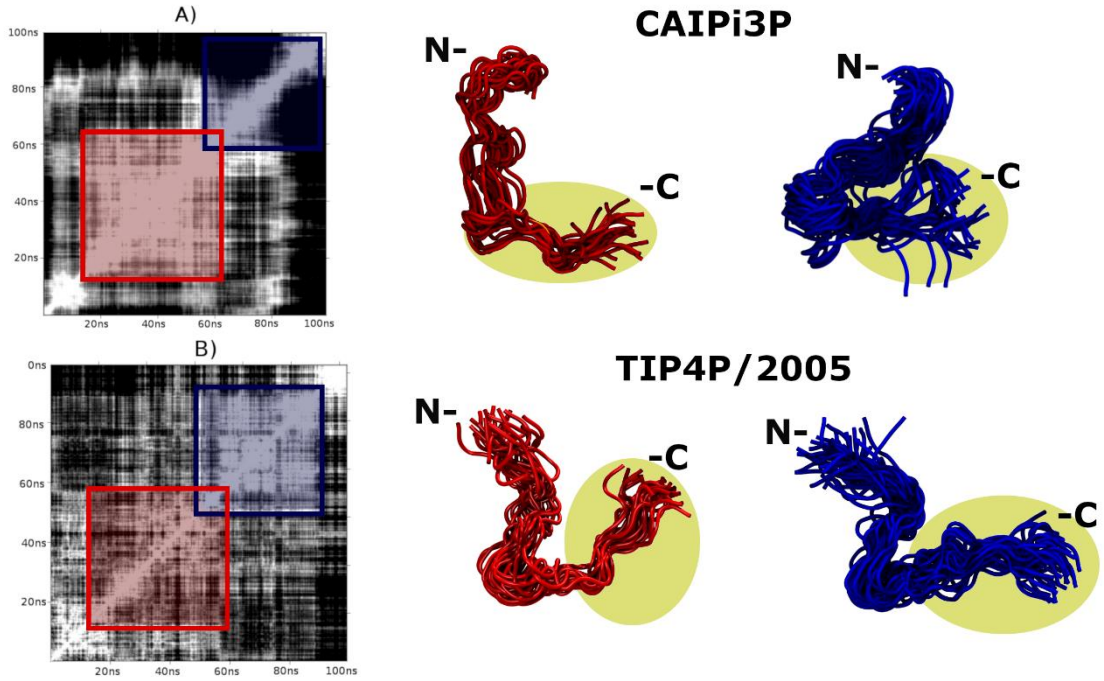

Figure 32: RMSD matrices for the R/S peptide and their respective clusters for AMBER03ws. A) CAIPi3P matrix B) TIP4P/2005 matrix. R/S repeat is highlighted yellow.

Since the R/S repeat region is very polar (Figure 32; highlighted regions in yellow), it might be expected for this region to interact with favourably with water. The glycine residue, which is adjacent to the R/S repeat, acts as a “hinge”, partially collapsing the ensemble (blue clusters in Figure 32) in simulations using both solvent models. Considering this structural peculiarity, R/S peptide presents itself a challenge for modelling and suffers more from the force-field selection from the solvation model, since the force-fields are known to be directly affected by the accuracy of the calculated charges[113]. The results show that there are improvements still to be made on the AMBER force field and CAIPi3P parameters, yet the sampling achieved by the application of CAIPi3P model outperformed that of TIP4P/2005, as it was shown with the sampled SAXS curves.

4.3.4 The effects of CAIPi3P on partially disordered structures

The solution NMR structure of the partially disordered protein *At2g23090* from *Arabidopsis thaliana* has been deposited in the RCSB PDB Data Bank (PDB code: 1VWK[207]). It was used to assess the accuracy of the CAIPi3P water model for very flexible and partially disordered proteins since it has a C-terminal globular region and a long loop formed by 46 residues. While *At2g23090* presents itself as a challenging and exciting benchmarking test, the NMR ensemble was used to study the possible dynamics. When analysed, the AMBER99SB-ILDN protein force field combined with CAIPi3P water model significantly outperformed all the other combinations of protein force fields and solvent models (Figure 33 .A), ith an RSD = 0.003 for CAIPi3P+AMBER99SB-ILDN.

As expected, trajectories obtained with AMBER03ws (Figure 33) showed a distribution of sampled radius of gyration with a higher average value.

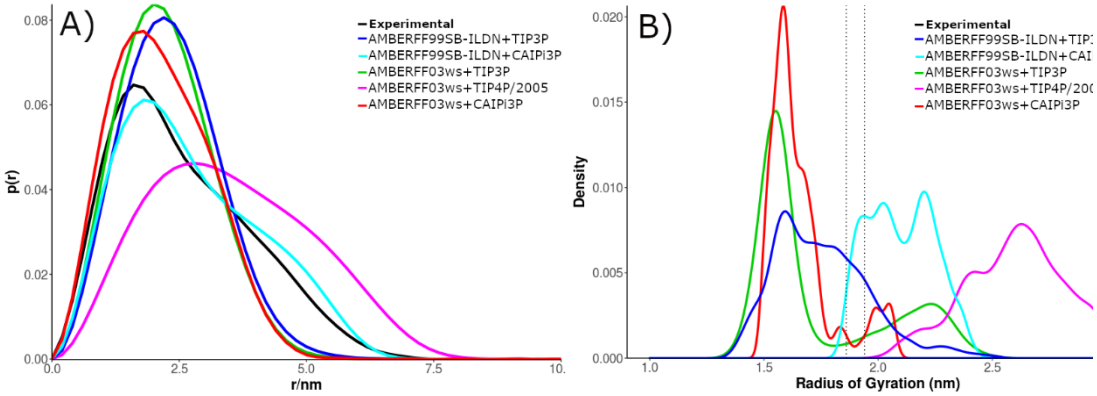

Figure 33: Small-angle X-ray scattering (SAXS) radial distributions and calculated radii of gyration of *At2g23090* protein: A) SAXS distributions for five chosen combinations of protein and water parametrisations; B) Distributions of the sampled radius of gyration for the combinations shown in panel A; the radius experimental interval is highlighted using dashed lines.

The AMBER03ws+CAIPi3P combination (Figure 33 .B; red) showed a bimodal radial distribution, with the highest peak around 1.7 nm, in a collapsed configuration. The protein collapsed on itself, resulting in a Gaussian radial distribution with experimental RSD = 0.014, and a dispersed bimodal radial density (Figure 33.B) with the highest peak in 1.6 nm, which is lower than the range calculated from the solution NMR ensemble.

In contrast, the application of the TIP4P/2005 model with the AMBER03ws force field resulted in the unfolding of the small globular C-terminal domain (Figure *34*), with a radius of gyration centred around 2.7 nm, which is much higher than the

experimental range. This demonstrates the limitations of the applicability of AMBER03ws force field in the simulations of multi-domain proteins containing globular domains connected by intrinsically disordered regions (IDRs). The C-terminal domain remained folded in simulations using AMBER99SBN-ILDN (Figure 34). The reverse happened for AMBER03ws+TIP4P/2005: the radii of gyration were outside of the experimental range, resulting in the average conformations that were too stretched in comparison to the experimental data.

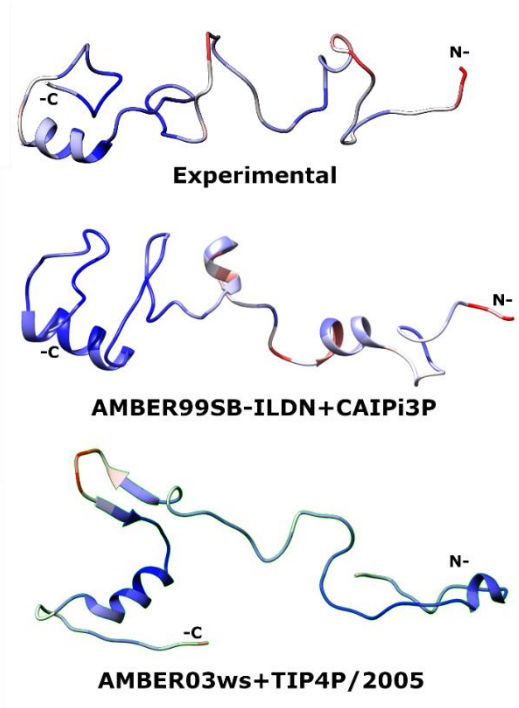

Figure 34: Average structures for the At2g23090: Highly flexible regions (high per-residue RMSF) are coloured red, while more rigid regions with lower per-residue RMSF are coloured blue.

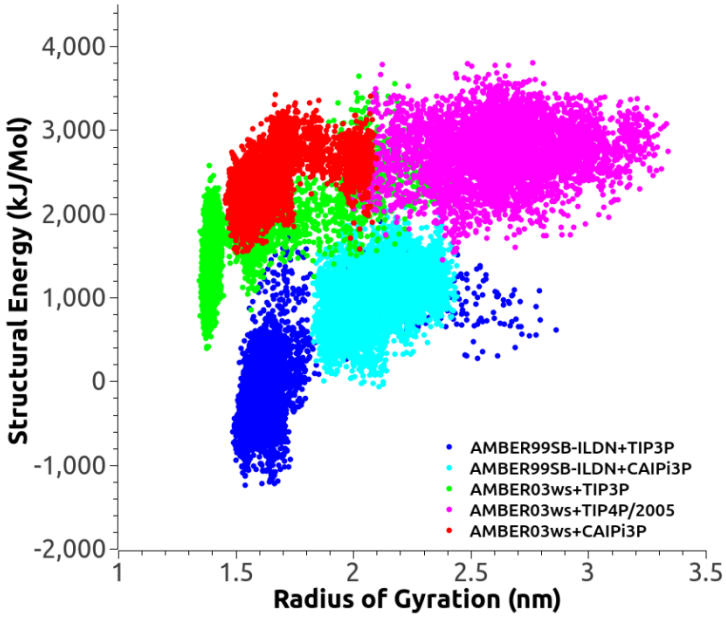

Figure 35: *At2g23090* structural energy versus radius of gyration. In contrast to Histatin5 energetics, the *At2g23090* requires partial internal interactions. T

Unfolding of the globular C-terminal domain was coupled with an increase in the structural potential energy in AMBER03ws simulations. Figure 35 shows that the structural energy obtained for the ensemble in the simulations using AMBER99SB-ILDN+CAIPi3P was more favourable. The increase of structural energy resulting from AMBER03ws (red and magenta clusters) unfolds the structured region, stretching the average configuration. When using AMBER99SB-ILDN with CAIPi3P, the energy can be kept within the collapsed AMBER99SB-ILDN+TIP3P (blue cluster in Figure 35). The solvent model stabilizes the unstructured sequences, and a structured biased force-field holds the intra-protein interactions reasonably. This favourable energy which arises from the intramolecular interactions retained within the folded C-terminal domain, resulting in the cyan cluster in Figure 35, which is an equilibrated cluster between both magenta and blue cluster.

4.3.5 Applicability of CAIPi3P solvation model to globular proteins

To benchmark the CAIPi3P model, two model globular proteins were simulated; lysozyme and ubiquitin, using the same protocol as described for IDPs and partially unfolded *At2g23090*. For lysozyme, the residual root-mean-square fluctuations (RMSF) obtained for both water models when applying established

AMBER99SB-ILDN force field are shown in Figure *36*. Simulations with TIP3P resulted in higher RMSF values for most of the residues. In simulations with CAIPi3P, stronger electrostatic solvent-solute interactions increased the stability of most of the residues, decreasing the overall RMSF.

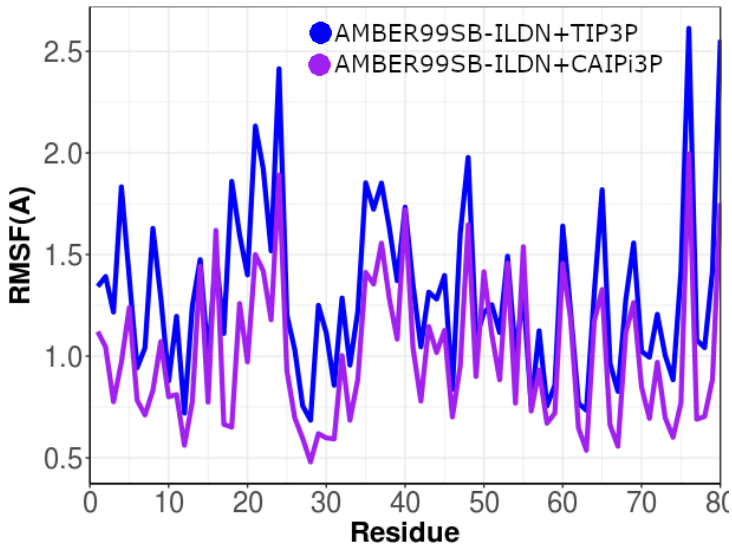

Figure 36: Lysozyme RMSF per residue. In Blue, the simulation of the lysozyme with the usual combination of force-field/water. In purple, the simulation of using CAIPi3P.

For ubiquitin, the root-mean-square fluctuations (RMSF) per residue obtained for both water models when applying established AMBER99SB-ILDN force field are shown in Figure *37*. Simulations with both water models achieved very similar results, with only one region (loop 40-50) with markedly increased RMSF when the TIP3P model was applied in comparison to CAIPi3P. Again, this difference can be attributed to stronger electrostatic solvent-solute interactions in CAIPi3P, which increased the stability of the protein region, decreasing its overall per-residue RMSF. Yet the effect was much less pronounced for ubiquitin than for lysozyme.

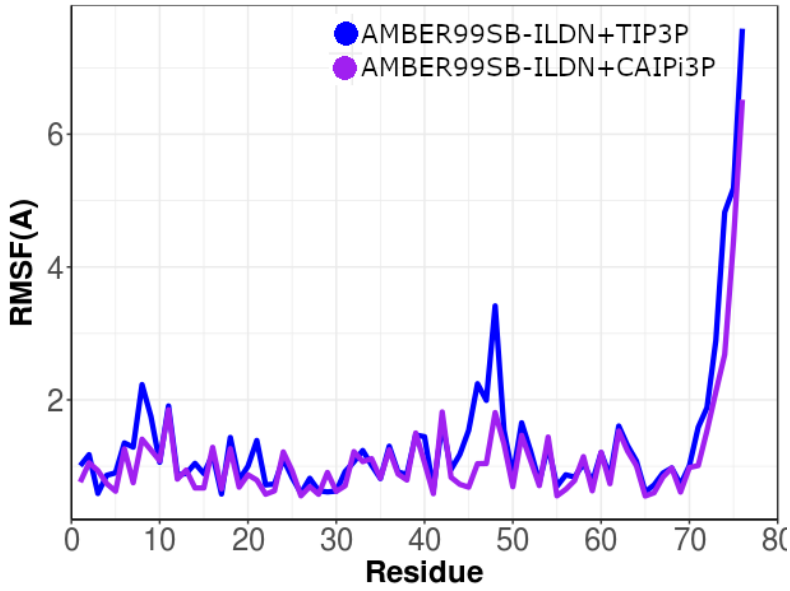

Figure 37: Ubiquitin RMSF per residue. In Blue, the simulation of the lysozyme with the usual combination of force-field/water. In purple, the simulation of using CAIPi3P.

4.4 Discussion

This chapter focused on the development of the new solvation model, denoted CAIPi3P. Compared to the established and popular TIP3P water model, CAIPi3P, which is based on the same framework, considerably improved the sampling of intrinsically disordered model peptides. All-atom MD simulations using CAIPi3P improved the SAXS scattering profile for two model IDPs: R/S peptide and histatin 5, and partially disordered *At2g23090* from *A. thaliana* with the central IDR. The improvement was evident for all force fields used for the protein, although, the selection of the most appropriate force field plays a vital role in the sampling improvement.

For the R/S peptide, the improvement was evident in simulations with AMBER03ws force field. Application of the CAIPi3P model resulted in a better agreement for the radius of gyration since the framework prevented the artificial collapse of the polypeptide chain, which is a common pitfall of atomistic simulations of IDPs. CAIPI3P, due to modified electrostatics, maintained the generated conformations stretched, which resulted in better agreement with the experimental data.

It is essential to focus on the differences in primary sequence between these two model IDPs. Histatin 5 has several polar residues dispersed throughout the length

of the peptide, resulting in an overall uniform polar distribution. This homogeneous distribution helps the polypeptide chain to maintain favourable interactions with the solvent, resulting in the overall expanded structure.

The R/S peptide is polar and charged, with the charged residues located within the 8 C-terminal arginine – serine (R/S) repeats, as shown in Figure 32 (highlighted regions). The obtained ensemble was affected by the C-terminal charge distribution, which facilitated the collapse of the polypeptide chain. Such a collapse was markedly reduced when the AMBER03ws force field was applied. The sampling was further improved when CAIPi3P water was used since it favoured the solute-solvent electrostatic interactions due to increased dipole moment of the water molecule. Solvent-solute interactions thus competed with excessive intramolecular solute-solute interactions, which lead to the collapse. In this part of the work, mainly AMBER forcefields were tested and assessed their accuracy. Rauscher and coworkers[175] used R/S peptide to assess the accuracy of the CHARMM36m, obtaining accurate results for SAXS curve. Nonetheless, Rauscher and coworkers did not test AMBER force-fields, therefore, a comparison between CHARMM36m and AMBER03ws with CAIPi3P should be made in the future.

For *At2g23090*, MD simulations showed a good agreement with experimental data when using the CAIPi3P water model in combination with AMBER99SB-ILDN protein force field. Differences between AMBER99SB-ILDN and AMBER03ws laid within the side chain charge distribution and the values of backbone torsional angles for a specific set of residues[174,202]. These changes affected the solvent-solute interactions. Consequently, in the *At2g23090* simulations, the compact globular C-terminal domain unfolded, increasing the interactions with the solvent molecules and the internal structural energy. In contrast, AMBER99SB-ILDN force field held the globular domains folded. This resulted in a similar radial SAXS distribution between the resulting ensemble and the experimental data when used CAIPi3P model. CAIPi3P water molecules interacted with the polar regions of the protein, improving the local sampling within the intrinsically disordered region and shielding the long-range interactions, avoiding the artificial collapse of the polypeptide chain.

The average radius of gyration also was closer to the experimental value when CAIPi3P was used. Table 6 shows all the calculated and experimental values for all tested systems. Given the high structural fluctuations in IDPs, the errors bars have a significant intersection. Hence, there is no statistical difference in this subject when TIP4P/2005 and CAIPi3P are compared.

Table 6: Radius of gyration in Å for all molecules with all used combinations. The error values are the average standard deviation of the replicas

	AMBER99SB-ILDN+TIP3P	AMBER99SB-ILDN+CAIPI3P	AMBER03WS+TIP3P	AMBER03WS+TIP4P/2005	AMBER03WS+CAIPI3P	EXP
HISTATIN5	7±1	12±2	9±2	11±2	12±1	12±0.5
R/S-PEP	10±1	9±2	9±2	12±1	13±2	13 ±0.5
ATG	13 ± 3	19 ± 3	17±3	22±4	13±2	20±2

Nonetheless, there is a considerable improvement in the accuracy of the sampled conformations when simulations were carried out with CAIPi3P solvation model. Table 7 shows that systems simulated with CAIPi3P showed the lowest difference between the calculated SAXS distribution curve and the experimental distribution.

Table 7: Root-mean-square difference between experimental SAXS and calculated SAXS radial densities. The RSD was calculated between the average distribution between replicas and the experimental curve.

	AMBER99SB-ILDN+TIP3P	AMBER99SB-ILDN+CAIPi3P	AMBER03ws+TIP3P	AMBER03ws +TIP4P/2005	AMBER03ws +CAIPi3P
Histatin 5	0,0190	0,0064	0,0120	0,0110	0,0064
R/S peptide	0,0340	0,0330	0,0082	0,0080	0,0070
At2g23090	0,010	0,003	0,011	0,014	0,014

It is essential to discuss the bulk water parameters calculated for CAIPi3P, which are summarised in Table 8. By changing the dipole moment of the TIP3P water model (Figure 27), most of the bulk water parameters were improved for CAIPi3P in comparison to the standard TIP3P model. However, several significant changes need to be addressed, such as the average oxygen-oxygen radial density distance ($R_{O\text{-}O}$) and the density. The $R_{O\text{-}O}$ distance for CAIPi3P was lower than the experimental distance, resulting in a higher density of 1.05 g/cm$^{3.}$ This results in a more compact water configuration, increasing the water-water correlation and decreasing the overall potential energy of the bulk water.

