UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE QUÍMICA E BIOQUÍMICA

Study of the interaction between the amyloid β peptide and the S100B protein using computational techniques

Filipe Eduardo Pequito Rodrigues

MESTRADO EM BIOQUÍMICA
Especialização em Bioquímica

Dissertação orientada por:
Doutor Miguel Ângelo dos Santos Machuqueiro
Doutor Cláudio M. Gomes

2020

Acknowledgments

I would like to start by acknowledging my two supervisors, Miguel Machuqueiro and Cláudio Gomes for their key role in guiding me during this work. A special thanks to Miguel Machuqueiro for walking me through and showing me a glimpse of the Molecular Modelling and Simulations world, with unending patience and determination to always teach me something more. I would also like to thank Bruno Victor, Paulo Costa, Pedro Reis, Tomás Silva, Diogo Reis, Nuno Oliveira, Sara Ferreira, Bernardo Henriques and Andreia Fortuna for all the fun lunches and coffee breaks before the pandemic hit, as well as interesting and enlightening discussions and advises. I would also like to thank Pedro Suzano, João Vitorino and Maria Carolina, as well as Nuno Oliveira once again for accompanying me during Bachelor's and Master's degrees, for putting up with my sometimes stubborn personality. Especially to João Vitorino for also having the idea of trying to do a sort of volunteer internship in this computational research group.

I would also like to thank some of my closest friends, Marlene Cantante, Bruno Santos and Susana Godinho for all their support, company, long serious and not so serious talks and all the trips to the cafe, moments which I always enjoy, even if I am quiet most of the time. Also for always cheering me up.

Lastly, but not least, I would like to give a special acknowledgment to my family, namely my parents, Eduardo Rodrigues and Albertina Rodrigues for their great efforts in both economical and supportive sides, not only during Master's and Bachelor's, but during my growth up to this point. I am very aware of all the sacrifices you have done, and I'm extremely thankful to you for your unshakable presence and support. There is only you two and my grandparents to thank for, for teaching me many of the science unrelated things I know, and for shaping me into the person I am today. Also to my uncle and aunt, Francisco Pequito and Joaquina Pequito, and to my cousin, Francisco Pequito, for their support and always amusing me.

Abstract

Amyloid beta peptides (Aβ) are associated with several neurodegenerative diseases, notoriously with Alzheimer's disease (AD). In the AD brain, Aβ peptides are produced as result of misprocessing of APP which results in the production of the variants with multiple lengths, including the highly amyloidogenic Aβ_{42}. The latter aggregates extracellularly, forming the so called amyloid plaques, which constitute one of AD's landmarks. The formation of these aggregates, which are neurotoxic, leads to an inflammatory response releasing several inflammatory citokynes, among which is S100B. S100B is a small homodimeric protein from the S100 protein family that binds four Ca^{2+} per dimer in EF-hands motifs and which also comprises an interfacial metal binding site for Cu^{2+} and Zn^{2+}. It has signaling and pro-inflammatory functions in late AD stages but recent findings implicate this protein as a regulator of proteostasis and metallostasis in the brain. S100B was described to act as an extracellular chaperone counteracting Aβ_{42} aggregation in a Ca^{2+} dependent manner. Binding of Ca^{2+} to S100B induces conformational changes which result in the exposure of a binding cleft within the dimer interface that is able to mediate protein:protein interactions. Experimental data from structural NMR experiments revealed that Aβ_{42} interacts with S100B through this region, involving Lys28 and Ile31. However, the atomic details of this interaction are not well known. The main goal of this thesis is therefore to elucidate these details, through the use of computational techniques, namely docking calculations and molecular dynamics simulations, guided by experimental evidence.

Initially, quantum mechanics calculations were performed to parametrize the metal coordination centers of S100B. At the start of this work, we used molecular docking calculations to obtain starting configurations of the complex for molecular dynamics simulations. As such, we used HADDOCK, a docking tool developed to handle protein-protein interactions. After obtaining and selecting promising structures of the complex, we used them as starting configurations of long (1 μs) molecular dynamics simulations. Most simulations converged and the conformational space obtained for the different binding modes (BMs) was explored using several analysis. We focused our attention at the interfaces formed by the α-helix rich region of Aβ42, what we named the *head* group, where we identified hotspot interactions. We quantified the hydrophobic/hydrophilic nature of the interactions, using a newly devised SAS^{hydro} index property, and calculated the binding affinities between S100B and the Aβ42 *head* region using molecular mechanics energies combined with Poisson-Boltzmann and solvation surface area (MM/PBSA).

Two BMs (BM1 and 2) showed very stable interfaces, where the two key residues were packed inside the S100B interface and with Lys28 stabilized by neighboring acidic residues, which provided an increase in the Coulombic contribution to the binding energy. The main difference between BM1 and BM2 came from a the topological arrangement of the Aβ_{42} *tail* region, where in BM1 it was mostly unstructured, while in BM2, it retained a large meta-stable α-helix. The fact that the peptide is only partially folded in solution creates a huge challenge to model its conformational space which is inaccessible to most computational techniques. However, with the help of all experimental evidence available, we were able to devise a protocol to direct both the conformational space of the peptide and the geometry of the final complex configurations. Altogether, we were able to explore successful binding modes which provided atomistic information about the specific interactions that are stabilizing the complex.

Keywords: Aβ, S100B, HADDOCK, Molecular dynamics, binding configuration

Resumo

Péptidos beta amiloides (Aβ) estão associados a várias doenças neurodegenerativas, notoriamente com a Doença de Alzheimer (AD). A nível cerebral, os péptidos Aβ são produzidos como resultado de um processamento proteolítico que deixa o segmento que origina os péptidos Aβ intacto, ao invés de sofrer um processamento proteolítico que trunca este mesmo segmento, o que conduz à perda da sua capacidade de agregar. Estas duas vias de degradação da APP competem, diferindo apenas num dos dois enzimas necessários para o processamento. Do processamento amiloidogénico resultam variantes do péptido Aβ com vários comprimentos, incluindo o Aβ_{42}, com elevada propensão amiloidogénica. Este agrega, a nível extracelular, formando placas amiloides, que constituem um dos grandes marcos da AD. A formação destes agregados amiloides, que são neurotóxicos, leva a uma resposta inflamatória, que leva à libertação de várias citocinas inflamatórias, de entre as quais se destaca a S100B. A S100B é uma pequena proteína homodimérica da família de proteínas S100 que têm a capacidade de se ligar a 4 Ca^{2+} por dímero em motivos *EF-hand* e que também contem centros de ligação a Cu^{2+} e Zn^{2+} localizados na interface entre monómeros. Esta proteína possui funções de sinalização e de pró-inflamação em estados tardios da AD, contudo descobertas recentes implicam esta proteína como um regulador da proteostase e metalostase a nível cerebral. A S100B possui também atividade de chaperona, contrariando a agregação de Aβ_{42} de uma maneira dependente de Ca^{+2}. Ligação de Ca^{2+} à S100B induz alterações conformacionais na mesma, que resultam na exposição de uma cavidade de ligação localizada na interface do dímero, que é capaz de mediar interações proteína:proteína. Dados experimentais de experiências de ressonância magnética nuclear (NMR) revelaram que o Aβ_{42} interage com a S100B através desta cavidade, envolvendo a lisina 28 e a isoleucina 31 do péptido. Contudo, os detalhes a nível atómico desta interação não são bem conhecidos. O principal objetivo desta tese é de elucidar estes detalhes, através do uso de técnicas computacionais, nomeadamente cálculos de *docking* e dinâmica molecular, guiadas por evidências experimentais.

Um dos desafios neste trabalho prende-se pelo facto do péptido utilizado, o Aβ_{42}, ter uma estrutura em solução aquosa bastante dinâmica, o que corresponde a uma enorme variabilidade conformacional. Como tal, combinando esta variabilidade com todas as possibilidades de encaixar o péptido na superfície da S100B, mesmo considerando apenas a cavidade de ligação localizada na interface exposta após ligação de iões metálicos, obtem-se um problema de dificil resolução com ferramentas de *docking* com o poder computacional atual. Como tal, para além de se limitar a região da superfície da S100B à qual o péptido pode estar ligado, foi também tido em consideração que o segmento do péptido que se encontra nessa região deverá ter uma estrutura em hélice. Considerando ainda os dados de NMR acima referidos, que sugerem que a lisina 28 e a isoleucina 31 do péptido devem ter especial relevância na ligação com a S100B, optámos por definir que o segmento dos resíduos 25 a 35 do péptido, que contem os resíduos 28 e 31, deverá ser o segmento com estrutura em hélice. De forma a simplificar ainda mais o sistema, como a lisina 28 e isoleucina 31 estão contidas no segmento 25-35, considerámos que para os primeiros cálculos de *docking*, este segmento será o que apresenta um maior impacto neste complexo. Como tal, truncou-se o péptido e para estes primeiros cálculos, utilizou-se apenas este segmento, sendo que nos últimos cálculos de *docking* foi utilizado o péptido completo. Todas estas considerações permitiram simplificar e reduzir bastante a quantidade e variedade de possibilidades para a configuração deste complexo,

acelerando assim a obtenção de configurações iniciais do mesmo.

Inicialmente, cálculos de mecânica quântica foram realizados com o intuito de parametrizar os centros de coordenação metálicos da S100B. No ínicio deste trabalho, cálculos de *docking* molecular foram utilizados de modo a obter configurações iniciais do complexo para serem utilizadas em simulações de dinâmica molecular. Como tal, utilizámos o HADDOCK, uma ferramenta de *docking* desenvolvido para lidar com interações proteína:proteína. Após obtenção e seleção de estruturas do complexo promissoras, estas foram utilizadas como configurações iniciais em simulações longas de dinâmica molecular, com cerca de 1 μs. A maior parte das simulações convergiram e o espaço conformacional obtido para os vários modos de ligação (BMs) foi explorado utilizando várias análises. Nós focámos a nossa atenção nas interfaces formadas pela região do Aβ_{42} enriquecida em hélice α, a qual nomeámos de grupo cabeça, no qual foi possível identificar interações importantes. Nós quantificámos a natureza hidrófoba/hidrófila destas interações, utilizando uma propriedade recentemente desenhada, denominada de indice SAS^{hydro}, e calculámos a afinidade de ligação entre a S100B e a região da cabeça do Aβ_{42} utilizando cálculos de energia de mecânica molecular combinados com análise de Poisson-Boltzmann e de área superficial de solvatação (MM/PBSA).

Foram obtidos cinco BMs distintos para o complexo S100B:Aβ_{42} com este procedimento. Com este plano, nós amostrámos o espaço conformacional destes cinco modos de ligação, cada um com três replicados, atingindo um total de tempo de simulação de 15 μs. Nós aplicámos esforços significativos na análise da equilibração das simulações de dinâmica molecular. Isto é de grande importância, dado que uma grande porção do péptido Aβ_{42}, a qual apelidámos de região da cauda, foi particularmente difícil de convergir e, ao correr apenas simulações triplicadas, não temos certeza de ter sido atingido equilíbrio nesta região do péptido. Contudo, o nosso foco, desde o início, esteve sempre na região da cabeça, que foi consideravelmente mais fácil de equilibrar nos modos de ligação que se mostraram mais estáveis. Nós observámos que, em várias configurações, o péptido perdeu a maior parte da dua helicidade, o que conduziu a que fossem descartadas. Nenhum dos replicados simulados para o modo de ligação 4 foram mantidos, o que indica que esta configuração obtida pelo HADDOCK não era estável. Apesar deste modo de ligação ter sido obtido a partir de uma solução particularmente boa dos cálculos de *docking*, o restante do péptido, nomeadamente a região da cauda, após reconstrução posterior do péptido, pode ter sido colocada numa região particularmente desfavorável da superfície da S100B. Um dos replicados do modo de ligação 3 e dois do modo de ligação 5 foram também descartados devido a instabilidades no conteúdo de estrutura secundária do péptido. Para além disto, no replicado restante da configuração 5, foram amostradas configurações nas quais a isoleucina 31 e, parcialmente, a lisina 28 não estavam completamente dedicadas à interface com a S100B, o que vai contra as evidências experimentais. É interessante que este tenha sido o único modo de ligação obtido de cálculos com o Aβ_{42} completo, em vez de apenas um segmento do mesmo. Isto confirma as nossas restrições iniciais relativamente ao uso do péptido completo, dado que isto cria um grande espaço configuracional para os cálculos, nos quais a informação experimental que utilizámos para guiar o procedimento de *docking* ficou diluída. No fim foi possivel atingir um modo de ligação no qual o péptido Aβ não estava nem instável nem estabilizado numa configuração enviesada.

Finalmente, é ainda de referir que os nossos esforços computacionais para estudar os detalhes moleculares do complexo S100B:Aβ_{42} estão significativamente limitados pela natureza complexa das interações em questão. Apesar da proteína S100B ser particularmente estável e com um espaço conformacional facilmente capturável com poucas centenas de nanosegundos, o péptido Aβ_{42} é particularmente problemático, devido à sua grande variabilidade conformacional previamente referida. O facto deste péptido se encontrar apenas parcialmente *folded* em solução cria um enorme desafio para modelar o seu espaço conformacional vasto, que é inalcançável pela maior parte das técnicas computacionais. Contudo, com a ajuda de evidências experimentais disponíveis, foi possível desenhar um protocolo para guiar tanto o espaço conformacional como a geometria das configurações do complexo finais. No geral, foi pos-

sível explorar modos de ligação plausíveis, o que fornece informação a nível atomístico relativamente às interações específicas que estabilizam o complexo.

Palavras-chave: Aβ, S100B, HADDOCK, Dinâmica Molecular, Configuração de ligação

Contents

List of Figures

List of Tables

List of Abbreviations

Asp	Aspartic acid
Glu	Glutamic acid
His	Histidine
Met	Methionine
Thr	Threonine
Leu	Leucine
Asn	Asparagine
Gln	Glutamine
Phe	Phenylalanine
Ile	Isoleucine
Lys	Lysine
Ser	Serine
AD	Alzheimer's Disease
Aβ	Amyloid beta peptide
APP	β-amyloid precursor protein
l-BFGS	limited memory - Broyden–Fletcher–Goldfarb–Shanno
MD	Molecular dynamics
MM	Molecular mechanics
NMR	Nuclear magnetic resonance
PB	Poisson-Boltzmann
PBC	Periodic boundary conditions
PDB	Protein data bank
PEF	Potential energy function
RMSD	Root mean square deviation
SEM	Standard error of the mean
SPC	Simple point charge
APP	amyloid precursor protein
QM	Quantum Mechanics
HF	Hartree–Fock
DFT	Density Function Theory
ESP	Electrostatic Potential
MK	Merz-Kollman
RESP	Restrained Electrostatic Potential
PME	Particle-Mesh Ewald
SASA	Solvent Acessible Surface Area
WW	Wimley–White
MM/PBSA	Molecular Mechanics with Poisson-Boltzmann and Surface Area solvation
MM/GBSA	Molecular Mechanics with generalized Born and Surface Area solvation
BM	Binding Mode

Chapter 1

Introduction

1.1 Protein Aggregation

In general, the structure of a protein is paramount to the performance of its functions. Proteins present several levels of structure, from primary, the main sequence, to secondary, mainly consisting of α-helix and β-sheet due to interactions between main chain elements, whereas tertiary structure is based of the spatial arrangement the protein exhibits (Figure 1.1). The quaternary structure is associated with a possible formation of complexes by two or more monomers, which is not characteristic to all proteins in their folded native state [1]. The complex process of a protein acquiring its native structure is called folding and is mainly driven by physico-chemical principles [3–5]. It is worth mentioning that, besides its complex nature, the folded state is only marginally stable (Figure 1.2) [6]. This is essential for the majority of non-structural proteins, since this partial "instability" allows for an extra flexibility which is required for these proteins to perform their roles. The marginal stability observed in the folded states

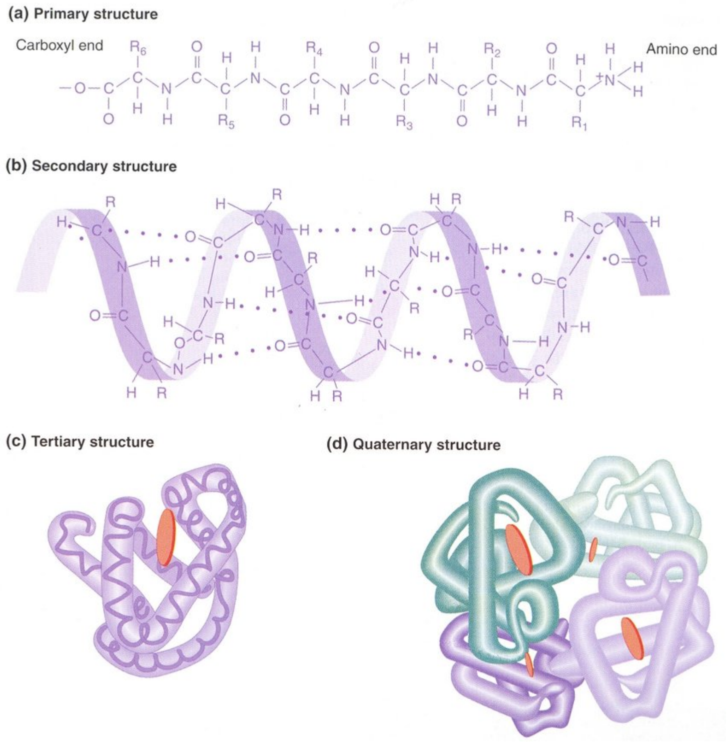


Figure 1.1: The four levels of protein structure (a-d). Adapted from Reference [2].

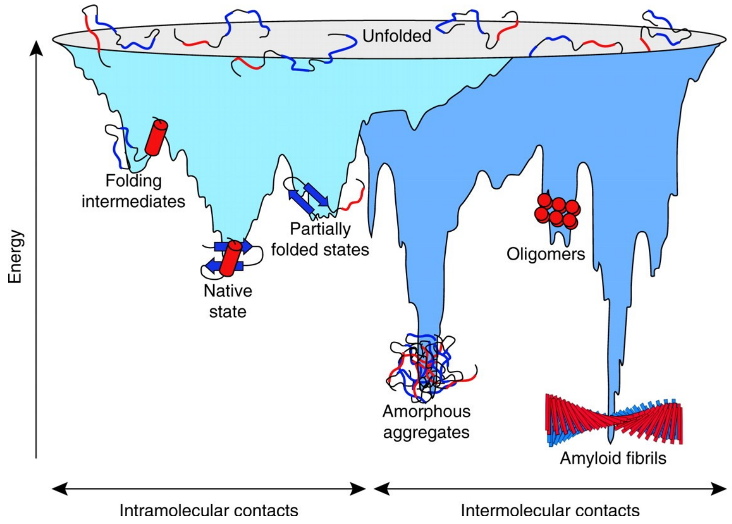


Figure 1.2: Energy landscape scheme illustrating the protein folding, misfolding and aggregation states. Adapted from [7].

means that they do not represent a global energy minima. Realistic folding energy landscapes are usually formed by rugged funnel shape profiles where the native state is made of an ensemble of native-like conformations. These lie only at energetic local minima, as previously mentioned, whereas the global minima often correspond to aggregation states [8].

This process, in particular amyloid aggregation, is characterized by several steps (Figure 1.3), starting with the exposure of aggregation-prone regions, due to conformational fluctuations, which may lead to the formation of non-native dimers rich in cross β-sheet structure. Then, several of these units can interact to form oligomers, which continued linear growth leads to protofilaments, and their association forms fibrils [8, 9]. These structures, namely the amyloid fibrils, are highly stable structures, requiring very harsh conditions to disassemble. These can be very problematic, since they can destabilize membranes and quelate metal ions, which are essential to the homeostasis of the cell [10, 11]. Furthermore, the amyloid oligomers can induce the aggregation of other soluble proteins, further enhancing the amyloid fibril growth and disabling other proteins that end up aggregating [12].

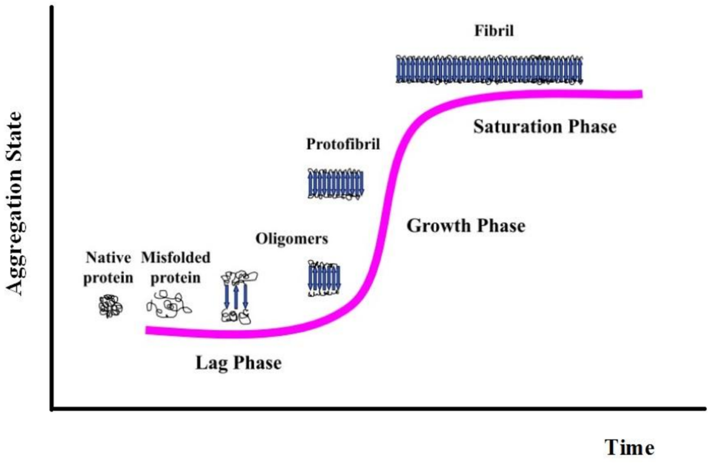


Figure 1.3: The several stages of the amyloid aggregation phenomena. Adapted from [13].

1.2 Aggregation and Alzheimer's disease

Alzheimer's disease is an example of a neurodegenerative disorder deeply connected with the aggregation phenomena. In this particular case, the disease is associated with two distinct aggregation processes, the tau protein at the intracellular level, and Aβ peptides at the extracellular level [14]. This work will focus on the later.

Alzheimer's disease is amongst the more common neurodegenerative diseases. Some of its characteristics consist of metal ion imbalances, chronic neuroinflammation and formation of protein aggregates, neurofibrillary tangles at the intracellular level (mainly made of tau protein) and amyloid plaques at the extracellular level (Aβ peptides) [14, 15]. The presence of exogenous Aβ promotes the release of great amounts of pro-inflammatory cytokines and further production of Aβ peptides. These phenomena result in a loop of positive autocrine regulation, which leads to the previously mentioned chronic inflammation, as well as to increased levels of Aβ [15].

1.3 Aβ peptides

The several different sized Aβ peptides all derive from a type I transmembrane protein, β-amyloid precursor protein (APP), more precisely, from a proteolytic pathway of this protein. APP has two major secretase dependent degradation pathways (Figure 1.4). One of these is anti-amyloidogenic and goes through by α- and γ-secretase, while the amyloidogenic pathway goes through the β- and γ-secretase [16].

