Modeling the Stability of Protein Solutions and of Hepatitis B Virus-Like Particles

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN (DR. RER. NAT.)

von der KIT-Fakultät für Chemieingenieurwesen und Verfahrenstechnik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION
von

M.Sc. Srđan Pušara

aus Sarajevo

Tag der mündlichen Prüfung:	24.07.2023
Erstgutachter:	Prof. Dr. W. Wenzel
Zweitgutachter:	Prof. Dr. J. Hubbuch

Contents

List of Figures

List of Tables

List of Abbreviations

- **PPI:** Protein-protein interactions
- **MD:** Molecular dynamics
- **DLVO theory:** Derjaguin-Landau-Verwey-Overbeek theory
- **VLP:** Virus-like particle
- **HBV:** Hepatitis B virus
- **Cp:** Core protein of Hepatitis B virus
- **HBcAg:** Core antigen protein of Hepatitis B virus
- **HBsAg:** Core surface protein of Hepatitis B virus
- **C-ter:** C-terminus side of protein
- **xDLVO theory:** Extended DLVO (Derjaguin-Landau-Verwey-Overbeek) theory
- **PB:** Poisson Boltzmann
- **NVE:** Microcanonical ensemble
- **NVT:** Canonical ensemble
- **NPT:** Isobaric-isothermal ensemble
- **PBC:** Periodic boundary conditions
- **CG:** Coarse-grained
- **US:** Umbrella sampling
- **WHAM:** Weighted Histogram Analysis Method
- **LYZ:** Lysozyme
- **Subs:** Subtilisin
- **BSA:** Bovine serum albumin
- **IgG1:** Human immunoglobulin type I
- **LJ:** Lennard-Jones
- **SASA:** Solvent Accessible Surface Area
- **COM:** Center of mass
- **SBCG:** Shape-based coarse-grained
- **BPTI:** Bovine pancreatic trypsin inhibitor

- **RbnA:** Ribonuclease A
- **ChymA:** α-Chymotrypsinogen
- **APBS:** Adaptive Poisson Boltzmann solver
- **BLG:** Beta-lactoglobulin
- **PME:** Particle mesh Ewald
- **AUC:** Analytical ultracentrifugation

1. Introduction

Proteins are complex macromolecules that play a critical role in biology. They are made up of long polypeptide chains of aminoacid residues that fold into a unique three-dimensional structure.[1] This folding is driven by a complex network of van der Waals interactions, electrostatic, steric, and hydrophobic origin. Despite the fact that only twenty naturally occurring amino acids are involved in the production of proteins, they can have many different sizes, shapes, and biological functions, making them the most diverse biological macromolecules. Proteins serve a variety of functions in the body, including acting as structural building blocks, molecular machines for synthesizing new molecules, and nutrient transporters.[2] They are also used in a variety of biotechnology applications, including food, agriculture, and pharmaceuticals.[3, 4] The production of pharmaceutically important recombinant proteins (biopharmaceuticals) has become easier with the advancement of recombinant DNA technology. The therapeutic proteins show promise in a variety of clinical treatments, including vaccines, hormones, growth factors, enzymes, blood factors, cytokines, and anticoagulants.[5, 6] Despite their potential, protein processing remains

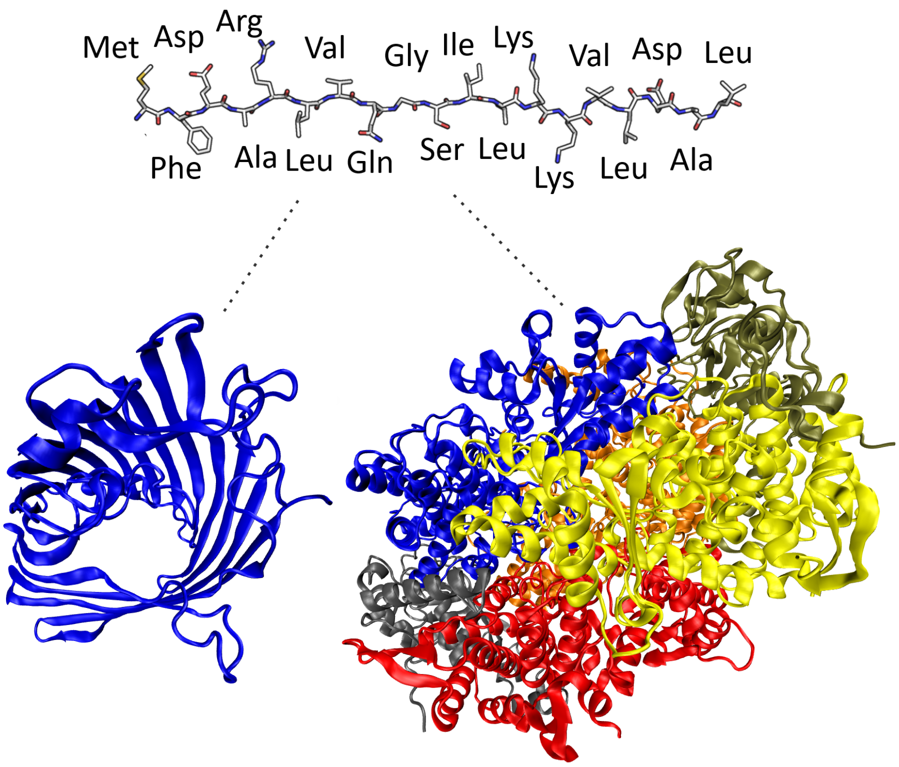


Figure 1.1.: Illustration of the various sizes and shapes that proteins can adopt. Part of this illustration was adapted from a public domain source under the Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license.

a challenge due to their chemical and physical instability. Protein solutions are prone to aggregation or precipitation, which is one of the major processing challenges.[7] The development of new bioprocessing technologies will be critical in overcoming these issues and paving the way for new applications.

Protein properties such as solubility, aggregation, precipitation, and crystallization are primarily driven by protein-protein interactions (PPIs),[8, 9, 10] which are modulated by the nature of proteins as well as solution properties such as ionic strength, pH, and temperature.[11, 12] PPIs can be classified into specific and nonspecific interactions. Weak noncovalent interactions, such as attractive van der Waals and hydrophobic interactions, as well as attractive or repulsive electrostatic interactions, govern nonspecific interactions. The balance of repulsive and attractive forces experienced by macromolecules in solution determines the stability of protein solutions.[11, 12] Specific PPIs are more directional and powerful, regulating processes such as oligomerization, specific recognition, substrate to enzyme binding, and macromolecular self-assembly. Theoretical models can be extremely useful in understanding protein dispersity and phase behavior, as well as in reducing and directing experimental effort to design new processing technologies.

1.1. Modeling Techniques for Understanding Proteins

PPIs are the primary driving force behind many cellular biological processes such as signal transduction, transport, metabolism, and transcription.[13, 14] Understanding PPIs is therefore important from both a fundamental biology standpoint and for designing new therapies once the molecular level of diseases is elucidated. Furthermore, understanding PPIs is necessary for rational design of biotechnology applications and protein processing. As a result, efforts were made to develop experimental techniques for characterising PPIs; however, these techniques are still limited, time consuming, and incapable of providing a complete atomistic level description.[15, 16] The experimental determination of PPIs is further complicated by the fact that proteins can take on a wide range of shapes, posing new challenges and increasing complexity. As a result, modelling of PPIs is critical for bridging the gap between macroscopically measured experimental observables and microscopic structure function relationships.[17] Molecular dynamics (MD) simulations are powerful tools for simulating the movement of biomacromolecules, investigating various processes, and studying protein structure and function.[18, 19] MD simulations frequently employ an all-atom representation of proteins, allowing for a detailed representation of molecular systems and providing insight into the atomistic origins of various macroscopic phenomena. However, the main limitation of MD simulations is that they are computationally quite expensive; as a result, simulated systems are small, and simulations can only be carried out for a few hundred nanoseconds or several microseconds depending on computational power.

Many biochemical processes, on the other hand, are much slower and occur on much longer time scales ranging from milliseconds to seconds. Many coarse-grained force fields have been developed to allow for the simulation of larger systems over longer timescales.[20] These force fields allow for faster simulations by reducing the number of degrees of freedom, but at the expense of reduced accuracy. Furthermore, accelerated MD techniques were developed to simulate rare events that are normally inaccessible to conventional MD.[21] Despite these advances, MD simulations still cannot capture all macroscopic processes. Aside from system size and simulation timeline (typically in range up to 100 ns and 1 ms), another limitation of these techniques is the accuracy of force fields, which are a set of parameter and equations for use in molecular mechanics simulations parametrized on a limited number of examples to reproduce specific experimental values. As a result, applying the force fields to different systems may result in larger error. Because the majority of simulated atoms in MD simulations are solvent molecules, continuum models that average out the effective solvent interaction and model a solvent-like continuum were developed.[22] These models are more computationally efficient while providing comparable accuracy. Using implicit solvent models[22] based on Poisson-Boltzman theory[23, 24, 25, 26] or more simplified Born models are two examples.[27, 28] Furthermore, Brownian dynamics simulations that treat proteins as rigid molecules and implicitly model the solvent were reported.[29]

Proteins, on the other hand, can be modelled by mean field colloidal theories, which treat proteins as simplified uniform shapes (usually spherical) and ignore protein dynamical behavior. The most commonly used model is based on DLVO (Derjaguin, Landau, Verwey, and Overbeek) theory,[30] which models PPIs as a sum of repulsive electric double-layer forces and attractive ven der Waals interactions.[31] These models were successful in explaining some general colloidal properties of protein solutions, but they were ineffective in differentiating specific protein systems due to their omission of structural details.[32, 33] Given the large number of proteins that can participate in a variety of processes, there is no ideal method for modeling PPIs. Multiscale models that take advantage of different levels of theory are frequently required.