Therefore, the usage of a higher dipole yields higher barriers to reorganise the solvent surrounding the solute, which contributes to the better sampling of the protein observed in CAIPi3P simulations.

The differences between experimental and CAiPI3P bulk water parameters shows that the latter requires structural changes, as shown in table 7. These modifications may come in tuning the vibrational frequency of the H-O-H angle to modify water -water interactions to decrease the hydrogen bonds, which should result in better or the position of the charges in the molecular scaffold, following the method used in the parametrisation of the OPC model.

Table 8: Bulk water parameters calculated for CAIPi3P and TIP3P water model. These were calculated using the methods explained in Izadi and coworkers[124]

	CAIPi3P	**TIP3P**	**Experimental**
Dipole moment (μ (D))	2.69	2.34	2.5-3
Density (g/cm³)	1.05 ± 0.05	0.980	0.997
ΔH_{vap}[kcal/mol]	10.6 ± 0.005	10.26	10.52
Isobaric compressability C_p (cal/(K.mol))	23.7 ± 0.05	18.72	18
Thermal expansion α[10^−4*K^−1]	5.4 ± +0.1	9.2	2.56
O-O First peak distance(Å)	2.7	2.77	2.8
Static dieletric constant (ϵ_0)	74.5 ± 1	94	78.4
Self diffusion coefficient (m²/s)	4.67±0.2	5.5	2.3

To summarise, the reparametrised dipole moment and partial atomic charges for the TIP3P water model generated a new solvation model denoted Charge-Augmented 3 Point Water model for Intrinsically disordered Proteins (CAIPi3P). This model is transferrable, robust, and suitable for the atomistic MD simulations of IDPs, resulting in ensembles with a considerably better agreement with experimental data (SAXS). For the IDP models (histatin 5 and R/S peptide), simulations using CAIPi3P resulted in better agreement between calculated SAXS radial densities and experimental data. CAIPi3P is also applicable to studies of globular proteins and – most importantly – functionally relevant multidomain proteins bearing globular domains and intrinsically disordered regions.

Chapter 5 – Hybrid_FF and intrinsically disordered regions

Following the improvements that emerged from using CAIPI3P with known force-fields for IDPs, a framework to parametrise flexible regions in proteins has been developed. IDRs are challenging in several different aspects for protein biochemistry: their structural plasticity directly affects the dynamics of the structured domains, and this long-range effect is difficult to quantify and emulate. To improve MD simulations on this aspect, Hybrid_FF was developed. It applies AMBER03ws topological parameters into residues without secondary structures, and AMBER99SB-ILDN parameters are used for the remaining residues. This novel force-field should be directly compatible with CAIPi3P and should be used together. This framework was tested on six different proteins, which all had their structures solved via solution NMR with a diverse range of secondary structure percentage. For proteins containing a shorter terminal IDRs, the difference between using an established force field and Hybrid_FF was not substantial. Simulations of systems with lower percentages of secondary structure and longer loops showed that Hybrid_FF significantly improved the radius of gyration and the residual fluctuation. Moreover, the sampled states using CAIPi3P, coupled with Hybrid_FF, are in better agreement with the solution NMR data.

5.1 Disorder within organised systems

Intrinsically disordered regions (IDRs) are protein sequences that lack a higher organisational degree. These regions vary in length and may be functionally relevant, i.e. involved in biological roles exerted by the protein, such as signalling and molecular recognition. An example is IDRs in the proteins with established roles in autophagy. The IDR1 region of the prototypical autophagy receptor p62 promotes the oligomerisation which is important for the function of the protein in autophagy[1,10] (Figure 38 A). LC3- and Keap1-interacting regions (LIRs and KIRs) of p62 are embedded in IDRs[20] and the model of full-length p62 is consistent with experimental data available for these functional motifs (Figure 38.B). It has been also shown that functional motifs responsible for protein-protein interactions within autophagy initiation signalling complex (Atg13, Ulk1), autophagy receptors (p62, Smurf1, Nbr1), autophagosome nucleation (Becn1) and expansion complex (Atg3) are embedded within multiple IDRs, and the overall architecture resembles

“bead on elastic string” model of autophagy proteins (Figure 38.C). Literature suggests that these IDRs may serve as interaction hubs for other proteins and may adopt diverse conformations upon binding[14,20], but the thermodynamics of those regions and their influence on the biological function of the adjacent domains are currently underexplored.

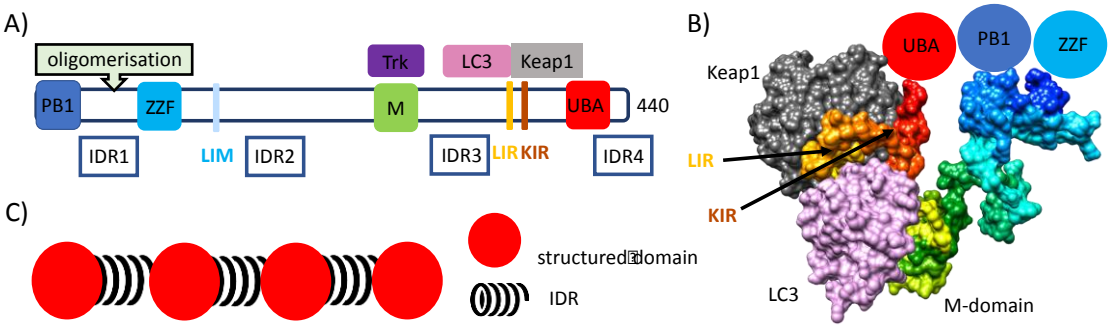

Figure 38: **A)** Diagram representation of p62 receptor. Structural domains, IDRs, key interactors, and functional sequence motifs LIM, LIR and KIR are highlighted. **B)** All-atom model of p62 IDR2-IDR3 region in complex with Keap1 (grey; PDB: 3WDZ) and LC3 (pink; PDB: 2K6Q). Relative positions of globular domains PB1 (PDB: 4MJS), ZZF (PDB: 5YPC), and UBA (PDB: 2JY7) domains are marked by circles. LIR (orange) and KIR (red) motifs are highlighted. **C)** The “beads on the elastic string” model of autophagy proteins containing IDRs. Made by Dr Agnieszka Bronowska,

Another class of proteins with IDR linkers connecting globular domains are the RNA binding proteins such as the La protein (LaP). Present in eukaryotic cells, these proteins bind to nascent RNA molecules, which protect the RNA against degradation and help throughout its maturation process. Being comprised of several different domains[208], LaP has two domains that are directly related to the RNA recognition and binding: the RNA recognition motif (RRM) and the La type-RNA binding motif (LaM). These two domains are linked by a short IDR region (12 residues long) that controls the intercommunication between domains.

Several proteins have similar domains in their sequence, such as the polypyrimidine tract binding protein isoform 1 (PTBP1) and the La-related proteins (LARPs). PTBP1 is an important regulator of RNA alternative splicing, which also affects localisation and stabilisation of the RNA. It is composed of several RRM domains, which are linked by IDR regions. Alfano and coworkers[209]

obtained the structures of RRM1 and RRM2 of PTPB1, part of the IDR that connects both of them.

The La related protein (LARP) family contains the La motif and several RNA recognition motifs. LARP6 is one of such proteins, on which aside of the similarities with LaP, regulates the expression of collagen in the cell. Martino and coworkers[210] obtained the structures of both LARP6-RRM and LaM, with terminal IDR regions being partially solved.

Because IDRs are highly disordered, they commonly represent elusive regions to study using experimental and computational methods alike. Computational studies have proved to be a crucial tool to analyse proteins and their dynamics at atomic resolution. However, IDRs simulations often result in discrepancies between the obtained data and its experimental counterpart. As discussed in section 4.4, these differences are usually attributed to bias towards globular structures. This, similarly, to IDPs, causes simulations of IDRs to be problematic for MD.

Typically, force fields are biased towards the formation of stable secondary structures, with a high propensity toward α-helices[202]. Studies have shown how these generalist descriptors such as AMBER99SB, AMBER99SB-ILDN AMBER14SB drives simulation in the direction of a compact structure[211]. The formation of stable secondary components is controlled by the strong hydrogen bonds formed between the polar backbone atoms coupled with dihedral parameters that help the helical ϕ/ψ distributions to be achieved. This is in accordance with most of the available structural data obtained by X-ray crystallography. On the other hand, with the increase of structures solved by the solution NMR, this paradigm started to shift[212]. Solution NMR studies enabled modelling and description of the dynamics of the flexible region with an atomic resolution. Hence, new force fields, such as the AMBER03, AMBER03ws and CHARMM36m, were developed. For AMBER03ws, the solution proposed was based on the modification of the LJ parameters to increase the water-protein interactions.

In principle, this should steer the protein towards a more open configuration, resulting in a less contracted conformation. As shown in Chapter 5, this is optimal for fully disordered proteins, but it facilitates the unfolding of structured domains.

Residue-specific parameters based on both AMBER99SB-ILDN and AMBER03ws were applied to a small curated dataset of six proteins solved using solution NMR and containing intrinsically disordered termini of variable lengths. Four proteins in this dataset were RNA binding domains, which required their termini to be stretched to assist with the future recognition of RNA. Upon the application of the Hybrid_FF, a series of improvements were observed. As expected, for molecules with shorter termini and higher helical/sheet content, the results resembled the results obtained by the simulations with the AMBER99SB-ILDN force field. When the termini were longer, Hybrid_FF combined with CAIPi3P solvation model sampled the protein radius of gyration within the experimental range. This resulted from the structural stabilisation caused by the application of AMBER99SB-ILDN parameters into the structured core and the scaled interaction with water in flexible regions which were parametrised using AMBER03ws.

5.2 Development of Hybrid_FF

Wei *et. al.* developed ff99FFIDP, a force field that adds CMAP corrections to specific non-structured sequences in a protein[213]. Influenced by this work and by the preliminary data, Hybrid_FF was developed.

By using DSSP to assign the secondary structure content per residue, Hybrid_FF software can assign different parameters per residue. Since previous simulations showed that AMBER03ws has a propensity to unfold stable regions and AMBER99SB-ILDN forces unstructured regions to a molten globule state, applying AMBER99SB-ILDN to structured regions and AMBER03ws to unstructured regions should improve local sampling for IDRs.

The compatibility between AMBER99SB-ILDN and AMBER03ws parameters was assured by two factors. First, both force fields have their roots found in the same force field family: AMBER96. The most significant difference between these

force fields arises from their solute-solvent interactions, given different LJ parameters for the water oxygen sigma values.

5.3 Methodology

To benchmark Hybrid_FF, a set of RNA binding protein domains with experimental structures solved by the solution NMR were selected. The set had different lengths of disordered termini and different percentages of structured components. The benchmark set is showed in Table 9.

Table 9: Molecules used to benchmark the accuracy of Hybrid_FF against known force fields.

Molecule Name	PDB code
PTB1-RRM1	**1SJQ**
PTB1-RRM2	**1SJR**
LaP-LaM	**1S7A**
LaP-RRM	**1S79**
LARP6-LaM	**2MTF**
LARP6-RRM1	**2MTG**

Each protein was parametrised with a different combination of the protein force fields and the water models, using the following pairs: AMBER99SB-ILDN+TIP3P, AMBER99SB-ILDN+CAIPi3P, AMBER03ws-ILDN+CAIPi3P, AMBER03ws+TIP4P/2005, Hybrid_FF+CAIPi3P, Hybrid_FF+TIP3P, Hybrid_FF +TIP4P/2005. For each combination, a 1 nm cubic box was centred on the structure.

Each simulation system was solvated and subsequently, sodium and chloride ions were added to the system at a concentration of 0.1 M to mimic the “physiological” concentration and to neutralise the simulation box. The bonds were constrained using the LINCS algorithm[200], setting a 2 fs time step. The electrostatic interactions were calculated using particle-mesh Ewald method[116], with a non-bonded cut-off set at 0.1 nm. All structures were energy minimised using the steepest descent algorithm for 20,000 of 0.02 nm steps. The minimisation was stopped when the maximum force fell below 1000 kJ/mol/nm using the Verlet cutoff scheme. This was followed by an NVT equilibration of 20

ps with a time step of 2 fs and positional restraints applied to the protein backbone (k = 1000 kJ/mol/nm), and a subsequent NPT equilibration (20 ps, 2 fs time step) with backbone positional restraints applied. Finally, the production simulations were run for 100 ns in triplicates, for each force field/solvent model combination. The temperature was set constant at 300 K by using an alternative Berendsen[128] thermostat (τ = 0.1 ps). The pressure was kept constant at 1 bar by using a Parrinelo-Rahman barostat with isotropic coupling (τ = 2.0 ps) to a pressure bath[201]. The results were analysed using GROMACS tools.

5.4 Results

5.4.1 Dynamic properties of IDR-containing proteins – PTBP1 RNA recognition motif domains

Two of the benchmark proteins were the RNA recognition motif domain 1 and 2 of the polypyrimidine tract binding protein isoform 1 (PTBP1-RRM1 and PTBP1-RRM2, UniProt code: P26599).

PTBP1 is ~57 kDa (531 residues) protein that plays an important regulatory role in pre-mRNA splicing and in the regulation of alternative splicing events. It binds to the polypyrimidine tract of introns and may promote RNA looping when bound to two separate polypyrimidine tracts in the same pre-mRNA. It contains four RNA-recognition motifs (RRM) domains, connected by IDRs of variable lengths.

PTBP1-RRM1 shows a canonical RRM motif with four β-strands interacting with two α-helices. It has a high percentage of IDRs (40%; 60% of the secondary structure assignment being alpha-helices and beta-sheets). The C-terminus of this domain is comprised of an unstructured motif which is 15 residues long (I84-S99). A small pocket formed by the residues forming the central β-sheet (V60, L89 and F98) is pivotal for stabilisation of that disordered C-terminus. Therefore, it could be expected that the stability of the whole protein should be directly correlated with the beta-sheet – loop interactions maintained throughout the simulation.

MD simulations of the PTBP1-RRM1 applying different force fields and water models gave very different outcomes. As showed in Figure *39*, parametrisation with AMBER99SB-ILDN+TIP3P resulted in an inaccurate collapsed structure,

reflected by a smaller radius of gyration when compared to the distribution given by either AMBER03ws or Hybrid_FF (Figure 39 .A). AMBER99SB-ILDN was unable to adequately sample the IDRs, including the C-terminus since its parameters were not set to describe highly flexible regions properly. This also affected the convergence of the radius of gyration (Figure 39 .B). Given the biased nature of the generalist combinations, such as AMBER99SB-ILDN with TIP3P, they were unable to sample experimental ranges of the radii of gyration. This force field did not reach a reasonable convergence in the simulation time scale. However, when AMBER99SB-ILDN or AMBER03ws were combined with CAIPi3P, both sampled average values and convergence was markedly improved

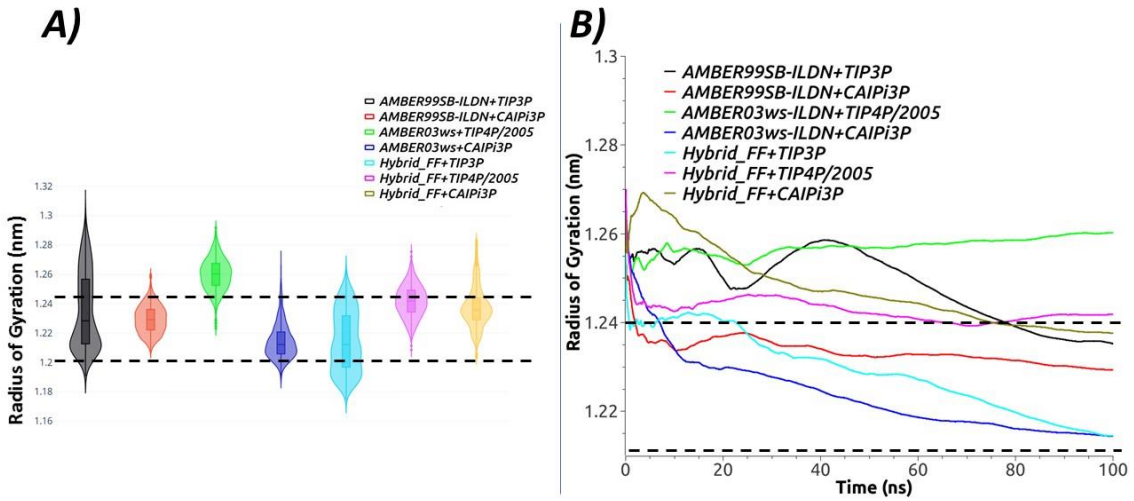

Figure 39: Radii of gyration of the PTBP1-RRM1: A) Radius of gyration distribution for different combinations of force field/water model combinations; B) Radius of gyration cumulative convergence for different force fields/water models.

Simulation with AMBER03ws protein parameters with TIP4P/2005 solvation model showed the fastest convergence. This was unexpected since AMBER03ws is prone to attain expanded conformations with high radial fluctuations. Two combinations with the best agreement with experimental data were the Hybrid_FF with TIP3P and AMBER03ws with CAIPi3P. The accuracy of the first combination can be explained by AMBER03ws assigning higher flexibility to the unstructured loops, while CAIPi3P ensuring the structure stability given its higher dipole moment. Good agreement between the Hybrid_FF with TIP3P and

experimental data can be rationalised by Hybrid_FF being able to assign residue-specific parameters which can simulate flexible regions adequately and TIP3P, which is prone to “collapse” unfolded sequence motifs.

The C-terminus of the PTBP1-RRM1 is comprised of polar residues. Since CAIPi3P is highly polarised, the charge-charge interaction between the solute and solvent is increased. This can explain why the structural averages (Figure 40) of PTBP1-RRM1 simulated using CAIPi3P show the C-terminal region extended, resulting in a higher radial distribution.

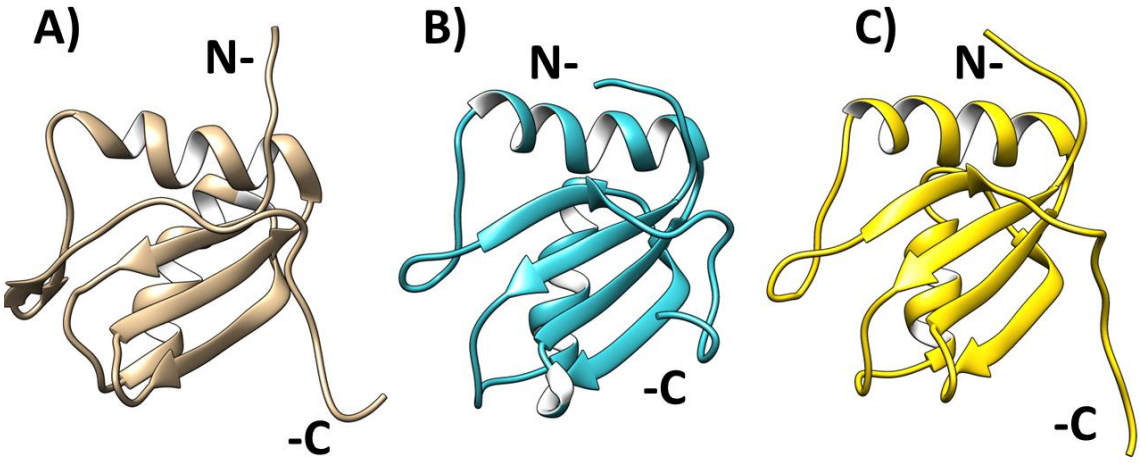

Figure 40: Structural averages for PTBP1-RRM1: A) Experimental average structure; B) Hybrid_FF with TIP3P; C) Hybrid_FF with CAIPI3P. The termini were sampled in a closed conformation to the experimental conformation when TIP3P was used. CAIPi3P overstretched the C-terminus, resulting in a larger radius of gyration.

PTBP1-RRM2 structure shows a spherical, globular core with disordered termini. The RRM2 domain is a critical component of the PTB1P-RNA interaction. While RRM3 and RRM4 mediate direct RNA-protein interactions, RRM1 and RRM2 contribute significantly to the allosteric stabilisation of the RNA binding event[214].

PTBP1-RRM2 sequence contains 47% of unstructured residues, which is higher than for the RRM1. These unstructured residues concentrate predominantly in the region located between residue S121 and K134, which is the longest IDR in the whole protein.

