**Lucie da Rocha**

Licenciatura em Bioquímica

# Molecular modeling study of β-lactoglobulin dimerization: a first step to hypoallergen design for immunotherapy

Dissertação para obtenção do Grau de Mestre em

Bioquímica para a Saúde

Orientador: António Baptista, PhD, ITQB-NOVA

Co-orientador: Sara R. Campos, PhD, ITQB-NOVA

**Abril, 2018**

**Lucie da Rocha**

Licenciatura em Bioquímica

# Molecular modeling study of β-lactoglobulin dimerization: a first step to hypoallergen design for immunotherapy

Dissertação para obtenção do Grau de Mestre em

Bioquímica para a Saúde

Orientador: António Baptista, PhD, ITQB-NOVA

Co-orientador: Sara R. Campos, PhD, ITQB-NOVA

Juri

Presidente: Prof. Doutor Pedro Manuel H. M. Matias

Arguente: Prof. Doutor Miguel Ângelo dos Santos Machuqueiro

Vogal: Prof. Doutora Margarida Archer Franco Frazão

**Instituto de Tecnologia Química e Biológica António Xavier da Universidade Nova de Lisboa**

**Abril, 2018**

# Molecular modeling study of β-lactoglobulin dimerization: a first step to hypoallergen design for immunotherapy

## ACKNOWLEDGMENTS

At first, I would like to share that I feel very grateful to have been integrated in the Molecular Simulation group at ITQB in order to develop the work of this thesis. I have finally found the path that makes me academically happy and that I truly want to follow. For this I owe Dr. Antonio Baptista and Dr. Sara Campos for giving me the opportunity to work with them and opening me the doors that lead to the world of computational simulation.

Starting with the main people responsible for this work, I would like to express my gratitude to Dr. António Baptista and Dr. Sara Campos, for their support, guidance and for all of their patience and availability during this last year. I thank Dr. Antonio Baptista for all the devotion and time spent teaching me so many different things in the most various areas and for making me questioning everything, leaving me curious and always wanting to know more, and Dr. Sara Campos for all her attention, correctness and insightful comments, without them this experience would not have been as pleasant and stimulant.

I would also like to acknowledge everyone in the Protein Modelling group for always being available for any help and for all their great disposition and fellowship.

To my mother, a big thank you, for all the support along these years, both moral and financial, this was indeed crucial for the accomplishment of this year duration work. But mostly for always believing in me and in my capabilities.

A special thank you to all my friends and family, for all their support and always trying to cheer me up when things were getting a little more stressful.

Last but not least, I would like to thank in particular to Ruben for being my biggest support, for all the love and continuous encouragement and for always knowing how to make me smile and see the brigtside of things when they were getting tougher.

Thank you all!

## ABSTRACT

Milk and its derivatives are important worldwide food sources, particularly for infant nutrition, but face a major health complication: some of their proteins are allergens, especially β-lactoglobulin (BLG), a major component of bovine milk.

The fate of BLG upon ingestion remains unsettled, being unclear how extensive BLG proteolysis is and how it affects allergenicity. The fact that proteolytic resistance and antigenic response remain related even for non-oral administration suggests that they are not causally related but rather reflect an underlying common feature.

This feature may be the formation of dimers, which can hinder proteolysis and seems to facilitate the binding of protein allergens to IgE antibodies; indeed, BLG is dimeric when complexed with IgE Fab fragments and shows lower antigenicity when in the monomeric form. As shown in experimental studies, this form is predominant at pH below 2 and above 9 and between these there is the formation of a reversible dimer at a moderate ionic strength.

Previous studies have shown that the dimerization involves electrostatic interactions, for which a better understanding at a molecular-level is essential. In this study, we used constant pH molecular dynamics method, for the monomer and dimer, and found that BLG dimer presents higher stability than the monomer. Also, when dimeric, it presents two different states: a compact state, that resulted from the tight association of the two monomers, observed at pH 3, and a relaxed state, with a less tight interface, observed for pH $\geq$ 5. The potential of mean force obtained from the umbrella sampling method corroborates the presence of these two states, and is in agreement with experimental dissociation rates. From the analysis of protonation features at the dimer interface, we propose two new monomer-inducing mutants, D137N and R148Q, that should disrupt an interface salt bridge and, thus, be good hypoallergen candidates.



**Keywords**: β-lactoglobulin, allergen, dimerization, pH-sensitive behaviour, constant-pH molecular dynamics, umbrella sampling

## RESUMO

O leite e os seus derivados são fontes alimentares importantes a nível mundial, particularmente para a nutrição infantil, mas enfrentam uma complicação importante para a saúde: algumas de suas proteínas são alergeneas, especialmente a β-lactoglobulina (BLG), um dos principais componentes do leite bovino.

Após ingestão, não é claro o destino da BLG, quão extensa é a sua proteólise, nem qual a origem da sua alergenicidade. A resistência proteolítica e a antigenicidade estão relacionadas, mesmo em administração não oral, o que sugere que não estão casualmente relacionadas mas que têm uma caraterística em comum.

Esta característica pode ser a formação de dímeros, que podem impedir a proteólise e facilitar a ligação de anticorpos IgE; aliás, a BLG encontra-se na forma dimérica quando ligada a IgEs e apresenta uma menor antigenicidade como monómero. Dados experimentais mostram que a pH abaixo de 2 e acima de 9 encontra-se na forma de monómero, e que neste intervalo de pH, forma um dímero reversível, a uma força iónica moderada.

Estudos anteriores mostraram que a dimerização envolve interações eletróstaticas, sendo que um melhor entendimento destas a nível molecular é essencial. Neste trabalho, utilizámos o método de dinâmica molecular a pH constante para estudar o monómero e o dímero e descobrimos que a BLG é mais estável quando dimérica. Além disso, o dímero apresenta dois estados: um estado compacto, que resultou de uma estreita associação entre os dois monómeros, observado a pH 3, e um estado relaxado, com uma menor interface de contato, observado a pH $\geq$ 5. O potencial de força média, obtido pelo método de *umbrella sampling*, corrobora a presença destes dois estados, estando também de acordo com as taxas de dissociação obtidas experimentalmente. Analisando a protonação dos resíduos na interface do dímero, propomos dois mutantes que induzem o monómero, D137N e R148Q, quebrando uma ponte salina na interface e, assim, podendo ser considerados para imunoterapia.



**Keywords**: β-lactoglobulina, alergénio, dimerização, comportamento sensível ao pH, dinâmica molecular a pH constante, *umbrella sampling*

# LIST OF CONTENTS

# FIGURE INDEX

# TABLE INDEX

# 1 INTRODUCTION

Milk and dairy products are a major component in the daily diet of humans, satisfying the nutritional requirements of children. Thus, by the 2015-2020 Dietary Guidelines for Americans recommendation [1], a consumption of three portions per day of milk, or equivalent milk products, is essential to promote a healthy life [2]. It is well known that milk and dairy products can cause an allergic reaction in susceptible individuals, especially in children under three years of age [3], whose origin, in some individuals, is due to the presence of the protein allergen β-lactoglobulin, the major whey protein of bovine milk. Since 84% of the milk production is bovine, this allergen must be paid attention to [2][4].

β-Lactoglobulin (BLG) is a small protein (18.3 kDa) containing 162 aminoacids [2][5] that can exist in the form of monomer, dimer or higher order aggregates and can be found in the milk of many ruminant species [5]. This protein, that has been widely studied over the past 60 years, belongs to the lipocalin superfamily of transporter molecules, binding small hydrophobic molecules (such as retinol, fatty acids, progesterone and others) into a hydrophobic cavity. Its function is not totally understood, but, due to this property, it is believed to act as a specific transporter [2][5]. This protein can be found in the milk of most mammals, but has not been detected in the milk of humans, rodents or lagomorphs [2].

Bovine BLG has two main genetic variants, A and B, that differ by two aminoacids in the polypeptide chain, Asp 64 → Gly and Val 118 → Ala [6], and a rarer variant, C, where Glu 59 of variant B is substituted by histidine [7]. In the present study, the focus will be on the A variant, since it is the most abundant variant in bovine milk [6][8].

## 1.1 Molecular Structure

BLG has been structurally characterized in both monomeric and dimeric forms by NMR [9][10][11][12] and X-ray diffraction studies [13][14][15][16]. Most NMR studies have been performed at pH ~2.5, very low ionic strength and moderate protein concentrations, conditions in which the protein tends to be monomeric, whereas X-ray diffraction studies where done over the pH range ~5.5-8.2, at low to high ionic strengths and high concentrations.

As shown in **Figure 1.1**, this revealed, for each monomer, a fold-up into an eight-stranded (strands A-H) antiparallel β-barrel forming a calyx structure, flanked by a three-turn α-helix on the outer surface. A ninth β-strand (I) forms an antiparallel interaction at the dimer interface [5][17],

stabilizing the two monomers by polar peptide-peptide interactions. Another stabilizing interaction comes from the external loop of AB strands, where a hydrogen bonded ion pair connects the residues of each monomer (e.g. Asp33 with Ala34 and Asp33 with Arg40) [17]. The α-helices that are antiparallel on the outer surface of the dimer have no direct monomer-monomer interaction, the closest elements being the flexible side chains of Asp137A and Arg148B, and vice versa [5].

The calyx structure is characterized by two β-sheets, where strands A-D form one sheet and strands E-H form a second [5]. It is in the hydrophobic calyx that the binding of ligands occurs [5][17]. The loops connecting the β-strands at the closed end of the calyx, BC, DE and FG are quite short, where those at the open end are longer and more flexible. The EF loop acts as the gate over the binding site, where the residue Glu89 is found. This residue was established to have a particularly interesting function associated with the Tanford transition [18], discussed below.

In the CD loop, Cys66 is linked with Cys160 near the C terminus, and Cys106 on strand G with Cys119 on strand H, leaving Cys121 as a free reactive thiol group [5]. This buried cysteine has a pH-dependent activity, being associated with denaturation and aggregation, particularly at low pH. It appears that reaction of Cys121 with almost any thiol reagent enhances dissociation into monomers [3][19][20].

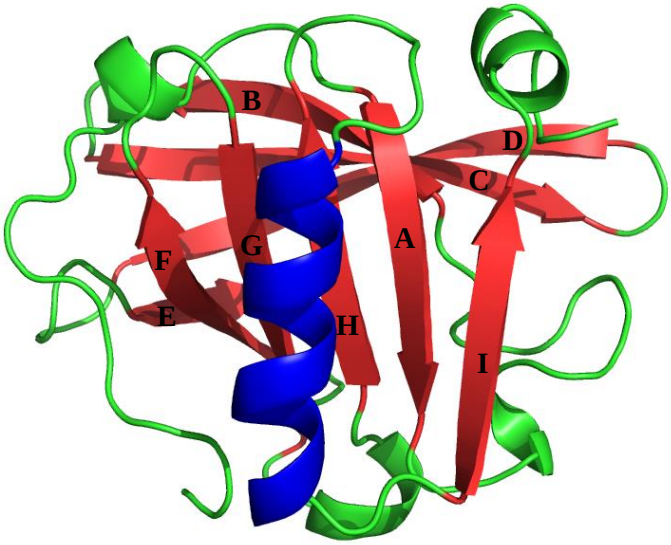


**Figure 1.1** Structural representation of BLG monomer. The β-strands are colored in red and labelled. The α-helix is shown in blue.

The anion-binding residues that are thought to be responsible for the dimer formation [21] are shown in **Figure 1.2.**.

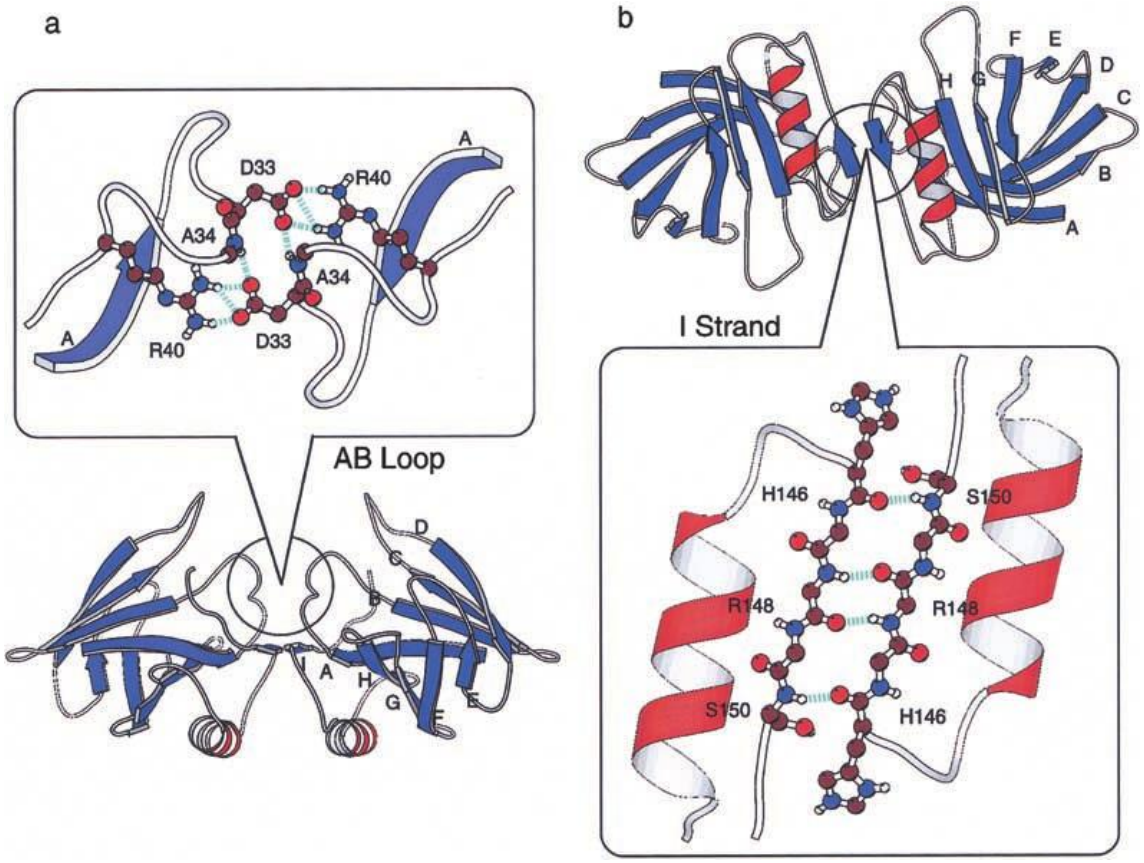


**Figure 1.2** Dimer interface of BLG. Side view (a) with respect to AB loops and top view (b) focusing on the βI strands, with hydrogen bonds interaction between side chains and main chains. Taken from[21].

## 1.2 Open/Closed binding site

In the late 1950´s, Tanford et al. [18] predicted the presence of a carboxylic acid residue with an anomalously high $pK_a$ value around 7.5, related to a conformational change at that pH. Later, this was attributed to a reversible pH-dependent transition of the EF loop associated with the protonation of Glu89 [16][13]- **Figure 1.3**. This transition is now called the Tanford transition [22], and, as shown by X-ray structure analyses, [13] at pH 6.2, the EF loop is in the "closed" position, that is, it covers the top of the barrel burying Glu89 inside the calyx, whereas, at pH 8.1, the glutamic acid is in its ionized form, exposing its buried chain and moving the loop away from the barrel. NMR studies at pH ~7.0 using an Ala34Cys mutant that forms a covalently linked dimer, were unable to detect this conformational change. However the timescale at which it occurs is probably too slow to be detected [7].

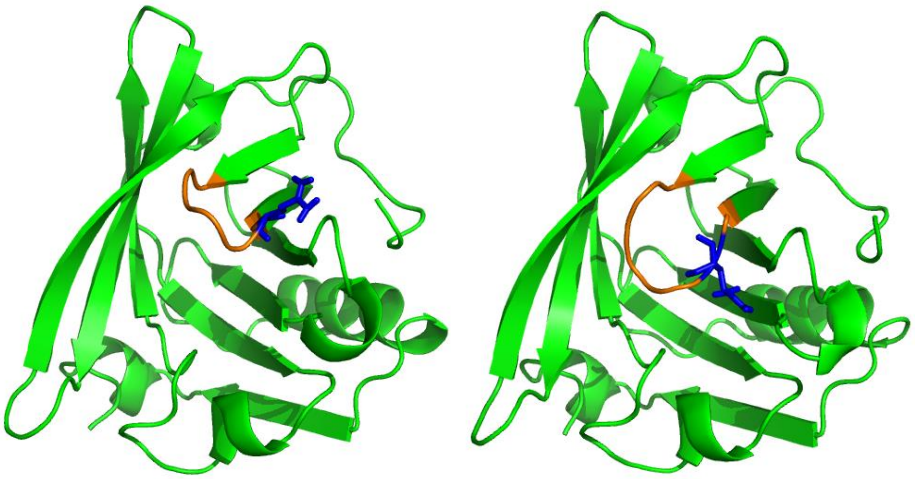

**Figure 1.3** Open (left) and closed (right) conformation of BLG. Orange: loop near the binding site; Blue: Glu89.

## 1.3 Monomer or Dimer

Protein association occurs in many cellular processes to various extents and under different conditions, and understanding this phenomenon constitutes an important part in a variety of pharmaceutical processes.

NMR and X-ray crystallography studies have elucidated the tertiary and quaternary structure of BLG and have found that physical as well as chemical modifications induce changes in the protein structure, affecting their interfacial properties. A presence of two predominant forms is observed under physiological conditions: the monomeric and dimeric forms [7][23]. At pH below 2 and above 9, BLG tends to preserve its secondary and tertiary structure while being monomeric, whereas between these pH values, its quaternary structure adopts a reversible dimer at moderate ionic strength (~0.1M) [23].

### 1.3.1 Monomer-Dimer Equilibrium

The equilibrium between these two forms has been widely studied over the past 60 years using essentially every possible biophysical technique (SAXS, light scattering, sedimentation and diffusion, isothermal titration calorimetry, etc) [21][17][24], but with somewhat unclear results when comparing the literature data due to differences in experimental conditions (see table S1 in reference [7]). Several researchers have determined that this association is mainly affected by pH, ionic strength, protein concentration and temperature, all these being factors correlated with each other.

The pH effect seems to have an essential role in this association. At moderate ionic strength (~0.1 M) and room temperature (20 ºC), the dimer predominates in pH range 2-9, whereas, at lower and higher pH values its monomeric form is predominant [7][21][25][26][9], suggesting that electrostatic interactions play a significant role in BLG association. Indeed, alterations in the ionic strength reinforce the effect of these interactions. Increasing the ionic strength (0.001-1 M), especially at low pH, will cause a significant increase in the fraction of BLG in the dimeric form [25][21][7][27][26]. This can be explained by a shielding effect caused by the ions in solution on the interaction of the two identically charged proteins. It is worth mentioning that Majhi et al. [23] also observed that the aggregation rates depend significantly both on the ionic strength and the pH; although this rate is not directly related to the dimerization constant, it shows the marked dependence of the association process on these two parameters. At pH values near the isoelectric point (~5.2) [24][18][28], the association is maximized, decreasing as |pH-pI| increases. Conclusive evidence was reported for an octameric form [23][24] near the isoelectric point (pH ~5.2) at a moderate ionic strength.

In one of the most recent experiments, Mercadante et al [7] have measured dissociation constants for form A at an ionic strength of 0.1 M, temperature of 25 ºC, and pH values of 2.5, 3.5, 6.5 and 7.5, obtaining values of 14.8, 4.0, 4.0 and 10.8 μM, respectively. In addition, they also measured the rate constants for BLG dimer dissociation and association ($k_{off}$ and $k_{on}$), under the same conditions, and demonstrated that they were lower at low pH and greater at higher pHs: $k_{off}$ is 0.008 and 0.007 $s^{-1}$ at pHs 2.5 and 3.5, respectively and >0.1 $s^{-1}$ at pHs 6.5-7.5; $k_{on}$ is 540 $M^{-1}$ $s^{-1}$ at pH 2.5, > 25,000 $M^{-1}$ $s^{-1}$ at pH 6.5 and > 9,200 $M^{-1}$ $s^{-1}$ at pH 7.5.

BLG aggregation is also tied in with temperature effects. An increase in the temperature to about 50ºC-80ºC exposes the hydrophobic surfaces that are buried in the protein interior due to a partial unfolding, including the already mentioned Cys121 (section **1.1**). This protein reorganization alters the monomer-dimer equilibrium in a pH-dependent way [20][21][29]. The free reactive thiol of Cys121 buried in the protein, when reacting with a thiol-reactive agent, promotes the unfolding of the structure at pH 7, producing a monomeric species [19]. This free group also plays a crucial role when induced by heat at different pH values. Some studies have shown that the monomeric form is enhanced at low pH [21][29], as opposed to neutral and high pH values [20] where the formation of thiol-disulfide exchange reactions, potentially with other monomers, occurs.

Other subtle features in BLG conformation have an impact on its dimer formation. The conformation of the α-helix is known to change at low pH [11] which may therefore affect the protein stability and quaternary structure, but this result was observed in a recombinant protein and by comparing NMR and X-ray structures, which should be taken into account.

The protein concentration also affected the monomer-dimer equilibrium, being that an increase in concentration promotes dimerization [30][25].For example, using Mercadante et al [7] dissociation constants, we can see in **Figure 1.4**, that the fraction of dimer is very abundant at concentrations near the concentration of BLG in the milk [2], while it decreases considerably with a decrease in protein concentration. Despite this, the BLG concentration used in the literature varies widely between 5 µM and 1 mM, including concentration in milk (~ 3 g/L ≈ 160 µM) and 0.001M.

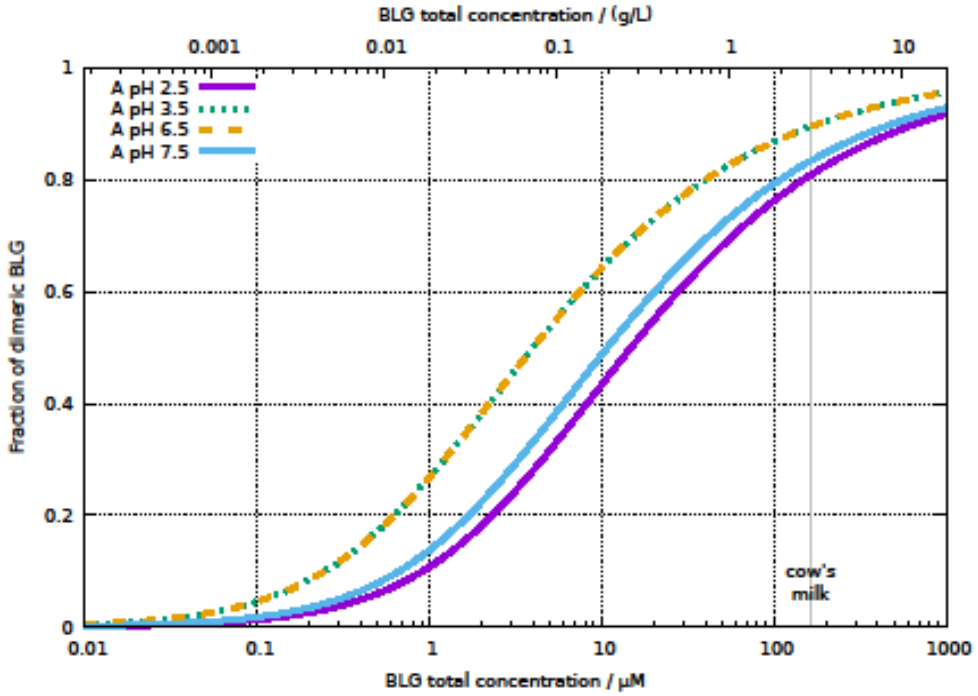


**Figure 1.4** Fraction of dimeric BLG as a function of its total concentration. The ionic strength used was 0.1M. Adapted from [7]

## 1.4 Allergenicity

The type of allergic response that BLG causes in susceptible individuals belongs to the class of type I hypersensitive reactions [30] that are based on the binding, in a cross-linked way, of IgE antibodies to epitopes, i.e., minimal peptides in the surface of the protein that can be recognized by the immune system [4].

These antibodies are secreted by mature B cells and although they bind to the epitopes they also bind to receptors, FcεRI, on the mast cells or basophil surface which, upon cross-linking, will trigger an allergic reaction by degranulation of biological mediators, like histamine and lipid mediators [4][30]- **Figure 1.5**.

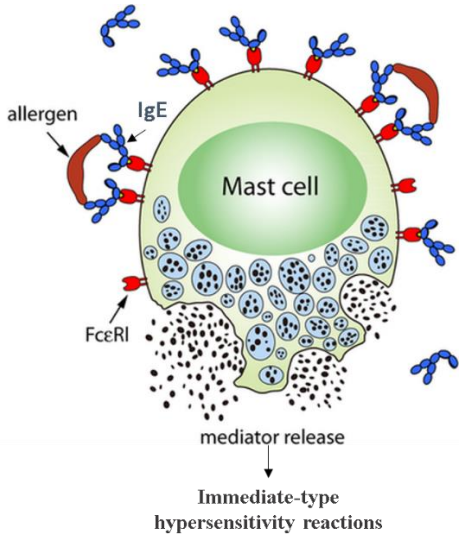


**Figure 1.5** Type I hypersensitivity reaction mediated by factors released by mast cells as a response to the binding of IgE antibodies. The FcεRI molecules make the connection between antibody-epitope. Adapted from [4].

The characteristics of an allergen are difficult to predict and, although the presence of epitopes is essential, this alone is not enough and it seems that the contribution of other factors is significant, like resistance to proteolysis, post-translational glycosylation and enzymatic activity [4]. These factors may be involved in the well-known allergenicity of BLG, although the exact causes remain unclear.

### 1.4.1 Resistance to Proteolysis

Since protein allergens are stable in the gastrointestinal track long enough to provoke an immune response, resistance to proteolysis and allergenicity seems to be correlated [4]. Indeed, the resistance to proteolysis of BLG seems to affect its allergenicity. Due to its stability at acidic pH, BLG seems to resist digestion in the stomach, and so, is able to act as a transporter of bound ligands, like vitamin A and retinol to the small intestine [3]. Uhrinova et al. [11], suggests that due to its closed conformation at the pH of the stomach the ligands would be protected and could be released upon the opening of the barrel within the basic small intestine. Therefore, intact BLG is absorbed in the small intestine, triggering the allergenic reaction.

This stability is thought to be caused by its two disulfide bonds, one located close to the C-terminal and accessible to the surface (Cys66-Cys160) and the other located between 2 β-strands (Cys106-Cys119). Indeed, the study of del Val et al. [3] proved that for the BLG dimer, the reduction of these bonds increased its sensitivity to pepsin digestion and reduced its allergenicity. However, this instability only diminishes the allergic reaction, since some epitopes may remain intact and still trigger an immune response in predisposed individuals [4].

### 1.4.2 Dimerization

The availability of epitopes is often directly related to dimerization. When in the monomeric form, the allergenicity appears to be absent or reduced, but it increases in the dimeric form, most probably due to an increase in the number of epitopes, facilitating the cross-linking of IgE antibodies to protein allergens [30]. Indeed, BLG is dimeric when complexed with an IgE antibody as observed in a X-ray crystal [31]. In this study, the epitopes are found in a flat area of the BLG surface in each monomer, consisting of six different short fragments of the polypeptide chain, located especially in the β-strands- **Figure 1.6**.

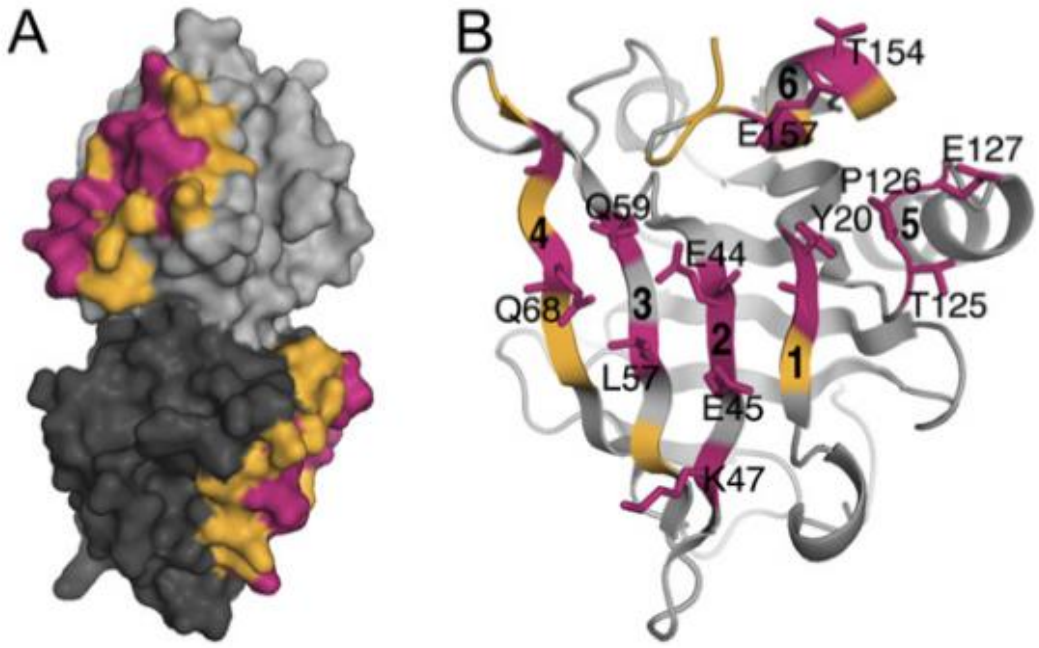


**Figure 1.6 A)** Location of two epitopes in the BLG dimer. **B)** The IgE-binding epitope of BLG. The residues making contacts are shown in purple, and the residues that are buried upon the binding of IgEs are shown in orange. The different segments of the epitope are numbered 1-6. Taken from [31]**.**

The first four β strands are where the epitopes are mostly located: epitope residues W19-Y20 in strand A, V43-K47 in strand B, L57-Q59 in strand C and C66-Q68 in strand D. Part of the epitope is also formed by the loop between strand H and the following helix, P126-E127, and the C-terminal α-helix, T154-E157. The antibody binding does not affect the binding of ligands [31].

Dimerization is generally not a pre-requisite for antigenicity, as monomers can trigger the cross-linking of FcεRI with two different IgE antibodies, if there is a sufficient number of epitopes. However, we do not know if it is the case for BLG. In case of its allergenicity being correlated with its dimerization, BLG can be subjected to different physiological conditions that affect both its stability and dimerization and, consequently its allergenicity.

Despite the usually low gastric pH (~pH 2), there is a sharp increase to ~pH 5 upon meal ingestion, that gradually returns to its standard value after some hours [32]. Thus, after ingestion, BLG passes from a pH 6.36 in the milk [33] to a pH near its isoelectric point, where maximum dimerization is expected. In the small intestine, the pH is around 6.5 in the fasted state and slightly decreases during digestion [32], meaning that BLG would also be at a pH close to its isoelectric point, thus favoring the dimeric form. This is further affected by BLG concentration, which would be close to the milk´s concentration after ingestion (see Figure 1.4), and may subsequently change during digestion.

### 1.4.3 Reducing allergenicity- Mutations

An approach to overcome the problem of allergenicity could be the production of mutants that could work as hypoallergens. Indeed, it is known that site-directed mutations can alter structural features of BLG.

It is assumed that the two I strands and the two AB loops play an important role in dimer formation, thus, mutations in sites belonging to the interface and that are thought to contact with each other could reduce the formations of dimers. Evidence of mutations of I-strand residues to Pro and disruption of the salt bridge between Asp 33 and Arg 40 in the AB loops (by mutating these residues to oppositely charged residues) is known to weaken the dimer stability [34] (see **Figure 1.2**); this also reinforces the role of electrostatic attraction between these residues. When in the monomeric form, the number of epitopes may be insufficient, preventing the IgE cross-linking and allergic reaction. This is indeed corroborated by the fact that a mutation in His 146 to Pro (strand I) reduces the dimerization, and also the allergic response [30]. Site-directed mutations of the IgE-binding epitopes are also an alternative, which will consequently reduce the binding of IgEs antibodies [4].

As described previously apropos of the reduction of the disulfide bonds, computer modelling studies suggested that site-directed mutations disrupting these bonds affect the accessibility of epitopes and the stability of BLG [3]. Indeed, these effects can be seen in the mutation of free Cys121 by Ala, Ser and Val, being more pronounced with C121S, due to the introduction of a polar group in the hydrophobic environment [35]. This mutant is more susceptible to temperature and disulfide reducing agents, and, probably because of that, also to peptic digestion [36], potentially reducing its allergenicity.

## 1.5 Studying monomer-dimer equilibrium by molecular modelling methods

Taking all these observations into account, we can identify the pH and electrostatic interactions as an essential contributing factor for BLG aggregation. However, this mechanism has not been fully elucidated and comparison and interpretation of the literature data is very difficult, due to the variety of experimental conditions used.

BLG presents a pH-dependence of the monomer-dimer equilibrium, related to allergenicity, and of the regulation of a gating mechanism involving the EF loop associated to ligand binding. Computational studies have been used to understand the molecular phenomena underlying experimental observations, to complement and interpret their results. However, these methods have focused on electrostatics [27][7][23] or on structural dynamics separately [37][38], thus ignoring the direct effect that pH has on these two factors when combined. Constant-pH molecular dynamics (CpHMD) is able to overcome this problem by simulating the coupled association to changes of protonation and conformation at a given pH. This work intends to use CpHMD to understand the molecular events underlying the pH-sensitive behaviour of BLG at an atomic detail. This can be combined with umbrella sampling (US), providing the free energy at each distance between the two monomers.

The main goal is to study the dimerization of BLG at different pH values by performing these steps:

1) Study the monomeric and the dimeric forms of BLG in solution, using continuum electrostatics methods;

2) Study the monomeric and the dimeric forms of BLG in solution, using CpHMD methods;

3) Study the gradual dimerization process of two BLG molecules, combining the CpHMD method with umbrella sampling;

4) Propose BLG mutants likely to produce monomeric variants to be used as hypoallergens.

These studies also allowed to investigate the pH-dependent gating mechanism of the EF loop associated with the Tanford transition.

# 2 MOLECULAR SIMULATION

When addressing a biological problem, a usual approach is to try to solve it by experimental methods. These methods are known to be fast and efficient, however studying some properties of a system can be challenging in various aspects. Some biological systems are very complex and measuring some details at an atomic level can be challenging, not only in terms of complexity but also on the resources and time spent. Computer simulations can provide an insight on that [39]. They are able to give additional information on microscopic properties, by analyzing dynamics behaviour, thermodynamics patterns, and others, thereby predicting also the macroscopic behaviour [40]. The system is represented by a model, the initial configuration of which can be obtained from different sources (e.g. X-ray diffraction) and, with the appropriate energetic description chosen and a time of simulation long enough, the properties of interest can be studied [41]. The computed properties obtained can be compared with experimentally measured properties or, if computing unmeasurable properties, they can be used to approximate theories applied to the same system, thus testing the theory.

At first sight it may not be obvious why these studies are needed, but its prediction of specific properties of a system can become a powerful tool, e.g. measurements in certain conditions in laboratory can become challenging, expensive or even impossible when we want to work at high temperature and pressure [42]. Computer simulations can do that type of measurements [42].

Despite the increasing computer power and greater accuracy, it is not to be forgotten that computational methods have an associated error and the results that we obtain are never in direct relationship to the experimental or theoretical ones [42]. Therefore, although experimental studies are always preferable, computer simulations are a good alternative when they become too challenging. This type of simulations should be seen as a good complement that together with experimental measurements provide a way to solve problems in such approaches that were unthinkable in the past and that will make a difference in the scientific progress [42].

This chapter describes the early history of computer simulation and its progress followed by a brief description of the theory involved as well as the methods used in our studies to perform molecular simulations.

## 2.1 Historical overview

Computers appeared during the Second World War as a tool to help develop nuclear weapons and code breaking. In the beginning of 1950's they became available for non-military use being that scientists showed an interest on these machines, marking the beginning of the field of computer simulation [42]. The first computer simulation was performed in 1953 at the Los Alamos National Laboratories in the United States on a liquid. This work considered molecules as rigid spheres and hard-disks, introducing the "Monte Carlo" method that we know today and still use in simulations [43]. Until then, liquids were studied by mechanical simulation in which molecules were manipulated as large packed gelatine balls, but the gravity effect could not be eliminated and some interactions between them were neglected. It was only a few years later that the Lennard-Jones potential was implemented by Wood and Parker [42].

The first proper molecular dynamics (MD) simulation was accomplished in 1957, by solving the Newton's equation of motion for an assembly of rigid spheres and then, by using Lennard-Jones particles [44]. In 1964, the first MD simulation of liquid argon was reported by Rahman, being the first real liquid do be used in this type of simulations [44].

The following years were marked by a large increase in computational power and, in the early 1970s, a study on molecular dynamics of a diatomic molecular liquid appeared followed by two studies of water [41]. Until then, only monatomic studies had been performed [41].

Analyses of more complex molecules began to take place: small rigid molecules, hydrocarbons and macromolecules, such as proteins, allowing a better understanding of their behaviour. With that, new methods were implemented, e.g. "stochastic dynamic" methods and quantum mechanical effects were incorporated [44]. In the following years, molecular mechanics force fields with energy minimization were implemented, leading to modern versions making use of simulated annealing and related methods.

Despite the increase in computational power, it was difficult for molecular simulation to gain some credibility and only in the late 1970s it was recognized and able to be used as an open door for problems solving. The 80's were marked by simulations of biomolecules of increasing size [41].

Nowadays, computer simulations allow us to analyze larger molecules and for a longer time scales than decades before and with high predictability of their atomic properties, but, comparing to nature, it is still many orders of magnitude slower [41]. In hindsight, we can see that much of the methodology used in this type of simulations has been developed since then, although their basic algorithms have changed little [42] since the first MD simulation appeared. The next years will

probably be characterized by faster simulation times and more complex biological systems without the need of sacrificing other simulation aspects.

## 2.2 Statistical Mechanics

Typical experimental studies measure thermodynamic properties that are defined by a small set of parameters (such as the temperature, the pressure and the molecular concentrations), reflecting a global behaviour that is averaged over a large number of particles and over the time of the measurement [42], obtaining the macroscopic properties of a system. But atoms are constantly interacting with each other and with their environment, thus it is that dynamic motion that allows to range all thermally-accessible space, thereby obtaining information about the structure and dynamic properties of a system [40].

Computer simulation studies provide information about each individual atom or molecule using a microscopic approach, but, what we are usually interested in is an average or a fluctuation of a property (e.g. the fraction of secondary structure, the ratio between two molecular forms, the distance between two molecules and the protonation of titrable sites) of the simulated system [45]. But calculating the whole real trajectory for observing this property is unfeasible. Instead, shorter trajectories when taken together should be equivalent to the whole trajectory. The bridge that makes the connection between the microscopic behaviour and macroscopic properties is made by the laws of statistical mechanics and it allows us to express thermodynamic properties in terms of microscopic quantities [44]. These laws use different representative ensembles, where each ensemble gives a probability distribution of the possible microstates states of the system (positions and velocities of the atoms).

These ensembles are generated using several copies of the system, where some thermodynamic properties are kept constant, [45] and since each copy has the same probability distribution as the others, a thermodynamics consistency between those simulations is attained [44]. Such a system is called “ergodic”.

Macromolecular systems can be studied by computer simulation using methods that generate the proper ensemble distribution. Indeed, we can see this applied in MD with the coupling of the temperature and pressure bath (as described in section **3.4**) and with the sampling of protonation states using the Monte Carlo approach (as described in subsection **2.5.3**).

## 2.3 Periodic Boundary Conditions

In computer simulations the molecular system is placed in a space-filling shaped box (e.g. cubic) of finite size. This will have an effect on the molecule due to the edges of the box, the interactions felt there will be different than in the center where it does not have a wall and it is only surrounded by the solvent [45]. This is an artifact of this type of simulations and to overcome these edge effects periodic boundary conditions are taken into account [41].

A simulation is performed using a small number of particles that experience forces as if they were in a bulk fluid [46]. For that, the box containing the molecular system is treated as if it were surrounded by an infinite number of copies of identical images of itself, simulating an infinite system and thus eliminating the boundaries. During the simulation time, the atoms move inside the box, where each atom experiences the same changes as its image in every copy, so, when one atom leaves the central box by crossing the boundary on one side it will enter on the other side of the translated image position (shown by the vectors in **Figure 2.1 a)**).

These atoms are restricted by a cut-off radius (this topic will be addressed in subsection **2.4.1)** applied on each atom so that they only take into account the interactions with nearest neighbors, i.e. closest periodic image [40]. An atom should not simultaneously interact with another atom and its periodic image, so the length of the box is chosen to be at least twice the cut-off radius of an atom. The grey atom in **Figure 2.1 b)** will only interact with atoms or images of atoms that lie within the cutoff [41][45].

The type of box to be chosen depends on the system under study. It should adequately accommodate the number of solvent molecules that is needed to fill the remaining empty space in the box using the minimum number of molecules possible, thus optimizing the time of simulation needed. For example, if simulating a spherical solute, the use of a spherically shaped computational box, like a truncated octahedron, is more adequate that a rectangular or cubic [41].

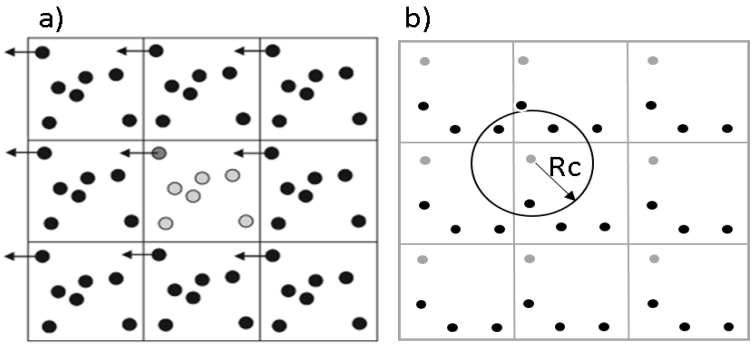


**Figure 2.1 a)**- This figure shows the central box, where the central atoms are colored in grey, and their copies in black. Each copy is identical at an atomic level. **b)**- Interactions of the black atom that are possible within the cut-off radius applied. Adapted from [41][45].

## 2.4 Molecular mechanics

Molecular mechanics (MM), is the description of molecular systems using classical mechanics, in order to treat them using Newton´s equation of motion (as described in **section 2.5**) [42]. The molecules are treated as balls (atoms) connected by springs (bonds) that interact with each other by bonded and non-bonded interactions with their nearest neighbours where physical forces predict its structure and dynamics, i.e. the molecules stretch, bend and rotate about those bonds [45].

The system is described using the Born- Oppenheimer approximation as an empirical function of potential energy where the electrons are implicit variables of this function and only nuclei are considered. This simplification allows to attain the computer speed required to perform this type of calculations and, more important, to achieve the required chemical accuracy [39]. This overcomes the problem in quantum mechanics of addressing a significant number of atoms.

This type of mathematical equations, which depend on the nuclear atomic coordinates and describe all intra and inter molecular interactions, are called force fields and it is their use in computer simulation that defines a realistic atomic model [39][47]. For different types of atomic models, different force fields are specified.

A typical energy function for a system is shown in **equation 2.1** where the total potential energy, $V(\mathbf{R})_{\text{total}}$, is decomposed into two terms [39] where **R** represents the coordinates of all atoms.

$$V(\mathbf{R})_{\text{total}} = V(\mathbf{R})_{\text{internal}} + V(\mathbf{R})_{\text{external}} \qquad \textbf{2.1}$$

The total potential energy accounts for the contribution of the covalently bonded atoms, $V(\mathbf{R})_{\text{internal}}$, and the non-bonded interactions, $V(\mathbf{R})_{\text{external}}$ [39].

These two terms can be described as the sum of the following parameters- **equation 2.2** and **2.3**.

$$V(\mathbf{R})_{\text{int}} = \sum_{\text{bonds}} \frac{1}{2} K_b (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi \left[1 + \cos\left(n_\chi \, \delta_\chi - n_\chi \sigma_\chi\right)\right] \qquad \textbf{2.2}$$

$$V(\mathbf{R})_{\text{ext}} = \sum_{\substack{\text{nonbonded} \\ \text{atom pairs}}} \left( \varepsilon_{ij} \left[ \left( \frac{r_{m,(ij)}}{\mathbf{r}_{ij}} \right)^{12} - 2 \left( \frac{r_{m,(ij)}}{\mathbf{r}_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r \mathrm{r}_{ij}} \right) \qquad \textbf{2.3}$$

For the internal potential energy, the first and second term describe the covalent stretching along bond b and the angle bending (three-body) interaction. These two terms are treated as harmonic potentials where the bond (b) and angles (θ) are kept near their equilibrium values, $b0$ and θ0, respectively. $K_b$ and $K_\theta$ correspond to the force constant of the first and second term, respectively.

The third term represents the dihedral or torsion angles, that is, the rotations that occur around a bond including the parameters for the force constant, $K\chi$, that dictates the barrier of energy of the rotation, the periodicity or multiplicity, $n$, indicating the number of cycles per 360º rotation and the phase $\sigma$ which stands for the location of the maxima in the dihedral energy surface [39].

Due to the impact of the environment on the properties of macromolecules and to the large number of non-bonded interactions that occur in biological molecules (e.g., hydrogen bonds involved in the secondary structures of proteins), the external potential energy terms (**equation 2.3**) is considered one of the most important energy term for computational studies. It accounts for the contribution of van der Waals and Coulombic (electrostatic) interactions between non-bonded pairs of atoms.

The parameters to treat this external term are the well depth, εij, between atoms i and j, the minimum interaction radius, $r_{m,(ij)}$, and the partial atomic charge, qi. The term in square brackets is used to treat the van der Waals (VDW) interactions, which include the force between instantaneously induced dipoles (London´s dispersion force). These interactions are treated as a Lennard-Jones (LJ) 6-12 potential, which represents the interaction when the electron clouds responsible for the attractive and repulsive components of the interatomic interactions remain localized close to the individual atoms [39].

The other term that contributes to the external potential energy is the Coulombic interactions. These interactions describe the interaction between partial atomic charges, $q_i$ and $q_j$, on atoms i and j divided by the distance $r_{ij}$, with a dielectric constant taken into account [39].

Overall, **equation 2.1** describes the physical interactions, shown in **Figure 2.2**, that dictate the structure and dynamic properties of biological molecules [39], although some variants do exist.

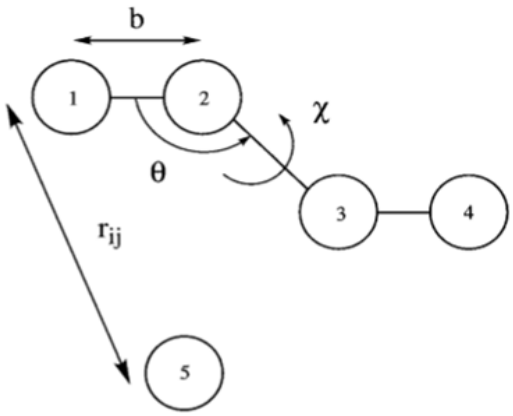


**Figure 2.2** Representation of the four types of interactions defined in equations 2.2 and 2.3. Molecule A comprises atoms 1-4 and molecule B comprises atom 5. Between the atoms 1 and 2, 2 and 3 and 3 and 4 are the bonds, b. The angles θ involve atoms 1-2-3 and 2-3-4 and there is a dihedral angle χ between atoms 1-2-3-4. Molecule B is interacting with molecule A with all four atoms by VDW and Coulombic interactions which can occur between all pairs of atoms within an interatomic distance rij [39].