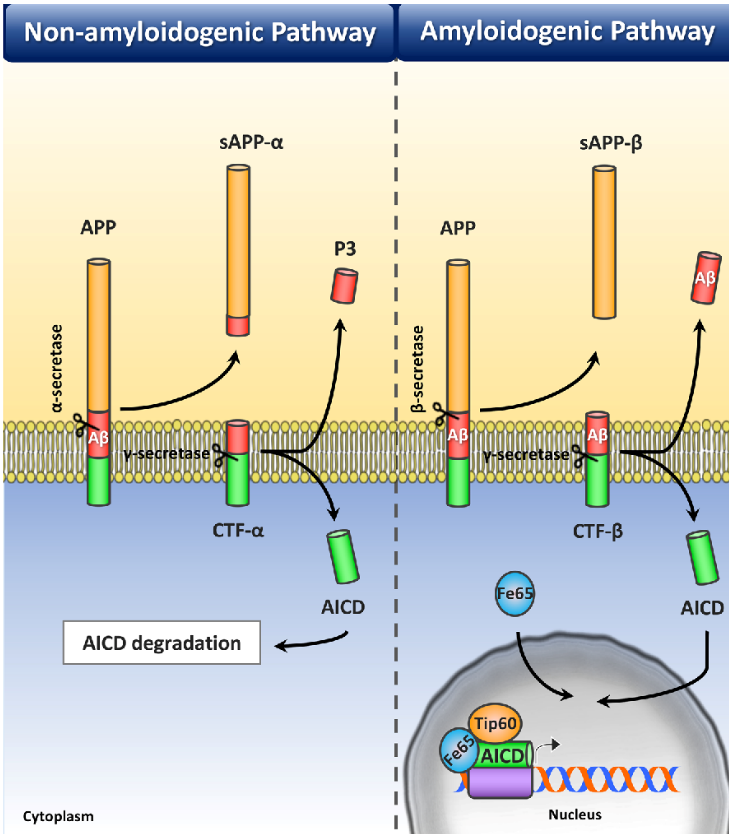


Figure 1.4: Degradation pathways of APP. Adapted from [17].

γ-secretase is present in both pathways and exerts its activity in the hydrophobic region of the cellular membrane, which releases the resulting peptide to the extracellular domain. α- and β-secretases compete, to some degree, for the cleavage of APP. α-secretase cleaves APP in the middle of the Aβ region and, after γ-secretase action, a truncated form of Aβ, called p3 is released, which does not appear to show any pathological relevance. On the other hand, β-secretase cleaves the majority of APP ectodomain, and after γ-secretase action, the Aβ peptide is released. It is noteworthy that γ-secretase cleavage can occur more than once on the same substrate, and also, the final hydrolysis is imprecise, which leads to Aβ peptides of varying lengths [16].

The two most prominent Aβ peptides are Aβ_{40} and Aβ_{42}. Even though these peptides only differ in their size by two residues, they have several different characteristics. For instance, in the bodily compartments where these peptides co-exist, they are present in different concentrations. They also display distinct sub-cellular origins, uptake pathways and aggregation mechanisms, as well as different levels of toxicity, where Aβ_{42} is more toxic [18]. Despite their differences, both peptides are thought to lack a well defined structure in aqueous media [19]. When in membrane or membrane-like environments, they take mainly a α helix conformation (Figure 1.5), In more aqueous media, it can adopt more β-sheet prone conformations, which, if not stabilised, can trigger the amyloid formation and propagation [19–21]. Regardless, due to its prominence on the disease onset and higher cytotoxicity, the peptide mainly addressed on this work is Aβ_{42}.

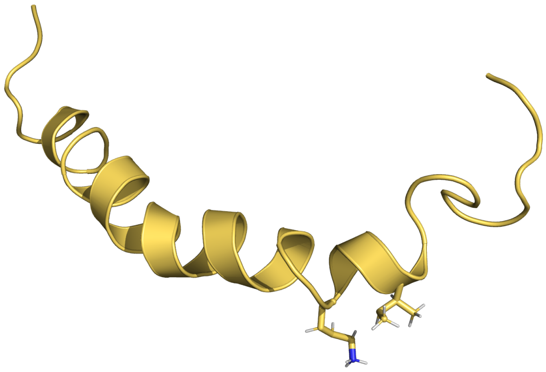


Figure 1.5: Aβ_{42} peptide structure in a membrane-like environment. Adapted from frame 26 of PDB entry 1Z0Q, a structure obtained through NMR in a 30/70 ratio solution of Hexafluoro-2-propanol/water ([22]

1.4 The S100B possible role in aggregation

One of the cytokines affected by the levels of Aβ is S100B, which belongs to the S100 chaperone family and corresponds to one of the more abundant soluble proteins in cerebral tissue [15]. This protein is a homodimer containing binding sites for different ions [15, 23]. Each monomer contains one "pseudo EF-hand" domain near the N-terminal, with the backbone of residues Ser18, Glu21, Asp23, Lys26 and the side chain of Glu31, and a EF-hand domain near the C-terminal, with the backbone of Glu67 and side chains from Asp61, Asp63, Asp65 and Glu72, to bind calcium. The domain closer to the N-terminus establishes a weak bind to a calcium ion [23]. Additionally, the dimer possesses two coordination centers for zinc ions, both shared between the monomers, with His15 and His25 from one monomer, and His85 and Glu89 from the other monomer. It is worth noting that this is true at roughly neutral pH (6.5), while at more basic pH (9.0), a ligand swap occurs, with Glu89 being replaced by His90 (which should now be mostly deprotonated), with no impacting changes to the overall structure of the coordination center

[23]. The binding of zinc, leads to an increase in the affinity towards calcium [15, 23–25] and, when both ions are complexed, it induces a significant conformational change exposing a hydrophobic region in the dimer interface (Figure 1.6).

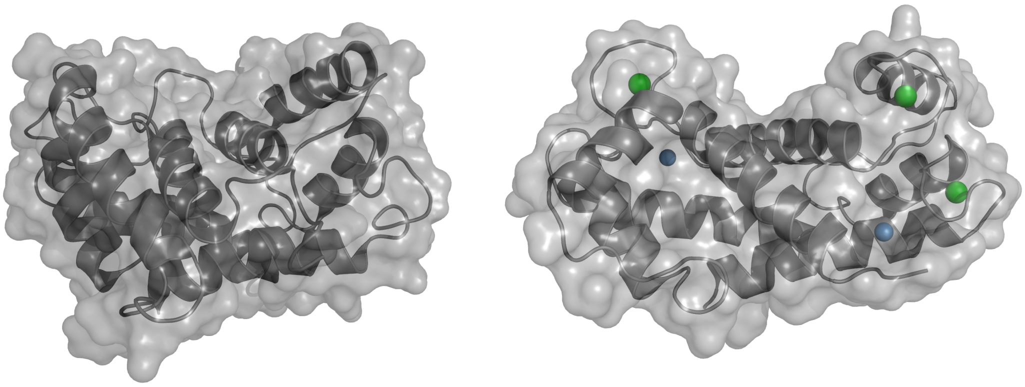


Figure 1.6: Structure of S100B in the apo (left) and in the ion-bound (right) states. The apo state was obtained from the PDB entry 2PRU [26], while the ion-bound state was obtained from PDB entry 3D0Y [27].

This region corresponds to a promiscuous binding site for peptides [25]. Even though the exposure of this region may play a role in some native functions of S100B, it also has been associated with increased aggregation potential [28]. Along these lines, it was proposed that S100B may have a role in the aggregation of Aβ [15, 28]. Indeed, recent studies show that S100B can interact with Aβ peptides, particularly with Aβ_{42}, which has a strong connection to the formation of amyloid plaques (Figure 1.7) [15]. This interaction occurs at the mentioned binding site (top and middle panels of Figure 1.7) and seems to increase the α-helix content in the Aβ_{42} structure [15]. This Aβ sequestration promoted by S100B may be the molecular basis for the slower formation of amyloid aggregates [15]. Furthermore, when present in high concentrations (μM), S100B has a neurotoxic effect, which highlights the complex regulation revolving around the S100 protein family and its interacting partners [15, 25]. Also, it is worth pointing out that implicit to the chaperone activity of S100B comes the necessity to establish reversible interactions with several different proteins and peptides, which requires a promiscuous binding site to their *clientome*, the range of proteins on which the chaperone can act, with only a transient interaction [29]. This is achieved by binding with frustrated regions of the target protein [29]. This frustration is due to the conflict between the constraint imposed on the mobility of amino acids, particularly relevant when in structured motives, and the necessity to minimize the free energy of contacts between amino acids [29, 30]. Altogether, it is clear that many of the important questions remain unanswered and the precise mechanism through which this protein interacts with Aβ is still unknown.

1.5 Current limitations using computational techniques

An important obstacle in fighting Alzheimer's disease is the lack of consensus relative to the definition of ideal bio-markers to detect the early onset of the pathology. This is aggravated by a lack of drugs to help prevent or treat the disease. In fact, the focus has been on treating the symptoms, probably due to a lack of understanding of the dynamic behaviour of molecules, such as Aβ_{42}, which may have a role on the development of the disease [31]. The apparent intrinsic lack of structure by the Aβ peptides in aqueous media, and the fact that they gain different structures when interacting with specific partners, makes these molecules very difficult to study from a structural biology point of view. Furthermore, several important inter molecular events involving Aβ, relevant to processes such as folding or binding, happen on a time

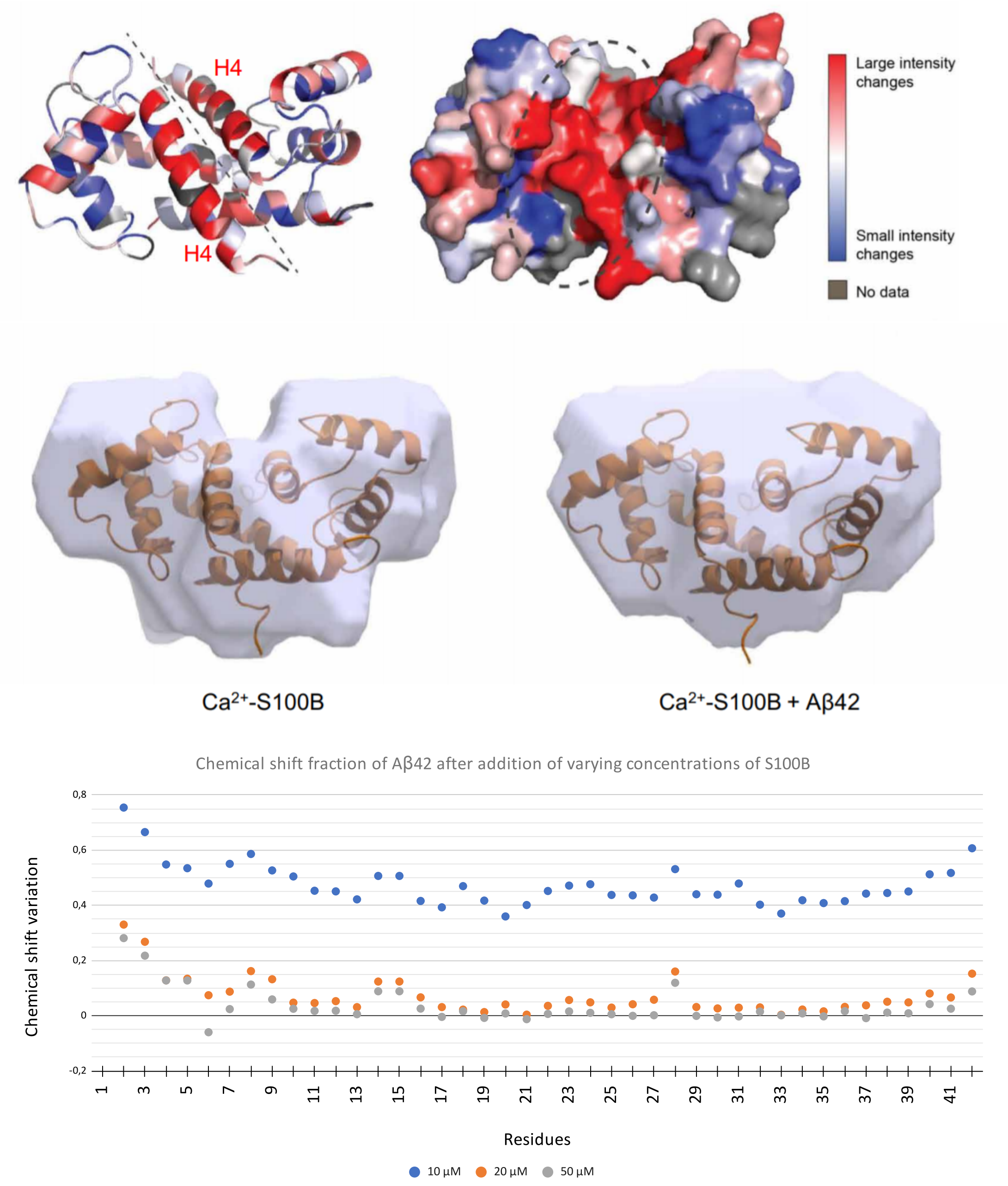


Figure 1.7: Experimental evidence supporting the binding of Aβ_{42} to S100B. Top panel: secondary structure (left) and surface (right) representations of S100B colored by Aβ_{42} interaction heat map (red corresponds to more interaction); Middle panel: surface filling representation of a structural model of S100B in the absence (left) and presence (right) of Aβ_{42}; Bottom panel: chemical shift variation per Aβ_{42} residue at different concentrations of S100B. All images were adapted from data present in Reference [15].

scale not easily reached by conventional experimental methods. These difficulties have also been holding back many computational efforts to address this system [32]. Nevertheless, methods such as Molecular Dynamics (MD) simulations allow to study the conformational space of many biomolecules at atomic resolution and up to the μs timescale. Also, these techniques can be complemented with Molecular Docking calculations, which is another computational technique used to obtain starting configuration of macromolecular complexes, that would otherwise take too long to obtain with MD simulations. This is pivotal when only indirect biophysical data is available without any NMR or X-Ray Crystallography structural information. The biophysical data available can be used to guide these computational methods and help design protocols tackling complex systems, like the study of the interaction mechanism between $A\beta_{42}$ and S100B.

1.6 Goals of this work

The main goal of this study is to elucidate the molecular details of the interaction between $A\beta_{42}$ and S100B. Understanding this interaction will allow us to comprehend the mechanism of action of S100B over the aggregation process of $A\beta$ in the context of Alzheimer's disease. We will use experimental data to guide a docking protocol in order to obtain starting configurations of the $A\beta$:S100B complex for further study using MD simulations. The HADDOCK web server will perform the docking between the $A\beta$ peptides and S100B and several long MD simulations will help us study these complexes in the following aspects:

- to evaluate the overall stability of the complex formed;
- to identify the key residues and residue segments involved in the complex stability;
- to understand the structural impact of the binding phenomena to both molecules;
- to define a plausible preferred binding mode in the $A\beta$:S100B complex.

Chapter 2

Theory and Methods

We will start by laying the theoretical foundations of the several computational techniques employed in this thesis. All methods used in this work are numeric in nature, hence, we think that a brief context introduction is in order. In order to use computational structural biology techniques, like molecular docking or molecular dynamics simulations, we rely on the availability of structural information about the system in study. The experimental data on protein structures is very important, but sometimes we need to use quantum mechanics calculations to obtain information, like electron density or the electrostatic potential, that is not readily available from NMR or X-ray crystallography. In this chapter, we will address each technique in both their theoretic foundations and the settings adopted to perform the calculations.

2.1 Quantum Mechanics

The S100B protein used in this work, contains metal binding sites for which the parameters are not available in regular protein force fields. Hence, we resorted to Quantum Mechanics (QM) techniques to obtain bonded parameters and atomic partial charges for these metal complex groups. The most detailed and precise model of a molecule is derived from QM. This is due to the fact that the molecular orbitals and the electrons are used directly with few approximations. The main disadvantage of dealing with such high level of detail is that only very small systems can be tackled with our current computational power.

2.1.1 Solving Schrodinger's equation

Computational techniques based on QM use the Schrodinger's equation to compute the time-independent interactions of particles. By solving this equation, we can describe the energy of a given system, as well as perform geometry optimization on its structure. After we obtain an optimized (minimized) structure, we can then derive bonded parameters, such as bond length, bond angles and dihedrals, as well as perform electrostatic potential calculations to help fit the atomic partial charges. The solution of Schrodinger's equation is only possible for monoeletronic systems [33], while for polyeletronic systems, only an estimation of the solution can be achieved, for which an assortment of theories can be applied. The time-independent Schrodinger's equation takes the form of,

$$H\Psi = E\Psi \tag{2.1}$$

where E represents the energy of the system, Ψ represents the wave function of a given atom and H represents a Hamiltonian operator which may be described as follows,

$$H = -\frac{\hbar^2}{2\mu}\sum_i \nabla_i^2 + \sum_i V(r_i) + \sum_{i \neq j} V(r_{ij}) \tag{2.2}$$

where $\hbar$ is Planck's constant, μ is the reduced mass, which corresponds to a normalized mass, where M is the nuclear mass and m the mass of the electron. ∇ corresponds to the Laplacian term applied to systems with more than one dimension, and V corresponds to a Coulombic term to describe the potential between two charges, between each electron and the nucleus ($V(r_i)$) or between each electron pair ($V(r_{ij})$).

$$\mu = \frac{mM}{m+M} \tag{2.3}$$

$$V(r_i) = -\frac{Z_e^2}{4\pi\varepsilon_0 r_i} \tag{2.4}$$

$$V(r_{ij}) = -\frac{Z_e^2}{4\pi\varepsilon_0 r_{ij}} \tag{2.5}$$

where Z_e is the charge of the nucleus, e is the charge of a single electron, ε_0 is the permittivity in vacuum and r is the distance between the two particles in consideration [33, 34].

The Hamiltonian operator for a hydrogen-like system, with one nucleus and one electron, takes the following form,

$$H = -\frac{\hbar^2}{2\mu}\nabla^2 - \frac{Z_e^2}{4\pi\varepsilon_0 r} \tag{2.6}$$

for which, an exact solution can be found. On systems with more than one electron, a Coulombic term to describe the interactions between electrons is added to take into account inter electronic repulsion, which results in wavefunctions that depend simultaneously on the coordinates of all the electrons. This makes it impossible to find an exact solution, only an approximation can be achieved [33, 34]. In order to achieve such approximation, several theories may be applied. One of such is the Hartree–Fock (HF) theory, which makes the assumption and approximation that the motion of a given electron is independent of the motion of every other, while still feeling a repulsive term due to the average positioning of the other electrons [35]. Even though this is a bit outdated by nowadays standards, the HF theory was still very important for developing its successors. More recent than the HF method, the Density Function Theory (DFT) took a different approach, bypassing the determination of the wave function. Instead, this method calculates the density distribution function using a functional set, which circumvents the multi-electronic problem [36]. Nevertheless, the DFT method does not rely on a theoretical basis as robust as the HF method, which led to the development of hybrid methods which combine these two distinct approaches.

2.1.2 Functionals and basis sets

As previously mentioned, in order to obtain an approximated solution of Schrodinger's equation for polyeletronic systems, different theory levels can be used. Both DFT and hybrid methods make use of functionals, which can be derived *Ab initio* or also possess empirically derived parameters. Functionals are functions of another function and, in the case of DFT theory, are function of the spatially dependent electron density. With the development of hybrid DFT methods, which helped improve the model of the exchange and correlation interactions, this approach became an important tool to obtain an approximate solution for multi electronic systems. On the other hand, a basis set, which is used by all the previously addressed methods, corresponds to a set of mathematical functions used to describe the atomic orbitals of the atoms.

2.1.3 Geometry Optimization

As previously mentioned, QM is used, among other ends, to perform geometry optimization. We used the software package Gaussian [37] to perform the QM calculations. This protocol consists of an energy minimization of the system until it achieves a local energetic minimum, which translates into an optimized structure. After obtaining this structure, we compared its bond lengths, angles and dihedrals to the ones present in our forcefield database. When similar parameters are available in the forcefield, they are adopted, otherwise, new parameters were created in the forcefield, following the geometric information from the QM calculation. The final parameters used for Molecular Dynamics are present in following link: `http://mms.rd.ciencias.ulisboa.pt/files/FilipeRodrgues_G54A7-ModFF.zip`

2.1.4 Electrostatic Potential (ESP) Charge Calculation

The QM calculations generate optimized structures of the metal coordinating sites and, from the final wave-functions, it is possible to calculate the electrostatic potential over the complexes surfaces. With relatively simple Merz-Kollman (MK) fitting procedures, we can obtain atomic partial charges that are compatible with the GROMOS forcefields. We have performed these MK fits using Gaussian [37], where no restrictions were in place. Alternatively, we also tested the use of restrained electrostatic potential (RESP) [38] charge calculations, without observing significant differences. The only restrictions imposed by the GROMOS forcefields are related with the charge groups overall neutrality. Therefore, we used the MK-derived partial charges from the QM calculations and manually curated them to obtain GROMOS-compatible charge groups. It is worth mentioning that all QM techniques were mainly used in this work to parameterize the metal coordination centers and allow performing Molecular Dynamics simulations in this protein.

2.1.5 QM Settings

In this work, the definition of the several ion coordination centers of the S100B dimer was required, in order to obtain parameters to properly describe these groups using MD techniques. This was performed by building small model systems surrounding each metal center, which could be optimized using QM calculations. Considering that some of the interacting residues were coordinating the metal ions with elements from their main chain, we simply mutated them to ALA residues (with the exception of Ser18), for computational efficiency. When residues are not interacting with the metals or coordinating only with their main chain, we opted by building the QM system using them whole. This procedure created 3 small model systems which attempt to mimic the 3 different metal centers in S100B protein (Table 2.1 and Figure 2.1).

Table 2.1: Coordination centers of S100B dimer. Being a homodimer, this protein contains six coordination centers, two copies of each. ^{a}) Residues which coordinate with their main chain; ^{b}) these two residues belong to the other monomer.