1.2. Virus-Like Particles Based on Hepatitis B Core Proteins: A Promising Tool for Therapeutic Applications

Virus-like particles (VLPs) have emerged as a widely accepted technology in recent decades, particularly as vaccines, among many possible protein-based therapeutics.[34, 35] Many are already in use as commercial medical products, while others are in various stages of clinical research. VLPs are nanoparticles formed by self-assembly of viral capsid coat proteins that have high morphological similarity to natural viruses but lack the natural viral genome and are thus noninfectious. Currently, over 110 viral proteins from 35 different viral families have been shown to assemble into VLPs.[36] VLPs can be expressed recombinantly in a wide range of bacterial, yeast, and insect host systems. Because of their inherent immunogenicity and ability to encapsulate therapeutic nucleic acids, VLPs are also excellent candidates for next-generation vaccines or nanocarriers for targeted drug delivery.[37, 35, 38, 39, 40] VLPs can be engineered to efficiently target specific tissues, penetrate cells, and deliver therapeutic agents at the desired site of action while requiring much lower doses than traditional oral therapies would require. VLPs have the potential to replace currently used liposomal or polymer-based nanoparticles for targeted drug delivery due to their low toxicity.[35] Furthermore, VLPs are being researched for use in gene therapies to treat cancer or genetic disorders such as mitochondrial disorders and Parkinson's disease. Chemotherapeutic drugs, siRNA, RNA aptamers, proteins, and peptides can all be packed and delivered by VLPs.[41, 42, 43, 44] Furthermore, VLP properties can be further tailored by introducing surface modifications or inserting epitope sequences, providing additional design flexibility in the pursuit of new therapies.[35]

VLPs based on hepatitis B virus (HBV) core proteins are especially good candidates for next-generation VLP therapeutics agents because they can self-assemble in almost all expression systems.[45] HBV is a small DNA virus from the Hepadnaviridae family, with an internal protein capsid composed of core proteins (Cp) and a lipid envelope containing other types of proteins.[45] HBV infection continues to be a major cause of transient and chronic liver disease worldwide, with over 360 million people chronically infected and one million deaths per year.[46] A virus capsid is a protein shell made up of viral structural proteins that encapsulates and protects a virus's genome. Two different types of HBV based VLPs can be produced from viral proteins: using the core antigen protein (HBcAg, also known as Cp) which form the internal capsid, or using the surface antigen (HBsAg), which requires lipids to spontaneously assemble into nanoparticles, with the first being the more common. Several HbsAg VLP-based vaccines, including Recombivax HB and Engerix-B, have been approved for use in humans to treat hepatitis HBV infection.[45] Aside from that, HBV-based VLPs are being intensively investigated as promising nucleic acid nanocarriers for use in nucleic acid-based therapeutics.[45, 35, 47] The mature HBV capsid is made up of 120 dimers of Cp that form an icosahedral structure with T = 4 quasi-symmetry, while a smaller fraction of capsids (5%) have T = 3 symmetry, which is formed by 90 copies of dimers.[45, 48] These two capsids have overall diameters of 36 nm and 32 nm, respectively, as shown on Figure 1.2. The core Cp protein has 183 amino acids and exists in solution as a homodimer.[49] It is made up of a rigid assembly domain (1-140 aa) and a flexible C-terminal region (150-183 aa).[50] A hinge peptide at 141-149 aa links these domains and it can perform morphogenic functions.[45, 51] The C-terminal (C-ter) domain contains multiple positively charged arginine residues that allow negatively charged nucleic acids to bind.[52] Unlike the assembly domain, which has a stable tertiary structure and is responsible for the capsid's surface charge, the C-ter has no defined tertiary structure, making it extremely flexible.[50] During the HBV life cycle, it is mostly found in the capsid interior, but it can also be found on the capsid's outer surface penetrating through pores formed by the pentamers of Cp dimers.[53] Cp dimers can self-assemble in vitro, resulting in capsids with indistinguishable morphology from natural capsids. Full length Cp183 dimers can assemble at higher ionic strengths (>=0.25 M), which are required to overcome the strong electrostatic repulsion caused by the positively charged C-ter domain.[47] Apart from natural Cp183 dimers, truncated proteins can also be assembled, as long as the truncation is less than 140 aa. The distribution of T = 3 and T = 4 isomorphic capsid structures is influenced by the truncation point, with longer proteins favouring the T = 4 configuration.[45, 52] Cp149 dimers truncated at residue 149 form only 5% of T=3 capsids, whereas Cp140 capsid proteins form 85% of T=3 capsids. Further C-terminal residue deletions inhibit capsid protein assembly, and Cp139 dimers are unable to aggregate into VLPs. In vitro assembly begins with a slow nucleation step (formation of a trimer of dimers), followed by a rapid elongation phase.[54, 49] Weak hydrophobic interactions at dimer-dimer interfaces drive assembly, which is balanced by electrostatic repulsion.[55, 56] These weak interactions reduce the possibility of mis-assembly, because they allow incorrectly bound subunits to dissociate and reassemble correctly.[57, 58, 59] The assembly is allosterically controlled, and its efficiency is typically modulated by changes in solution conditions such as pH, temperature, or increasing ionic strength, which reduces electrostatic repulsions between dimers (and possibly also induces conformational change of Cp into assembly active state). 5.1. Recombinantly expressing Cp in host cells (for example, Escherichia Coli), followed by cell lysis and purification, is a common method for producing HBV VLPs.[60, 61, 62] After being expressed in the host cell, Cp_2 dimers spontaneously assemble into capsids encapsulating random nucleic acids from host cells.[63, 60] Thus, it is critical to obtain purified dimers that are free of host nucleic acids and other impurities. Precipitation and disassembly are two key components of the purification stages that are essential for removing encapsulated contaminants and enhancing structural integrity. To encapsulate the therapeutic cargo, purified Cp_2 dimers are mixed with therapeutic nucleic acids and then reassembled into capsids.[47, 64, 60] VLPs made of wild-type Cp183 proteins, as well as variants with a few replaced amino acids, have previously been demonstrated.[65, 66, 67] Purification of these systems, however, is difficult due to the low solubility of the Cp183 construct.[47] As a result, HBV VLPs made of Cp with shorter nucleic acid binding regions are being studied for effective nucleic acid loading.[65, 66] It has been reported that different nucleic acid binding lengths can affect HBcAg VLP capsid stability, as well as the phase behavior and purification process of HBcAg VLPs. Factors influencing the purification process, such as encapsulated nucleic acids or general Cp protein properties, are critical to the effective large-scale production of VLPs for nucleic acid delivery. For this, protein-protein and protein-nucleic acid interactions, which influence the disassembly and assembly

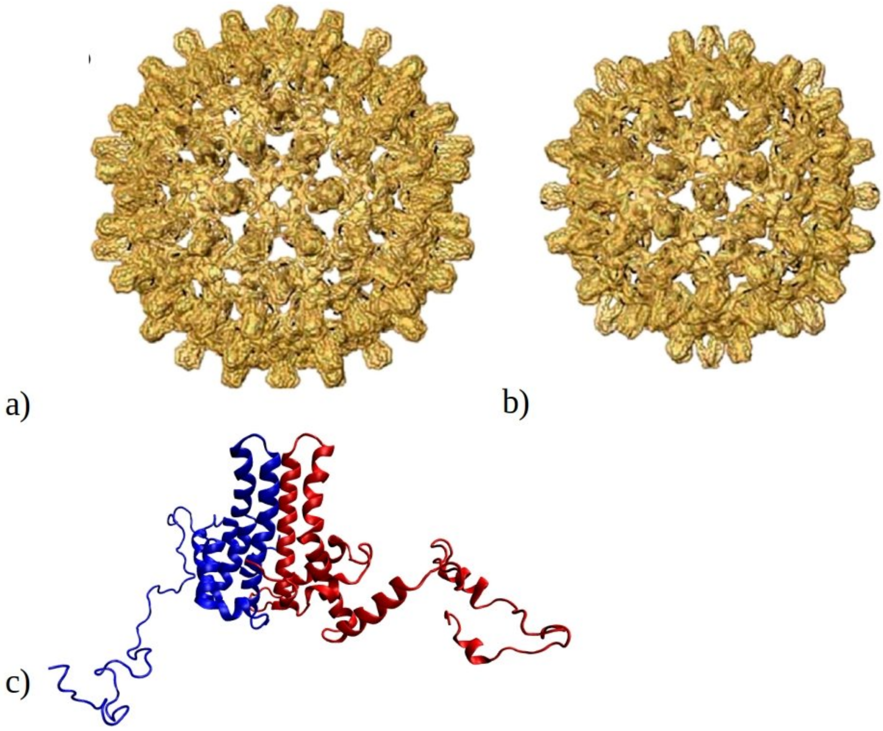


Figure 1.2.: Structure of HBV capsid of a) T=4 and b) T=3 symmetry, along with c) structure of core protein dimer

processes during the VLP preparation, have to be better understood.