The potential energy functions shown in **equations 2.2** and **2.3** are the base of energy functions that are included in the several force fields used in computational studies. A force field can be defined by two routes of parameters: fit the parameters to results of ab-initio quantum calculations on small molecular clusters or fit them to experimental data (X-ray data on small molecules, free energies of solvation, spectroscopic data, etc) [41].

### 2.4.1 Cut-off, neighbours list and twin-range methods

The last term in the potential energy function, the non-bonded interactions, accounts for the van der Waals and Coulombic interactions and it runs over all atom pairs in the molecular system. Calculating these non-bonded interactions is the most time consuming part of the potential energy function. Thus, applying a cut-off radius Rc (see **Figure 2.1 b)**) that only takes into account the atoms interactions lying within it, reduces the time spent in the calculation without introducing any errors, as long as the range of the non-bonded interactions is smaller than Rc. These neighbour atoms, i.e., those atoms within the cutoff distance, are added to a neighbour list, whereby only the distance to atom pairs included there are calculated [46], being updated every so many simulation steps (about every 5 or 20 MD time steps) [41].

Since Coulombic interactions are long-ranged, its interactions that lie outside this cutoff are taken into account by a dielectric constant where its value is given according to the reorganization of favorable new interactions, unlike van der Waals interactions that are short-ranged.

One of the methods that allows to do that is the twin range method, illustrated in **Figure 2.3**, in which two cut-off radiis are used: the interactions below the first cutoff are evaluated as normal at each step. Interactions between the first and second cut-off are recalculated only when their neighbour list is updated (about every 5 or 20 MD time steps) and are kept constant between these updates[46].

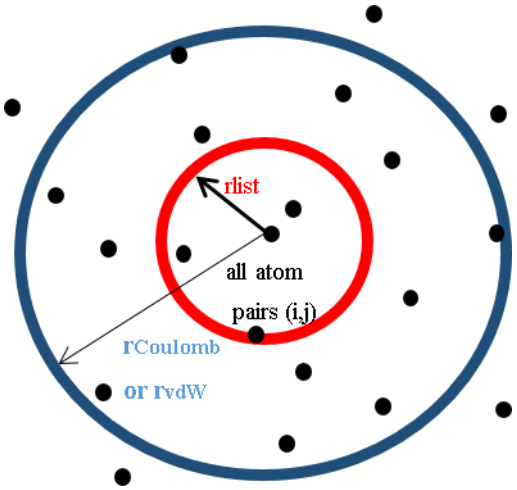


**Figure 2.3** Twin range method applying a cutoff for the van der Waals interactions, rvdW, and for the Coulombic interactions, rCoulomb. Adapted for [41].

### 2.4.2 Constraints and Restraints

Constraints and restraints are defined in specific atoms of the molecular system with the purpose of restricting some degrees of freedom, affecting its conformational/rotational geometry. They are two different concepts. On one hand, a constraint is a requirement that the system is forced to satisfy throughout a simulation, like specific values for bonds or angles. On the other hand, when a restraint is applied it encourages the system to adopt a particular value by imposing penalties when it deviates from that desired value [46].

## 2.5 Molecular Dynamics

After an appropriate force field for the model of interest has been chosen, the dynamics of the system must be simulated to obtain the average of the properties of interest. For that, molecular dynamics (MD) methods are used [46], integrating Newton ´s equations of motion.

The state of the system at any future time can be predicted from its current state with the sequential derivation of sets of atomic positions and velocities, generating new conformations [46]. These new conformations are obtained according to the potential energy of the preceding states, which

is calculated using a MM function that is time-independent, accounting only for the interatomic distances of all the atoms in the system. Thus, the potential energy of the system is obtained using a Molecular Mechanics force field while the sampling of different conformations is obtained using Newton´s equation, obeying the laws of classical mechanics [46].

For a system consisting of *N* particles, a force field is used to obtain the gradient of potential energy, *V*, calculating the forces acting on each atom, $\mathbf{F}_i$, which are given by the negative of the first derivative of potential energy, *V*, with respect to positions ($\mathbf{r}_1$,$\mathbf{r}_2$,…$\mathbf{r}_N$), as shown in **equation 2.4**:

$$\mathbf{F}_i = -\left(\frac{\partial V}{\partial \mathbf{r}_i}\right), i = 1, \ldots, N \qquad \textbf{2.4}$$

The forces acting on each atom are then integrated in Newton´s second law to calculate the new atomic positions:

$$\mathbf{F}_i = m_i \mathbf{a}_i \leftrightarrow \mathbf{F}_i = m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} \qquad \textbf{2.5}$$

where m*i* is the mass of the particle *i* and **a***i* is its acceleration [45].

After the assignment of initial coordinates and velocities at time zero, the potential energy is computed (**equation 2.4**) and the **equation 2.5** is integrated performing the calculation in small time steps, where the forces acting on the atoms are kept constant. The atoms then have new molecular positions and velocities and another potential energy calculation and integration step takes place. These two equations are repeatedly solved until the system has performed a time of simulation long enough that the system reaches equilibrium resulting in a trajectory along time [46] where its dynamic motions provide useful structural and dynamic information, i.e. concerning molecular geometries and energies, atomic fluctuations, etc [41].

For the system to have a good integration of the Newton´s equations of motion, the choice of an accurate algorithm that predicts the trajectory of all atoms is crucial. There are several types of algorithms but one of the most used and accurate is the Leap Frog algorithm [45]:

$$\mathbf{v}_i\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}_i\left(t - \frac{1}{2}\Delta t\right) + \frac{\Delta t}{m_i}\mathbf{F}_i(t) \qquad \textbf{2.6}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \Delta t \mathbf{v}_i \left(t + \frac{1}{2}\Delta t\right) \quad \textbf{2.7}$$

The atomic positions, **r**, at time *t*, and the velocities, $\mathbf{v}_i$, at time $t + \frac{\Delta t}{2}$ are used to obtain the new positions and velocities using the forces, $\mathbf{F}_i(t)$, on each atom.

This will allow for the velocities to "leap over" the positions and then the positions to "leap over" the velocities [46] as it is shown in **Figure 2.4**.

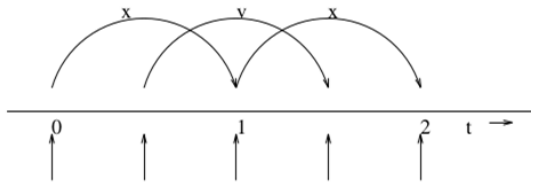


**Figure 2.4** The Leap-Frog integration method [68].

The choice of the time interval, Δt, must be chosen carefully taking into account the properties of the molecular system and it must be considerably smaller than the characteristic time of the fastest motion studied [39].

## 2.6 Protein electrostatics

Electrostatic interactions are long-ranged with a large force constant and fundamental in all biomolecular systems due to charged and polar groups found ubiquitously in biological macromolecules. Among these, proteins are particularly relevant due to their charged and/or polar aminoacids in which electrostatics will play an important role in its structure and function. Indeed, electrostatic effects are involved in many biological phenomena like protein stability (much due to the presence of hydrogen bonds), enzyme catalysis (lowering the energy barrier), biomolecular recognition, protein-protein association, among others [48]. Changes in pH will alter electrostatic interactions due to variation in the protonation/deprotonation in titrable sites, changing the protein´s total net charge [47]. In solution, these variations are due to the pH of the medium and the tendency of their ionizable groups to receive/lose protons.

The ionization state of a titrable aminoacid depends on its equilibrium acidity constant, $K_a$, [49] that can be determined by a titration of an aqueous solution containing the aminoacid of interest:

$$AH \rightleftharpoons A^- + H^+ \quad \textbf{2.8}$$

$$K_a = \frac{[A^-][H^+]}{[AH]} \quad \textbf{2.9}$$

After obtaining the equilibrium constant, the p$K_a$ can be calculated:

$$pK_a = -\log K_a \quad \textbf{2.10}$$

Where its variation according to the pH solution is given by the Henderson-Hasselbalch equation:

$$p\mathrm{H} = pK_a + log\frac{[A^-]}{[HA]} = pK_a + \log\frac{f}{1-f} \quad \textbf{2.11}$$

The term *f* corresponds to the fraction of deprotonated groups. In the case of a simple group in solution this p$K_a$ is independent on pH and its value can also be given by the pH value at the midpoint titration, where $[A^-]=[HA]$ and f=1/2.

For determining the p$K_a$ in titrable aminoacids in proteins, this procedure is not reliable due to two main reasons:

1) The charged groups in the protein interior have usually a poor stabilization due to the limited reorganization of their environment, compared to the aminoacids in solution where the free rotation of water molecules allows for a stabilization of the aminoacid´s charged groups. This makes the p$K_a$ in the protein different from the one in solution.

2) Neighbour titrating groups can affect the charge of each other by direct (electrostatic interactions between them) or indirect (electrostatic interactions with other groups) interactions. This makes the p$K_a$ in the protein pH dependent. Thus, given the importance of these pH effects at a molecular basis, a proper treatment of electrostatic interactions is essential [47]. For that, simple models were developed [50][51][52][53].

### 2.6.1 Continuum electrostatics and Poisson Boltzmann Equation

Using reasonable simplifications of these models, an “average” description of the system is obtained with the proper electrostatic aspects. One of these types of models is the continuum electrostatics (CE) model- **Figure 2.5** and it is based on the following assumptions:

- The system is described at the macroscopic level while reflecting the microscopic properties. That is, each time a group in the protein moves or receives/loses a ligand it will change the reorientation of its dipoles, and the solvent and the protein will try to reorganize to compensate any unfavorable new interactions. This reorganization of dipoles is captured by a dielectric constant ε. The higher the rearrangements of polar and charged groups, i.e. reorientation of the molecule, the higher is its dielectric constant, ε. The solute atoms are usually between 2-4 and the solvent is typically ~80, due to a larger reorientation. [48][47].

- Ions in solution (Na+, Cl-, etc) are attracted by opposite charges at the solute surface and they are taken into account in the model by the ionic strength (*I*) which is a measure of the counterions concentrations [54].

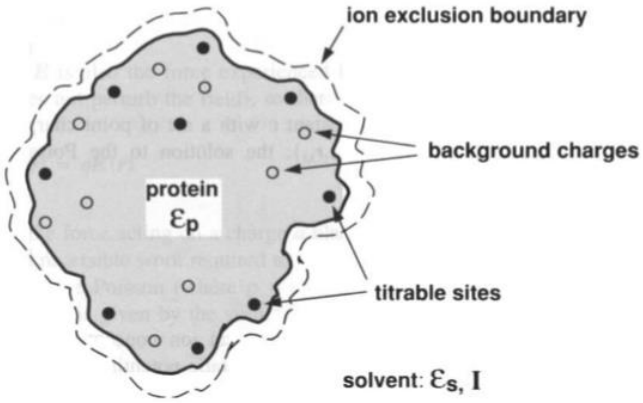


**Figure 2.5** Continuum electrostatics model of a protein in solution. $\varepsilon_p$ and $\varepsilon_s$ represent the dielectric constant of the protein and the solvent, respectively and *I* is the ionic strength [47].

There are several CE methods but the classical one to treat electrostatic interactions in solution is the Poisson equation:

$$-\nabla \cdot [\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})] = \rho_{tot}(\mathbf{r}) \qquad \textbf{2.12}$$

where $\varepsilon(\mathbf{r})$ is the system dielectric constant at point r, $\varphi(\mathbf{r})$ is the electrostatic potential at point **r** and $\rho_{tot}(\mathbf{r})$ (= $\rho_{prot}(\mathbf{r})$ + $\rho_{ions}(\mathbf{r})$) is the total charge (density) at point **r**.
The effect of counterions can be described by the Debye-Hückel approximation [55]. Including this in the Poisson equation, we get the Poisson-Boltzmann (PB) equation:

$$-\nabla \cdot [\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r})] + \bar{k}^2(\mathbf{r})\varphi(\mathbf{r}) = \rho(\mathbf{r}) \qquad \textbf{2.13}$$

where $\bar{k}^2(\mathbf{r}) = 8\pi e^2 I(\mathbf{r})/k_B T$, $I(\mathbf{r})$ is the ionic strength at **r**, $k_B$ is the Boltzmann constant and $\rho(\mathbf{r})$ is the protein charge at **r**.

In the PB equation both shape and charge distribution effects of the molecules are taken into consideration, where solute atoms are considered explicitly with point partial charges at atomic positions with a low dielectric constant (~2), unlike the solvent (~80). Solving this equation, the potential $\varphi$ is given through the space, as a function of $\varepsilon$, $\rho$, $I$ and $T$[48].

In practice, the PB equation is solved by mapping the system ($\rho + \varepsilon + I$) into a 3D grid, using a finite difference method (a more detailed description in subsection **3.5.3**) This will allow to solve the PB equation and to estimate the electrostatic potential at all points, obtaining the electrostatic energy:

$$U = \sum_i q_i \, \varphi(\mathrm{r}_i) \qquad \textbf{2.14}$$

By obtaining the electrostatic energy $U$, we obtain the electrostatic part of the free energy of a certain change that occurs in the system.

### 2.6.2 Free energies of protonation states- Thermodynamic cycle

Protonation free energies can be obtained by considering a thermodynamic cycle- **Figure 2.6** involving the titrable sites in the protein and in solution.

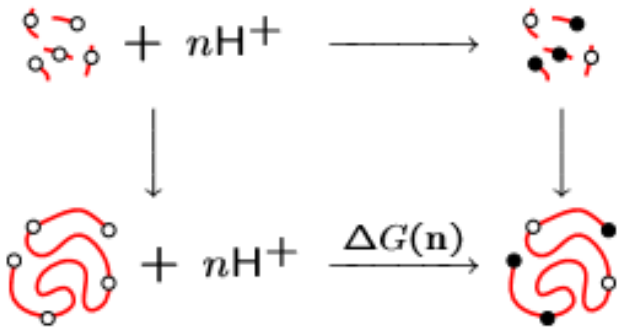


**Figure 2.6** Thermodynamic cycle of the protonation free energies taken from a CE model. Image from Dr. António Baptista.

This thermodynamic cycle describes the calculation of the protonation free energy of each titrable site in the protein, ΔG(**n**), by using the protonation free energy of the same titrable site in the solvent and taking into account the free energies associated with the change of the titrable aminoacid from the solvent to the protein in the protonated and deprotonated forms.

The free energy of ionization, ΔG, can be written as a sum of an internal part (bond breaking and other electronic structure changes) and an external part (steric changes in protonation/deprotonation). As the internal part is the same for both the protein and model compound, its contribution can be ignored, only the external part is considered. The steric changes are similar in both the protein and model compound, so they are neglected, thus, the remaining difference is the electrostatic part [56], i.e. a shift from the model compound and the titrable site in the protein is calculated.

By decomposing each term in the thermodynamic cycle, the sum of individual and pairwise terms to compute the free energy of each protonation state **n** is given by **equation 2.15**:

$$\Delta G(\mathbf{n}) = -2.3RT\sum_i \mathrm{n}_i pK_i^{int} + \sum_i \sum_{j<i} \left(\mathrm{n}_i \mathrm{n}_j + \mathrm{n}_i z_j^0 + \mathrm{n}_j z_i^0\right) W_{ij} \qquad \textbf{2.15}$$

where **n**= ($n_1$,$n_2$,..) represents the global protonation state, $n_i$ is the occupancy of site *i* in the protonation state n (n=0 deprotonated, n=1 protonated), $pK_i^{int}$ is the intrinsic $pK_a$ of site *i* (its $pK_a$ when all the others are kept neutral), W*ij* is the interaction between sites *i* and *j* and $z_i^0$ is the charge of site *i* when deprotonated [56].

### 2.6.3 Sampling of protonation states- Monte Carlo (MC)

By using the PB approach we are able to obtain the free energy contributions of a particular protonation state, but these states have to be properly sampled. From statistical mechanics, we know that the probability of the occurrence of each state **n** is proportional to its Boltzmann factor (**equation 2.16**):

$$\exp\left[-\frac{E(\mathbf{n})}{RT}\right] = \exp\left[-2.3pHN(\mathbf{n}) - \frac{\Delta G(\mathbf{n})}{RT}\right] \qquad \textbf{2.16}$$

where E is the "energy" of the system in the state **n**, R is the gas constant, ΔG is given by **equation 2.15** and N(**n)** is the total number of protons.

There is the need for some additional method that produces this correct distribution of states, the most general being the Monte Carlo method. This method can properly sample those states in a stochastic manner, generating a sequence of random states during the simulation in order to model the protonation equilibrium [57].

Given an initial protonation state, the binding sites are randomly flipped between protonated and deprotonated states in such a way that the protonation space of all sites is sampled [57].

A modification from a protonation state **n** to a new state **n**′ is either accepted or rejected based on the Metropolis criterion [43], which involves the change of the energy E:

$$\Delta E = E(\mathbf{n}') - E(\mathbf{n}) = 2.3RTpH[N(\mathbf{n}') - N(\mathbf{n})] + [\Delta G(\mathbf{n}') - \Delta G(\mathbf{n})] \qquad \textbf{2.17}$$

The new protonation state is accepted when $\Delta E \leq 0$, or if $\Delta E > 0$ when $e^{-\Delta E/RT} > A$, where A is a random number over the interval [0,1]. Upon acceptance, the new state is counted and used as a starting point for the next random flip. Otherwise, the previous state is used again for another random flip, the new state being rejected [41].
This method, in comparison to other methods available, has the advantage that most of the new system´s states are of low energy, ensuring that they are populated at equilibrium [41].

## 2.7 Constant-pH molecular dynamics

The effect of changes in pH is of great importance to understand, since it leads to protonation/deprotonation of titrable sites of solute species affecting their structure, stability, biomolecular recognition and other biomolecular properties [48]. However, addressing these changes has always been particularly challenging. Understanding pH effects means to understand the electrostatic changes associated with them, but treating pH by computational studies has been done, until recently, without coupling conformational and protonation aspects.
By performing MM/MD simulations, the molecule undergoes conformational changes with the assignment of fixed protonation states to the titrable sites. This can become problematic due to the choice of a typical state for the sites that is actually not present at a given pH: the protonable groups can exhibit atypical $pK_a$ values e.g. when in the protein they can vary their $pK_a$ by several pK units in relation to the $pK_a$ in solution. There are also the effects caused by the protein environment, site-site interactions and, more importantly, the impossibility of changing the protonation states during the MD simulation [58].
Another approach was the development of electrostatic theoretical models, like the PB model previously discussed in subsection **2.5.1**. This model treats the solute as rigid and, by MC methods, samples the protonation states, whereby it cannot capture the conformational reorganization associated to changes in charges.
Numerous procedures have been done to overcome the problems in these two approaches and, over the last decades, an alternative methodology was developed that combines the MM/MD and the PB/MC methods [58].

### 2.7.1 The stochastic titration method

The combination of these two approaches is the base of the stochastic titration method for constant-pH MD (CpHMD) simulations, allowing each approach to do what they do best: MM/MD does the conformation sampling at fixed protonation state, and PB/MC does the protonation sampling at fixed conformation. The pH is used as an external thermodynamic parameter, like the temperature or pressure. Several other CpHMD methods have meanwhile been developed [59][60][61][62][63][64][65], but the stochastic titration method, developed at the Molecular Simulation group at ITQB NOVA [58][66] is the only that is able to include both the use of explicit solvation models and ionic strength [66].

This CpHMD method uses discrete protonation states periodically updated from PB/MC calculations performed along the MD simulation. Its general algorithm is presented in the next figure [58].

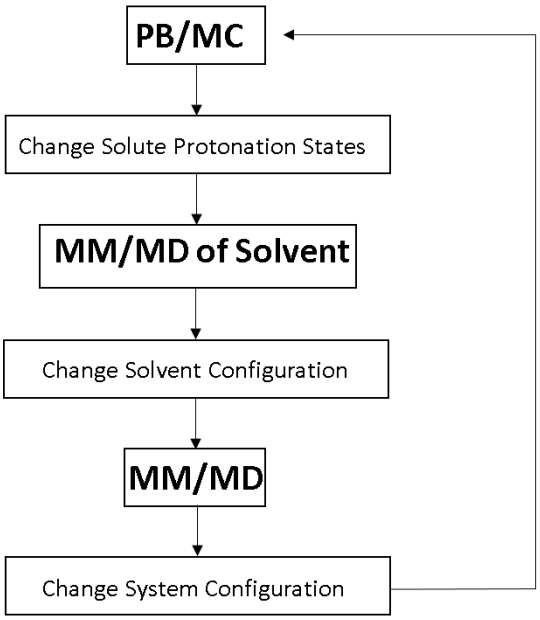


**Figure 2.7** Algorithm used for the stochastic titration method. Adapted from [58].

This method runs in different sequential blocks. The first block is a PB/MC step for the computation of the protonation states at a given pH and sampling of these states according to the MC calculation. The second block is a short MM/MD run of the solvent while the protein is kept rigid, allowing the relaxation of the solvent to adapt to the new protonation states [58][66]. The time allowed for this relaxation is one of the most critical parameters for this method: it must be enough to adapt to the electrostatic changes but not too high which may result in overadapted

solvent configurations (around 0.2 ps seems to be appropriate). The last block is a MM/MD of the full system [58][66].

A variation of this method is the addition of reduced titration, that is, an exclusion list is created with sites that are titrating too far away from the pH we are interested in. A threshold is defined and the sites for which either protonated or deprotonated frequencies are less than this threshold are put in the exclusion list. This will not only reduce the overall computing cost but it will also automatically select the sites that are titrating at the pH chosen.

This method has the additional step of a first full PB/MC on the system, in every $m^{th}$ step, to update the exclusion list of sites whose mean occupancies fall outside a defined threshold (see **Figure 2.8**) [66].

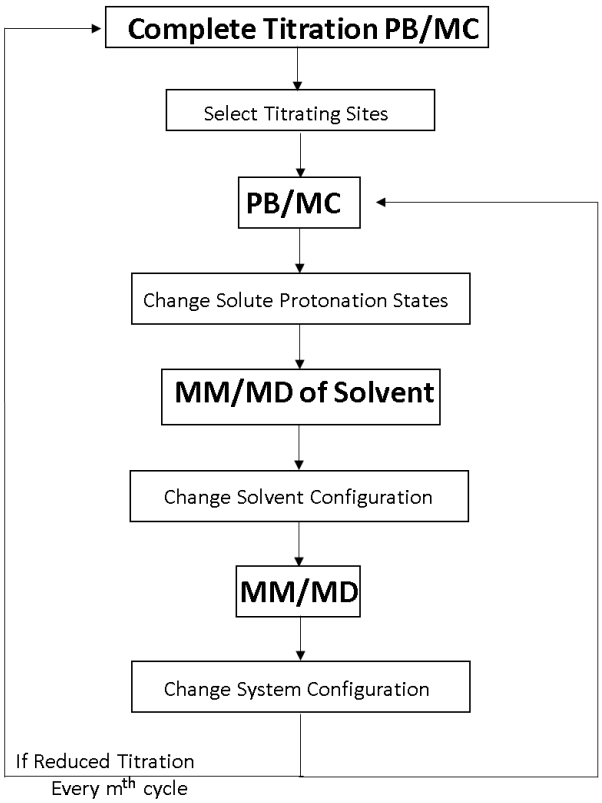


**Figure 2.8** Algorithm used for the stochastic titration method including the reduced titration step. Adapted from [66].

The stochastic titration method properly couples the protonation-conformation changes that occur in a biological system, accounting for the ionic strength and the explicit inclusion of solvent, which makes it suited to study the pH effects on β-lactoglobulin.

## 2.8 Umbrella Sampling

Sometimes just knowing that "two molecules spend a lot of time together" is not enough, and quantifying the driving force of the interaction between these two molecules becomes relevant, i.e. calculating a free energy for the interaction. For a given reaction coordinate (e.g. a distance between two monomers) the free energy measures how likely is a value of that coordinate, so that the lower the free energy of a state (e.g. two monomers at a specific distance) the more favoured it is. However, its calculation is a challenge in computational biology. Umbrella sampling is one of the methods able to calculate this free energy along a reaction coordinate [42].

The idea is to do multiple umbrella sampling calculations by splitting the reaction coordinate into several windows, sampling all regions. In each window, an MD simulation is performed and an additional energy term, a bias potential, is applied, keeping the system in the region of interest:

$$V'(\mathbf{R}) = V(\mathbf{R}) + U(\mathbf{R}) \qquad \textbf{2.18}$$

where $V(\mathbf{R})$ is the potential energy of coordinate state **R** (coordinates of all atoms) and $U(\mathbf{R})$ is the new added biasing term. This term is generally a harmonic force that tries to keep the system close to the reference point, $\xi_i^{ref}$, of the window *i*:

$$U(\mathbf{R}) = 1/2K\left[\xi(\mathrm{R}) - \xi_i^{ref}\right]^2 \qquad \textbf{2.19}$$

The multiple bias simulations thus obtained can then be combined using the Weighted Histogram Analysis Method (WHAM) [67], and an estimate of the free energy can be calculated. The sampling of the configurations has to be the best possible, so overlap between windows is required. Each of these umbrella sampling simulations generates a biased histogram (i.e. the probability distribution of finding a particular ξ value) and then WHAM works by transforming all these biased histograms into a single unbiased one. This unbiased histogram, P(**ξ**), is then used to compute a Potential of Mean Force (PMF), i.e. the free energy:

$$W(\xi) = -RT\ln\frac{P(\xi)}{P^*} \qquad \textbf{2.20}$$

Where $P^*$ is a reference value for which W=0. This PMF plays the role of an "effective energy" as a function of ξ (e.g. of the distance between two monomers).

# 3 METHODS

## 3.1 Box shape

The shape of a simulation box should be chosen according to the minimum number of solvent molecules needed to fill the box.

Among the several possible shapes for space-filling unit cells (see **Figure 3.1**), in this present work, for CpHMD simulations, a rhombic dodecahedron was chosen, due to its shape resembling more of a sphere than a cube and therefore saving computational time [68].

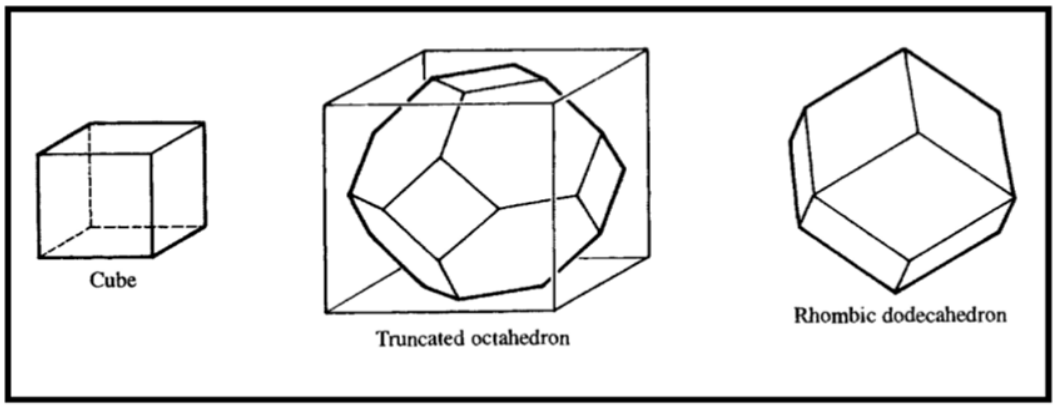


**Figure 3.1** Different periodic cells used in computer simulations. Adapted from [46].

## 3.2 Constraints and Restraints

The use of constraints was applied on all bonds using the LINCS (Linear Constraint Solver) [69] algorithm for the biological system and the SETTLE [70] algorithm for the water. When applied, positions restraints used harmonic forces.

## 3.3 Treatment of long range non-bonded interactions

For treating nonbonded interactions during the MM/MD steps a neighbour list was created and they were treated using a twin-range method (as described in subsection **2.4.1**) with two cutoffs: 0.8 and 1.4 nm.

The neighbour list, being updated every 5 simulation steps, uses a grid in the box to calculate the neighbour atoms. For the van der Waals and electrostatic interactions, the cutoff used was in a 1.4 nm radius.

The treatment of electrostatic interactions was made using the Reaction Field method [71] during the minimization energy steps and using the Generalized Reaction Field method [72] when performing an MD simulation of the system before and during CpHMD runs. The reaction field method uses a dielectric continuum representation with a defined dielectric constant beyond a cutoff sphere, wherein electrostatic interactions are treated explicitly. The generalized reaction field is an extension of the reaction field where ionic strength is considered beyond the cutoff radius.

The dielectric constant used was 54 [73] and the ionic strength was set to 0.1M.

## 3.4 Temperature and Pressure Control

When simulating biological systems, it is necessary to control the temperature and the pressure of the system, keeping them constant during all the simulation. Before performing the CpHMD runs, an initiation phase is performed where a MD simulation is run, allowing the temperature and pressure to stabilize. For the temperature, the system was put in an external bath at a fixed temperature using the velocity rescaling algorithm [74] during all MM/MD runs.

This algorithm is essentially a Berendsen thermostat [75] with an additional stochastic term which ensures a correct kinetic energy distribution. In the Berendsen thermostat, the bath acts as a source of thermal energy, supplying or removing heat (i.e. kinetic energy) from the system, that is, when a deviation of the system temperature from $T_0$ occurs, it is slowly corrected according to **equation 3.1** [46][68]:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \qquad \textbf{3.1}$$

where $T_0$ is the temperature of the bath, $T$ the instantaneous temperature and $\tau_T$ is a coupling parameter that defines the strength of the coupling between the temperature and the kinetic energy at a time constant, i.e. when $\tau_T$ is small, the coupling is strong.

The heat flow of the system is given by scaling the velocities every simulation step by a time-dependent factor λ:

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\left\{\frac{T_0}{T\left(t - \frac{\Delta t}{2}\right)} - 1\right\}\right]^{\frac{1}{2}} \qquad \textbf{3.2}$$

However, this method does not generate a proper canonical ensemble, which is solved with the velocity rescaling method, that has an additional parameter, a random seed, providing a correct ensemble of the kinetic energy [68].

The methods used for the pressure are the Parrinello-Rahman approach [76] in the initiation phase and the Berendsen barostat [75] in the CpHMD runs. In the Berendsen barostat, the pressure is kept fixed in similar way to the Berendsen thermostat, where an external pressure bath corrects the system´s pressure deviations in each time step, by applying the **equation 3.3** [68]:

$$\frac{dp}{dt} = \frac{p_0 - p}{\tau_p} \qquad \textbf{3.3}$$

where $p_0$ is the pressure of the bath, $p$ the instantaneous pressure and $\tau_p$ is the coupling constant between the bath and the pressure.

In this case, the coordinates and box vectors are rescaled every step with a scaling parameter μ given by **equation 3.4**:

$$\mu = 1 - \frac{\beta \Delta t}{3\tau_p}(p_0 - p) \qquad \textbf{3.4}$$

where $\beta$ is the isothermal compressibility of the system.

The Parrinello-Rahman method takes into account the fluctuations in pressure or volume, changing the shape and volume of the cell [76].

In our simulation, both the solvent and the protein were put in separate temperature baths of 300 K, due to this being the temperature used in experimental studies, and to a pressure bath of 1 atm during the MM/MD runs.

## 3.5 Preparing the simulation

Every molecular dynamics simulation follows a protocol involving three phases: the preparation of the system, the minimization/relaxation, where initial velocities are assigned, and the production phase, where a trajectory is generated and stored for further analysis [46][39].

### 3.5.1 Preparation phase

In a MM/MD simulation, the choice of a molecular model with the initial coordinates of the system and an appropriate force field is essential for a proper prediction of the properties of a system.

Structural system

The structural models of choice were the monomer and dimer of β-lactoglobulin variant A.

For the monomer, two different conformations were chosen from the protein data bank (PDB): one where the conformation of the EF loop was in the "open position" (1BSY) and the other where it is in the "closed position" (3BLG). Both conformations were obtained by X-ray crystallography at pH 7.1 and pH 6.2 for 1BSY and 3BLG, respectively [13]. For choosing the dimer, there were some available structures in the PDB [13][16] however their sequence was incomplete and either the two variants were mixed or the assembling of the two monomers from the crystal was different from the one believed to occur biologically. Thus, the dimer for both the open and closed conformation were constructed manually using the molecular visualization software PyMOL [77]. By making use of the structure for the monomer 1BSY, symmetry mates within 20Å were generated and aligned with the dimer 1BEB [16]. While this dimer was thought to have the correct association it was not used due to the factors described above. The overlapped symmetry mate was the chosen one, being then concatenated with the other monomer. For the conformation 3BLG the procedure was the same. The water at the interface of the two monomers was removed by a cutoff of 0.5 nm.

For better conformational sampling, one replicate of each conformation was generated by changing the initial velocities of the original ones.

Aqueous system

For 1BSY, the box was filled with 9760 solvent molecules for the monomer and with 37730 solvent molecules for the dimer.

For 3BLG, the box was filled with 9445 solvent molecules for the monomer and with 36306 solvent molecules for the dimer.

Both simulations used the simple point charge (SPC) model [78].

The force field model

All simulations were performed using GROMOS 54A7 force field [79] with the GROMACS simulation package [80], version 4.0.7, modified in house to include ionic strength (version 4.0.7_pH_I) [66].

### 3.5.2 Minimization/Relaxation phase

It is good practice to submit the initial structure to an energy minimization before performing the simulation. This will refine the initial coordinates obtained by experimental studies (mainly X-ray and NMR) avoiding local stresses due to non-bonded overlaps and relaxing bond lengths and angles [39].

The system was first minimized with ~$10^3$ steps of steepest descent with position restraints on all atoms, with a force constant of 1000 kJ $mol^{-1}$ $nm^{-2}$, followed by another ~$10^3$ steps of steepest descent with no restraints, or until it reaches an energy minimum.

The initial velocities ($v_x$, $v_y$, $v_z$) where then randomly assigned by a Maxwellian distribution, at a temperature *T*, as shown in **equation 3.5** [39]:

$$P(v) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{1}{2}} \exp\left[\frac{-mv^2}{2k_B T}\right] dv \qquad \textbf{3.5}$$

where *m* is the atomic mass of an atom and $k_B$ is the Boltzmann´s constant.

The relaxation phase consisted of 50 ps of MD in NVT ensemble with a $\tau_T$ of 0.01 ps, restraining harmonically all the atoms (1000 kJ $mol^{-1}$ $nm^{-2}$), followed by another 50 ps of MD in NVT ensemble with a $\tau_T$ of 0.1 ps, where only the Cα atoms were restrained and ending with another 100 ps of MD in NPT ensemble with a $\tau_T$ of 0.1 ps and a $\tau_P$ of 0.5 ps keeping the restriction on the Cα atoms. The temperature was performed at 300 K and the pressure at 1 atm.

### 3.5.3 Production phase

After the system being well equilibrated the production phase can begin. The simulations of 100 ns were made using the CpHMD method at pHs: 3, 4, 5, 6, 7 and 8. The choice of these pHs intended to include the pH of the stomach [32], pH close to the pI [5] and neutral pH, where the EF loop shows variations [18].

The conformations from 1BSY and 3BLG structures were used in the monomeric and dimeric form producing two simulations at each pH. This is represented in the following scheme, where the abbreviations used hereafter for each simulation are shown.


**Figure 3.2** Representation of the used abreviations for the two conformations.

The reduced titration approach [66] was used considering a threshold of 0.001, where an exclusion list was generated and updated every 50 cycles of CpHMD.

The step of the solvent relaxation was 0.2 ps of MM/MD, followed by an MM/MD run for all the system of 10 ps for both the monomer and dimer. For a better understanding of these segments see **Figure 2.8** in section **2.6**.

Protonation states: PB/MC calculations

All the electrostatic interactions for each protonated state were calculated using the version 2.2.9 of the MEAD (Macroscopic Electrostatics with Atomic Detail) program that solves the Poisson-Boltzmann equation [81]. This equation gives as solution the electrostatic potential and, in practice, it can be obtained using the finite difference method. The idea is to approximate derivatives as differences between function values sampled at points a finite distance apart [56]. The protein is mapped into a three-dimensional grid distributing the atomic charges, the dielectric constant and the ionic strength over the neighbouring grid points- **Figure 3.3**. A Stern layer (see **Figure 2.5**) is assigned between the protein and the solvent, indicating a narrow region where the ionic strength is zero and only the dielectric constant of the water is considered [39]

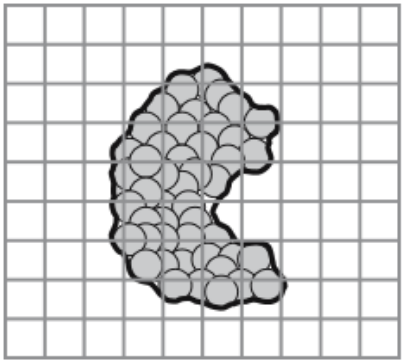

**Figure 3.3** Finite difference method. Adapted from [82].

First, the electrostatic potential at the points on the grid wall is computed using some approximations: the Coulombic equation is used assuming a single dielectric constant, corresponding to the solvent. Then, an approximated discretized version of the PB equation is solved at each grid point to update its potential, and this is repeated until the potential converges for all grid points.

A focusing approach was used. First a large coarse grid is used for the whole system, followed by a second finer grid, applied at a region of interest (e.g. titrable site), and using the potential from the previous step on its walls. A third grid can be used, and so on, until the potential reaches the required accuracy [39]. These successive levels of focusing allows to obtain highly accurate local solutions to the PB equation with reduced levels of computational effort [82].

A two-step focusing procedure was used in the protein with an initial cubic grid of 121x121x121 with a mesh size of 1 Å and a second cubic grid of 81x81x81 with a mesh size of 0.25 Å focused on the titrable residues.

Solving the thermodynamic cycle in **Figure 2.6** of the subsection **2.5.2** gives us the protonation energies. This can be done by MEAD, where a model compound of each titrable site with known $pK_a$ value is used.

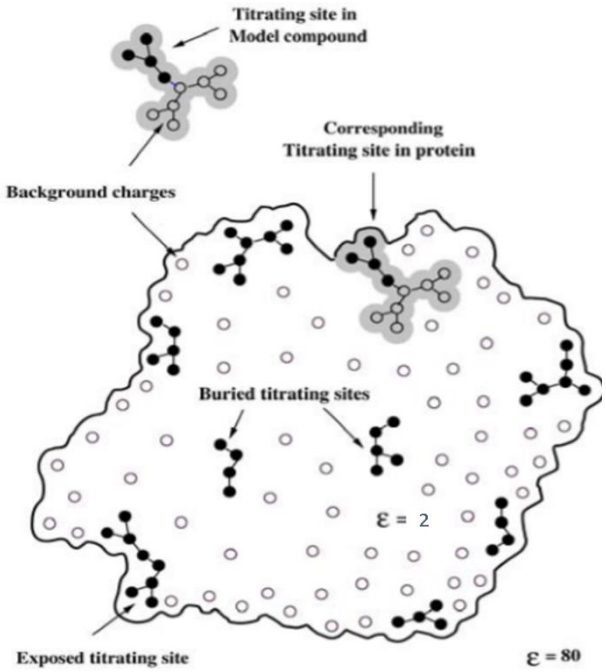


**Figure 3.4** MEAD model for $pK_a$ calculations of proteins. Adapted from [56].

For the model compound, it was applied a first focus with a 61x61x61 grid spaced 1 Å and then a second one with an 61x61x61grid spaced 0.25 Å. The structure of the model compound is

obtained by taking the corresponded fragment in the protein- **Figure 3.4**. The parametrization of the model compounds was made by Carvalheda et al [83].

For all the PB calculations, the atomic charges and radii were obtained from the GROMOS 54A7 force field [79] as described elsewhere [84], the molecular surface was defined with a solvent probe of radius 1.4 Å and a Stern layer of 2.0 Å, the dielectric constant set to 2 for the molecular interior and 80 for the solvent. The temperature was 300 K and the ionic strength 0.1 M.

The sampling of protonation states (see subsection **2.5.3**) was done using a MC method as implemented in the version 1.6.0 of the in house PETIT (Proton and Electron TITration) program [57][85].$10^5$ MC cycles were performed, with each cycle consisting of random choices of state according to the Metropolis rule [43] for all individual sites and pairs of sites with a coupling above 2.0 $pK_a$ units [57].

PB/MC had a proper consideration of proton isomers for each titrable site [85].

MM/MD segments

All the MM/MD segments were performed using the in-house version 4.0.7_pH_I of the GROMACS software. A NPT ensemble was used with constraints in all atoms. The temperature was set to 300 K and the pressure to 1 atm coupled to a temperature and pressure bath with $\tau_T$ of 0.1 ps and a $\tau_P$ of 2 ps, respectively.

Newton´s equations of motion were integrated using the Leap-Frog algorithm with a time step of 2 ps.

## 3.6 Umbrella Sampling

Before performing umbrella sampling, a series of configurations must be generated along a reaction coordinate ξ (center of mass distance between monomers), by pulling one monomer away from the other monomer.

Only the 3BLG conformation of the dimer was used at pHs 3, 4 and 5, with two simulations, due to its closed conformation which is typical of these pHs. The choice of these pHs was due to interesting properties observed in the CpHMD results (chapter 4). The replicates were generated by performing different pullings with different initial velocities.

The parameters from the GROMOS 54A7 force field were applied, the short and long-range nonbonded interactions were cut-off at 1.4 nm using the Generalized Reaction Field. In order to generate equilibrated starting structures for the pulling simulations, the system was placed in a triclinic box of simple point charge (SPC) water long enough for the pulling simulations to take

place and to satisfy the minimum image convention, being filled with 106648 solvent molecules. The minimization and equilibration phase followed the same procedure as described above.

### 3.6.1 Pulling

All pulling simulations were conducted with the GROMACS package, in-house version 4.0.7_pH_I, using fixed protonation states typical of pH 4.

The monomer B was pulled away from the monomer A, from an initial distance of 3 nm between the centers of mass, generating a series of configurations where some were used as starting configurations for the umbrella sampling windows, which are run independently from each other.

The monomer A was restrained along its x, y, z axis, being used as an immobile reference for the pulling simulations and the monomer B was pulled along the z-axis over the course of 400 ps of MD, using a spring constant of 1000 kJ $mol^{-1}$ $nm^{-2}$ and a pull rate of 0.01 nm $ps^{-1}$. A final center-of-mass (COM) distance between monomers A and B of approximately 6.8 nm was achieved. From these trajectories, snapshots were taken to generate the starting configurations for the umbrella sampling windows.

### 3.6.2 Production phase

The distribution of sampling windows used was of 0.1nm between the COM distance 3.0-4.0 nm resulting in 11 windows for pHs 3, 4 and 5. This spacing allowed for overlap of the windows covering a reasonable distribution of configurations. The chosen initial configurations of the sampling windows were submitted to a brief NPT equilibration inside a rectangular box built with length long enough to avoid PBC artifacts. This length corresponds to twice the center of mass distance (d) plus twice the radius of each monomer (r) plus a safeguard distance (s): (2d + 2r + s), where r = 2 nm and s= 2 nm.

The umbrella sampling simulations were run without position restraints during 50 ns using the CpHMD method, with a spring constant of 1000 kJ $mol^{-1}$ $nm^{-2}$ for the COM distance restraint.

Analysis of the results was performed with the weighted histogram analysis method (WHAM) [67].

## 3.7 Rigid PB/MC calculations

Rigid PB/MC calculations were also done for single structures, namely 1BSY and 3BLG, using crystallographic waters with relative solvent accessibility lower than 0.5. For 1BSY, this corresponds to 59 solvent molecules for the monomer and 118 for the dimer. For 3BLG, this corresponds to 115 solvent molecules for the monomer and 230 for the dimer. The parameters for water were taken from the simple point charge (SPC) model [78].

These calculations were done using the same parameters as in the CpHMD simulations described above, but with a dielectric constant set to 10 for the molecular interior. The range of pH used was from -5 to 20, with a 0.2 increments.

Arg residues were considered to be always charged. Proton tautomerism was used for all other titrable sites (Asp, Glu, free Cys, Tyr, Lys, Nter and Cter) and non-titrable sites with orientable protons (Ser, Thr and waters).

## 3.8 Analyses

All the analyses were performed using the GROMACS software or in-house tools. When computing averaged protein properties only the contribution of the equilibrated system was accounted, being considered equilibrated after 30 ns of simulation time, for the CpHMD simulations, and after 10 ns for the umbrella sampling.
For the dimer, some of the analyses were performed for each individual chain.

### 3.8.1 Protein analysis

The root mean square deviation (RMSD) was determined for the protein backbone. For each case, the backbone atoms were fitted to the respective atoms in the initial BLG conformation of each replicate ($t_2$=0). The distances between the initial atomic positions of an atom i, $\mathbf{r}_i(t_2)$, and its atomic positions at a given time $t_1$,$\mathbf{r}_i(t_1)$, were used to calculate the RMSD with the following equation:

$$RMSD(t_1, t_2) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i \parallel \mathbf{r}_i(t_1) - \mathbf{r}_i(t_2) \parallel^2\right]^{\frac{1}{2}} \quad \textbf{3.6}$$

where $M = \sum_{i=1}^{N} m_i$ and $m_i$ is the atomic mass of atom $i$ [68].

The root mean square fluctuation (RMSF, i.e. standard deviation) for each aminoacid residue was calculated by doing a least-square fitting of all structures to the reference frame BLG

structure at 30 ns, followed by a calculation of an average structure (with coordinates $\overline{\mathbf{r}_i}$), relative to which the atomic fluctuations are computed as:

$$RMSF(t_1) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i \parallel \mathbf{r}_i(t_1) - \overline{\mathbf{r}_i} \parallel^2\right]^{\frac{1}{2}} \quad \textbf{3.7}$$

where the summation is over the atoms of the residue in question.

The protein radius of gyration, $R_g$, was calculated using the following equation:

$$Rg = \left(\frac{1}{M}\sum_{i}^{N} d_i^2 m_i\right)^{\frac{1}{2}} \quad \textbf{3.8}$$

where $d_i$ is the distance between the atom *i* and the center of mass of the protein [68].
The secondary structure of the protein was assigned using the DSSP criterion defined by Kabsch and Sander [86]. The fraction of each type of secondary structure was obtained by averaging at each pH by the total number of residues over time. In the results presented, the different types of helix are reported as a single helix content, and the β-bridges and β-sheets are reported as a single β-structure content.

The geometrical criteria used to define the hydrogen bonds (H-bonds) between all possible donors and acceptors were [68]:

$$r \leq r_{HB} = 0.35\ nm$$
$$\alpha \leq \alpha_{HB} = 30°$$

where r represents the donor-acceptor distance and α the hydrogen-donor-acceptor angle.