Coordination centers	Coordinating residues
CA1	Ser18^{a}, Glu21^{a}, Asp23^{a}, Lys26^{a}, Glu31
CA2	Asp61, Asp63, Asp65, Glu67^{a}, Glu72
ZN1	His85, Glu89, His15^{b}, His25^{b}

All missing hydrogen atoms in these systems were added and a molecular mechanics energy minimization step was performed using Avogadro [39]. All resulting structures were geometry optimized at the QM level using Gaussian, using the B3LYP functional and the 6-31G basis set. At the end of the optimization, the vibrations frequencies were also calculated to ensure that real minima were obtained. The Van der Waals radii of 0.24 and 0.21 nm were used for calcium and zinc ions, respectively [40]. The bonded parameters and the atomic partial charges obtained were used to define the newly created force field building blocks (Table 2.2). These building blocks are also available from our wesite at: `http://mms.rd.ciencias.ulisboa.pt/files/FilipeRodrgues_G54A7-ModFF.zip`

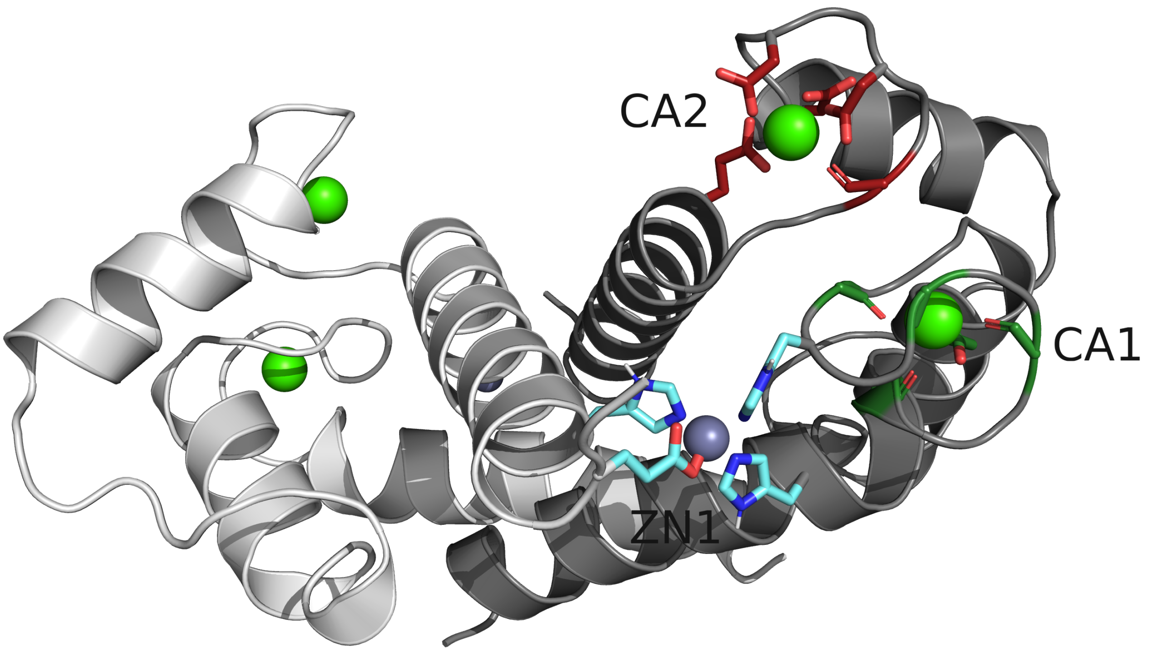


Figure 2.1: Structural representation of the three distinct metal coordination centers. The names of each coordination center correspond to the ones presented in Table 2.1.

Table 2.2: Description of the new GROMOS 54A7 building blocks created to describe the metal coordination centers.

Block names	Composition
CAS	Full serine residue coordinating with main chain
CAE	Full glutamate residue coordinating with main chain
CAD	Full aspartate residue coordinating with main chain
CAK	Full lysine residue coordinating with main chain
EEX	Glutamate residue without side chain
DDX	Aspartate residue without side chain
HSX	Histidine residue without side chain

2.2 Molecular Mechanics/Molecular Dynamics

In Molecular Mechanics/Molecular Dynamics (MM/MD), we use classic mechanics and Newtonian physics to describe the systems. This is possible due to the Born-Oppenheimer approximation, which allows the nuclear motions to be dissociated of the electronic motions, since the last ones adjust to the movements of the heavier nuclei. Even though this approach is less precise, it allows the study of more complex systems when compared to QM techniques, which trade computational efficiency for higher precision. MD simulations will be the adopted and used throughout this work in order to study the interactions between two relatively large macromolecular entities during an extended period of time.

2.2.1 Potential Energy Function

MD simulations rely on classic mechanics and describe the total potential energy of a given system using a function which depends on several contributions.

$$U = bonded \begin{cases} \frac{k}{2}(l-l_0)^2 + \\ \frac{k}{2}(\sigma-\sigma_0)^2 + \\ \sum_{n=0}^{N} \frac{V_n}{2}[1+cos(n\omega-\gamma)] + \\ k(1-cos2\omega) \end{cases} + nonbonded \begin{cases} \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} + \\ \sum_{i=1}^{N}\sum_{j>i}^{N}(C_{ij}^{12}/r_{ij}^{12}) - (C_{ij}^{6}/r_{ij}^{6}) \end{cases} \tag{2.7}$$

From these, we can distinguish 2 groups. The contributions associated with the local chemical topology and bonded interactions: the bond lengths, angles, and dihedral torsions. In the bond energy term,

$$U(l) = \frac{k}{2}(l-l_0)^2 \tag{2.8}$$

l is the distance between a given pair of atoms, l_0 is a reference bond length between said pair which minimizes the energy, and k is a elastic force constant of the atomic pair. The bond length term refers to the stretching and compression movements. A simple way to mathematically describe these movements is by a harmonic potential based on Hooke's Law for elastic behaviours. A similar harmonic potential approach can be used to model the angle amplitude parameter, associated with angle bending movements,

$$U(\sigma) = \frac{k}{2}(\sigma-\sigma_0)^2, \tag{2.9}$$

where σ is the angle between a given trio of atoms, σ_0 is a reference angle between said trio which minimizes the energy, and k is a elastic force constant. In the dihedral torsion angles, the energy term

$$U(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1+cos(n\omega-\gamma)] \tag{2.10}$$

$$U(\omega) = k(1-cos2\omega) \tag{2.11}$$

ω is the torsion angle or the angle between the 123 and the 234 planes. Equation 2.10 corresponds to proper dihedrals, which cover the normal torsions in the molecules bonds, while equation 2.11 corresponds to improper ones. Improper dihedrals are meant to keep planar groups (e.g. aromatic rings) in a planar geometry, or to prevent chiral molecules from flipping over to their mirror images.

The second set of energy terms is related to contributions of the non-bonded interaction. These can be further distinguished between electrostatic and Van der Waals interactions. It is worth mentioning that the strength of these interactions decreases with an increase of the distance between atom pairs. The electrostatic interactions may be described for a given atom pair (or between two distinct regions of a molecule) that possess at least a partial atomic charge. Therefore, these interactions may be described with Coulomb's law for the description of electrostatic forces between two point charges.

$$U(r) = \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}, \tag{2.12}$$

where r_{ij} is the distance between a given atom pair, q_i and q_j are the partial atomic charges of each atom, ε_0 is the permittivity in vacuum and ε_r is the dieletric constant of the environment. The Lennard-Jones potential between two atoms can be used to calculate the repulsion and dispersion terms:

$$U(r) = \sum_{i=1}^{N}\sum_{j>i}^{N}(C_{ij}^{12}/r_{ij}^{12}) - (C_{ij}^{6}/r_{ij}^{6}) \tag{2.13}$$

where C_{ij}^{12} and r_{ij}^{12} refer to the interaction parameter and distance between atoms for the short distance repulsion forces, and C_{ij}^{6} and r_{ij}^{6} correspond to the same parameters but for long distance attractive dispersion forces.

Equation 2.7 allows the calculation of the potential energy of every system configuration. It should be noted that this is a fairly simple representation of the potential energy function, and it can be more or less precise, depending on the use of higher order and complexity terms for each of the parameters previously described. Nevertheless, an increase in the complexity of these terms does not readily imply an increased accuracy, but will certainly lead to an increased computational cost. This is a factor that requires careful evaluation, considering the study case.

2.2.2 Force field

After defining the PEF, MD simulations require a set of rules and parameters (which are present in the energy function as variables) to properly describe each system. These parameters are present in a database, usually called force field. The force field can be fully atomistic, which translates into describing every atom explicitly, as suggested by the name. This type of force fields allow for a higher level of detail in terms of atomic motions, while implying higher computational costs. Another type corresponds to the coarse-grained force fields, on which several atoms, or even whole amino acid residues are represented as sphere-like single particles, or beads. This type of representation allows for the study of more complex or larger systems, trading this and lighter computational costs for a lower detail level, generally. The final force field category corresponds to the united-atom force fields, the type that has been adopted in this work. Here, some atoms, usually aliphatic/apolar hydrogen atoms, are collapsed to the heavy atom they are bound, therefore, treated implicitly. This is due to the fact that these atoms usually are not relevant with little contribution to the system properties we are interested.

It should be noted that none of the different force field categories serves as a for-all-purposes solution, since they all have different applications. Besides this, there are also several force fields that, even if belonging to the same category, may be more appropriate for simulating different systems such as DNA, lipid bilayers and proteins. This is due to the fact that each force field is empirically parametrized from small molecules, following different recipes, and then extrapolated for other systems with higher size and/or complexity. This serves to reinforce the idea that each problem requires a carefully thought approach, and, consequently, an adequate force field.

2.2.3 Energy Minimization

One of the major goals of MD simulations is too study the conformational space of a given system. Each of the possible conformations has a given energy value associated, which can be determined using the PEF. Before starting MD simulations, all systems should always be subject to energy minimization. This is a useful procedure to avoid high energy interactions which could lead to numerical instabilities and simulation crashes. Due to the complexity of the energy landscape of a protein system (Figure 1.2), the global energy minimum is very hard to find. Nevertheless, the main objective of the energy minimization procedure is to find a local minimum, which in theory should be stable enough to serve as a starting point for simulations. Some examples of energy minimization methods are the conjugate gradient, the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newtonian, also known as l-BFGS, and the steepest descent algorithms. In this section we will present the algorithm used for this work, the steepest descent.

The steepest descent algorithm is a simple yet robust and efficient energy minimization method. This algorithm, as the name suggests, minimizes the energy in the direction of the highest energy decrease, until it reaches a minimum. This will most likely be only a local minimum and, even if it is not a particularly good minimum, will most likely remain trapped in it. In other context, this could be a large problem, but for our purpose, the steepest descent run is perfect since it corrects all high-energy interactions created in the system setup in a highly efficient computational calculation. This method achieves this by defining a vector $\vec{r}$ for a set of coordinates and a maximum displacement, h_n, and calculating new positions using:

$$\vec{r}_{n+1} = \vec{r}_n + \frac{F_n}{max(|F_n|)} h_n, \tag{2.14}$$

where F_n is the force applied to the atoms and $max(|F_n|)$ is the maximum force applicable. After obtaining the new positions, the potential energy for the new position is calculated and compared to energy of the previous position. If $U_{n+1} < U_n$, the new position is accepted and the maximum displacement is increased, allowing a higher change of positions for consecutive steps. If $U_{n+1} > U_n$, the new position is rejected and h_n is lowered. Hence, as the structure gets closer to a energetic minimum, the displacement gets progressively lower and the new positions differ less from previous ones, leading to convergence to a minimized conformation. This whole process is iterated for a number of steps defined by the user, and stops once that number of steps is reached, or if F_n is higher than $max(|F_n|)$.

2.2.4 Molecular Dynamics

One of the main advantages of MD simulations is the study the conformational space of a given system in a timescale unattainable by most experimental techniques. This is achieved by integrating time over the system forces derived from the potential energy. The force derives from the gradient of potential energy of the system and from the potential energy function:

$$F_i = -\nabla_{ri} U, \tag{2.15}$$

where U is the potential energy function and ∇_{ri} is the gradient of the potential energy calculated for a given set of coordinates. Given that each particle and atom has a mass and that we have the coordinates of every particle in our system, then Newton's Second Law can be applied to obtain the acceleration,

$$\frac{dr_i}{dt^2} = \frac{F_i}{m_i} \tag{2.16}$$

where r_i is the coordinates of atom i, m_i its corresponding mass, and F_i the force applied to the atom at a given instant.

By integrating equation 2.16, velocities and position after dt can be inferred. By iterating this procedure a n number o steps, a trajectory which contains a simulation of the motion of the system is obtainable. However, since the atoms of the system are connected, which implies that the motion of each atom is coupled to the motion of its neighbours, then this leads to a problem impossible to solve analytically. Therefore, the use of numerical methods is necessary and work on two main assumptions, that the force is constant throughout the time step and that the integration step is very small, allowing for precise estimations of accelerations, velocities and positions. One of the commonly used integration methods is the leap-frog algorithm, which uses the following two equations to determine new positions and velocities, respectively:

$$r_i(t+\Delta t) = r_i(t) + v_i(t+\frac{\Delta t}{2})\Delta t, \tag{2.17}$$

$$v_i(t+\frac{\Delta t}{2}) = v_i(t-\frac{\Delta t}{2}) + \frac{F_i(t)}{m_i}\Delta t, \tag{2.18}$$

2.2.5 Temperature Coupling

In order to better simulate the *in vivo* conditions, parameters such as temperature and pressure must be defined. Taking into account that, physiologically, these parameters typically have a very narrow fluctuation range, on our simulations these can be set to a reference value. To achieve this, algorithms must be employed to adjust the values throughout the simulations to match these reference values. For the temperature, the adjustments consist on adding or removing kinetic energy to the system in order to simulate temperature changes. This change in energy translates into changes in velocities, which will impact the system temperature. The temperature deviation from the reference value will define the magnitude of the correction, following the formula:

$$\frac{dT}{dt} = \frac{(T_{ref} - T)}{\tau_t}, \tag{2.19}$$

where T_{ref} is the reference temperature of the thermal bath and τ_t is the coupling constant. With a higher coupling constant, the coupling is weaker and will lead to less frequent temperature adjustments, while with a lower constant, the opposite occurs. In this work, the algorithm used for temperature coupling was the velocity-rescale (v-rescale) thermostat. This works by coupling the system in study with a external thermal bath which can add or remove energy, similar to a Berendsen termostat, but with an additional stochastic term to correct the obtained fluctuations.

2.2.6 Pressure Coupling

Similarly to the temperature treatment, the system pressure also needs to be coupled to a reference value using an algorithm. Here, to adjust the pressure, the volume size is changed with the concomitant changes in the positions of atoms. For this work, the algorithm used for pressure coupling was the Parrinello-Rahman barostat. Analogously to the temperature coupling, a reference pressure value is defined, and the pressure fluctuations will be proportional to the difference between the system pressure and the reference:

$$\frac{dP}{dt} = \frac{P_{ref} - P}{\tau_p}, \tag{2.20}$$

where P_ref is the reference pressure and τ_p is the coupling constant, which functions similarly to the temperature coupling constant.

2.2.7 Periodic Boundary Conditions

Most MD simulations are performed on small finite size system box systems resembling crystallography unit cells. These would be very unrealistic if the boundaries of the box would be modeled with non-interacting walls, leading to artifacts and other phenomena impacting the quality of the results obtained through the simulation. To circumvent this, periodic boundary conditions (PBC) can be used, which consists in creating infinite periodic images of the simulation system (Figure 2.2).

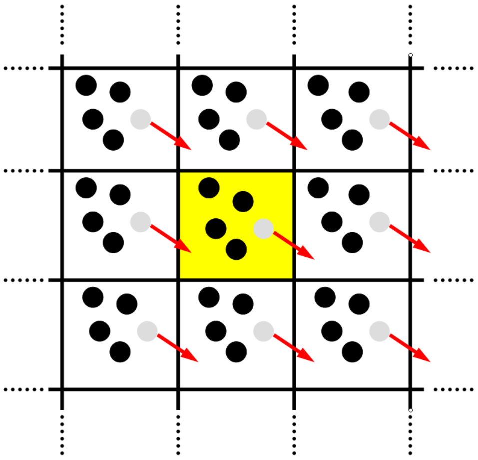


Figure 2.2: Two-dimensional representation of periodic boundary conditions. Adapted from [41].

This allows the system to be simulated in an infinite environment and, when ensuring that the solute does not interact with its periodic image, it models an infinite dilution. This is a realistic representation of most physical systems at lower concentrations. With PBCs, the solute can diffuse out of the unit cell, only to appear on the opposite side of the same box. This phenomena can be handled during the simulations, but when running analysis over the trajectory of a simulation, a PBC correction is usually required. This is particularly important in systems where more than one solute molecule are involved and, if PBC correction is not adequately performed, the system may mistakenly appear to be broken.

2.2.8 Non-bonded interactions treatment

Performing calculations for all possible the non-bonded interactions would have a prohibitive computational cost, in particular, if this would also include the interactions between periodic images. A simple solution to speed up these calculations consists on the definition of the neighbour lists for all atoms, using a cutoff. These lists are updated with a user-defined frequency and all atoms contained in it are treated explicitly, while the atoms outside of the cutoff are ignored (Van der Waals) or treated using a specific method (Coulombic). In this work, a PME method was applied to deal with the long range electrostatic interactions. This method splits the Coulombic interactions into two components, a short distance part (direct space), where explicit calculations are performed, and a long distance part (reciprocal space), which can be solved by representing each particle on a grid, which can then have a Fourier transform applied.

2.2.9 Constraints

One of the main concerns of MD simulations is the computational costs, and by consequence, the time consumed. One of the ways of decreasing these concerns is to increase the time step of the integration. However, as discussed previously, this can lead to numerical instabilities, since fast structural fluctuations, like X–H bonds, are changing at very fast speeds, near the time step. A way of circumvent this limitation and increase the time step is by using constraint algorithms. These tools work by returning the bonds and/or angles to reference values after a integration time step, in order to prevent excessive and unrealistic bond stretching and angle bending due to higher time jumps. Also, note that if the bond and/or angle changes between steps is larger than a user-defined limit, these algorithms give a warning message, since it may be headed to a simulation crash. In this work, two algorithms were used, LINCS, to constraint proteins and peptides, and SETTLE, which was used to do the same task for water molecules.

2.2.10 Position and Distance Restraints

In our simulations, position and distance restraints were used to hold the solute in space or to keep two atoms from going apart. Restraints work by applying an energy penalty proportional to a constant defined by the user. In the case of position restraints, this penalty is applied when specific atoms, which can be defined by the user, change their position. In the case of distance restraints, this is applied when a distance between two groups of atoms becomes higher than a threshold defined by the user. These types of restraints are commonly used during the initialization process. This allows the solvent to adjust to the system and also for a smooth equilibration of the system to the target temperature and pressure conditions without causing artificial structural deformations. Examples of applications of distance restraints are for instance keeping together the contact surfaces of protein-ligand or protein-protein interactions, or to maintain a specific portion of a secondary structure element during equilibration steps.

2.3 MD settings

For this work, several distinct MD simulations were employed. All were done using GROMACS v.2018.6 [42, 43] and the GROMOS 54a7 force field [44, 45]. Firstly, simulations using S100B in its ion-bound state were done. The initial dimer was the human S100B crystal structure, available in the Protein Data Bank (PDB ID: 3D0Y [27]). This structure was striped of waters, had the initial residue of the first chain, an N-formylmethionine, mutated to a methionine, and the two final residues of each monomer, which were missing, added. Test simulations with the $A\beta_{25-35}$ were also done, with its initial structure being built with PyMOL [46], in a helical conformation. After this process, both systems were built with a dodecahedral box, with 0.9 nm for S100B and 1.2 nm for $A\beta$ between protein and box limits. These were then solvated with $\sim$9900 SPC waters [47] for S100B and $\sim$1800 for $A\beta$, and had ions added to achieve charge neutrality, 16 Na^+ for S100B and 1 Cl^- for $A\beta$. Both these setups were then submitted to a two-step minimization procedure, both steps with the steepest descent minimization algorithm, with constraints on all atoms turned on only for the second one. The LINCS algorithm [48] was used to constraint all protein bonds, while the SETTLE algorithm [49] was used to constraint all water molecules.

After the minimization procedure, each system was divided in 3 identical replicates, and these were submitted to a three-part initialization procedure. In the fist part, with 100 ps, a v-rescale heat bath [50] was turned on at 310 K with separate couplings for the protein and the solvent (water and ions) with a relaxation time of 0.02 ps exclusively for this first step, and 0.1 ps for the remaining simulations. Also, initial velocities were randomly generated using the Maxwell velocity distribution at 310 K exclusively for this step, on the following ones, the velocities are taken from the last step. On the second part, with 200 ps, additionally to the v-rescale heat bath, the Parrinello-Rahman isotropic pressure coupling [50–52] was used to keep the pressure at 1 bar, with isothermal compressibility of $4.5\text{x}10^{-5}$ bar^{-1}, and a relaxation time of 0.5 ps. The final step has the same parameters as the previous one. Additionally, each step has position restraints with decreasing force, of 1000, 100 and 10 $\text{kJ.mol}^{-1}\text{.nm}^{-2}$ for the first, second and third initialization steps, respectively. After the initialization procedure, the systems were submitted to MD production. The equations of motions were integrated using a 2 fs time step, with the lists of neighbors being updated every 10 ps. The Particle-Mesh Ewald (PME) electrostatics [53] was applied using 0.12 nm for the maximum grid spacing of the Fast Fourier Transform and a cutoff distance of 1.4 nm for Lennard-Jones and Coulomb interactions. The interpolation order for PME was 4. These parameters were applied to the systems. The S100B dimer ran for 200 ns, while the $A\beta_{25-35}$ system ran for 100 ns.

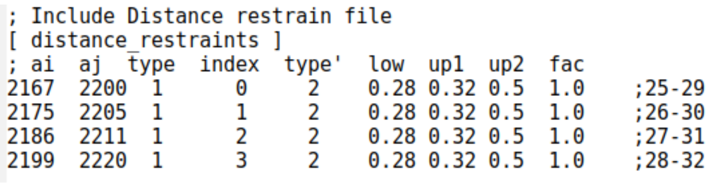
After obtaining and selecting several configurations for the S100B-$A\beta_{25-35}$ complex (see below, Section 3.2 and Figure 2.4), the peptide segment was reconstructed to form $A\beta_{42}$. For the Binding Modes 1 and 2, this was achieved by using PyMOL to build the remainder of the peptide in a helical conformation, adjusting the direction of the added regions to optimize the complex volume, hence, the system size. For Binding Modes 3 and 4, this was achieved by using the pair fitting feature of PyMOL starting from an experimental NMR structure of the peptide (PDB ID: 1Z0Q) [22]. In the Binding Mode 5, this was unnecessary since the Aβ structure provided to Haddock was already the experimental NMR structure of the full peptide. The resulting structures were then submitted to a minimization, initialization and production MD protocol identical to the one previously described.