1.3. Motivation and Thesis Outline

Given the broad range of processes in which proteins can participate, there is no universally ideal method for simulating these processes and modeling the protein interactions that drive them. The primary objective of this thesis is to develop and apply computational models to investigate protein-protein interactions and their impact on the stability of protein solutions. Specifically, we aim to explore factors that contribute to the stability of a wide range of proteins, as well as to analyze the stability of virus-like particles composed of the core proteins of hepatitis B viruses. By simulating these systems and analyzing the resulting data, we aim to better understand the underlying mechanisms governing protein stability and provide insights into strategies for enhancing protein stability in biopharmaceutical applications. To improve computational efficiency, we use coarse-grained representations of proteins and a range of theoretical levels, from mean-field continuum theories to advanced techniques like accelerated molecular dynamics with free energy calculations. Our approach aims to accurately capture the complex behavior of protein-protein interactions and elucidate the underlying mechanisms governing protein stability. This research contributes to the development of more efficient and effective computational tools for investigating protein systems, with potential applications in the biopharmaceutical industry and beyond.

The structure of this thesis is as follows. Chapter 2 provides a concise overview and literature review of the theory that serves as the foundation for the development of new models (outlined in Chapters 3 and 4) and for the recapturing of existing methods used to study protein dimerization and assembly processes. While not an exhaustive review of all aspects of the theory, Chapter 2 offers a comprehensive explanation of the fundamental concepts required to understand the results presented in this thesis. It provides readers with a clear understanding of the theoretical framework employed in our research and its significance in elucidating the underlying mechanisms governing protein stability.

In Chapter 3, we introduce a novel developed computational model called the xDLVO-CG model, which predicts the stability of protein solutions and the dependence on pH of solution and salt concentration by calculating second osmotic virial coefficients. The second virial coefficients serve as a measure of the effective interaction between two proteins in solution, and the xDLVO-CG model is based on the extended DLVO theory, which includes a new term for ion-protein dispersion interactions. To model protein interactions, the xDLVO-CG model uses a coarse-grained representation that is shape-based, allowing for the consideration of anisotropic protein-protein interactions. By employing a shape-based coarse-grained representation, the model eliminates or reduces the need for fitting experimental data. We have validated the xDLVO-CG model with experimental data for several benchmark proteins and demonstrated its potential for predicting the stability of a wide range of proteins. Moreover, we have used the model to compute osmotic second virial coefficients for hepatitis B virus core protein dimers, providing insights into the stability and phase behavior of these complexes under different conditions.

Chapter 4 introduced an enhanced model for predicting protein solution stability called the xDLVO-CGhybr model. This new model builds upon the xDLVO-CG model from the Chapter 3 by incorporating a hybrid approach to calculate the electrostatic potential of mean force. At short protein separations, the Poisson-Boltzmann theory is applied to all-atom structures of proteins, while at larger separations, the Debye-Hückel theory is applied to coarse-grained structures for accurate calculation. Additionally, a coarse-grained Lennard-Jones potential was introduced and parametrized from all-atom potentials. This improved model was tested on six different proteins and demonstrated enhanced accuracy in predicting second osmotic virial coefficients when compared to the xDLVO-CG model.

Chapter 5 of the thesis explores the molecular dynamics of beta-lactoglobulin proteins and fragments of hepatitis B capsids using an accelerated technique called umbrella sampling. The study aims to investigate the free energy of beta-lactoglobulin dimerization and the free energy of dissociation of trimers of dimers of hepatitis B core proteins. To increase computational efficiency, a coarse-grained SIRAH force field is utilized, which retains the positions of backbone atoms while preserving high structural details of protein structure. The study investigates a delicate dependence of the monomer-dimer equilibrium of beta-lactoglobulin on the pH and ionic strength of the solution. Furthermore, our research on hepatitis B capsids focuses on the stability of trimers of dimers through the attachment of DNA to core proteins. We aim to investigate the dependency of the free energies of trimer binding on the length of the C-terminus side of the core protein. Additionally, we examine the influence of the attachment of nucleic acid molecules as an additional stabilizing factor for the trimers of dimers. This research has implications for understanding the main factors contributing to stability and processability of VLP capsids.

Chapter 6 provides a summary of the key outcomes of this thesis, presenting the main findings and conclusions drawn from the research. The chapter also highlights the strengths and successes of the theoretical tools used throughout this work, including the xDLVO-CG and xDLVO-CGhybr models, and the SIRAH force field in combination with umbrella sampling. Furthermore, the chapter identifies areas that require further improvement and investigation can be made to enhance the accuracy and applicability of the models.

2. Theoretical Background

2.1. The Second Osmotic Virial Coefficient: A Measure of Protein-Protein Interactions

The effective intermolecular interactions that occur between protein molecules determine protein solubility and propensity to aggregate in a solution. Therefore, it is necessary to have theoretical tools that can estimate these interactions and compare them to experimental observations. The second osmotic virial coefficient B_{22}, which is defined by the virial equation of state, is a widely-used thermodynamic quantity for estimating protein-protein interactions.[68, 69, 70] Specifically, B_{22} defines the difference between the osmotic pressure of a solution and that of an ideal solution, and it can be expressed using the following equation:

$$\Pi = RTc_p(\frac{1}{M_W} + B_{22}c_p + B_3c_p^2 + ...) \tag{2.1}$$

where Π is the osmotic pressure, c_p denotes the protein concentration (in mass units), R the gas constant, T the absolute temperature, and M_W the molecular weight of the protein. The equation 2.1 offers a molecular interpretation of B_{22}, in which positive values of B_{22} correspond to repulsive protein interactions. This leads to osmotic pressure higher than that of an ideal gas. Conversely, if B_{22} is negative, the overall interactions are attractive.

When two solutions with differing solute concentrations are separated by a semipermeable membrane, the solvent molecules will tend to diffuse spontaneously from the region of higher solvent potential (i.e., an area with lower solute concentration) to the region of lower solvent potential, in order to balance out the solute concentrations on both sides of the membrane. The osmotic pressure can be defined as the minimum pressure required to prevent the pure solvent from spontaneously crossing a semipermeable membrane and flowing into the solution when the two are in contact. The equation that calculates the osmotic pressure of a solution can be derived in analogy to the ideal gas equation of state, which was originally developed for monoatomic gases and assumes that intermolecular potential is negligible.[71] In reality, most solutions do not behave ideally, especially at higher solute concentrations where solute-solute interactions become significant and cannot be disregarded, thus osmotic pressure is modelled by virial expansion.[71]

McMillan-Mayer used statistical thermodynamics concepts to derive an exact relationship between intermolecular potential and osmotic pressure virial expansion. Specifically, by integrating out the solvent degrees of freedom, the grand canonical partition function for a solution was reduced to one with an effectively solute-only form.[71, 69] In the effective solute grand partition function, the total effective solute potential can be decomposed into components that are mean force potentials for isolated groups of one, two, three, and so on solute molecules.[71] This allowed for an expansion of osmotic pressure in powers of solute concentration, analogous to the virial expansion of gas pressure in powers of density. The relationship between the effective interaction potential between two molecules in solution, W_{22} and the osmotic second virial coefficient can be expressed using the following formula:

$$B_{22} = \frac{1}{2}\frac{N_\mathrm{A}}{M_W^2}\int_0^\infty (1 - e^{-\frac{W_{22}(r)}{k_BT}})4\pi r^2 dr \tag{2.2}$$

where r is the intermolecular distance between two molecules, N_A is the Avogadro constant and k_B is the Boltzmann constant. The equation 2.2 shows that if the intermolecular potential $W_{22}(r)$ is known, it is possible to compute the second virial coefficient as a function of temperature. The function f(r), which appears within the brackets of equation 2.2, is commonly referred to as the Mayer f-function, named after Mayer who first recognized its significance.

Higher virial *Nth* virial coefficients can be calculated by analogy using the interactions of N molecules in a volume V. For example, the third virial coefficient B_3 describes the system's effective three-body interactions. Thus, the nonideal N-body problem can be reduced to a series of one-body, two-body, three-body problems, and so on. Only the second virial coefficient influences the total deviation of osmotic pressure in diluted protein solution, so higher body interactions can be ignored.[71]

Although initially derived for globular molecules, the equation 2.2 can be extended to encompass molecules of arbitrary shape, regardless of whether they are spherical or not,[71, 69] according to:

$$B_{22} = \frac{1}{2}\frac{N_\mathrm{A}}{M^2}\int_{\Omega_1}\int_{\Omega_1}\int_0^\infty (1 - e^{-\frac{W_{22}(r,\Omega_1,d\Omega_2)}{k_BT}})4\pi r^2 drd\Omega_1 d\Omega_2 \tag{2.3}$$

$W_{22}(r, \Omega_1, \Omega_2)$ represents the effective interaction potential (potential of mean force, PMF) between two proteins as a function of intermolecular center-of-mass distance r and relative orientation Ω_1 and Ω_2. Equation 2.3 takes into account the PMF for all possible orientations of the proteins with respect to each other, including the angular variables and intermolecular distances.