The protein solvent accessible surface area (SAS) was calculated using a rolling probe with a radius of 0.14 nm. The contact surface was calculated considering the following equation:

$$\frac{(SAS_A + SAS_B - SAS_{dimer})}{2} \quad \textbf{3.9}$$

where A and B, represents chain A and chain B of the dimer.

Titration curves

The titration curves were obtained by averaging at each pH value the occupancy states of each titrable site over the equilibrated system. A Hill equation [87] was fit to the data:

$$f(\mathrm{pH}) = \left[1 + 10^{n(\mathrm{pH}-\mathrm{p}K_a)}\right]^{-1} \quad \textbf{3.10}$$

to get the p$K_a$ and $n$ (Hill coefficient) values for each titrable site. These fits were done using a nonlinear least-squares algorithm [88].

PMF

The potential of mean force was calculated using the following equation:

$$W(\xi) = -RT\ ln\ \frac{P(\xi)}{P^*} \quad \textbf{3.11}$$

where P(ξ) is the probability of finding the center of mass (COM) of the two BLG monomers at a distance ξ, and $P^*$ is a reference value, for which $W = 0$. For each PMF calculation an average P($\xi$) histogram was obtained from the histograms of the two replicates, after which this equation was applied.

Calculation of averages and errors

The results obtained from each one of the properties measured in the protein analysis were averaged considering the contribution of the 2 replicates for each pH value. For a sample $x_1, x_2, \ldots x_N$ of independent values of the variable $x$, the average ($\bar{X}$) is:

$$\bar{X} = \frac{1}{N}\sum_i x_i \quad \textbf{3.12}$$

For most average properties, the global average is represented together with the partial averages obtained with each replicate. This provides a good measure of the uncertainty, thus replacing the use of error bars in plots.

For the computed p$K_a$ values, errors were calculated using a bootstrap method [89]:

1. At each pH value, we randomly sampled (with replacement) two new average occupancies from the original set of two values (one per simulation) and computed their pH-specific global average;

2. These global averages (one per pH) were used to perform the fit of the Hill equation;

3. Steps 1 and 2 were repeated 1000 times, producing 1000 values for each parameter ($pK_a$ and n);

4. The error of each parameter was computed as the standard deviation of its 1000 values.

For the PMFs, errors were obtained as follows:

1. The error δP at each bin of the average histogram P ($\xi$) was computed as the standard deviation over the two replicates;

2. The error δW of the PMF at each bin was then computed from δP using the differential formula for the propagation of errors [90]:

$$\delta W = \left|\frac{dW}{dP}\right| \delta P = \frac{RT}{P} \delta P \qquad \textbf{3.13}$$

### 3.8.2 Plots and representations

All the presented plots were made using the gnuplot software [88] and all the molecular representations were done using the PyMOL software [77].

# 4 RESULTS AND DISCUSSION

In this chapter we will discuss a wide range of properties of our simulated systems in an attempt to clarify the questions proposed as part of the goal of this work. The following results were obtained from the analysis of molecular trajectories from MD simulations, and will help us to identify the characteristic processes of the system.

Before performing these simulations, rigid PB/MC calculations were made, i.e., calculations in which the conformation of the protein is fixed, changing only the protonation states. These had the intention of elucidating us about the residues that will be considered titrable in the pH range used in CpHMD simulations. So, the results obtained from these calculations will be first presented and discussed.

## 4.1 Rigid PB/MC Calculations

The data obtained from these calculations include the mean protonation state of each titrable site in a pH range from -5 to 20, which was used to follow the global titration of the proteins and its individual titration of the residues. Only the global titration will be shown.

Global titration curve

The global titration curve for both the monomer and dimer, in the open and closed conformation, can be seen in **Figure 4.1.**

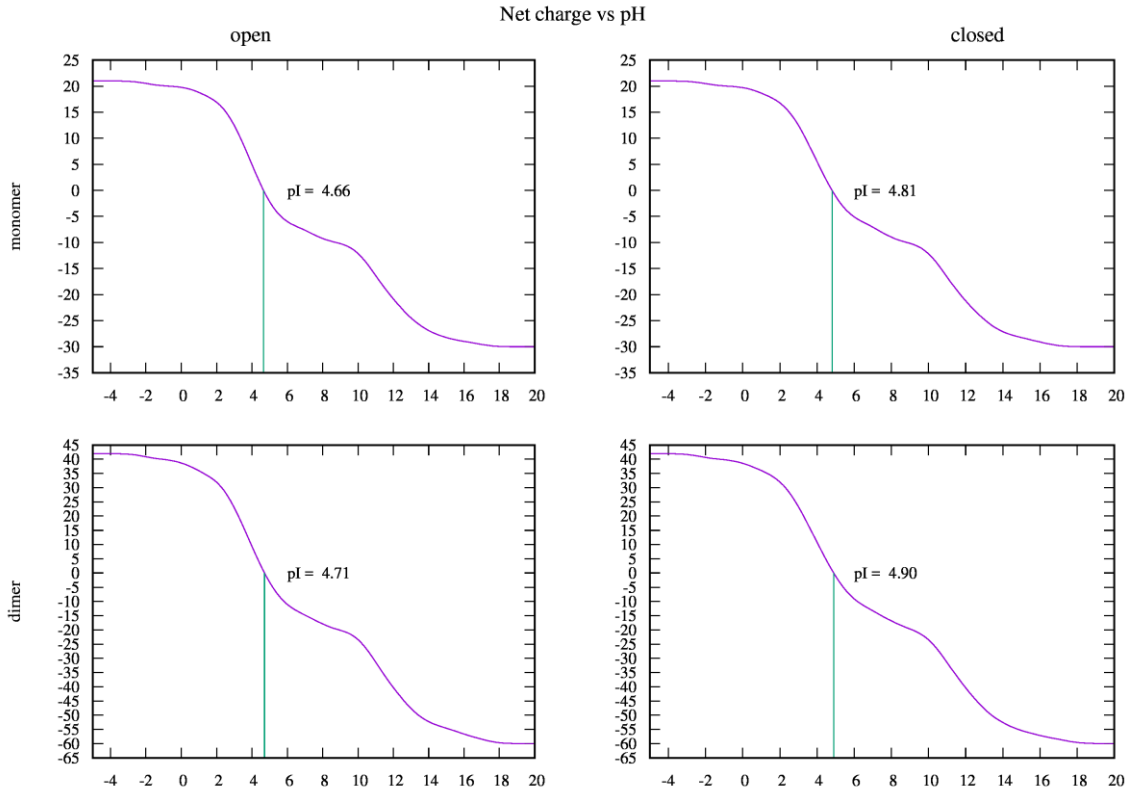


**Figure 4.1** Mean protein net charge as a function of the pH, considering both the monomer and dimer, with open and closed conformations. The pI is also shown.

The protein net charge as a function of the pH conditions shows a similar profile for the four conformations with slight differences that can be mostly observed in the isoelectric point, pI, computed neglecting any bound ions.

These differences in the pI observed in the four conformations can be due to two factors, one depending on the difference from the open to the closed conformation and the other arising from the difference between the monomer and the dimer.

This pI appears to be higher for the closed conformations when compared to the open ones, in both the monomer and dimer. The most likely reason seems to be that, in the closed conformation, the Glu89 associated with the Tanford transition can be protonated at higher pH values, as opposed to the open conformation, resulting in an extra charge. Remarkably, Tanford et al. [22] were able to infer the presence of these two curves by analyzing the experimental titration curve, concluding that the conformational transition was due to a (then unidentified) carboxyl group.

By comparing the pI of the monomer with the dimer, the pI of the dimer is slightly higher. An explanation for this could be the presence of titrable sites at the interface that are interacting with each other, changing the global charge and, consequently, the pI.

When comparing the obtained pIs with the experimental pI (~5.2), a small difference is encountered, but, when comparing it to other computed pIs, 4.76 [91], the value seems very similar. This difference could originate from the partial binding of a cation to BLG in the experimental calculation of the pI that goes by unnoticed, or, to the error associated from the computational calculations that is generally less than one unit.

Computed pKas

The $pK_a$ values for each titrable site were obtained and are shown in **Table 4.1**. These values will allow us to determine which sites titrate too far from the pH range used in CpHMD simulations and thus will not be considered, saving some computational time.

The typical $pK_a$ values for titrable sites in a solvated peptide environment ($pK_{typ}$) were determined experimentally by Pace and co-workers using eight blocked pentapeptides ($Ala_5$-$NH_2$, Ac-$Ala_5$ and Ac-$Ala_2$-X-$Ala_2$-$NH_2$, where Ac denotes acetyl and X=Asp, Glu, His, Cys, Tyr and Lys) [92][93].

**Table 4.1** Computed $pK_a$ values for the open and closed conformations and $pK_{typ}$ values.

| Residue | Computed $pK_a$ | | | | | | $pK_{typ}$ |
|---|---|---|---|---|---|---|---|
| | Open | | | Closed | | | |
| | Monomer | Dimer | | Monomer | Dimer | | |
| | | Chain A | Chain B | | Chain A | Chain B | |
| Nter | 7.93 | 8.04 | 8.07 | 7.29 | 7.36 | 7.39 | 8.00 |
| **Lys8** | 10.55 | 10.57 | 10.57 | 10.64 | 10.66 | 10.65 | 10.40 |
| Asp11 | 3.10 | 3.90 | 3.07 | 3.17 | 3.15 | 3.17 | 3.94 |
| **Lys14** | 11.13 | 11.15 | 11.14 | 11.14 | 11.17 | 11.16 | 10.40 |
| **Tyr20** | 12.88 | 13.95 | 13.95 | 13.86 | 13.92 | 14.02 | 9.84 |
| Asp28 | 3.93 | 3.59 | 3.56 | 3.89 | 3.80 | 4.26 | 3.94 |
| Asp33 | 3.98 | 4.72 | 4.74 | 3.91 | 4.32 | 4.15 | 3.94 |
| **Tyr42** | 13.81 | 15.50 | 15.61 | 12.64 | 14.68 | 14.40 | 9.84 |
| Glu44 | 6.86 | 6.81 | 6.78 | 6.65 | 6.57 | 6.53 | 4.25 |
| Glu45 | 3.86 | 3.85 | 3.82 | 3.85 | 3.82 | 3.82 | 4.25 |
| **Lys47** | 11.71 | 11.74 | 11.74 | 11.88 | 11.87 | 11.84 | 10.40 |
| Glu51 | 4.27 | 4.30 | 4.36 | 4.13 | 4.10 | 4.14 | 4.25 |
| Asp53 | 4.09 | 4.07 | 4.08 | 4.33 | 4.27 | 4.36 | 3.94 |
| Glu55 | 2.83 | 2.80 | 2.85 | 2.59 | 2.60 | 2.58 | 4.25 |
| **Lys60** | 11.86 | 12.61 | 12.46 | 11.11 | 11.73 | 11.79 | 10.40 |
| Glu62 | 3.48 | 3.52 | 3.49 | 2.85 | 2.75 | 2.82 | 4.25 |
| Asp64 | 3.88 | 3.90 | 3.90 | 3.87 | 3.90 | 3.89 | 3.94 |
| Glu65 | 4.00 | 4.01 | 4.01 | 4.16 | 4.18 | 4.18 | 4.25 |
| **Lys69** | 11.33 | 11.36 | 11.41 | 12.42 | 12.39 | 12.28 | 10.40 |
| **Lys70** | 12.09 | 12.10 | 12.09 | 11.67 | 11.68 | 11.74 | 10.40 |
| Glu74 | 5.05 | 5.06 | 5.14 | 4.51 | 4.70 | 4.52 | 4.25 |
| **Lys75** | 11.59 | 11.55 | 11.55 | 11.38 | 11.44 | 11.40 | 10.40 |
| **Lys77** | 10.47 | 10.48 | 10.49 | 10.26 | 10.24 | 10.27 | 10.40 |
| **Lys83** | 11.51 | 11.60 | 11.62 | 11.22 | 11.47 | 11.35 | 10.40 |
| Asp85 | 3.35 | 3.22 | 3.31 | 3.50 | 3.53 | 3.50 | 3.94 |
| Glu89 | 3.66 | 3.65 | 3.64 | 7.83 | 8.14 | 8.21 | 4.25 |
| **Lys91** | 11.02 | 11.10 | 11.23 | 11.44 | 11.41 | 11.57 | 10.40 |
| Asp96 | 3.50 | 3.48 | 3.38 | 3.14 | 3.07 | 3.23 | 3.94 |
| Asp98 | -1.93 | -2.11 | -2.06 | 2.02 | -2.42 | -2.23 | 3.94 |
| **Tyr99** | 12.83 | 12.93 | 12.94 | 12.49 | 12.49 | 12.58 | 9.84 |
| **Lys100** | 10.85 | 10.90 | 10.89 | 10.65 | 10.59 | 10.76 | 10.40 |
| **Lys101** | 12.80 | 12.97 | 11.96 | 13.36 | 13.50 | 13.08 | 10.40 |
| **Tyr102** | 16.82 | 17.05 | 16.97 | 15.78 | 15.88 | 15.95 | 9.84 |
| Glu108 | 0.72 | 0.54 | 0.52 | 0.59 | 0.33 | 0.42 | 4.25 |
| Glu112 | 4.61 | 4.63 | 4.64 | 4.96 | 5.08 | 4.89 | 4.25 |
| Glu114 | 4.23 | 4.06 | 4.00 | 4.85 | 4.94 | 4.82 | 4.25 |
| **Cys121** | 15.18 | 16.13 | 15.96 | 15.93 | 17.28 | 17.30 | 8.55 |
| Glu127 | 4.41 | 4.39 | 4.38 | 4.45 | 4.42 | 4.43 | 4.25 |
| Asp129 | 2.65 | 2.53 | 2.69 | 1.19 | 1.16 | 1.75 | 3.94 |
| Asp130 | 3.86 | 3.69 | 3.58 | 3.94 | 3.56 | 3.51 | 3.94 |
| Glu131 | 1.49 | 1.37 | 1.22 | 4.13 | 4.05 | 3.97 | 4.25 |
| Glu134 | 4.16 | 3.91 | 4.02 | 3.43 | 3.40 | 3.26 | 4.25 |
| **Lys135** | 12.23 | 12.43 | 12.30 | 11.25 | 11.56 | 11.62 | 10.40 |
| Asp137 | 2.34 | 0.59 | 0.59 | 2.22 | 0.41 | 0.67 | 3.94 |
| **Lys138** | 10.43 | 10.67 | 10.61 | 10.78 | 11.09 | 11.05 | 10.40 |
| **Lys141** | 10.28 | 10.93 | 11.01 | 10.87 | 11.79 | 11.77 | 10.40 |
| His146 | 7.41 | 7.38 | 7.56 | 7.68 | 8.17 | 8.24 | 6.54 |
| Glu157 | 5.37 | 5.36 | 5.35 | 5.06 | 5.05 | 5.06 | 4.25 |
| Glu158 | 4.83 | 4.92 | 4.91 | 5.30 | 5.36 | 5.35 | 4.25 |
| His161 | 5.60 | 5.93 | 6.16 | 5.05 | 5.18 | 5.46 | 6.54 |
| Cter | 3.66 | 3.57 | 3.63 | 3.79 | 3.62 | 3.71 | 3.94 |

Marked in bold, are the residues that were not included in CpHMD calculations due to a titration at pHs too far from the pH range performed in our simulations. The residues considered were: Asp, Glu and His. Tyr and free Cys were kept neutral, and Arg (not shown in the table) and Lys were kept charged.

## 4.2 CpHMD simulations

### 4.2.1 Monomer/dimer and open/closed analysis

The results obtained from the simulations performed for the monomer and dimer of the PDB structures 1BSY (open) and 3BLG (closed) at 6 different pH conditions (3, 4, 5, 6, 7, and 8) will be presented and discussed in this section. In these analyses we will evaluate how the pH has affected the conformational behaviour of BLG and what residues were involved in these changes. We will start by analysing several general properties to identify the main monomer/dimer and open/closed differences, followed by analyses focusing on some protein domains, the dimer interface and the protonation features.

For some cases only the average behaviour at each pH value will be shown and analyzed, taking into account the equilibration time of 30 ns, chosen by observation of the temporal analyses.

The simulations ran during 100 ns for the monomer, however, for the dimer, due to a separation distance between the two monomers larger than what was allowed by the box size, for some replicates, these simulations were stopped before reaching 100 ns. **Table 4.2** shows the simulation time for each replicate of the dimer for the open and closed conformation.

**Table 4.2** Simulation time for the open and closed conformations for the two replicates of the dimer.

| | **Simulation time (ns)** | | | |
|---|---|---|---|---|
| | Open 1 | Open 2 | Closed 1 | Closed 2 |
| **pH 3** | 100 | 100 | **73** | 100 |
| **pH 4** | 100 | 100 | 100 | 100 |
| **pH 5** | 100 | 100 | 100 | 100 |
| **pH 6** | 100 | 100 | 100 | 100 |
| **pH 7** | **19** | **68** | **57** | 100 |
| **pH 8** | **92** | 100 | 100 | 100 |

This table shows that at pH 7 the dimer tends to dissociate into monomers for three replicates, this effect also being present but less pronounced at pH 3 and 8, since only one replicate for each pH was affected.

After analyzing the results for the replicate2 at pH7 for the monomer in the closed conformation, an error was noticed and this simulation was discarded. It was not possible to repeat this simulation in time for this thesis.

Root Mean Square Deviation

The root mean square deviation (RMSD) analysis allows us to evaluate the deviation of the BLG structure from its initial conformation during the simulation time. The RMSD of each structure shows the deviations against time relative to the backbone from the initial BLG structure. The obtained results for each simulated pH value, considering both open and closed conformations for the monomer and dimer, each one with two replicates, are depicted in **Figure 4.2**.

The results are presented and discussed together for both the monomer and dimer at each pH value.

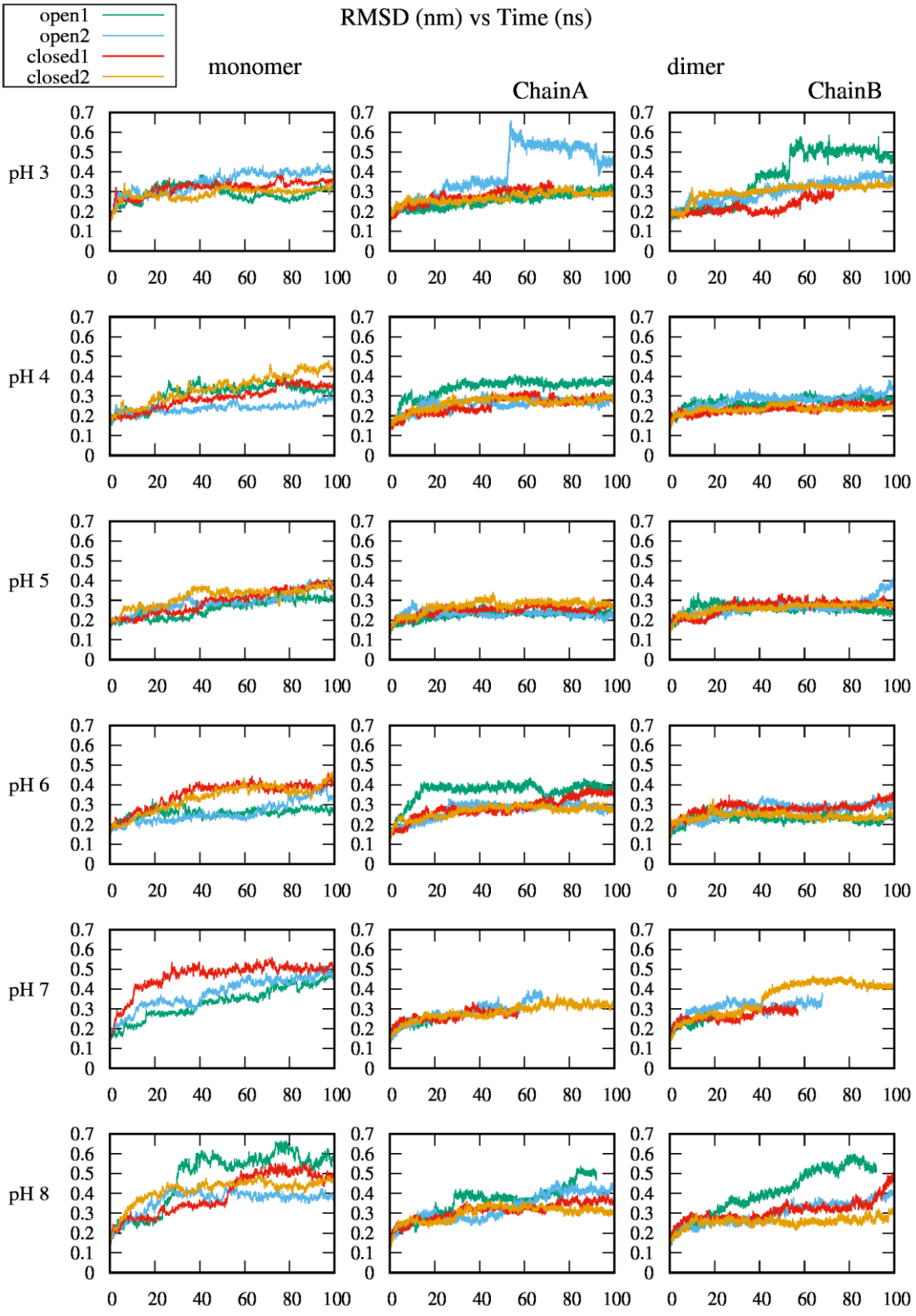


**Figure 4.2** RMSD vs time considering the monomeric and dimeric forms in the open and closed conformations, for each pH value. The protein backbone was fitted to its respective backbone in the initial BLG conformation.

The RMSD for both the monomer and dimer shows a similar profile, with higher deviations at higher pH values, especially at 7 and 8, indicating a deviation from its initial structure. These deviations are more pronounced in the monomeric form, which could correspond to a destabilization of the structure, while in the dimeric form the obtained results can point to a structure that is more stable in relation to the monomer.

For the dimer, an exception is seen at pH 3, for both chain A and chain B, corresponding in each chain to one replicate of the open conformation, where there is a high deviation occurring at ~55 ns until the end of the simulation time, indicating some variation in the structure that remained.

A minimum deviation from the initial structure seems to occur at pHs 4, 5, and 6, especially in the dimer case, which is concordant with the fact that the dimer is more stable at pHs near the isoelectric point (~5.2) [5]. Indeed, it is at pH 5 that the structure has the lower RMSD. At pH 7 a detailed analysis of the dimer is not possible, due to shorter run times.

Radius of Gyration

The radius of gyration is a measure of the protein compactness, wherein variations in this property are usually related with conformational changes.

The results are shown in **Figure 4.3** and correspond to the mean average of the radius of gyration of the monomer, dimer and its two chains that were calculated independently for both the open and closed conformations. Not all the replicates have the same simulation time and there are missing two replicates at pH 7: replicate 2 on the closed monomer and replicate 1 of the closed dimer.

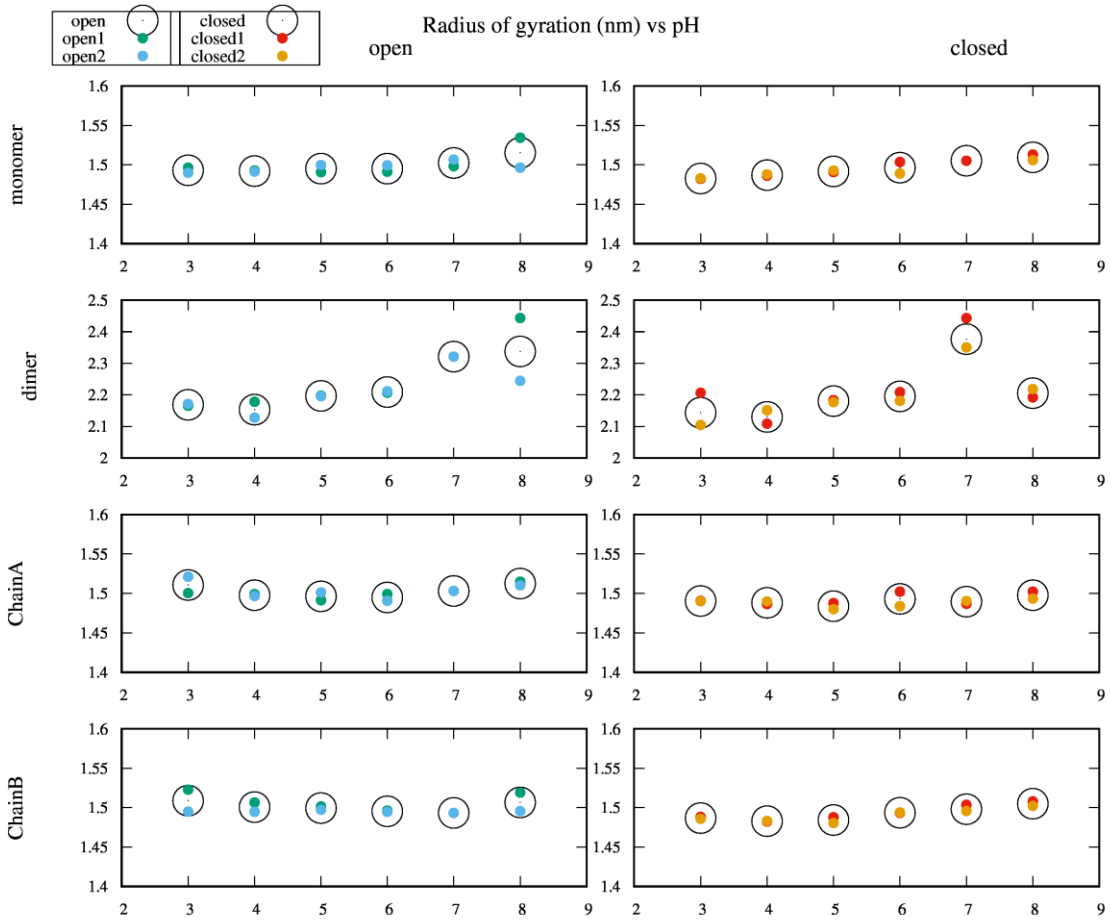


**Figure 4.3** Mean radius of gyration as function of the pH considering the two replicates, for the open and closed conformations. Each replicate is shown with a filled smaller circle and its average their represented with a larger empty circle.

For the monomer, both in the open and closed conformations, the radius of gyration is always approximately 1.5 nm, with a very slight increase at pH 8 in the open conformation. At pH values where BLG appears to be more stable, as seen in the RMSD, the value 1.5 nm is observed. At pH 8, this small increase could be related with its corresponding RMSD, that shows higher deviations from its initial structure, but, more analysis must be done for a correct interpretation.

Variations of this radius are more pronounced in the dimer. As for the monomer, the radius of gyration of the dimer is stable at a pH range from 3 to 6, but in this case the value is higher due to the two chains (2.1-2.2). The difference starts at pH 7, where an increase in the radius of gyration is noted, both in the open and closed conformations, indicating some conformational variability. At pH 8, this increase continues in the open conformation, as opposed to the closed conformation where this radius goes back to its standard value.

Analyses of the two chains independently, shows a radius that remains in a value of approximately to 1.5 nm, both in chain A and B, similar to the monomer. This reflects some stability of these

three structures (chain A, B and monomer) that is less observed in the overall dimer, indicating that the variations that occurred in the dimer are not related to monomers individually but to the overall dimer, which could correspond to its dissociation that is known to occur at higher pH values. Indeed, as seen in **Appendix I**, section **A1**, the radius of gyration as function of time shows that the replicates that suffered dissociation are those that had an increase in the radius of gyration.

Hydrogen bonds

By analyzing the number of main-chain hydrogen bonds, we are able to follow the loss or gain of the backbone structure. The overall results for the number of hydrogen bonds are depicted in **Figure 4.4**.

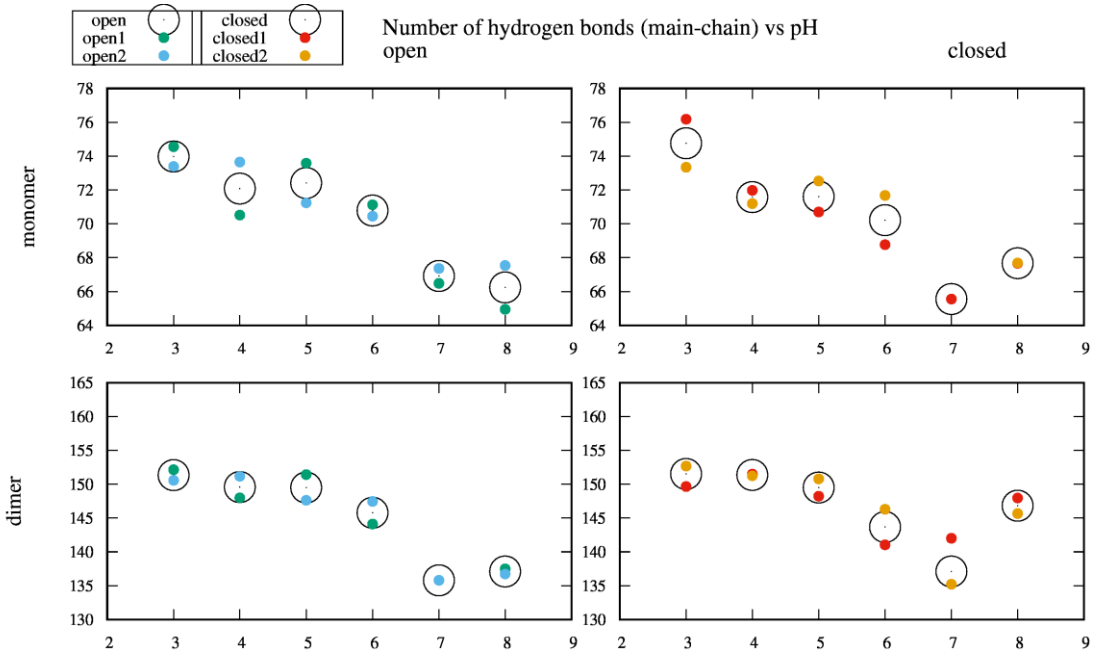


**Figure 4.4** Mean number of main-chain hydrogen bonds as function of the pH considering the two replicates for the open and closed conformation. Each replicate is shown with a filled smaller circle and their average is represented with a larger empty circle.

Considering both the open and closed monomer, there is a significant decrease in the number of hydrogen bonds at pH 7 and 8. At the pH range from 3 to 6 this decrease is observed but less accentuated, being at pH 3 that BLG has the maximum number of H-bonds. The decrease observed at higher pHs appears to be correlated with the RMSD analysis, wherein higher deviations at these pHs could correspond to the loss of some backbone structure.

The dimer shows a similar profile, with a decrease in the number of hydrogens bonds for higher pH values, with exception of pH 8 for the closed conformation. At pH 7, BLG has the lowest value of the number of hydrogen bonds of all the pHs, being this decrease maintained in the open

conformation at pH 8. For the closed one, at that pH, there is a significant increase, that is similar to the H-bonds of the more stable pHs (3, 4 and 5).

The observed decrease at higher pHs could correspond to the loss of some structure or/and, for the dimer, to the separation of its monomers, where the loss of hydrogen bonds occurs at the interface. The higher stability observed at pH 8 in the closed dimer is also observed in the RMSD analysis. At that pH, the open dimer appears to have higher deviations as opposed to its closed conformation, suggesting that there is an additional factor stabilizing the interactions at the interface and/or its secondary structure. However, in the RMSD analysis, a slight increase appears in the last ns of the simulation, which could indicate that the closed dimer was starting to dissociate. Further analysis must be conducted in order to understand where this loss of hydrogen bonds occurs and what it affects.

By analyzing the behaviour of the number of hydrogen bonds as a whole in both the monomeric and dimeric form, the decrease in the number of hydrogen bonds at higher pHs seems to be correlated with experimental studies. For the dimer, at pH values apart from the pI, BLG becomes less stable, this effect being more pronounced at higher pHs [7]. Thus, predominantly at pH 7, this separation was well observed in our simulations, as reflected in shorter simulation times. The comparison of the results at pH 8 in the closed conformation with experimental data, should take into account that, at that pH, the closed conformation is less populated in the real system, since the closed-open transition occurs around 7.5 [18].

Secondary structure

In order to understand if there is any relation between the pH and BLG secondary structure, the β-sheets and α-helices were followed during the simulation time, their fractions being represented as an average of the two replicates as function of pH. These analyses will also elucidate if the loss of hydrogen bonds was at the secondary structure level.

The fraction of β-structure and α-helix for each simulated pH value, considering the monomer and dimer for the open and closed conformation, each one with two replicates, is shown in **Figure 4.5** and **Figure 4.6**.

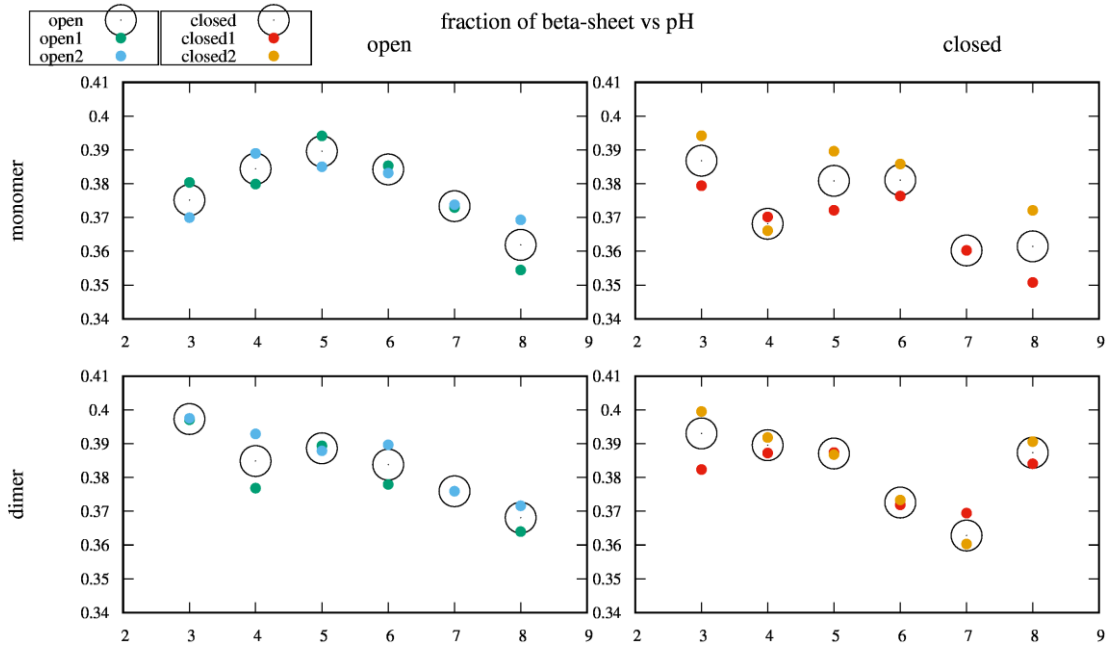


**Figure 4.5** Mean fraction of the residues in β -structure as function of the pH, considering the two replicates for the open and closed conformation. Each replicate is shown with a filled smaller circle and their average is represented with a larger empty circle.

Focusing on the monomer in the open conformation, there is a clear trend with a gain of β-sheet from pH 3-5, and progressive loss from pH 6-8. In the closed conformation, the profile is quite different, and there is some disagreement between the two replicates. At pH range from 3 to 6 the fraction of β-sheet is approximately 0.39, similar to the open conformation, but with a slightly smaller value at pH 4. These values are the pH values where BLG structure has the maximum of β-sheets, and it is also at these pHs that BLG appears to be more stable with less oscillations in the previous analyses. At pH 7 and 8, a higher decrease in the β-sheet structure is observed.

The dimer shows that, when in the open conformation, there is a progressive decrease from pH 3 to 8, being its maximum observed at pH 3. The closed conformation shows a similar profile but the decrease is more pronounced starting at pH 6, as opposed to the open conformation. Timasheff et al [94] observed a loss of the closed dimer β-structure between pH 4 and 6, which is in agreement with these results. At pH 8, it appears that BLG gains again its β content.

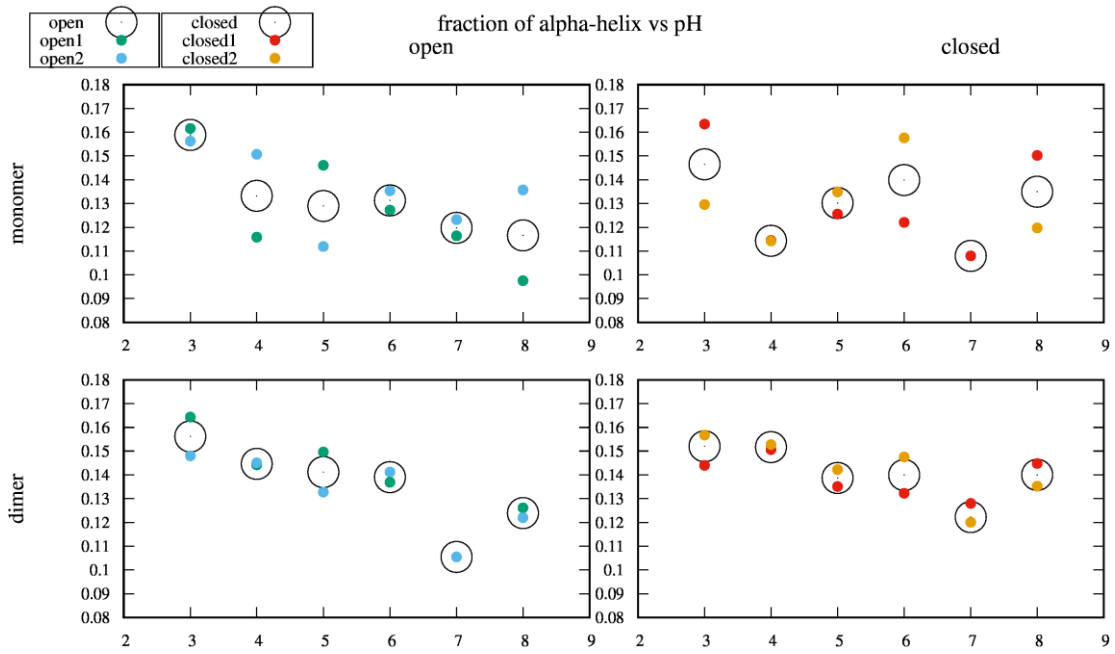


**Figure 4.6** Mean fraction of the residues in helix as function of the pH, considering the two replicates for the open and closed conformation. Each replicate is shown with a filled smaller circle and their average is represented with a larger empty circle.

BLG has one helix in its native structure and small transient helices may form during the simulations; the fraction of residues in helix as function of the pH is represented in **Figure 4.6**.

Considering the monomer in the open conformation, a progressive loss of the α-helix seems to occur, as in the β-sheet, decreasing considerably after pH 3. In the closed conformation this profile changes, having also a maximum at pH 3 but with a decrease at pH 4 followed by a gain of helix structure at pH 5 and 6. pH 7 shows a loss of the helix, in the same range value as pH 4, but pH 8 regains that helix. No trend is observed and we have a high dispersion of values which makes it difficult to reach a conclusion.

As for the dimer in the open conformation, a progressive loss of its helix fraction is present, very similar to the one observed for that conformation of the monomer, but with a minimum at pH 7. However, it should be noted that this minimum is influenced by the lack of one replicate. In the closed dimer, the helix fraction is maintained at all pHs, with a slight decrease at pH 7.

Taking into account the global behaviour of the content in β-sheets and helix, this loss of structure is similar to the profile of the number of hydrogen bonds, showing that the decrease in the number of backbone hydrogen bonds corresponds also to a decrease in the secondary structure at the same pHs. In fact, this can be observed in **Figure A2.1**, in **Appendix I** where the sum of the two types of secondary structure is depicted.

The peculiar stability of the closed dimer at pH 8 observed in the H-bonds analysis is also present in the secondary structure plots.

RMSF analysis

The root mean square fluctuations (RMSF) measures the fluctuations around its average structure, capturing the flexible regions of the protein.

The overall results for the RMSF of the monomer and dimer, with its two chains calculated independently, are depicted in **Figure 4.7**.

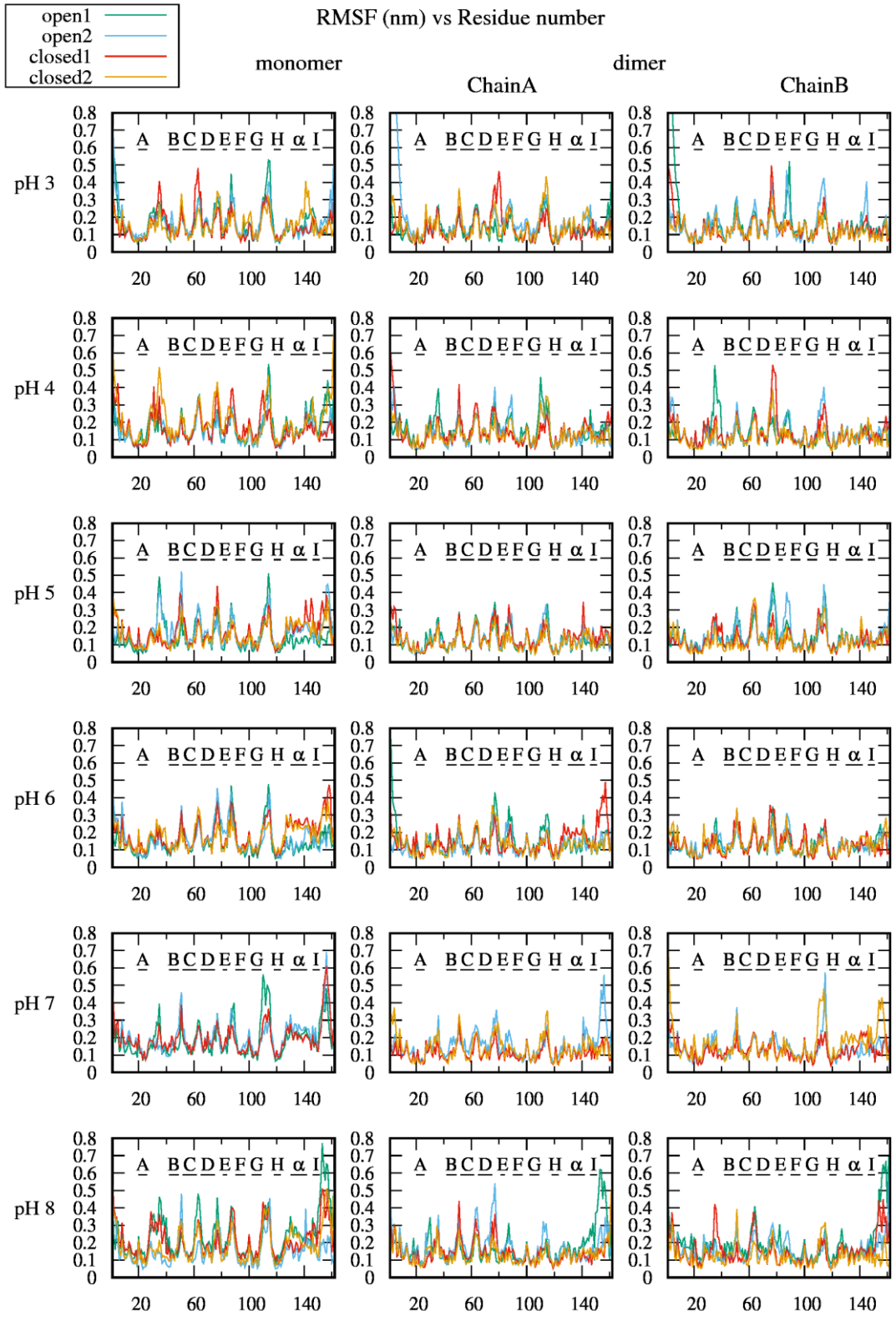


**Figure 4.7** Root mean square fluctuations for each residue, considering the monomer and dimer simulated at the pH range from 3 to 8. The position of the β-strands (A-I) and $\alpha$-helix in the native structure is indicated

.

An overview of the RMSF, for the monomer and dimer, shows that both seem to have more flexible regions in the loops, as expected. A difference stands out at the N-terminal of the chains A and B of the dimer at pH 3, corresponding to different replicates of the open conformation. This sharp increase suggests some loss of structure which is, indeed, correlated to the high deviations observed at this pH in the RMSD of the dimer for the same replicates. Another difference happens at higher pHs, 7 and 8, which are the same pHs that have marked differences in all the precending analyses. At these pHs, for both the monomer and dimer, the region close to the C-terminal becomes more flexible, with higher fluctuations than at lower pHs, although that region is always somewhat flexible in the monomer. In agreement with the RMSD and H-bond analyses, this suggests some loss of structure at higher pHs that does not happen in the known β-sheets (A-H) or in the α-helix. By comparing the dimer with the monomer, in general lower fluctuations are seen in the dimer, suggesting that association helps stabilizing the dimer.

Further analyses were conducted to better understand the processes that happened at the N-terminal and close to the C-terminal region.

### 4.2.2 Protein Domains

N-terminal region

By visual inspection of the trajectories displaying high RMSF in the N-terminal at pH 3 (chain A of replicate 2 and chain B of replicate 1 in the open dimer) using PyMOL software, a coil was observed in approximately the first 10 residues that suffers some alterations. Starting from ~55 ns for replicate 1, as observed in the RMSD analysis, this coil turns to the opposite side and becomes more exposed, as seen in **Figure 4.8**. This disorder could facilitate the binding of proteases, thus contributing to the high proteolytic susceptibility of BLG at low pH.

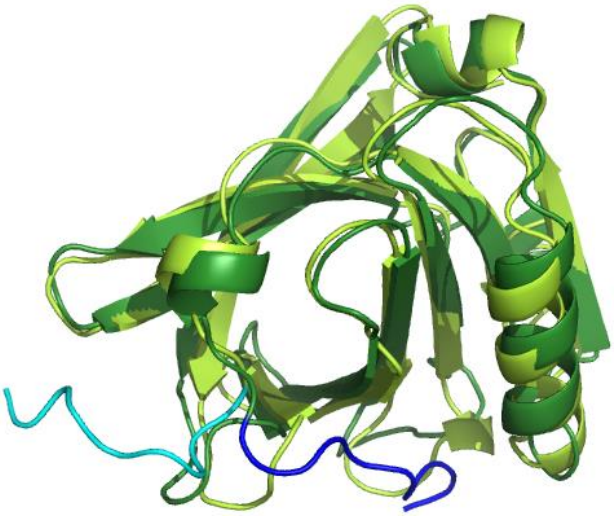

**Figure 4.8** Fitting of two selected BLG structures for the replicate 1 of chain B (open dimer at pH 3). Dark green shows the structure before the change in the coil (dark blue) occurs, taken at 1ns. Light green shows the representation of BLG structure after the alterations suffered by the coil (light blue), captured by the snapshot at 60 ns. A similar alteration is observed for replicate 2 of chain A.