The five S100B-Aβ systems had the initial methionine from both chains of S100B removed. After this, the systems were built with a dodecahedral box, with ~1.0 nm to 1.2 nm between protein and box limits (Table 2.3). These were then solvated with ~ 15000 to ~ 18000 SPC waters, and had 19 Na$^+$ added to achieve charge neutrality. These setups were then submitted to a two-step minimization procedure, both steps with the steepest descent minimization algorithm, with constraints on all atoms turned on only for the second one. The LINCS algorithm was used to constraint all protein bonds, while the SETTLE algorithm was used to constraint all water molecules.

Table 2.3: Summary table of the setup parameters which change between Binding Modes.

Configuration	1	2	3	4	5
Distance to box (nm)	1.2	1.0	1.2	1.1	1.2
Solvent molecule #	15000	15000	18000	18000	16000
Na$^+$ counter ions #	19				

After the minimization procedure, the systems were submitted to a three-part initialization procedure. In the fist part, with 100 ps, a v-rescale heat bath was turned on at 310 K with separate couplings for the protein and the solvent. In the Binding Mode 4 we performed a smoother heating protocol with a 250 K intermediate simulation, prior to the final 310 K equilibration. The relaxation time of 0.02 ps was used exclusively for the first step, while 0.1 ps was used for the remaining simulations. Also, initial velocities were randomly generated using the Maxwell velocity distribution at 310 K exclusively for this step, on the following ones, the velocities are taken from the last step. On the second part, with 200 ps, additionally to the v-rescale heat bath, the Parrinello-Rahman isotropic pressure coupling was used to keep the pressure at 1 bar, with isothermal compressibility of 4.5x10^{-5} bar^{-1}, and a relaxation time of 0.5 ps. The final step has the same parameters as the previous one. Additionally, each step has position restraints with decreasing force, of 1000, 100 and 10 kJ.mol^{-1}.nm^{-2} for the first, second and third initialization steps, respectively. After the initialization procedure, each system was then submitted to a pre-equilibration procedure, which consisted on a MD production run with distance restraints (Figure 2.3) on the helical segment of residues 25-35 of Aβ. This was done to perform long equilibrations on Binding Modes 1 and 2 and remove the bias of building the peptide in a complete helical conformations, while still maintaining the 25-35 region intact and interacting with the S100B protein. More importantly, these runs were also done to improve the stability of the complex interface. These runs were of 200 ns for the configurations 1 and 2, and 100 ns runs for the remaining configurations. After this pre-equilibration, the three final frames, each from one of the three final blocks of this procedure, were used as the starting point for MD production runs, therefore forming three replicates for these MD runs. The equations of motions



```
; Include Distance restrain file
[ distance_restraints ]
; ai   aj   type  index  type'  low  up1  up2  fac
2167  2200  1      0      2     0.28 0.32 0.5  1.0     ;25-29
2175  2205  1      1      2     0.28 0.32 0.5  1.0     ;26-30
2186  2211  1      2      2     0.28 0.32 0.5  1.0     ;27-31
2199  2220  1      3      2     0.28 0.32 0.5  1.0     ;28-32
```

Figure 2.3: Distance restraints applied to residues 25-35 of Aβ peptide in order to keep its secondary structure

were integrated using a 2 fs time step, with the lists of neighbors being updated every 10 steps. The Particle-Mesh Ewald (PME) electrostatics was applied using 0.12 nm for the maximum grid spacing of the Fast Fourier Transform and a cutoff distance of 1.4 nm for Lennard-Jones and Coulomb interactions. The interpolation order for PME was 4. These parameters were applied to both systems. Each of these simulations ran for 1 μs.

2.4 MD Analysis

2.4.1 RMSD

The Root Mean Square Deviation (RMSD) is a fairly common method of evaluating the similarity between two structures, one of which being the reference structure. This method consists off, firstly, superimposing the atomic coordinates by trying to fit the two structures by means of translations and rotations. This is then followed by applying the following calculation:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}, \tag{2.21}$$

where N is the total number of atoms of each structure and δ_i^2 is distance between a given atom and its equivalent in the reference structure. These calculations can be done for specific groups of atoms, not necessarily the whole protein, like in our case, where we performed the calculations only for the Cα of the protein/peptide. The main chain atoms are usually selected since they usually fluctuate their positions less when compared to the side chains of the amino acid residues. All RMSD calculations were done using the Gromacs 2018.6 package.

2.4.2 SASA

The Solvent Acessible Surface Area (SASA) is a measure of the area of the protein surface which is exposed to the solvent. This may be useful to detect and follow conformational changes which involve changes in solvent exposure on a given area or specific residues of a protein. The values obtained using this calculation correspond to the area values drawn by the center of the spherical probe, as it passes through the surface of the protein. Similarly to RMSD, these calculations can be done for the whole protein or for specific groups of aminoacid residues. By doing these calculations for each residue separately, it was possible to obtain a hydrophobicity/hydrophilicity index for each residue [54]. This can be calculated by summing the product between the solvent-accessible surface area and the hydrophobicity (Wimley–White) scale [55] of each residue. The information can then be averaged out over time and plotted per residue or can be averaged out over residue and plotted over time. It is also possible to estimate the interfacial area of the S100B-Aβ complexes, by calculating the SASA of the S100B protein in a complex both in the presence and absence of the Aβ partner. A difference between these 2 values in each conformation will provide the desired interfacial area. All SASA calculations were done using the GROMACS 2018.6 package.

2.4.3 DSSP

The DSSP program calculates the secondary structure content of a protein or of a group of amino acid residues according to a set of criteria. This program calculates angles, distances, and hydrogen bond energies, from the atom coordinates, and assign each residue to a secondary structure type. With this analysis, we can follow the protein/peptide structure stability over time and easily identify the more drastic conformational changes. All DSSP calculations were done using the GROMACS 2018.6 package.

2.4.4 Gyration Radius

The radius of gyration provides a measure of how compact a protein is, which is usually related to its volume. This information complements the data obtained by SASA and RMSD and is important to assess the equilibration in MD simulations. The computation is done by first determining the center of mass of the protein. After that, the following calculation is applied:

$$R_g = \sqrt{\frac{\Sigma_i ||r_i^2|| m_i}{\Sigma_i m_i}}, \tag{2.22}$$

where r_i is the distance between a given atom and the center of mass, and m_i is the mass of the atom. All gyration radius calculations were done using the GROMACS 2018.6 package.

2.4.5 MM-PBSA Binding Free Energy

The best way to assess the stability between two interacting macro molecules is to calculate their binding free energy (ΔG). This property is particularly useful to evaluate different binding modes, since, due to error cancellation, the relative free energies will be more reliable than absolute values. This can be very useful in this particular work, and will help us rank the different binding modes tested for the S100B-Aβ_{42} complex.

Amongst the several methods that exist to estimate the ΔG, we opted to use one of the most popular methods, the Molecular Mechanics with Poisson-Boltzmann and Surface Area solvation (MM/PBSA) method [56]. It is worth noting that this particular method was originally developed in the late 90s by Kollman *et al.* [56] and since then, has seen many different applications and gone through several adjustments. This approach lacks consensus in terms of details, since the performance of these many variants seems to depend on the systems where it is applied [56]. As such, in this section we will address the specific method used for this work, as well as its approximations.

In MM/PBSA method, the binding free energy is estimated from the following relation in which both the complex and the individual molecular entities are considered:

$$\Delta G_{bind} = G_{complex} - (G_{S100B} + G_{A\beta_{42}}) \tag{2.23}$$

Each of the free energy components previously presented ($G_{complex}$, G_{S100B}, and $G_{A\beta_{42}}$) is estimated from the following equation:

$$G_x = < E_{MM} > -TS + < G_{solv} > \tag{2.24}$$

where G_x corresponds to any of the free energy components, $< E_{MM} >$ corresponds to the molecular mechanics energy term in vacuum, TS corresponds to the entropic contribution and $< G_{solv} >$ corresponds to the solvation contribution. The $<>$ symbols represent averages over several conformations. One of the major approximations of this method is to assume that the entropic contributions will cancel out, which might not be realistic for many systems. Nevertheless, some approaches to include these contributions in the MM-PBSA calculation were not successful [56], which only supported the most simple approach of ignoring it. In systems where the degrees of freedom do not change significantly between the complexed and its free forms, this might be the best decision.

The $< E_{MM} >$ energy term can be further broken down into its components:

$$E_{MM} = E_{bonded} + E_{nonbonded} \tag{2.25}$$

where E_{bonded} is respective to a energetic term associated with bonds, angles and dihedrals, while $E_{nonbonded}$ refers to the Coulombic and Van der Waals interactions. These parameters can be calculated using a PEF and a force field (Section 2.2.1). However, it is worth noting that the calculation of some of these energy terms (the self terms) are not necessary, since they cancel out between free and complex forms.

$$E_{MM} = (E_{bonded} + E_{nonbonded})_{self} + (E_{bonded} + E_{nonbonded})_{cross} \tag{2.26}$$

The $E_{bonded/cross}$ contribution is zero since there are no physical bonds between partners, while the $E_{bonded/self}$ will cancel out, as the bonded parameters do not change with complexation. Since the $E_{nonbonded/self}$ are also not affected by the complex formation, they will also cancel out, which leaves only $E_{nonbonded/cross}$ to be calculated (we named it $E_{nonbonded}$, for simplicity).

$$E_{nonbonded} = E_{vdw} + E_{Coul} \tag{2.27}$$

The E_{vdw} corresponds to the apolar term calculated using the Lennard-Jones potential part of the PEF, and E_{Coul} corresponds to the polar part, calculated using a Coulomb potential.

The solvation energy (G_solv) can broken down into:

$$G_{solv} = G_{polar} + G_{apolar}, \tag{2.28}$$

which corresponds to the polar (G_{polar}) and apolar (G_{apolar}) parts of the electrostatic energy of changing the solute from vacuum to a solvent (water in our case). The polar part is calculated by solving the Poisson-Boltzmann equation. The G_{apolar} term can still be further broken down into two distinct terms.

$$G_{apolar} = G_{cavity} + G_{vdw} \tag{2.29}$$

where G_{vdw} refers to the energy associated with Van der Waals interactions with water and the G_{cavity} accounts for the energy associated with creating a space/cavity in the solvent. Some approaches require this separation, while other (like the one used in this work) estimate the full G_{apolar} term from multiparametric calibrations using the SASA and/or volume differences upon complex formation [57]. In this work, we used a GROMACS external tool g_mmpbsa [57] to calculate all these contributions of the MM-PBSA method.

2.4.6 Statistical Analysis

All analysis which involved the average of a given measure, presented with error values, were estimated using standard error of the mean (SEM) method. In our work, the formula used to estimate the SEM is:

$$SE = \left(\frac{n-1}{n} \sum \left(\frac{\bar{x}_i - \bar{x}}{2} \right)^2 \right)^{1/2}, \tag{2.30}$$

where n, as previously, is the number of replicates, $\bar{x}_i$ is the average of each subset combination of replicates and $\bar{x}$ is the average of all replicates.

2.5 Molecular Docking

As previously mentioned, for MD simulations an initial structure of the system in study is required. Since we only had available the structures of metal-bound S100B protein and Aβ_{42} individually, not in complex, then a molecular docking procedure was required. The main goal of docking techniques is to sample the configuration on which two or more molecules interact and bind. These techniques can usually be divided in two components, the method used to sample several conformations for the binding phenomena and the scoring function, used to evaluate and qualify the configurations obtained through the sampling method [58].

The sampling algorithm can be, for example, geometry/spatial features based, where the program calculates several distances between the atoms of the different molecules to dock, to create a matrix, which is used to generate different configurations. Another approach is to use a incremental construction algorithm, which separates one of the molecules into different fragments. Then, one of these is selected, typically the largest or the one with a important group for the interaction to be docked to the other molecule. After this, the remaining fragments are incrementally added. This allows the docking procedure to test different orientations of the fragmented molecule, mimicking its flexibility. Yet, other sampling methods could be stochastic based ones, such as Monte Carlo or Molecular Dynamics simulations, which allow for flexibility of the molecules to generate different configuration. Here, it is worth noting that even though these procedures represents flexibility in a more robust way, in the case of MD the simulations, they are very time consuming and may not be able to pass energy barriers, which may compromise the sampling. This latter problem is not shared by Monte Carlo methods.

The scoring functions are usually based on force fields and complemented with empirical data [58]. They are used as an energy function to calculate the different binding energy terms and estimate the complex binding affinity. Due to the fact that several approximations are normally employed, the scoring functions may have performance issues which stem from not properly dealing correctly with some factors, such as the solvation effects. To improve this, methods such as MM-PBSA or MM-GBSA can be used to perform rescoring and do a more robust analysis off the binding affinities [58].

The system flexibility is also an important aspect to docking procedures. Depending on the type of interactions which are required for the binding phenomena, and whether effects such as induced-fit and allosteric modulation are expected to take place, the introduction of flexibility to one or more molecules may be required [58]. The main disadvantage of adding flexibility to a docking calculation is that we are adding more degrees of freedom, which will make the complete sampling of the system more computational intensive, if not impossible. When dealing with a peptide/protein or a protein/protein docking problem, this issue becomes highly problematic and conventional docking fails. For this particular reason, the HADDOCK webserver was used to perform molecular docking for this work.

2.5.1 HADDOCK

The HADDOCK webserver is an online docking server which allows the users to run information-driven docking procedures. The HADDOCK webserver possesses different interfaces with distinct levels of access. The easy interface, available to all users, the expert interface and the guru interface, which both require the user to make a request to access these interfaces. The main difference between these three interfaces is an increasing level of control over the parameters of a given docking run, presented in a increasing order of control. There is also a fourth interface which allows the user to submit a docking run by uploading a single file to the webserver [59–61]. For this work, only the easy and expert interfaces were used.

The HADDOCK docking procedure consists of three distinct steps, named it0, it1 and water. The first step consists of a randomization stage, on which the molecules are separated by at least 25 Å and are randomly rotated around their center of mass and are translated within a 10 Å cube, in order to randomize the starting orientations. This is followed by a rigid body energy minimization by freezing all molecular bonds, angles and rotations around bonds. From this procedure, 1000 model results are obtained, which are saved to the disk. This procedure is repeated several times, 5 by default. Their respective HADDOCK scoring function, which results from the application of the HADDOCK scoring function, is also calculated. From these structures, typically the best 200 are selected for the following steps, it is noteworthy that these following steps do not have a drastic effect on the HADDOCK score, and are only used to refine the structures of the solutions.

The second step consists of a semi-flexible refinement using a high temperature MD simulation in torsion angle space, which by increasing the flexibility of the structures, improves the interaction interface. This is done in four stages, a track of high temperature rigid-body dynamics, a rigid body annealing, annealing with flexible side-chains on the interface and annealing with flexible backbone and side-chains on the interface. The flexible regions are defined for all residues within 5 Å of another molecule, on the interface. It is typical that all the solutions used on this step continue to the next and final one.

The third and final step consists of a flexible refinement in explicit solvent, TIP3P water. The solutions are solvated in a 8.5 Å solvent shell and submitted to molecular dynamics simulations in Cartesian space, only with weak restraints on backbone atoms outside the interface, which is followed by a final minimization.

The HADDOCK score is calculated for each of the previously described steps, and several terms have different weights on different steps:

$$E_{it0} = 0.01E_{air} + 0.01E_{vdw} + 1.0E_{elec} + 1.0E_{desolv} - 0.01BSA, \tag{2.31}$$

$$E_{it1} = 0.1E_{air} + 1.0E_{vdw} + 1.0E_{elec} + 1.0E_{desolv} - 0.01BSA, \tag{2.32}$$

$$E_{water} = 0.1E_{air} + 1.0E_{vdw} + 0.2E_{elec} + 1.0E_{desolv}, \tag{2.33}$$

where E_{air} is the ambiguous interaction restraints energy component, E_{vdw} is the intermolecular Van der Waals energy term described by a 12-6 Lennard-Jones potential, E_{elec} is the intermolecular electrostatic energy term described by a Coulomb potential, E_{desolv} is an empirical desolvation energy term, and BSA is the buried surface area, in Å^2. It is worth noting that the non-bonded energies are calculated using a 8.5 Å cutoff, using the OPLS force field parameters [59, 62, 63].

2.5.2 HADDOCK Settings

In this work, several docking runs using the HADDOCK webserver were used in order to obtain complex S100B-Aβ_{42} structures. Due to the vast conformational variability of the peptide, we started by only using the peptide segment relative to which we have experimental evidences. Four distinct docking runs led to structures with reasonable complex configurations. The first of these (Trial 1) was done using the Easy Interface of HADDOCK webserver 2.2. In this run, a helical conformation of Aβ_{25-35} was used, selecting all 11 residues of the segment as active residues. The crystal structure of S100B was also used, with the residues experimentally identified to be a part of the binding, marked as active residues (Table 2.4).

Table 2.4: Haddock active residues defined as input for each calculation. Passive residues were always defined automatically. Note that the residue numbering changes between Aβ_{25-35} (trials 1–3) and Aβ_{42} (trial 4).

Trial	S100B active residues	Aβ active residues
1	62, 71, 74, 75, 78, 79, 82, 83, 171, 172, 175, 178, 179, 182, 186	1-11
2		4, 7
3		4, 7
4		28, 31

In both cases, passive residues were assigned automatically. After building the complete Aβ_{42} from the 25-35 segment, as previously described, this solution led to binding modes 1 and 2.

The third and fourth complex configurations were obtained from docking runs using the Expert Interface (Trials 2 and 3), with the same parameters except for the input Aβ_{25-35} segment structure which differs between these two runs. Here, only residues Lys28 and Ile31 of the peptide were marked as active residues. Also, these calculations had increased sampling parameters, namely the number of structures for rigid body docking, number of trials for rigid body minimization, number of structures for semi-flexible refinement and number of structures for explicit solvent refinement all increased to 10000, 50, 1000 and 1000, respectively. The selection of the configurations was based on both the electrostatic energies and the haddock scores.

Finally, the fifth complex configuration was obtained from a docking run (Trial 4) with all parameters identical to the Trials 2 and 3, with the exception that the previously input Aβ_{25-35} segment was now substituted with the complete structure of the Aβ_{42} (PDB ID: 1Z0Q [22]). Therefore, the active residues had their number also replaced by the respective numbering on the full peptide. The final solution was selected based on the electrostatic energies and haddock scores.

Figure 2.4 summarizes all systems studied and provides an illustrative view on the binding modes and complex configurations generated by Haddock and the MD simulations. We opted to show here this scheme to help the reader to better understand the workflow of the upcoming Results section.

Figure 2.4: Scheme with the molecular representation of Aβ peptides and S100B protein used throughout this work. Each row correspond to a different binding mode (BM) and the different columns correspond to the Haddock and MD inputs/output configurations. The Aβ peptides are shown in cartton representation, while the S100B structures are shown as grey surface. The MD input column corresponds to the Haddock solutions with the peptide reconstructed, when necessary, which were submitted to initialization and pre-equilibration MD stes. The MD production start column shows the initial configurations for production run of each of the replicates (purple, green and blue). The MD production end corresponds to final frame of each simulation, at 1 μs.

Chapter 3

Results and Discussion

3.1 MD simulations of free S100B and Aβ_{25-35} peptide

Initial simulations were performed using either the S100B protein in the Ca^{2+} and Zn^{2+} bound state (PDB ID: 3D0Y) or the free Amyloid-beta 25-35 (Aβ_{25-35}), built in a helical conformation. The metal-coordinated S100B was simulated for 200 ns to attest its overall stability, in particular, the quality of the metal center parameters derived from Quantum Mechanics. The 100 ns MD simulations of the Aβ_{25-35} peptide fragment were performed to check the stability of the helical structure, when the peptide is free in solution. There are experimental data hinting that the Aβ peptide should have some helicity content when bound to S100B [15]. Therefore, we used these simulations to evaluate the helix propensity of the shorter fragment, which will also be used in this work.

Several structural parameters calculated for the S100B system showed that the system is stable and that the simulations converged within the 200 ns runs (Figure 3.1). We note that the RMSD and the gyration radius values converged to plateaus ~ 0.27 nm and ~ 1.7 nm, respectively, for all the replicates. For the gyration radius in particular, replicates 1 and 3 showed a more significant adjustment between the initial and the final converged values, which often happens with initial structures obtained from X-ray crystallography. These experimental structures are sometimes packed in conformations which are not representative of the protein in solution. Most structural properties will capture this equilibration process differently. An initial increase in the RMSD is expected, to capture the structural deviation from the crystal structure. Also, a small loss of helical content can also be expected, which was really not observed in our system. Finally, the observed variations in the radius of gyration are also typical of this process, which sometimes take longer equilibration times. All these results attest the strong stability of S100B protein in solution. Additionally, we also calculated the asymmetry RMSD values between the two monomers of S100B. This property quantifies the conformational deviations between the two subunits and capture conformational transitions when they happen only on one side of the dimer. However, the relatively low values of $RMSD^{Sym}$ confirm that there is little asymmetry in this protein. From these results, we considered that equilibrium is well achieved after the first 50 ns of simulation, which were discarded when calculating equilibrium properties.

A similar equilibration analysis was also calculated for the simulations using the Aβ_{25-35} peptide. Taking into consideration that the system simulated in these simulations was a small peptide, only the secondary structure and the gyration radius will provide meaningful information (Figure 3.2).