2.2. Fundamentals of DLVO Theory

DLVO theory, developed in the 1940s by Derjaguin and Landau and by Verwey and Overbeek,[30, 72, 31] explains the stability of colloidal dispersions through an interplay between attractive van der Waals forces $W_{vdW}(r)$ and repulsive electrostatic forces $W_{el}(r)$.[31] The van der Waals force is attractive and depends on the distance between the particles, while the electrostatic force is repulsive and depends on the surface charge of the particles and the ionic strength of the surrounding medium. DLVO theory assumes that the electrostatic double layer forces and the van der Waals forces are independent and can be simply summed to obtain the total interparticle forces between two colloidal particles, as described by the following equation:

$$W_{22}(r) = W_{vdW}(r) + W_{el}(r) \tag{2.4}$$

The DLVO theory was successful in explaining an essential physics of colloidal phenomena, including coagulation, flocculation, and surface tension for colloidal particles of various shapes and sizes.[72] While DLVO theory has been successful in describing the stability of colloidal systems at intermediate and large interparticle distances, it becomes less accurate at shorter separations due to the presence of non-DLVO forces, such as hydration forces, depletion forces and so on.[32, 72] Moreover, DLVO theory does not account for the effects of ion-specific interactions, such as ion pairing or Hofmeister effects, which can significantly influence the stability and behavior of colloidal systems.[33]

DLVO theory has also been used to model protein interactions and calculate second osmotic virial coefficients, where proteins are typically represented as ideal spheres. B_{22} in aqueous protein solutions can be calculated using the DLVO model as a function of pH, salt type, salt concentration, and temperature.[**?**] The DLVO model, however, has several limitations, such as proteins being represented by a simplified spherical shape which limits its ability to model proteins with more complex shapes. Moreover, PPIs can involve other types of interactions beyond electrostatic repulsion and van der Waals attraction. As a result, several extended DLVO models with additional interactions have been reported.[73, 74]

2.2.1. Extension to xDLVO Theory

The most commonly used model for calculating B_{22} coefficients of proteins is the extended DLVO (xDLVO) model,[**?**, 75] which includes an additional term for osmotic attraction depletion. In particular, the xDLVO model incorporates a hard sphere potential $W_{hs}(r)$, a dispersion potential $W_{disp}(r)$, an electrostatic potential $W_{el}(r)$, and an osmotic attraction potential $W_{osm}(r)$:

$$W_{22}(r) = W_{hs}(r) + W_{disp}(r) + W_{el}(r) + W_{osm}(r) \tag{2.5}$$

The hard sphere potential is used to describe repulsive forces between proteins at short distances due to their excluded volume:

$$W_{HS}(r) = \begin{cases} 0, \ r > 2(R_p + \sigma) \\ \infty, \ r \leq 2(R_p + \sigma), \end{cases} \tag{2.6}$$

where R_p is the protein's spherical radius and σ is the thickness of the water layer surrounding the protein surface (estimated as 0.1 nm).

The dispersion potential, which describes the attractive forces resulting from electromagnetic quantum fluctuations between molecules, has been calculated using the Hamaker formula.[76] This formula is derived by the integration of the attractive component of the Lennard-Jones potential, which is also referred to as London dispersion forces, between two homogeneous spheres:

$$W_{disp}(r) = -\frac{A_H}{12}\left[\frac{(2R_p)^2}{r^2-(2R_p)^2} + \frac{(2R_p)^2}{r^2} + 2\ln(1 - \frac{(2R_p)^2}{r^2}\right], \quad r > 2(R_p + \sigma) \tag{2.7}$$

Here, A_H is the Hamaker constant for the system of proteins, solvent, and salt, which is usually determined by fitting from experimentally measured B_{22} values and, in some cases, theoretical models. Despite its widespread use, the Hamaker formula has some limitations, particularly in systems where the geometry and composition of the particles are complex or heterogeneous. In these cases, more sophisticated theoretical and computational methods may be required to accurately predict the dispersion forces between the particles. Nonetheless, the Hamaker formula remains a valuable tool for understanding the fundamental physics of intermolecular interactions in a wide range of physical and biological systems.

Electrostatic interactions are particularly important in colloidal systems, and the stability of such systems can be altered by adding salt, which modifies the strength of the electrostatic interactions. When charged particles are placed in a solution containing electrolytes, they are surrounded by a diffuse cloud of counter-ions and co-ions, known as the electrical double layer. This double layer plays a crucial role in determining the strength and range of the electrostatic interactions between the particles, as shown in Figure 2.2. As similarly charged particles approach each other, their double layers begin to overlap, resulting in a net repulsive force that prevents their contact. The magnitude of this repulsion depends on the surface charge density and the ionic strength of the surrounding medium.

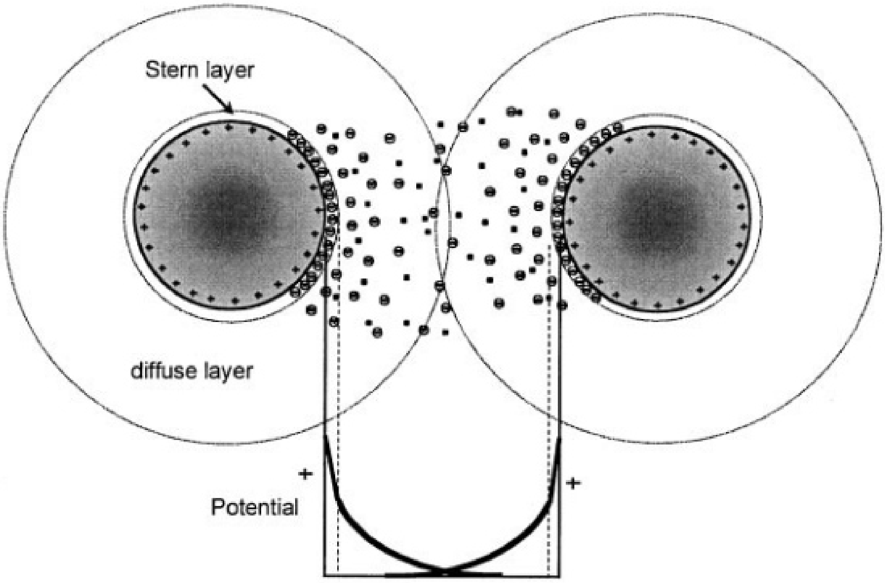


Figure 2.1.: Illustration of the repulsion between two particles caused by the overlap of their electrical double layers. As similarly charged particles approach each other, their double layers begin to overlap, resulting in a net repulsive force that prevents their contact. The magnitude of the repulsion depends on the surface charge density and the ionic strength of the surrounding medium. Reprinted with permission from [72]. Copyright 2001 Elsevier Science Ltd.

Electrostatic interactions in the xDLVO model are calculated using Debye-Hückel theory, which accounts for electrostatic potential screening in the presence of electrolytes:[77]

$$W_{el}(r) = \frac{Z^2 e^2 \exp(\kappa(2R_p - r))}{4\pi\varepsilon_0\varepsilon_r r(1+\frac{\kappa R_p}{2})^2}, \quad r > 2(R_p + \sigma), \tag{2.8}$$

where ϵ_r is relative permittivity, Z is protein charge, and κ is the inverse Debye length given by:

$$\kappa = \sqrt{\frac{2N_A e^2 I}{\varepsilon_0\varepsilon_r k_B T}} \tag{2.9}$$

Debye-Hückel theory is a mathematical model that describes the screening of electrostatic potentials in the presence of electrolytes. This theory is based on the assumption that the ions in solution are point charges and the resulting electrostatic potential is computed solving the linear Poisson Boltzmann equation, which takes into account the effects of ionic concentration, temperature, and solvent dielectric constant. Debye-Hückel theory allows the calculation of the electrostatic potential in the electrical double layer as a function of distance from the particle surface, taking into account the concentration and valence of the electrolyte ions. The Debye length is a measure of the thickness of the electrical double layer that forms around colloidal particles in a solution, and it determines the range over which the electrostatic interaction has effect. It is defined as the distance over which the electrostatic potential decreases by a factor of e from its value at the particle surface. Limitation of Debye-Hückel theory theory is that the equation 2.8 cannot accurately account for the effect of electrolytes at ionic strengths greater than 0.1 M.

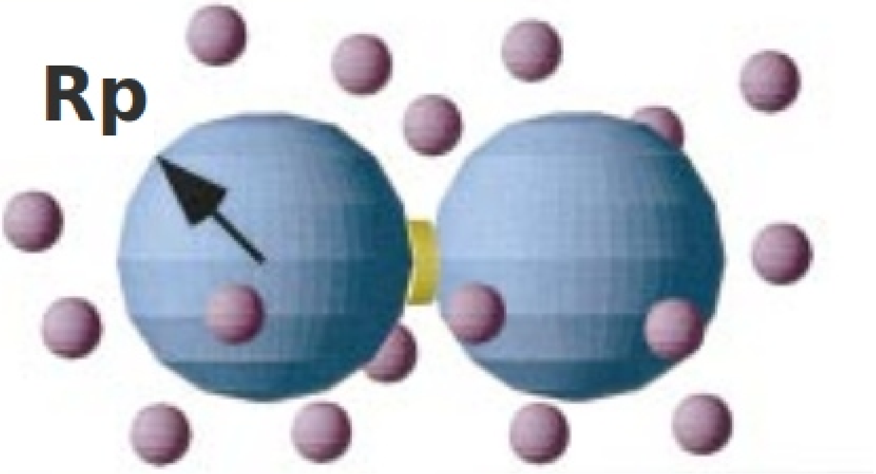


Figure 2.2.: Schematic representation of the attractive osmotic depletion potential. The large blue spheres represent proteins, while the small purple spheres represent salt ions. The yellow region represents the excluded volume between the proteins, where ions are excluded due to steric hindrance. The depletion force arises due to the resulting decrease in ion concentration in this region, which leads to a local difference in osmotic pressure and a corresponding attractive force between the proteins. Adapted with permission from [78]. Copyright 2002 American Physical Society.