Single-turn α-helix

The higher fluctuations at pH 7 and 8, close to the C-terminal region, were investigated by **DSSP** analysis (see **Appendix I**, section **A2**) and by visual inspection using the PyMOL software. With these, we have found the presence of a single-turn α-helix composed by four residues (153-156). To see if the RMSF observations were indeed correlated with this helix, RMSD analysis of this helix was performed. The results obtained for each simulated pH value, considering both open and closed conformations, for the monomer and dimer, are depicted in **Figure 4.9**.

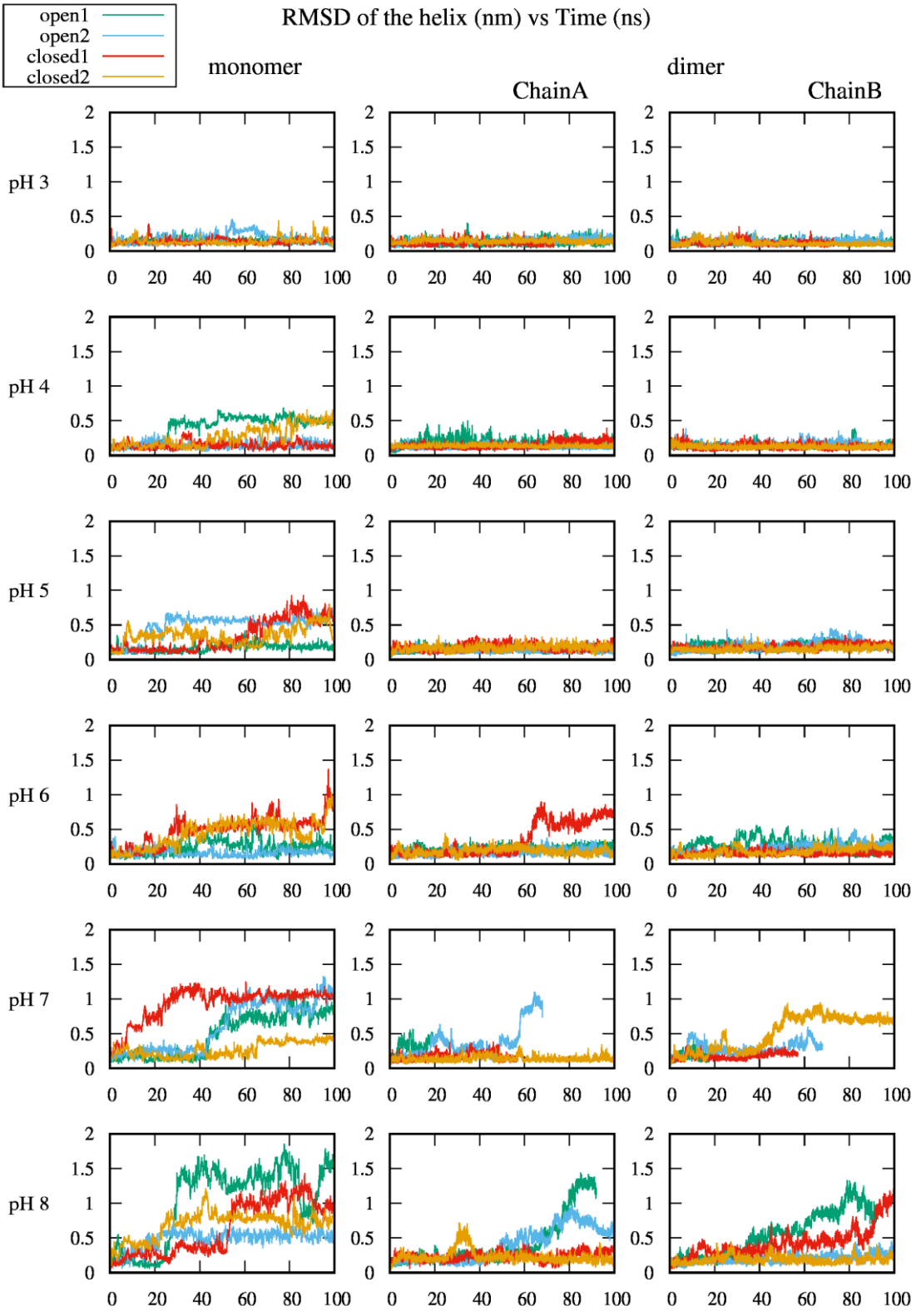


**Figure 4.9** RMSD of the helix formed by the residues 153-156 considering the monomeric and dimeric forms, in the open and closed conformations, for each pH value. All protein backbone atoms were fitted to the respective atoms in the initial conformation. This helix was already present in the X-ray structure.

The RMSD for both the monomer and dimer appears to have higher values at higher pHs, especially at pH 7 and 8, this effect being much more pronounced in the monomer. Indeed, the deviations in the monomer start at a lower pH than in the dimer, suggesting some additional stabilization of this helix by dimerization at low pH, with a higher destabilization occurring at pH 8. Due to shorter runtimes, it is harder to reach conclusions about the dimer in pH 7, but the same observation can also be noticed for some replicates.

When the helix segment starts to suffer deviations during the simulation, they are maintained and appear to be irreversible during the course of the simulation.

By comparing the RMSF results with these, the higher fluctuations indeed correspond to the same replicates that have higher deviations. This observation is evident at pH 8, especially for the replicate 1 in both the monomeric and dimeric form, that being the replicate with higher fluctuations and deviations from its initial structure in both analyses. **Figure 4.10** shows the pH-dependent unfolding of this helix for the replicate 1 of the monomer in the open conformation.

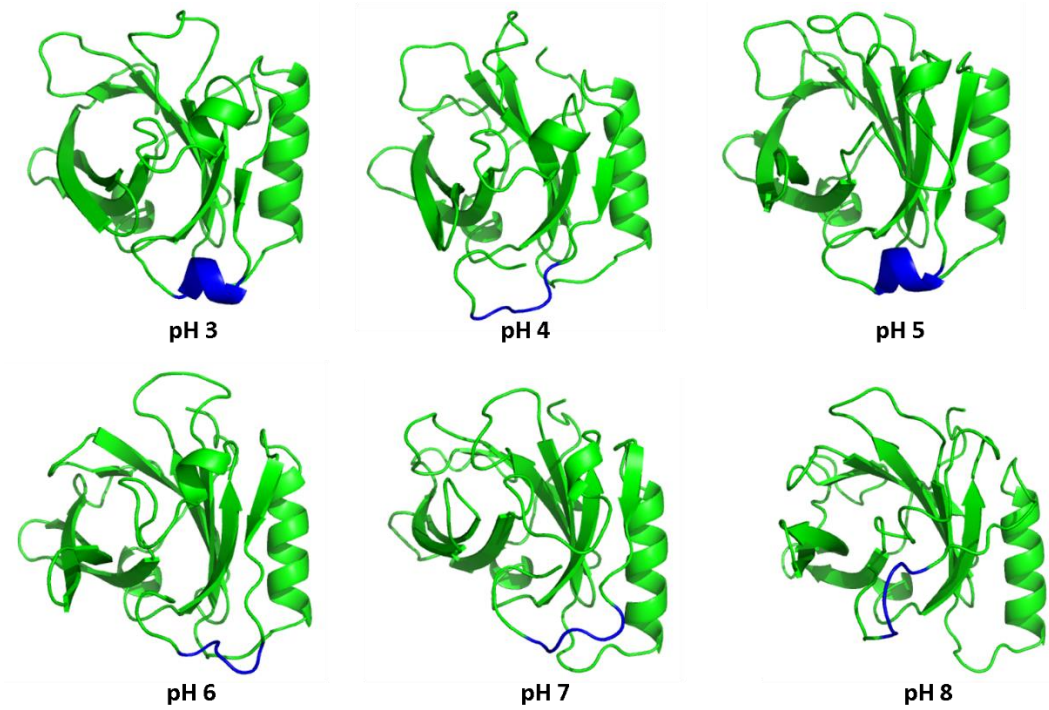


**Figure 4.10** BLG structure (in green) for the replicate 1 in the open monomer for all pHs. In blue, the helix segment with residues 153-156 is shown. These structures were taken from the last ns of the simulation.

Therefore, we can conclude that the alterations observed were due to an unfolding in a pH-dependent way, especially at higher pHs.

EF loop

The EF loop in the BLG structure is known to be involved in a gating mechanism associated with ligand binding. In this loop, Glu89 is responsible for this reversible pH-dependent transition. When ionized it becomes exposed, allowing the binding of ligands, and becomes buried when protonated. To identify if this transition was observed in the simulations runtime, the trajectories were visually inspected using PyMOL. However, we did not observe the overall open/closed transition. Despite that, to better characterize the movement of this loop and to see if there was any pH-dependence, some analyses were performed (RMSD of the EF loop, distance between Glu89 and Ser116, and the solvent accessible surface area of the loop), but these were too inconclusive. In future work a more detailed analysis of this loop will be performed.

### 4.2.3 Interface

The study of dimerization is our main task and understanding its subsequent dissociation is essential. In this section the focus will be the interface between the two monomers, by analyzing the contact surface, the number of hydrogen bonds at this interface, and also what are the residues involved.

Contact surface

A property that can follow the dissociation of the dimer is the contact surface between the two monomers. Our trajectories (**Table 4.2**) showed a dissociation predominantly at pH 7, but also at pH 3 and 8, thus, it is expected that the contact surface decreases at these pHs. The results from the contact surface are shown in **Figure 4.11**.

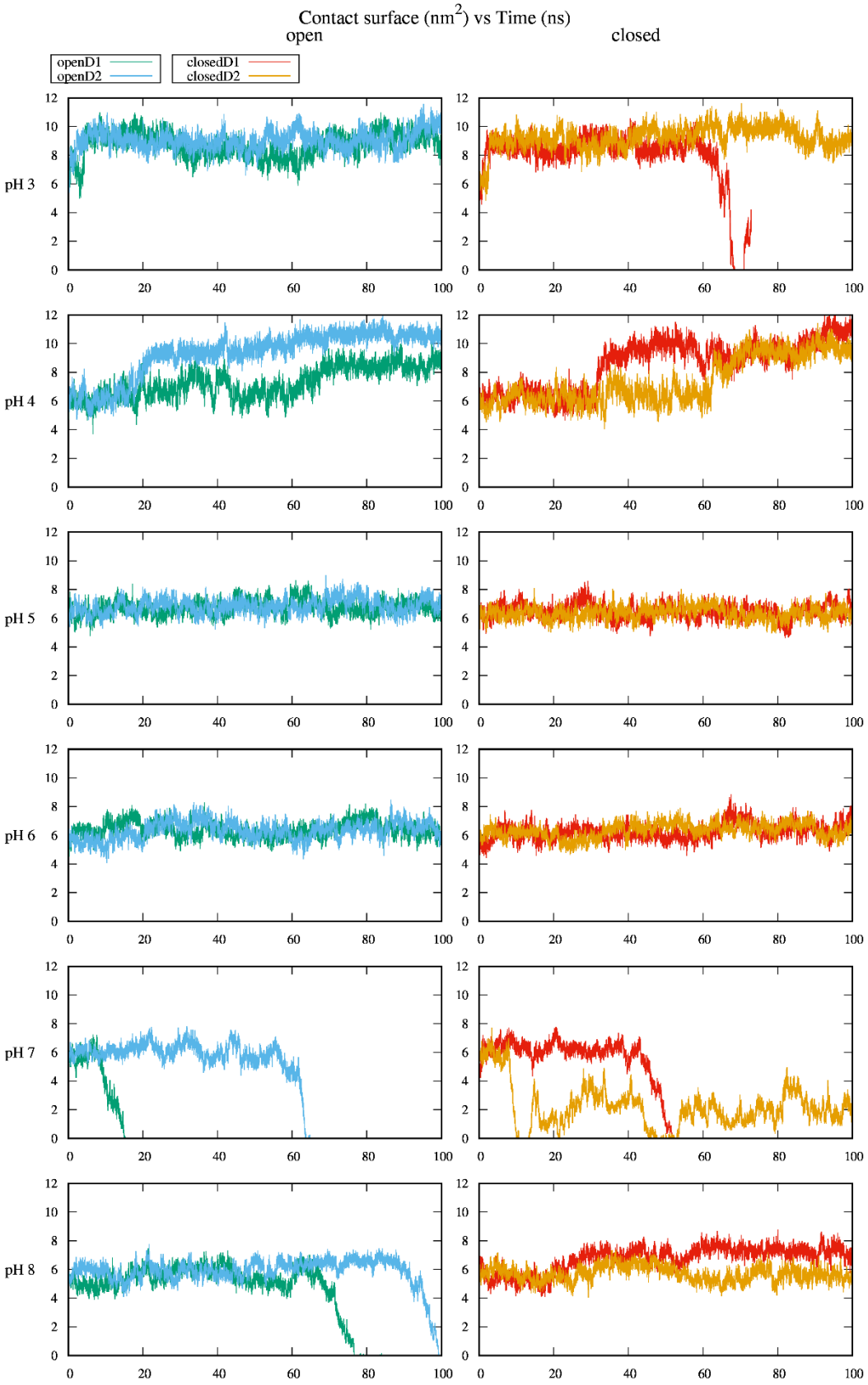


**Figure 4.11** Contact surface of the dimer in the open and closed conformation, for each pH value.

The contact surface clearly shows us the dissociation of the monomers when their contact area drops to zero, that being particularly evident for the closed dimer at pH 3 (replicate 2), both conformations at pH 7, and the open conformation at pH 8. The contact surface results are correlated with the results of the distance between the center of mass of chain A and chain B (see **Appendix I** in section **Distances**).

At pH values apart from the pI, the interactions at the interface are known to become weaker and the dimer starts to dissociate [5]. In our simulations, this dissociation is more pronounced at higher pHs, but especially at pH 7, where the dissociation is already present shortly after the start of the simulation. It is also interesting to note that, for the replicate 2 of the closed conformation at pH 7, the dimer dissociates almost completely but there are still some remaining interactions. At pH 8 in the closed conformation, the dimer seems stable, which is indeed correlated with the previous analyses, appearing to exist an additional factor that stabilizes this closed dimer at this pH, or that more simulation time was needed to obtain a similar profile to the one in the open conformation. Nonetheless, as previously noted, this closed conformation is not expected to be the predominant one at that pH.

At pHs were the dimer is known to be more stable (5 and 6) both conformations have a similar profile, but with a lower contact surface than at lower pHs. This is a bit counterintuitive and surprising, since it is at these pHs that its aggregation is known to be highest. In fact, looking at the overall picture, it appears that the area of contact reveals two states: a more compact one predominant at pH 3, with a contact surface around 9-10 $nm^2$, and a more relaxed one predominant at pHs 5-8, with a contact surface around 6-7 $nm^2$. There also appears to be a combination of these two trends at pH 4. It should also be noted that the initial configurations seem to correspond to the state with smaller contact surface.

To understand the changes at the structural level, BLG structure at pHs 3, 4 and 5 was inspected using PyMOL software, and the fitting of all three is represented in **Figure 4.12**.

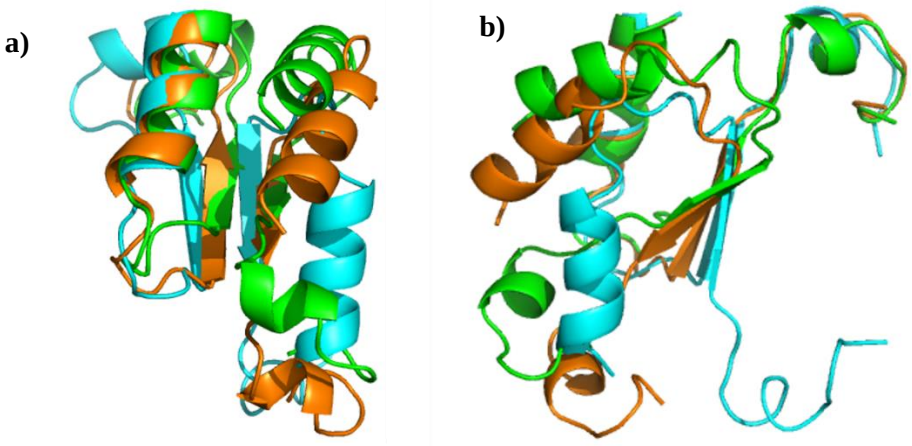


**Figure 4.12** Superimposed BLG structures at pH 3 (green), pH 4 (orange) and pH 5(blue). Front view (a) with respect to the alterations in $\alpha$-helix and side view (b) focusing on the βI strands. For clarity, only the interface of the dimer is represented, including the helix and β -strand I of each monomer. These structures were taken at the 40 ns snapshot from the closed dimer.

This visual inspection allowed us to notice some changes in these α-helices and β-sheet at the interface (strand I). For a better understanding of the following explanation, regularly spaced snapshots of the helices, along the simulation, are represented in **Figure 4.13** and the conformational changes of the βI strands in **Figure 4.14**.

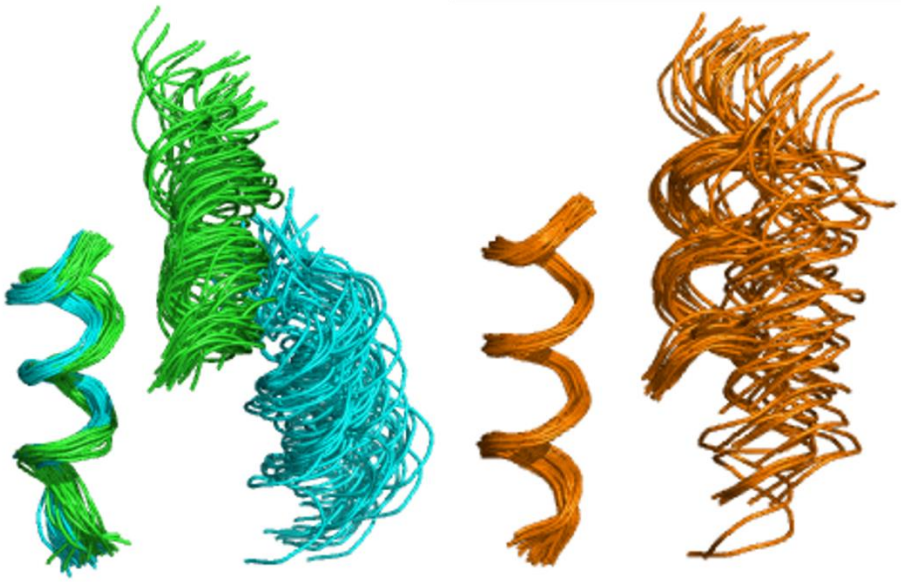

**Figure 4.13** BLG helix structures at pH 3 (left green), pH 5 (left blue) and pH 4 (right orange). Regularly spaced snapshots are shown by fitting the helix on chain A (left helix).

In **Figure 4.13**, the green and blue representations correspond to pH 3 and 5, respectively, their fitting being made with the helix of chain A (on the left), showing the structural alterations in the α-helix of chain B. We observed that at pH 5, the simulation snapshots are very similar to the X-ray crystallographic structures used in this work (1BSY and 3BLG) which were indeed determined at pH > 5. At pH 5, the helices in both chains are parallel to each other, which is not observed at pH 3. A rotation of the chains relative to each other appears to change the structural arrangement of this helix, making it perpendicular to the other and closer. Indeed, this approximation seems to be the cause of the higher contact area between the two monomers seen above, since residues that could not interact directly before [5] are now capable. At pH 4, we can visualize a combination of the conformational arrangements at pH 3 and 5, corresponding to the two plateaus observed previously in the contact surface area plots (**Figure 4.11**).

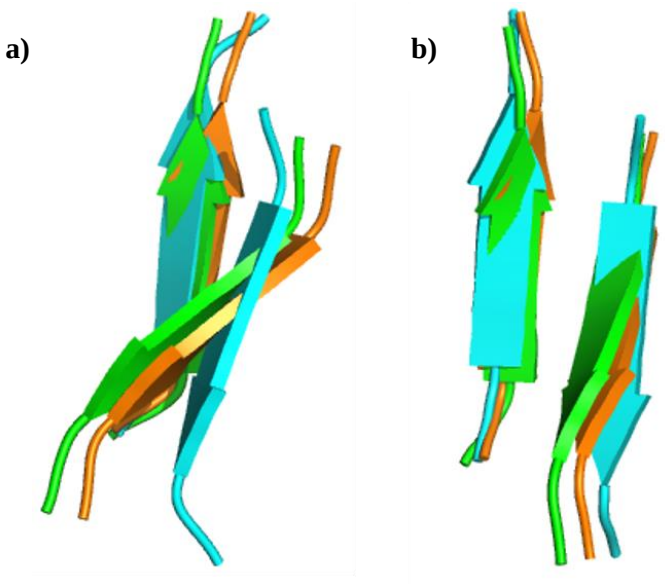


**Figure 4.14** Fitting of BLG structure at pH 3 (green), pH 4 (orange) and pH 5 (blue) of the β -sheet in I strand in chain A (left strand). Only one snapshot (40 ns) is represented to ease visual observation, but all the snapshots have a similar behaviour.

By looking at the βI strands behaviour, a relative reorientation of the β-strands occurs at pH 3 and 4. This is related to the approximation and rotation of the α-helices at these pHs, changing the behaviour from two parallel βI strands (pH 5) to two "crossed" βI strands (pH 3 and 4).

Some authors [12][9][11] have verified a difference in the orientation of this helix when comparing their NMR structure, obtained at low pH, with crystallographic data, in agreement with our results. In the compact state, the helices of the two monomers are closely interacting in a region where several carboxyl and amino groups exist, suggesting a possible cause for the pH-dependence, which will be investigated in subsection **4.2.4**.

Hydrogen bonds

The number of hydrogen bonds involving the main-chain at the interface was measured, a correlation being expected with the previous contact surface analysis. The number of H-bonds as a function of time, for the dimeric form in the two conformations, is depicted in **Figure 4.15.**

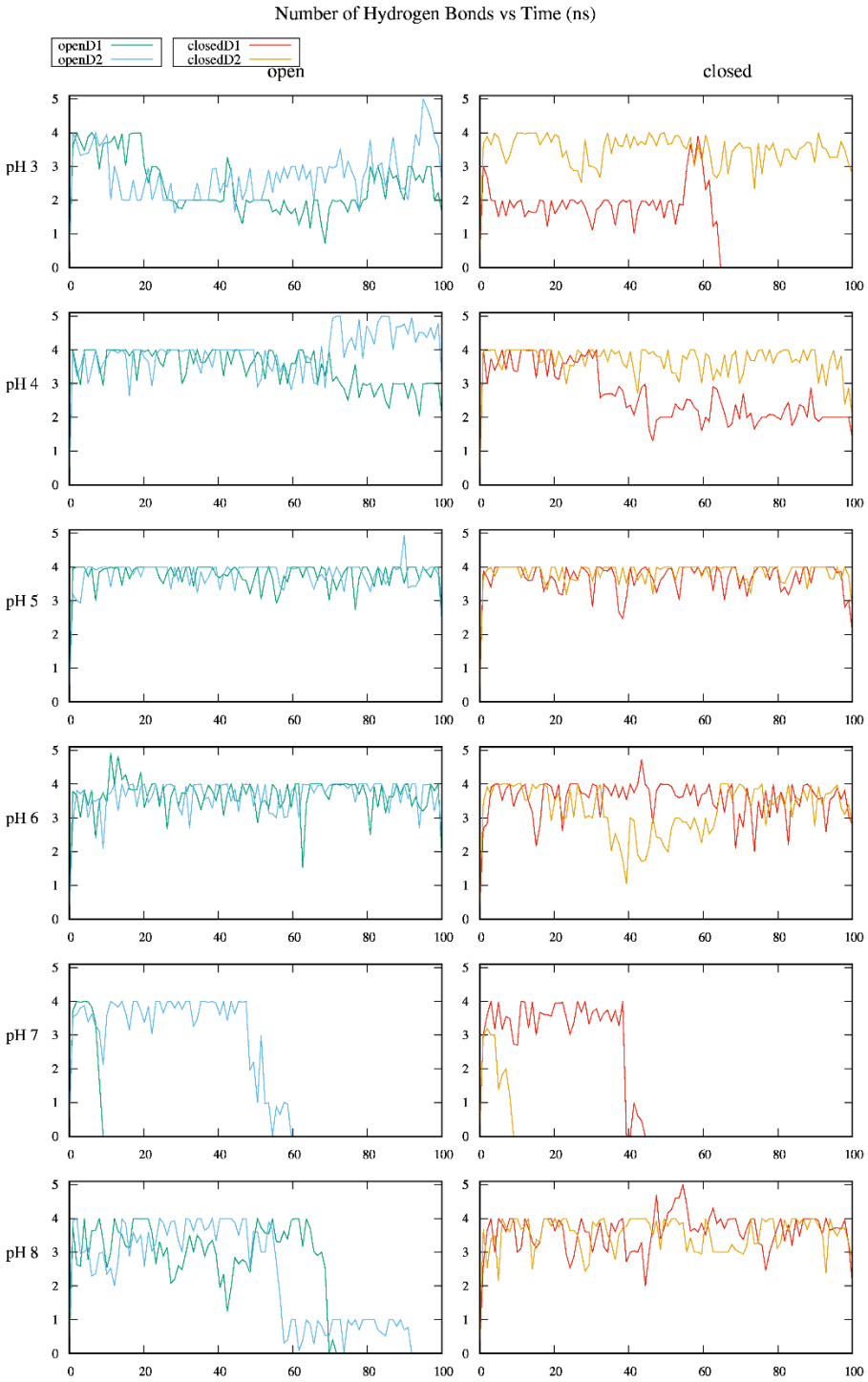


**Figure 4.15** Number of hydrogen bonds between the main-chain of the two monomers, considering the open and closed dimer, for each pH value. Smoothing was done using a kernel function.

The decrease in the number of H-bonds in the main-chain is very clear for the replicates that have dissociated. For the replicate 2 of the closed dimer at pH 7, whose monomers moved apart but still with some remaining interactions, these interactions are not captured in this analysis, indicating that they are not due to main-chain hydrogen bonds but to other type of interactions. For the pHs where the structure was seen to remain associated, a number of 3-4 H-bonds is observed, except at pH 3 and 4. As mentioned above, pH 3 was seen to be in a compact state, also present at pH 4, accompanied by a distortion of the βI strands at the interface (**Figure 4.14**). This distortion disrupts the H-bonds present in these βI strands, lowering the number of H-bonds at these pHs. This effect is more pronounced for the majority of the replicates at pH 3 while, again, the profile at pH 4 appears to be a combination of the compact and relaxed state profiles. For the replicate 2 of the open dimer at pH 8, we can see a decrease in the number of H-bonds that precedes the dissociation that occurs in the last few nanoseconds, as seen in the contact surface analysis. By visual inspection, we observed a slight rotation of the two βI strands relative to each other.

Interactions at the interface

The distances between some residues that are thought to interact at the interface of the dimer [34] were considered, their calculation being made between atoms at the end of the side chains. Very small initial distances are expected with a progressive increase for the monomers that move apart. The chosen residues were Asp33 and Arg40, located in the AB loops, Arg148, located in the I strand, and Asp137, located in the α -helix, their pairs being described in **Table 4.3**.

**Table 4.3** Residue pairs selected between Chain A and Chain B at the dimer´s interface.

| **Chain A-Chain B** | **Atom pair** |
|---|---|
| Asp33-Arg40 | Cγ-Cζ |
| Arg40-Asp33 | Cζ-Cγ |
| Asp137-Arg148 | Cγ-Cζ |
| Arg148-Asp137 | Cζ-Cγ |
| Asp137-Asp137 | Cγ- Cγ |
| Arg148-Arg148 | Cζ- Cζ |

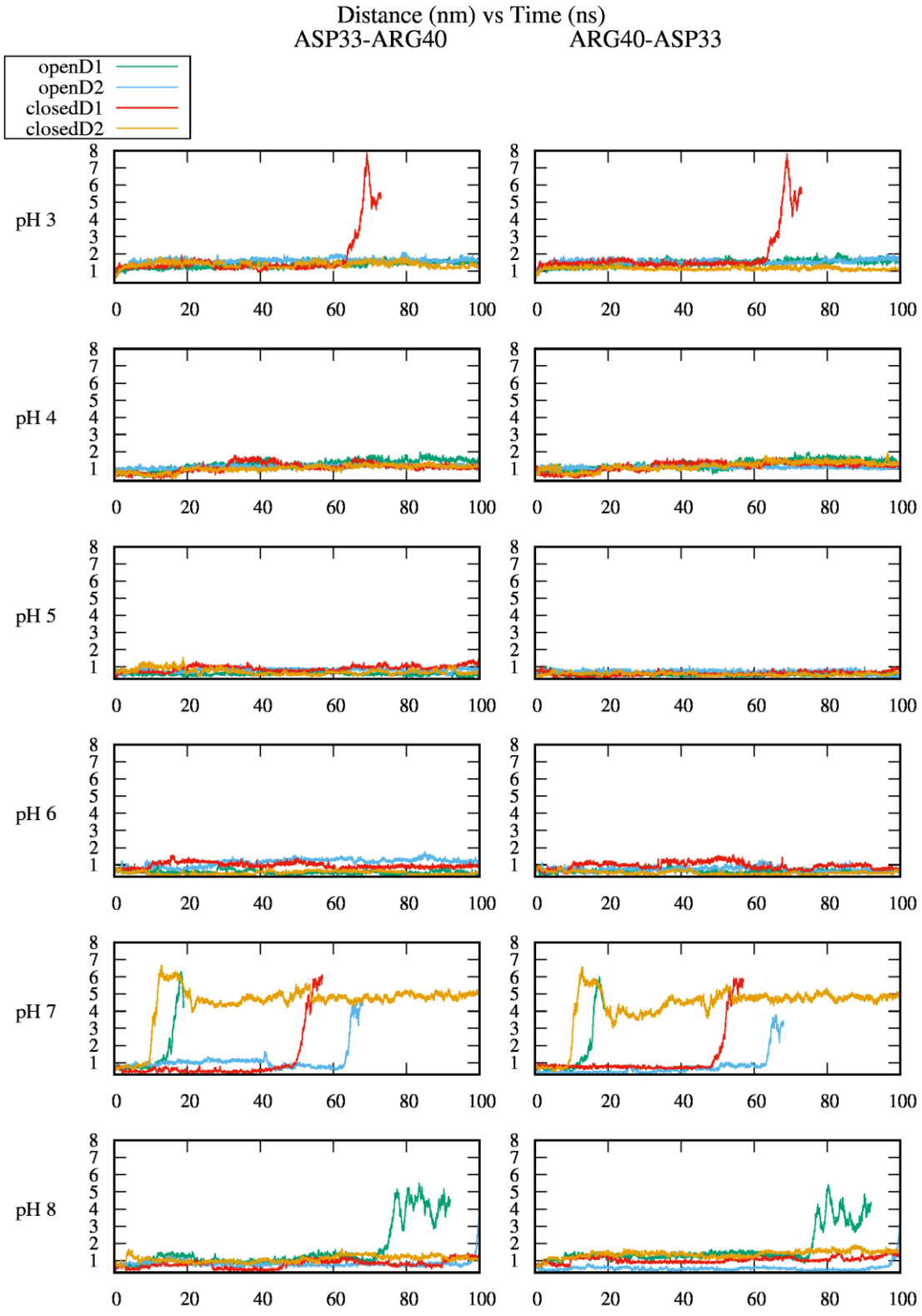


**Figure 4.16** Distance between Asp33 and Arg40 of different chains, considering the open and closed dimer, for each pH value. For details see Table 4.3.

As shown in **Figure 4.16**, the profile of the distances between Asp33 of chain A with Arg40 of chain B, and vice versa, is practically the same, so these two distances will be analysed together.

The visible increase in distance observed in some of these plots, show us a direct correlation with the decrease in contact surface for the replicates that had been seen to dissociate. For some of these dissociated replicates, the distance between Asp 33-Arg40 is higher than for others. This is not necessarily related with a higher overall separation of the two monomers, but it can also be due to the rotation of the monomers relative to each other, changing the distance between these two residues but not necessarily between the COM of the monomers. Indeed, this rotation was observed by visual analysis of the trajectories, and a more detailed analysis will be performed in future work.

At pH 8, the monomers move apart in both replicates of the open conformation , but the distance between Asp33 and Arg40 is not maintained only in one replicate, due to the dissociation of the other happening almost at the end of the simulation runtime. For pHs were the dimer is known to be more stable, a short distance between the residues in the two monomers is maintained.

Similar distance profiles were observed for the several pairs involving Arg148 (βI strand) and Asp137 (α-helix), and these results are presented in **Appendix I**, section **Distances**.

As already mentioned above (subsection **4.2.3**, contact surface), the simulations at pHs 3, 4 and 5 have shown some differences in the protein structure, more exactly, in the βI strands and α-helices. Since Arg148 and Asp137 are located in these regions, a focused analysis at small distances was conducted- **Figure 4.17**.

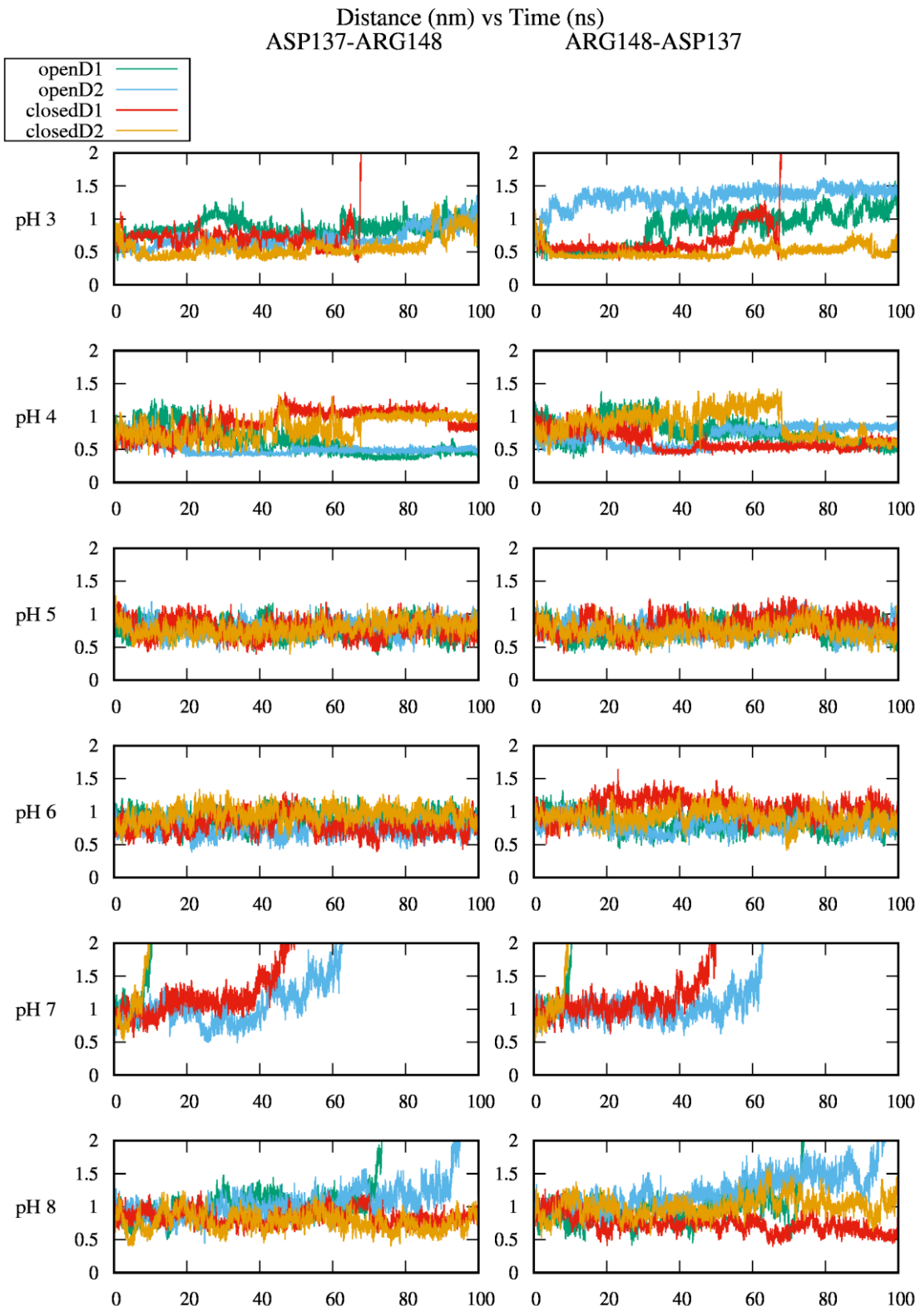


**Figure 4.17** Distance between Asp137 and Arg148 of different chains, considering the open and closed dimer, for each pH value. For details see Table 4.3.

At lower pHs, more precisely 3 and 4, for some replicates there are very low distances (0.4-0.5 nm) with small oscillations, indicating that the two residues get closer to each other, forming a salt bridge or H-bond. The same is observed for the interactions between corresponding residues: Asp137-Asp137 and Arg148-Arg148 (**Appendix I**, section **Distances**).

### 4.2.4 Protonation

Global titration curve

The global titration curve for both the monomer and dimer, in the open and closed conformation, can be seen in **Figure 4.18**.

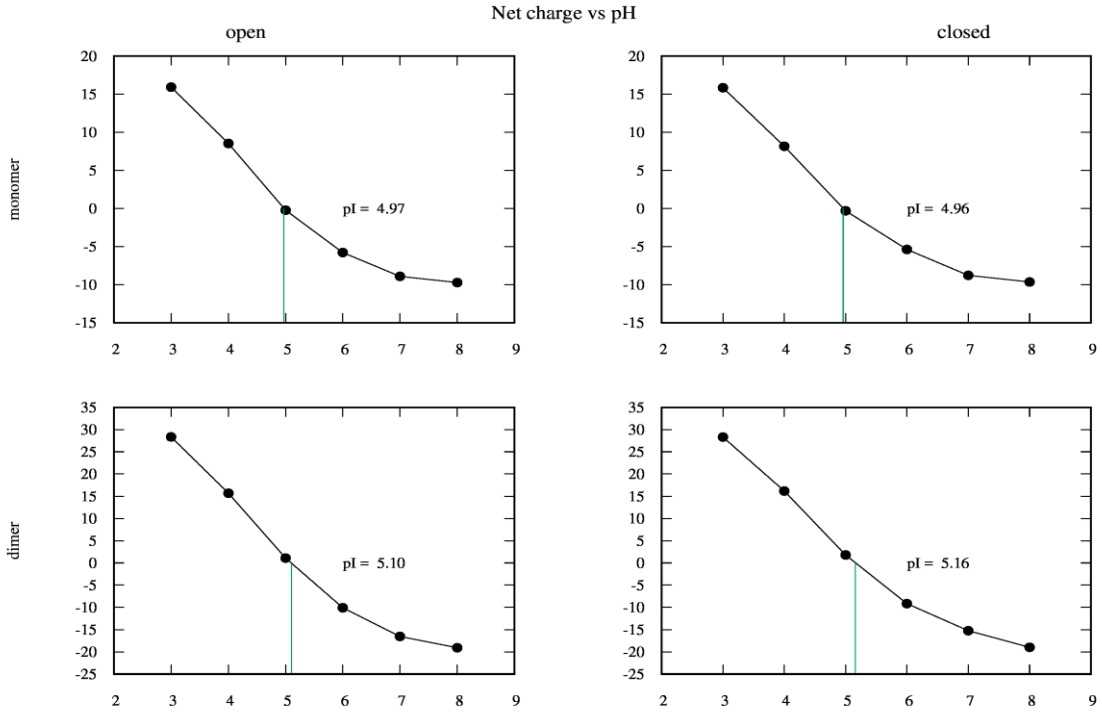


**Figure 4.18** Mean protein net charge as a function of the pH, considering both the monomer and dimer, with open and closed conformations. The pI is also shown.

By comparing the obtained pI with the ones from PB/MC calculations **(Figure 4.1** in section **4.1**), the trend is very similar. The difference between the open and closed structures is not as large as in rigid PB/PC calculations, which could be due to the EF loop, but further analysis must be made on this loop. When comparing the dimer with the monomer, its pI value is slightly higher, maybe due to the presence of titrable sites at the interface that are interacting with each other. These pI values, obtained with CpHMD, are in principle more precise than the ones obtained from rigid PB/PC calculations, since it is a more realistic method. Indeed, they are more alike the values obtained with experimental data (~5.2).

Titration of the individual groups

The pKa values for each titrable site were obtained from the fit to a Hill equation and are shown in **Table 4.4**.

**Table 4.4** Computed pKa values (with errors) for the open and closed conformations of the monomer and dimer, and typical values in solution ($pK_{typ}$) [93][92].

| Residue | Computed pKa | | | | | | pKtyp |
|---|---|---|---|---|---|---|---|
| | Open | | | Closed | | | |
| | Monomer | Dimer | | Monomer | Dimer | | |
| | | Chain A | Chain B | | Chain A | Chain B | |
| **Nter** | 6.93±0.05 | 6.85±0.03 | 7.39±0.21 | 7.33 ± 0.13 | 7.11 ±0.09 | 7.22 ± 0.03 | 8.00 |
| **Asp11** | 3.34±0.04 | 3.33±0.02 | 3.05±0.02 | 3.38 ± 0.06 | 3.30 ± 0.03 | 3.36 ± 0.03 | 3.94 |
| **Asp28** | 3.43±0.03 | 3.35±0.07 | 3.28±0.24 | 3.39 ± 0.28 | 3.13 ± 0.38 | 3.36 ± 0.08 | 3.94 |
| **Asp33** | 4.29±0.09 | 4.45±0.05 | 4.67±0.06 | 3.83 ± 0.03 | 4.61 ± 0.34 | 4.34 ± 0.05 | 3.94 |
| **Glu44** | 4.42±0.12 | 4.04±0.49 | 4.33±0.22 | 4.47 ± 0.12 | 4.52 ± 0.11 | 4.39 ± 0.06 | 4.25 |
| **Glu45** | 5.46±0.04 | 5.47±0.03 | 5.55±0.05 | 5.53 ± 0.04 | 5.65 ± 0.14 | 5.45 ± 0.09 | 4.25 |
| **Glu51** | 3.69±0.04 | 3.52±0.09 | 3.64±0.09 | 3.78 ± 0.03 | 3.77 ± 0.10 | 3.57 ± 0.02 | 4.25 |
| **Asp53** | 4.68±0.15 | 4.62±0.21 | 4.73±0.19 | 4.81 ± 0.03 | 4.54 ± 0.10 | 4.89 ± 0.34 | 3.94 |
| **Glu55** | 5.62±0.16 | 6.15±0.22 | 6.02±0.18 | 5.94 ± 0.22 | 6.78 ± 0.11 | 6.04 ± 0.04 | 4.25 |
| **Glu62** | 4.30±0.16 | 4.24±0.04 | 4.37±0.05 | 4.32 ± 0.02 | 4.20 ± 0.04 | 4.26 ± 0.06 | 4.25 |
| **Asp64** | 4.35±0.01 | 4.24±0.02 | 4.20±0.01 | 4.27 ± 0.03 | 4.34 ± 0.04 | 4.34 ± 0.01 | 3.94 |
| **Glu65** | 4.68±0.02 | 4.73±0.04 | 4.79±0.02 | 4.71 ± 0.01 | 4.70 ± 0.01 | 4.64 ± 0.02 | 4.25 |
| **Glu74** | 4.91±0.04 | 4.85±0.03 | 4.79±0.07 | 4.16 ± 0.29 | 4.68 ± 0.09 | 4.44 ± 0.08 | 4.25 |
| **Asp85** | 3.85± .07 | 4.11±0.17 | 4.19±0.08 | 3.51 ± 0.03 | 3.67 ± 0.07 | 3.74 ± 0.08 | 3.94 |
| Glu89 | 4.08±0.01 | 4.02±0.02 | 4.15±0.04 | 5.07 ± 0.14 | 5.48 ± 0.11 | 6.04 ± 0.77 | 4.25 |
| Asp96 | <3 | <3 | <3 | 3.10 ± 0.36 | <3 | <3 | 3.94 |
| Asp98 | <3 | - | <3 | - | - | <3 | 3.94 |
| Glu108 | 4.70±0.13 | 5.49±0.08 | 5.05±0.28 | 5.03 ± 0.54 | 6.59 ± 0.13 | 6.41 ± 0.11 | 4.25 |
| **Glu112** | 4.12±0.30 | 4.47±0.10 | 4.03±0.27 | 4.17 ± 0.42 | 3.69 ± 0.03 | 3.85 ± 0.16 | 4.25 |
| **Glu114** | 3.76±0.24 | 3.95±0.05 | 4.20±0.20 | 4.47 ± 0.05 | 4.36 ± 0.03 | 4.37 ± 0.09 | 4.25 |
| **Glu127** | 4.45±0.03 | 4.21±0.04 | 3.95±0.12 | 4.46 ± 0.03 | 3.99 ± 0.18 | 3.92 ± 0.05 | 4.25 |
| **Asp129** | 3.31±0.11 | 3.13±0.12 | 3.39±0.17 | 3.44 ± 0.15 | 3.40 ± 0.08 | 3.42 ± 0.06 | 3.94 |
| **Asp130** | 3.96±0.02 | 3.45±0.10 | 3.41±0.06 | 3.86 ± 0.07 | 3.04 ± 0.26 | 3.36 ± 0.46 | 3.94 |
| **Glu131** | 4.10±0.02 | 3.68±0.07 | 3.95±0.05 | 4.05 ± 0.06 | 4.08 ± 0.07 | 3.97 ± 0.03 | 4.25 |
| **Glu134** | 4.60±0.01 | 4.60±0.09 | 4.49±0.09 | 4.50 ± 0.03 | 4.61 ± 0.03 | 4.63 ± 0.04 | 4.25 |
| Asp137 | 3.49±0.05 | - | <3 | 3.09 ± 0.18 | <3 | - | 3.94 |
| His146 | 6.53±0.01 | 6.04±0.46 | 6.30±0.015 | 6.23 ± 0.03 | 5.56 ± 0.16 | 5.52 ± 0.13 | 6.54 |
| Glu157 | 5.07±0.10 | 5.29±0.03 | 5.16±0.18 | 4.96 ± 0.04 | 5.28 ± 0.01 | 5.34 ± 0.00 | 4.25 |
| Glu158 | 4.87±0.20 | 5.65±0.05 | 5.62±0.18 | 4.91 ± 0.07 | 5.56 ± 0.07 | 5.59 ± 0.11 | 4.25 |
| His161 | 5.98±0.40 | 4.72±0.99 | - | 4.86 ± 1.06 | - | - | 6.54 |
| Cter | 4.48±0.27 | 4.71±0.53 | 5.22±0.28 | 4.47 ± 0.13 | 5.41 ± 0.28 | 4.95 ± 0.38 | 3.94 |

The residues that had atypical titration curves or pKa values (not marked in bold) and that were selected to be presented and discussed in this subsection: Glu89, Asp98, Glu108, Asp137, Glu157, Glu158, His161 and Cter. Their titration curves are depicted in **Appendix I** section **A4**, and for additional analysis their protein environment was analysed using PyMOL software.

For the majority of the sites, the differences with its $pK_{typ}$ are within one pH unit, suggesting a generally solvated residue, as observed by visual inspection. However, there are some interesting sites, such as Glu89, the residue associated with the gating mechanism- **Figure 4.19**.