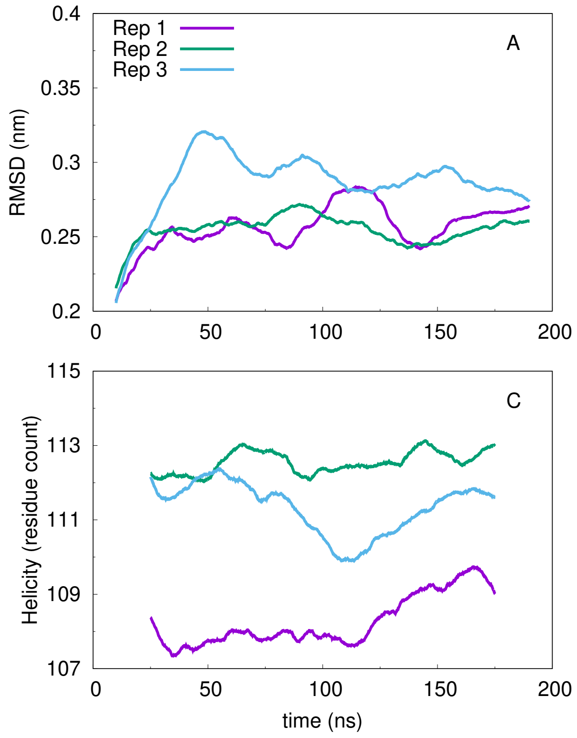


Figure 3.1: Structural properties from MD simulations of free S100B ion-bound dimer. The properties include: RMSD (A), gyration radius (B), helicity (C) and monomer asymmetry (D). A floating window of 20 ns was used for RMSD and gyration radius, while for monomer asymmetry and helicity a floating window of 50 ns was used in order to reduce the local fluctuations of each data series.

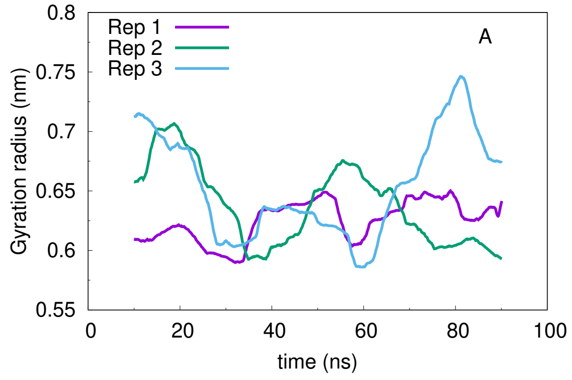


Figure 3.2: Structural properties properties from MD simulations of free Aβ_{25-35} peptide fragment. The properties include: the gyration radius (A) and Helicity (B).A floating window of 20 ns was used for Gyration Radius, while for Helicity a floating window of 10 ns was used in order to reduce the local fluctuations of each data series.

The gyration radius values showed a few large fluctuations but, since we expected a significant loss of the secondary structure, we focused our attention to the helicity data. Indeed, in all three simulation replicates the peptide completely looses its helical structure after $\sim$ 20 ns. Nevertheless, as previously mentioned, these simulations were done to confirm that the Aβ_{25-35} peptide would require a partner to be able to stabilize its helical structure. Furthermore, the initial steps of these MD simulations, while still in non-equilibrium, we sampled several meta-stable conformations where the peptide exhibited a helical core. These were subsequently used in the S100B docking procedure using the HADDOCK webserver.

3.2 The binding modes between S100B and Aβ

As described previously in chapter 2.5.1, the online platform Haddock was used to obtain configurations of S100B bound to Aβ_{42} (PDB ID: 1Z0Q [22]) or Aβ_{25-35} peptide. For this procedure, we used a partially helical Aβ_{25-35} structure, which was selected from our equilibration MD simulations. In this peptide conformation, only the central core of the fragment are fully helical. We have also considered performing the HADDOCK runs using a selected S100B structure from MD simulations. However, since the X-ray structure seemed very representative of the MD ensemble, we deemed it unnecessary for now and adopted the experimental structure. There are clear advantages of using a smaller fragment of Aβ_{42}, like the 25-35, to find the best binding modes to S100B. In particular, since the experimental data at our disposal provided clues about the significant role of Lys28 and Ile31 residues [15], which are included in the shorter peptide. Furthermore, taking into consideration the complex conformational space of a larger peptide, such as the Aβ_{42}, this would result in a significantly more difficult task for the computational docking protocol. Therefore, in our docking procedure, we initially used only the short fragment, aiming at solutions from which we can expand and rebuild the complete Aβ_{42}.

Our first Haddock calculation using the Aβ_{25-35} interacting with S100B resulted in a series of solutions, two of which located in the expected binding cavity (Figure 3.3). Since S100B is a homodimer, these two clusters (7 and 9) are completely symmetric, hence, we proceeded using only cluster 7 to obtain the first two binding mode configurations (BM1 and BM2). Interestingly, this cluster also showed a very high contribution of its electrostatic energy component (Table 3.1).

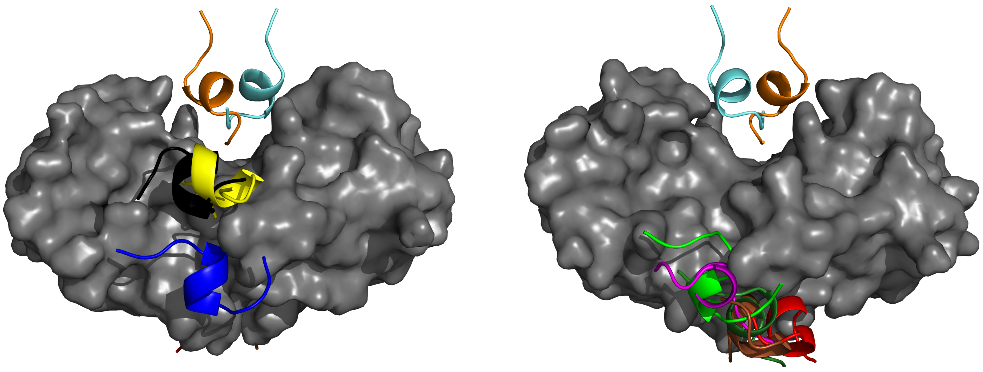


Figure 3.3: Structural representation of the Haddock trial 1 solutions. The Aβ_{25-35} peptide structures are shown in cartoon colored in different colors. The colors red, green, blue, yellow, magenta, light blue, orange, dark green, brown, and black follow the numeric order of the clusters in Table 3.1.

Table 3.1: Haddock scores and energy results for trial 1. The clusters were ordered by their Electrostatic energy. This Haddock run originated two seemingly symmetrical solutions, clusters 7 and 9. Due to the apparent symmetry, only one of these was used. Cluster 7 originated binding modes 1 and 2.

Cluster	Haddock Score	E_{elec}	E_{VdW}
9	**-43**	**-138**	**-16**
7	**-52**	**-134**	**-21**
20	-50	-91	-22
1	-59	-90	-24
2	-64	-64	-28
5	-70	-59	-31

Since the available experimental data hints at a significant interaction by Lys28 of Aβ_{42} with S100B [15], possible ionic interactions could play a role. In fact, it was expected that this particular residue could establish electrostatic interactions with negatively charged residues (Figure 3.4). The first two binding modes were created by rebuilding the remaining parts of Aβ_{42} in a helical configuration in two different directions, one pointing downwards towards S100B and the other one pointing upwards, more towards the solvent, and then manually moved and adjusted to optimize the size of the simulation box, therefore, optimizing the computational costs within reason, as well as try to sample distinct regions of the conformational landscape.

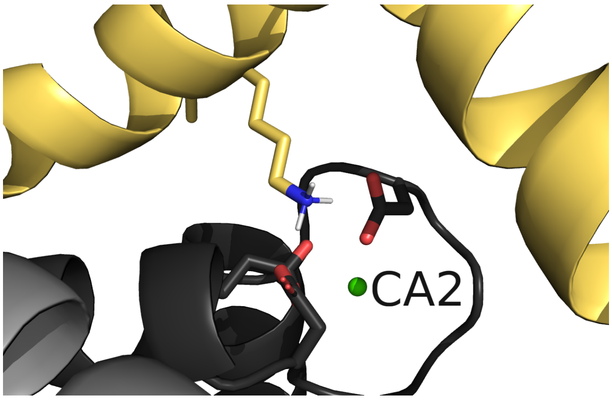


Figure 3.4: Structural representation of the ionic interaction between Lys28 of Aβ and an aspartate residue of S100B, which is coordinated with a calcium ion. This configuration corresponds to the BM1, where the Aβ_{42} was rebuilt from cluster 7 of Trial 1 haddock calculation.

Our second and third docking calculations had increased sampling parameters when compared to the first calculation, as described in Section 2.5.2. The Aβ_{25-35} peptide was used with the same S100B experimental structure and many solutions were obtained, where only one, from the second calculation, was not located in the expected binding cavity (Figures 3.5 and 3.6). Therefore, to select the best binding modes, we used the Haddock cluster score and its electrostatic energy component (Tables 3.2 and 3.3). The binding modes 3 and 4 (BM3 and BM4) were obtained from cluster 1 and cluster 6 from the second and third calculation, respectively. It should be noted that in all runs using the Aβ_{25-35} peptide, some cluster solutions were discarded due to clear steric clashes in the final rebuilt Aβ_{42}. This was done mainly through visual inspection, by evaluating if the resulting solutions had a reasonable orientation to accommodate the remaining of the peptide. Many solutions were indeed compromised (Figure 3.7) and were ignored, regardless of their docking scores and electrostatic energy component.

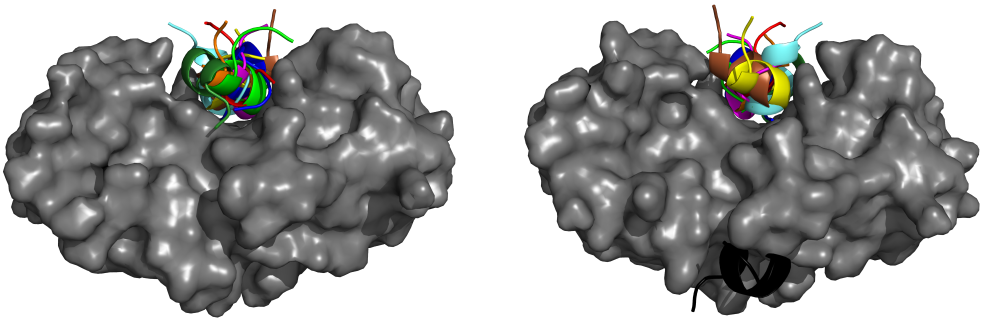


Figure 3.5: Structural representation of the Haddock trial 2 solutions. The Aβ peptide structures are colored following the color code from Figure 3.3 and the numeric order of the clusters in Table 3.2, respectively.

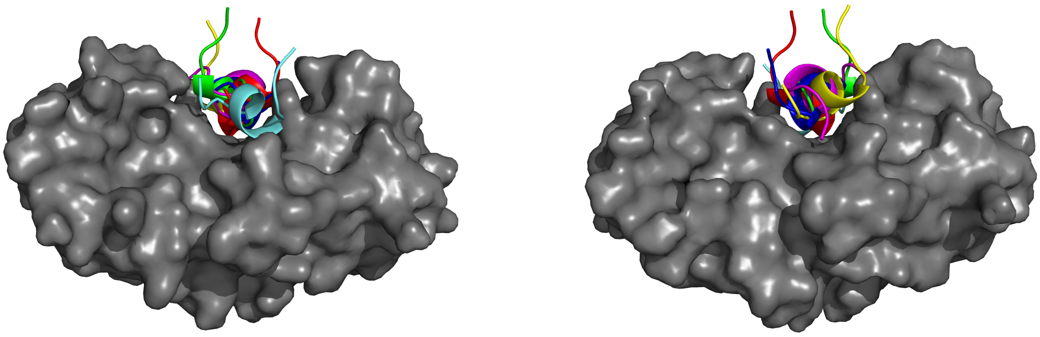


Figure 3.6: Structural representation of the Haddock trial 3 solutions. The Aβ peptide structures are colored following the color code from Figure 3.3 and the numeric order of the clusters in Table 3.3, respectively.

Table 3.2: Haddock scores and energy results for trial 2. The clusters were ordered by their Electrostatic energy. Cluster 1 originated binding mode 3.

Cluster	Haddock Score	E_{elec}	E_{VdW}
13	-48	-127	-28
1	**-56**	**-109**	**-29**
5	-48	-99	-28
7	-50	-94	-30
3	-57	-88	-36
4	-58	-86	-31
17	-61	-73	-30
2	-48	-48	-32
6	-49	-47	-33
10	-44	-41	-27

Table 3.3: Haddock scores and energy results for trial 3. The clusters were ordered by their Electrostatic energy. Cluster 6 originated binding mode 4.

Cluster	Haddock Score	E_{elec}	E_{VdW}
6	**-46**	**-164**	**-23**
4	-35	-119	-21
5	-43	-82	-28
1	-67	-68	-38
2	-60	-64	-33
3	-55	-58	-35

In our fourth and final docking calculation with S100B, we used a complete Aβ_{42} structure and obtained several solutions in the expected binding cavity (Figure 3.8). Most of these solutions could have been selected, since they show good Haddock scores and significant electrostatic energy contributions (Table 3.4). Due to computer limitations, we opted simply to select the solution with the best electrostatic energy contribution to the binding. Therefore, the fifth binding mode (BM5) was obtained directly from cluster 16 of the Haddock run.

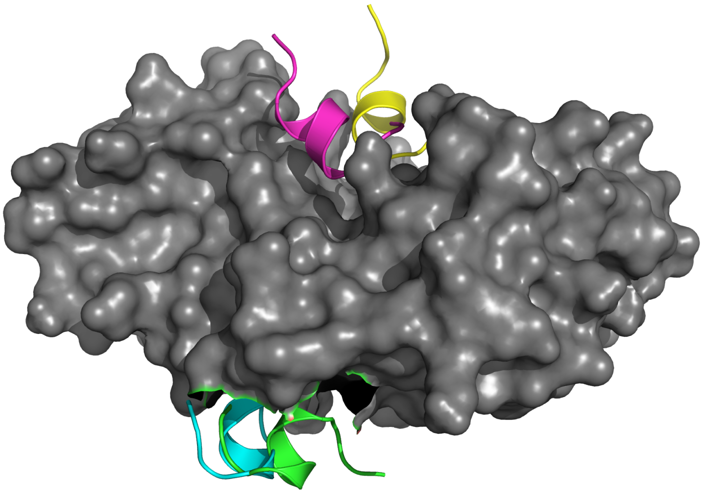


Figure 3.7: Structural representation of several Aβ_{25-35}:S100B complex Haddock solutions which were discarded after visual inspection. S100B displayed in a gray surface. The Aβ_{25-35} peptide solutions are shown in cartoon with colors. Clusters 1 and 20 from trial 1 are displayed in green and blue, respectively.These two solutions were docked in a region for which there is no experimental evidence supporting it. Clusters 13 (trial 2) and cluster 4 (trial 3) are represented in magenta and yellow, respectively. Despite being in the correct region of S100B, these solutions lack a proper orientation for rebuilding the final Aβ_{42} peptide.

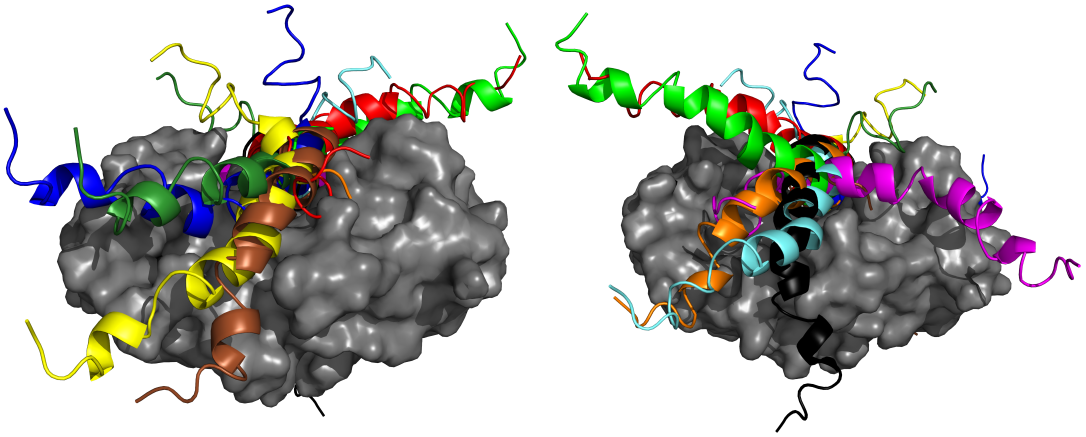


Figure 3.8: Structural representation of the Haddock trial 4 solutions. The Aβ peptide structures are colored following the color code from Figure 3.3 and the numeric order of the clusters in Table 3.4, respectively.

Table 3.4: Haddock scores and energy results for trial 4. The clusters were ordered by their Electrostatic energy. Cluster 16 originated binding mode 5.

Cluster	Haddock Score	E_{elec}	E_{VdW}
16	**-120**	**-235**	**-60**
13	-127	-165	-60
22	-88	-161	-44
7	-87	-139	-41
14	-94	-130	-46
8	-97	-120	-48
5	-94	-107	-46
2	-94	-106	-48
1	-86	-106	-49
10	-92	-50	-44

For a summary view of all Haddock solutions used and how they resulted in the 5 binding modes to be explored in the next section using MD simulations, please check out Figure 2.4 at the end of the Methods section.

3.3 MD simulations of the S100B:Aβ complex

3.3.1 Equilibration of Aβ helicity

After obtaining and choosing the starting binding modes described previously, the five systems were then set up and initialized as described in Section 2.3. After setup and initializing the system, a long pre-equilibration run was performed (200 ns for BM1 and 2; 100 ns for BMs 3–5), using distance restraints holding the α-helix hydrogen bonds in the 25-35 segment of the Aβ_{42} peptide. The experimental data available pinpoints Lys28 and Ile31 as the two main residues involved in the S100B binding (Figure 1.7). The 3-residue difference in the sequence between these two residues suggests that if this portion of the peptide is in α-helix conformation, the two residues would be pointing to the same region, which would be the S100B binding cavity. Following this reasoning, we stabilized the helicity of this segment using distance restraints and performed long equilibration of the remaining segments of Aβ_{42}. Following this reasoning, we also decided to select two specific segments of this peptide, a *head* and a *tail* region, to simplify the analysis and the discussion. The *head* region, consisting of residues 24-36, comprises what we expect is a more stable αhelix and contains the two key residues known to interact strongly with S100B. The *tail* region, which consists of residues 2-22, will also undoubtedly interact with S100B. Nevertheless, having no experimental data to guide our model, we simply let it equilibrate in order to minimize instabilities coming from this portion of the peptide. This region distinction is also supported by the fact that the N-terminus region Aβ peptides (our *tail*) has a clear less propensity to form α-helical structures, even when lowering the dielectric of the solvent media [64].

We used the MD pre-equilibration step to follow and evaluate the secondary structure and the complex interface of Aβ_{42}:S100B (Figure 3.9). As can be seen on these results, in all systems simulated, the helical content of the peptide converged to a plateau within the initial 100 ns ($\sim$27 residues for BM3 and $\sim$20 residues for the remaining BMs). As expected, the peptide lost in most cases half of its helical content in this pre-equilibration run. We should note that $\sim$ 10 residues out of the total observed are in the *head* group region, which was kept helical using distance restraints. The interface SAS from the *head* of the peptide also converges promptly, which confirms that 100 ns was enough for this pre-equilibration step. Furthermore, this stable interfaces in the *head* region suggest a good complementarity in all BMs obtained from the docking procedure. Based on the interfacial SAS from the *tail* of Aβ_{42}, we can see a wide variety amongst the several binding modes, and this property does not seem to converge. This was expected and, as mentioned before, we only rely on this portion of the peptide to adsorb to S100B, even only transiently, in order not to destabilize the *head* group region. From this pre-equilibration runs, we stabilized the interfaces of the 5 BMs *head* group regions, which allowed the identification of all

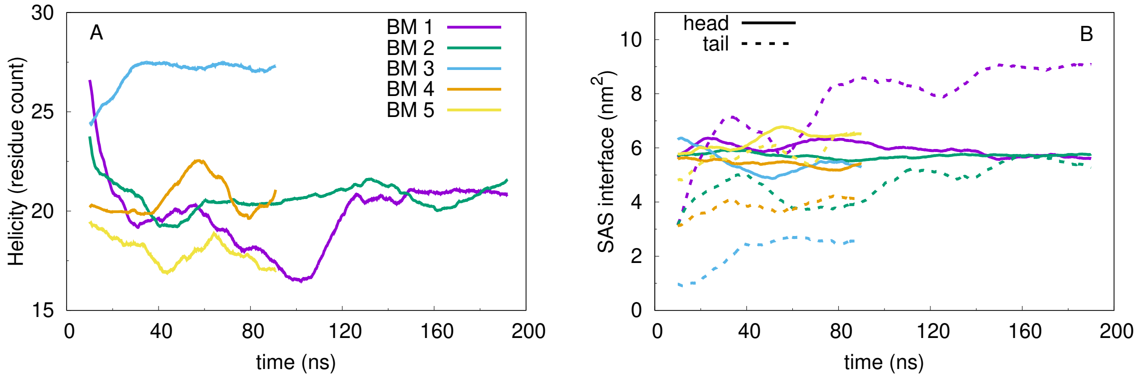


Figure 3.9: Helical content of Aβ_{42} (A) and the Interfacial SAS of the complex (B) during pre-equilibration MD. The 5 binding modes are represented and the interface SAS was also split between *head* and *tail* regions. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series.

key residues involved in the interface Table 3.5). The residues identified in each BM at the end of the pre-equilibration runs will define the molecular details of those interfaces and can later be used to map major conformational transitions using simple SAS calculations.

Table 3.5: Summary table with the average Interfacial SASA (SASA^{int}) between the S100B and the Aβ_{42} *head* region, and key residues involved in that interface in each BM. These were identified using SASA interface calculations for each residue in the initial conformations of each BM (conformations selected for each replicate). The average SASA^{int} values were also calculated using the 3 starting configurations (replicates) of each BM.