The osmotic depletion potential is a non-electrostatic and non-van der Waals force that arises in solutions containing non-adsorbing polymers or other molecules. The depletion force was first theoretically described by Asakura and Oosawa in 1954,[79] and it is particularly strong when colloidal particles or proteins are immersed in polymer solutions. This effect is often used to induce flocculation or phase separation of colloid particles or proteins. In protein solutions with higher ionic strengths, the osmotic depletion potential can become significant, making its inclusion in the extended DLVO theory essential. When two protein surfaces are brought closer together, the gap between them can become smaller than the diameter of salt ions, leaving only space for solvent molecules in between. This reduction in available space results in a decrease in configurational entropy, causing salt ions to be excluded. As a result, an osmotic pressure gradient is created between the interstitial space and the surrounding solution, causing colloidal particles to be pushed into contact with one another. This phenomenon is illustrated in Figure 2.2. The osmotic attraction potential can be calculated using the following formula:[79]

$$W_{osm}(r) = -\frac{4\pi k_B}{3} T r_{23}^3 \rho_3 \left(1 - \frac{3r}{4r_{23}} + \frac{r^3}{16 r_{23}^3}\right), \quad 2(R_p + \sigma) \leq r \leq 2r_{23} \tag{2.10}$$

where $r_{23} = R_p + R_3 + \sigma$ is the sum of mean hydrated radius of a protein R_p, the salt R_3 and the water layer σ, while ρ_3 is the salt density. This force has a magnitude on the order of the osmotic pressure of the macromolecule solution and a range on the order of the diameter of the macromolecules. The strength and range of this force depend on salt concentration and ion size. The range of the force increases with the size of non-adsorbed species (ions or molecules), which is why the effect is particularly strong in polymer solutions.

Furthermore, various modifications to the xDLVO model were introduced in order to model interactions in protein solutions containing mixtures of more than one excipient or in the presence of polymer (by adding PRISM potential to total potential of mean force which describes equilibrium properties of polymer solutions).[73, 74] The xDLVO model is capable of modeling and predicting second osmotic coefficients for a variety of protein systems, as well as fitting experimental data and extracting Hamaker constants and other parameters.[3] This model, however, models protein as a a simple sphere, whereas many proteins have irregular non-globular shapes. Finally, another significant limitation of the xDLVO theory is that it does not account for ion specificity, which is known to influence protein interactions in solution (i.e. Hofmeister series).[33, 80]

2.3. Poisson-Boltzman theory

The Poisson-Boltzmann (PB) equation is used for calculating electrostatic potential in a biomolecular system by using a continuum model of the surrounding solvent and counterion environment.[25, 26, 81] It takes into account the shape and charge distribution of the protein and is used for various purposes, including biomolecular structural analysis, pKa calculations, and modeling of processes where global electrostatic properties are crucial.[23, 82, 83, 84] The full PB equation can be derived starting from the Poisson equation, which is used to calculate the electric potential for a given charge distribution:[23]

$$\Delta \cdot \epsilon(x)\Delta\phi(x) = \rho(x) \ , \quad x \in \Omega, \quad where \quad \phi(x) = g(x) \quad for \quad x \in \partial\Omega \tag{2.11}$$

where $\phi(x)$ denotes the dimensionless electrostatic potential produced by a charge distribution $\rho(x)$ in a polarizable continuum with a dielectric constant $\epsilon(x)$. This equation is solved in a finite domain Ω with Dirichlet boundary conditions, where a fixed potential $g(x)$ is applied on the boundary $\partial\Omega$.[81, 23]

In a biomolecular system, there are two types of charges: fixed charges $\rho_f(x)$, which are associated with proteins, and mobile charges $\rho_m(x)$, representing the counterions found in the surrounding electrolyte. Interior of protein is described by a fixed charge distribution, $\rho_f(x)$, where M partial atomic charges are represented by delta functions:

$$\rho_f(x) = \frac{4\pi e_c^2}{k_B T} \sum_{i=1}^{M} Q_i \delta(x - x_i) \tag{2.12}$$

where x_i and Q_i are the coordinates and partial charges of each atom, respectively.[23]

The mobile counterion charges $\rho_m(x)$, which surround the protein in a continuous manner, are modeled using a Boltzmann distribution:

$$\rho_m(x) = \frac{4\pi e_c^2}{k_B T} \sum_{j}^{m} c_j q_j e^{-[q_j\phi(x) + V_j(x)]} \tag{2.13}$$

where m is the number of counterion species, c_j and q_j are their bulk concentrations and charges, and V_j is the steric potential preventing steric clash with protein.[24]

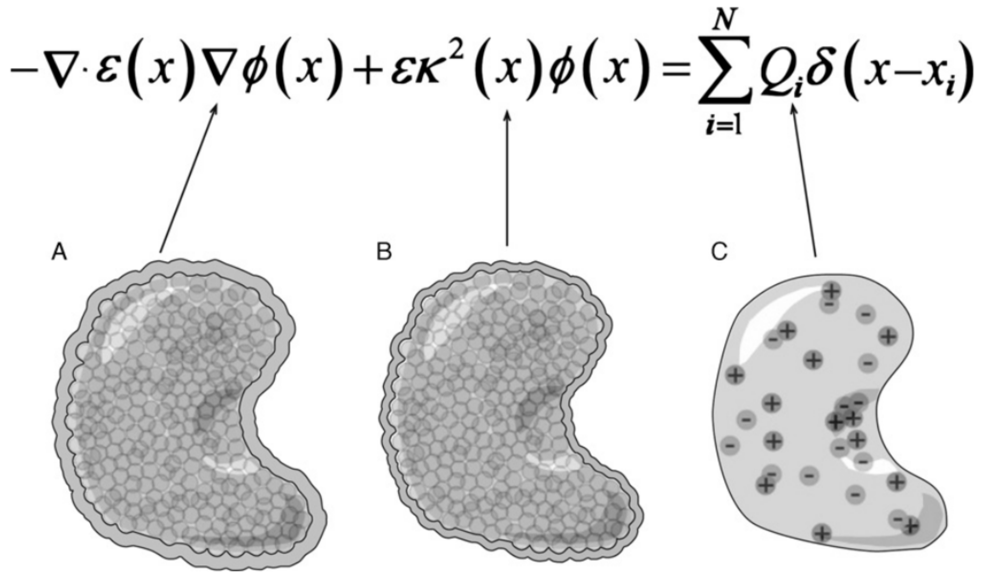
For a 1:1 electrolyte, the equation 2.13 can be simplified to:

$$\rho_m(x) = \overline{\kappa}^2 \sinh\phi(x) \tag{2.14}$$

where $\overline{\kappa}$ is a coefficient that describes ion accessibility and is affected by ionic strength.[23]

When equations 2.12 and 2.13, which describe the charge distribution of proteins and counterions, are incorporated into the Poisson equation 2.11, the resulting equation becomes a complete nonlinear Poisson-Boltzmann equation:

$$\begin{gathered} -\Delta \cdot \epsilon(x)\Delta\phi(x) + \overline{\kappa}^2 \sinh\phi(x) = \frac{4\pi e_c^2}{k_B T} \sum_{i=1}^{M} q_i \delta(x - x_i) \ , \ \text{for} \quad x \in \Omega, \\ \text{where} \quad \phi(x) = g(x) \ \text{for} \ x \in \partial\Omega \end{gathered} \tag{2.15}$$


$-\nabla\cdot\varepsilon(x)\nabla\phi(x)+\varepsilon\kappa^2(x)\phi(x)=\sum_{i=1}^{N}Q_i\delta(x-x_i)$
A
B
C

11

Figure 2.3.: Illustration of key terms in the Poisson–Boltzmann equation. (A) Dielectric permittivity ($\epsilon(x)$) exhibits a sharp transition across the solvent-accessible surface, with the interior of the biomolecule having a significantly lower permittivity than the surrounding environment. (B) The ion-accessibility parameter ($\kappa(x)$) correlates with the bulk ionic strength outside the ion-accessible biomolecular surface. (C) Biomolecular charge distribution is represented by point charges located at the center of each atom within the biomolecule. Reprinted with permission from [81]. Copyright 2008 Elsevier Inc.