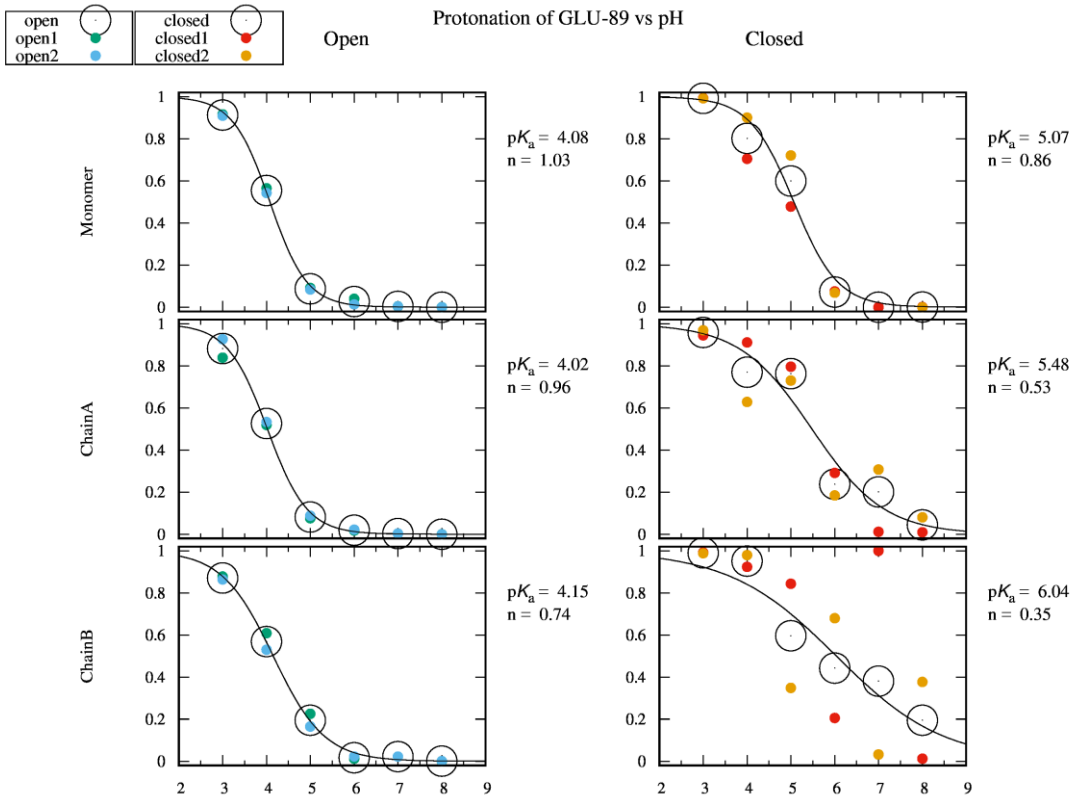


**Figure 4.19** Protonation of Glu89 considering the monomer and dimer, in the open and closed conformation, for each pH value.

In the open conformation, its pKa value appears to be normal, while in the closed conformation its pKa value is slightly higher, probably due to the more difficult accommodation of its ionized form when buried in that conformation. When comparing these values for the closed form with the ones in rigid PB/PC calculations, they are lower and with some dispersion of the averaged protonation for the replicates. As previously mentioned, a more detailed analysis of the EF loop will be necessary, which, in particular, will help to understand the protonation behaviour of Glu89.

Asp96 and Asp98, located in the FG loop have an unusually low pKa value, being ionized at almost all pH values. This could be due to their interaction with several nearby Lys residues, as observed in various snapshots, which should stabilize the ionized carboxyl form.

Lower pKa values are also observed for Asp137 in the dimer. This residue, analysed in the previous section, was thought to form a salt bridge with Arg148, or with its equivalent residue in the other chain at low pH, when at least one of them is neutral. Thus, by observing its pKa values in the dimer- **Figure 4.20**, much lower values are present when comparing to its pKtyp and with the monomer.

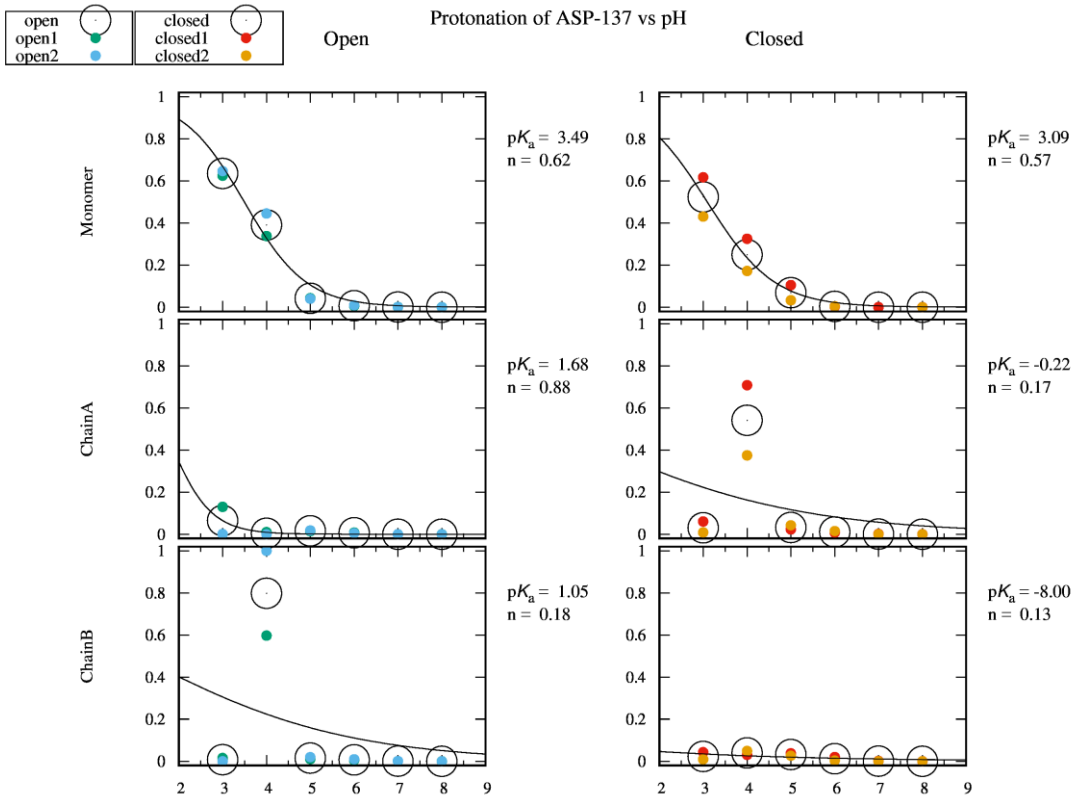


**Figure 4.20** Protonation of Asp137 considering the monomer and dimer, in the open and closed conformation, for each pH value.

It appears that in the dimer this residue is always charged, with exception at pH 4 of the Chain A in the closed dimer and chain B in the open one, where it has a tendency to be protonated. This rather atypical behaviour seems to be due to network of interactions already mentioned, but further analyses must be conducted.

The titrable site Glu108 appears to be partially covered by the N-terminal and has a high dispersion of points, when observing its titration curve. Two types of populations seem to be present, which could indicate that sometimes it is exposed while others internalized. When

exposed, it can be interacting with a positive charge in its proximity, explaining the low pKa value estimated in the PB/PC calculations.

At the interface periphery we can find the site His146. Its titration curve, has a high dispersion of the averaged protonation for the replicates. This dispersion could be related with its interaction with nearby residues or due to becoming buried upon dimerization. Eventually, the conformational alterations seen at pH 3, 4 and 5 may also affect its titration.

In the C-terminal region, some titrable sites with an atypical behaviour can be found. Glu157 and Glu158 are located after the small helix (153-156) and have slightly high pKa values, which may be due to the presence of several carboxylic acids in that region. Since this helix is known to have pH-dependent conformational changes (sub section **4.2.2**, Single-turn α-helix), their pKa alterations may be related to this, but further analyses are needed. His161 has a very high point dispersion in its titration curve- **Figure 4.21**, being protonated in some replicates, while deprotonated in others, suggesting that it is conformational trapped.

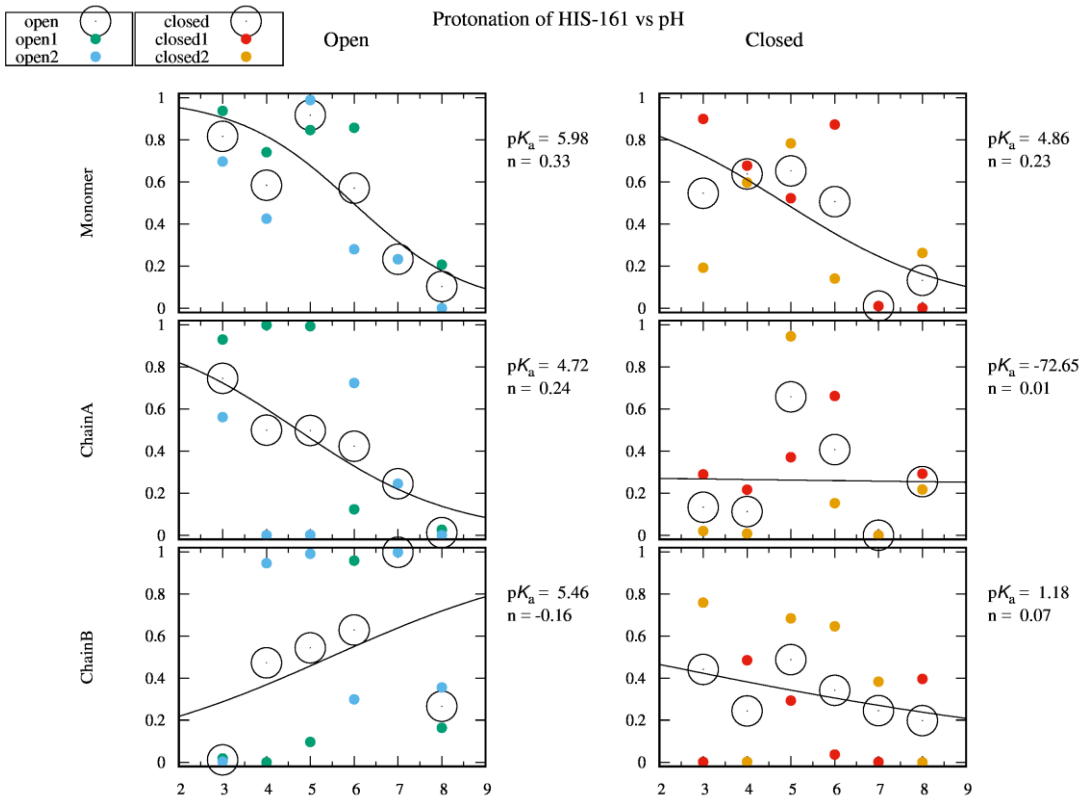


**Figure 4.21** Protonation of His161 considering the monomer and dimer, in the open and closed conformation, for each pH value.

In initial structures, it was observed that His161 was internalized in a hydrophobic area close to three phenolic rings, which can explain a trapped deprotonated state. In the Cter titrable site,

some bumps and a shoulder are present in the titration curve of the dimer, with a high dispersion of the replicates, which may again be related with transient internalization of the site.

In other residues, Nter and Asp33, shoulders were also observed in the titration curves, but, after visual analysis we found that these residues were too distant from each other, indicating that these changes are most probably not due to a coupling between those sites but rather to conformational alterations.

When comparing the $pK_a$ values obtained in PB/MC calculations with the ones obtained by CpHMD, we found some residues that had an abnormal $pK_a$ only in PB/MC calculations, most probably due to an artifact of the crystallographic structure. However, there were others where their $pK_a$ or protonation pattern became abnormal also (or only) in CpHMD simulations. In these cases, more investigation is necessary.

Further analyses in future work will be performed to better understand these changes.

### 4.2.5 Overall Discussion

By looking at the overall behaviour of the CpHMD results analysed so far in this section, both the monomer and dimer have a pH-dependent profile, appearing that the BLG structure is more stable as a dimer than as a monomer. The dimerization and its subsequent dissociation are well evidenced at pH values away from the pI (pHs 3, 7 and 8), as reported in experimental studies [5]. This is more pronounced at higher pHs, especially at pH 7, since the dissociation begins to occur shortly after starting the simulation. At this pH, this is observed in both the open and closed conformations, but at pH 8 only the open conformation seems to prefer the monomeric form, appearing that there is an additional factor stabilizing the closed conformation or that this structure needed more simulation time until starting to dissociate. Trying to compare this with experimental data is not feasible, since at that pH the closed conformation is less populated, as the closed-open transition occurs around 7.5 [18]. A clear transition was not observed in our timescale, probably occurring in a longer timescale.

Despite the dissociation of one replicate shown at pH 3, BLG dimer appears to be in a compact state at that pH, having a higher contact area between the two monomers, than at other pHs. Thus, BLG appears to have two states: one compact state at pHs 3 and 4 and a relaxed state at the other pHs. pH 3 and 4 have suffered conformational alterations in the βI strands and α -helices relative orientations, resulting in an approximation of the two helices that enables them to interact, directly or indirectly, which is not observed at the other pHs. This showed that, despite the already reported higher association and stability at pH 5 [7] this is not directly related with a higher contact

area. The results of the following section, obtained using the umbrella sampling method, will help elucidating this.

At pHs where the dissociation occurs it seems that there is also some instability in the secondary structure, some unfolding having been observed in a small helix close to the C-terminal and a more exposed coil in the N-terminal.

When comparing the open with the closed conformation it appears that there is not a clear trend in terms of dimerization and stability.

The protonation of several titrable sites seems to be associated with structural changes, either localized or related to the global compact/relaxed transition of the dimer. A better characterization of the interaction involved in those changes requires a more in-depth analysis of the simulation data, which will be pursued in future work.

## 4.3 Umbrella Sampling

As already seen in the previous section, the dimerization of BLG affects several properties at a molecular level. Thus, for a more detailed and controlled analysis of this association we decided that it would be interesting to use umbrella sampling (US) to study the closed conformation at pHs 3, 4 and 5, where interesting features were observed, including a compact/relaxed transition of the dimeric form.

In addition to making possible the calculation of a potential of mean force (PMF), the US data can be used to analyse several properties such as: the distance between the two monomers, the contact surface, the number of hydrogen bonds and the protonation. For the results where only the average behaviour is considered an equilibration time of 10 ns was taking into account.

The US simulations were performed for several target distances between the COM of the monomers (3.0, 3.1…4.0 nm) and a simulation time of 50 ns for each window. This should provide a reasonable sampling, with no strong interactions expected between the monomers beyond 4.0 nm.

Some of the simulations had a shorter runtime due to a problem with the box centering algorithm used, their simulation time being depicted in **Table 4.5,** for replicate 1 and **Table 4.6**, for replicate 2**.** Although that algorithm problem has been fixed meanwhile, it was not possible to continue these simulations in time for the delivery of this thesis.

**Table 4.5** Simulation time (ns) at different target distances for replicate 1.

| | **Distance** | **3.0** | **3.1** | **3.2** | **3.3** | **3.4** | **3.5** | **3.6** | **3.7** | **3.8** | **3.9** | **4.0** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation Time (ns)** | pH 3 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | **48** | **40** | 50 |
| | pH 4 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| | pH 5 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |

**Table 4.6** Simulation time (ns) at different target distances for replicate 2.

| | **Distance** | **3.0** | **3.1** | **3.2** | **3.3** | **3.4** | **3.5** | **3.6** | **3.7** | **3.8** | **3.9** | **4.0** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation Time (ns)** | pH 3 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | **48** | **48** | 50 |
| | pH 4 | 50 | 50 | 50 | 50 | **45** | 50 | 50 | **48** | 50 | 50 | **47** |
| | pH 5 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |

### 4.3.1 Distance between the two monomers

We started by analyzing the distances between the center of mass (COM) of the monomers sampled in each US window, in which a distance restraint force constant of 1000 kJ $mol^{-1}$ $nm^{-2}$ was used. The obtained results for both replicates at each pH value are depicted in **Figure 4.22**.

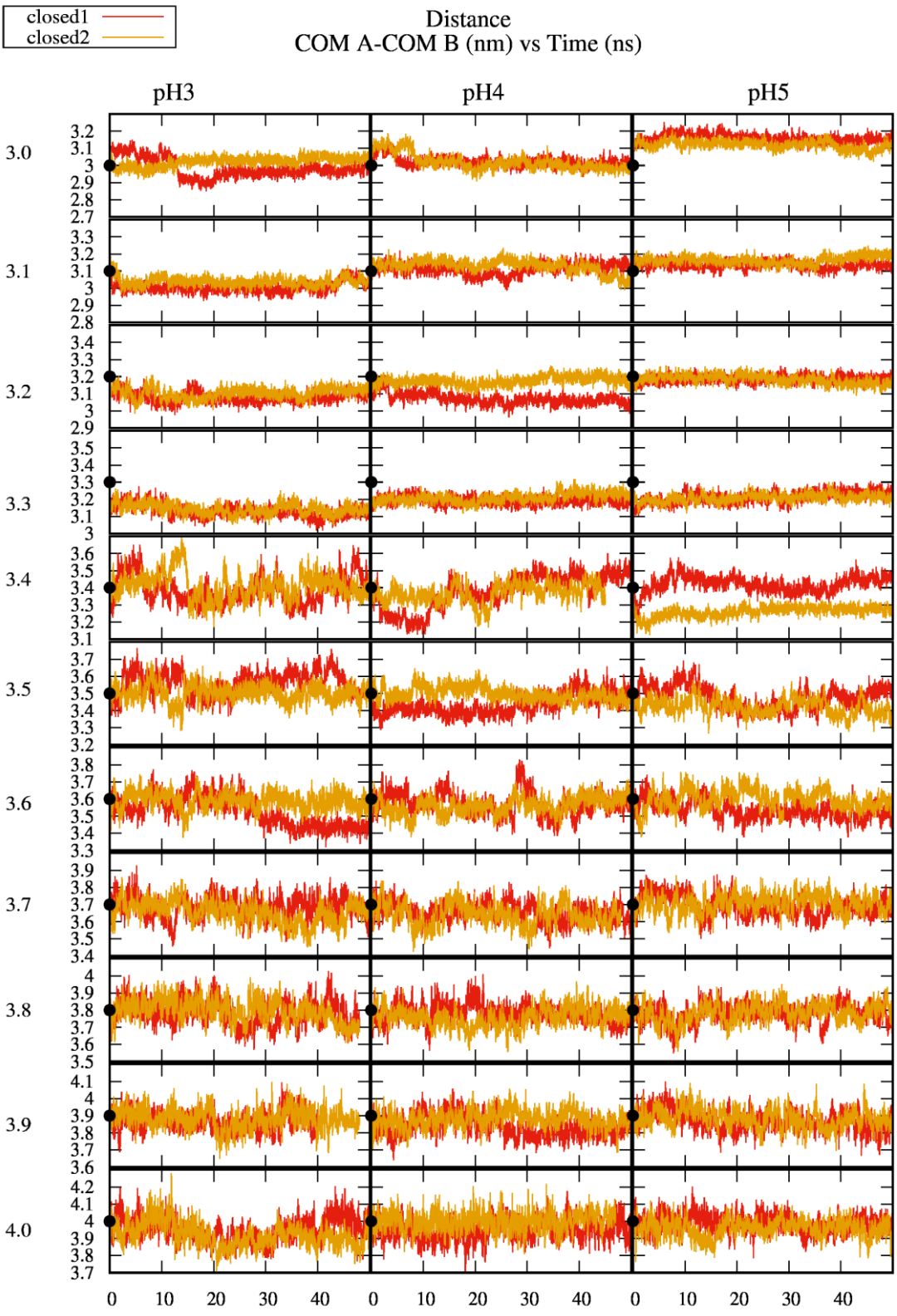


**Figure 4.22** Center of mass (COM) distance between the two monomers, considering both replicates for each pH value and US window. Each US window is labeled with the value of its target distance, also shown as a black point (t=0).

By looking at the overall picture, the first 10 ns are considered to be their equilibration time, and so only the results after this time are used for computing averages.

Considering all the target distances at which the two monomers were placed, lower fluctuations are observed for shorter distances. Since these fluctuations are directly related with the force applied to keep the monomers at that distance and with the intrinsic force between the monomers, it appears that, at those distances, the intrinsic force is larger than the restraint force imposed, pulling the monomers closer or pushing them further away. For greater target distances this effect is not observed, which means that the intrinsic force is weaker and less significant. Overall, the amplitude of these fluctuations does not exceed 0.5 nm, being even lower at low target distances.

At pH 3, from windows 3.0 to 3.3 nm, the attractive forces between the two monomers are stronger than the force applied, such that the dimer tends to be closer than the distance imposed. It is just from 3.4 nm on, that the force is sufficient to maintain them at the target distance, with exception of one replicate of the window 3.6, which remains at a closer distance over the simulation time.

For the pH 4 profile, we can notice some variations. When placed at a distance of 3.0 or 3.1 nm, the force applied is enough to keep the monomers at that distance. However, when increasing the target distance to 3.2 or 3.3 nm, they tend to attract each other, showing that the intrinsic interaction is stronger than the force imposed. From 3.4 nm on, the imposed force seems enough to keep the monomers at the target distance.

At pH 5, a clear repulsion is present at the target distance 3.0 nm, unlike anything observed at the other two pHs. When increasing the target distance up to 3.3 nm, an actual COM distance from 3.1 to 3.2 nm continues to be preferred. At 3.4 nm it seems that the replicate 2 has attractive forces higher than the applied force.

The overall behaviour at three pHs shows that, when at a very low target distance (3.0-3.2 nm), there is an intrinsic attractive force at pH 3, that is less present at pH 4 and even becomes repulsive at pH 5. Indeed, this observation is strongly correlated with the results obtained in the previous section (subsection **4.2.3**, contact surface), where it seems that at pH 3 the dimer is in a compact state, with a higher contact surface between the two monomers, as opposed to pH 5, where a relaxed state is found. pH 4 presents a combination of these two states.

### 4.3.2 Contact surface

The averaged results for the contact surface at each different target distance are presented in **Figure 4.23**. It should be noted that the averages were computed for each US window and are assigned to their corresponding target distance, although as seen in the previous section, the actual sampled distances fluctuate around and may deviate from the target ones.

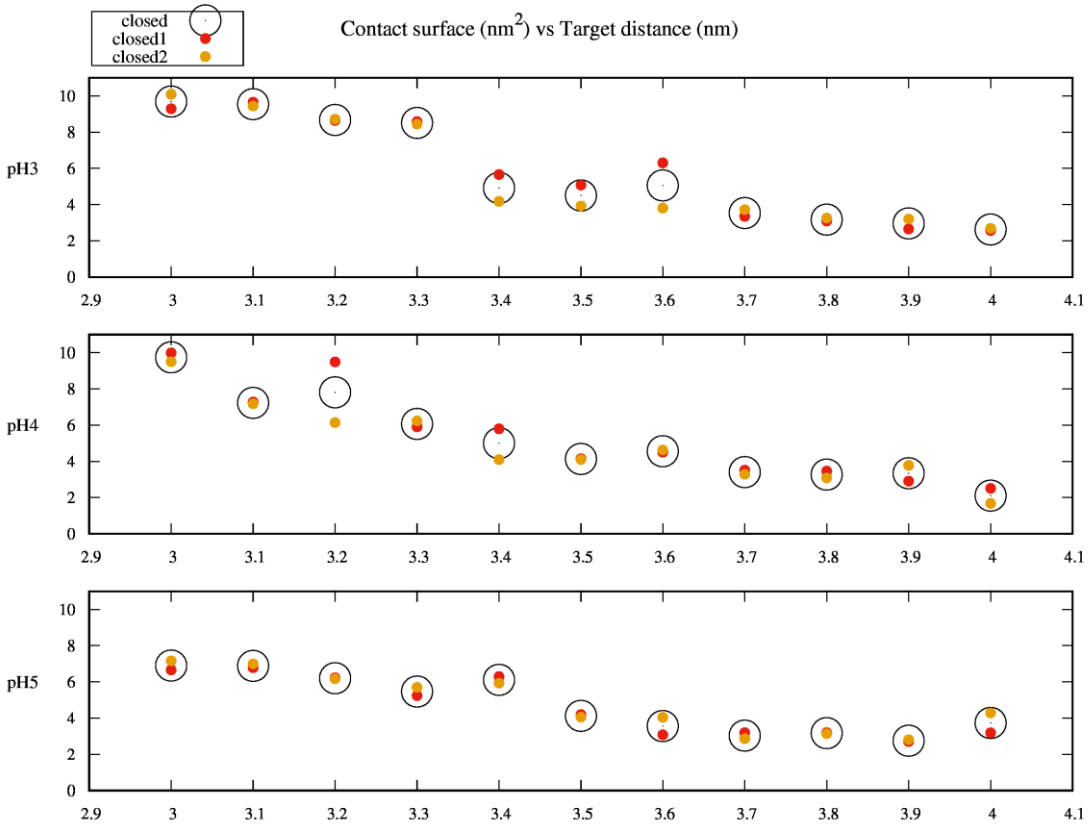


**Figure 4.23** Mean contact surface between the two monomers, as function of their target distance, considering both replicates. Each replicate is shown with a filled smaller circle and their average is represented with a larger empty circle.

At first sight, the results obtained for the contact surface at the three pH values are as expected. As the target distance between the two monomers increases, their contact surface consequently decreases. However, a more detailed analysis shows that, at pH 3, from 3.0 to 3.3 nm, the dimer has a higher contact surface, with a more pronounced decrease starting from 3.4 nm, when comparing with the other pHs. The trend previously shown in the distance analysis, from a compact state to a relaxed one, with a combination of both at pH 4, is also present here, being in fact, at pH 5 that the contact surface is lower for smaller distances, reflecting the fact that the actually populated distances are a bit higher, as seen in the previous subsection (see **Figure 4.22**).

For pH 3, at 3.6 nm, one replicate had a higher contact area than the overall behaviour, but as seen in the previous results of the distance analysis, this replicate deviates from its target distance. The same seems to happen for pH 4 at a 3.2 nm distance.

### 4.3.3 Hydrogen bonds

The number of main-chain hydrogen bonds was calculated, their averaged results for each target distance being depicted in **Figure 4.24**.

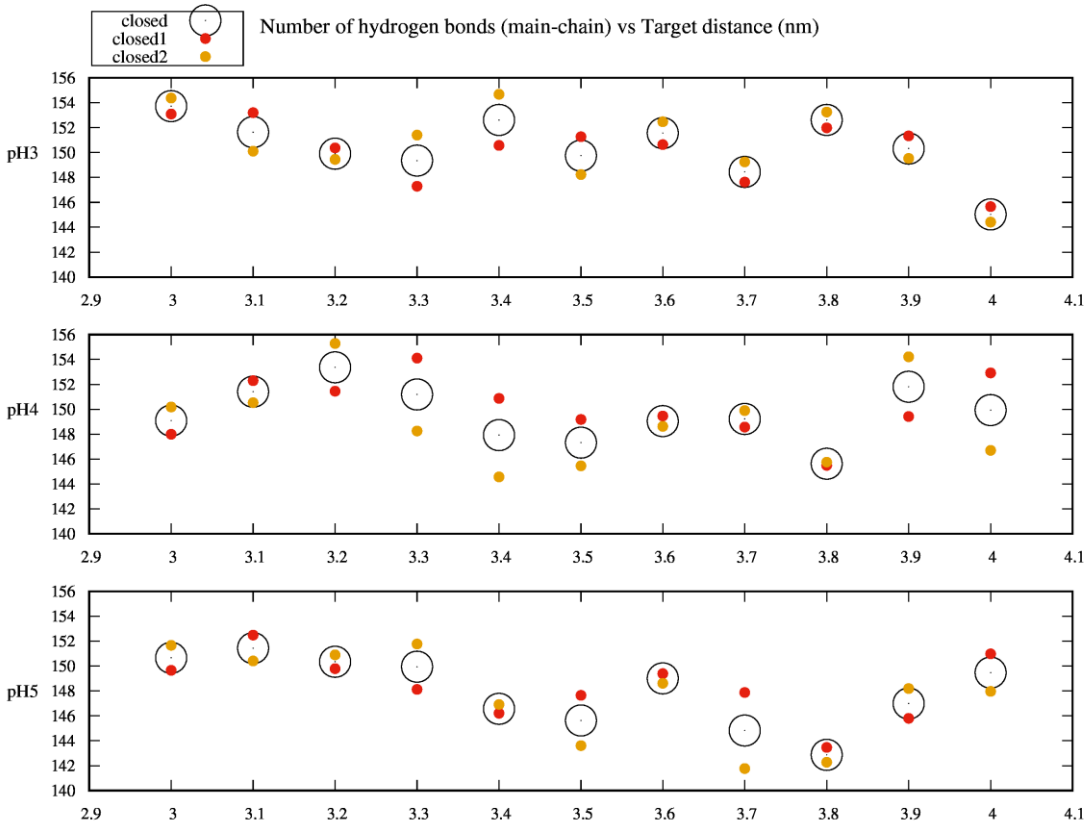


**Figure 4.24** Mean number of main-chain hydrogen bonds for each target distance and pH. Each replicate is shown with a filled smaller circle and their average is represented with a larger empty circle.

No clear trend is observed for any of the pH values, suggesting no loss or gain of backbone structure can be correlated with the distance between the monomers. However, the dispersion of values between the two replicates and the fact that the actually populated distances do not, in some cases, correspond to the target ones (see **Figure 4.22**- subsection distance between the monomers), makes it difficult to reach any conclusion.

### 4.3.4 Potential of mean force (PMF)

To understand how "strong" is the interaction between the two monomers, the free energy of the system was measured by means of a potential of mean force (PMF) as shown in **Figure 4.25**. Here, the effect of the applied biasing force has been removed, that is, the PMF value at a given distance measures how likely it is to find the two BLG monomers at that distance if no force were applied.

The reference probability value (see section **3.8**- Analyses) chosen for all PMF calculations was the maximum of the three probability densities. So, the alignment of the three different PMFs is made by this probability maximum (PMF=0).

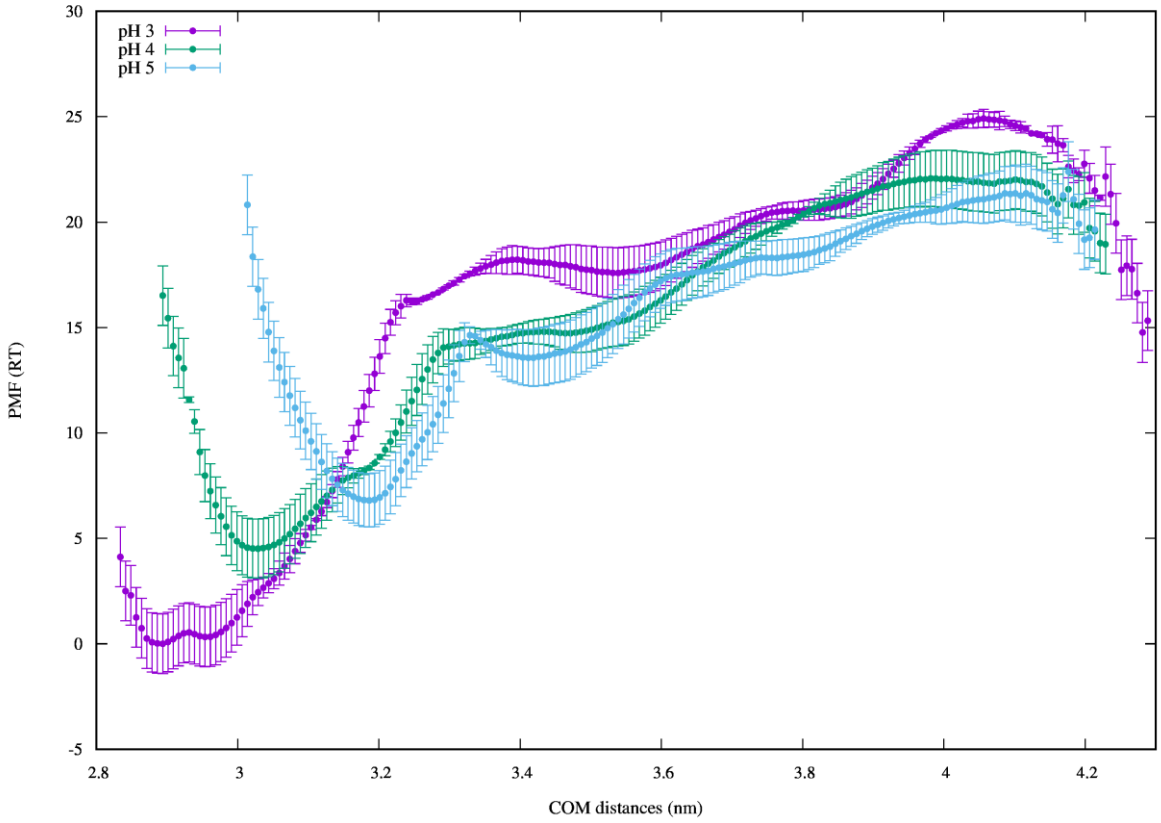


**Figure 4.25** PMF as a function of the distance between the two monomers COM, considering pHs 3, 4 and 5, showing error bars computed as described in section 3.8- Analyses.

Before evaluating the individual behaviour at each pH, a brief description of these type of plots follows. The depth of the well represents the maximum of the probability to find the two monomers at a certain distance, i.e. the lower the bottom of the well the more favoured is the interaction. As the distance increases, so does the free energy, until reaching a baseline where no more variations in this energy are observed, meaning that the two monomers no longer "feel" each other.

In our results, the PMF profile is quite different for each pH value. At pH 3, we can see a deeper well and located at lower distances than at pH 4 and 5, showing more favoured interactions at those distances. Indeed, this well becomes less deep as the pH increases and it shifts to longer distances. These observations are related to what we previously found in the unbiased CpHMD simulations. At pH 3, the dimer is found to be in a compact state and, in fact, it is at that pH that a favoured interaction is observed at very low distances (2.8-3.0). With an increase in pH the dimer clearly prefers to be at higher distances, especially at pH 5, where BLG is found to be in a relaxed state. And once again, pH 4 appears to have an intermediate behaviour that reflects a combination of the structural forms present at the two other pHs.

As the distance increases, we would expect to obtain a baseline with no changes in the free energy but, unfortunately, this was not observed and the sampling of longer distances seems to be necessary. Despite that, we verified that our sampling is satisfactory for temporal convergence (**Appendix I**, section **A5**) and that, for the replicates with shorter simulation times, this restriction was not significant. Nonetheless, there are some differences between the replicates, suggesting that more of them should probably be used.

Mercadante et al. [7] experimentally determined the rate constants, $k_{off}$ and $k_{on}$, for BLG dimer dissociation. Their obtained $k_{off}$ values show that BLG has a very low dissociation constant at ~pH 3 (0.008) and higher $k_{off}$ (> 0.1) at higher pHs. This is in agreement with our results, being at pH 3 that BLG has a stronger interaction, seen by the depth of the well, as opposed to the other two pHs, where a smaller difference is observed between the minimum and the baseline values. As for the $k_{on}$ values, it is not possible to relate them with our results, since more sampling with monomers in different orientations would be necessary to capture the possible approximation of the two monomers from different angles. Ideally, a perfect sampling would allow us to compute these two rate constants and the dimerization equilibrium constant.

### 4.3.5 Protonation

Protein net charge

The protein net charge for the different target distances can be seen in **Figure 4.26**.

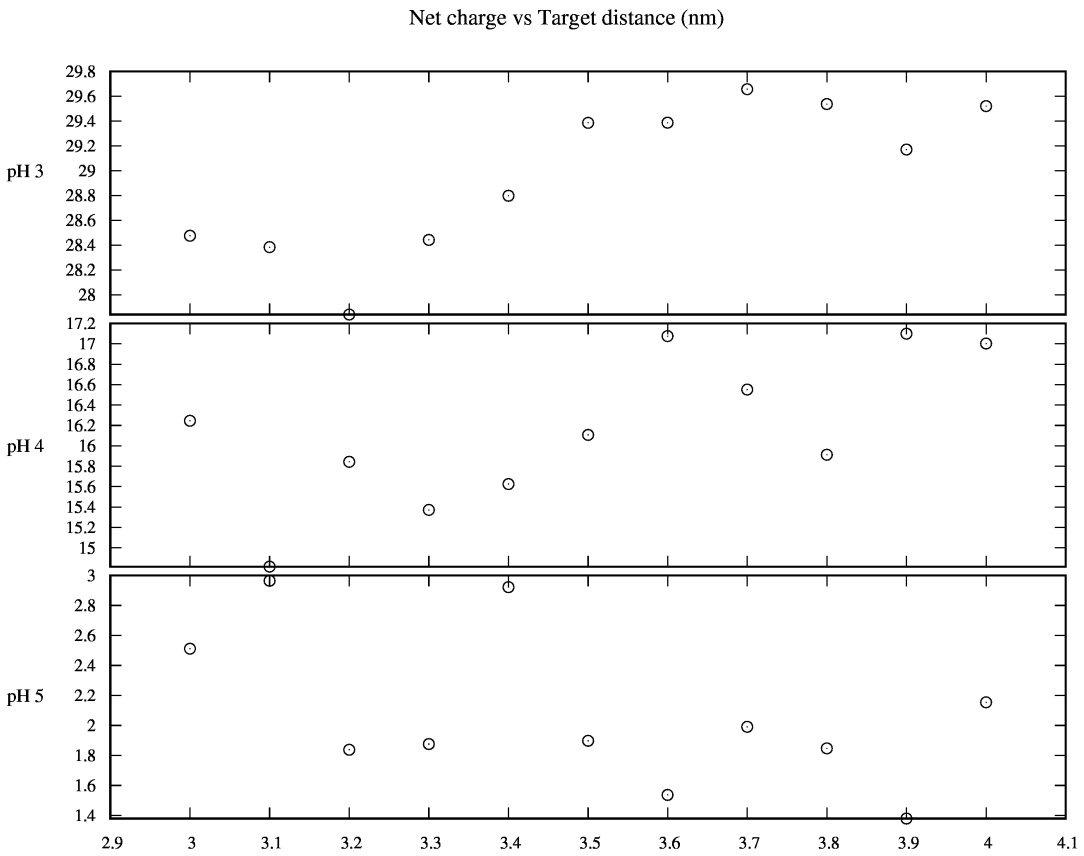


**Figure 4.26** Mean protein net charge as function of the target distance, considering the average (empty circle) of the two replicates, at pHs 3, 4 and 5.

By analyzing the mean protein net charge for the three pHs, we can see that as the target distance increases at pHs 3 and 4, the total charge also tends to increase. This trend is more pronounced at pH 3, where the protein is more positively charged. At pH 5, near the isoelectric point, it appears that the changes in protonation are not directly related with the monomers distance. The decrease in charge as the monomers approximate, at pHs 3 and 4, is probably a direct consequence of the overall charge repulsion, with each positive monomer promoting the dissociation of protons on the other one.

Results of the protein net charge of each individual chain of the dimer were also observed, but no trend was established.

These analyses were performed on a per window basis, with the averages assigned to the corresponding target distances, while in fact, as seen previously, the actually populated distances

are sometimes different. Therefore, for doing a more detailed analysis, a correct attribution of the distances is necessary; unfortunately, this will not be done in time for this thesis.

Protonation of individual groups

For most of the titrable sites, the averaged protonations show no clear trend with the target distances, therefore only the ones with a visible trend will be presented in this section.

As previously mentioned, the titrable site Asp137 is found in the α -helix and is thought to form a salt bridge with Arg148 or an H-bond with its corresponding residue in the other chain. Its average protonation for the different distance windows is presented in **Figure 4.27**.

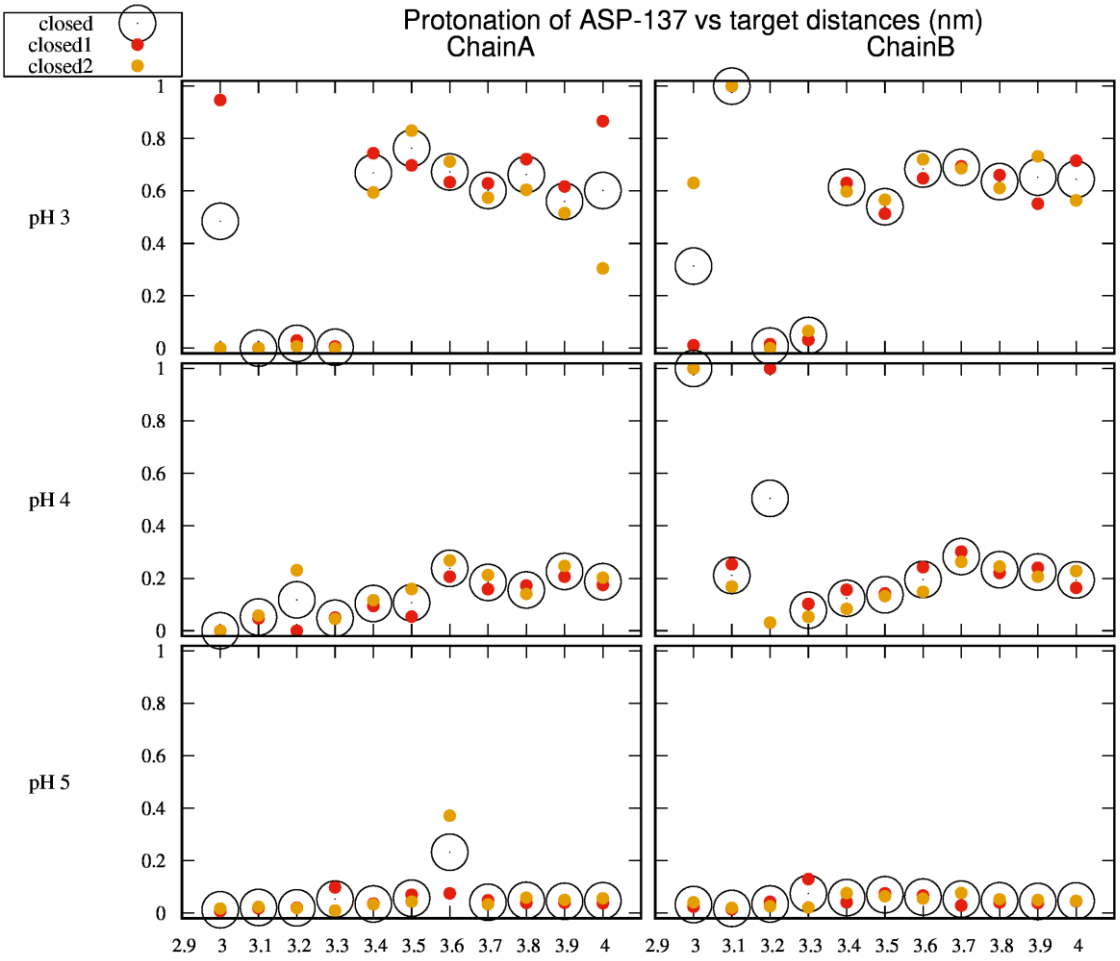


**Figure 4.27** Protonation of Asp137 for each target distance, at pHs 3, 4 and 5. Each replicate is shown with a filled smaller circle and its average is represented with a larger empty circle.

For the results at pH 3, a clear trend is observed. At lower target distances, Asp137 tends to be deprotonated but, as the distances increases from 3.4 to 4.0, this residue tends to become partially protonated. This protonation dependence with the distance is also present at pH 4, but with much

smaller variations. At pH 5, no trend seems to be present. This can be explained by the interaction with Arg148, which tends to induce the ionized form of Asp137 at close distances. At higher distances the protonation is the one expected at those pHs.

The residue His146, located at the interface, also appears to have a pH-dependent trend, its protonation profile being shown in **Figure 4.28**.

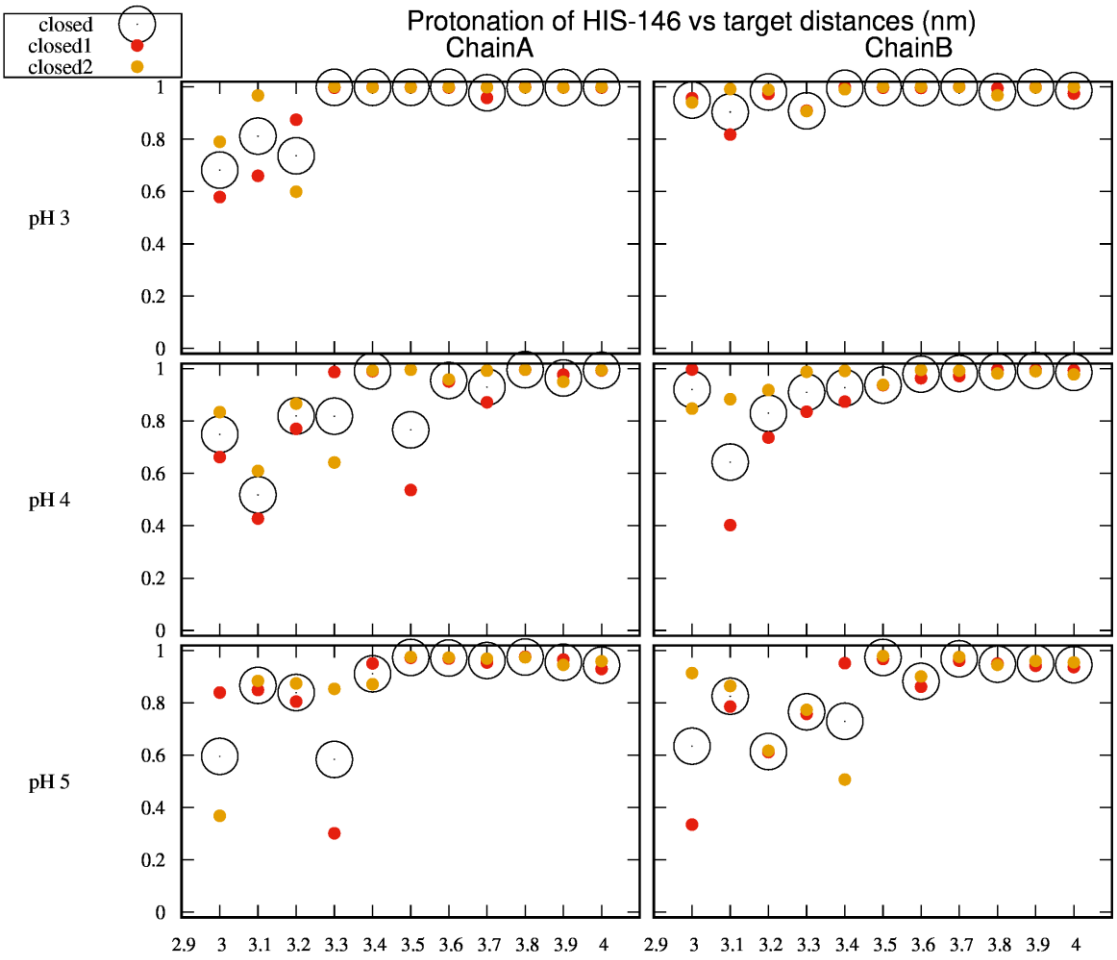


**Figure 4.28** Protonation of His146 for each target distance, at pHs 3, 4 and 5. Each replicate is shown with a filled smaller circle and its average is represented with a larger empty circle.