The force field choice affected heavily the simulation outcomes. Figure 41 shows that the usage of AMBER03ws, regardless of the solvent selection, unfolded the protein. For the radius of gyration convergence, five out of seven force field-water combinations attained a reasonable converged state in 100 ns time scale. These resulted in the ensemble which was very different from the solution NMR ensemble due to AMBER03ws properties that drive the ensemble towards more unfolded conformations. This is beneficial in simulations of IDPs, but not multidomain proteins, wherein globular domains are connected by with IDRs.

AMBER99SB-ILDN simulations achieved good agreement with the experimental data, which is consistent with the known good performance of AMBER99SB-ILDN in simulations of globular proteins[215].

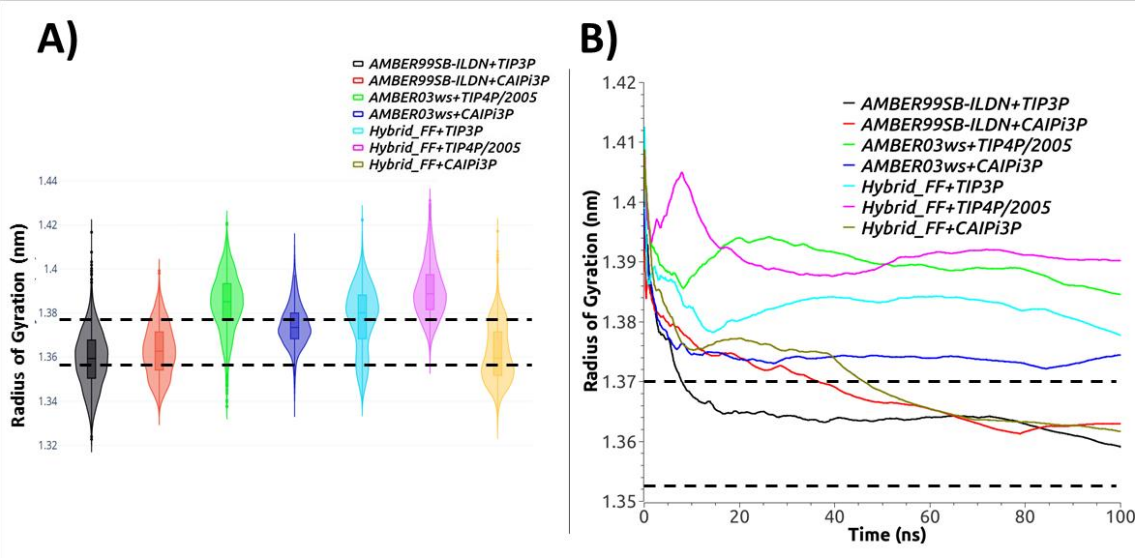

Figure 41: Radii of gyration for the PTBP1-RRM2: A) Radius of gyration distribution for different combinations of force field/water model combinations; B) Radius of gyration cumulative convergence for different force fields/water models. PTBP1-RRM2 showed a highly organised and globular structure with short loop regions.

Since RRM2 is pivotal for protein-RNA interaction and molecular recognition, it has a high concentration of superficial polar residues. This explains a good agreement with the experimental radius of gyration when CAIPi3P was used in combination with AMBER99SB-ILDN and/or Hybrid_FF. CAIPi3P by itself was not able to maintain the internal hydrophobic core collapsed, as could be observed when AMBER03ws has been used. Therefore, even though the simulation results showed more unfolding, this was not related to the water model

used. The Lennard-Jones radius differences between AMBER99SB-ILDN and AMBER03ws for the protein atoms affects the electrostatic interactions that hold the secondary structures together. This makes the intra-chain solute-solute interactions more prone to unfold. This feature is important in simulations of IDPs, but it is not suitable for handling folded domains.

5.4.2 Hybrid_FF retains the dynamics of both core and loop residues – La protein RNA binding mediators

The La protein (LaP) is a phosphoprotein located in the nucleus. Specifically, human LaP (Uniprot ID: P05455) is a multidomain protein which mediates the interaction with the RNA through its N-terminal domain. Within this region, there are two subdomains: La motif (LaP-LaM) and central RRM domain (LaP-RRM), both interact directly with RNA strands. Alfano and coworkers[209] solved the structures of both domains by solution NMR, showing a high percentage of residues without assigned secondary structure.

LaM shows 54% of secondary structure assignment. The N-terminal region of LaM shows a high spatial fluctuation of residues N7, E8 and K9 and high propensity to an intrinsic disorder. The C-terminal region is comprised of two short β-strand sheets spanning residues S75-E103. E78, which is shown to be one of the critical residues for the RNA binding event, is located within this region.

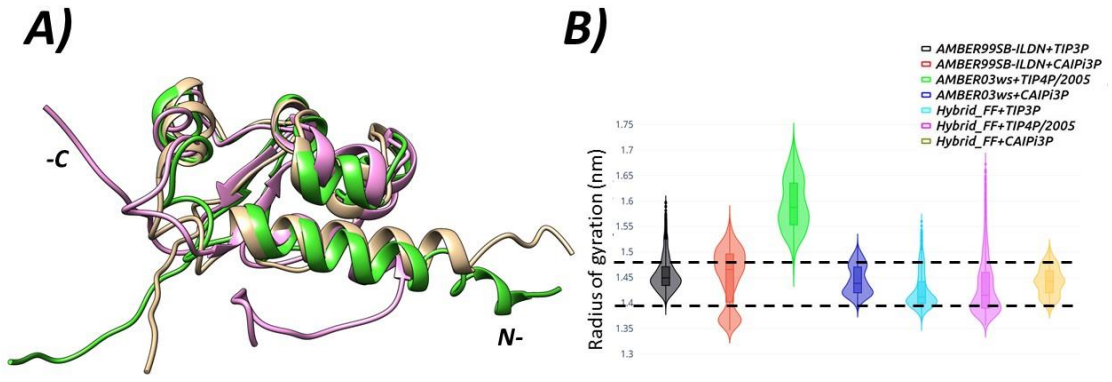

Figure 42: LaP-LaM simulations and N-terminal stability. A) Structural averages of the obtained ensembles: AMBER03ws+TIP4P/2005 (green), Hybrid_FF+TIP4P/2005 (purple), and experimental structure (ochre). B) The radii of gyration distributions obtained for different combinations of force field/water model tested. AMBER03ws+TIP4P/2005 was unable to maintain

scaffold cohesion, unfolding the terminal regions, resulting in a higher average radius distribution. Hybrid_FF combined with TIP4P/2005 unfolded the N-terminus, which collapsed towards the alpha-helical structure, resulting in a lower radius of gyration.

MD simulations of LaP-LaM using six combinations of force fields and solvation models (Figure 42) showed a good agreement with the experimental ensembles. This was not unexpected, considering a high secondary structure content of LaP-LaM. In particular, Hybrid_FF combined with CAIPi3P solvation model, generated a configuration with a distribution which showed a good agreement with the experimental data (Figure 42). The only combination which yielded results with a poor agreement with experimental data was the AMBER03ws+TIP4P/2005. In simulations carried out employing AMBER03ws+TIP4P/2005, the C-terminal region unfolded (Figure 42.A). This strongly indicated that AMBER03ws was not appropriate to simulate globular domains, especially when combined with the TIP4P/2005 water model. In the previous study, AMBER03ws+TIP4P/2005 combination yielded good agreement with experimental data, but that study focused on a set of full-length IDPs[176].

Considering the lower content of structured residues in the RRM domain of La (48% of structured residues), it was expected that the parametrisation of less ordered configurations would yield better agreement with the experimental data. Simulations with either AMBER03ws or Hybrid_FF protein force fields and TIP4P2005 solvation model over-sampled extended configurations, as shown in Figure 43. TIP4P2005 water model was developed to scale the water-protein interactions less favourably than the protein-protein interactions, causing a deformation of the hydrophobic core. Over-sampling of extended configurations occurred also in the Hybrid_FF simulations, even though Hybrid_FF assigns AMBER99SB-ILDN parameters to residues in the structured core. Thus, this shows that the TIP4P/2005 water model should not be used in simulations of multidomain proteins containing IDRs, as it was unable to capture the LaP-RRM molecular dynamics accurately. AMBER03w+TIP4P/2005 and Hybrid_FF+TIP4P/2005 were unable to maintain the N-terminal experimental conformations. When Hybrid_FF was used combined with CAIPi3P solvation model, the resulting ensemble reached the distribution which fitted well within the experimental range

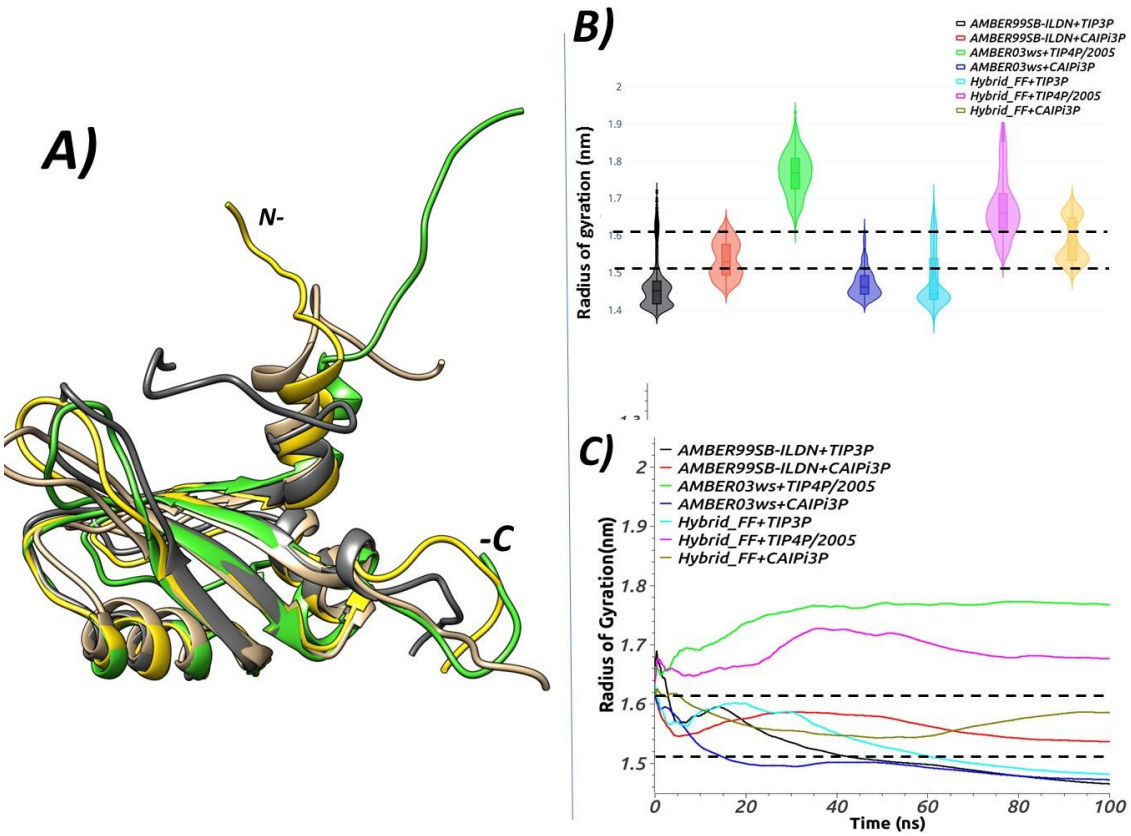

Figure 43: LaP-RRM simulations with different combinations of the protein force field and solvation models a) Structural averages of the attained ensembles for the LaP-RRM: AMBER99SB-ILDN+TIP3P as grey, Hybrid_FF+CAIPi3P as yellow, AMBER03ws+TIP4P/2005 as green, and experimental structure as ochre. B) The radius of gyration distribution for the investigated force fields/water models combinations C) Convergence of the radii of gyration for all different force fields/water models combinations.

In contrast, simulations which employed Hybrid_FF jointly with CAIPi3P sampled experimental configurations accurately. These resulted in the best quartile distribution of radii values within the experimental values.

The convergence curves for the LaP-RRM simulations showed a good convergence for most of the force field/solvation combinations tested. Similarly, to the other RRM domains investigated in this work, CAIPi3P water model showed better agreement with the experimental data than the other water models. The canonical RRM secondary structural topology showed a mixed α-β structure, with the rigid hydrophobic core. This may explain the improvement in the experimental agreement for the RRMs since Hybrid_FF assigned AMBER99SB-ILDN parameters to structured regions of the protein and CAIPI3P

enhanced the stabilisation of the disordered regions and an overall improvement of the sampling.

5.4.3 LARP6-LaM and LARP6-RRM1

La-related proteins (LARPs) form a large family of RNA-binding eukaryotic proteins, involved in cell growth and proliferation primarily through the regulation of protein synthesis.

All LARPs are comprised of seven distinct protein families. Others than LARP6, investigated in this chapter, are LARP1, LARP1B, LARP3 (aka genuine La or SSB), LARP4A, LARP4B, and LARP7. All LARPs contain the La module, which is a conserved domain for RNA binding. The La module is assembled by two domains: the RNA recognition motif 1 (RRM1) and the La motif (LaM). Their synergistic work regulates the interaction with RNA and the dimer-nucleotide configuration.

Martino and collaborators resolved the structures of both domains of LARP6 separately using solution NMR. This lowers the structural content of LARP6-LaM to 34%, which is the lowest of all domains in this work and among all LARPs. LARP6-LaM has the longest solved C-terminal IDR, containing 30 residues length IDR located between T70-E90.

LARP6-RRM1 has two intra-domain short loops: one spanning A199-G206 residues, and another one between residues Y250 and E257. Since both RRM domains (PTBP1 and LARP6) investigated in this study have a high secondary structure content, the overall effect on the protein dynamics should be similar. LARP6-RRM1 was less prone to unfold, and it maintained a stable radius of gyration in simulations using AMBER03ws combined with CAIPi3P water model, as showed in Figure 44. Both AMBER03ws with TIP4P/2005 and the Hybrid_FF+TIP3P showed an expanded radius of gyration. Hybrid_FF+TIP3P simulations yielded expanded conformations, which showed that Hybrid_FF framework works even with the TIP3P water model. The region between A199-G206 expanded after 60 ns of simulation (Figure 44), which resulted in a more extended radius of gyration distribution. The ability to sample open and closed conformations for the Hybrid_FF+TIP3P showed that the residues parametrised

with AMBER03ws can fluctuate between loop-core and loop-solvation interactions even when TIP3P is used.

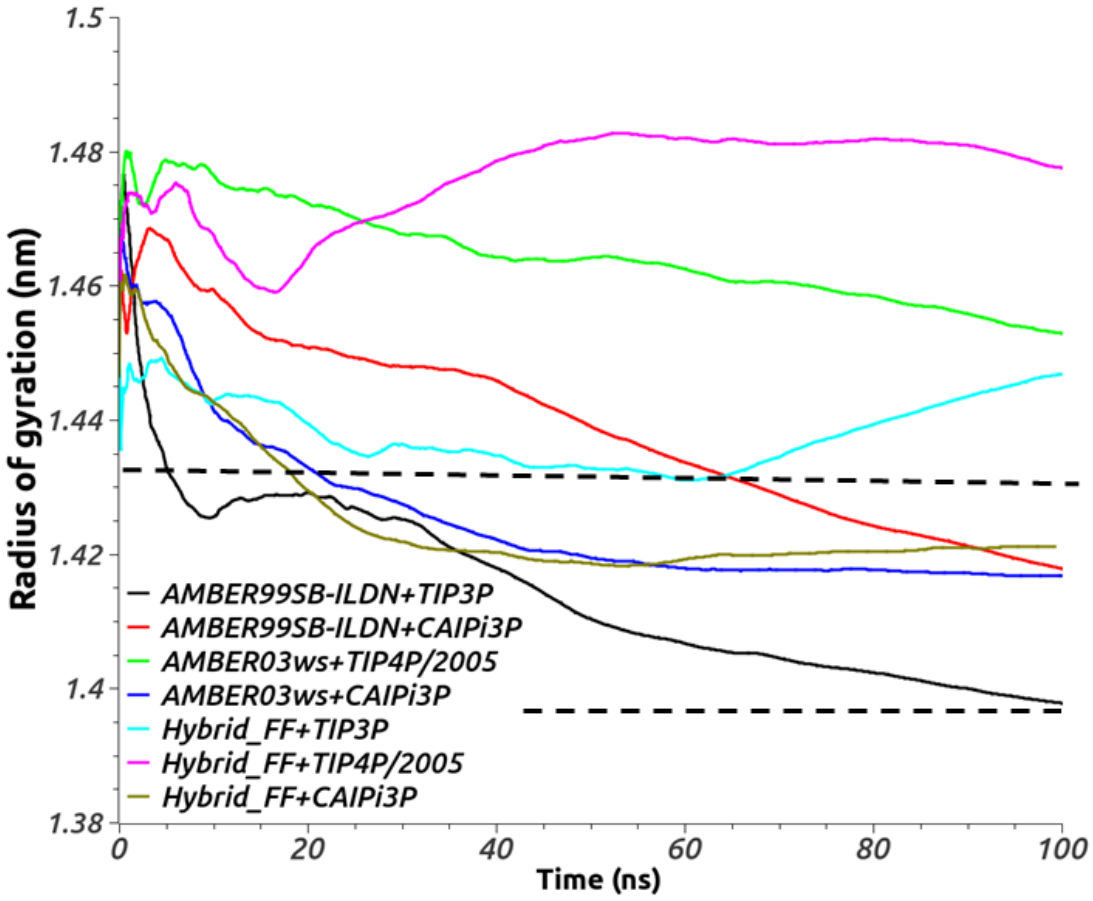

Figure 44: Cumulative convergence of LaP-RRM1 radius of gyration: Hybrid_FF with CAIPi3P converged faster in comparison to AMBER03ws+TIP4P/2005 and AMBER99SB+CAIPI3P.

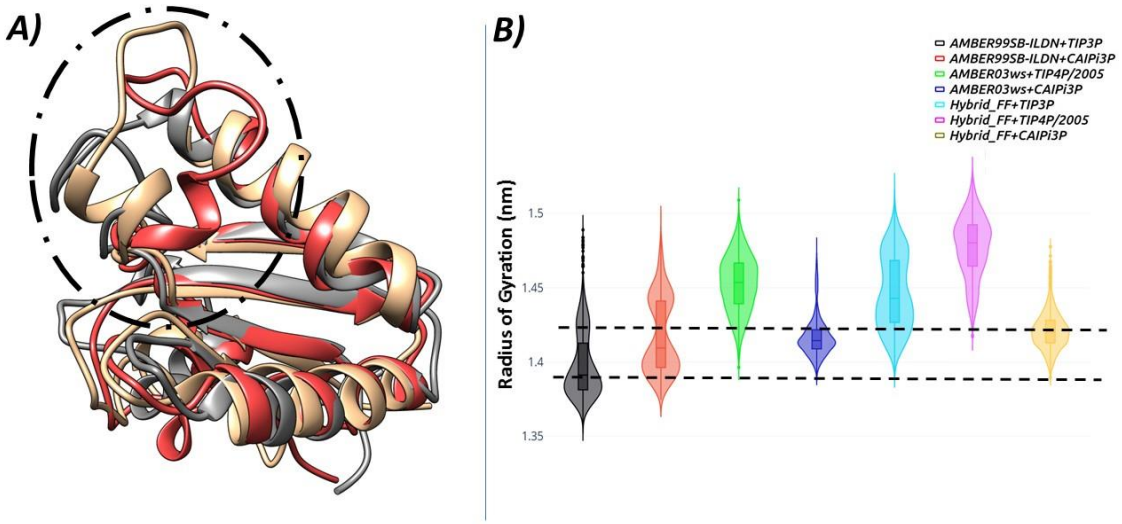

Figure 45: Conformations of LARP6-RRM1 and their respective radius of gyration. A) Representative average structures for different simulations of the LARP6-RRM: Experimental structure is coloured ochre, AMBER99SB-ILDN+TIP3P is coloured grey, and AMBER99SB-ILDN+CAIPi3P is coloured red. The loop between residues A199-G206 is highlighted with dashed lines. B) The radius of gyration distributions for different force field/water model combinations. The highlighted loop is the region with the highest RMSF between different combinations of force-field/water models.