Figure 2.3 is provided to illustrate the terms that are involved in the PB equation. Aside from the partial charge distribution described by Equation 2.12, atomic positions and radii are also included in the coefficients $\epsilon(x)$ and $\kappa(x)$. The dielectric constant is discontinuous along the biomolecular surface, taking solute dielectric values inside the protein and bulk solvent values outside the surface, and modeling crossing between dielectric interfaces should be given special attention. Various models can be used to represent the dielectric function $\epsilon(x)$, such as discontinuous transitions at the molecular surface, smooth spline-based definitions, and Gaussian-based descriptions.[85]

Assuming that $\sinh\phi(x)$ can be approximated as $\phi(x)$, the complex full nonlinear PB Equation 2.15 can be simplified into a linearized form that is comparatively easier to solve:

$$\begin{aligned}-\Delta\ \cdot\ \epsilon(x)\Delta\phi(x)+\overline{\kappa}^2\phi(x)=\frac{4\pi e_c^2}{k_BT}\sum_{i=1}^{M}q_i\delta(x-x_i)\ ,\ \text{for } x\in\Omega,\\ \text{where}\quad \phi(x)=g(x)\text{ for } x\in\ \partial\Omega\end{aligned}\tag{2.16}$$

Linearization of Equation 2.15 provides the most accurate results when the nonlinear contributions of $\sinh\phi(x)$ can be disregarded, which is typically the case for small electrostatic potential values.[23] Analytical solutions to the PB equation are limited to a few simple cases, and for complex biomolecules like proteins, numerical solutions based on finite difference methods are employed. This involves mapping the macromolecule and a region of the surrounding solvent onto a cubic grid, with each grid point assigned values for the charge density, dielectric constant, and ionic strength.[24]

After solving the PB equation (2.15 or 2.16) to determine the electrostatic potential $\phi(x)$, the free energy can be computed by integrating the potential over the domain of interest:

$$G(\phi) = \int_{\Omega} \left[\rho_f \phi - \frac{\epsilon}{2} (\Delta\phi)^2 - \overline{\kappa}^2 (\cosh\phi - 1) \right] dx \tag{2.17}$$

In Equation 2.17, the first term represents the energy required to insert the protein charges into the electrostatic potential, and can be interpreted as the energy of interaction between the fixed charges. The numerical solution of the PB equation, however, also yields significant "self-energy" terms that indicate the energy of charge interaction on its own.[24, 82] The "self-energy" terms are eliminated by performing reference computations using the same discretization (mesh spacing) and subtracting the results from the numerical solution The second term in Equation 2.17 represents the energy of polarization in the dielectric medium, while the third term represents the energy of the mobile charge distribution and can be interpreted as the excess osmotic pressure of the system.[23]

Despite its success in modeling and elucidating electrostatic phenomena in biomolecular interactions, the PB theory is still approximate and has several limitations, particularly for highly charged systems or higher ionic strengths.[86, 87, 88] One of the main approximations in the PB theory is the mean-field approach used to treat ions, which assumes that each ion is surrounded by a uniform cloud of counterions and coions. This approach ignores counterion correlations and fluctuations, which are especially important in highly charged systems such as DNA and RNA.[23] As a result, the PB theory can underestimate the repulsive electrostatic interactions between highly charged particles and lead to incorrect predictions of the behavior of charged biomolecules in solution.[87] Regardless of these approximations, the PB model remains a useful tool for describing basic solvation behavior and can be applied to a variety of biomolecular systems. It can be applied to a variety of biomolecular systems, such as protein-ligand binding, protein-protein interactions, and ion channels.[84]

2.4. Molecular dynamics

MD is a powerful simulation method that allows us to gain insights into the movements and behavior of atoms and molecules.[18, 19] By using sophisticated computational techniques, MD can provide a detailed picture of chemical, physical, or biological processes and their structure-to-function relationships. Unlike many experimental techniques, which may not be able to access processes on atomistic scales, MD can provide us with an unprecedented level of detail. This is especially important in fields such as materials science, biophysics, and chemistry, where a deep understanding of the underlying molecular processes is essential. With MD, researchers can explore the intricate dynamics of individual molecules, uncovering the underlying mechanisms that govern their behavior. The first computer simulation of a liquid was carried out over 60 years ago at the Los Alamos National Laboratories in the United States. This groundbreaking research was performed on the Los Alamos computer, known as "MANIAC" which was one of the most powerful computers of its time.[19] This was the first Monte-Carlo simulation, and represented molecules as idealised geometries like spheres and disks. In 1957, the first MD simulation was performed, which simulated a system of hard spheres. It wasn't until 1964, several years later, that the simulation of more complex systems, such as Lennard-Jones particles, became possible.[19] Since then, computational power has significantly increased enabling simulating larger and more sophisticated systems.

MD is a computational technique that models the behavior of atoms and molecules in a system over time. This deterministic approach adheres to the principles of classical mechanics and involves integrating Newton's equations of motion to track the movements of a group of atoms as they interact with one another.[18, 19] During these simulations, discrete time steps are utilized, with each step typically lasting around one femtosecond (fs). At each time step, the force F_i acting on each atom is calculated according to the following equation:

$$F_i = - \bigtriangledown_{r_i} U(r_1, ..., r_N) = \dot{p}_i \tag{2.18}$$

In this equation, the symbols r_i and p_i represent the Cartesian coordinates and momenta, respectively, of atom i. The term $U(r_1, ..., r_N)$ refers to the potential energy of the system, which is a function of the positions of all atoms in the system.

These forces are used to update the positions and velocities of each atom according to Newton's second law:

$$\dot{r}_i = \frac{p_i}{m_i} \tag{2.19}$$

where m_i denotes mass of atom i.

A simulated system can be represented by a set of particle positions and momenta. The Hamiltonian, denoted as H, represents the total energy of a system of N atoms and is expressed as the sum of the kinetic and potential energy functions:

$$H(q, p) = K(p) + U(q) \tag{2.20}$$

Here, q and p are sets of generalized coordinates and momenta that describe the positions and velocities of each atom in the system, respectively:

$$\begin{aligned} q &= (q_1, q_2, ..., q_N) \\ p &= (p_1, p_2, ..., p_N) \end{aligned} \tag{2.21}$$

The kinetic energy function, $K(p)$, is dependent only on the momenta, while the potential energy function, $U(q)$, is dependent only on the positions of the atoms.

Given the knowledge of the potential energy and coordinates of a system, Equation 2.18 can be employed to compute the forces experienced by each atom. At each time step, 3N position coordinates and 3N momenta must be calculated for a system of N atoms, resulting in a set of 6N first-order differential equations that need to be solved numerically. By recalculating the forces using the updated coordinates, a trajectory that follows the temporal evolution of the system can be obtained.[18, 19]

2.4.1. Finite Difference Methods for Accurate Time Integration in Molecular Dynamics Simulations

The finite-difference approach is a widely used method for solving ordinary differential equations, including equations 2.18 and 2.19. This is achieved through integration algorithms that use finite difference methods to divide time into discrete time-steps δt. With information about the particles' positions and derivatives at a given time, these algorithms can accurately predict their positions at a later time $t + \delta t$. The timestep δt is chosen to preserve system dynamics and energy, and should be shorter than the fastest movement in molecules, which is usually the atomic vibrations. An ideal integration algorithm for MD simulations should possess several key qualities. Firstly, it should be efficient and capable of using a relatively long timestep to speed up the simulation process. Additionally, it should accurately reproduce the classical trajectory of the particles, while also satisfying the conservation laws for energy and momentum.[19, 18] By meeting these criteria, the integration algorithm can produce reliable and accurate results while minimising computational resources. So far, various algorithms have been developed and used for MD simulations, including the Verlet, Leapfrog, and Velocity Verlet algorithms, among others.[89, 90, 91, 19]

The Verlet algorithm[89] is based on a Taylor expansion of the particle coordinate r(t) with respect to time:

$$\begin{aligned} r_i(t+\delta t) &= r(t) + \Delta t v(t) + \frac{\delta t^2}{2} a(t) + \mathcal{O}(\delta t^3) \\ r_i(t-\delta t) &= r(t) - \Delta t v(t) + \frac{\delta t^2}{2} a(t) + \mathcal{O}(\delta t^3) \end{aligned} \tag{2.22}$$

where $v(t)$ and $a(t)$ denote particle velocity and acceleration at a given time. When we subtract the expansion of $r_i(t + \delta t)$ from the expansion of $r_i(t - \delta t)$ and rearrange the resulting expression, we obtain:

$$r_i(t+\delta t) = 2r(t) - r_i(t-\delta t) + \delta t^2 a(t) + \mathcal{O}(\delta t^4) \tag{2.23}$$

The Verlet algorithm has limitations because it doesn't directly calculate velocities (as shown in equation 2.23). Although velocities are not required for computing the time evolution of a system, they are often necessary to calculate the kinetic energy, which is needed to evaluate whether the total energy of the system is conserved.

As a solution, the Velocity Verlet scheme was developed, which directly calculates velocities. In this scheme, positions are computed using a half-step approach, as demonstrated by:

$$\begin{aligned} v(t+\frac{\delta t}{2}) &= v(t) + \frac{\delta t}{2} a(t) \\ r_i(t+\delta t) &= r(t) + \Delta t v(t+\frac{\delta t}{2}) \\ v(t+\delta t) &= v(t+\frac{\delta t}{2}) + \frac{\delta t}{2} a(t+\delta t) \end{aligned} \tag{2.24}$$

The Velocity Verlet scheme is a reliable and straightforward method that has several advantages. It is numerically stable and convenient to use, as well as being precisely reversible in time. Moreover, if conservative forces are present, the scheme is guaranteed to conserve linear momentum.[19]

Exact solutions for long timescales in MD simulations are unattainable with integration algorithms. However, accurate solutions are only necessary for relevant timescales when calculating time correlation functions. The primary aim of molecular dynamics simulations is generating states sampled from the microcanonical ensemble. Although exact classical trajectories are not essential for this purpose, conserving energy is critical to ensure the trajectories remain on the appropriate constant-energy hypersurface. Deviations from this hypersurface can compromise the accuracy of the results, affecting the correct ensemble averages.[19]

2.4.2. Statistical ensembles

MD simulations provide valuable information about the microscopic behavior of atomic and molecular systems, including their positions, velocities, and other properties. However, to understand the macroscopic behavior of these systems and compare simulation results with experimental data, the principles of statistical mechanics must be applied. An important goal of MD simulations is to obtain meaningful and reliable results that can be compared with experimental observations. To achieve this, simulations must adhere to the ergodic hypothesis.[19, 71] Since it is impractical to visit all possible states of a system in a finite simulation time, we must sample relevant properties over a sufficient duration. If a system is ergodic, the time average of a property over a long simulation time is equivalent to its ensemble average.[71] This means that we can obtain reliable macroscopic information