In these protonation plots, despite some dispersion of replicate points, an increase in protonation is apparent as the distance between the monomers increases. This is observed for all pHs with some slight differences but, overall, at a distance from approximately 3.3 until 4.0 this residue becomes protonated. His146 is located close to the interface, and the observed trend may result either from burying upon dimerization or from interaction with nearby residues, as already mentioned in section **4.2.4**- Protonation.

We expect more insightful results when regrouping these distances into their “real” distances.

### 4.3.6 Overall discussion

Taken as a whole, the results for the Umbrella Sampling section reinforce our previous conclusions regarding the observation of two dimeric states: compact and relaxed. At pH 3, stronger interactions and at lower distances are present between the two monomers, characteristic of a compact state. This contrasts with pH 5, where repulsion occurs at these small distances and a favoured interaction is only seen for larger distances, corresponding to a relaxed state. Again, the behaviour at pH 4 agrees with the fact that its structure is a combination of the compact and relaxed states. These conclusions are in good agreement with the $k_{off}$ values of Mercadante et al [7], where a very low rate dissociation is observed at pH 3, as opposed to the higher rates reported for higher pHs; as pointed out in the Introduction, the dimerization equilibrium constant follows a different trend, reaching its peak near the isoelectric point. Unfortunately, due to the lack of replicates with different orientations, the $k_{on}$ values were not possible to infer and, thus, the dimerization equilibrium constant, $K_D$ could not be obtained.

Despite the sampling being enough for temporal convergence, the lack of rotational sampling and at larger distances between the monomers does not allow a deeper discussion, but reveals the importance of showing these effects, as they will surely have a pH-dependent behaviour with interesting features.

The regrouping of the computed properties (contact surface, H-bonds, protonation, etc...) according to their actual distances is a priority, since it will likely reveal currently hidden trends.

# 5 PROPOSED MUTATIONS

As discussed in the Introduction, the well-known allergenicity of BLG in susceptible individuals after milk ingestion is thought to be due to its dimerization. Thus, including site-directed mutations that alter BLG structure and favors the monomeric form can be a good option for an immunotherapy, since lower amounts of dimers will be present in the human body and, thus, the mutant could work as a hypoallergen [30].

Our observations along this work are in agreement with the already proposed mutations [30][34], that mostly focus in modifying some residues that are located in the two βI strands and AB loops at the interface, in order to destabilize the dimeric form.

Here, we propose additional mutations in two residues that are also located at the interface, since destabilizing the interactions in this region enhances the monomeric form [21]. Asp137 is thought to interact with Arg148 or with its corresponding residue in the other chain, so, modifying this residue into an Asn would disrupt the salt bridge and/or the H-bond formed. The choice of this new residue is mainly due to its similarity in structure, with a polar group that it is not charged. Another option could be a point mutation in Arg148 to Gln, for the same reasons as Asp137, with Gln being chosen because, like Arg148, it has a long chain but no net charge.

# 6 CONCLUSION

At the start of this work, we knew from experimental data that BLG had a monomer-dimer equilibrium, related to allergenicity, and a gating mechanism involving the EF loop associated to ligand binding, both pH-dependent. However, very little insight about the molecular phenomena underlying these observations existed.

Our work was innovative by using a computational molecular approach that takes into account the direct effect of pH, by combining different protonation states with changes in BLG conformation, thus allowing us to understand the pH-sensitive behaviour at an atomic detail.

The principal conclusion that we can draw from this work, is that pH effects are a contributing factor for BLG dimerization, but the pH-dependence of this monomer-dimer equilibrium is far from straightforward. The dissociation is indeed more noted at pH values away from the pI, as already reported in experimental data [5], but this is most observed at pH 7, and not as much at pH 3 and 8; this may reflect a higher dissociation rate constant, $k_{off}$, at pH 7, although this has not been captured by the (low) precision of the existing experimental data [7]. In addition, at pHs where BLG appears to be in stable association (pHs closer to the pI), this does not necessarily mean that they have a low $k_{off}$, as observed by Mercadante et al. [7] and corroborated by our umbrella sampling simulations. Indeed, conformational alterations in the βI strands and α -helices showed us that BLG structure presents two different states:

- A **compact** state, that resulted from the tight association of the two monomers, observed at pH 3;
- A **relaxed** state, with a less tight interface, similar to what is found in crystallographic structures, observed for pH ≥ 5.

An apparent combination of these two states is observed at pH 4. Although some authors have already noticed some changes in this interface region [16][11][9], the atomic details underlying these changes were not known. By analysis of the distance-dependent potential of mean force, we have found that these states are highly related to how "tight" is the interaction between the two monomers. At pH 3, a deep well at small distances (compact state) corresponds to low $k_{off}$ values, while at pH 5, a shallower well at higher distances (relaxed state) corresponds to higher $k_{off}$ values. However, as already stated, the dimer is more favoured at pH 5, since the equilibrium depends not only on the $k_{off}$ value, but also on the association rate constant, $k_{on}$. Unfortunately, a relation of our results with $k_{on}$ is more difficult due to insufficient rotational sampling.

Also, by comparing the monomeric with the dimeric form, we can conclude that the dimer BLG structure is more stabilized and that its dissociation is accompanied by a slight loss of the secondary structure.

The Tanford transition associated to the EF loop was not observed, and probably occurs at a longer timescale.

Since the dimer is suspected to be allergenic, we propose a mutation in Asp137 to Asn and in Arg148 to Gln, with the goal of reducing the dimer population, and thus creating a hypoallergen for immunotherapy.

Last but not least, it should be noted that the monomer-dimer equilibrium of BLG is highly associated with its concentration (see **Figure 1.4** in Introduction) and the solution ionic force, and that several of the experimental articles that we found do not seem to take this factor into account or be aware of it, referring their observations to the wrong form of BLG (monomer vs dimer), which consequently affects further conclusions.

# 7 FUTURE PERSPECTIVES

In order to fully understand the biological impact that BLG protein has in the human body more structural details with a pH-dependence must be studied. For that, additional analyses on our trajectories can be done (some of them were already mentioned along this thesis):

- Calculate the electrostatic potential (isopotential contours, surface mapping, etc), in order to gain insight into the association process.
- Characterize the EF loop in detail, not only to observe the overall Tanford transition, but also minor movements of this loop that could be significant in the binding of ligands;
- Analyse the relative rotation of the two monomers;
- Analyse in more detail the titrable residues with atypical $pK_a$ values and/or titration curves, to better understand the relation of their protonation with conformational alterations.
- Regroup the computed properties in US according to their actual distances;

An improvement of the current simulation data can also be obtained by:

- Fixing some of the missing simulation data: complete unfinished simulations, rerun problematic ones (with grid-centering problems), etc.
- Increasing CpHMD simulation time, since it appears that at pH 8 the closed dimer structure has not yet stabilized.
- Using more replicates and more US simulation time, with longer distances, to better estimate the $k_{off}$ of the dimer.
- Improving US sampling with monomers in different orientations, to capture the dimer association from different angles, and thus estimate the $k_{on}$;

New studies can also be performed:

- Repeat unbiased CpHMD and US simulations at different ionic strengths;
- Perform US only for the EF loop, in order to overcome its slow conformational transition;
- For a better understanding of the transient compact/relaxed transition of the dimer, it will be interesting to do experimental studies using circular dichroism (CD) or optical rotatory dispersion (ORD) at a low pH and with a typical ionic strength (~0.1 M), since experimental studies at very low pH, seem to use very low ionic strength probably to favor the monomeric form.

These new analyses and studies will not only be relevant for trying to better understand the biological problem, but also to answer some of the questions left open by the work developed during this master thesis.

# 8 BIBLIOGRAPHY


[1] D. KB, R. Olson, and C. KO, "Dietary guidelines for americans," *JAMA*, vol. 315, no. 5, pp. 457–458, Feb. 2016.

[2] S. Harjinder, M. Boland, and A. Thompson, *Milk Proteins*. 2014.

[3] G. Del Val *et al.*, "Thioredoxin treatment increases digestibility and lowers allergenicity of milk," *J. Allergy Clin. Immunol.*, vol. 103, no. 4, pp. 690–697, 1999.

[4] R. D. J. Huby, "Why Are Some Proteins Allergens?," *Toxicol. Sci.*, vol. 55, no. 2, pp. 235–246, 2000.

[5] G. Kontopidis, C. Holt, and L. Sawyer, "Invited Review: β-Lactoglobulin: Binding Properties, Structure, and Function," *J. Dairy Sci.*, vol. 87, no. 4, pp. 785–796, 2004.

[6] P. Martin, M. Szymanowska, L. Zwierzchowiski, and C. Leroux, "The impact of genetic polymorphisms on the protein composition of ruminant milks," *Reprod. Nutr. Dev.*, vol. 44, no. 6, pp. 433–459, 2002.

[7] D. Mercadante *et al.*, "Bovine β-lactoglobulin is dimeric under imitative physiological conditions: Dissociation equilibrium and rate constants over the pH range of 2.5-7.5," *Biophys. J.*, vol. 103, no. 2, pp. 303–312, 2012.

[8] G. Robitaille, M. Britten, J. Morisset, and D. Petitclerc, "Quantitative analysis of beta-lactoglobulin A and B genetic variants in milk of cows beta-lactoglobulin AB throughout lactation," *J. Dairy Res.*, vol. 69, no. September 1998, pp. 651–654, 2002.

[9] F. Fogolari, L. Ragona, L. Zetta, S. Romagnoli, K. G. De Kruif, and H. Molinari, "Monomeric bovine β-lactoglobulin adopts a β-barrel fold at pH 2," *FEBS Lett.*, vol. 436, no. 2, pp. 149–154, 1998.

[10] L. Ragona, F. Pusterla, L. Zetta, H. L. Monaco, and H. Molinari, "Identification of a conserved Hydrophobie cluster in partially folded bovine β-lactoglobulin at pH 2," *Fold. Des.*, vol. 2, no. 5, pp. 281–290, 1997.

[11] S. Uhrínová, M. H. Smith, G. B. Jameson, D. Uhrín, L. Sawyer, and P. N. Barlow, "Structural changes accompanying pH-induced dissociation of the beta-lactoglobulin dimer.," *Biochemistry*, vol. 39, no. 13, pp. 3565–3574, 2000.

[12] K. Kuwata, M. Hoshino, V. Forge, S. Era, C. A. Batt, and Y. Goto, "Solution structure and dynamics of bovine beta-lactoglobulin A.," *Protein Sci.*, vol. 8, no. 11, pp. 2541–5, 1999.

[13] B. Y. Qin, M. C. Bewley, L. K. Creamer, H. M. Baker, E. N. Baker, and G. B. Jameson, "Structural basis of the tanford transition of bovine β-lactoglobulin," *Biochemistry*, vol. 37, no. 40, pp. 14014–14023, 1998.

[14] B. Y. Qin, G. B. Jameson, M. C. Bewley, E. N. Baker, and L. K. Creamer, "Functional implications of structural differences between variants A and B of bovine β-lactoglobulin," *Protein Sci.*, vol. 8, no. 1, pp. 75–83, 2008.

[15] K. M. G. Oliveira, V. L. Valente-Mesquita, M. M. Botelho, L. Sawyer, S. T. Ferreira, and I. Polikarpov, "Crystal structures of bovine β-lactoglobulin in the orthorhombic space group C2221: Structural differences between genetic variants A and B and features of the Tanford transition," *Eur. J. Biochem.*, vol. 268, no. 2, pp. 477–483, 2001.

[16] S. Brownlow *et al.*, "Bovine β-lactoglobulin at 1.8 Å resolution — still an enigmatic lipocalin," *Structure*, vol. 5, no. 4, pp. 481–495, 1997.

[17] G. Graziano, “Role of hydrophobic effect in the salt-induced dimerization of bovine β-lactoglobulin at pH 3,” *Biopolym. - Pept. Sci. Sect.*, vol. 91, no. 12, pp. 1182–1188, 2009.

[18] C. Tanford, L. G. Bunville, and Y. Nozaki, “The Reversible Transformation of β-Lactoglobulin at pH 7.5,” *J. Am. Chem. Soc.*, vol. 81, no. 15, pp. 4032–4036, 1959.

[19] K. Sakai, K. Sakurai, M. Sakai, M. Hoshino, and Y. Goto, “Conformation and stability of thiol-modified bovine beta-lactoglobulin.,” *Protein Sci.*, vol. 9, no. 9, pp. 1719–1729, 2000.

[20] M. A. M. Hoffmann and P. J. J. M. Van Mil, “Heat-Induced Aggregation of β-Lactoglobulin: Role of the Free Thiol Group and Disulfide Bonds,” *J. Agric. Food Chem.*, vol. 45, no. 8, pp. 2942–2948, 1997.

[21] K. Sakurai, M. Oobatake, and Y. Goto, “Salt-dependent monomer – dimer equilibrium of bovine b-lactoglobulin at pH 3,” *Protein Sci.*, vol. 10, pp. 2325–2335, 2001.

[22] C. Tanford, “Ionization-linked Changes in Protein Conformation. I. Theory,” *J. Am. Chem. Soc.*, vol. 83, no. 7, pp. 1628–1634, 1961.

[23] P. R. Majhi, R. R. Ganta, R. P. Vanam, E. Seyrek, K. Giger, and P. L. Dubin, “Electrostatically Driven Protein Aggregation : Low Ionic Strength -Lactoglobulin at,” *Langmuir*, no. 6, pp. 9150–9159, 2006.

[24] R. Townend, C. A. Kiddy, and S. N. Timasheff, “Molecular interactions in b-lactoglobulin. VII. The hybridization of b-lactoglobulins A and B,” *J. Am. Chem. Soc.*, vol. 83, no. 1958, pp. 1419–1423, 1961.

[25] G. Baldini *et al.*, “Salt-induced association of β-lactoglobulin by light and X-ray scattering,” *Macromolecules*, vol. 32, no. 19, pp. 6128–6138, 1999.

[26] D. Renard, J. Lefebvre, M. C. A. Griffin, and W. G. Griffin, “Effects of pH and salt environment on the association of β-lactoglobulin revealed by intrinsic fluorescence studies,” *Int. J. Biol. Macromol.*, vol. 22, no. 1, pp. 41–49, 1998.

[27] F. Fogolari *et al.*, “Electrostatic properties of bovine beta-lactoglobulin.,” *Proteins*, vol. 39, no. 4, pp. 317–330, 2000.

[28] Y. Nozaki, L. G. Bunville, and C. Tanford, “Hydrogen-ion titration curves of β-lactoglobulin,” *J. Am. Chem. Soc.*, vol. 81, no. 5, pp. 5523–9, 1959.

[29] X. Lin, C. Holt, D. Mcnulty, D. Clarke, S. Brownlow, and G. Jones, “Effect of temperature on the secondary structure of β-lactoglobulin at pH 6.7, as determined by CD and IR spectroscopy: a test of the molten globule hypothesis,” *Biochem. J.*, vol. 324, pp. 341–346, 1997.

[30] J. Rouvinen1 *et al.*, “Transient Dimers of Allergens,” *PLoS One*, vol. 5, no. 2, pp. 1–9, 2010.

[31] M. Niemi *et al.*, “Molecular Interactions between a Recombinant IgE Antibody and the β-Lactoglobulin Allergen,” *Structure*, vol. 15, no. 11, pp. 1413–1421, 2007.

[32] J. B. Dressman *et al.*, “Upper gastrointestinal (GI) pH in young, healthy men and women,” *Pharmaceutical Research*, vol. 7, no. 7. pp. 756–761, 1990.

[33] A. J. Rugg-Gunn, G. J. Roberts, and W. G. Wright, “Effect of human milk on plaque pH in situ and enamel dissolution in vitro compared with bovine milk, lactose, and sucrose,” *Caries Res.*, vol. 19, no. 4, pp. 327–334, 1985.

[34] K. Sakurai and Y. Goto, “Manipulating monomer-dimer equilibrium of bovine β-lactoglobulin by amino acid substitution,” *J. Biol. Chem.*, vol. 277, no. 28, pp. 25735–25740, 2002.

[35] M. Yagi, K. Sakurai, C. Kalidas, C. A. Batt, and Y. Goto, “Reversible Unfolding of Bovine

β-Lactoglobulin Mutants without a Free Thiol Group," *J. Biol. Chem.*, vol. 278, no. 47, pp. 47009–47015, 2003.

[36] D. Jaya *et al.*, "A recombinant C121S mutant of bovine β-lactoglobulin is more susceptible to peptic digestion and to denaturation by reducing agents and heating," *Biochemistry*, vol. 43, no. 20, pp. 6312–6321, 2004.

[37] M. Bello, M. D. C. Portillo-Téllez, and E. García-Hernández, "Energetics of ligand recognition and self-association of bovine β-lactoglobulin: Differences between variants A and B," *Biochemistry*, vol. 50, no. 1, pp. 151–161, 2011.

[38] 3 Martiniano Bello, 1 Gerardo Pe´rez-Herna´ndez, 2 D. Alejandro Ferna´ndez-Velasco and 1 and Enrique Garcı´a-Herna´ndez1* Roberto Arreguı´n-Espinosa, "Energetics of protein homodimerization: Effects of water sequestering on the formation of b-lactoglobulin dimer," *Proteins*, vol. 70, no. 2, pp. 311–319, 2008.

[39] M. W. Oran M. Becker, Alexander D. MacKerell, Jr. , Benoît Roux, *Computational Biochemistry and Biophysics*. 2001.

[40] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, vol. 21. 2010.

[41] B. Wilfred, V. Gunsteren, and C. Berendsen, "Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry," *Anger,. Chem. lnr. Ed. Engl*, vol. 29, no. Md, pp. 992–1023, 1990.

[42] D. Frenkel and B. Smit, *Understanding Molecular Simulation*. 2002.

[43] A. H. T. Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, vol. 21, no. 6, 1953.

[44] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*. 1987.

[45] A. Hinchliffe, *Molecular Modelling for Beginners*. 2008.

[46] Andrew R. Leach, "Molecular Modelling." 2001.

[47] P. J. Martel, A. Baptista, and S. B. Petersen, "Protein Electrostatics," *Biotechnol. Annu. Rev.*, vol. 2, no. C, pp. 315–372, 1996.

[48] F. Fogolari, A. Brigo, and H. Molinari, "The Poisson-Boltzmann equation for biomolecular electrostatics: A tool for structural biology," *J. Mol. Recognit.*, vol. 15, no. 6, pp. 377–392, 2002.

[49] M. M. Cox and D. L. Nelson, *Lenninger- The principles of Biochemistry*. 2008.

[50] A. Warshel, F. Sussman, and G. King, "Free Energy of Charges in Solvated Proteins: Microscopic Calculations Using a Reversible Charging Process," *Biochemistry*, vol. 25, no. 26, pp. 8368–8372, 1986.

[51] A. Warshel, "Calculations of chemical processes in solutions," *J. Phys. Chem.*, vol. 83, no. 12, pp. 1640–1652, 1979.

[52] J. G. Kirkwood, "Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions*," *J. Chem. Phys.*, vol. 2, pp. 251–361, 1934.

[53] Y. Y. Sham, Z. T. Chu, and A. Warshel, "Consistent Calculations of p $K_a$ 's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches," *J. Phys. Chem. B*, vol. 101, no. 22, pp. 4458–4472, 1997.

[54] A. Ben-Naim, "Statistical thermodynamics for chemists and biochemists," *Plenum Press. New York*, 1992.

[55] E. H. P. Debye, "De La Theorie Des Electrolytes. I. Abaissament Du Point De Congelation Et Phenomenes Associes." pp. 185–206, 1983.

[56] D. Bashford, “Macroscopic Electrostatic Models for Protonation States in Proteins,” *Front. Biosci.*, vol. 9, pp. 1082–1099, 2004.

[57] A. M. Baptista, P. J. Martel, and C. M. Soares, “Simulation of electron-proton coupling with a Monte Carlo method: Application to Cytochrome c3 Using Continuum Electrostatics,” *Biophys. J.*, vol. 76, no. 6, pp. 2978–2998, 1999.

[58] A. M. Baptista, V. H. Teixeira, and C. M. Soares, “Constant- *p* H molecular dynamics using stochastic titration,” *J. Chem. Phys.*, vol. 117, no. 9, pp. 4184–4200, 2002.

[59] A. M. Walczak and J. M. Antosiewicz, “Langevin dynamics of proteins at constant pH,” *Am. Phys. Soc.*, vol. 66, 2002.

[60] J. Mongan, D. A. Case, and J. A. McCammon, “Constant pH molecular dynamics in generalized Born implicit solvent,” *J. Comput. Chem.*, vol. 25, no. 16, pp. 2038–2048, 2004.

[61] M. S. Lee, F. R. Salsbury, and C. L. Brooks, “Constant-pH molecular dynamics using continuous titration coordinates,” *Proteins Struct. Funct. Genet.*, vol. 56, no. 4, pp. 738–752, 2004.

[62] J. Khandogin and C. L. Brooks, “Constant pH molecular dynamics with proton tautomerism,” *Biophys. J.*, vol. 89, no. 1, pp. 141–157, 2005.

[63] M. Dlugosz and J. M. Antosiewicz, “Constant-pH molecular dynamics simulations: A test case of succinic acid,” *Chem. Phys.*, vol. 302, no. 1–3, pp. 161–170, 2004.

[64] M. Długosz, J. M. Antosiewicz, and A. D. Robertson, “Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 69, no. 2 1, pp. 1–10, 2004.

[65] A. M. Baptista, P. J. Martel, and S. B. Petersen, “Simulation of protein conformational freedom as a function of pH: Constant pH molecular dynamics using implicit titration,” *Proteins Struct., Funct., Genet.*, vol. 27, no. 4, p. 523, 1997.

[66] M. Machuqueiro and A. M. Baptista, “Constant-pH molecular dynamics with ionic strength effects: Protonation-conformation coupling in decalysine,” *J. Phys. Chem. B*, vol. 110, no. 6, pp. 2927–2933, 2006.

[67] J. Kästner, “Umbrella sampling,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 1, no. 6, pp. 932–942, 2011.

[68] D. van der Spoel *et al.*, “Gromacs User Manual version 4.0,” *Manuals*, p. 308, 2005.

[69] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, “LINCS: A Linear Constant Solver for Molecular Simulations,” *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997.

[70] P. A. Miyamoto, S., Kollman, “SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models,” *J. Comp. Chem.*, vol. 13, pp. 952–962, 1992.

[71] J. A. Barker and R. O. Watts, “Monte carlo studies of the dielectric properties of water-like models,” *Mol. Phys.*, vol. 26, no. 3, pp. 789–792, 1973.

[72] I. G. Tironi, P. E. Smith, W. F. Van Gunsteren, and I. Introduction, “A generalized reaction field method for molecular dynamics simulations,” vol. 102, no. April, 1995.

[73] P. E. Smith and W. F. van Gunsteren, “Consistent dielectric properties of the simple point charge and the extended simple point charge water models at 277 and 300K,” *J. Chem. Phys.*, vol. 100, no. February, pp. 3169–3174, 1994.

[74] G. Bussi, D. Donadio, and M. Parrinello, “Canonical sampling through velocity rescaling,” *J. Chem. Phys.*, vol. 126, no. 1, 2007.

[75] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak,

"Molecular dynamics with coupling to an external bath," *J. Am. Chem. Soc.*, vol. 81, no. 8, pp. 3684–3690, 1984.

[76] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, 1981.

[77] W. L. DeLano, *The PyMOL molecular graphics system*. 2002.

[78] J. A. N. Hermans and W. F. Van, "A Consistent Empirical Potential lor Water-Protein Interactions," *Biopolymers*, vol. 23, no. 8, pp. 1513–1518, 1984.

[79] N. Schmid *et al.*, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7," *Eur. Biophys. J.*, vol. 40, no. 7, pp. 843–856, 2011.

[80] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, 2008.

[81] D. Bashford and K. Gerwert, "Electrostatic Calculations of the pKa Values of Ionizable Groups in Bacteriorhodopsin," *J.Mol.Biol.*, pp. 473–486, 1992.

[82] N. A. Baker, D. Bashford, and D. A. Case, "Implicit Solvent Electrostatics in Biomolecular Simulation," *New Algorithms Macromol. Simul.*, vol. 49, no. 10, pp. 263–295, 2006.

[83] C. A. Carvalheda, S. R. R. Campos, M. Machuqueiro, and A. M. Baptista, "Structural effects of ph and deacylation on surfactant protein c in an organic solvent mixture: A constant-pH MD study," *J. Chem. Inf. Model.*, vol. 53, no. 11, pp. 2979–2989, 2013.

[84] V. H. Teixeira *et al.*, "On the use of different dielectric constants for computing individual and pairwise terms in poisson-boltzmann studies of protein ionization equilibrium," *J. Phys. Chem. B*, vol. 109, no. 30, pp. 14691–14706, 2005.

[85] A. M. Baptista and C. M. Soares, "Some Theoretical and Computational Aspects of the Inclusion of Proton Isomerism in the Protonation Equilibrium of Proteins," *J. Phys. Chem. B*, vol. 105, no. 1, pp. 293–309, 2001.

[86] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[87] M. Machuqueiro and A. M. Baptista, "Acidic range titration of HEWL using a constant-pH molecular dynamics method," *Proteins Struct. Funct. Genet.*, vol. 72, no. 1, pp. 289–298, 2008.

[88] "http://www.gnuplot.info/." [Online]. Available: http://www.gnuplot.info/.

[89] C. A. Carvalheda, S. R. R. Campos, and A. M. Baptista, "The Effect of Membrane Environment on Surfactant Protein C Stability Studied by Constant-pH Molecular Dynamics," *J. Chem. Inf. Model.*, vol. 55, no. 10, pp. 2206–2217, 2015.

[90] J. R.Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. 1997.

[91] I. Eberini *et al.*, "Electrostatics of folded and unfolded bovine β-lactoglobulin," *Amino Acids*, vol. 42, no. 5, pp. 2019–2030, 2012.

[92] R. L. Thurlkill, G. R. Grimsley, J. M. Scholtz, and C. N. Pace, "pK values of the ionizable groups of proteins," *Protein Sci.*, vol. 15, no. 5, pp. 1214–1218, 2006.

[93] G. R. Grimsley, J. M. Scholtz, and C. N. Pace, "A summary of the measured pK values of the ionizable groups in folded proteins," *Protein Sci.*, vol. 18, no. 1, pp. 247–251, 2008.

[94] S. N. Timasheff, L. Mescanti, J. J. Bash, and R. Townend, "Conformational A , B , and C Transitions of Bovine," *J. Biol. Chem.*, vol. 241, no. 11, pp. 2496–2501, 1966.

# A. APPENDIX I

## A1. Radius of gyration

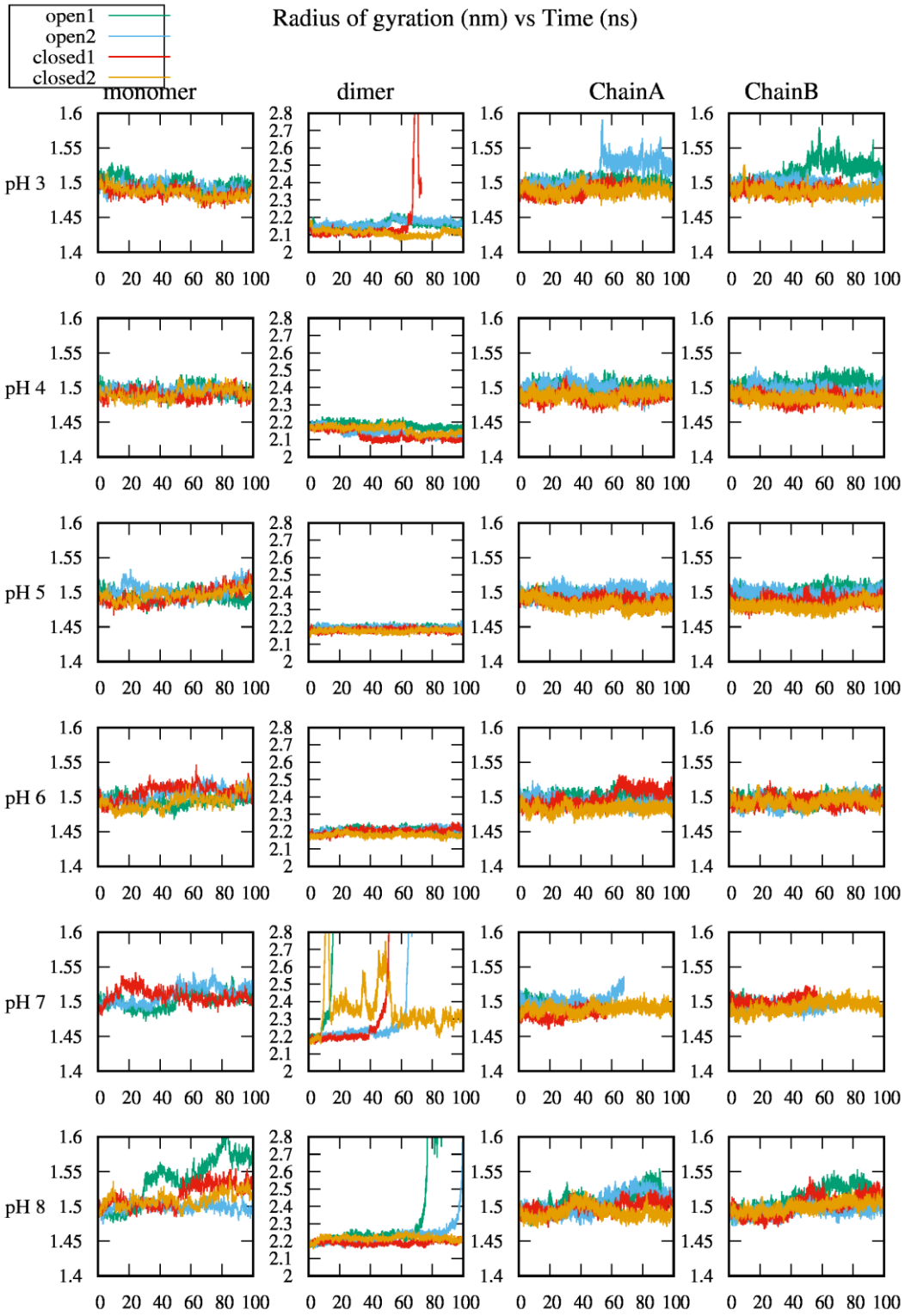


**Figure A1.1** Radius of gyration vs time considering the monomeric and dimeric forms in the open and closed conformations, for each pH value.

## A2. Secondary Structure

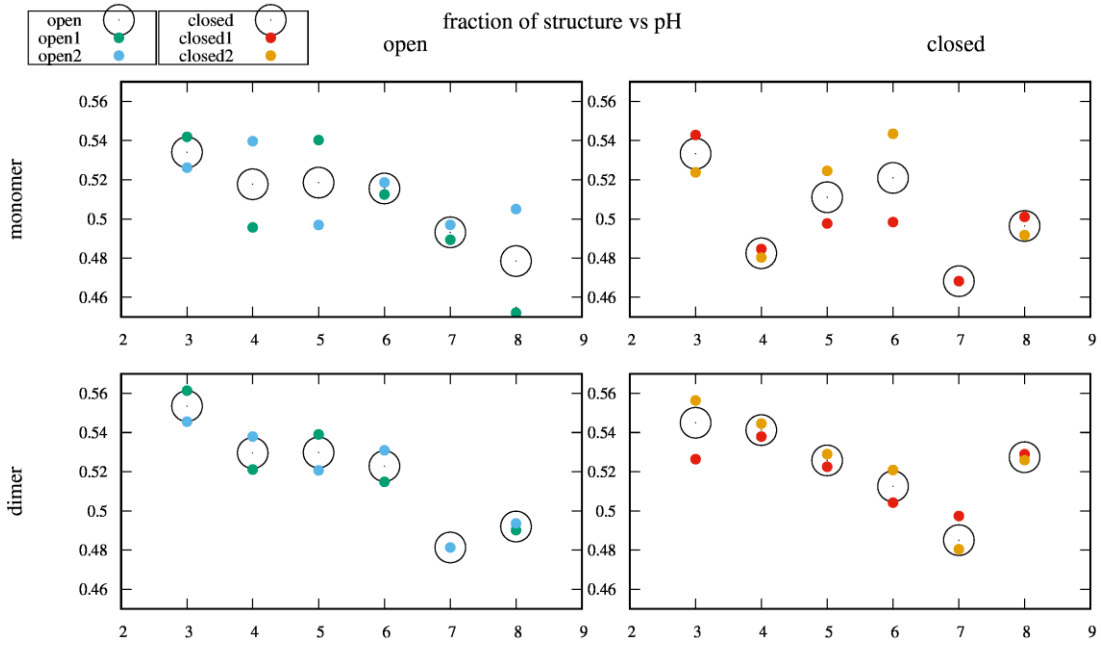


**Figure A2.1** Mean fraction of the secondary structure as function of the pH, considering the two replicates for the open and closed conformation.

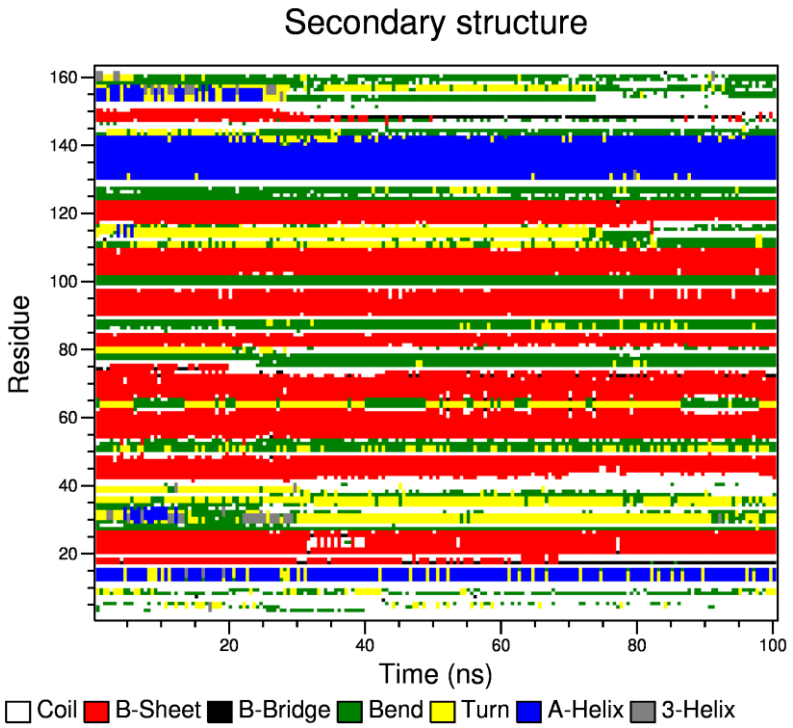


**Figure A2.2** Secondary structure of the open monomer, replicate 1, at pH 8 during the 100 ns of simulation time.

## A3. Distances

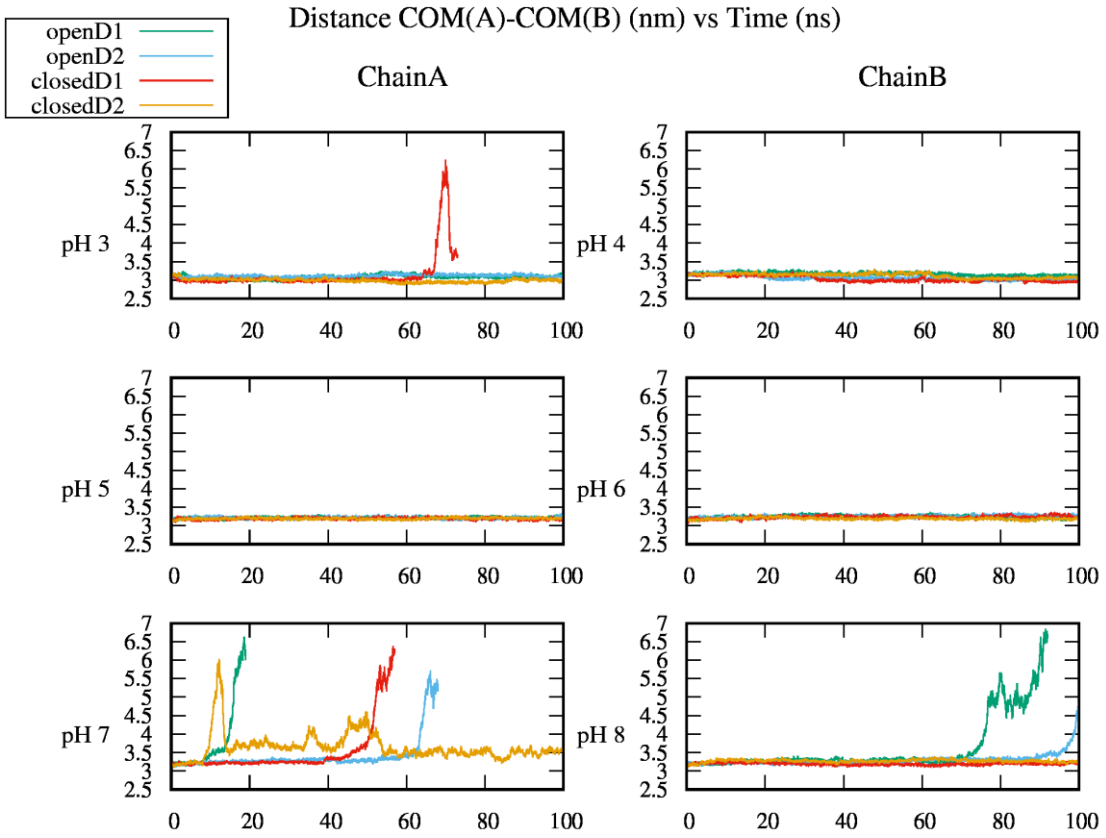


**Figure A3.1** Distance between the COM of chain A to the COM of chain B, considering the open and closed dimer, for each pH value.

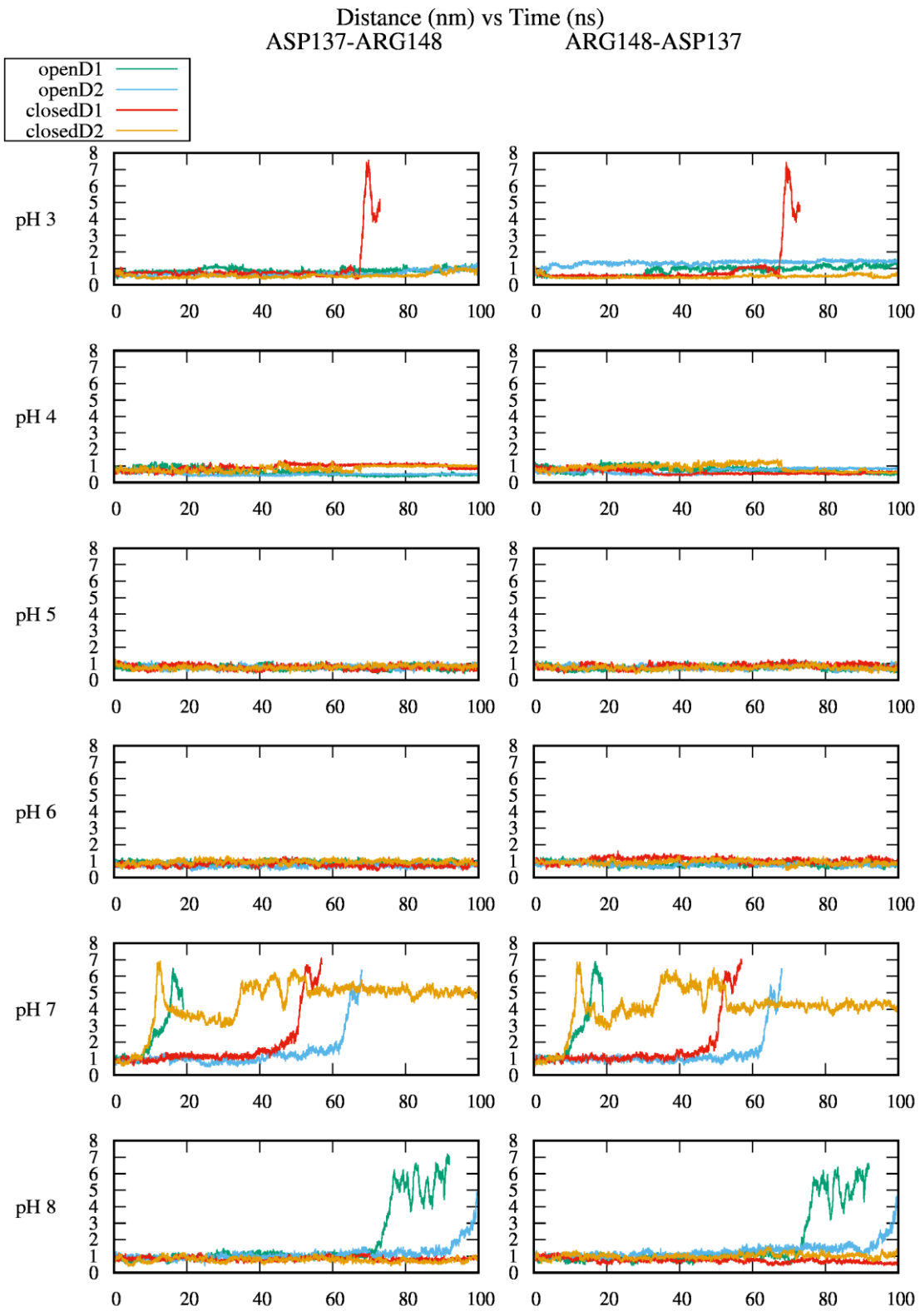


**Figure A3.2** Distance between Asp137 and Arg148 of different chains, considering the open and closed dimer, for each pH value.

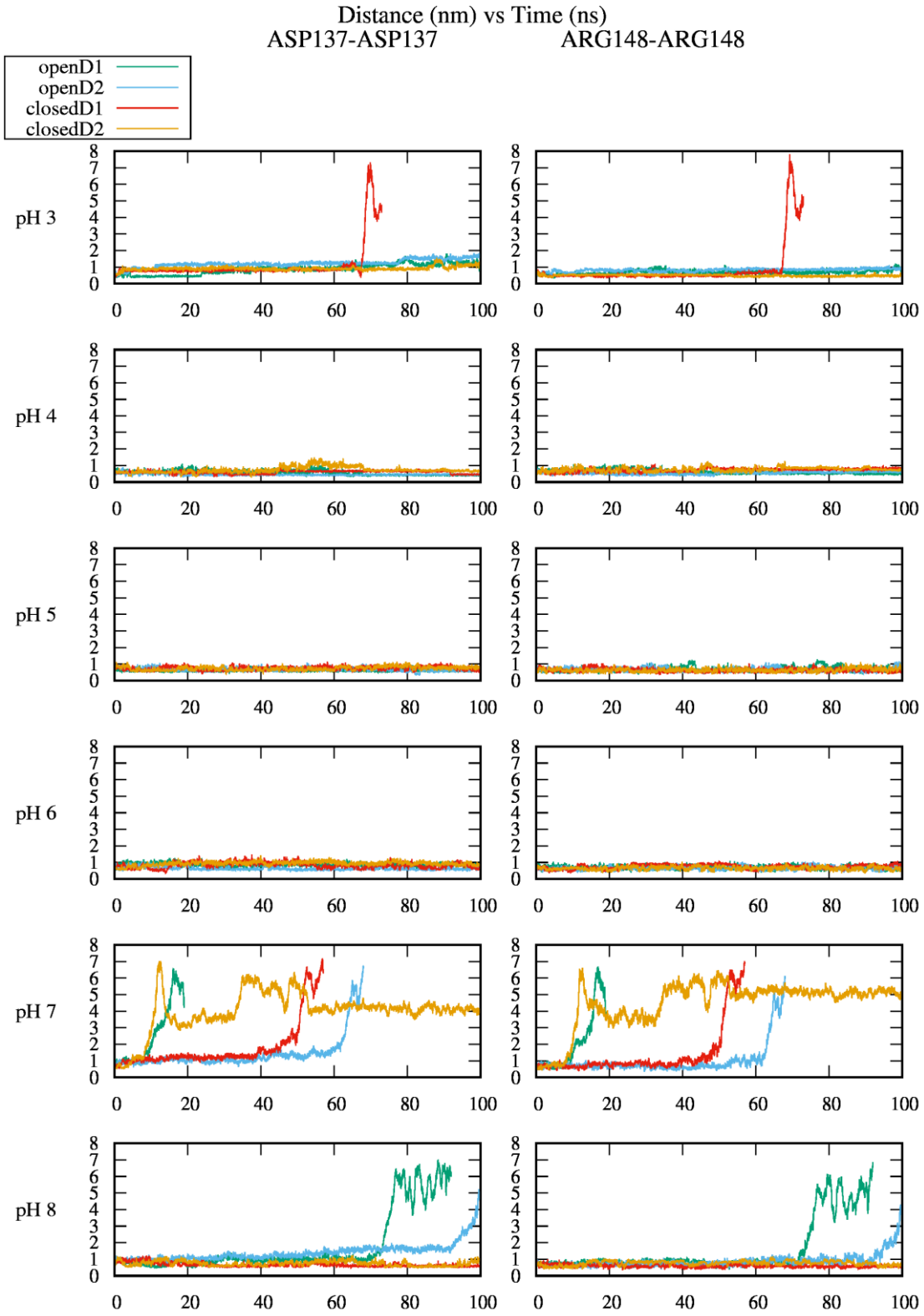


**Figure A3.3** Distance between Asp137-Asp137 and Arg148-Arg148 of different chains, considering the open and closed dimer, for each pH value.