Hybrid_FF with CAIPi3P achieved not only a good distribution of radii of gyration within the experimental range but a faster convergence towards an accurate value (Figure 45). Interestingly, the ensembles of a loop located between A199-G206 residues generated by different force fields varied considerably. In AMBER99SB-ILDN simulations, the A199 unfolded from the helical conformation, forming a series of semi-structured loops (Figure 45 A) for both water models tested, which contrast with the open-loop configurations sampled by both AMBER03ws and Hybrid_FF.

LARP6-RRM1 showed a different convergence profile in comparison to the PTBP1-RRM2. PTBP1-RRM2 has a lower structural content (44%) in comparison to LARP6-RRM1 (53%), but it converged faster compared to LARP6-RRM1, particularly in AMBER99SB-ILDN simulations. This strongly indicates the erroneous bias applied to flexible regions when using a generalist force field such as AMBER99SB-ILDN. In contrast, Hybrid_FF, combined with either TIP3P or CAIPi3P, showed an improvement in the convergence, especially in LARP6-RRM1 simulations. This force field framework impeded a fast collapse, allowing the protein to equilibrate in a more extended, near-native ensemble, as shown in Figure 45.

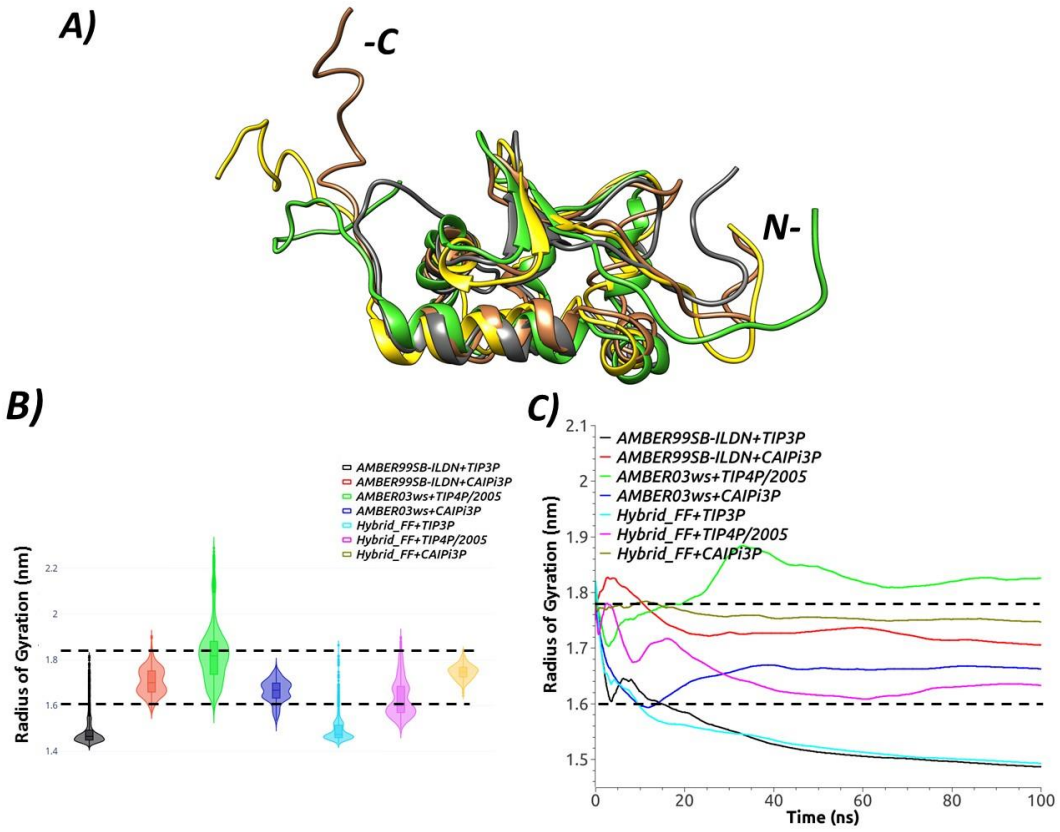

Figure 46: LARP6-LaM simulation results. A) Ensemble averages: the experimental structure is coloured ochre; AMBER99SB-ILDN+TIP3P is coloured grey, AMBER03ws+TIP4P/2005 is coloured green, and Hybrid_FF+CAIPi3P is coloured yellow. B) The radii of gyration distribution force field/water model combinations C) Convergence of the radii of gyration for all investigated combinations of the force field and water models.

All simulations that used CAIPi3P model attained values within the experimental range of the radius of gyration. In particular, Hybrid_FF+TIP4P/2005 achieved an average value within the experimental radius boundaries (Figure 46 B). The disordered residues in the N-terminus were stretched and retained an extended conformation (Figure 46.A), without collapsing, in contrast to the structured core, which retained the globular conformation expected for proteins parametrised with AMBER99SB-ILDN simulations.

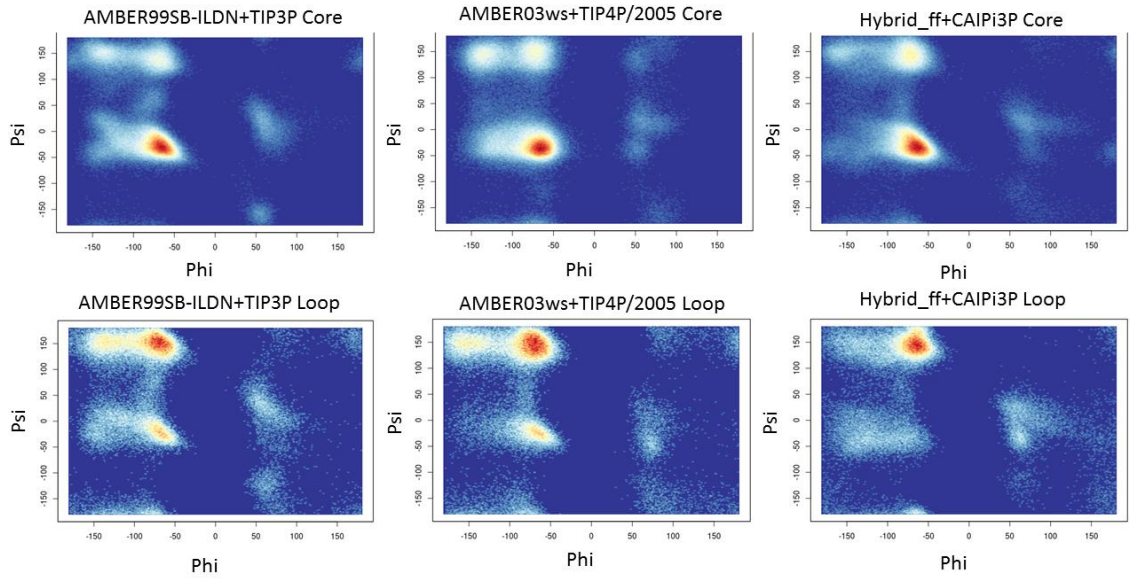

Figure 47: Ramachandran plots of different areas of LARP6-LaM: Hybrid_FF kept the structured core configuration, similar to the TIP3P+AMBER99SB-ILDN, it also improved sampling of the disordered region compared to AMBER03ws+TIP4P/2005.

As shown in Figure 47, the Ramachandran distributions for both structured core and unstructured N-terminus are significantly different between force-fields and water models. Hybrid_FF+CAIPI3P sampled Φ/Ψ dihedrals similar to AMBER99SB-ILDN in the globular core, and similar to AMBER03ws in the loops, which resembled disordered backbone Φ/Ψ distributions[212].

6.4.4 Residual flexibility in comparison to experimental data

The root-mean-square deviation between simulated and experimental residual spatial fluctuation was calculated to assess how the force field/water combinations affected the protein residual flexibility. As shown in Figure 48, Hybrid_FF+CAIPi3P is the most accurate combination regarding the residual fluctuations. For LARP6-LaM, all force fields resulted yielded high values of radius of gyration average values, but it is notably higher for the AMBER03ws+TIP4P/2005 combination.

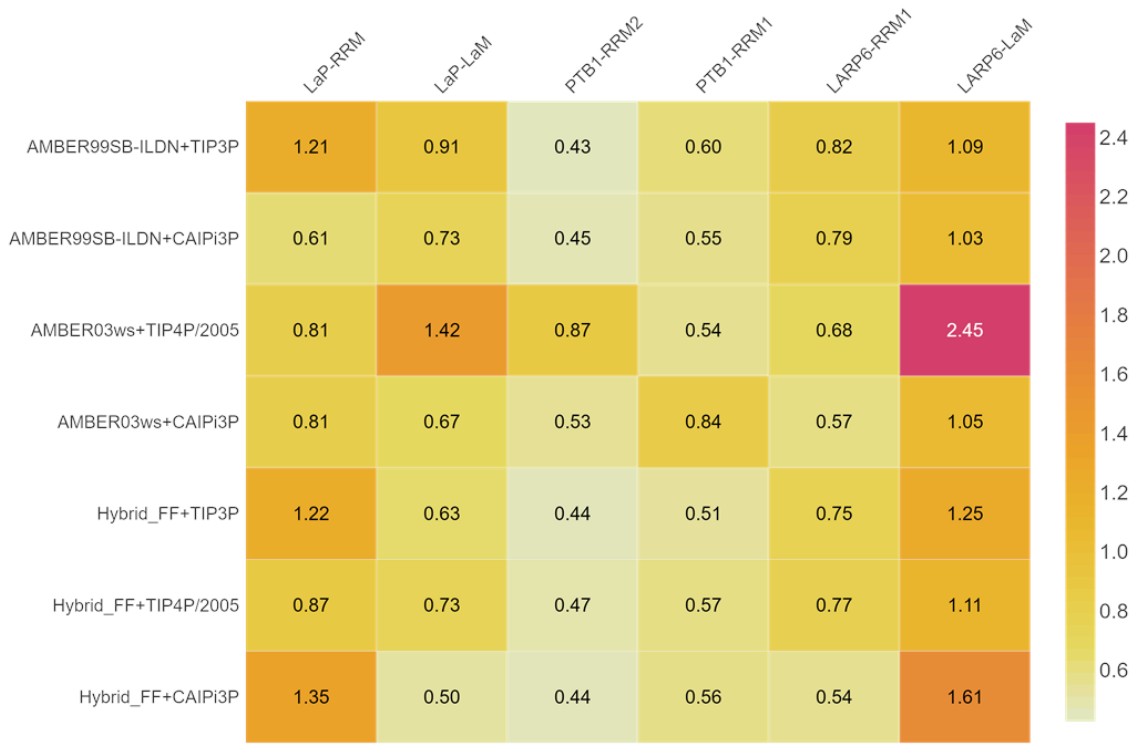

Figure 48: Spatial fluctuation RMSD [in angstroms] for all proteins using different force field/water model combinations. RMSD gradient is colour-coded: highest – red and lowest – white. AMBER03ws+TIP4P/2004 shows the highest RMSD fluctuations for all proteins. Hybrid_FF+CAIPi3P shows the lowest RMSD fluctuations values.

The parameters for Hybrid_FF applied to disordered regions improved the sampling of the system, compared to the AMBER99SB-ILDN force field. Also, Hybrid_FF achieved good agreement with experimental data and improved sampling when compared to simulations using AMBER99SB-ILDN for proteins with low content of disordered residues. Interestingly, CAIPi3P coupled with Hybrid_FF obtained results for LAR6-LaM, the domain with the longest disordered region. Thus, even with accurate predictions of the radii of gyration, it failed to grasp the residual fluctuation in the N-terminus. This can be corrected by increasing the length of the simulations, which should result in a better sampling of disordered regions.

5.5 Discussion

In this part of work, it has been shown how parametrisation affects the molecular dynamics of six RNA-binding proteins with a diverse range of structural disorder. The RRM1 and RRM2 domains in PTBP1 showed that AMBER99SB-ILDN generalist force field and AMBER03ws force field developed for IDPs were unable to sample the correct radius of gyration. The RRM1 domain, with 60% folded structural content, achieved the best results when the Hybrid_FF framework was applied. Hybrid_FF parametrised a protein ensuring that more flexible, disordered regions such as the RRM1 termini do not over-collapse on the protein globular core. For more flexible PTBP1-RRM2 domain, Hybrid_FF combined with CAIPi3P water model resulted in the most accurate sampling and the best agreement with the experimental radii of gyration

The simulations of LaP-LaM domain yielded the N-terminus region prone to partial unfolding. This unfolding characteristic can be observed in simulations using the AMBER03ws+TIP4P/2005 combination. As works by Best *et. al*.[174] and Henriques *et al*.[176] demonstrated, scaling of solute-solvent interaction through modifications of Lennard-Jones parameters diminished solute-solute interactions, resulting in a more unfolded configuration for the partially structured termini. The application of structure biased force fields prevented the N-terminus unfolding, which resulted in the improved agreement with the experimental data of the radii of gyration, sampled in simulations using AMBER99SB-ILDN and Hybrid_FF.

The LaP-RRM domain has a long partially unfolded N-terminal region. Due to its inherent flexibility, this model proved a challenging test set, given most of the force field/water model combinations failed to sample experimental conformations correctly. The best outcome was reached by using CAIPi3P+Hybrid_FF. This combination sampled conformations that converged quickly within the experimental radius of gyration,

The LARP6-LaM domain has the lowest secondary structure content. The best results were achieved via two protocols: Hybrid_FF+CAIPi3P, and Hybrid_FF+TIP4P/2005. Results obtained with CAIPi3P can be explained by the favourable solute-solvent interactions, while the Hybrid_FF+TIP4P/2005 can be

explained by the AMBER99SB-ILDN parameters applied to the protein core. The core was tightly held together as the loops favourably interacted with the solvent, given the compatibility between TIP4P/2005 and AMBER03ws.

Importantly, Hybrid_FF retained unstructured characteristics of long disordered loops without losing the structured characteristics of the hydrophobic core. When the results of simulations using AMBER03ws+TIP4P/2005, AMBER99SB-ILDN+TIP3P and Hybrid_FF+CAIPi3P had their backbone dihedrals compared, only the Hybrid_FF+CAIPi3P retained terminal residues within disordered dihedral distributions while keeping backbone torsions of the folded core within structured values for φ/ψ.

All macromolecules used in this study are specific domains of RNA binding proteins, showing similar fold overall and the variable content of disordered residues. Therefore, to further validate the methodology, expanding the dataset towards different protein classes will be done. Future works will include LC3- and Keap1-interacting regions of p62 autophagy receptor.

It is important to note that Hybrid_FF requires the definition of structured and unstructured regions to be made *a priori,* similar to ff99FFIDPs[213]. Therefore, if the initial secondary structure assignment is incorrect, simulation artefacts will be very likely observed. This should be minimised by the compatibility between AMBER03ws and AMBER99SB-ILDN parameters, but this initial step should proceed with caution.

In summary, two established force fields were tested as benchmarks against a hybrid framework built upon their parameters, resulting in the Hybrid_FF force field. The Hybrid_FF was showed to be accurate for the simulation of IDRs, resulting in an ensemble with a better experimental agreement for the radius of gyration, and without losing the globular core cohesion. Hybrid_FF integrates the set of tools designed in this work, which should help to correctly model flexible regions of structured and partially structured proteins.

The source code for Hybrid_FF software can be found at github.com/ammvitor

Chapter 6 – Stability and long-range effect of disordered loops

Following the development of the CAIPi3P water model (Chapter 5) and the Hybrid_FF (Chapter 6) force field framework to improve sampling in flexible regions, the focus shifted towards the assessment of the overall dynamics, entropic effects of flexible regions in protein, and their quantification. Recently, studies on the entropic force caused by flexible region within protein structures have shown that long-range entropic effect may be essential to understand the protein dynamics[216,217].

This chapter describes the development of a method for configurational entropy calculations using a first-order approximation approach, denoted SQuE (Structural Quantifier of Entropy). SQuE is rigorous, computationally efficient, robust and numerically reliable for calculating configurational entropy contribution to free energy in protein-ligand complexes and other macromolecular interactions. It has been validated by the test case of the UDP-glucose 6-dehydrogenase (UGDH) for benchmarking and a series of PAS-B domains to show the applicability and predictive ability of SQuE.

6.1 Established methods for the configurational entropy calculations of macromolecules

Entropy calculations represent one of the most challenging steps in assessing the thermodynamics and obtaining the binding free energy in proteins and their complexes, which is a grand challenge in computational biology [218,219]. The free energy landscape defines all the thermodynamic characteristics of a protein system, as the binding affinities of protein-ligand[217,220] and protein-protein interactions[221]. It also governs other vital processes such as solvation of small molecule ligands[222] and enzymatic reactions [223].

Unfortunately, calculation of absolute entropy by atomistic molecular simulations such as molecular dynamics (MD) and Monte Carlo is challenging, if not impossible[224]. This difficulty arises due to entropy being a function of the multi-dimensional configurational partition function[35,114], which is not attainable by this simulation methods[113]. Atomistic simulations are, however, widely used by the

computational biology community and considerable attention has been devoted to approximating the configurational entropy from the atomistic simulations.

One of the most popular approaches, which is based on quasiharmonic (QH) analysis, has been introduced by Schlitter[225]. He proposed estimating absolute and relative entropies of a macromolecule based on the evaluation of the covariance matrix of Cartesian positional coordinates obtainable by atomistic molecular dynamics (MD) simulation. Schlitter's approach requires only a calculation of a determinant, which was the reason for its popularity among computational biologists. In Schlitter's framework, the approximation for the absolute entropy represents an upper limit for the quantum mechanical entropy[225], however, it may be slow and may be difficult to converge. Another popular and well-established approach is based on the normal mode (NM) analysis[226]. Both methods and their shortcomings were explained in detail in chapter 4.

6.2 Long-range entropy effects: Development of SQuE

Alternatively to QH approach and NM analysis, entropic force contributed by a given IDR can be calculated using Structural Quantifier of Entropy (SQuE) approach, developed in this work: from the ensemble obtained by MD simulation, one can define an orthogonal basis set K, by finding the linear uncorrelated motions. K is comprised of m eigenvectors K_i as such; the ensemble configuration can be described as (Eq. 39)[227]:

$$R(x, y, z, t) = R(k_1, k_2, k_3, \dots, k_m) \tag{39}$$

By definition, the motions in the basis K are uncorrelated. Therefore, the probability function per state is uncorrelated as well, hence (Eq. 40):

$$\begin{aligned} &\rho\big(R(x, y, z, t)\big) = \rho\big(R(k_1, k_2, k_3, \dots, k_m)\big) \\ &= \rho\big(R(k_1)\big)\rho\big(R(k_2)\big)\rho\big(R(k_3)\big) \dots \rho\big(R(k_m)\big) \end{aligned} \tag{40}$$

For this ensemble, one can define Shannon Entropy[228] as Eq. 41:

$$S = -\int \rho(R(x,y,z,t))\, ln(\rho_i(R(x,y,z,t))dR \tag{41}$$

applying equation 34 to equation 35 (Eq. 42):

$$S = \sum_{i=1}^{m} -\int \rho(R(k_i))\, ln(\rho(R(k_i))dR \tag{42}$$

so, the entropy per basis will be (Eq. 43):

$$S_i = -\int \rho(R(k_i))\, ln(\rho(R(k_i))dR \tag{43}$$

6.3 The importance of intrinsically disordered regions for structural entropic compensation

Initially, the performance of SQuE was tested on UDP-glucose 6-dehydrogenase (UGDH), which was recently extensively studied by Keul *et. al*. [216] and focused on the entropic effects of UGDH 30 residues long intrinsically disordered C-terminus (ID-tail). The study reported that the entropic effect caused by the ID-tail affected UGDH its function and modulated its allosteric switch.

UGDH is allosterically regulated by UDP-α-D-xylose, a negative feedback system. Unusually, the same active site serves both purposes: the enzyme forms an inactive hexamer that can break up into three active dimers, and this switch depends on competition between the UDP-α-D-glucose substrate and the UDP-α-D-xylose allosteric ligand. While the substrate behaviour remained unchanged in the absence of the ID-tail, the affinity for the allosteric ligand was tenfold lower. The authors of the study tried a whole range of mutations in the ID-tail, including switching all lysines to serines, swapping out all the prolines, and introducing thirty serines in a row. None of these had any effect on the affinity. Varying the length of the ID-tail showed a simple exponential-decay relationship with the

UDP-α-D-xylose affinity, with the most favourable length of that tail to be around 30 residues. ID-tail, regardless of its sequence, seemed to be tied to entropic effects.