BM	SASA^{int} (nm)	Key Residues	
		A$_\beta$ *head*	S100B
1	5.8±0.0	24, 25, 28, 29, 31, 32, 33, 34, 35, 36	62, 71, 74, 75, 78, 79, 82, 153, 160, 162, 166, 170
2	5.1±0.4	27, 28, 31, 32, 35, 36	71, 78, 150, 151, 153, 154, 160, 162, 166, 169, 170, 173
3	5.2±0.4	26, 27, 28, 30, 31, 32, 33, 34, 35	59, 62, 75, 78, 79, 82, 153, 154, 162, 163, 166, 170
4	5.0±0.2	24, 25, 26, 29, 32, 33, 34, 35, 36	25, 71, 74, 150, 153, 165, 166, 169, 170, 173, 176, 177, 181
5	6.3±0.1	24, 25, 26, 28, 29, 30, 32, 34, 35, 36	62, 63, 71, 75, 79, 146, 147, 150, 153, 165, 166, 169, 170, 173, 174, 177

3.3.2 Convergence of the MD production runs

When performing production MD simulations, a very important initial quality control procedure is to attested their convergence. To do this, we calculated RMSD values for S100B and for Aβ_{42}, separately, and the gyration radius of the complex during the simulations for each replicate (Figure 3.10). We can observe that, the S100B RMSD values are very low and the profiles converged swiftly. The RMSD^{Sym} variations were also small and are probably just capturing weak interactions with the Aβ_{42} peptide, which are asymmetric by nature. This also confirms the lack of allosteric behaviour from S100B, where both monomers continue significantly synchronised in their structural fluctuations, mostly unaffected by the presence of the peptide. The peptide RMSD values are quite high and showcase their large movements before adsorbing and stabilizing at the S100B surface. In the gyration radius, all replicates for each binding mode converged, which indicates that no major desorption events take place throughout the simulations. Replicate 2 of BM2 is quite interesting, since at around $\sim$500 ns it became more compact, reaching gyration radius values achieved by no other complex. We compared the conformations at the beginning and end of the simulation for this particular replicate (Figure 3.11) and observed that a significant movement of an helix towards the center cavity, which results in the observed decrease in the radius of gyration.

We also calculated the secondary structure over time for the S100B protein in all BMs (data not shown). In all systems, this protein showed almost unperturbed helicity profiles, confirming once again that it is very stable.

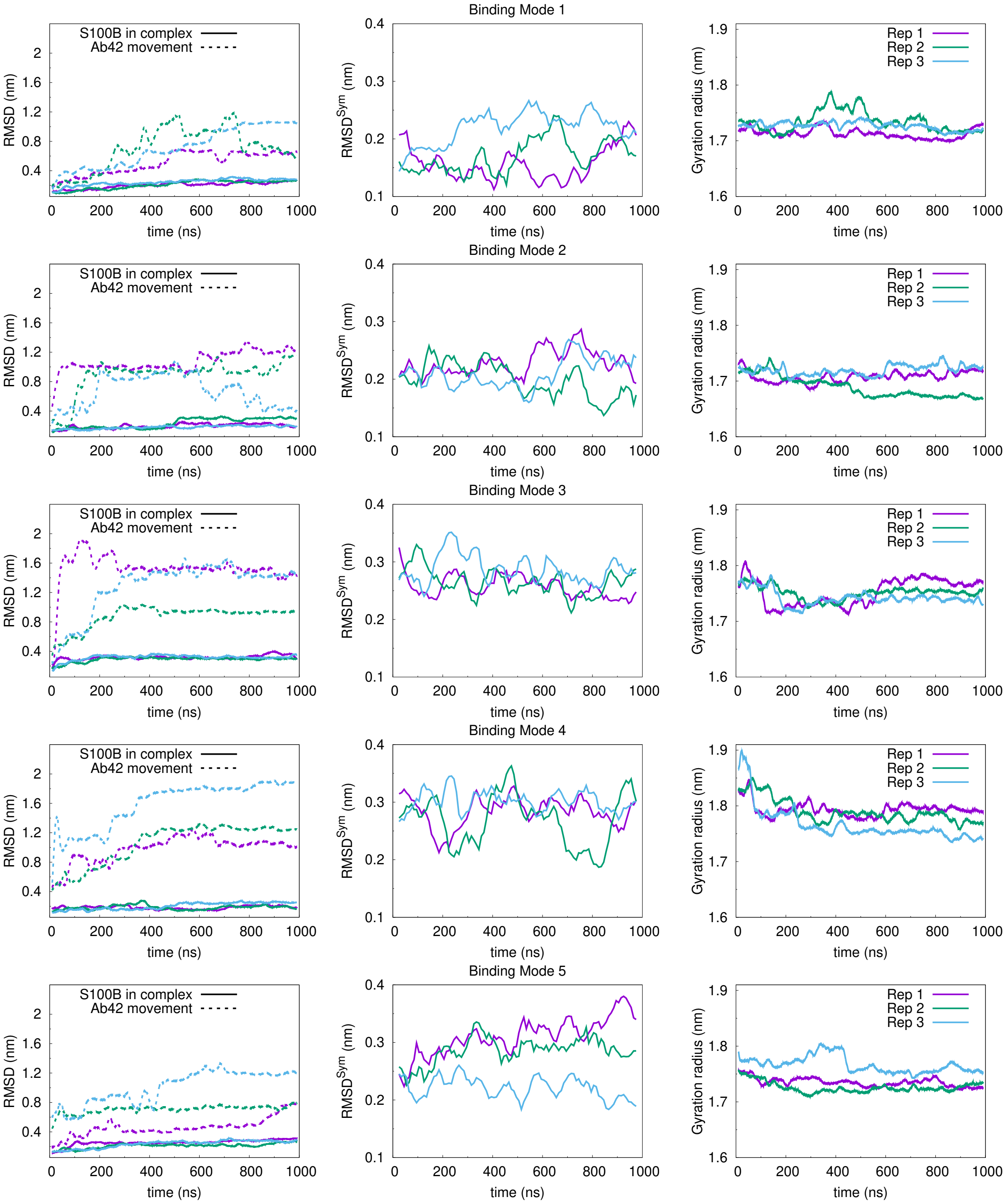


Figure 3.10: RMSD values of S100B or Aβ_{42} (left), asymmetry RMSD (RMSD^{Sym}) between S100B monomers (center) and the gyration radius of the complex (right) for all BMs. For the RMSD calculations, only the S100B dimer was used in the conformational alignment. Each row corresponds to a binding mode. A floating window of 20 ns was used for gyration radius and RMSD, while a floating window of 50 ns was used for monomer symmetry in order to reduce the local fluctuations of each data series.

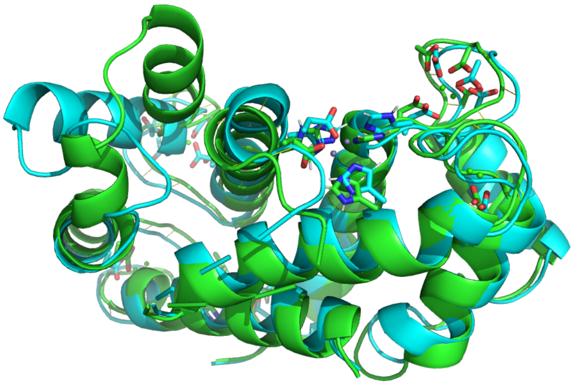


Figure 3.11: A conformational transition in S100B from replicate 2 of BM2. The green structure corresponds to the final frame of the simulation, while the cyan one corresponds to the initial frame. The image was obtained by aligning the two dimer structures.

The Aβ_{42} helical content convergence is an important objective and, while the *tail* region will be very dynamic and would probably need longer simulations, a good convergence of the *head* region is pivotal. Therefore, we calculated the helical content of the *head* and *tail* regions of Aβ_{42}, separately (Figure 3.12). The *head* group in BM1 and 2 are fully helical and seem very stable throughout the simulations. Similar behavior was observed for replicates 1 and 2 of BM3, and replicate 2 of BM5. Replicate 3 of BM3 and replicate 1 of BM5 converged to only a partially folded structure, suggesting that a significant conformational transition occurred in the binding pocket. In the remaining replicates, including all from BM4, there is a complete loss of helicity, indicating a high instability of these interfaces and a failure by S100B in stabilizing the α-helix content of the *head* group. The helicity of the *tail* region of Aβ_{42} seem to converge for all BM/replicates. The large conformational variability is expressed in the plateaus obtained, where some show only a residual amount of helical structure, which we theorize to be closer to equilibrium, while others retaining a significant amount of helical residues, probably still far from equilibrium. At this point, we use the experimental data available (Figure 1.7) and the information available in the literature [64] about this peptide to identify the outlier simulations and discard them. Therefore, the indicated replicates from BM3 (Rep 3) and 5 (Reps 1 and 3), as well as all replicates from BM4 will not be analysed further and nor be used to extract equilibrium properties.

Considering that the molecular details of the interaction established between S100B and Aβ_{42} is the focus of this work, a measure of the interface size can and should also be used to define the complex equilibration. After excluding the replicates with less stable secondary structure at the *head* region of the peptide, we calculated the interfacial area of the complex (Figure 3.13). The results show that the interfacial area of the peptide *head* region tended to stabilize between 5 and 7 nm^2 for all replicates and BMs. There are no significant deviations throughout the simulations. BM1 and BM2 showed the 3 replicates sampling the same area values. On the other hand, the two replicates of BM3 seem to be sampling slightly different configurations, but still showing high interface area values. When it comes to the interfacial area for the *tail* of the peptide (Figure 3.13E), it is clear that it is less stable than the interface established with the *head* region, due to the presence of larger fluctuations of this property. Nevertheless, this property does tend to converge in 200–400 ns of simulations. All simulations are sampling *tail* region interfaces of the same magnitude, which discards major configurational changes and the possibility of the tail detaching from S100B. Furthermore, it is interesting that all replicates of each BM are sampling the same region and that there are significant differences between BMs. This suggests some homogeneity within BMs, and some configurational variability between them, which was intended.

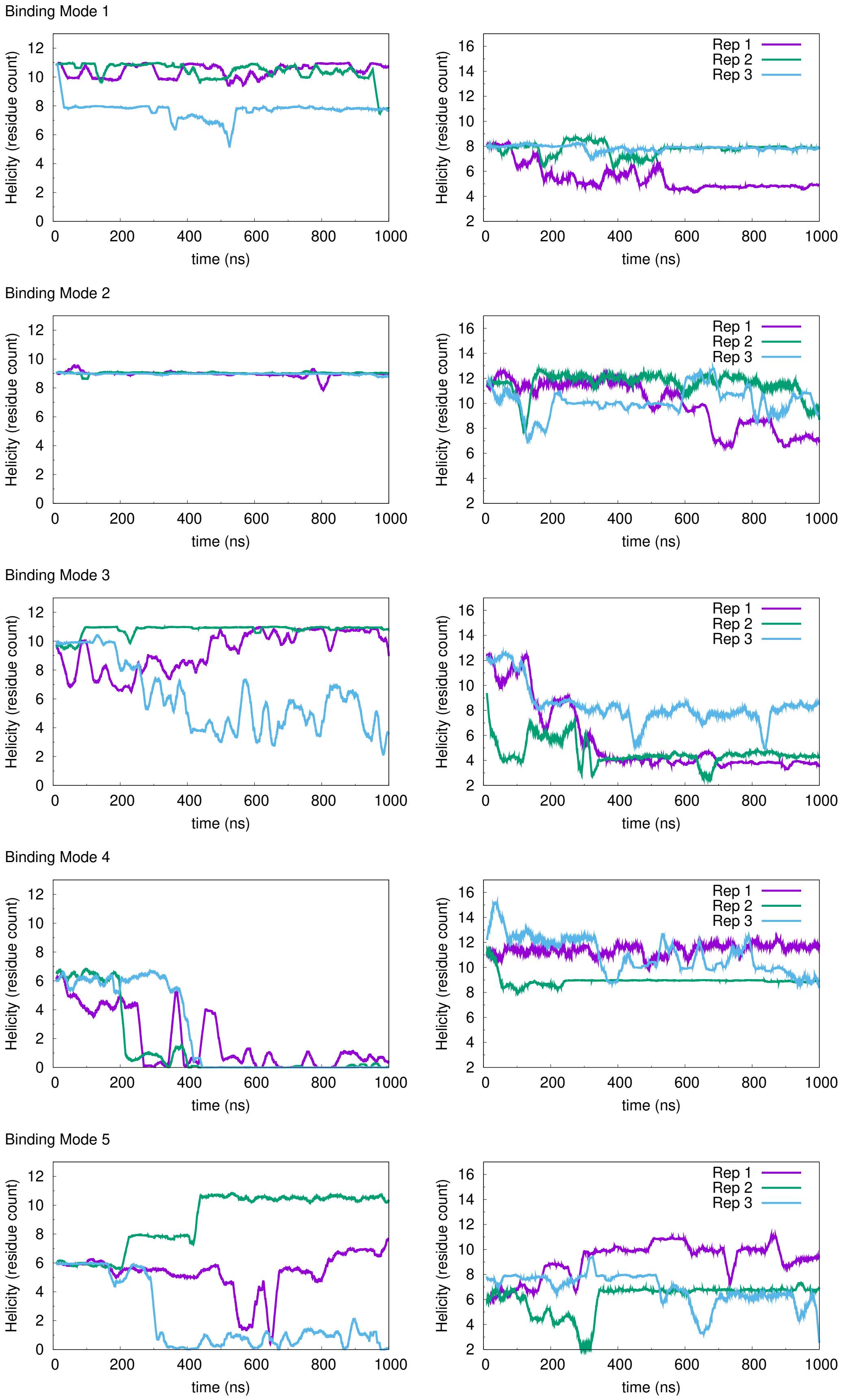


Figure 3.12: Helical content of Aβ_{42} *head* (left row) and *tail* (right row) regions during production MD runs. Each row corresponds to a binding mode. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series.

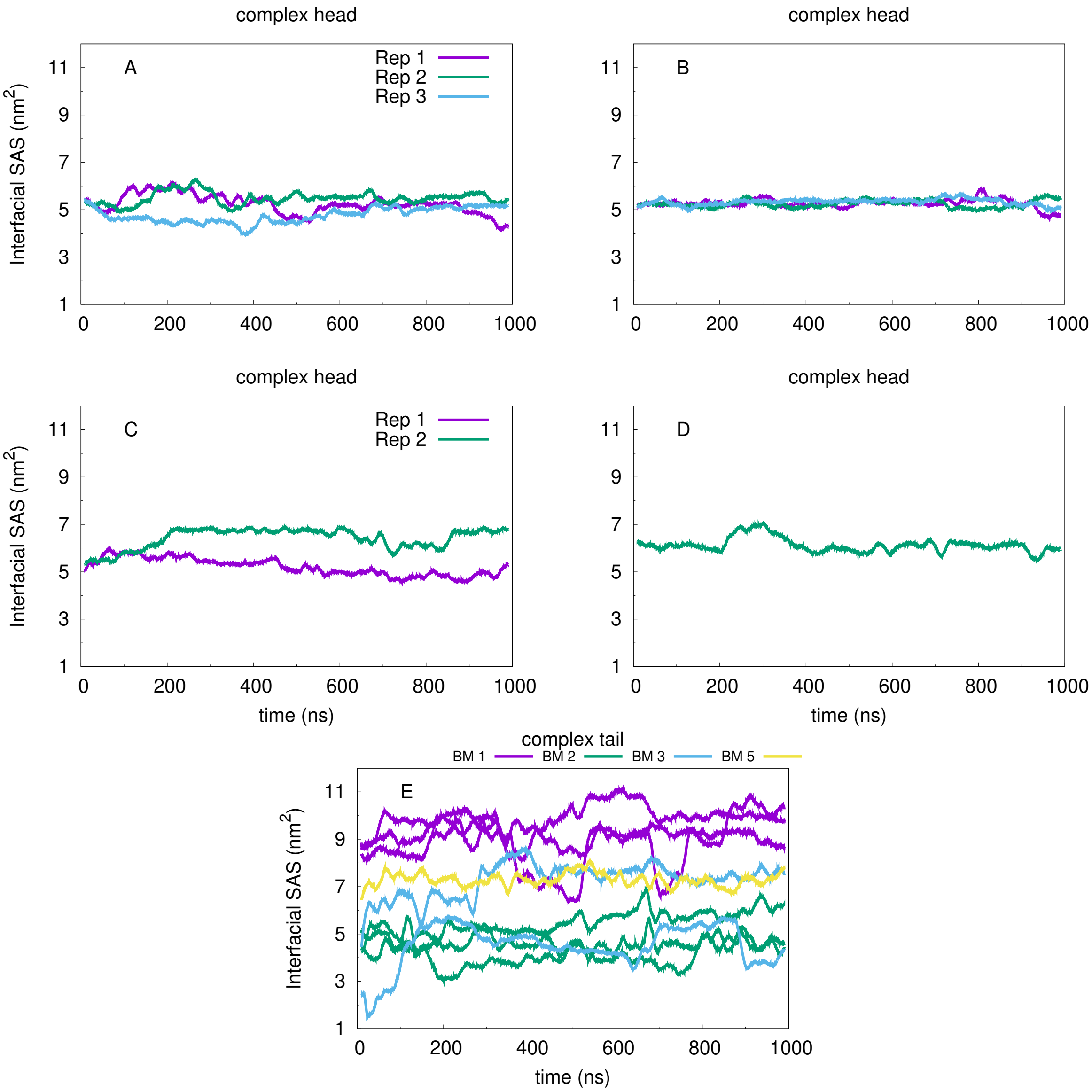


Figure 3.13: The interfacial area of S100B-Aβ_{42} *head* (BM1 A; BM2 B; BM3 C; BM5 D) and *tail* (E) regions. These final interfacial SASA values were obtained by averaging the two interfacial areas, one mapped on the protein and another mapped on the peptide surface. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series.

Looking at the residues involved in the initial and final interfaces (Tables 3.5 and Table 3.6), we can calculate the percentage of initial and final interfaces which are still formed at any given time. We can still calculate these values separately by focusing either on the *Aβ_{42} head* region (Figure 3.15) or on the S100B residues involved in the interface (Figure 3.14). These two analysis are important to capture those cases where the peptide either rotates in the same S100B binding region, or slides its own binding region along the protein surface. The Aβ_{42} *head* group residues interacting with the S100B protein are mostly retained throughout the simulation (∼60–80%), which was already hinted by the significant overlap between the list of amino acids from the initial and final interfaces. Furthermore, and also as a consequence, no completely new interfaces were created, since ∼60–80% of all final interfaces were already present at the start of the simulations. In overall, the second replicate of BM3 and 5 and all replicates of BM2 showed the most stable interfaces, where almost no variations were observed. On the S100B side, we note that most replicates tend to remain relatively stable along the simulations, retaining at least ∼60–80% of the overall interface. There are a few exceptions, like the replicate 3 of BM1, where we see a more pronounced decrease of the original interface and, at $\sim$ 500 ns, the final interface is formed. This structural rearrangement will most likely involve just a small Aβ_{42} *head* group movement along the binding cavity, since most of the original contacts are still in place. There are a few similar changes in other simulations but, more than being capturing any major configurational transitions, they illustrate the high sensitivity of the approach that is not commonly used in this type of structural analysis. Overall, these interfaces tended to be even more stable in the S100B protein than in the peptide, indicating that this dimer smaller conformational variability is playing a role and that the binding pocket which was hinted by the experimental data, is indeed complementary to the Aβ_{42} *head* region.

Table 3.6: Summary table of the key residues involved in the interface of each BM at the final conformation. These were identified using interfacial SASA calculations performed on the final frames of the stable replicates.

Binding Modes	Key Residues	
	Ab head	S100B
1	25, 26, 28, 31, 32, 35, 36	71, 75, 79, 82, 153, 154, 162, 166, 169, 170, 173
2	27, 28, 31, 32, 35, 36	71, 75, 78, 79, 82, 150, 151, 153, 154, 162, 163, 166, 170, 173
3	24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36	59, 60, 62, 71, 74, 75, 79, 150, 151, 153, 154, 162, 163, 166, 169, 170
5	24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36	62, 71, 75, 79, 82, 146, 147, 150, 153, 162, 165, 166, 169, 170, 173, 177, 182

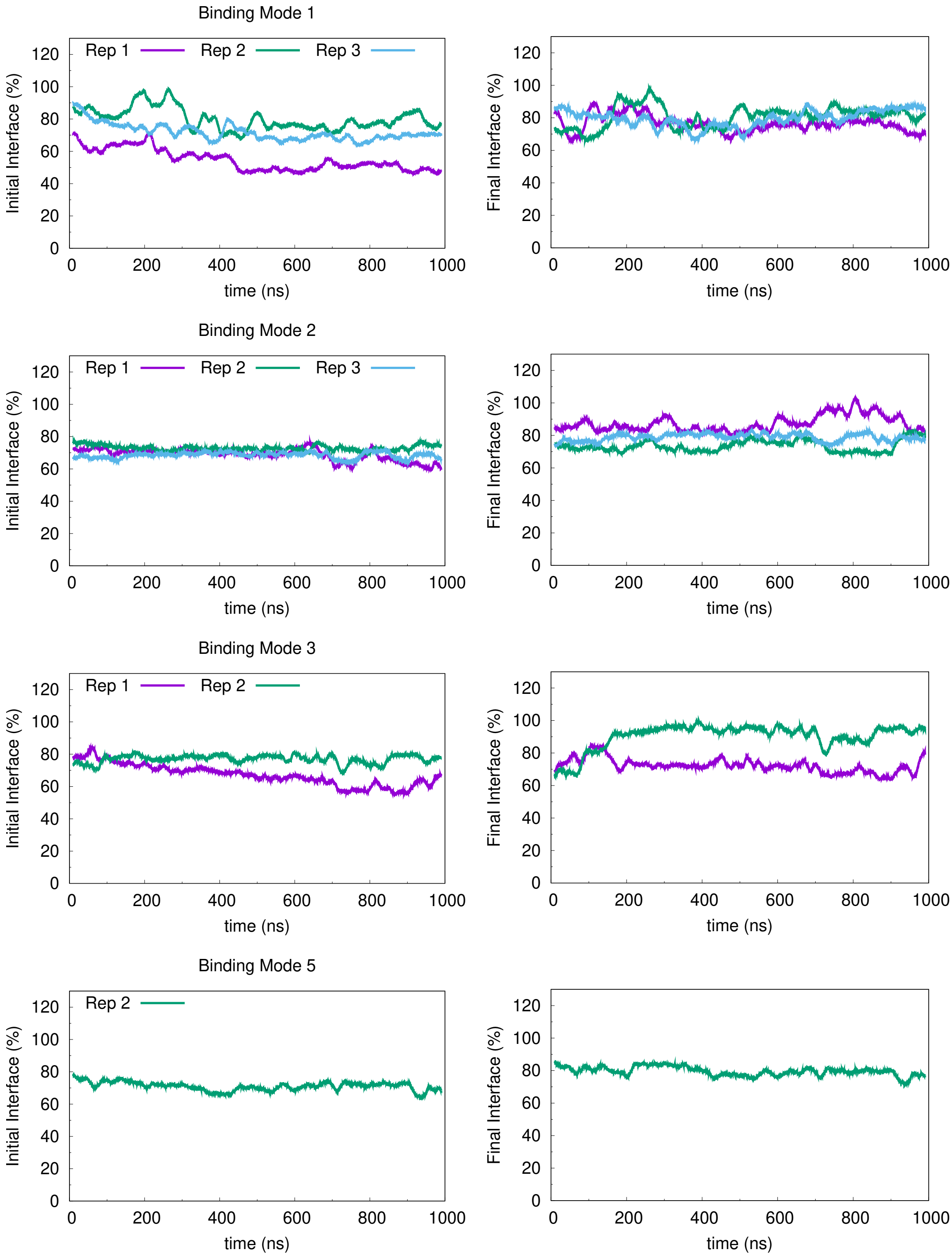


Figure 3.14: The initial (left) and final (right) Aβ_{42} *head* region interface percentage retained during production runs. The percentages were calculated using a sum of the interface area of each identified residue involved in the interface in the start or end of the simulations. This value is then divided by reference values obtained from the initial/final conformations of each BM. Each row represents a different BM. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series.