## A4. Individual titration curves from CpHMD

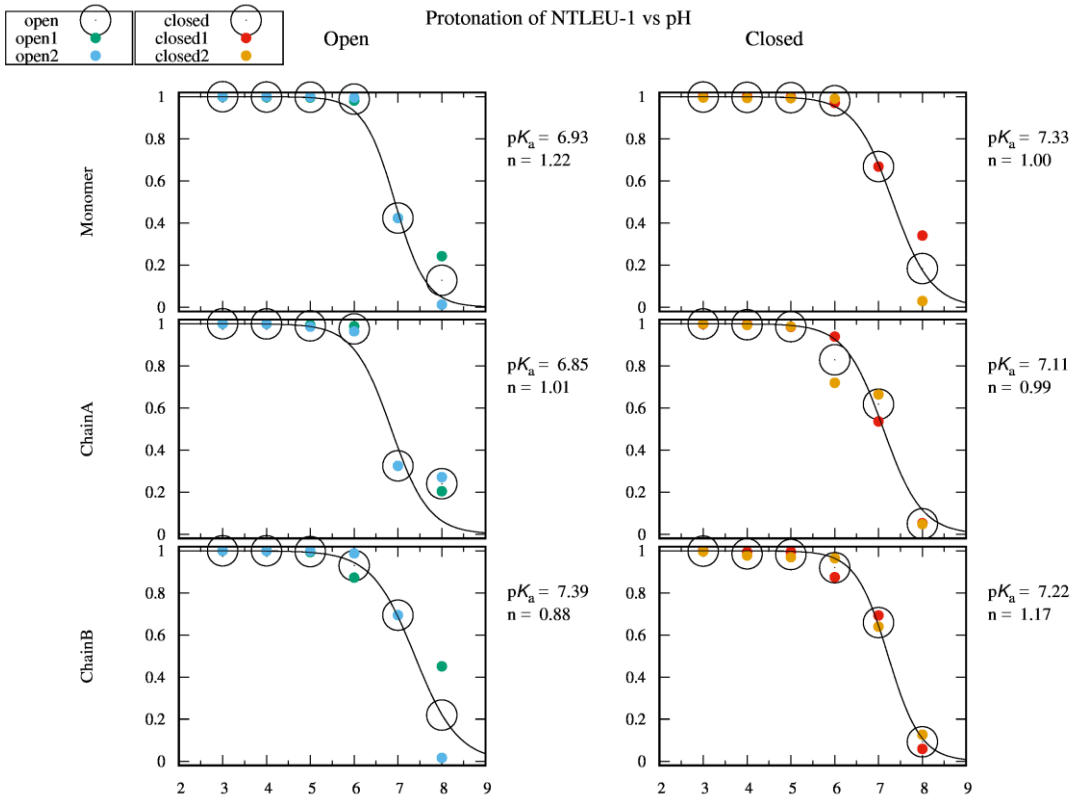


**Figure A4 1** Protonation of Nter as a function of pH.

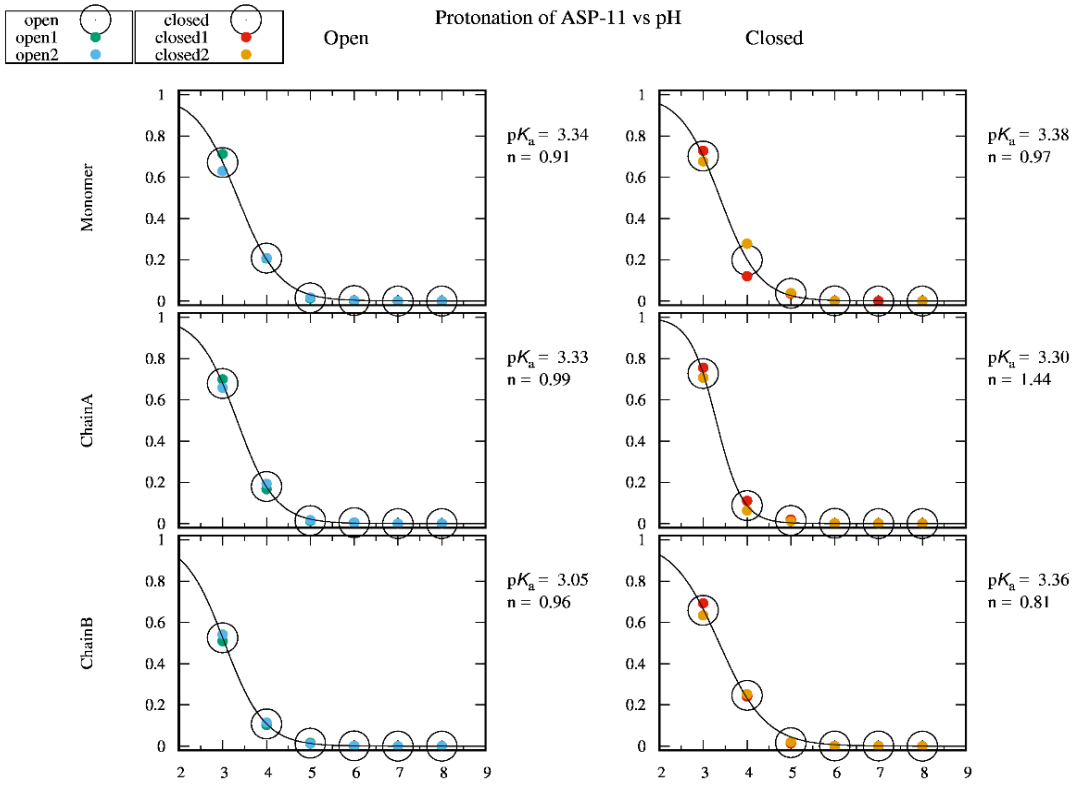


**Figure A4 2** Protonation of Asp11 as a function of pH.

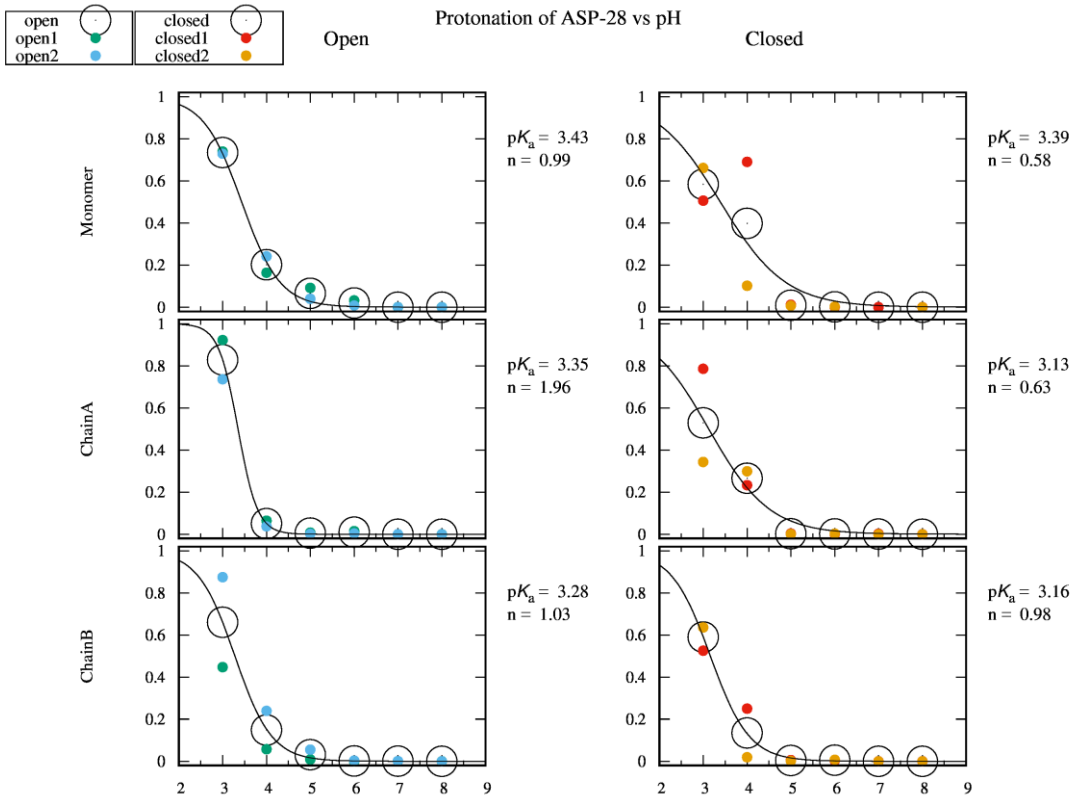


**Figure A4 3** Protonation of Asp28 as a function of pH.

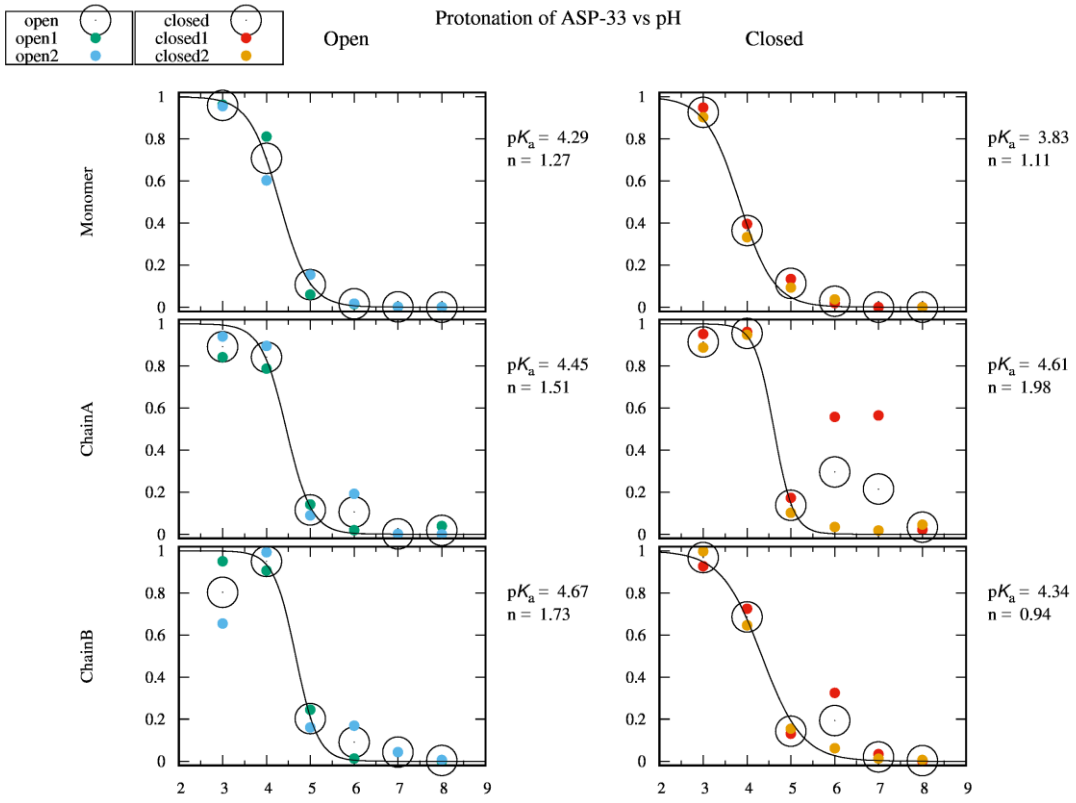


**Figure A4 4** Protonation of Asp33 as a function of pH.

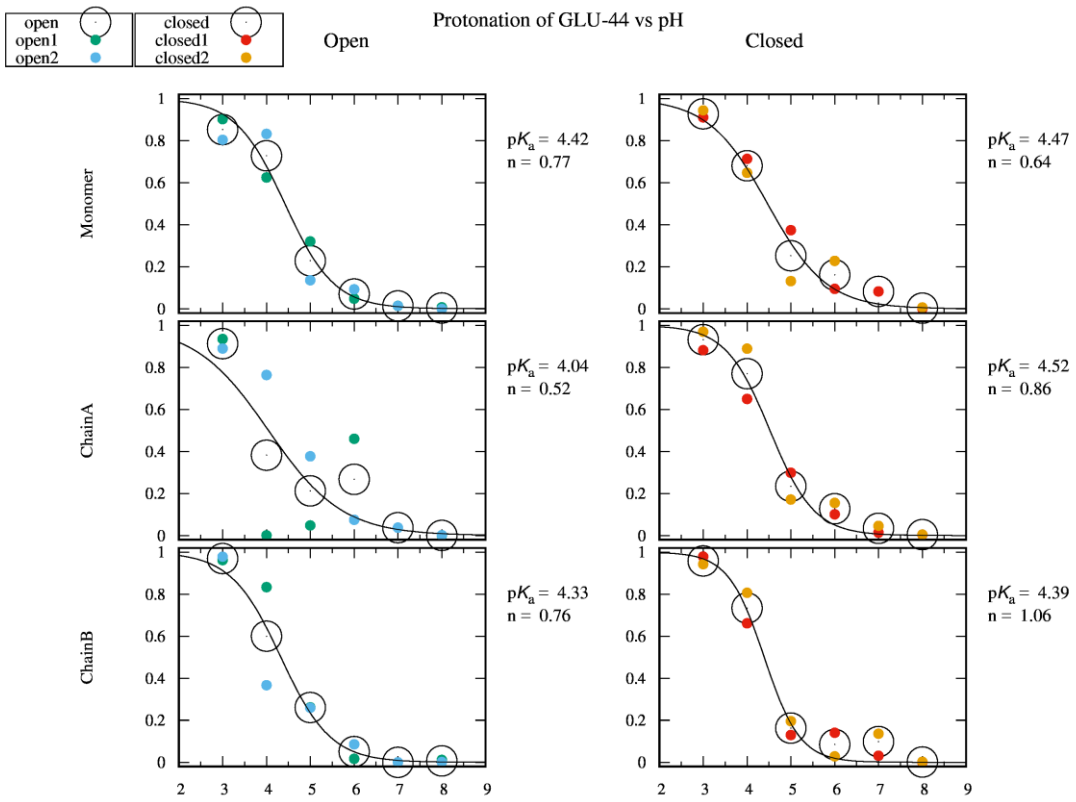


**Figure A4 5** Protonation of Glu44 as a function of pH.

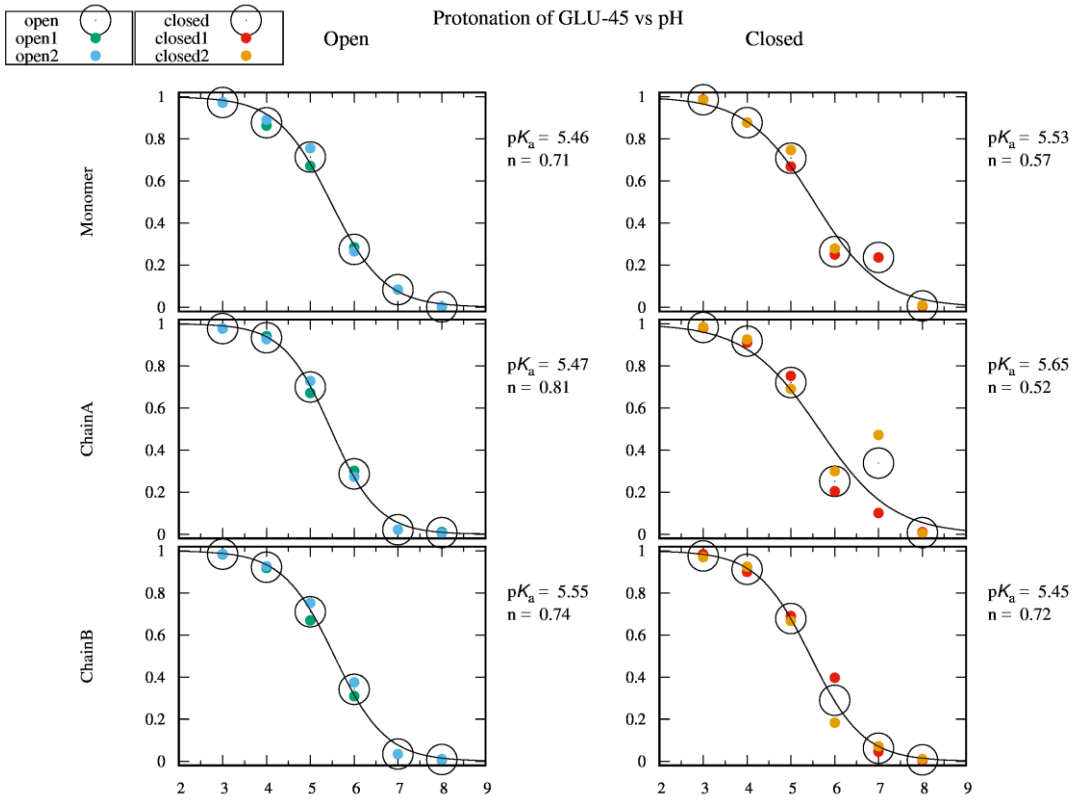


**Figure A4 6** Protonation of Glu45 as a function of pH.

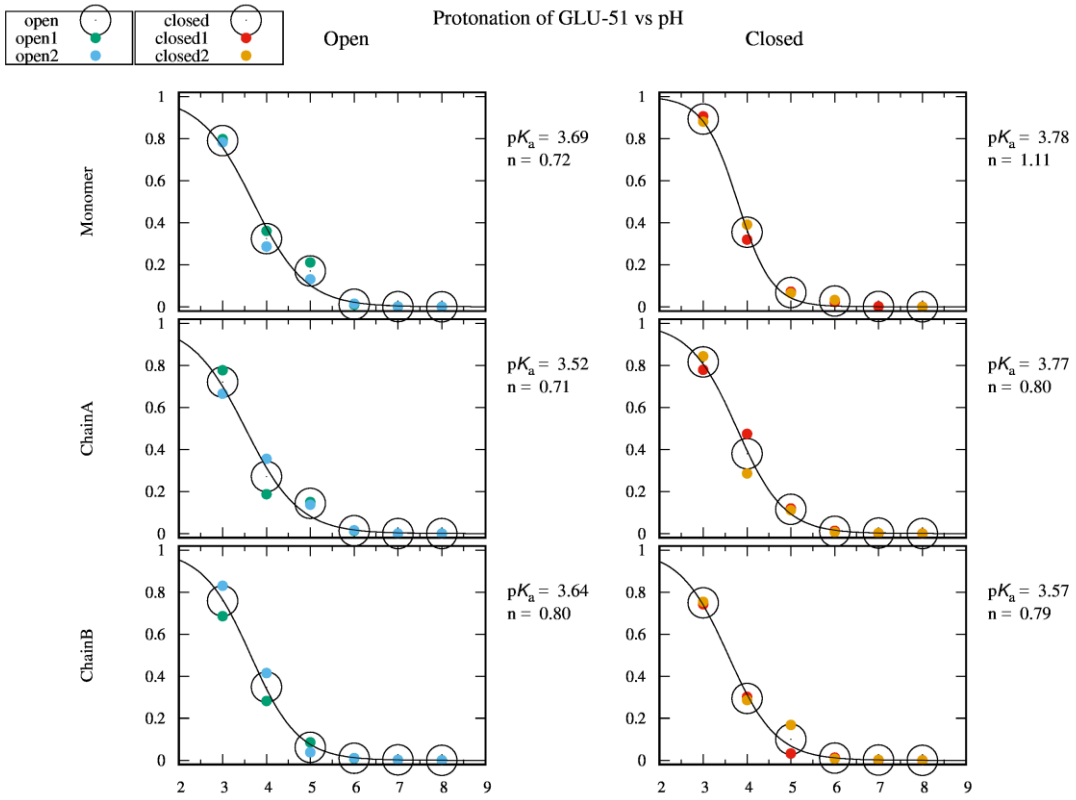


**Figure A4 7** Protonation of Glu51 as a function of pH.

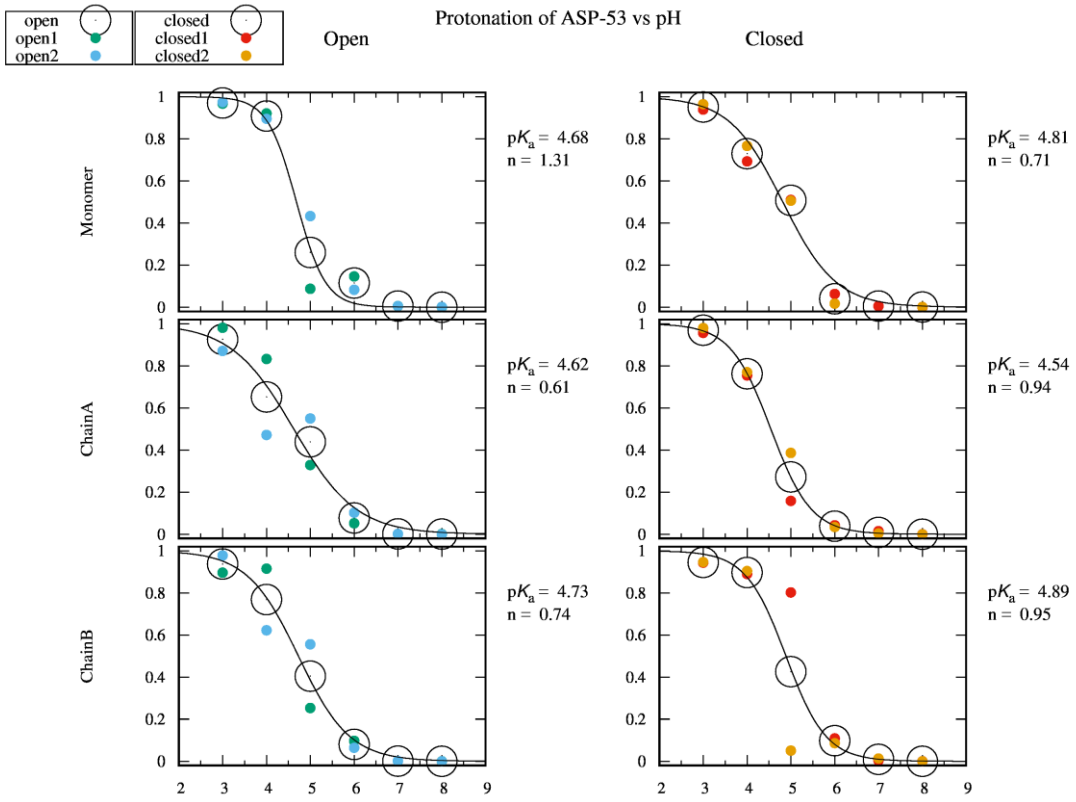


**Figure A4 8** Protonation of Asp53 as a function of pH.

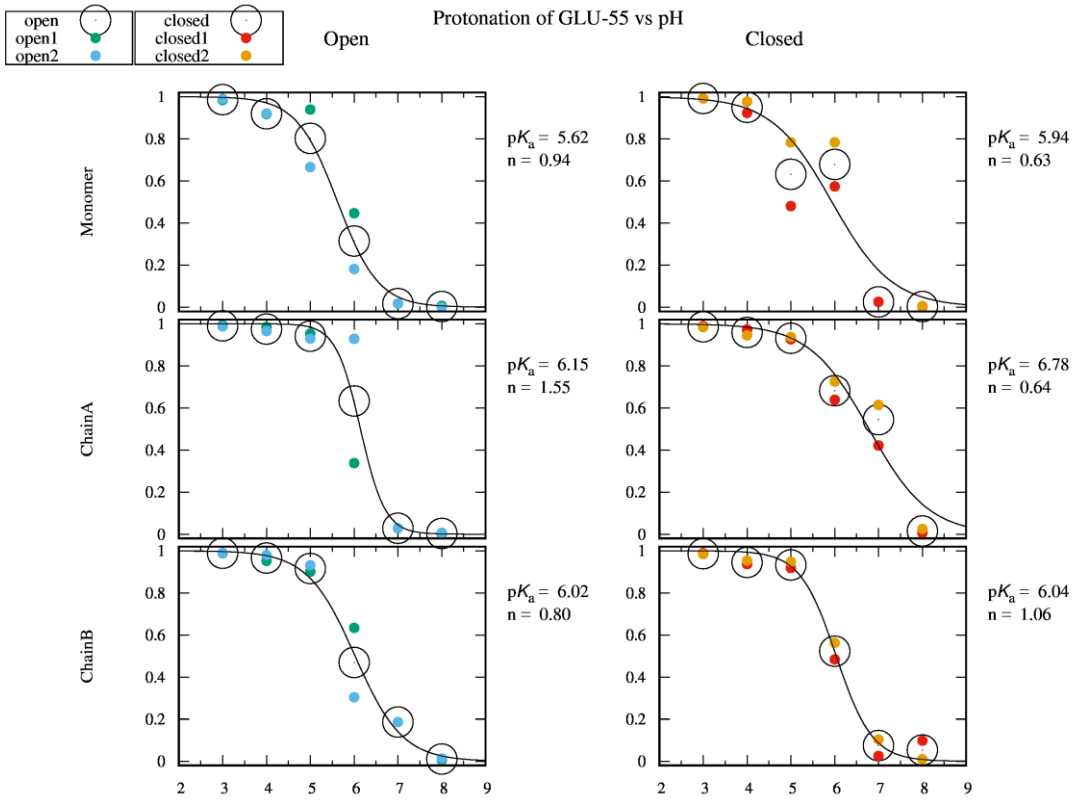


**Figure A4 9** Protonation of Glu55 as a function of pH.

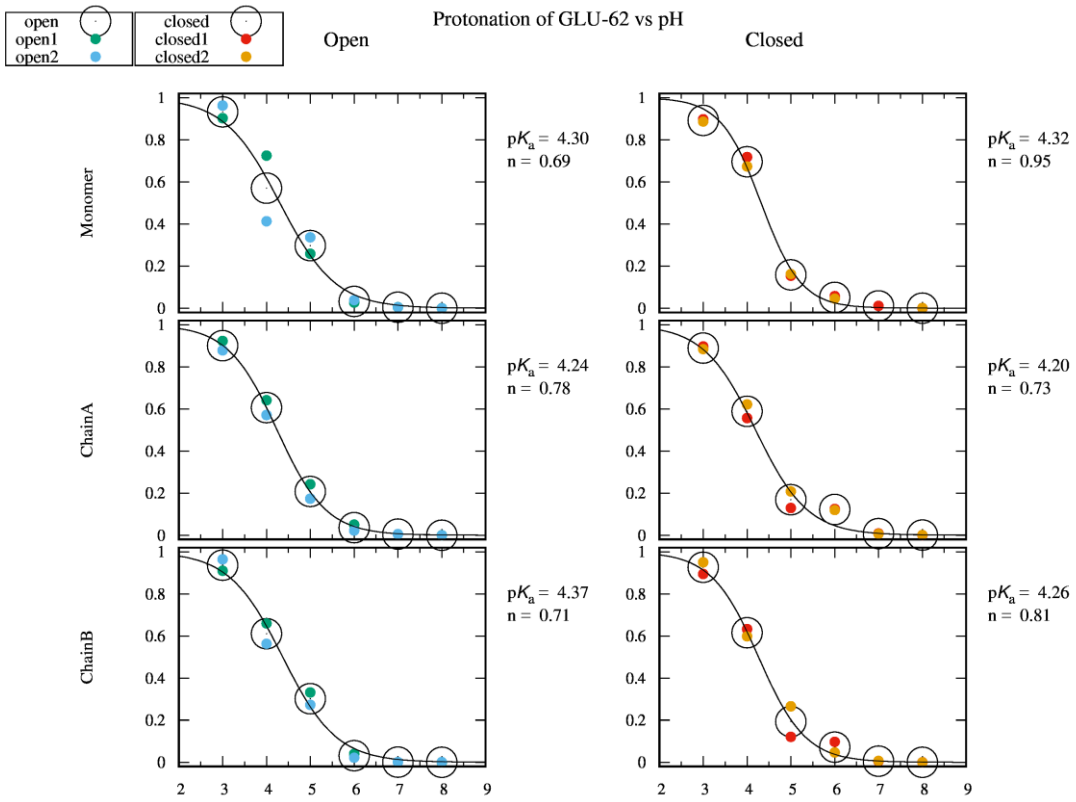


**Figure A4 10** Protonation of Glu62 as a function of pH.

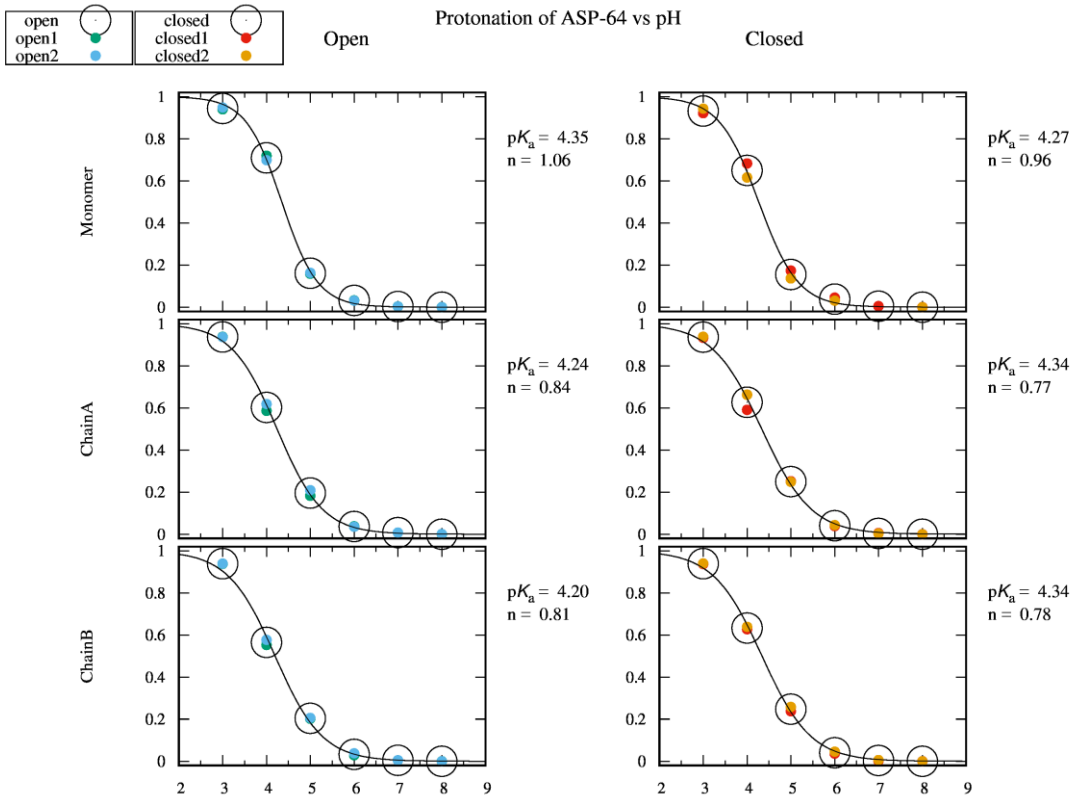


**Figure A4 11** Protonation of Asp64 as a function of pH.

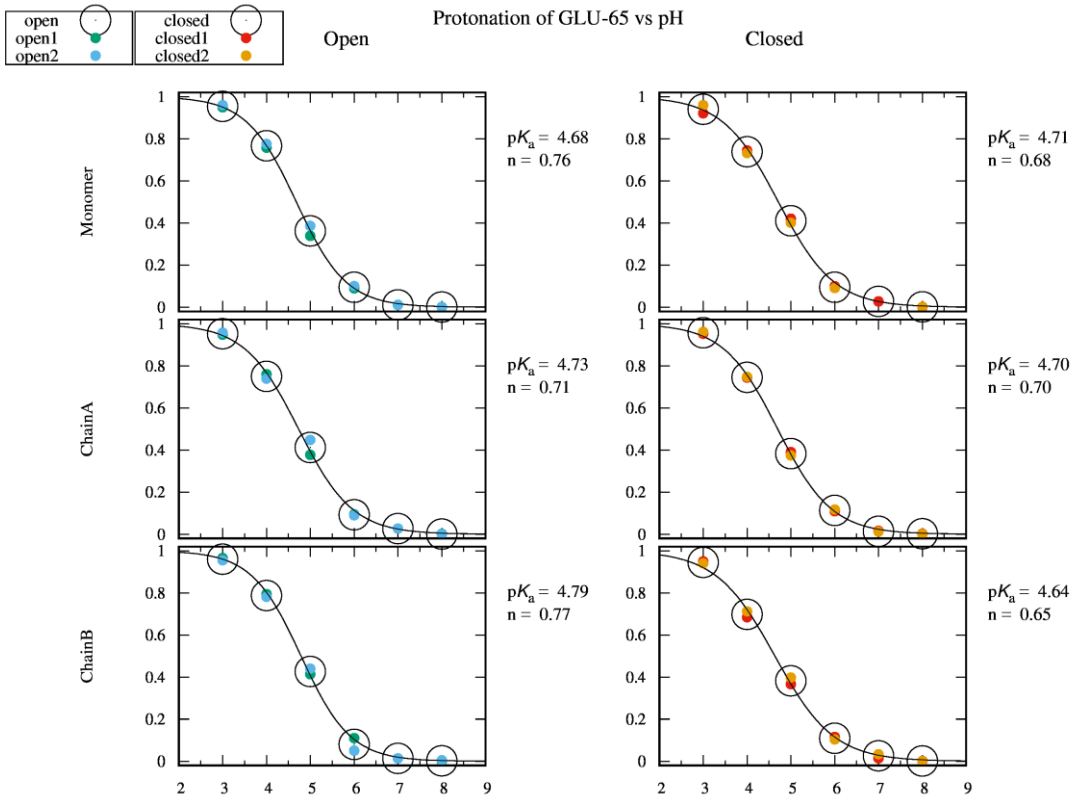


**Figure A4 12** Protonation of Glu65 as a function of pH.

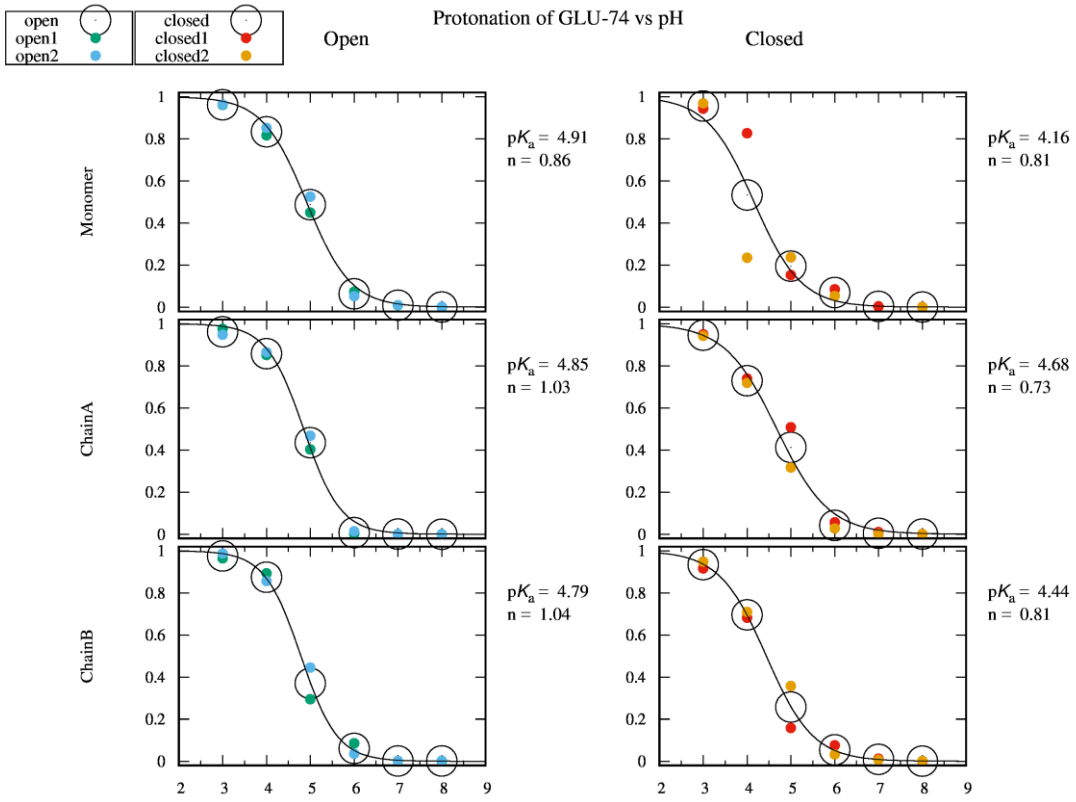


**Figure A4 13** Protonation of Glu74 as a function of pH.

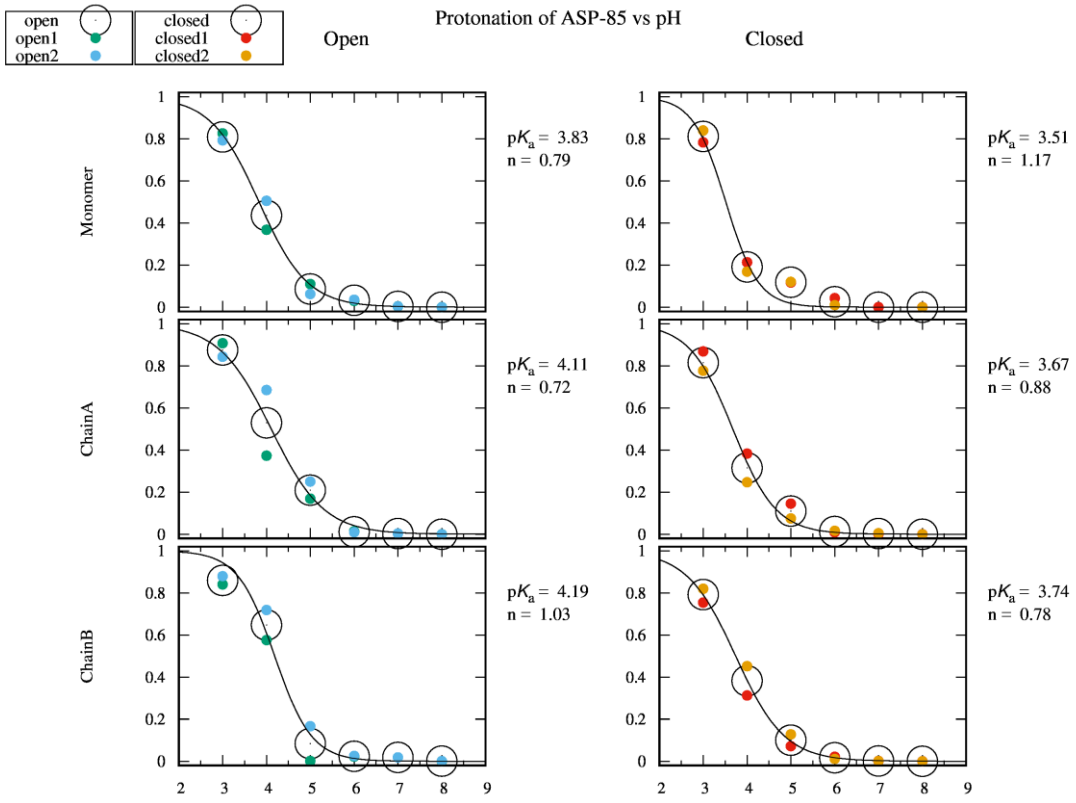


**Figure A4 14** Protonation of Asp85 as a function of pH.

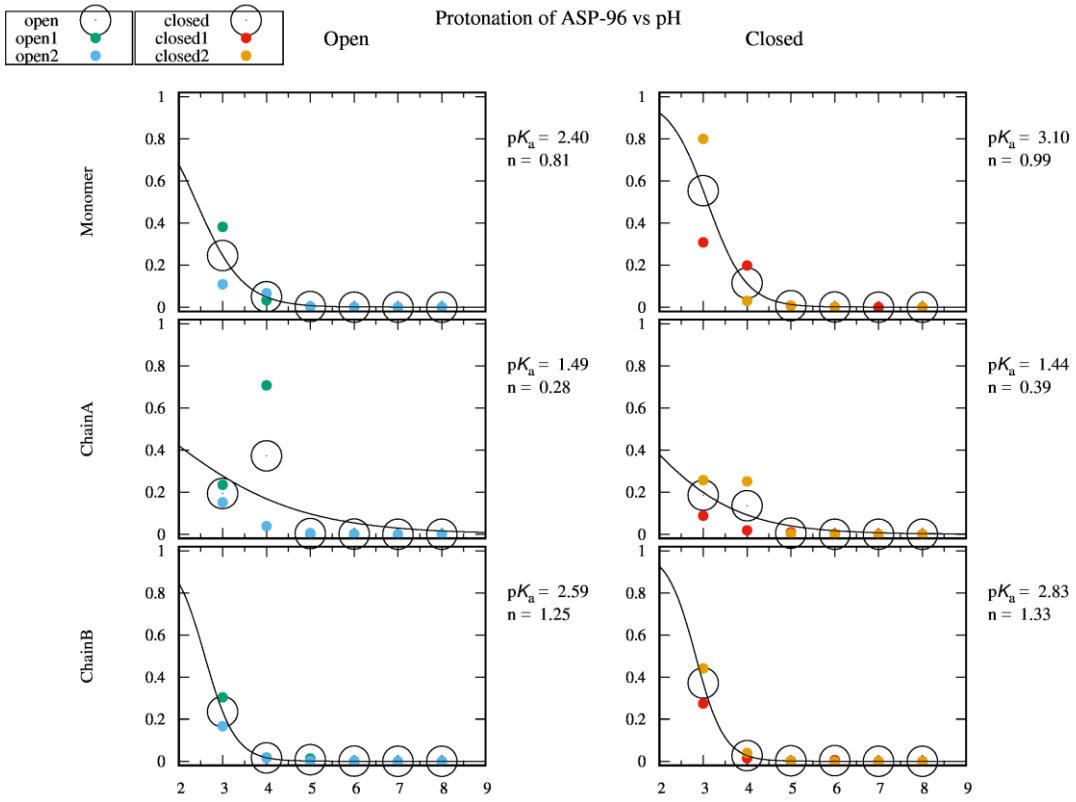


**Figure A4 15** Protonation of Asp96 as a function of pH.

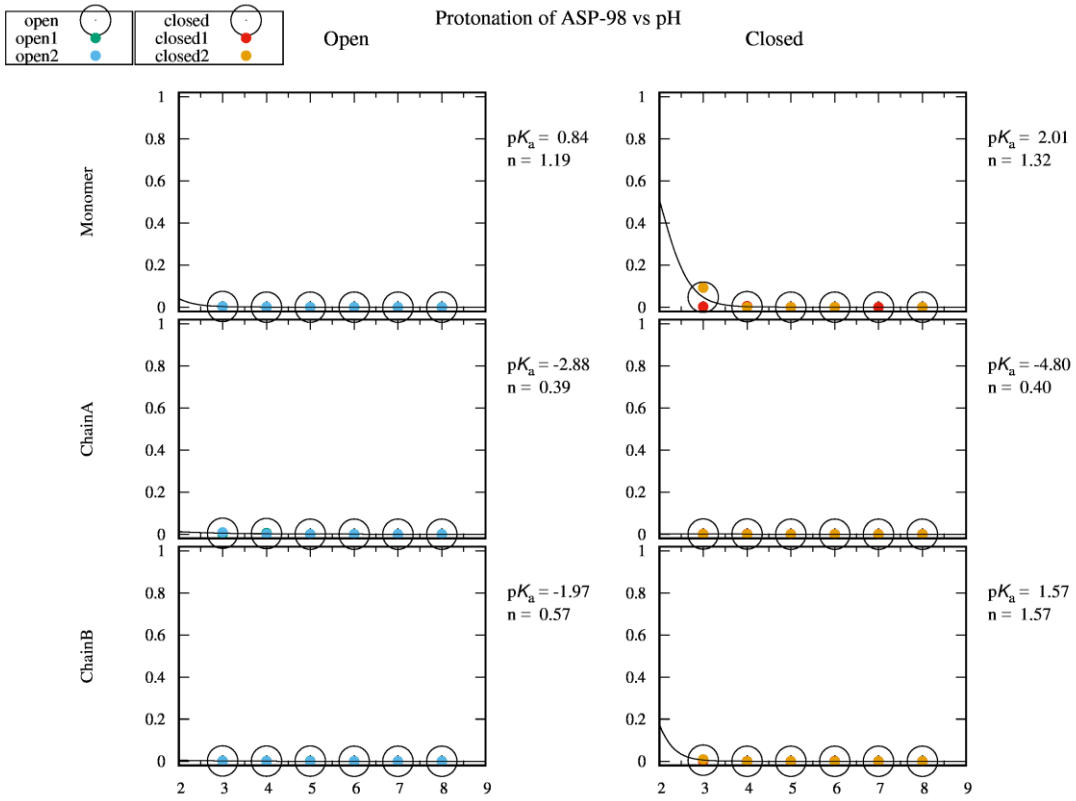


**Figure A4 16** Protonation of Asp98 as a function of pH.

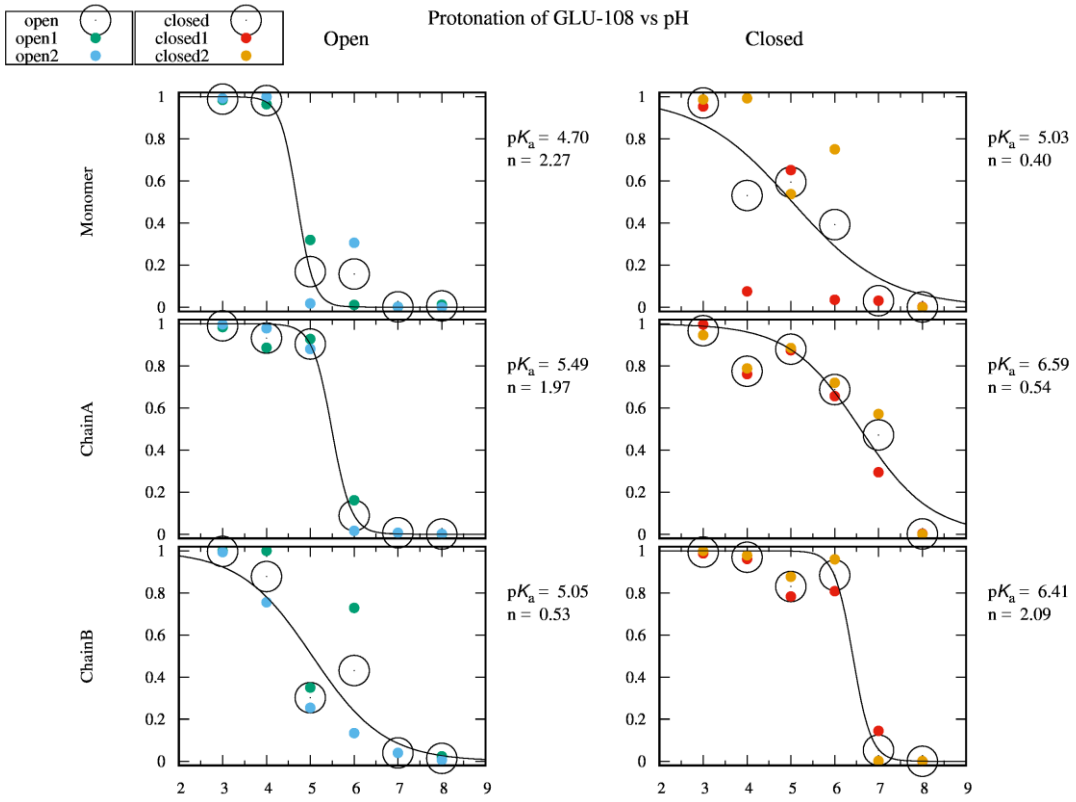