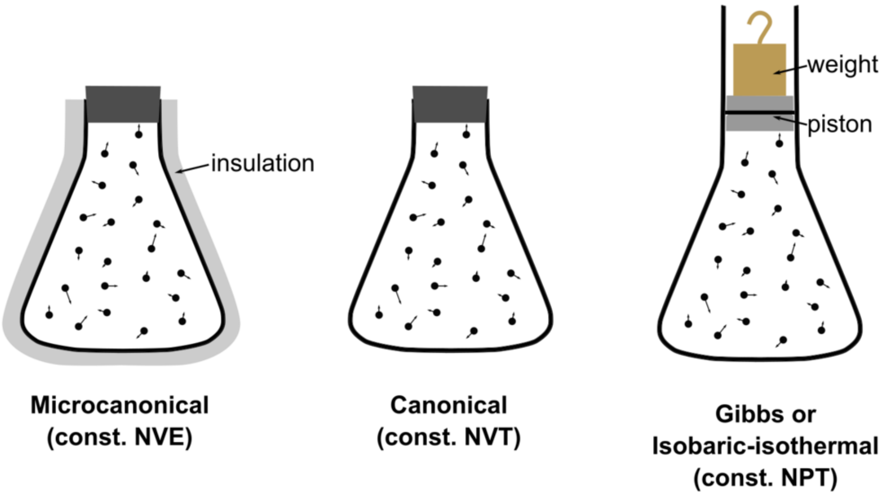


Figure 2.4.: Visual representation of the three statistical ensembles frequently utilized in molecular dynamics simulations: NVE, NVT, and NPT. Adapted from a public domain source under the Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license.

about the system from the microscopic MD simulations, as long as the simulations are conducted under appropriate conditions and for a sufficient duration.

In statistical mechanics, the properties of a system are studied in terms of its ensemble, which represents all possible microstates of the system. Different ensembles are used to describe the system under different conditions, such as fixed energy, volume, or temperature.[71] In MD simulations, three commonly used ensembles are the microcanonical (NVE), canonical (NVT), and isothermal-isobaric (NPT) ensembles, as shown of Figure 2.4. The microcanonical ensemble describes a system that is isolated and maintains a constant energy. In MD simulations, the equations of motion are modified to conserve energy. In the NVT ensemble, the system is allowed to exchange energy with a heat bath, while the total number of particles and volume remain constant.[92, 93] To simulate this ensemble, the Hamiltonian is modified to include a heat bath and the equations of motion are modified to maintain constant temperature. The NPT ensemble describes a system with constant temperature, volume, and pressure.[94] To simulate this ensemble, the Hamiltonian is modified to include a barostat, which maintains the desired pressure, and a thermostat, which maintains constant temperature. Commonly used thermostats in MD simulations include Velocity Rescaling, Nosé-Hoover and Langevin thermostats, along with barostats such as Berendsen, Parrinello-Rahman, and Nosé-Hoover barostats.[92, 94, 95, 93]

2.4.3. Periodic boundary conditions

In MD simulations, only a small number of molecules can be used in calculations due to computational limitations, typically placed in a simulation box. However, this presents a problem because molecules on the surface of the simulation box experience different forces than those in the bulk. To address this issue, periodic boundary conditions (PBCs) are used.[19] PBCs create an infinite lattice by replicating the simulation box throughout space. As a molecule moves in the original box (colored by grey on Figure 2.5), its periodic image in neighbouring boxes moves the same way. If a particle exits the simulation box on one side, it re-enters the box from the opposite side, as if it were entering from the periodic copy of the simulation box. This ensures that the simulation box effectively becomes an infinite lattice, which can be used to accurately simulate bulk properties of the material.[19]

The minimum image convention is an important concept related to PBCs, which dictates that a particle interacts only with the closest periodic images of other molecules, as

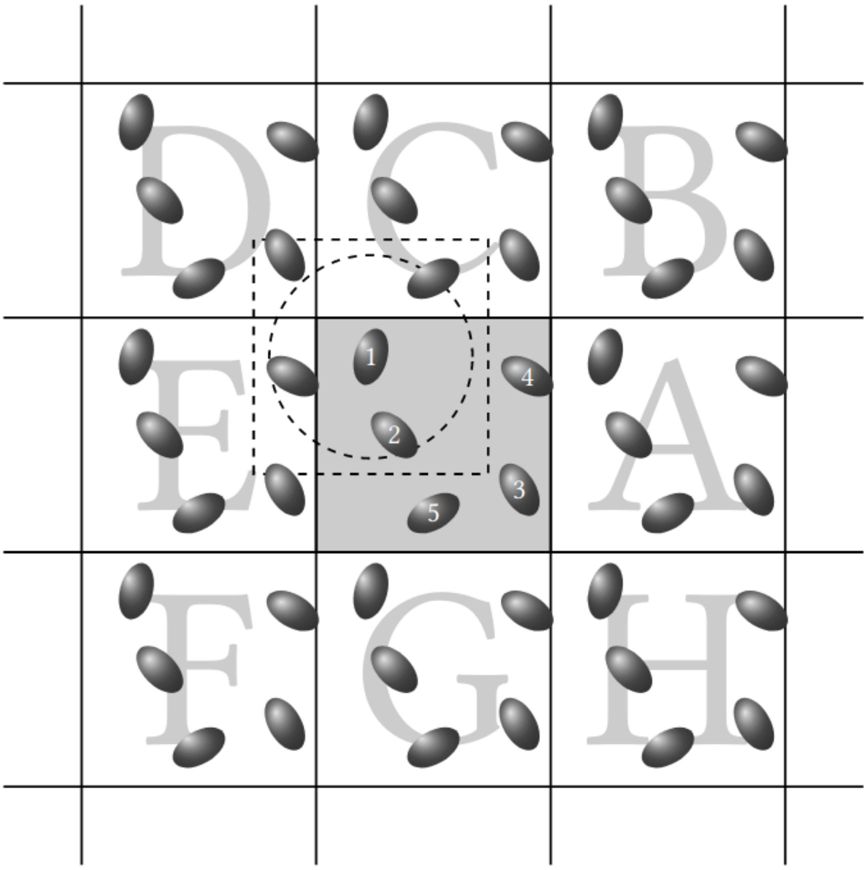


Figure 2.5.: Visual representation of the periodic boundary conditions used in molecular dynamics. For simplicity a two-dimensional periodic system is shown. Original box is depicted by grey colour, while replicated boxes are denoted by letters A, B and so on. Adapted with permission from [19]. Copyright 2017 Oxford University Press.

illustrated by the dashed rectangle in Figure 2.5. Furthermore, to optimize computational performance, particle interactions are considered only if the distance between them is smaller than a cutoff radius (depicted by a sphere in Figure 2.5). The cutoff radius must not exceed half of the simulation box to ensure that a particle interacts with only one image of any given particle.[19]

2.4.4. Force field

In context of MD simulations, a force field is a set of functional forms used to describe the potential energy of a collection of atoms, both within and between molecules. These functions and parameters are derived from experimental studies and from accurate quantum mechanical calculations.[19, 96, 97] MD simulations are then used to compare calculated properties with experimental results, allowing for refinement of the parameters. This process ensures that the force fields developed are transferable and can be applied to many different molecules.[19] As a result, individual force fields are regularly updated and extended to ensure their accuracy and relevance. Some force fields are specialised for certain types of molecules, such as proteins, lipids, or polymers, while others are more general and can be used for a wide range of systems.[97, 96]

While force fields may differ in their functional forms, the majority of force fields, including those utilized in this thesis, can be described by the following expression:

$$U = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsion} \frac{U_n}{2} [1 + \cos(n\Phi + \delta)] \\ + \sum_{LJ} 4\epsilon_{ij} [(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \sum_{el} \frac{q_i q_j}{4\pi\epsilon r_{ij}} \tag{2.25}$$

The bonded interactions (bonds, angles, and torsions) are described by the first three terms in Equation 2.25. The first term, which represents the sum over of all bonds with an equilibrium bond-length r_0 and bond force constant k_b, applies to each pair of chemically connected atoms. Some force fields use a more realistic functional form, such as the Morse potential, or simply fix the bonds at their equilibrium values. The second term, represents the sum oover all bond angles for each set of three connected atoms, where θ, θ_0 and k_a denote angle, reference angle and angle constant respectively. The third term represents the sum over all torsions involving four connected atoms, where U_n denote the dihedral force constant, and Φ , n and δ are the dihedral angle, the order constant, and the reference dihedral angle respectively. The final two terms in the equation represent the non-bonded interactions, the Lennard-Jones and the electrostatic interactions. Here, r_{ij} is the distance between atoms i and j, ϵ_{ij} is the potential well-depth of the interaction, and σ_{ij} is the distance at which the potential becomes zero and q_i, q_j are atom charges.