To assess this effect at the atomistic level of detail, MD simulations of UGDH were performed, followed by configurational entropy calculations. The performance of SQuE was compared with configurational entropy calculations based on quasiharmonic (QH) analyses. The results showed that SQuE outperformed QH approaches in terms of speed of calculations and the agreement with experimental data, particularly when the system was parametrised with Hybrid_FF framework combined with CAIPi3P solvation model.

As a prediction test of structural entropy barriers, SQuE was then applied to four different Per-Arnt-Sim B (PAS-B) domains: nuclear receptor coactivator 1 (NCOA1), hypoxia-inducible factors HIF-1α and HIF-2α, and aryl hydrocarbon receptor (AhR). PAS domains occur in proteins from all kingdoms of life. In animals, proteins containing PAS domains are regulating responses to hypoxia [229,230], circadian rhythms[230–232], hormonal stimuli[233], synaptic plasticity[234], memory[235], and xenobiotic stress[236], exerting their activity as transcription factors and nuclear receptor coactivators. The PAS domain is also a part of hERG potassium channel[237].

Within the conserved structural scaffold of PAS-B domains, there is a region known as loop1. Computational and structural studies have shown that the loop1 region located in the NCOA1-PASB domain may undergo through substantial conformational changes on nanoseconds time scale and adopt helical or partially unfolded configurations. This behaviour has implications for NCOA1 modulation and related coactivators by small molecule ligands, and we have suggested a binding mode for several confirmed NCOA1 binders. These characteristics made the PAS-B domains a case study for the predictive capabilities of SQuE.

Receptor-ligand interaction studies were carried out, via molecular docking, umbrella sampling and MMPBSA to provide insight on how the loop dynamics would affect the ability of PAS-B domains to bind small molecules

Five known small molecule NCOA1 binders were used to assess the binding energy landscape. Alongside it, binding site prediction tools were used on both conformations to study the formation of possible transient binding sites resulting from this conformation change.

The proposed framework introduces an intuitive approach to quantification of the configurational entropic effect in protein-ligand complexes and other macromolecular binding events, and it offers a reliable calculation of the configurational entropy per defined region.

6.4 Methods

6.4.1 UGDH setup and simulations

All UGDH simulations were performed using GROMACS 5.13 [199,238]. The UGDH monomer was parametrised with seven different combinations of force fields and water models: AMBER03ws[239] + TIP4P/2005[158], AMBER99SB-ILDN[215] + CAIPi3P, AMBER99SB-ILDN + TIP3P, AMBER03ws + CAIPi3P, Hybrid_FF + CAIPi3P, Hybrid_FF + TIP3P, and Hybrid_FF + TIP4P/2005. All simulated systems had the box distance set to 1 nm, and periodic boundary conditions were applied. The boxes were solvated with the necessary number of water molecules and Na+, and Cl- ions were added to achieve a 0.1 M concentration and maintain charge neutrality. The solvated systems were energy minimised and equilibrated. The minimisations ran using steepest descent for 1,000 cycles followed by the conjugate gradient. Energy step size was set to 0.001 nm, and the maximum number of steps was set to 50,000. The minimisation was stopped when the maximum force fell below 1000 kJ/mol/nm using the Verlet cutoff scheme. Treatment of long-range electrostatic interactions was set to Particle Mesh-Ewald (PME), and the short-range electrostatic and van der Waals cutoff set to 1.0 nm.

After the minimisation, NVT equilibration was performed for 20 ps with a time step of 2 fs and positional restraints applied to the backbone, while the system was heated from 0 to 300 K. The constraint algorithm used was LINCS, which was applied to all bonds and angles in the protein. With the Verlet cutoff scheme and the non-bonded short-range interaction, the cutoff was set to 1.0 nm. Long-range electrostatics were again set to PME. The temperature coupling was set between

the protein and the non-protein entities by using a Berendsen thermostat, with a time constant of 0.1 ps and the temperature set to reach 300 K with the pressure coupling off. NPT equilibration was then run at 300 K with a Parrinello-Rahman pressure coupling on and set to 1 bar, without positional restraints. The equilibration trajectories were set to 10 ns (discarded from the analysis), and the production MD simulations were performed for 100 ns, carried out in triplicates (300 ns cumulative production runs).

6.4.2 Modelling of disordered and helical conformations of PAS-B domains

In the investigations regarding conformation transitions of loop1 of PAS-B domains, alternate models of PAS-B domains of human aryl hydrocarbon receptor (AhR), and hypoxia-inducible factors HIF-1α, and HIF-2α, with loop1 adopting the helical conformation, were generated using the human nuclear receptor coactivator 1 (NCOA1) solution NMR structure as a template (PDB code: 5NWM). The alternate model of NCOA1, with loop1 adopting the partially unfolded conformation, has been generated using the human HIF-2α crystal structure as a template (PDB code: 5UFP). These models have been generated by SWISS-MODEL.

6.4.3 Conformational transitions of the loop1 within PAS-B domains

Four pairs of initial conformations of PAS-B domains were used in the analysis of the transitions between the loop and helical conformers of the loop1. The experimental structures of PAS-B domains with the partially disordered loop1 conformation of the HIF1-α and HIF2-α (PDB codes 4H6J and 5UFP, respectively) were matched with alternative structures, where loop1 adopted α-helical conformations. All simulations were performed using GROMACS 5.13. All four pairs of PAS-B conformations were parametrised using AMBER-ff03ws[43] force field and simulated with TIP4P-2005[44] water model. This combination was used to improve sampling of disordered protein regions.

Box distance was set to 1 nm from the edge of the protein, and periodic boundary conditions were applied. The box was solvated and Na+, and Cl- ions were added to achieve a 0.1M concentration and maintain charge neutrality. The solvated

systems were energy minimised and equilibrated. The minimisation ran using steepest descent for 1,000 cycles followed by the conjugate gradient. Energy step size was set to 0.001 nm, and the maximum number of steps was set to 50,000. The minimisation was stopped when the maximum force fell below 1000 kJ/mol/nm using the Verlet cutoff scheme. Treatment of long-range electrostatic interactions was set to Particle Mesh-Ewald (PME)[29], and the short-range electrostatic and van der Waals cutoff set to 1.0 nm.

After the energy minimisation, heating from 0 to 300 K was performed for 20 ps with a time step of 2 fs and position restraints applied to the backbone in an NVT ensemble. The constraint algorithm used was LINCS, which was applied to all bonds and angles in the protein[30]. With the Verlet cut-off scheme and the non-bonded short-range interaction, the cut-off was set to 1.0 nm. Long-range electrostatics were set to PME. The temperature coupling was set between the protein and the non-protein entities by using a Berendsen thermostat, with a time constant of 0.1 ps and the temperature set to reach 300 K with the pressure coupling off. Pressure equilibration was run at 300 K with a Parrinello-Rahman pressure coupling on and set to 1 bar[31] in an NPT ensemble. The equilibration trajectories were set to 10 ns (discarded from the analysis), and the production MD simulations were performed for 3 replicas of 100 ns each (for AhR, HIF-1α, HIF-2α and NCOA1).

For any simulations involving small molecule ligands, the ligands were parametrised using ACPYPE[240]. Partial atomic charges have been calculated using AM1-BCC level of theory[135] using the DockPrep tool in UCSF Chimera[198]. The equilibration runs were the same as for the *apo* systems. The production trajectories for protein-ligand simulations have been set to 100 ns, run in triplicates.

Analysis of the trajectories was performed using GROMACS tools, including root-mean-square deviation (RMSD) to assess overall stability, per-residue root-mean-square fluctuation (RMSF) to assess the local flexibility, solvent-accessible surface areas, dihedral angles, and principal component analysis (PCA). The coordinates were collected every 10 ps. Per-residue root-mean-square fluctuations (RMSFs) were calculated to assess the flexibility per region.

6.4.4 Molecular docking and druggable "hot spot" mapping

The mapping of potential druggable "hot spots"[35] was performed using FTMap[36] and Cryptosite[37] for a more reliable mapping.

Molecular docking was performed using the University of California, San Francisco DOCK 6.8 suite[39] with default grid scoring. For each system, the highest-populated cluster has been selected from the MD trajectory and assigned as a receptor. The grid spacing was 0.25 Å, and the grid included 12 Å beyond the geometric centre of the PAS-B inner cavity. The energy score was the sum of electrostatic and van der Waals contributions. The docking has been run for known HIF-2α small molecule binders, where the ligand has been re-docked to the protein, to check whether the docking procedure reproduced the native binding modes, as seen in the crystal structures. After the positive verification, the known ligands reported in ChEMBL have been docked to NCOA1 PAS-B sites detected by FTMap.

During the docking calculations, the ligands were subjected to 5,000 cycles of molecular mechanics energy minimisation. The number of maximum ligand orientations was 50,000. Solvent effects were modelled by implicit solvation (distance-dependent dielectric function). The best-scoring poses (ligand-protein complexes) were subjected to all-atom MD simulations, and the energetics has been assessed by PCA analysis, MMPBSA and umbrella sampling.

For the MMPBSA end-state interaction energy, the calculation was divided into three stages. First, the interaction energy term was calculated using the parameters used in the force field for both ligand and receptor. For the polar solvation term, the solute dielectric constant was set to 2, with a solvent dielectric constant as 80. The solvent probe radius was set to 1.4, with a density of grid points per A^2 of 10, using a linear Poisson-Boltzmann equation. For the apolar solvation term, we used a SASA model with a surface tension of 0.0226 kJ/molA^2 and an offset constant of 3.849 kJ/mol. All calculations were performed at a temperature of 300 K. These calculations were made using the Gromacs g_mmpbsa[43] module.

6.4.5 Umbrella sampling of the NCOA1 five best-scoring molecules

Umbrella sampling simulations were performed for best-scoring ligands binding to NCOA1 PAS-B domain to calculate the binding modes and their respective energetics more accurately. From the average configuration obtained during the three 100 ns simulation of equilibrium ligand-NCOA1 complex, an umbrella sampling run was made with a harmonic force applied over the ligand, pulling in the Z direction, as shown in Figure 49, increasing the distance from the binding site. The force constant was set as 1000 kJ mol^{-1} nm^{-2} with a pulling rate of 0.01 nm ps^{-1}, resulting in 30 windows per complex. The systems were parametrised with AMBER99SB-ILDN using the TIP3P solvation model, given it is a well-established set for free-energy calculations. Each umbrella window had restrained pressure equilibration of 10 ps and a 5 ns production length. The errors were calculated using a bootstrap methodology implemented in Gromacs.

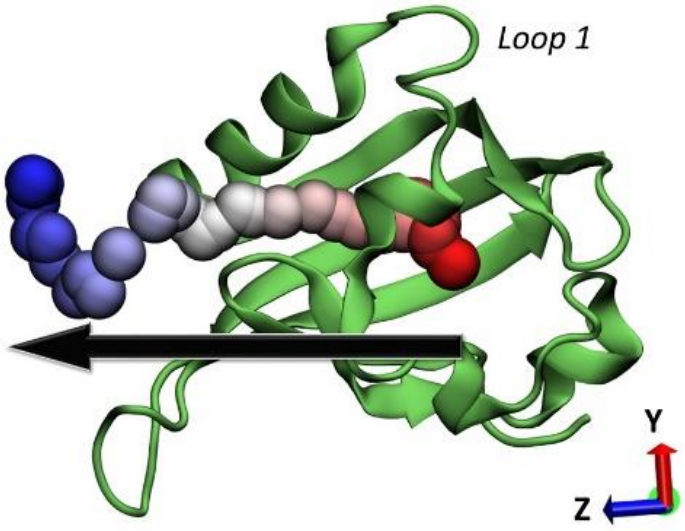

Figure 49: The black arrow indicates the direction the umbrella pathway took to calculate the affinity for NCOA1 ligands. The spheres represent the position of the centre of mass of the ligand throughout the pull simulation.

6.5 Results

6.5.1 UGDH ID-tail entropy affects the structured core configuration

The UGDH monomer was simulated in its truncated configuration (PDB 5W4X) and with the 30 residues modelled ID-tail to study the entropic effect of that tail. To ensure that the C-terminal ID-tail was properly sampled, configurational entropy was calculated from the MD trajectories obtained using seven different combinations of force-fields/water models listed in the Methods section and Table 10

Table 10: Entropy values calculated for the UGDH monomer

$T(S_{Truncated}$-$S_{Tail})$ (kcal/mol)	*SQuE*	*Quasi-harmonic*
AMBER03ws+CAIPi3P	0.5	-0.4
AMBER03ws+TIP4P/2005	1.1	- - - - - - - -
AMBER99SB-ILDN+CAIPi3P	1.1	-0.1
AMBER99SB-ILDN+TIP3P	-2.6	- - - - - - - -
Hybrid_FF+CAIPi3P	2.5	-0.14
Hybrid_FF+TIP3P	-1.7	0.08
Hybrid_FF+TIP4P/2005	0.1	-0.05

In the study by Keul and coworkers[218], the calculated "free energy" difference between a fully flexible and a rigidified C-terminal tail was 2.5 kcal/mol. As showed in Table 9, Hybrid_FF framework combined with CAIPi3P water model attained the same value. The quasiharmonic method was not able to calculate values for two of the seven combinations of the protein force field and water model (AMBER99SB-ILDN+TIP3P and AMBER03ws+TIP4P/2005), and for the others, the acquired values were very different from the free energy differences reported by Keul *et. al*. The failure on QH to calculate values resulted from a convergence problem for the said systems.

The residues within the globular core (residue 1 – 464) of UGDH changed their intrinsic flexibility depending on the presence of the C-terminal ID-tail in the structure. Figure 50 shows the residual root-mean-square fluctuation (RMSF) and the sequential summation of the RMSF curve for the structured core. Most of the residues were more flexible in the configuration without the ID-tail in comparison with the structure with that tail. This shows entropy-entropy compensation between the globular domain and the ID-tail. The entropy generated by the ID-tail compensated for the lower flexibility of the folded protein core, retaining them in a configuration closer to that observed in the crystal structure (Figure 50). The overall structured core residual flexibility is reduced on the system with the ID-tail, with a cumulative difference of 6.6 nm. Each system attained different final configurations, which depended on the presence of the ID-tail, which compensated for the loss of flexibility by the structured core.

.

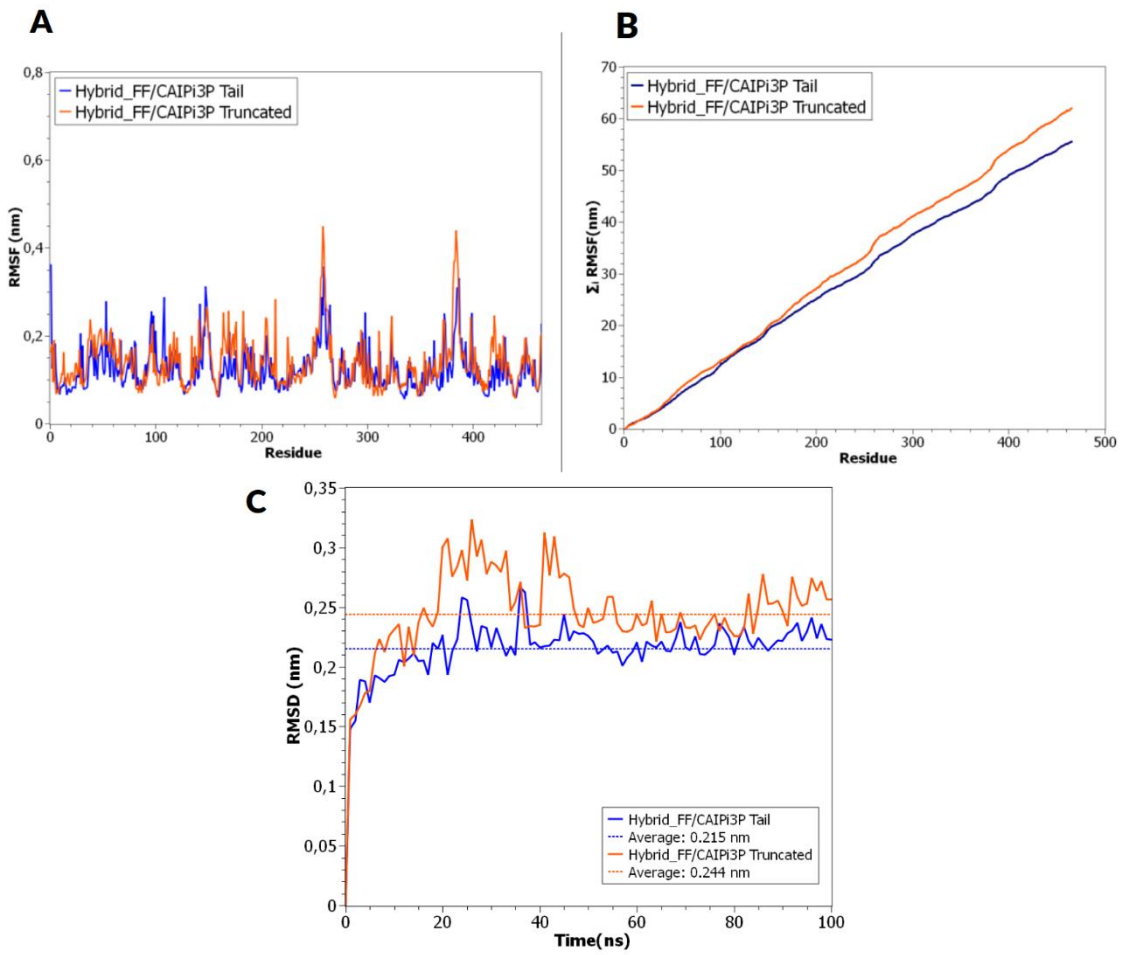

Figure 50: Analysis of intrinsic dynamics of the UGDH monomer using Hybrid_FF+CAIPi3P. A) RMSF per residue for the structured globular domain (residues 1 – 464). B) RMSF cumulative integral for the structured core domain C) RMSD curves for both structured core domains

6.5.2 The ID-tail directly affects the UGDH monomer allosteric switch

When calculated the essential dynamics through principal component analysis (PCA), it can be observed that the truncation of the ID-tail directly affects the allosteric α6 switch. As shown in Figure 51 A, the highlighted helix shifts inwards, burying deep into the nearby cavity. When the tail is truncated (Figure 51 B), the principal motion related to the switch is the outward movement, reducing the interactions between the helix and the remaining core.

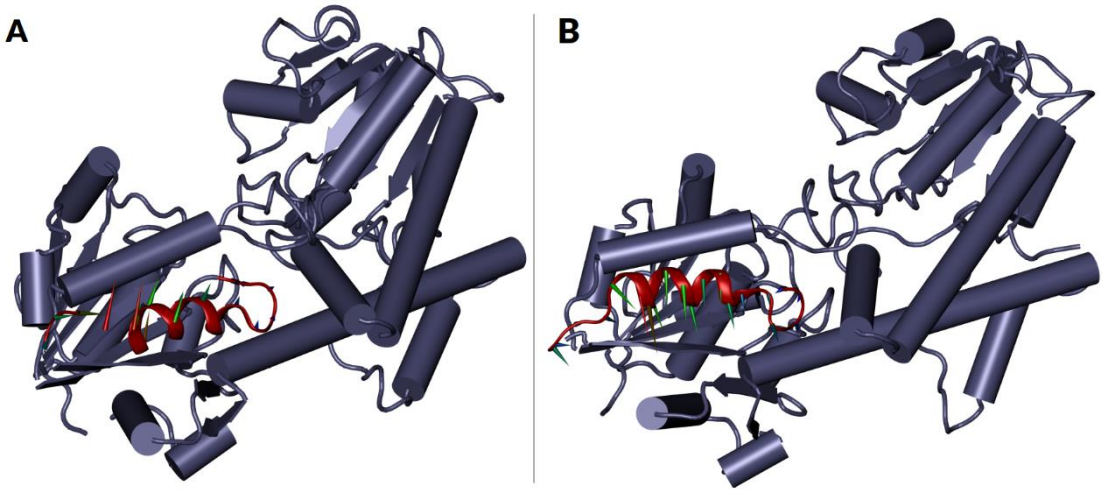

Figure 51: Porcupine plot of core domain for A) UGDH monomer containing the ID-tail B) Truncated UGDH model. The allosteric α6 switch is coloured red. The ID-tail affects the essential dynamics of the allosteric switch, which modulates the binding affinity to the allosteric ligand.