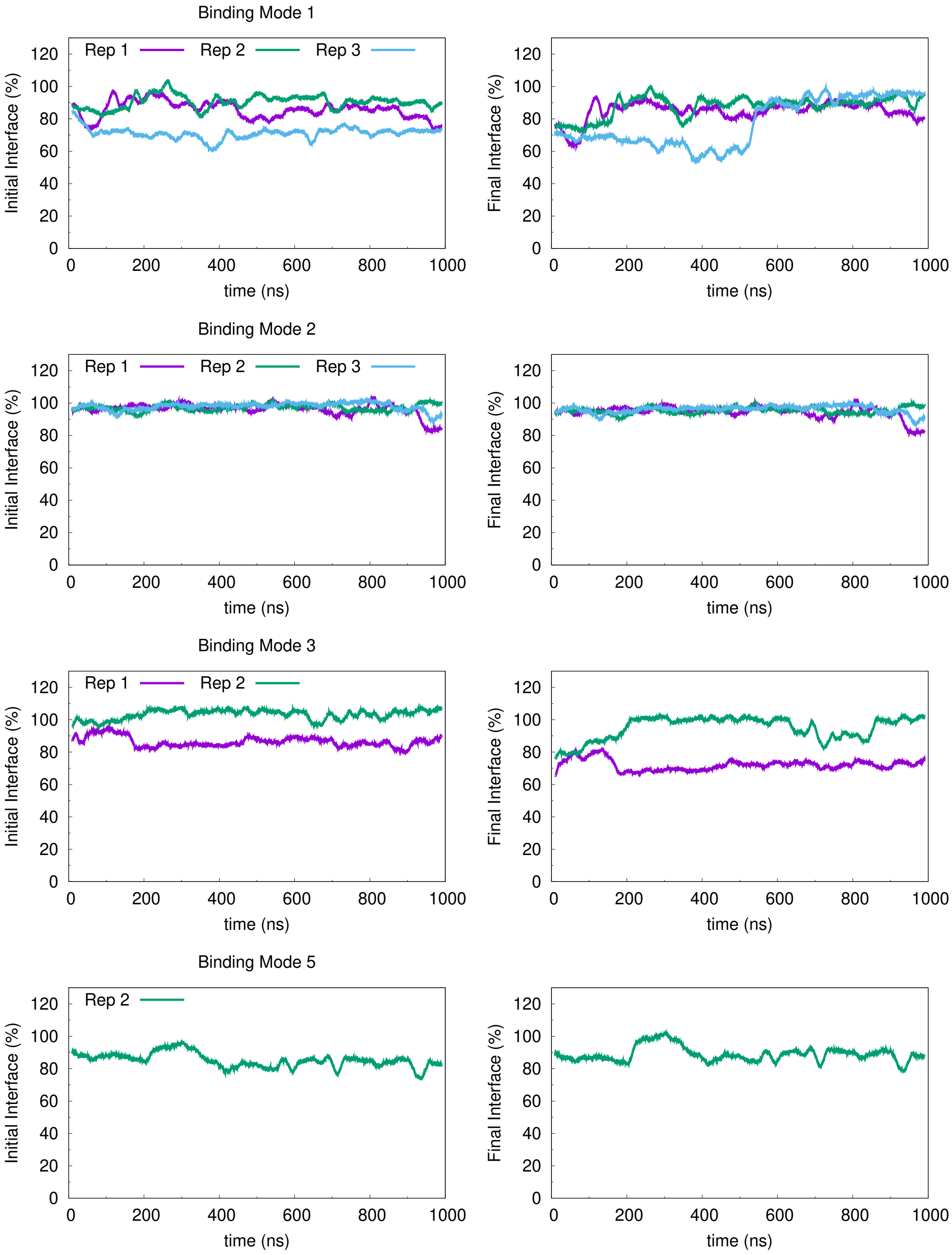


Figure 3.15: The initial (left) and final (right) S100B interface percentage retained during production runs. This interface was defined using only the Aβ_{42} *head* region as a partner. The percentages were calculated using a sum of the interface area of each identified residue involved in the interface in the start or end of the simulations. This value is then divided by reference values obtained from the initial/final conformations of each BM. Each row represents a different BM. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series.

From the SASA calculations, we can also evaluate separately the relevance of hydrophobic and hydrophilic interactions in the S100B:Aβ_{42} complex. These can be obtained by conjugating the SASA values of the residues at the interface with the hydrophobicity Wimley–White (WW) scale [55] and obtain what we call the hydrophobicity/hydrophilicity index (SAS^{hydro}) [54]. This index can be calculated using all residues at the interface or separating them between hydrophobic (negative values in the WW scale) and hydrophilic (positive values in the WW scale) residues (Figure 3.16).

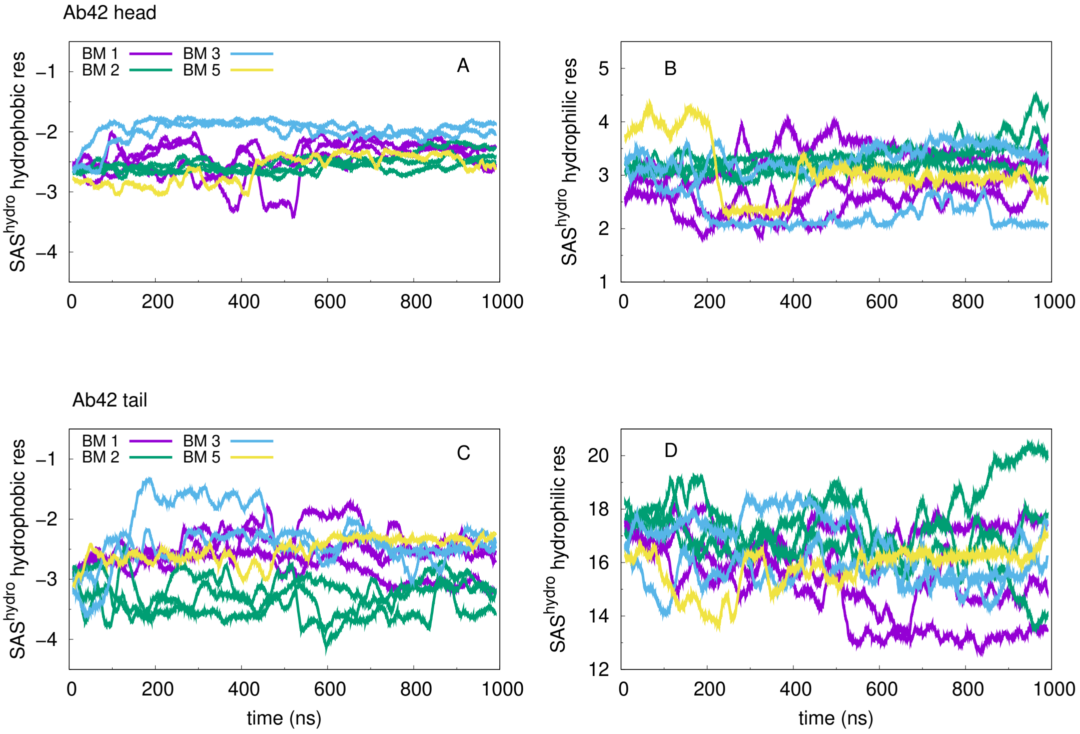


Figure 3.16: Hydrophobic/hydrophilic (SAS^{hydro}) indexes for the hydrophobic and hydrophilic residues in the *head* and *tail* Aβ_{42} regions at the interface. The hydrophobic and hydrophilic residues were separated based on their SAS^{hydro} signal. A floating window of 20 ns was used in order to reduce the local fluctuations of each data series. (hydrophobic index for the *head* A; hydrophilic index for the *head* B; hydrophobic index for the *tail* C; and hydrophilic index for the *tail* D)

From these results, we note that the *head* region of the peptide has a good balance between hydrophobic and hydrophilic residues at the interface. The SAS^{hydro} index values tend to converge at least around the half point of the simulations and there are no cases with clear shifts from one type of interactions to the other one. This results add to all previous observations regarding the stability of these interfaces. The SAS^{hydro} index values for the *tail* region hydrophilic residues are particularly high which can be explained due to the polar nature of this segment and the fact that there is more structural loss, allowing to establish more interactions with the surface of the dimer. Since many of these interactions, either polar or apolar, are transient in nature, we obtained larger fluctuations in this peptide region.

Taking into consideration all these properties, which were calculated over time for the different BMs, we have now to define which part of the simulations achieved equilibrium. Many properties were very stable over the simulations, specially those involving only the S100B. However, all those calculated with the Aβ_{42} peptide were more difficult to converge. Indeed, a simple radius of Gyration of the complex (Figure 3.10) seemed to only stabilize around ∼400 ns for most replicates, with a similar behavior observed with the helical structure content (Figure 3.12). Additionally, even in the S100B:Aβ_{42} *head* interfaces, which were particularly stable over our simulations, we observed small conformational transitions at ∼500 ns. Therefore, and to be on the safe side, we considered the equilibration time to be the first 500 ns of simulations. Henceforth, these first 500 ns of all simulations were discarded whenever equilibrium properties were calculated. Even though this equilibration period may be longer than necessary, it will assure that, after said period, at least the properties associated with the *head* region of the peptide will be equilibrated. As expected, for the peptide *tail*, some replicates would require longer simulation times. Furthermore, and probably more important, the high number of configurations on how to adsorb this region of the peptide to the S100B protein would require an impossibly large number of replicates to correctly capture its conformational space. Nevertheless, this should not significantly impair our study, since we will focus our analysis mainly on the *head* region, which contains the residues which are known to be very relevant to the binding, namely Lys28 and Ile31.

3.3.3 Evaluation of the binding modes

After defining the equilibration period for the production runs, we calculated some equilibrium properties to evaluate and compare between configurations and identify the most stable binding modes. We have characterized in detail the binding mode interfaces in previous sections, however, we have not yet quantified the role of each individual amino acid in establishing those interfaces. Before looking at any energetic criteria, we can analyse which residues are simply more abundant in each BM. To do this, we can extend our SAS^{hydro} index analysis to be performed per residue over the equilibrated segments of the simulations (Figure 3.17).

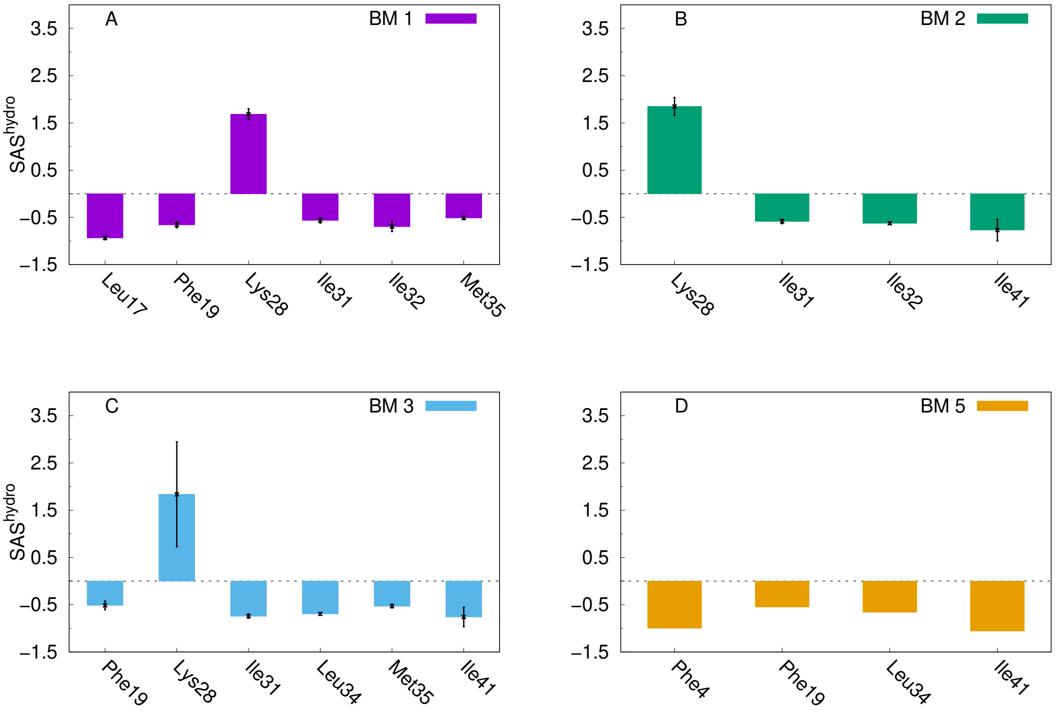


Figure 3.17: SAS^{hydro} index values for the most important residues of Aβ_{42} at each BM (BM1 A; BM2 B; BM3 C; and BM5 D). The positive/negative values stand for hydrophilic/hydrophobic residues, respectively. To select only the most persistent contacts, we have applied a cutoff of 1.5 and -0.5 to the hydrophilic and hydrophobic indexes, respectively. The error bars were calculated from the standard error of the mean between replicates, when available.

This persistence analysis using the SAS^{hydro} index, calculated for each amino acid separately, shows a few differences between binding modes. BM1–3 successfully identified Lys28 and Ile31 at the interface, which is in excellent agreement with the experimental data available [15]. These BMs show mostly *head* region residues, which indicates that the *tail* region is involved in more heterogeneous interfaces, resulting in less persistent interactions. In BM5, only hydrophobic interactions showed a relevant SAS^{hydro} index and the absence of the mentioned two key residues indicate that the Aβ_{42} *head* region, which is still mostly α-helical, changed its conformation to expose those residues to the solvent. Other hydrophobic residues, like Ile32, Met35, and Ile41 are consistently present at the interface and seem to constitute the hydrophobic side of the Aβ_{42} *head* region, suggesting a small amphiphilic character for this α-helix, when formed.

We know that the Lys28 residue of Aβ_{42} has a pivotal role in establishing electrostatic interactions with several partners from S100B. In order to identify those partners, we developed a thorough minimum distance procedure between Lys28, more specifically the side chain nitrogen atom, and all residues of S100B. This was done to find "hot spots" in the protein and compare them among BMs. After calculating these distances, we joined all data obtained from equilibrated segments and discarded the residues that were mostly at longer distances (>0.6 nm). For the few residues that were systematically interacting with Lys28, we calculated probability density distributions (Figure 3.18) and represented them for visual inspection (Figure 3.19). This procedure clearly identified several partners, which corresponded to very persistent interactions, like with Asp69 in BM2, or others more transient, like those in BM5. In BM1 and 2, Lys28 is directly facing the S100B interface (Figure 3.19A–B) and the minimum distance distributions (Figure 3.18A–B) perfectly describe the neighboring partners. The presence of Glu86 in BM1 and Asp69^B/Glu72^B in BM2 provide the negative electrostatic potential that stabilizes Lys28, working as anchors for these BMs. In BM3, some persistent interactions are polar in nature, but lack a counter-charge balance. This may explain why, in the absence of anionic residues, Lys28 is significantly rotated towards more solvent-exposed regions (Figure 3.19C). BM5 is particularly interesting, since in this configuration the Aβ_{42} *head* region is rotated almost 180^o along the helix axis, which partially exposed Lys28 to the solvent (Figure 3.19D). Another consequence of this setup is that Ile31 is also rotated along the same axis and ended being fully exposed to the aqueous medium. This is in disagreement with several experimental observations and we expect this BM to be less stable than the remaining ones.

Finally, in an attempt to quantify the stability of the different BMs, we estimated their binding energies using a MM-PBSA methodology [57]. These are still preliminary results, since the conditions and settings used in the calculations are very similar to the ones used by the authors of the *g_mmpbsa* GROMACS tool [57]. Since the *tail* region of Aβ_{42} showed a remarkable conformation heterogeneity with non-converged contributions to the overall binding phenomena, we opted to calculate the binding energies only for the *head* region (Figure 3.20). Most energy terms converge after 500 ns of simulations even though the differences in behaviour between replicates is significant. This will inevitably lead to large error values when calculating overall binding affinities (Table 3.7). The results confirm electrostatics as the main driver for the binding even in BMs which showed less marked ionic interactions involving Lys28 (BM3 and 5). These are the binding modes that show less negative Coulombic term values, while presenting the lowest VdW energies. This trade-off results in worse binding constants since the Coulombic contributions are significantly stronger. From the two solvation contributions, Solv^{Polar} and Solv^{Apolar}, only the Polar term is large and varies enough to influence the final binding energy. The Apolar term is very small and constant among the BMs. There is an interesting negative correlation between the Solv^{Polar} and the Coulombic terms, which can be reasoned from the position of Lys28. Whenever this residue moves away from the interface with S100B (BM3 and 5), the stabilizing Coulombic contribution to the binding energy decreases, but the polar solvation penalty also decreases, due to the presence of a more polar (charged) residue at the solvent surface. The final binding affinities are also obtained from a trade-off between these two contributions.

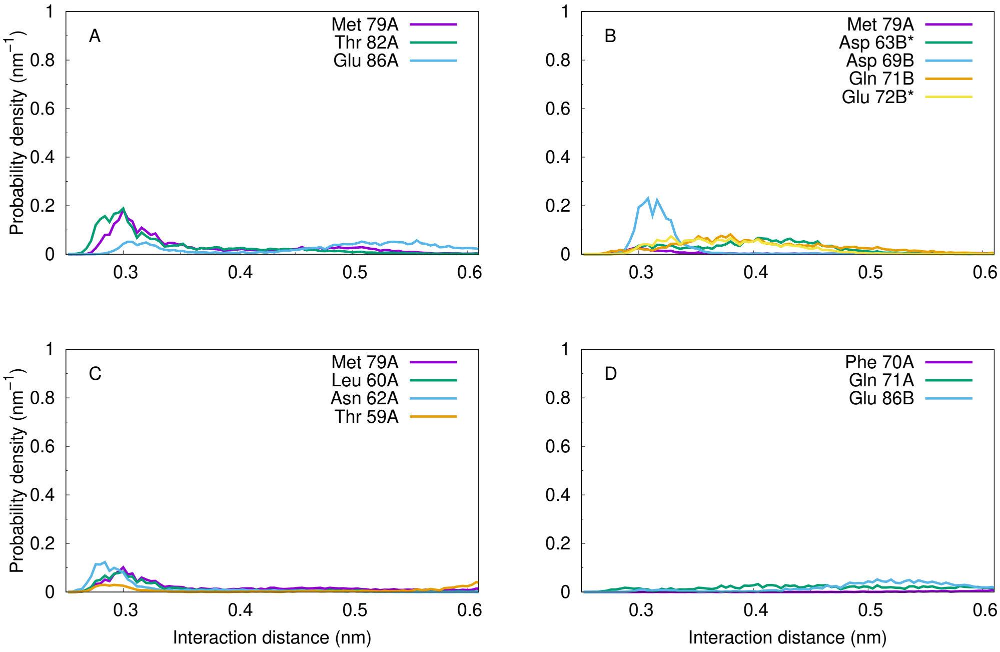


Figure 3.18: Probability density distributions of minimum distances between Lys28 and its hot spot partner residues in S100B for the different binding modes (BM1 A; BM2 B; BM3 C; and BM5 D). The residues marked with a * are involved in the metal coordination centers.

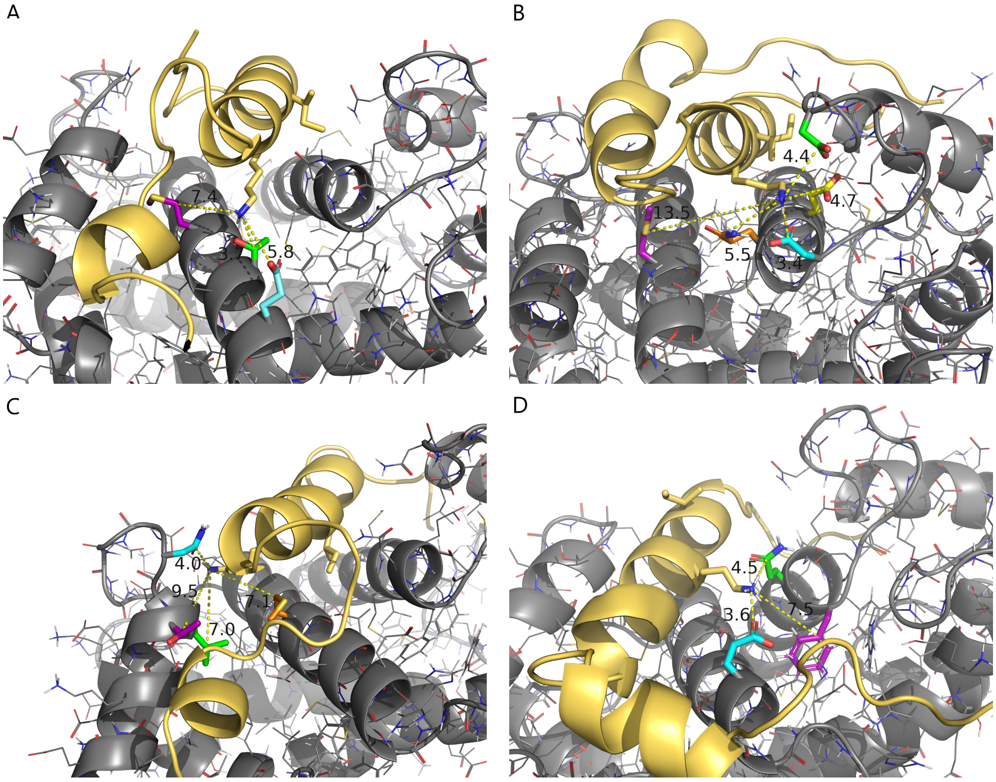


Figure 3.19: Structural representation of the Aβ_{42}:S100B complex for the different binding modes (BM1 rep1: A; BM2 rep2: B; BM3 rep2: C; and BM5 rep2: D). The focus on Lys28 of the Aβ peptide illustrates the different partners on the S100B side. The peptide and the protein are shown in yellow and gray cartoon, respectively. Both Lys28 and Ile31 of Aβ_{42} are shown using sticks. The protein residues were colored according to color code used in Figure 3.18. The distances between Lys28 and those residues, at the final snapshot, are also shown.