**Figure A4 17** Protonation of Glu108 as a function of pH.

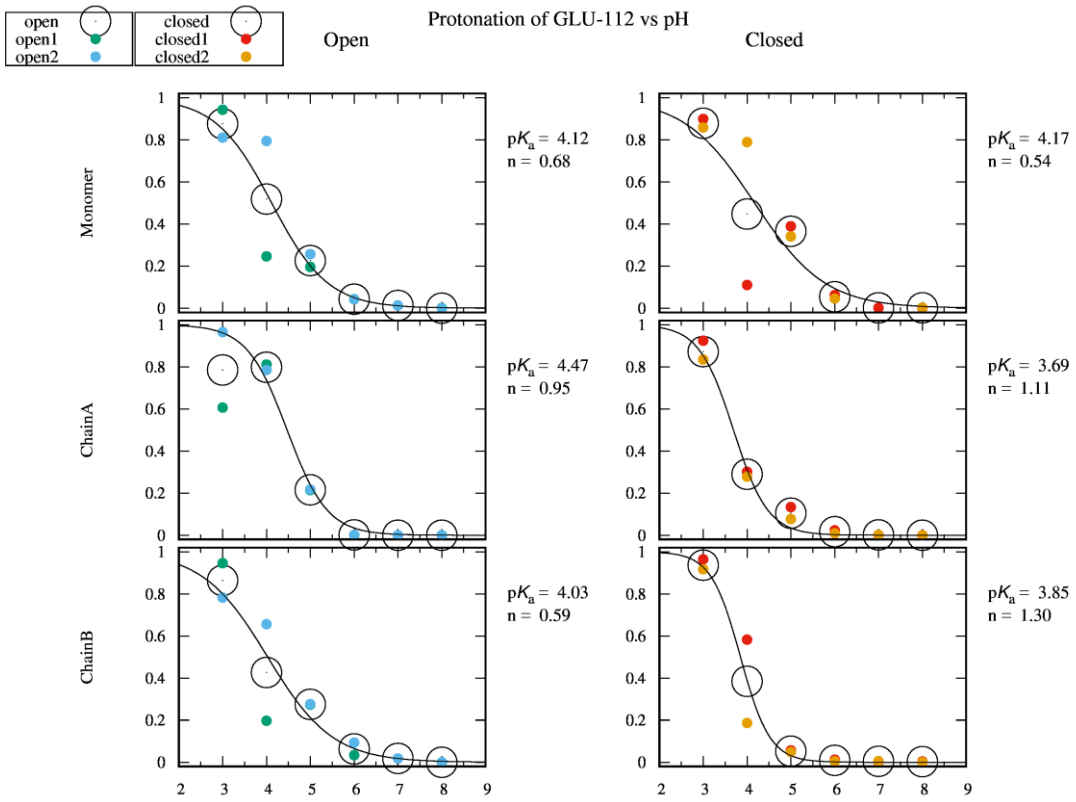


**Figure A4 18** Protonation of Glu112 as a function of pH.

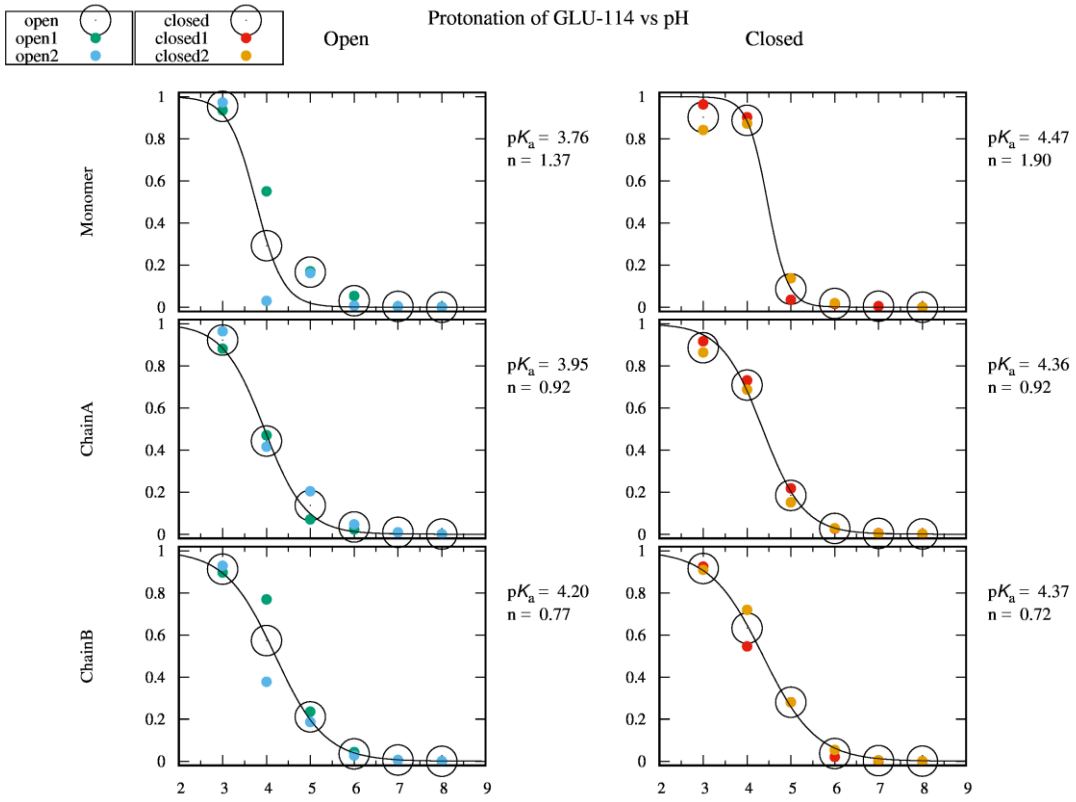


**Figure A4 19** Protonation of Glu114 as a function of pH.

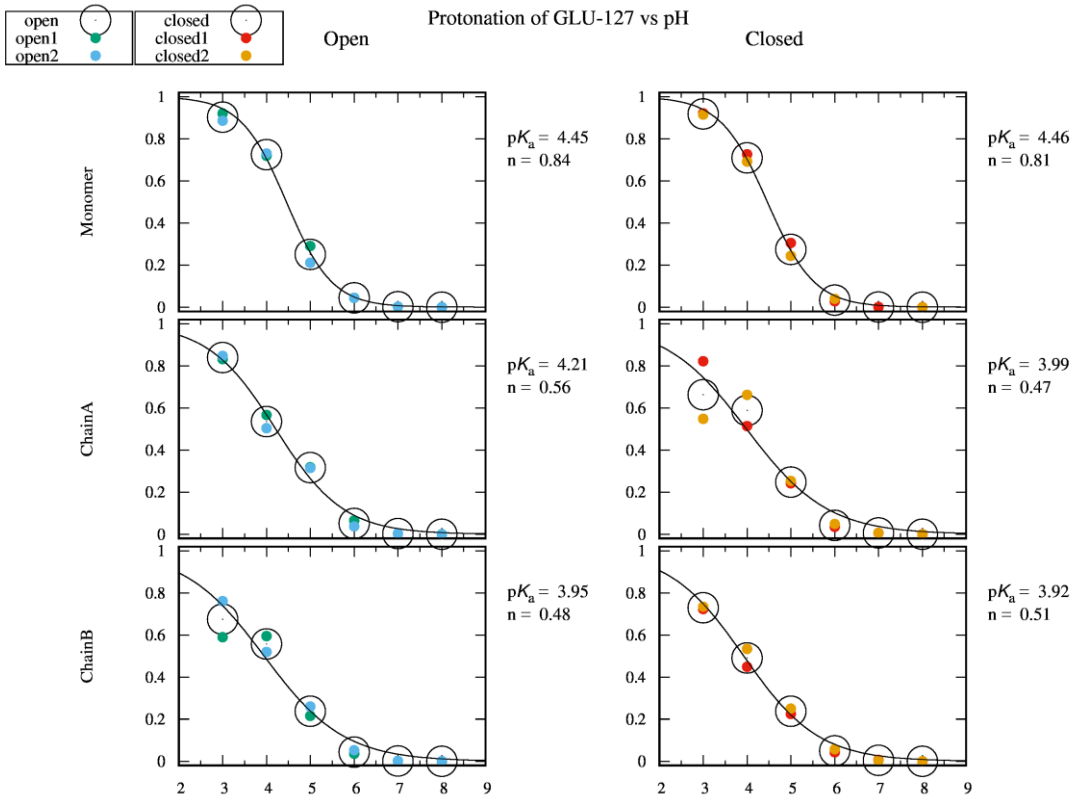


**Figure A4 20** Protonation of Glu127 as a function of pH.

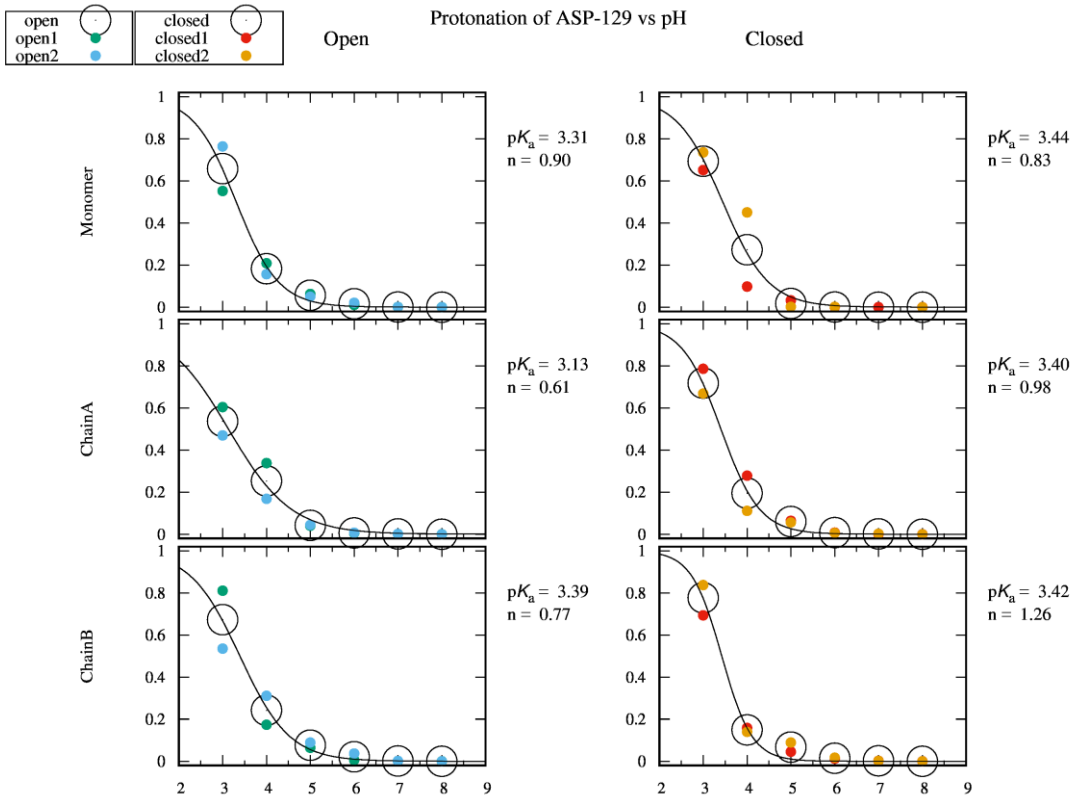


**Figure A4 21** Protonation of Asp129 as a function of pH.

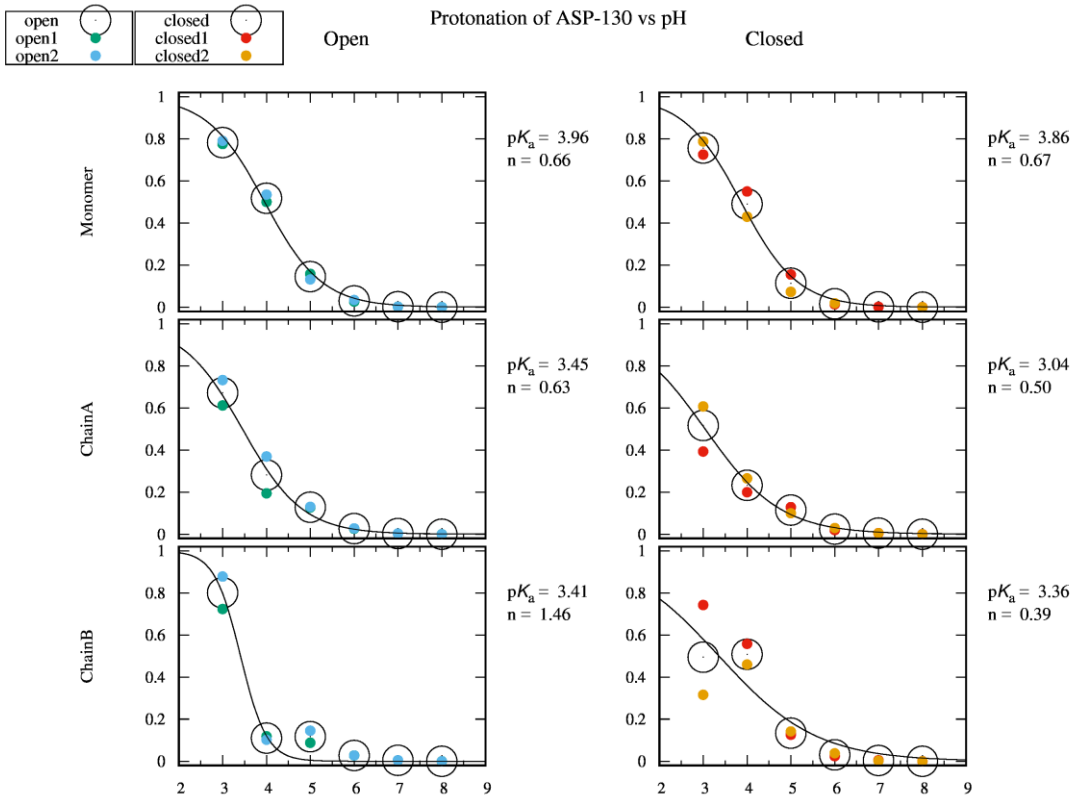


**Figure A4 22** Protonation of Asp130 as a function of pH.

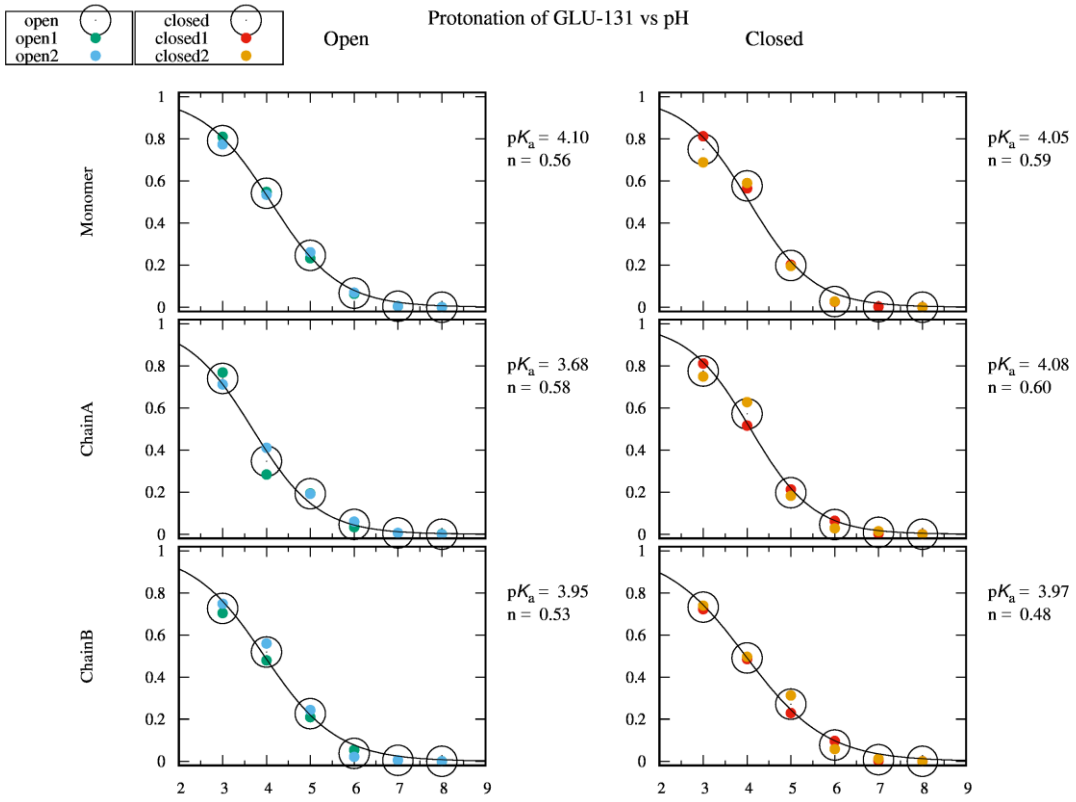


**Figure A4 23** Protonation of Glu131 as a function of pH.

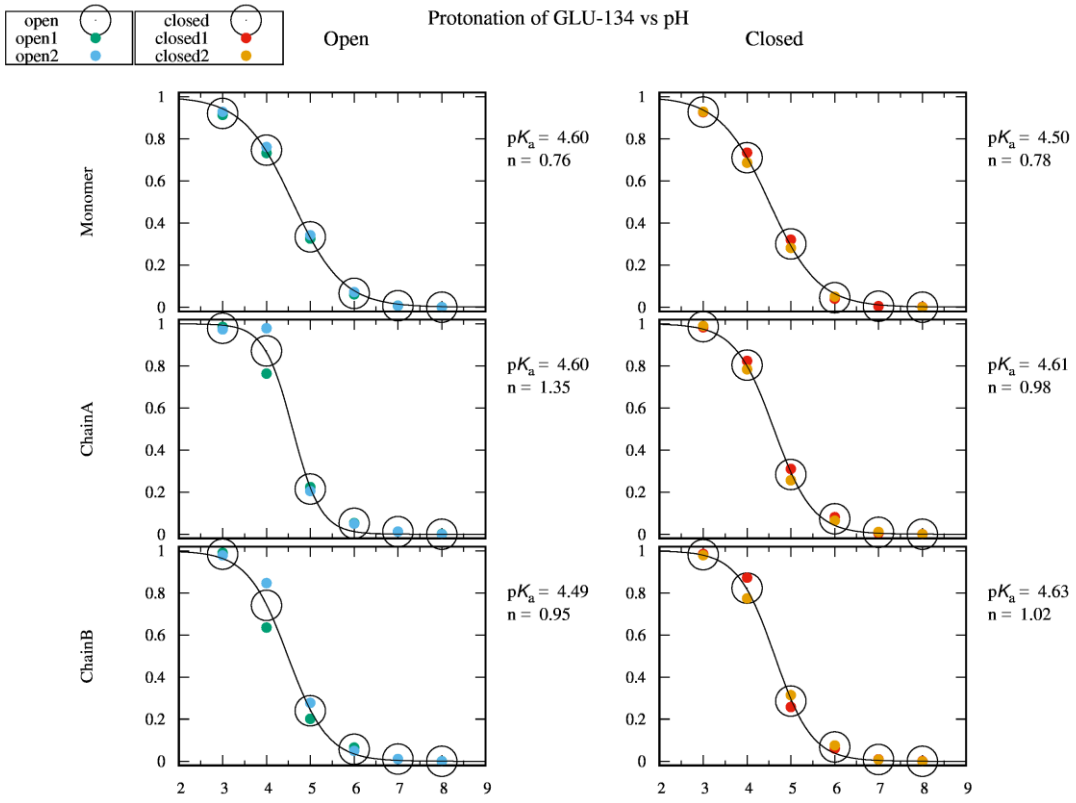


**Figure A4 24** Protonation of Glu134 as a function of pH.

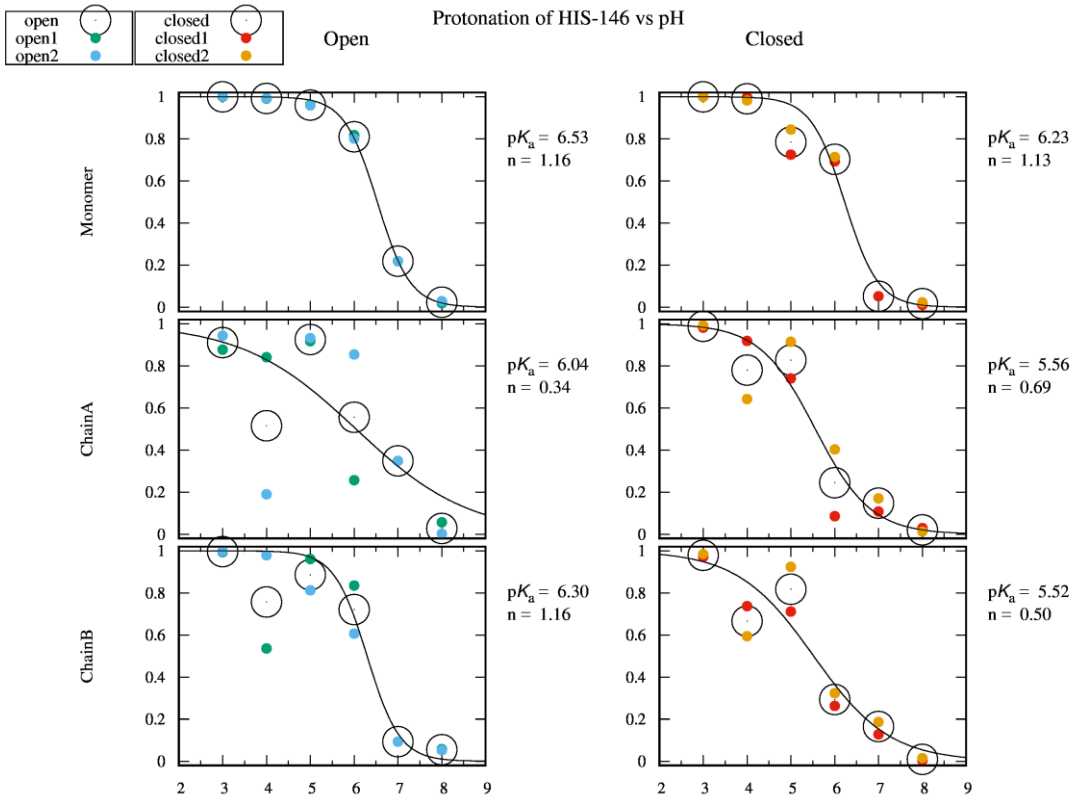


**Figure A4 25** Protonation of His146 as a function of pH.

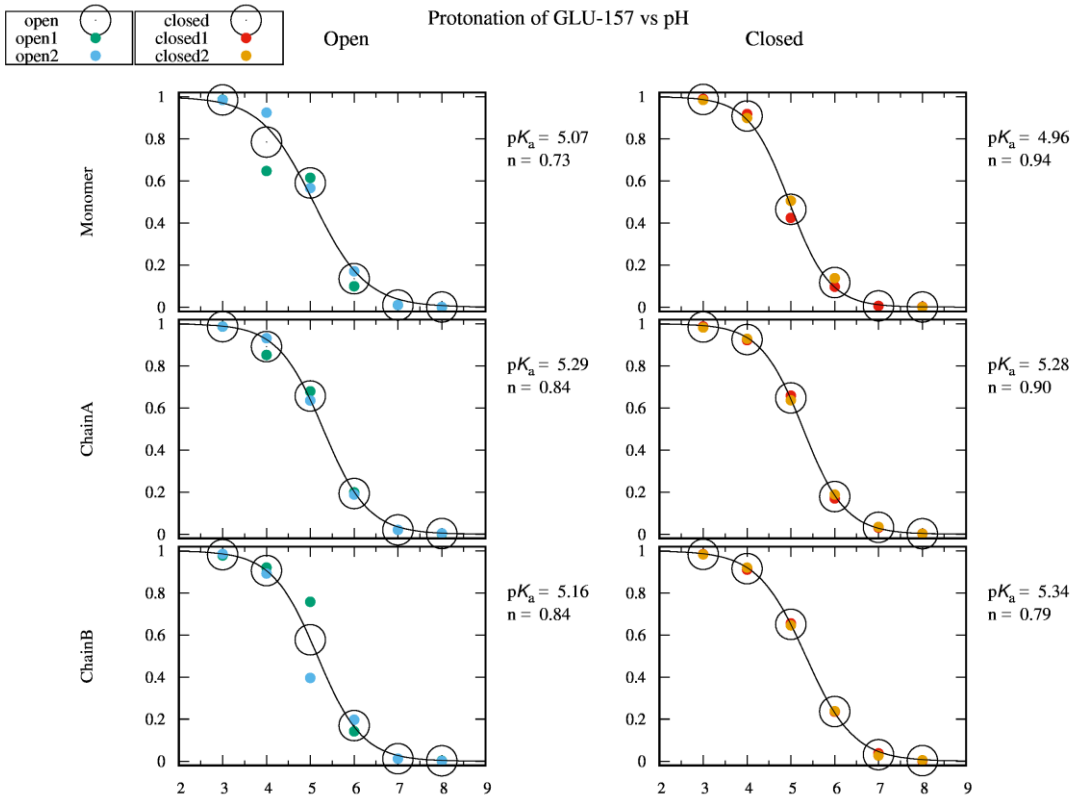


**Figure A4 26** Protonation of Glu157 as a function of pH.

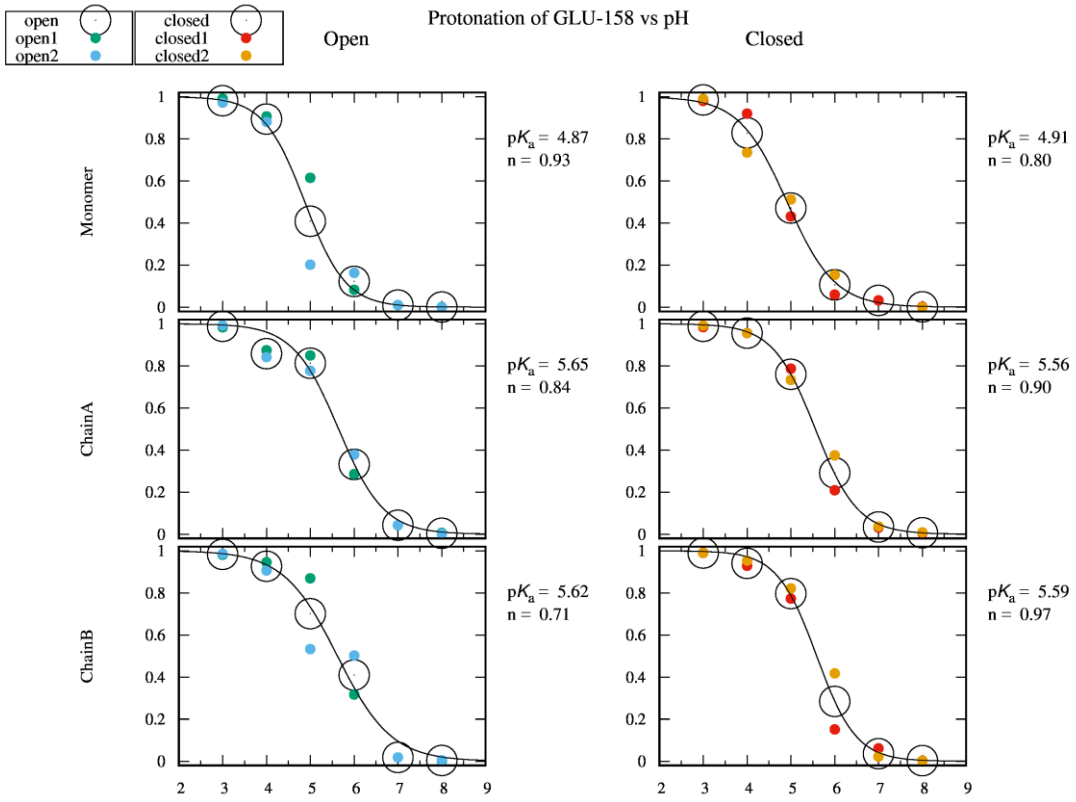


**Figure A4 27** Protonation of Glu158 as a function of pH.

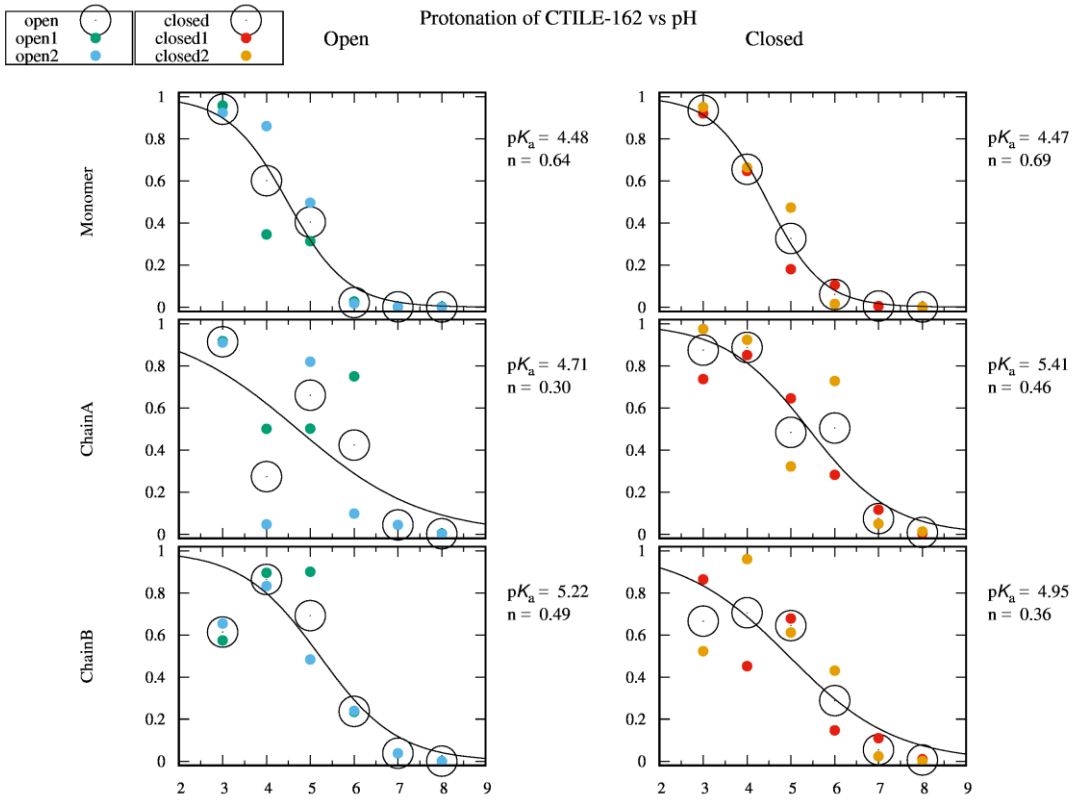


**Figure A4 28** Protonation of Cter as a function of pH.

## A5. PMFs

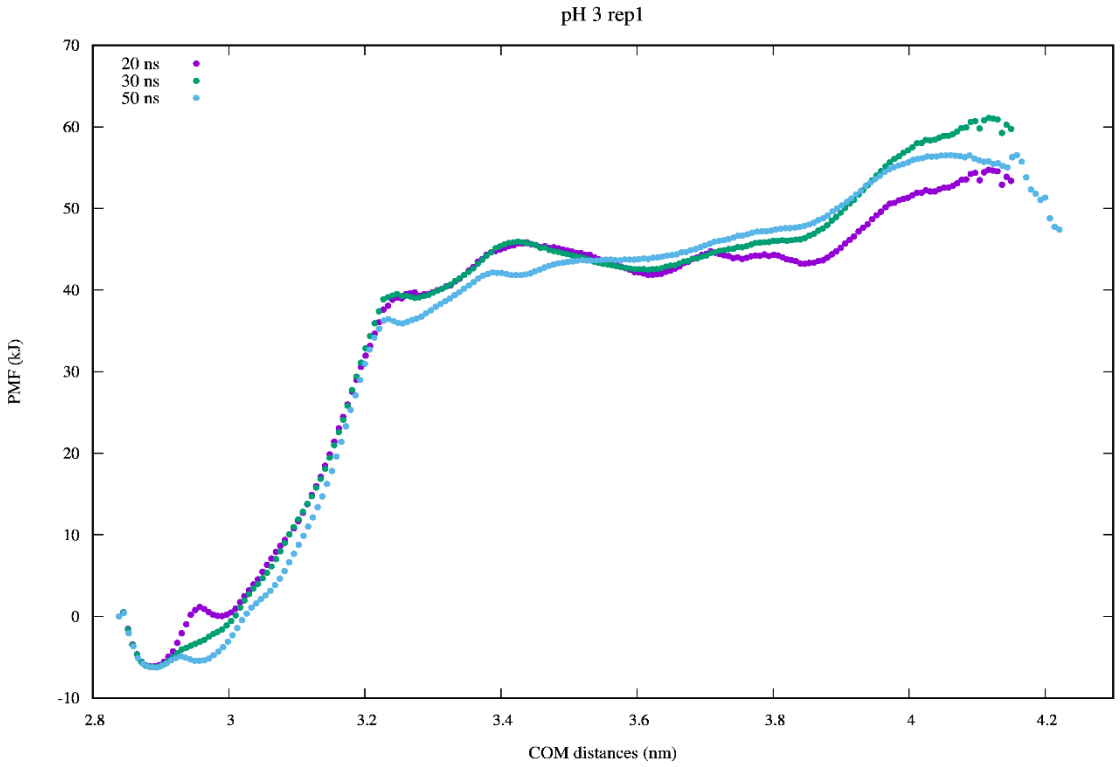


**Figure A5 1** PMF of pH 3, replicate 1.

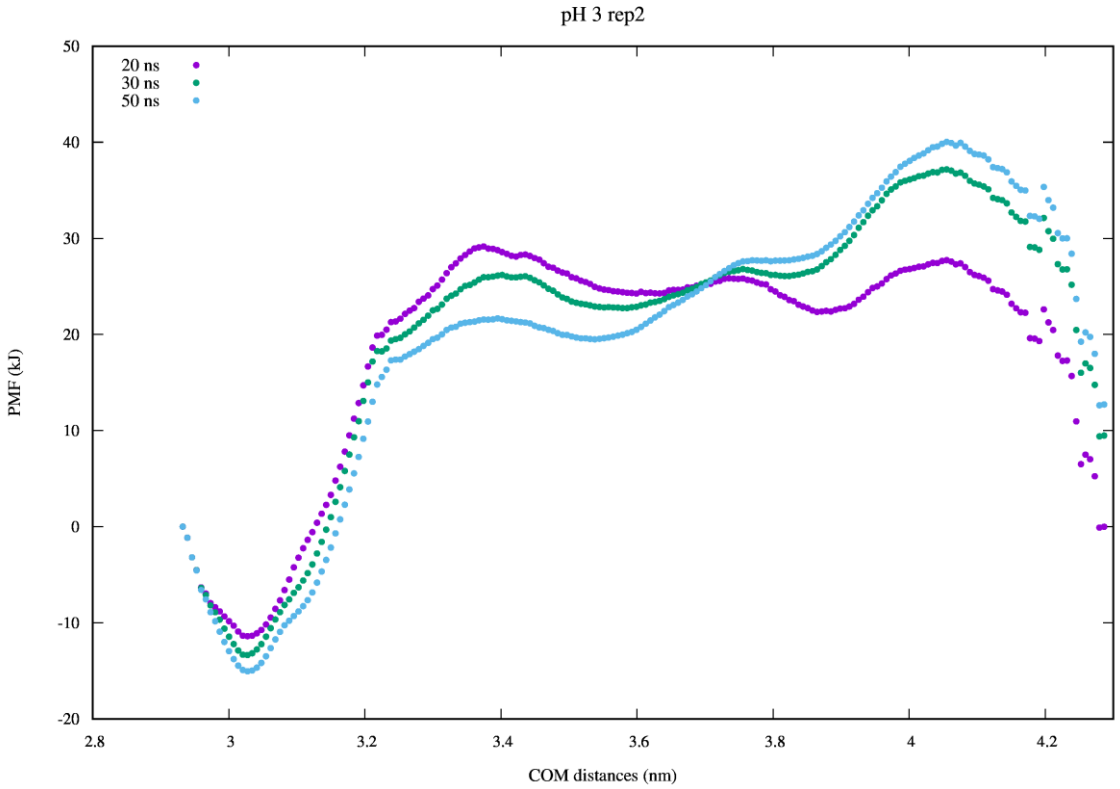


**Figure A5 2** PMF of pH 3, replicate 2.

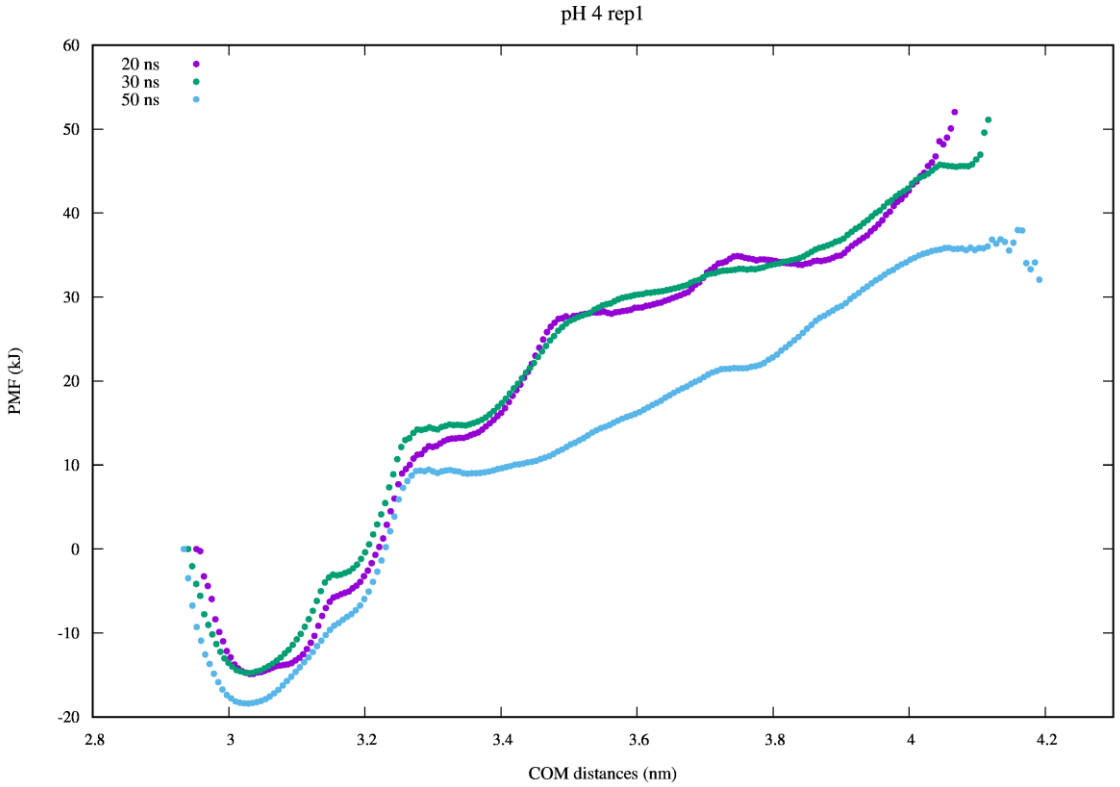


**Figure A5 3** PMF of pH 4, replicate 1.

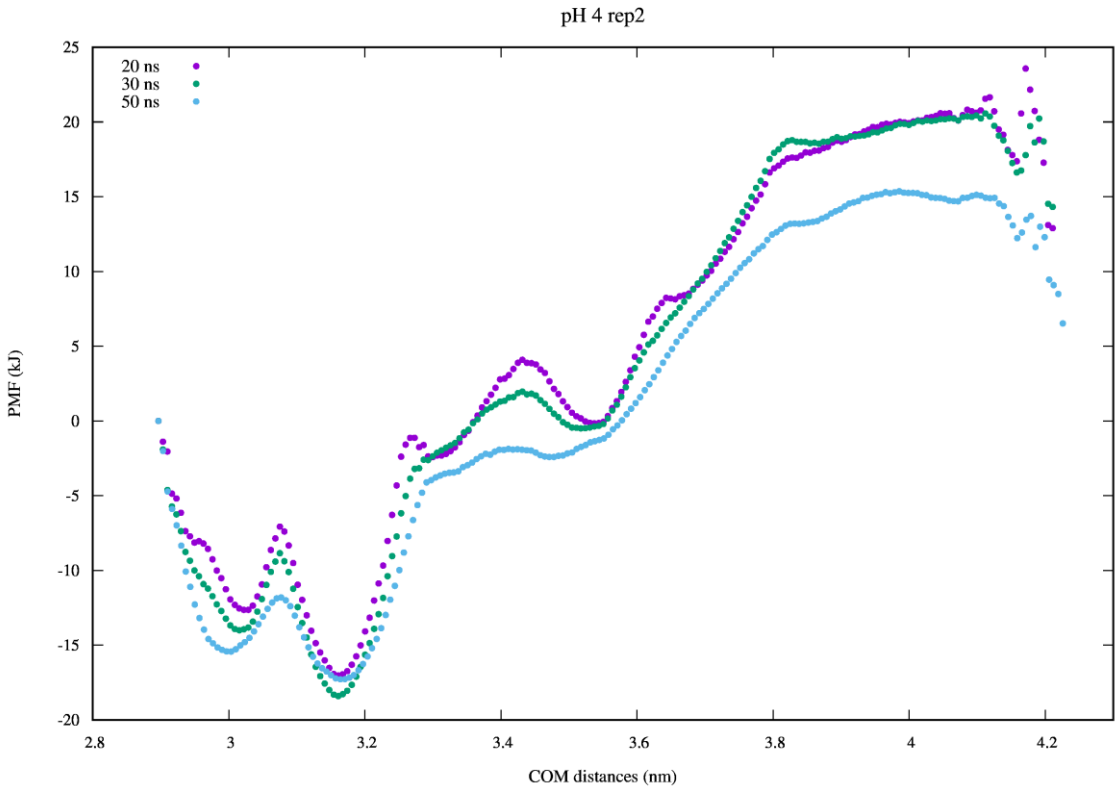


**Figure A5 4** PMF of pH 4, replicate 2.

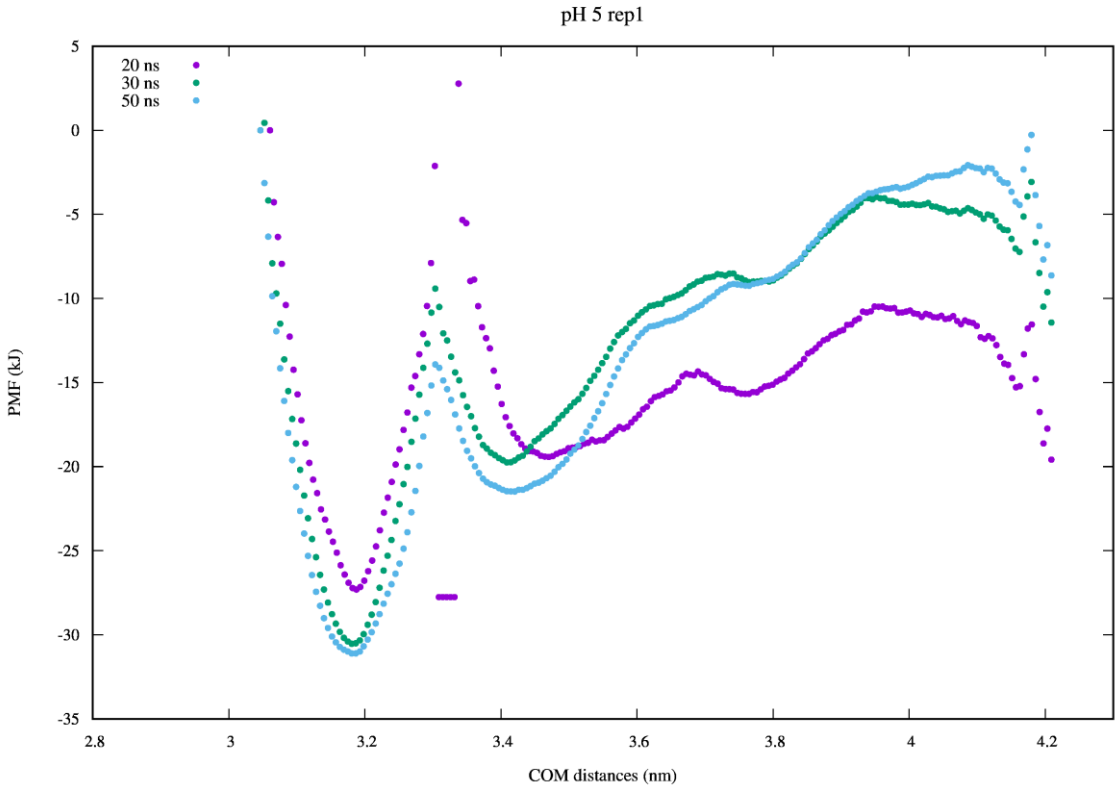


**Figure A5 5** PMF of pH 5, replicate 1.

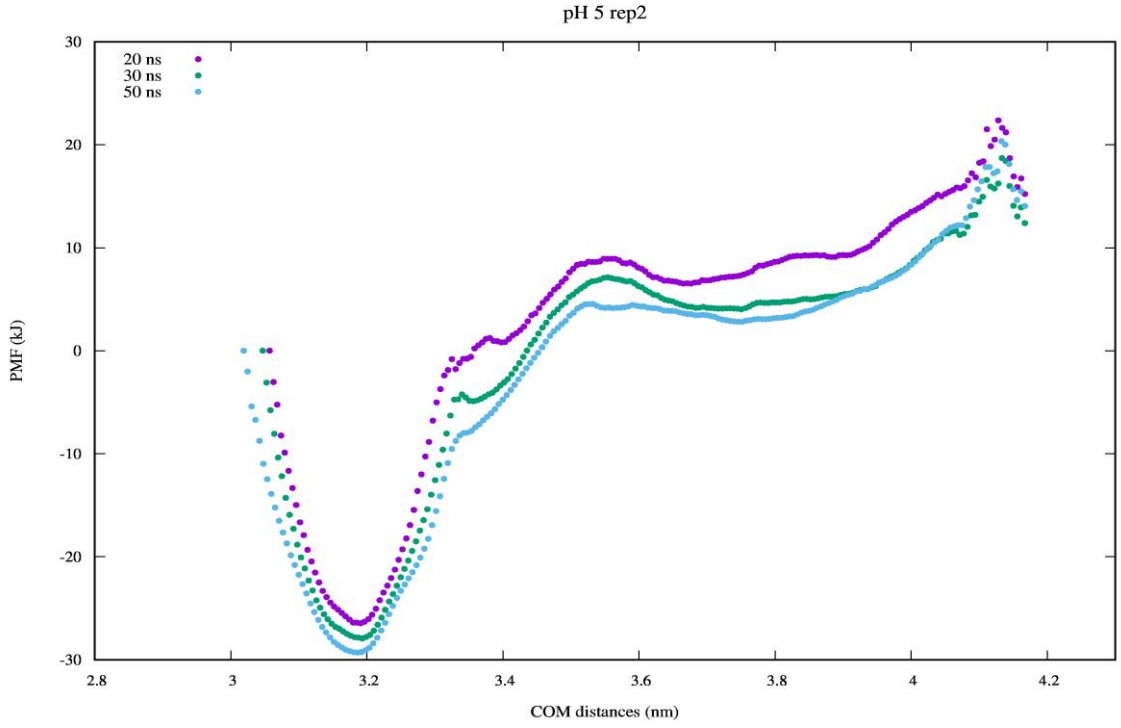


**Figure A5 6** PMF of pH 5, replicate 2.