The entropic force that emerges from the ID-tail affects the correlated motions of the UGDH monomer. The intramolecular atomic covariance calculated from the MD simulation of the full-length UGDH containing the ID-tail was generally lower than then its respective calculations for the truncated UGDH configuration. Counter-intuitively, the existence of the flexible ID-tail reduced the overall flexibility of the protein, and it reduced the correlation between the residue motions. This is evidence for the long-range entropic "quake" that the ID-tail generates throughout the protein structure, which is consistent with observations by Keul and coworkers.

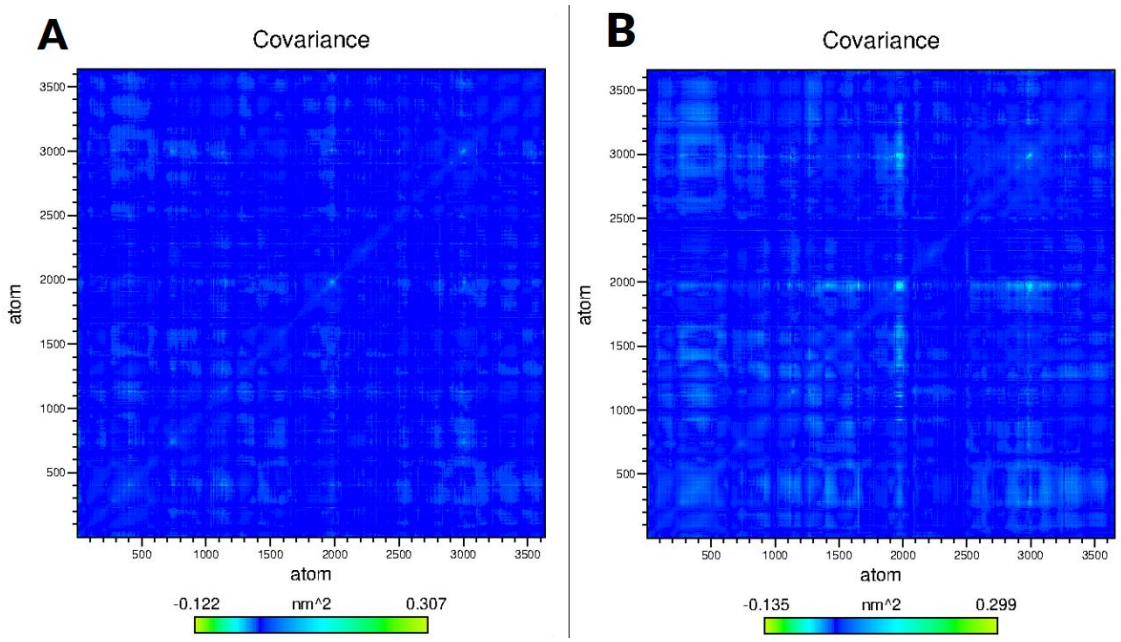

Figure 52: Covariance matrix for UGDH monomer A) ID-tail B) Truncated model. The ID-tail reduces the overall fluctuation of the core residues, reducing the intrachain atomic correlation for the structured residues.

Remarkably, in both configurations, the region corresponding to the α6 switch (Figure 52, atoms 2000-2300) had low intramolecular covariance. Therefore, the presence of the ID-tail affected the dynamics of the α6 helix, but the motions within the structured core did not correlate directly to the α6 switch moving upwards or downwards.

6.5.3 Conformational transitions within PAS-B domains

Since intrinsic dynamics are very likely to play an essential role in tuning the protein activity, as shown in the previous section. Now, the focus was on the most flexible part of the PAS domain, namely the loop1. This loop adopts extended and partially unfolded conformation in all PAS-B domains but nuclear receptor coactivators (which includes NCOA1), in which it adopts an extended α-helical conformation. Both experimental structures of NCOA1 domain reported to date solved PAS-B domain it as a complex with the STAT6 peptide, bearing LXXLL sequence motif, which defines a conserved nuclear receptor (NR) box[241]. This motif is a short α-helix, so it is plausible that the helical conformation of NCOA1 loop1 is induced by protein-protein interactions, and removing the interactor (herein – STAT6 peptide) may induce significant conformational changes in the

loop1 region, shifting the ensemble towards a more unfolded state, similar to that observed in other PAS-B domains.

To address this, an alternative conformation of NCOA1 PAS-B domain was modelled, with the loop1 adopting a partially disordered helix-loop-helix conformation, as showed in Figure 53.

The simulation results strongly indicated that NCOA1 may adopt more extended conformation of the loop1, typical for other PAS-B domains. Although the helical conformation was more stable than the extended loop, the energy difference was minor (0.3 kcal/mol, Figure 53). This suggested potential conformational changes and the existence of NCOA1 conformations that are “druggable” by small molecules.

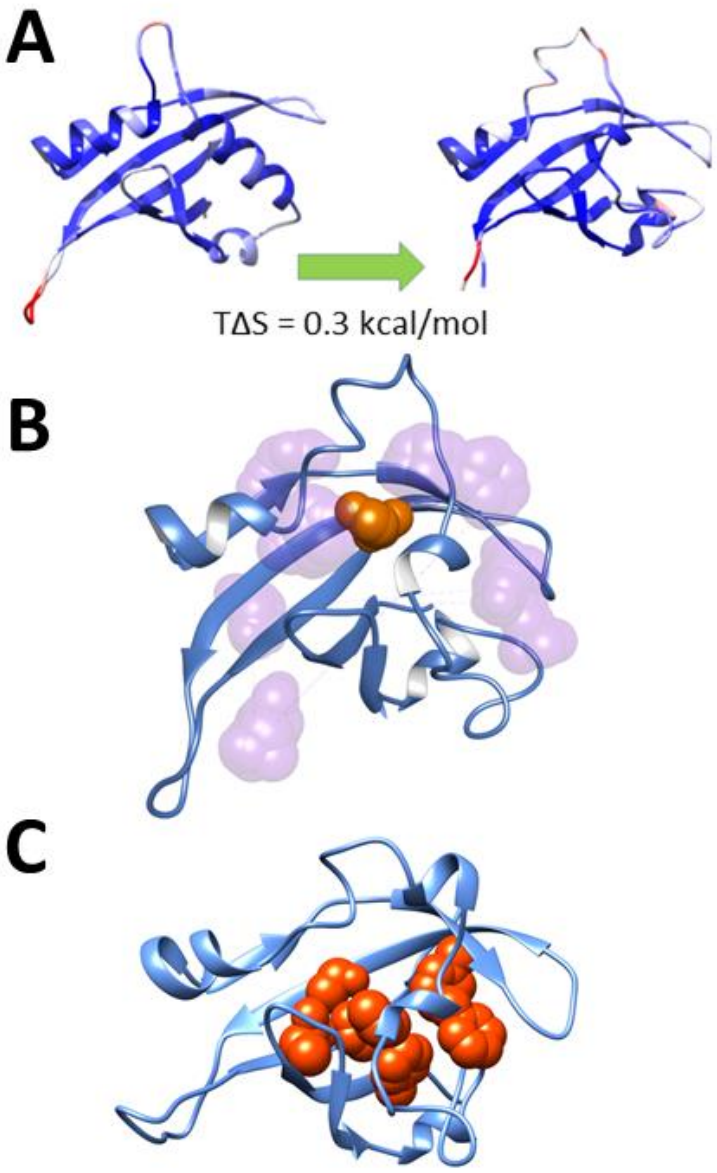

Figure 53: **A** Conformational change within the loop1 of NCOA1 PAS-B domain. Time-averaged structures are shown, indicating loop1 in the α-helix conformation (left panel) and the partially disordered conformation (right panel). The backbone is coloured by calculated B-factors: the most flexible parts are coloured red, while the less flexible is coloured navy blue. **B** FTMap scan of the NCOA1 PAS-B domain with the loop1 in partially disordered conformation. A "hotspot", detected in the centre of the PAS-B cavity, is rendered as spheres and coloured orange; the transparent spheres are the secondary binding areas. **C** Cryptosite analysis of the NCOA1 PAS-B domain with the loop1 in partially disordered conformation. The highest-scoring "hotspot", detected in the centre of the PAS-B cavity, is rendered as spheres and coloured orange.

To date, several small molecule NCOA1 ligands have been reported (Figure 54), but their binding sites remained unknown. Compound L1 was selected, deposited in ChEMBL, reported to inhibit human NCOA1 in a cell-based assay. The hotspot mapping, followed by molecular docking calculations and MD simulation, showed

that L1 might bind to the PAS-B domain of NCOA1 once loop1 adopts an extended conformation (Figure 53). The binding event induced further conformational changes in the domain, preventing the binding to LXXLL peptide of STAT6 (Figure 56B). Four other confirmed small molecule NCOA1 ligands with related structures (Figure 54) were also docked, their binding mechanics were evaluated, and the energy scores correlated with the experimentally determined IC_{50} values (Table 11).

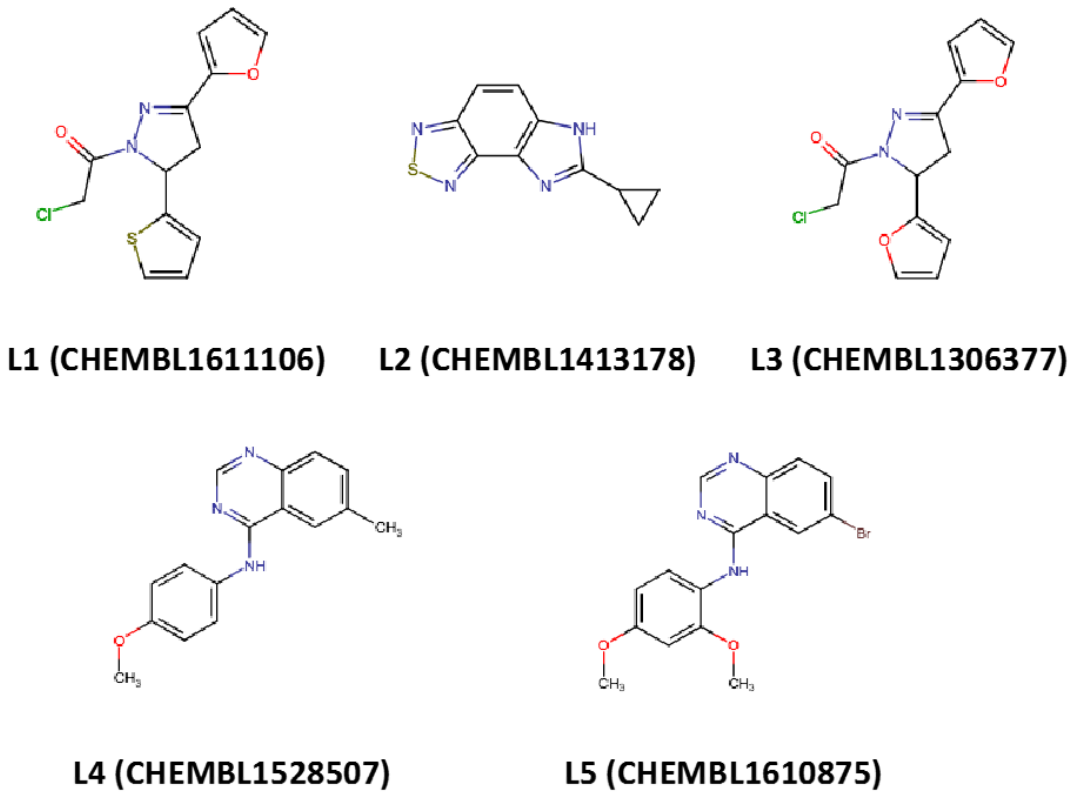

Figure 54: Chemical structures of five confirmed ligands of human NCOA1, used in this study.

The binding modes predicted for ligands L1-L5 may explain the observed differences in binding energetics. L1 and L3 are close analogues, differing only by the heteroatom in the five-membered ring, yet their Interaction energetics are markedly different: replacing the thiophene of L1 by furan has decreased the binding affinity from mid-nM range to 5.6 µM (Table 10). This decrease can be explained by in the desolvation of the oxygen atom O1, and increased desolvation

of the oxygen O2; contributing unfavourably to the binding affinity (Table 11**).** Likewise, compounds L4 and L5 are closely related analogues, and the preference of L5 (IC_{50} = 2.4 µM) over L4 (IC_{50} = 4.0 µM) can be explained by the favourable contribution of the second methoxy group in L5 (atom O2 and C2, Figure 56. D).

Table 11: Binding data for compounds L1-L5

	IC_{50} [nM]	**Estimated K_i Range****	**Calc. ΔG [kcal/mol]#**	**Calc. ΔH [kcal/mol]$**
L1	443	Mid/high nm	-12±2	-7±2
L2	1328	High nM to low µM	-14±2	-6±2
L3	5695	Mid µM	-4±2	-4±2
L4	3973	Low/mid µM	-11±2	-5±2
L5	2442	Low/mid µM	-13±2	-6±2
Pearson correlation between IC_{50} and calculated energy			0.83	0.97

***CHEMBL1201862 PubChem Bioassay data set**

**** Calculated by SeeSAR**

#Calculated by umbrella sampling MD simulation

$Calculated by MM-PBSA

The free energies calculated by umbrella sampling simulations correlated with the experimental IC_{50} values, showing favourable binding energies. L3 showed, in both umbrella simulations and MMPBSA calculations, the lowest affinity. This was particularly encouraging, considering L3 and L1 structural similarity. A reason for these differences may arise from the protein interaction after 0.5 - 0.7 nm, as shown in Figure 55. This region is the start of the energy barrier for the unbound PMF, on which L1 has the lowest barrier, in comparison to the other ligands, which may explain the favourable IC_{50}. This, coupled with the desolvation effects mentioned earlier, may explain the L3-L1 difference in terms of their binding affinity. It was not possible to achieve statistical significance to assess the affinity difference between L1, L2, L4 and L5 in respect to their binding free energies.

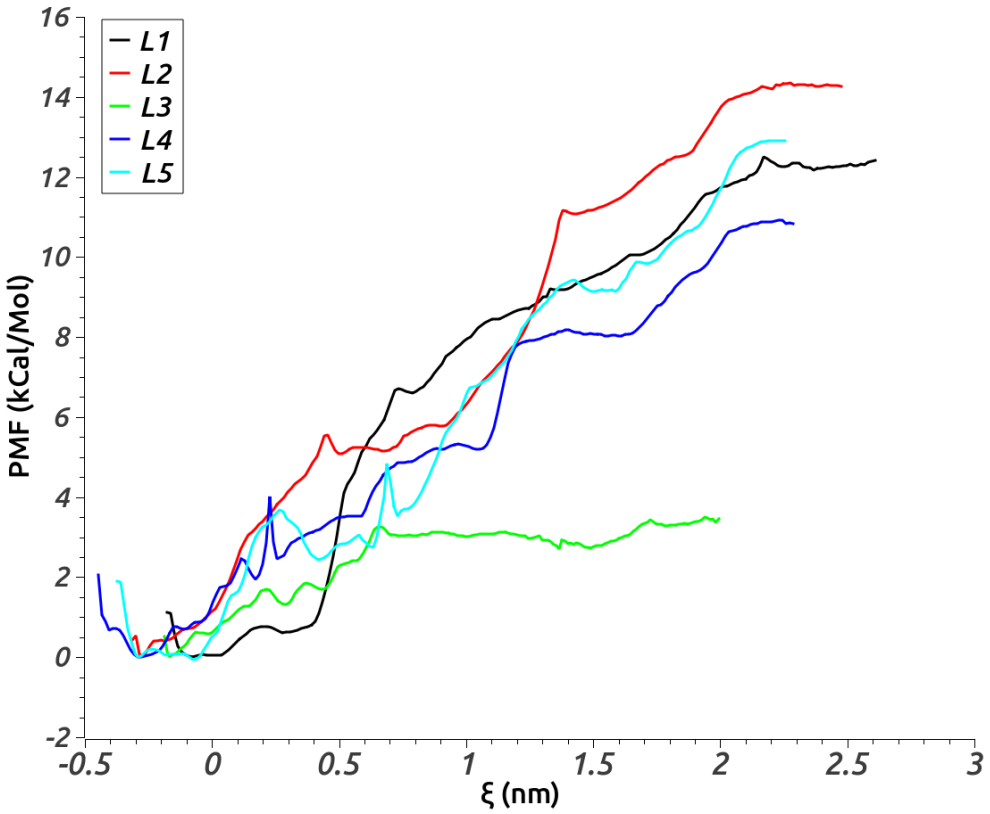

Figure 55: Potential of mean forces calculated by umbrella simulations for different ligands for NCOA1 PASB in the loop-helix-loop canonical conformation. L3 shows shallow binding energy in comparison to the other four ligands.

Given the plateau achieved by L3 (Figure 55, green), it could be concluded that a significant part of the binding free energy arose from intrinsic conformational effects of the protein. This was supported further by the protein RMSD calculations. For the L1, the RMSD of the ligand-protein complex converged to a 0.25 nm and the RMSD of the loop1 converged to 0.3 nm, while for the L3 complex RMSD converged to 0.3 nm and the loop1 RMSD converged to 0.35 nm.

The difference in their respective final configurations resulted from two effects: the thiophene moiety from the L1 bound better at the biding site (as shown by the MMPBSA), allowing loop1 to acquire a more stable conformation. This configuration was achieved given the thiophene moiety interaction with the loop1 being unfavourable, which was compensated by the favourable protein intra-side chain interactions.

Analysis of the pathway differences between L4 and L5 showed that the additional methoxyl group of L5 affected how the overall ligand was solvated in the unbound state, which can explain more favourable IC_{50}.

All five ligands showed favourable values of binding energy to the unstructured loop1 conformation by all three methods for interaction energy estimating: SeeSAR/HYDE, MMPBSA, and umbrella sampling. This indicates that NCOA1 is “druggable” by small molecules upon a configurational transition since the favourable binding site is found only in the canonical loop-helix-loop conformation. The structural entropy for the NCOA1 loop1 can be easily compensated by the binding energetics. This is shown in Table 10, as the entropic barrier is significantly lower (0.3 kcal/mol) in comparison to the binding energetics. Therefore, the results described in this section indicate that the loop1 of NCOA1 may adopt the “canonical” (disordered helix-loop-helix) conformation, which may be targeted by small molecules.