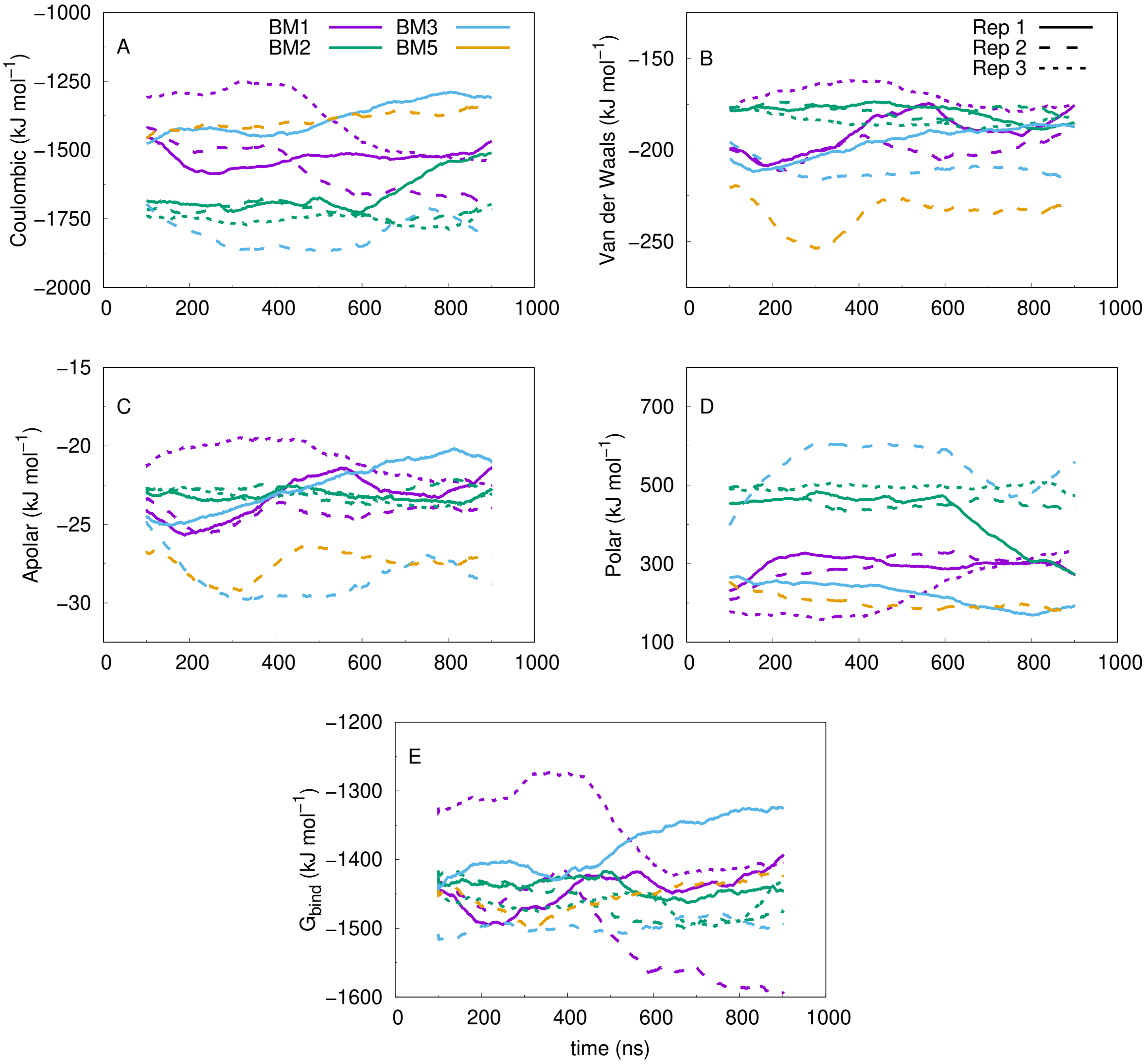


Figure 3.20: The different energetic contributions to the binding free energy (G_{bind}) for the different BMs of Aβ_{42} *head* and S100B over time. Coulombic (A) and Van der Waals (B) are the vacuum molecular mechanics contributions. The Apolar (C) and the Polar (D) refer to the solvation energy contributions. The sum of these 4 terms leads to the G_{bind} (E).

Table 3.7: Summary table of the average energy components calculated using a MM-PBSA method, for Aβ_{42} *head* complex with S100B. All energies are presented in kJ/mol units. The error values were calculated from the SEM using the different replicates, when available.

Energies	BM1	BM2	BM3	BM5
Coul	-1555.8±108.5	-1690.0±92.9	-1564.8±474.5	-1366.1
VdW	-184.9±13.6	-183.0±2.0	-200.2±23.2	-231.3
Solv^{Apolar}	-22.7±1.4	-23.2±0.1	-24.8±7.4	-27.1
Solv^{Polar}	297.0±15.4	430.6±79.1	372.1±352.5	188.5
G$_{bind}$	-1466.3±109.4	-1465.7±23.3	-1417.7±153.0	-1435.9

The final free energy of binding (G_{bind}) derives from the summation of all the other energetic contributions. The best configurations seem to be BM1 and 2, but the variability between replicates (Figure 3.20) leads to a large uncertainty (the error values in Table 3.7) making it not possible to discard the remaining ones. It is interesting that the magnitude of the binding constants (-1418 to -1466 kJ/mol, which gives -339 to -350 kcal/mol) suggests a remarkable interaction between the S100B protein and this portion of the Aβ_{42} peptide. Combining these values with the observation of very homogeneous interfaces and very stable complexes, it indicates that, indeed, this Aβ peptide can be stabilized in a partially folded structure (α-helical *head* group region) interacting with the S100B binding cavity on top of the C-terminal helices, as hinted by experimental data (Figure 1.7) [15]. This coincides with the fact that S100B, as a chaperone, binds with frustrated regions of the target protein, which is, in this case, Aβ. Considering the dynamic behaviour and conformational variability of the peptide, the *head* region of the peptide could be considered a frustrated segment when in a α-helical conformation.

Chapter 4

Concluding Remarks

Summing up, in this master thesis, we aimed to study in detail the interaction between S100B and the Aβ peptide in order to ascertain the stability of this complex and identify key features of the overall binding phenomena. To achieve this, we have parameterized several metal centers in S100B to provide a very accurate model of our template protein. Using the experimental data at our disposal, we performed guided docking calculations of the complete Aβ_{42} and a portion (Aβ_{25-35}) of what we called the *head* group. Five different binding modes (BM) for the S100B:Aβ_{42} system were obtained with this procedure and their stability and structural details were explored using long MD simulations. With this workflow, we sampled the conformational space of the five binding modes, each with three replicates, achieving 15 μs of total simulation time.

We put significant effort in the analysis of the MD simulations equilibration. This is of major importance since a large portion of our Aβ_{42} peptide, which we named the *tail* region, was very difficult to converge and, by running only triplicate simulations, we have no assurance that they achieved equilibrium. Nevertheless, our focus, from the start, was on the *head* group, which was significantly more easy to equilibrate in the BMs that showed to be stable. We did observe that, in several setups, the peptide lost most of its helicity, which led us to discard them. None of the replicates simulated for BM4 were kept, indicating that this configuration obtained from Haddock was not stable. Even though the BM4 was obtained from a very good solution from the docking calculation, the remaining of the Aβ_{42} peptide (the *tail* region), when was rebuilt *a posteriori*, may have been placed in an unfavorable location of the S100B surface Figure 2.4. One in BM3 and two replicates in BM5 were also discarded due to instabilities in their secondary structure content. Furthermore, BM5 showed a configuration where Ile31 and (partially) Lys28 were not fully committed to the interface with S100B, which is in disagreement with the experimental data. It is interesting that this was the only BM obtained from a molecular docking calculation using the complete Aβ_{42} peptide. This confirms our initial restraints about using the complete peptide, since this creates a large configurational space for the docking calculation, where the experimental information, which we used to guide the docking procedure, got diluted. In the end, we obtained a BM where the Aβ_{42} peptide was neither unstable or stabilized in a wrong configuration.

BM1 and BM2 showed very stable interfaces, where the two key residues were packed inside the S100B interface and with Lys28 stabilized by neighboring acidic residues, which provided a superior Coulombic contribution to the binding energy. The main difference between BM1 and BM2 comes from a the topological arrangement of the Aβ_{42} *tail* region, where in BM1 it is mostly unstructured, while in BM2, it retained a large α-helix. This *tail* helix was stabilized by adsorbing to the *head* helix in what can be interpreted as a meta-stable structure that would evolve towards the BM1 type of unfolded *tail* structure, given time.

Finally, our computational efforts to study the molecular details of the Aβ_{42}:S100B complex are significantly limited due to the complex nature of its interactions. Even though the S100B protein is very stable with a conformational space easily captured within a few hundred of nanoseconds, the Aβ_{42} peptide is highly problematic. The fact that the peptide is only partially folded in solution creates a huge challenge to model its conformational space which is inaccessible to most computational techniques. However, with the help of all experimental evidence available, we were able to devise a protocol to direct both the conformational space of the peptide and the geometry of the final complex configurations. Altogether, we were able to explore successful binding modes which provided atomistic information about the specific interactions that are stabilizing the complex.

Chapter 5

Future Work

In terms of future work, we plan to extend the calculations of the binding mode energies using another MM-PBSA method. The protocol used here limits the access to several important parameters in the method and was developed/optimized to study the interaction of small ligands with proteins. In our group, we are working on a new MM-PBSA method based on the DelPHI Poisson–Boltzmann solver which will most likely provide more reliable estimations of the binding energies. We will also extend the minimum distance protocol to other Aβ_{42} residues beyond Lys28, which was performed in this thesis.

In the future, we will also perform sequence alignments between S100B and several other S100 neuronal proteins, including S100A8, S100A9 S100A6 and S100A12, as well as S100A2, which is non-neuronal. The objective is to identify which residues will be the ideal partners for the Aβ_{42} peptide and their potential in sequestering this neurotoxic peptide from solution.

Bibliography

[1] A. Gruber. “Chapter A06 Protein Structure, Modelling and Applications”. In: *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*. National Center for Biotechnology Information (US), 2006, pp. 137–170.

[2] S. Schmid. “Immobilization of DNA and protein to polymerized SU-8 photoresist investigated with microarray assays”. Diploma Thesis. ETH Zürich, 2020, pp. 1–75.

[3] K. A. Dill. “Dominant forces in protein folding”. In: *Biochemistry* 29.31 (1990), pp. 7133–7155.

[4] C. N. Pace et al. “Forces contributing to the conformational stability of proteins”. In: *The FASEB journal* 10.1 (1996), pp. 75–83.

[5] C. N. Pace, J. M. Scholtz, and G. R. Grimsley. “Forces stabilizing proteins”. In: *FEBS letters* 588.14 (2014), pp. 2177–2184.

[6] D. M. Taverna and R. A. Goldstein. “Why are proteins marginally stable?” In: *Proteins: Structure, Function, and Bioinformatics* 46.1 (2002), pp. 105–109.

[7] R. M. Vabulas et al. “Protein folding in the cytoplasm and the heat shock response”. In: *Cold Spring Harbor perspectives in biology* 2.12 (2010), a004390.

[8] J. Adamcik and R. Mezzenga. “Amyloid polymorphism in the protein folding and aggregation energy landscape”. In: *Angewandte Chemie International Edition* 57.28 (2018), pp. 8370–8382.

[9] W. Wang and C. J. Roberts. “Protein aggregation–mechanisms, detection, and control”. In: *International Journal of Pharmaceutics* 550.1-2 (2018), pp. 251–268.

[10] C. J. Maynard et al. “Metals and amyloid-β in Alzheimer’s disease”. In: *International journal of experimental pathology* 86.3 (2005), pp. 147–159.

[11] S. A. Kotler et al. “Differences between amyloid-β aggregation in solution and on the membrane: insights into elucidation of the mechanistic details of Alzheimer’s disease”. In: *Chemical Society Reviews* 43.19 (2014), pp. 6692–6700.

[12] R. F. Sowade and T. R. Jahn. “Seed-induced acceleration of amyloid-β mediated neurotoxicity in vivo”. In: *Nature communications* 8.1 (2017), pp. 1–12.

[13] C. Iannuzzi et al. “Misfolding and amyloid aggregation of apomyoglobin”. In: *International journal of molecular sciences* 14.7 (2013), pp. 14287–14300.

[14] P. dos Santos et al. “Alzheimer’s disease: a review from the pathophysiology to diagnosis, new perspectives for pharmacological treatment”. In: *Current medicinal chemistry* 25.26 (2018), pp. 3141–3159.

[15] J. S. Cristóvão et al. “The neuronal S100B protein is a calcium-tuned suppressor of amyloid-β aggregation”. In: *Science advances* 4.6 (2018), eaaq1702.

[16] C. Haass et al. “Trafficking and proteolytic processing of APP”. In: *Cold Spring Harbor perspectives in medicine* 2.5 (2012), a006270.

[17] R. Coronel et al. “Role of amyloid precursor protein (APP) and its derivatives in the biology and cell fate specification of neural stem cells”. In: *Molecular neurobiology* 55.9 (2018), pp. 7107–7117.

[18] T. Qiu et al. “Aβ42 and Aβ40: similarities and differences”. In: *Journal of Peptide Science* 21.7 (2015), pp. 522–529.

[19] M. C. Owen et al. “Effects of in vivo conditions on amyloid aggregation”. In: *Chemical Society Reviews* 48.14 (2019), pp. 3946–3996.

[20] L. C. Serpell. “Alzheimer’s amyloid fibrils: structure and assembly”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1502.1 (2000), pp. 16–30.

[21] P. Nguyen and P. Derreumaux. “Understanding amyloid fibril nucleation and Aβ oligomer/drug interactions from computer simulations”. In: *Accounts of chemical research* 47.2 (2014), pp. 603–611.

[22] S. Tomaselli et al. “The α-to-β conformational transition of Alzheimer’s Aβ-(1–42) peptide in aqueous media is reversible: a step by step conformational analysis suggests the location of β conformation seeding”. In: *ChemBioChem* 7.2 (2006), pp. 257–267.

[23] J. Baudier, J. C. Deloulme, and G. S. Shaw. “The Zn2+ and Ca2+-binding S100B and S100A1 proteins: beyond the myths”. In: *Biological Reviews* 95.3 (2020), pp. 738–758.

[24] K. Prez and L. Fan. “Structural Basis for S100B Interaction with its Target Proteins”. In: *Journal of molecular and genetic medicine: an international journal of biomedical research* 12.3 (2018).

[25] J. S. Cristóvão and C. M. Gomes. “S100 proteins in Alzheimer’s Disease”. In: *Frontiers in neuroscience* 13 (2019), p. 463.

[26] S. Malik et al. “Analysis of the structure of human apo-S100B at low temperature indicates a unimodal conformational distribution is adopted by calcium-free S100 proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 73.1 (2008), pp. 28–42.

[27] T. Ostendorp et al. “The crystal structures of human S100B in the zinc-and calcium-loaded state at three pH values reveal zinc ligand swapping”. In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1813.5 (2011), pp. 1083–1091.

[28] S. B. Carvalho et al. “Intrinsically disordered and aggregation prone regions underlie β-aggregation in S100 proteins”. In: *PLoS One* 8.10 (2013), e76629.

[29] L. He and S. Hiller. “Frustrated interfaces facilitate dynamic interactions between native client proteins and holdase chaperones”. In: *Chembiochem* 20.22 (2019), pp. 2803–2806.

[30] D. U. Ferreiro, E. A. Komives, and P. G. Wolynes. “Frustration in biomolecules”. In: *Quarterly reviews of biophysics* 47.4 (2014), pp. 285–363.

[31] M. Pais et al. “Early diagnosis and treatment of Alzheimer’s disease: new definitions and challenges”. In: *Brazilian Journal of Psychiatry* AHEAD (2020).

[32] J. Nasica-Labouze et al. “Amyloid β protein and Alzheimer’s disease: When computer simulations complement experimental studies”. In: *Chemical reviews* 115.9 (2015), pp. 3518–3563.

[33] T. J. Zielinski et al. *Quantum states of atoms and molecules*. 2005.

[34] R. Winter. *Atomic Physics*. URL: `https://users.aber.ac.uk/ruw/teach/327/hatom.php`.

[35] C. D. Sherrill. “An introduction to Hartree-Fock molecular orbital theory”. In: *School of Chemistry and Biochemistry Georgia Institute of Technology* (2000).

[36] N. Harrison. “An introduction to density functional theory”. In: *Nato Science Series Sub Series III Computer and Systems Sciences* 187 (2003), pp. 45–70.

[37] M. J. Frisch et al. *Gaussian 09*. Gaussian, Inc., Wallingford CT. 2009.

[38] V. H. Teixeira, A. S. C. Capacho, and M. Machuqueiro. “The role of electrostatics in TrxR electron transfer mechanism: A computational approach”. In: *Proteins Struct. Funct. Bioinf.* 84.12 (2016), pp. 1836–1843.

[39] M. D. Hanwell et al. “Avogadro: an advanced semantic chemical editor, visualization, and analysis platform”. In: *Journal of cheminformatics* 4.1 (2012), p. 17.

[40] S. Batsanov. “Van der Waals radii of elements”. In: *Inorganic materials* 37.9 (2001), pp. 871–885.

[41] M. O. Steinhauser and S. Hiermaier. “A review of computational methods in materials science: examples from shock-wave and polymer physics”. In: *International journal of molecular sciences* 10.12 (2009), pp. 5135–5216.

[42] D. van der Spoel et al. “GROMACS: Fast, Flexible, and Free”. In: *J. Comput. Chem.* 26 (2005), pp. 1701–1718.

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

[44] N. Schmid et al. “Definition and testing of the GROMOS force-field versions 54A7 and 54B7”. In: *Eur. Biophys. J.* 40.7 (2011), pp. 843–856.

[45] W. Huang, Z. Lin, and W. F. van Gunsteren. “Validation of the GROMOS 54A7 Force Field with Respect to β-Peptide Folding”. In: *J. Chem. Theory Comput.* 7.5 (2011), pp. 1237–1243.

[46] Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 2.4*. 2020.

[47] J. Hermans et al. “A Consistent Empirical Potential for Water-Protein Interactions”. In: *Biopolymers* 23 (1984), pp. 1513–1518.

[48] B. Hess. “P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation”. In: *J. Chem. Theory Comput.* 4.1 (2008), pp. 116–122.

[49] S. Miyamoto and P. Kollman. “SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models”. In: *J. Comput. Chem.* 13.8 (1992), pp. 952–962.

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

[51] S. Nosé and M. Klein. “Constant pressure molecular dynamics for molecular systems”. In: *Mol. Phys.* 50.5 (1983), pp. 1055–1076.

[52] M. Parrinello and A. Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *J. Appl. Phys.* 52.12 (1981), pp. 7182–7190. ISSN: 0021-8979. DOI: `10.1063/1.328693`.

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

[54] J. S. Cristóvão et al. “Cu^{2+}-binding to S100B triggers polymerization of disulfide cross-linked tetramers with enhanced chaperone activity against amyloid-β aggregation”. In: *ChemComm* (2020). DOI: `10.1039/d0cc06842j`.

[55] W. C. Wimley and S. H. White. “Experimentally determined hydrophobicity scale for proteins at membrane interfaces”. In: *Nature structural biology* 3.10 (1996), pp. 842–848.

[56] S. Genheden and U. Ryde. “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities”. In: *Expert opinion on drug discovery* 10.5 (2015), pp. 449–461.

[57] R. Kumari et al. “g_mmpbsa: A GROMACS tool for high-throughput MM-PBSA calculations”. In: *Journal of chemical information and modeling* 54.7 (2014), pp. 1951–1962.

[58] X.-Y. Meng et al. “Molecular docking: a powerful approach for structure-based drug discovery”. In: *Current computer-aided drug design* 7.2 (2011), pp. 146–157.

[59] S. J. De Vries, M. Van Dijk, and A. M. Bonvin. “The HADDOCK web server for data-driven biomolecular docking”. In: *Nature protocols* 5.5 (2010), p. 883.

[60] G. Van Zundert et al. “The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes”. In: *Journal of molecular biology* 428.4 (2016), pp. 720–725.

[61] C. Dominguez, R. Boelens, and A. M. Bonvin. “HADDOCK: a protein- protein docking approach based on biochemical or biophysical information”. In: *Journal of the American Chemical Society* 125.7 (2003), pp. 1731–1737.

[62] C. Geng et al. “Information-driven, ensemble flexible peptide docking using HADDOCK”. In: *Modeling Peptide-Protein Interactions*. Springer, 2017, pp. 109–138.

[63] G. C. van Zundert and A. M. Bonvin. “Modeling protein–protein complexes using the HADDOCK webserver “modeling protein complexes with HADDOCK””. In: *Protein Structure Prediction*. Springer, 2014, pp. 163–179.

[64] M. Coles et al. “Solution Structure of Amyloid β-Peptide (1- 40) in a Water- Micelle Environment. Is the Membrane-Spanning Domain Where We Think It Is?” In: *Biochemistry* 37.31 (1998), pp. 11064–11077.