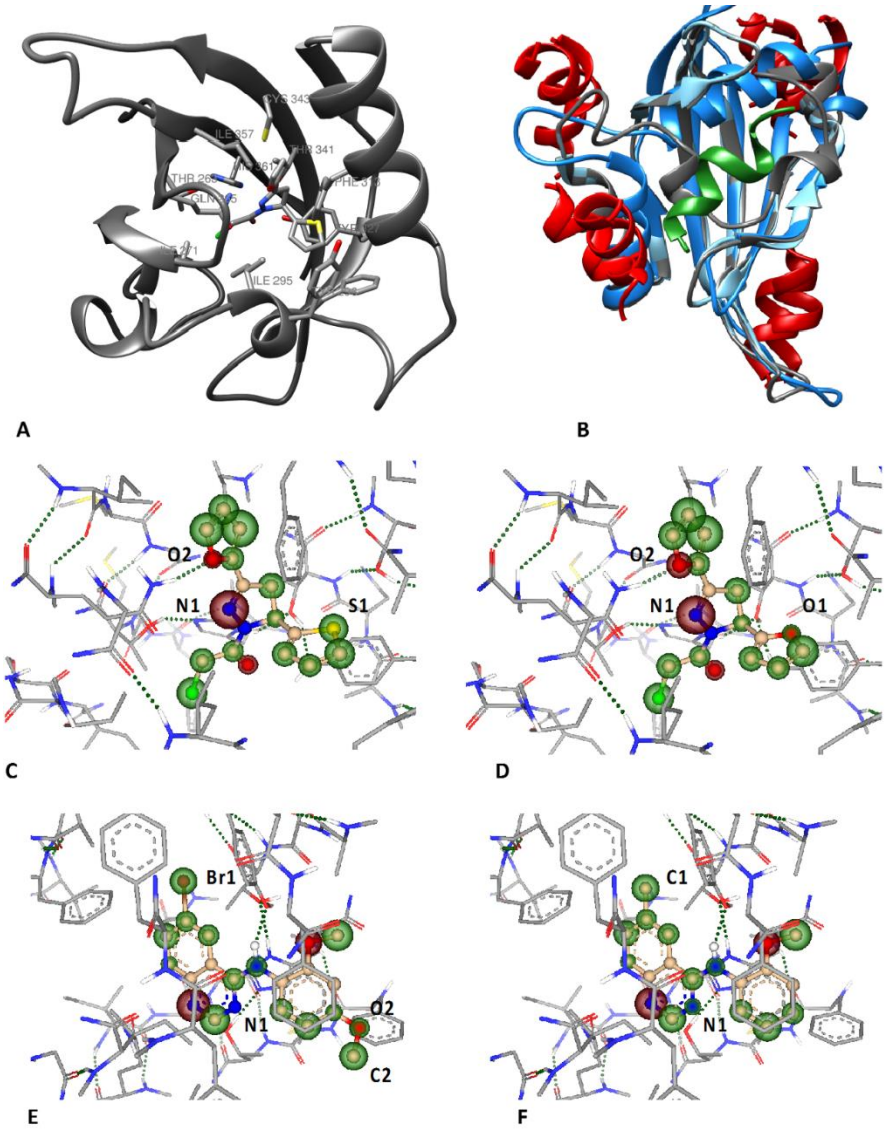

Figure 56: Binding poses for NCOA1. A) The binding poses for the compound L1, docked to the NCOA1 PAS-B domain. Protein is coloured grey, ligand is coloured by heteroatom. Residues critical for the binding, are showed and labelled. B) Docking of the LXXLL motif of STAT6 (red) to the PAS-B domain of NCOA1 with the loop1 in the “druggable” partially disordered conformation (grey), compared to the docking of the same motif (green) to the PAS-B domain of NCOA1 with the loop1 in the alpha-helical conformation (blue). The experimental structure is showed as light blue. C) Predicted binding mode for the compound L1, with per-atoms contributions of the ligand are mapped as the coronas, calculated by SeeSAR: green signifies favourable contribution, red signifies unfavourable contribution, and the magnitude of the contribution is proportional to the size of the corona. D) Predicted binding mode for the compound L2. E) Predicted binding mode for the compound L5. F) Predicted binding mode for the compound L5.

These findings are very encouraging for structure-based drug discovery efforts targeting steroid receptor coactivators for therapeutic interventions. PAS-B domain of NCOA1 has been shown to interact with several proteins linked to

cancer, including the transactivation domain of STAT6[242]. As such, disruption of STAT6-NCOA1 complex by targeting the PAS-B domain of NCOA1 could be an attractive therapeutic strategy. Also, NR boxes of NCOA1 and other coactivators interact with the PAS-B domain of NCOA1, promoting homo- and heterodimerisation of NCOA1 and this interaction may also be exploited therapeutically[243].

To assess the generality of the conformational behaviour observed for NCOA1, the same approach was taken as to other PAS-B domains: the HIF-1α, HIF-2α, and AhR; inducing α-helical conformation of the loop1, resembling that of NCOA1. The results, shown in Table 12 and Figure 53, indicated that the extended conformation of the loop1 was more entropically favourable for all three proteins. Therefore, the conformational transitions within these regions, in order to occur, must be driven by powerful enthalpic contribution (e.g. arising from protein-protein interactions). While this cannot be ruled out completely, particularly for AhR, such transitions are less likely for unbound PAS class I members, such as HIF-2α than for p160 transcription factors, such as NCOA1 (Table 12).

Table 12: Entropy values between unstructured to structured conformations, calculated by SQuE

Protein	**TΔS [kcal/mol]**
AhR	7.6
HIF-1α	11.3
HIF-2α	11.9
NCOA1	0.3

6.6 Discussion

In this chapter, the development and validation of SQuE methodology were described. An approximate calculation for entropy was devised using a simple decomposition of probabilities. A highly encouraging aspect of the SQuE approach is how accurate, robust and straightforward it is to both calculate and decompose for different regions and principal components. For these reasons, it

represents a powerful tool to quantify long-range entropic effects in macromolecular ensembles.

I showed that when using the Hybrid_FF combined with CAIPi3P solvation, the quantification of the configurational entropy for UGDH agrees with published data. Using SQuE, it was possible to show entropy-entropy compensation in UGDH and to demonstrate how the modulation of the allosteric switch functions at the atomistic level, when the C-terminal ID-tail of UGDH is truncated, opening a new area of understanding on how this molecule works.

As a prediction test of structural entropy barriers, SQuE was then applied to four different PAS-B domains: nuclear receptor coactivator 1 (NCOA1), hypoxia-inducible factors HIF-α1 and HIF- α2, and aryl hydrocarbon receptor (AhR). These PAS-B domains were simulated to quantify the effect of protein conformational changes and configurational entropy on their energy landscape. NCOA1 PAS-B, in particular, was the subject of a more in-depth study, given that its entropic barrier between the canonical loop-helix-loop and the full extended α-helical conformations showed the lowest value. Interaction studies between five reported NCOA1 PAS-B small molecule binders showed that the enthalpic gain arising from the binding event overcomes the free energy barrier.

As an upper limit approximation for smaller systems, SQuE predicts configurational entropy values within reasonable accuracy. The usage of principal components as the conjoined variables should minimise the problem resulting from SQuE using the first-degree approximation, but it may be an inaccurate approximation for more complex systems, like systems that require significant molecular motion to do they function, such as multi-domain complexes. One way to overcome this issue would be to implement a pairwise correlation correction, similar to the approach reported by Kilian *et. al.*[227,244]. This could be expanded to higher degrees of correlation, but it would inevitably increase the computational cost.

Chapter 7 – Conclusions

Throughout this work, a series of methods to improve the description of the dynamics of intrinsically disordered macromolecules by all-atom molecular dynamics simulations was developed and tested. Given the challenges that intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) face in experimental and computational studies alike, it is very rewarding and encouraging to see how simple corrections in the force field parameters and approximations used in thermodynamic parameters caused vast improvement in the sampled conformations.

The application of CAIPi3P solvation model to studies of intrinsically disordered systems resulted in a more realistic description of molecular motions in the system. CAIPi3P is based on three-point TIP3P framework and has dipole moment adjusted, which should favour solute-solvent polar interactions. Typically, when a generalist force field such as AMBER99SB-ILDN combined with popular TIP3P water model is being used, the protein collapses on itself and adopts a molten-globule state, which may not be realistic. CAIPi3P model prevents this collapse and helps the simulation to sample more realistic, extended configurations, regardless of the protein force field being used. For the histatin5 simulations, a clear improvement can be seen when CAIPi3P. This is expected, however, given the fact that histatin5 is was the parametrization goal used for the development of CAIPI3P. For the R/S peptide, a substantial improvement arises when AMBER03ws is used with CAIPi3P, showing the accuracy of CAIPi3P for IDPS with a higher concentration of polar/charged residues per area. Finally, the results for an Atg2 showed that CAIPi3P also improves the sampled states for the partially structured proteins, especially when coupled with AMBER99SB-ILDN. Nonetheless, CAIPi3P solvation model has a room for improvement: the parameters have several shortcomings regarding the bulk water properties, such as high density and low radial oxygen distance.

Hybrid_FF was developed as a force field framework to be used combined with CAIPi3P, with a goal to improve the sampling of disordered regions in simulations of multi-domain proteins, having globular domains connected by intrinsically disordered regions. For the test cases simulated, Hybrid_FF+CAIPi3P

combination showed an improvement for both sampled configurations and predicted radii of gyration, especially for proteins with a longer IDR terminus. These improvements have not been systematic throughout the test sets, but it is encouraging, nonetheless. Further tests are needed to ensure that desolvation energies and overall protein thermodynamics are accurate in comparison to the experimental values. More extended simulations should be run (on the microseconds to tens of microseconds) to assure structural cohesion. Another important point of consideration is the definition of structure-loop per residue: in this work, we used DSSP, a widely used algorithm to defines the secondary structure of each residue. However, the main parameters used for definition are hydrogen bonds with neighbours and backbone dihedrals. This might force loops as transition residues, assigning the wrong force-field. Also, the framework needs to be systematically tested on more structurally diverse classes of proteins – such a systematic study was beyond the scope of this dissertation.

Structural Quantifier of Entropy (SQuE) was developed to analyse the data generated by IDPs/IDRs simulations and extract information on configurational entropy. An interesting test case was the analysis of the UGDH enzyme, and the entropic force exerted by its intrinsically disordered C-terminus (ID-tail). SQuE correctly predicted, calculated and evaluated the configurational entropy changes showed how the ID-tail regulated the protein function via entropy-entropy compensation. As a case study, the helix-to-loop conformational changes were evaluated in several mammalian PAS-B domains, including NCOA1, and it was showed how these conformational changes affect the “druggability” of the protein. Coupled with docking and free energy techniques, SQuE showed that the helix-to-loop transition may happen only in specific PAS-B domains, namely NCOA1, and not in other proteins that contain this domain, such as HIF-1a transcription factor. SQuE uses a simple approximation to calculate the upper estimates of entropy, hence, should be used with caution: without the proper basis to define the dynamics and with a non-equilibrated system, it may yield inaccurate results.

To conclude, tools developed in this work are a useful toolset for molecular discoveries for IDPs/IDRs which can be transplanted to different areas of computational chemistry, biology and biochemistry and maybe even extended for different classes of proteins. We do expect that these tools are likely to pave a

way to improvements in robust and accurate molecular simulations of challenging IDPs/IDRs.

References

1. Lehninger, A. L., Nelson, D. L. & Cox, M. M. *Lehninger principles of biochemistry*. (W.H. Freeman, 2013).
2. Amino Acids | BioNinja. Available at: https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/amino-acids.html. (Accessed: 4th January 2020)
3. Liljas, A. *Textbook of structural biology*. (World Scientific, 2009).
4. Voet, D. & Voet, J. G. *Biochemistry 4e*. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki* (2010).
5. Rupp, B. *Biomolecular crystallography: principles, practice, and application to structural biology*. (Garland Science, 2010).
6. Poole, L. B. The basics of thiols and cysteines in redox biology and chemistry. *Free Radical Biology and Medicine* **80**, 148–157 (2015).
7. Alcock, L. J., Perkins, M. V. & Chalker, J. M. Chemical methods for mapping cysteine oxidation. *Chemical Society Reviews* **47**, 231–268 (2018).
8. Rajpal, G. & Arvan, P. Disulfide Bond Formation. in *Handbook of Biologically Active Peptides* 1721–1729 (Elsevier Inc., 2013). doi:10.1016/B978-0-12-385095-9.00236-0

9. Wedemeyer, W. J., Welker, E., Narayan, M. & Scheraga, H. A. Disulfide Bonds and Protein Folding †. *Biochemistry* **39**, 4207–4216 (2000).

10. Qin, M., Wang, W. & Thirumalai, D. Protein folding guides disulfide bond formation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11241–11246 (2015).

11. Berg, J. M., Tymoczko, J. & Stryer, L. *Protein structure , function and evolution Recommended reading material • Understanding of biological processes at*. (2013).

12. Martin, Y. C. Let's not forget tautomers. *Journal of Computer-Aided Molecular Design* **23**, 693–704 (2009).

13. Brink, T. Ten & Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J. Chem. Inf. Model.* **49**, 1535–1546 (2009).

14. Kim, M. O., Nichols, S. E., Wang, Y. & McCammon, J. A. Effects of histidine protonation and rotameric states on virtual screening of M. tuberculosis RmlC. *J. Comput. Aided. Mol. Des.* **27**, 235–246 (2013).

15. Varki, A. C. R. D. . E. J. D. . F. H. H. . S. P. . B. C. R. . H. G. W. . E. M. & E. *Essentials of Glycobiology, 3rd edition*. *Cold Spring Harbor (NY)* (Cold Spring Harbor Laboratory Press, 2015).

16. Pickart, C. M. Mechanisms Underlying Ubiquitination. *Annu. Rev. Biochem.* **70**, 503–533 (2001).

17. Kerscher, O., Felberbaum, R. & Hochstrasser, M. Modification of Proteins by Ubiquitin and Ubiquitin-Like Proteins. *Annu. Rev. Cell Dev. Biol.* **22**, 159–180 (2006).

18. Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **4**, (2002).

19. Chapter 2: Protein Structure – Chemistry. Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-2-protein-structure/. (Accessed: 9th April 2020)

20. Alberts, B. *Essential cell biology*. (Garland Science, 2013).

21. Perutz, M. F. *et al.* Structure of Hæmoglobin: A three-dimensional fourier synthesis at 5.5-. resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).

22. Kendrew, J. C. *et al.* Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. *Nature* **185**, 422–427 (1960).

23. Stanger, H. E. *et al.* Length-dependent stability and strand length limits in antiparallel β-sheet secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12015–12020 (2001).

24. Qian, H. & Schellman, J. A. Helix-coil theories: A comparative study for finite length polypeptides. *J. Phys. Chem.* **96**, 3987–3994 (1992).

25. Nesloney, C. L. & Kelly, J. W. Progress towards understanding β-sheet structure. *Bioorganic and Medicinal Chemistry* **4**, 739–766 (1996).

26. Chothia, C. Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* **75**, 295–302 (1973).

27. Richardson, J. S. & Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2754–2759 (2002).

28. Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **6**, 1425–1436 (1968).

29. Dobson, C. M. Principles of protein folding, misfolding and aggregation. in *Seminars in Cell and Developmental Biology* **15**, 3–16 (Elsevier Ltd, 2004).

30. Levinthal, C. Are there pathways for protein folding? *J. Chim. Phys. Physico-Chimie Biol.* **65**, 44–45 (1968).

31. Levinthal, C. How to Fold Graciously. *Mossbauer Spectrosc. Biol. Syst. Proc. a Meet. held Allert. House, Monticello, Illinois* 22–24 (1969).

32. Karplus, M. The Levinthal paradox: Yesterday and today. *Fold. Des.* **2**, (1997).

33. Rooman, M., Dehouck, Y., Kwasigroch, J. M., Biot, C. & Gilis, D. What is paradoxical about levinthal paradox? *J. Biomol. Struct. Dyn.* **20**, 327–329 (2002).

34. Dill, K. A. & Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology* . (Garland Science, 2003).

35. Reif, F. *Fundamentals of statistical and thermal physics*. (Waveland Press, 2009).

36. Jacobs, P. W. M. *Thermodynamics*. (Imperial College Press ; Distributed by World Scientific, 2013).

37. Lawden, D. F. *Principles of thermodynamics and statistical mechanics*. (Dover Publications, 2005).

38. Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 8721–8725 (1992).

39. Scheraga, H. A. *et al.* The Protein Folding Problem. *Lect. Notes Comput. Sci. Eng.* **49**, 90–100 (2006).

40. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).

41. Ellis, R. J. Molecular chaperones: assisting assembly in addition to folding. *Trends in Biochemical Sciences* **31**, 395–401 (2006).

42. Mashaghi, A., Kramer, G., Lamb, D. C., Mayer, M. P. & Tans, S. J. Chaperone Action at the Single-Molecule Level. *Chem. Rev.* **114**, 660–676 (2014).

43. Na, J. H., Lee, W. K. & Yu, Y. G. How do we study the dynamic structure of unstructured proteins: A case study on nopp140 as an example of a large, intrinsically disordered protein. *International Journal of Molecular Sciences* **19**, (2018).

44. Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **83**, 553–584 (2014).

45. Van Der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chemical Reviews* **114**, 6589–6631 (2014).

46. Kang, W., Jiang, F., Wu, Y. D. & Wales, D. J. Multifunnel Energy Landscapes for Phosphorylated Translation Repressor 4E-BP2 and Its Mutants. *J. Chem. Theory Comput.* (2019). doi:10.1021/acs.jctc.9b01042

47. Röder, K., Joseph, J. A., Husic, B. E. & Wales, D. J. Energy Landscapes for Proteins: From Single Funnels to Multifunctional Systems. *Adv. Theory Simulations* **2**, 1800175 (2019).

48. Chebaro, Y., Ballard, A. J., Chakraborty, D. & Wales, D. J. Intrinsically disordered energy landscapes. *Sci. Rep.* **5**, (2015).

49. Granata, D. *et al.* The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments. *Sci. Rep.* **5**, 15449 (2015).

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

51. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002).

52. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**, 18–29 (2015).

53. 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 Journal* **272**, 5129–5148 (2005).

54. Kim, P. M., Sboner, A., Xia, Y. & Gerstein, M. The role of disorder in interaction networks: A structural analysis. *Mol. Syst. Biol.* **4**, 179 (2008).

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

56. Tompa, P. & Fersht, A. *Structure and function of intrinsically disordered proteins*. (2009).

57. Tompa, P. & Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB Journal* **18**, 1169–1175 (2004).

58. Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L. & Rexach, M. Disorder in the nuclear pore complex: The FG repeat regions of nucleoporins are natively unfolded. *Proc. Natl. Acad. Sci. U. S. A.* **100**,

2450–2455 (2003).

59. Trombitás, K. *et al.* Titin extensibility in situ: Entropic elasticity of permanently folded and permanently unfolded molecular segments. *J. Cell Biol.* **140**, 853–859 (1998).

60. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049 (2004).

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

62. Breidenbach, M. A. & Brunger, A. T. Substrate recognition strategy for butulinum neurotoxin serotype A. *Nature* **432**, 925–929 (2004).

63. Shi, Y. Structural basis of Smad2 recognition by the smad anchor for receptor activation. *Science (80-.).* **287**, 92–97 (2000).

64. Sigalov, A. B., Zhuravleva, A. V & Orekhov, V. Y. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie* **89**, 419–21 (2007).

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

66. Chaturvedi, S. K., Siddiqi, M. K., Alam, P. & Khan, R. H. Protein misfolding and aggregation: Mechanism, factors and detection. *Process Biochem.* **51**, 1183–1192 (2016).

67. Uversky, V. Amyloidogenesis of Natively Unfolded Proteins. *Curr. Alzheimer Res.* **5**, 260–287 (2008).

68. Uversky, V. N. & Fink, A. L. Conformational constraints for amyloid fibrillation: the importance of being unfolded. (2004). doi:10.1016/j.bbapap.2003.12.008

69. Groh, N. *et al.* Age-dependent protein aggregation initiates amyloid-β aggregation. *Front. Aging Neurosci.* **9**, (2017).

70. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. (2008). doi:10.1146/annurev.biophys.37.032807.125924

71. Sunde, M. *et al.* Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **273**, 729–739 (1997).

72. Breydo, L. & Uversky, V. N. Molecular Mechanisms of Protein Misfolding. in *Bio-nanoimaging: Protein Misfolding and Aggregation* 1–14 (Elsevier Inc., 2013). doi:10.1016/B978-0-12-394431-3.00001-8

73. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition* **18**, 343–384 (2005).

74. Vousden, K. H. & Lu, X. Live or let die: The cell's response to p53. *Nature Reviews Cancer* **2**, 594–604 (2002).

75. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science (80-.).* **253**, 49–53 (1991).

76. Anil, B., Riedinger, C., Endicott, J. A. & Noble, M. E. The structure of an MDM2-Nutlin-3a complex solved by the use of a validated MDM2 surface-entropy reduction mutant. *Acta Crystallogr.,Sect.D* **69**, 1358–1366 (2013).

77. Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations. *Science (80-.).* **265**, 346–355 (1994).

78. Uhart, M. & Bustos, D. M. Protein intrinsic disorder and network connectivity. The case of 14-3-3 proteins. *Front. Genet.* **5**, (2014).

79. Simmons, L. K. *et al.* Secondary structure of amyloid β peptide correlates with neurotoxic activity in vitro. *Mol. Pharmacol.* **45**, 373–379 (1994).

80. Kirkitadze, M. D., Condron, M. M. & Teplow, D. B. Identification and characterization of key kinetic intermediates in amyloid β-protein fibrillogenesis. *J. Mol. Biol.* **312**, 1103–1119 (2001).

81. Crescenzi, O. *et al.* Solution structure of the Alzheimer amyloid β-peptide (1-42) in an apolar microenvironment: Similarity with a virus fusion domain. *Eur. J. Biochem.* **269**, 5642–5648 (2002).

82. Colvin, M. T. *et al.* Atomic Resolution Structure of Monomorphic A beta 42 Amyloid Fibrils. *J.Am.Chem.Soc.* **138**, 9663–9674 (2016).

83. Delacourte, A. & Buée, L. Normal and pathological Tau proteins as factors for microtubule assembly. *International Review of Cytology* **171**, 167–224 (1997).

84. Singh, T. J., Zaidi, T., Grundke-Iqbal, I. & Iqbal, K. Non-proline-dependent protein kinases phosphorylate several sites found in tau from Alzheimer disease brain. *Mol. Cell. Biochem.* **154**, 143–151 (1996).

85. Mart, M. J., Tolosa, E. & Campdelacreu, J. Clinical overview of the synucleinopathies. *Mov. Disord.* **18**, 21–27 (2003).

86. Morar, A. S., Olteanu, A., Young, G. B. & Pielak, G. J. Solvent-induced collapse of α-synuclein and acid-denatured cytochrome c. *Protein Sci.* **10**, 2195–2199 (2008).

87. Ulmer, T. S., Bax, A., Cole, N. B. & Nussbaum, R. L. Structure and Dynamics of Micelle-bound Human ⎵-Synuclein* □ S. (2004). doi:10.1074/jbc.M411805200

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

89. Guerrero-Ferreira, R. *et al.* Cryo-EM structure of alpha-synuclein fibrils. *Elife* **7**, (2018).

90. Stefanis, L. α-Synuclein in Parkinson's disease. *Cold Spring Harb. Perspect. Med.* **2**, (2012).

91. Uversky, V. N. A protein-chameleon: Conformational plasticity of α-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* **21**, 211–234 (2003).

92. Huber, R. Conformational flexibility in protein molecules. *Nature* **280**, 538–539 (1979).

93. DeForte, S. & Uversky, V. N. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. *Protein Sci.* **25**, 676–688 (2016).

94. Tamiola, K. & Mulder, F. A. A. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochemical Society Transactions* **40**, 1014–1020 (2012).

95. Sun, S. *et al.* Solid-state NMR spectroscopy of protein complexes. *Methods Mol. Biol.* **831**, 303–331 (2012).

96. Ladizhansky, V. Applications of solid-state NMR to membrane proteins. *Biochimica et Biophysica Acta - Proteins and Proteomics* **1865**, 1577–1586 (2017).

97. Fitzpatrick, A. W. & Saibil, H. R. Cryo-EM of amyloid fibrils and cellular aggregates. *Current Opinion in Structural Biology* **58**, 34–42 (2019).

98. Piovesan, D. *et al.* DisProt 7.0: A major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).

99. Oates, M. E. *et al.* D2P2: Database of disordered protein predictions. *Nucleic Acids Res.* **41**, (2013).

100. Potenza, E., Di Domenico, T., Walsh, I. & Tosatto, S. C. E. MobiDB 2.0: An improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* **43**, D315–D320 (2015).

101. Fukuchi, S. *et al.* IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* **42**, (2014).

102. Varadi, M. *et al.* PE-DB: A database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* **42**, (2014).

103. Linding, R. *et al.* Protein disorder prediction: Implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).

104. Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 (2003).

105. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta - Proteins Proteomics* **1804**, 996–1010 (2010).

106. Prilusky, J. *et al.* FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438 (2005).

107. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).

108. Comparison of Crystallography, NMR and EM - Creative Biostructure. Available at: https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm. (Accessed: 4th January 2020)

109. Wüthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **8**, 923–925 (2001).

110. De Rosier, D. J. & Klug, A. Reconstruction of three dimensional structures from electron micrographs. *Nature* **217**, 130–134 (1968).

111. Shoemaker, S. C. & Ando, N. X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. *Biochemistry* **57**, 277–285 (2018).

112. Guinier, A. La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys. (Paris).* **11**, 161–237 (1939).

113. Leach, A. R. *Molecular modelling: principles and applications*. (Prentice Hall, 2001).

114. Schlick, T. *Interdisciplinary Applied Mathematics*. **21**, (Springer Science Business Media, 2010).

115. Haile, J. M. *Molecular dynamics simulation: elementary methods*. (Wiley, 1997).

116. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

117. Honig, B. & Nicholls, a. Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149 (1995).

118. Liu, H. Y., Kuntz, I. D. & Zou, X. Q. Pairwise GB/SA scoring function for structure-based drug design. *J. Phys. Chem. B* **108**, 5453–5462 (2004).

119. Gilson, M. K. & Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42 (2007).

120. Qiu, D., Shenkin, P., Hollinger, F. & Still, W. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **101**, 3005–3014 (1997).

121. 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).

122. Ouyang, J. F. & Bettens, R. P. A. Modelling water: A lifetime enigma. *Chimia (Aarau).* **69**, 104–111 (2015).

123. Jorgensen, W. L. & Madura, J. D. Temperature and size dependence for monte carlo simulations of TIP4P water. *Mol. Phys.* **56**, 1381–1392 (1985).

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

125. Jensen, F. *Introduction to computational chemistry*. (John Wiley & Sons, 2007).

126. Press, W. H. *Numerical recipes : the art of scientific computing*. (Cambridge University Press, 1986).

127. Hünenberger, P. H. Thermostat Algorithms for Molecular Dynamics Simulations. in 105–149 (2005). doi:10.1007/b99427

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

129. Parrinello, M. & Rahman, A. Crystal structure and pair potentials: A molecular-dynamics study. *Phys. Rev. Lett.* **45**, 1196–1199 (1980).

130. Hartree, D. R. The Wave Mechanics of an Atom with a Non-Coulomb Central Field Part II Some Results and Discussion. *Math. Proc. Cambridge Philos. Soc.* **24**, 111–132 (1928).

131. Levine, I. N. *Quantum chemistry*.

132. Pople, J. A. *Approximate Molecular Orbital Theory (Advanced Chemistry)*. (Mcgraw-Hill (Tx), 1970).

133. Dewar, M. J. S., Zoebisch, E. G., Healy, E. F. & Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model1. *J. Am. Chem. Soc.* **107**, 3902–3909 (1985).

134. Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **13**, 1173–1213 (2007).

135. Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* (2002). doi:10.1002/jcc.10128

136. Ponder, J. W. & Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **66**, 27–85 (2003).

137. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

138. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666 (1988).

139. Van Gunsteren, W. F. & Berendsen, H. J. C. *The GROMOS Software for (Bio)Molecular Simulation GROMOS87 Groningen Molecular Simulation (GROMOS) Library Manual*.

140. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217

(1983).

141. Uversky, V. N. A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci.* **22**, 693–724 (2013).

142. Kuntz, I. D., Meng, E. C. & Shoichet, B. K. STRUCTURE-BASED MOLECULAR DESIGN. *Acc. Chem. Res.* 117–123 (1994).

143. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA–NA (2009). doi:10.1002/jcc.21334

144. Brozell, S. R. *et al.* Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput. Aided. Mol. Des.* **26**, 749–773 (2012).

145. BioSolveIT - SeeSAR. Available at: https://www.biosolveit.de/SeeSAR/. (Accessed: 6th January 2020)

146. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the Performance of the MM / PBSA and MM / GBSA Methods . I . The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model* **51**, 69–82 (2010).

147. Amber Advanced Tutorials - Tutorial 3 - MM-PBSA - Introduction. Available at: http://ambermd.org/tutorials/advanced/tutorial3/. (Accessed: 6th January 2020)

148. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **0441**, 1–13 (2015).

149. Lindström, A. *et al.* Postprocessing of Docked Protein-Ligand Complexes Using Implicit Solvation Models. *J. Chem. Inf. Model.* **51**, 267–282 (2011).

150. Sun, H. *et al.* Assessing the Performance of MM/PBSA and MM/GBSA Methods. 5. Improved Docking Performance by Using High Solute Dielectric Constant MM/GBSA and MM/PBSA Rescoring. *Phys. Chem. Chem. Phys.* **16**, 22035–22045 (2014).

151. Mezei, M. & Beveridge, D. L. Free energy simulations. *Ann. N. Y. Acad. Sci.* **482**, 1–23 (1986).

152. Doudou, S., Burton, N. a. & Henchman, R. H. Standard free energy of binding from a one-dimensional potential of mean force. *J. Chem. Theory Comput.* **5**, 909–918 (2009).

153. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F. & Hermans, J. *Intermolecular Forces*. (1981).

154. Jorgensen, W. L. Quantum and statistical mechanical studies of liquids. 11. Transferable intermolecular potential functions. Application to liquid methanol including internal rotation. *J. Am. Chem. Soc.* **103**, 341–345 (1981).

155. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in

effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).

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

157. 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).

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

159. 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).

160. Rick, S. W. A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums. *J. Chem. Phys.* **120**, 6085–6093 (2004).

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

162. Fellers, R. S., Leforestier, C., Braly, L. B., Brown, M. C. & Saykally, R. J. Spectroscopic determination of the water pair potential. *Science (80-.).* **284**, 945–948 (1999).

163. Mas, E. M. *et al.* Water pair potential of near spectroscopic accuracy. I. Analysis of potential surface and virial coefficients. *J. Chem. Phys.* **113**, 6687–6701 (2000).

164. and, P. R. & Ponder*, J. W. Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation. (2003). doi:10.1021/JP027815+

165. Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **42**, 147–154 (2017).

166. Baker, C. M. & Best, R. B. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 182–198 (2014).

167. Ueda, Y., Taketomi, H. & G?, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. *Biopolymers* **17**, 1531–1548 (1978).

168. Do, T. N., Choy, W. Y. & Karttunen, M. Accelerating the conformational sampling of intrinsically disordered proteins. *J. Chem. Theory Comput.* **10**, 5081–5094 (2014).

169. Silvestre-Ryan, J., Bertoncini, C. W., Fenwick, R. B., Esteban-Martin, S. & Salvatella, X. Average conformations determined from PRE data provide high-resolution maps of transient tertiary interactions in disordered proteins. *Biophys. J.* **104**, 1740–1751 (2013).

170. Ferrari, A. M., Wei, B. Q., Costantino, L. & Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **47**, 5076–5084 (2004).

171. Wang, W., Ye, W., Jiang, C., Luo, R. & Chen, H.-F. New Force Field on Modeling Intrinsically Disordered Proteins. *Chem. Biol. Drug Des.* **84**, 253–269 (2014).

172. Song, D. *et al.* {iff14IDPs/i force field improving the conformation sampling of intrinsically disordered proteins. *Chem. Biol. Drug Des.* (2016). doi:10.1111/cbdd.12832

173. Best, R. B. & Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* **114**, 14916–14923 (2010).

174. 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).

175. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).

176. 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).

177. Henriques, J., Arleth, L., Lindorff-Larsen, K. & Skepö, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* **430**, 2521–2539 (2018).

178. Smith, M. D., Rao, J. S., Segelken, E. & Cruz, L. Force-Field Induced Bias in the Structure of $A\beta_{21–30}$: A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields. *J. Chem. Inf. Model.* **55**, 2587–2595 (2015).

179. 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).

180. Ye, W., Ji, D., Wang, W., Luo, R. & Chen, H.-F. Test and Evaluation of *ff99IDPs* Force Field for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **55**, 1021–1029 (2015).

181. 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).

182. Liu, H., Song, D., Lu, H., Luo, R. & Chen, H.-F. Intrinsically disordered protein-specific force field CHARMM36IDPSFF. *Chem. Biol. Drug Des.* **92**, 1722–1735 (2018).

183. Bernetti, M. *et al.* Structural and Kinetic Characterization of the Intrinsically Disordered Protein SeV N_{TAIL} through Enhanced Sampling Simulations. *J. Phys. Chem. B* **121**, 9572–9582 (2017).

184. Do, T. N., Choy, W.-Y. & Karttunen, M. Binding of Disordered Peptides to

Kelch: Insights from Enhanced Sampling Simulations. *J. Chem. Theory Comput.* **12**, 395–404 (2016).

185. Han, M., Xu, J. & Ren, Y. Sampling conformational space of intrinsically disordered proteins in explicit solvent: Comparison between well-tempered ensemble approach and solute tempering method. *J. Mol. Graph. Model.* **72**, 136–147 (2017).

186. Duong, V. T., Chen, Z., Thapa, M. T. & Luo, R. Computational Studies of Intrinsically Disordered Proteins. *J. Phys. Chem. B* **122**, 10455–10469 (2018).

187. Cukier, R. I. Generating Intrinsically Disordered Protein Conformational Ensembles from a Database of Ramachandran Space Pair Residue Probabilities Using a Markov Chain. *J. Phys. Chem. B* **122**, 9087–9101 (2018).

188. Salvi, N., Abyzov, A. & Blackledge, M. Multi-Timescale Dynamics in Intrinsically Disordered Proteins from NMR Relaxation and Molecular Simulation. *J. Phys. Chem. Lett.* **7**, 2483–2489 (2016).

189. Papaleo, E., Camilloni, C., Teilum, K., Vendruscolo, M. & Lindorff-Larsen, K. Molecular dynamics ensemble refinement of the heterogeneous native state of NCBD using chemical shifts and NOEs. *PeerJ* **6**, e5125 (2018).

190. 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).

191. Bhattacharya & Lin. Recent Advances in Computational Protocols Addressing Intrinsically Disordered Proteins. *Biomolecules* **9**, 146 (2019).

192. Zerze, G. H., Zheng, W., Best, R. B. & Mittal, J. Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **10**, 2227–2234 (2019).

193. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).

194. Xie, M., Li, D.-W., Yuan, J., Hansen, A. L. & Brüschweiler, R. Quantitative Binding Behavior of Intrinsically Disordered Proteins to Nanoparticle Surfaces at Individual Residue Level. *Chemistry* **24**, 16997–17001 (2018).

195. Mercadante, D., Wagner, J. A., Aramburu, I. V., Lemke, E. A. & Gräter, F. Sampling Long- versus Short-Range Interactions Defines the Ability of Force Fields To Reproduce the Dynamics of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **13**, 3964–3974 (2017).

196. and, P. M. & Nilsson*, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. (2001). doi:10.1021/JP003020W

197. Niu, S., Tan, M.-L. & Ichiye, T. The large quadrupole of water molecules. *J. Chem. Phys.* **134**, 134501 (2011).

198. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for

exploratory research and analysis. *J. Comput. Chem.* (2004). doi:10.1002/jcc.20084

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

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

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

202. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

203. Petoukhov, M. V *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).

204. Svergun, D., Barberato, C. & Koch, M. H. J. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).

205. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building Water Models: A Different Approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).

206. Xiang, S. *et al.* Phosphorylation Drives a Dynamic Switch in Serine/Arginine-Rich Proteins. *Structure* **21**, 2162–2174 (2013).

207. Tyler, R. C., Tonelli, M., Lee, M. & Markley, J. L. NMR Solution Structure of the Partially Disordered Protein At2g23090 from Arabidopsis thaliana. *TO BE Publ.* doi:10.2210/PDB1WVK/PDB

208. Wolin, S. L. & Cedervall, T. The La Protein. *Annu. Rev. Biochem.* **71**, 375–403 (2002).

209. Alfano, C. *et al.* Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. *Nat. Struct. Mol. Biol.* **11**, 323–9 (2004).

210. Martino, L. *et al.* Synergic interplay of the La motif, RRM1 and the interdomain linker of LARP6 in the recognition of collagen mRNA expands the RNA binding repertoire of the La module. *Nucleic Acids Res.* **43**, 645–60 (2015).

211. Thompson, E. J., Depaul, A. J., Patel, S. S. & Sorin, E. J. Evaluating molecular mechanical potentials for helical peptides and proteins. *PLoS One* **5**, (2010).

212. Duong, V. T., Chen, Z., Thapa, M. T. & Luo, R. Computational Studies of Intrinsically Disordered Proteins. *J. Phys. Chem. B* **122**, 10455–10469 (2018).

213. Ye, W., Ji, D., Wang, W., Luo, R. & Chen, H.-F. Test and Evaluation of *ff99IDPs* Force Field for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **55**, 1021–1029 (2015).

214. Simpson, P. J. *et al.* Structure and RNA interactions of the N-terminal RRM domains of PTB. *Structure* **12**, 1631–43 (2004).

215. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* NA-NA (2010). doi:10.1002/prot.22711

216. Keul, N. D. *et al.* The entropic force generated by intrinsically disordered segments tunes protein function. *Nature* **563**, 584–588 (2018).

217. de Araujo, A. S., Martinez, L., Nicoluci, R. de P., Skaf, M. S. & Polikarpov, I. Structural modeling of high-affinity thyroid receptor-ligand complexes. *Eur. Biophys. J. with Biophys. Lett.* **39**, (2010).

218. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).

219. Meirovitch, H., Cheluvaraja, S. & White, R. P. Methods for calculating the entropy and free energy and their application to problems involving protein flexibility and ligand binding. *Curr. Protein Pept. Sci.* **10**, 229–243 (2009).

220. Gilson, M. K. & Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42 (2007).

221. Wereszczynski, J. & McCammon, J. A. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Q Rev Biophys* **45**, 1–25 (2012).

222. Zhou, H. X. & Gilson, M. K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **109**, 4092–4107 (2009).

223. *Protein simulations*. (Elsevier, 2003).

224. Head, M., Given, J. & Gilson, M. “Mining Minima”: direct computation of conformational free energy. *J. Phys. Chem. A* **5639**, 1609–1618 (1997).

225. Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **215**, 617–621 (1993).

226. Case, D. A. Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.* **4**, 285–290 (1994).

227. killian, B. J., Kravitz, J. Y. & Gilson, M. K. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **127**, (2007).

228. Leach, A. R. *Molecular modelling : principles and applications* . (Prentice Hall, 2001).

229. Feng, Z. *et al.* Modulation of HIF-2α PAS-B domain contributes to physiological responses. *Proc. Natl. Acad. Sci.* **115**, 13240–13245 (2018).

230. Depping, R. & Oster, H. Interplay between environmentally modulated feedback loops - hypoxia and circadian rhythms - two sides of the same

coin? *FEBS J.* **284**, 3801–3803 (2017).

231. Rojas-Pirela, M. *et al.* Structure and function of Per-ARNT-Sim domains and their possible role in the life-cycle biology of Trypanosoma cruzi. *Mol. Biochem. Parasitol.* **219**, 52–66 (2018).

232. Zhang, Y., Markert, M. J., Groves, S. C., Hardin, P. E. & Merlin, C. Vertebrate-like CRYPTOCHROME 2 from monarch regulates circadian transcription via independent repression of CLOCK and BMAL1 activity. *Proc. Natl. Acad. Sci.* **114**, E7516–E7525 (2017).

233. Culig, Z. & Santer, F. R. Studies on Steroid Receptor Coactivators in Prostate Cancer. in 259–262 (2018). doi:10.1007/978-1-4939-7845-8_15

234. Hartzell, A. L. *et al.* NPAS4 recruits CCK basket cell synapses and enhances cannabinoid-sensitive inhibition in the mouse hippocampus. *Elife* **7**, (2018).

235. Sun, X. & Lin, Y. Npas4: Linking Neuronal Activity to Memory. *Trends Neurosci.* **39**, 264–275 (2016).

236. Harmon, A. C. *et al.* Particulate matter containing environmentally persistent free radicals induces AhR-dependent cytokine and reactive oxygen species production in human bronchial epithelial cells. *PLoS One* **13**, e0205412 (2018).

237. Tang, X., Shao, J. & Qin, X. Crystal structure of the PAS domain of the hEAG potassium channel. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **72**, 578–585 (2016).

238. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).

239. Best, R. B. & Mittal, J. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. (2010). doi:10.1021/JP108618D

240. Sousa, A. W. & Vranken, W. F. Open Access ACPYPE - AnteChamber PYthon Parser interfacE. 1–8 (2012).

241. Labadie, B. W., Bao, R. & Luke, J. J. Reimagining IDO Pathway Inhibition in Cancer Immunotherapy via Downstream Focus on the Tryptophan–Kynurenine–Aryl Hydrocarbon Axis. *Clin. Cancer Res.* **25**, 1462–1471 (2019).

242. Koper, J. E. B. *et al.* Polyphenols and Tryptophan Metabolites Activate the Aryl Hydrocarbon Receptor in an in vitro Model of Colonic Fermentation. *Mol. Nutr. Food Res.* **63**, (2019).

243. Ge, L., Cui, Y., Cheng, K. & Han, J. Isopsoralen enhanced osteogenesis by targeting AhR/ERα. *Molecules* **23**, (2018).

244. Head, M., Given, J. & Gilson, M. “Mining Minima”: direct computation of conformational free energy. *J. Phys. …* **5639**, 1609–1618 (1997).