In real physical systems, the interactions between three or more molecules, known as many-body potentials, can have a substantial impact on the properties of liquids. Unfortunately, due to computational constraints, these non-additive terms are often not included in force fiels, and the pair potentials used are considered as effective pair potentials that account for all many-body effects. This approximation can lead to situations where the effective pair potential required to reproduce experimental results may vary with changes in temperature, density, and other factors, even the true two-body potential should remain constant.[19]

2.4.4.1. Coarse grained force fields

Due to limitations in computational resources, all-atom MD simulations are still only capable of modelling a relatively small number of systems for processes occurring on short time scales, despite the progress made in the field. As a solution to these limitations, researchers are constantly developing coarse-grained (CG) models that bypass the need for full atom representation of molecules.[20, 98] In this approach, groups of atoms are combined into a single 'virtual atom,' or bead, reducing the number of particles and explicit pairs required to calculate energy and force within a simulation box.[19] This results in a significant reduction in computational time. Furthermore, the use of a coarser representation of molecules increases the characteristic length scale of the system. This allows for the use of longer timesteps, effectively covering more real time in the simulation. This approach is particularly relevant for simulating large biomolecular systems, such as proteins, which undergo a variety of processes occurring over a wide range of time scales, ranging from nanoseconds to hours or longer, as illustrated in Figure 2.6. Given the significant surge in experimental data that requires interpretation, the development of CG models for proteins has become imperative.[98]

Nearly half a century ago, the initial CG protein models were proposed, but it wasn't until more recently that they gained widespread usage. In 2013, Michael Levitt, Ariel Warshel, and Martin Karplus were jointly awarded the Nobel Prize in Chemistry for their early achievements in developing multiscale models for complex chemical systems, including the crucial use ofv CG modeling to investigate large biomolecular systems such as proteins.[99, 100, 101, 102] Some examples of commonly used CG models for proteins

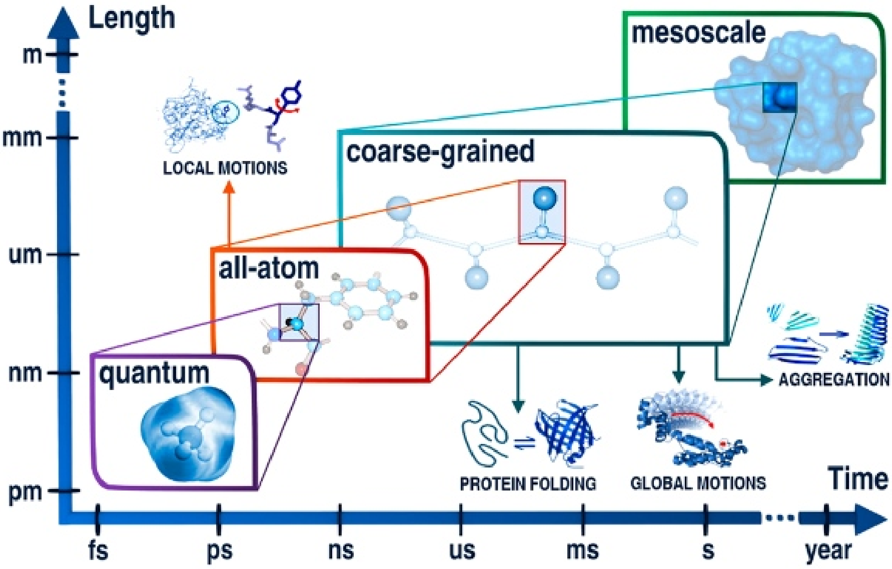


Figure 2.6.: Application of molecular modeling at different levels of resolution, including quantum, all-atom, coarse-grained, and mesoscale, showing approximate ranges of time scales and system sizes (lengths). Reprinted with permission from [20]. Copyright 2016 American Chemical Society.

include Martini,[103] UNRES,[104] PRIMO,[105] and the SIRAH force field,[106] which will be discussed in greater detail in Chapter 2.4.4.2 of this thesis. CG force fields can differ in their level of structural detail they provide, but a significant advantage of using them is the ability to retrieve higher resolution information through the performance of CG MD simulations. This allows for the recovery of full-atom resolution details, despite the initial use of a simplified CG model. The design of force fields for CG models is guided by the chosen level of resolution and underlying philosophy. The force field can be parametrized to reproduce reference all-atom MD simulations, or it can be estimated through statistical analysis of the structural and dynamic features observed in databases of experimental structures.[20, 19] Compared to all-atom force fields, CG force fields tend to "smoothen" the energy surface, which reduces the likelihood of becoming trapped in free energy minima. This smoothing occurs because in CG models, multiple atoms are typically grouped together

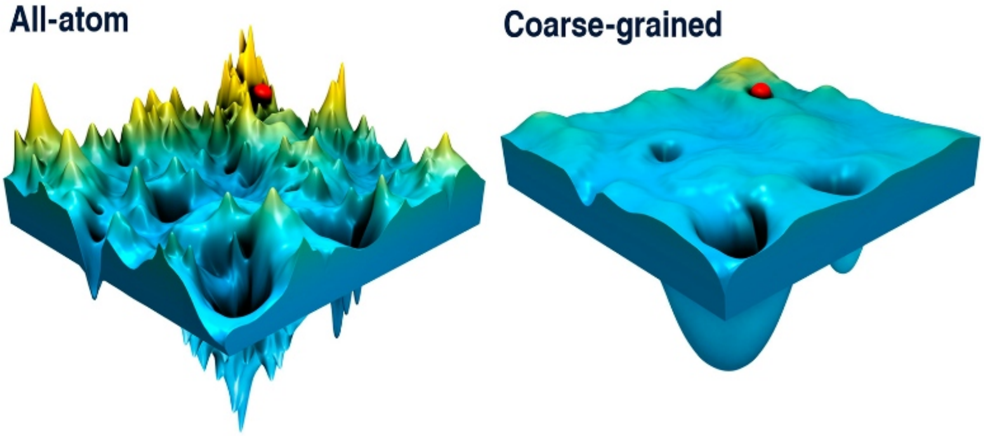


Figure 2.7.: Illustration the difference between all-atom and coarse-grained models in terms of the energy landscapes they generate, enabling more efficient exploration of the energy space. Reprinted with permission from [20]. Copyright 2016 American Chemical Society.

into a single bead, which can lead to an averaging of the energetic interactions between atoms.[20] This effect is illustrated in Figure 2.7, highlighting the advantage of using CG force fields in the study of complex biomolecular systems. Coarse-graining can affect the balance between entropy and enthalpy in a system by reducing the number of degrees of freedom. This can lead to a differences compared to all-atom models, even though the total free energy difference may still be accurately reproduced.

2.4.4.2. SIRAH Force Field

The SIRAH (Southamerican Initiative for a Rapid and Accurate Hamiltonian) is a CG force field capable of accurately modeling a range of complex biomolecules including proteins,[106, 107] DNA,[108, 109] lipids,[110] water,[111] and ions.[111] The SIRAH force field is a top-down approach to modeling biomolecules that is parametrized to fit their structural characteristics. By using a classical Hamiltonian to evaluate molecular interactions, similar to those found in commonly used atomistic models (represented by equation 2.25), it can be easily integrated with a range of popular MD software packages.[112] In summary, interactions between bonds and angles are described using harmonic terms, for example $\frac{1}{2}k_b(r - r_0)^2$, which involve reference force constants k_b and an equilibrium bond length r_0. Dihedrals, on the other hand, are represented using Fourier expansions. The non-bonded contributions are taken into account using Coulomb and 12-6 Lennard-Jones terms. The Lorentz-Berthelot combination rule is applied to all atom-type pairs, unless specified otherwise. In those cases, specific Lennard-Jones parameters are set. Finally, the 1-4 nonbonded interactions are scaled using the AMBER scaling factor.[107, 106] The SIRAH force field uses uniform masses of 50 amu for each bead, allowing for a time step of 20 fs to be used in molecular simulations. The all-atom to CG mapping method relies on physical and chemical intuition to simplify complex systems by reducing the number of atoms used and positioning CG beads based on the location of real atoms. The following paragraphs explains how this process is utilized for water, proteins, and DNA.

WT4 Model for Water

The accurate treatment of solvent effects is crucial for the successful modeling of biomolecular processes. The SIRAH force field employs a CG model for water called WatFour or WT4.[111] The model is designed to mimic transient water clusters that occur in pure water as a result of hydrogen bonding between water molecules, as depicted in Figure 2.8. One WT4 molecule consists of four beads that are linked together, each of which carries a partial charge. As a result, each WT4 bead corresponds to eleven water molecules. The bond stretching constant between beads was set to mimic the interaction strength of hydrogen bonds. The weak bonds present in WT4 molecules impart structural flexibility, causing deviations from the ideal tetrahedral configuration, a phenomenon that is observed in real water. The structure of WT4 is modeled using two types of beads, one "oxygen-like" with a negative charge (-0.41e) and the other "hydrogen-like" with a positive charge (0.41e). These beads mimic the charge distribution and hydrogen bond network observed in clusters of water molecules.[108] This unique configuration allows WT4 to generate its own dielectric permittivity, which is crucial in molecular simulations. Additionally, CG electrolytes are used to account for the effects of ionic strength and osmotic pressure. The WT4 model is capable of reproducing the more detailed tetrahedral organization of water through noncovalent interactions. The model parameters were optimized to reproduce bulk water properties, including density and diffusion coefficients, at physiological relevant temperature conditions in the range from 278 to 328 K.[111] The WT4 solvent model is also capable of accurately replicating important electrolytic properties such as screening, osmotic pressure, Bjerrum and Debye lengths. This results in correct concentration profiles, ion specificity, and local conformational changes that are observed in high-resolution X-ray structures of

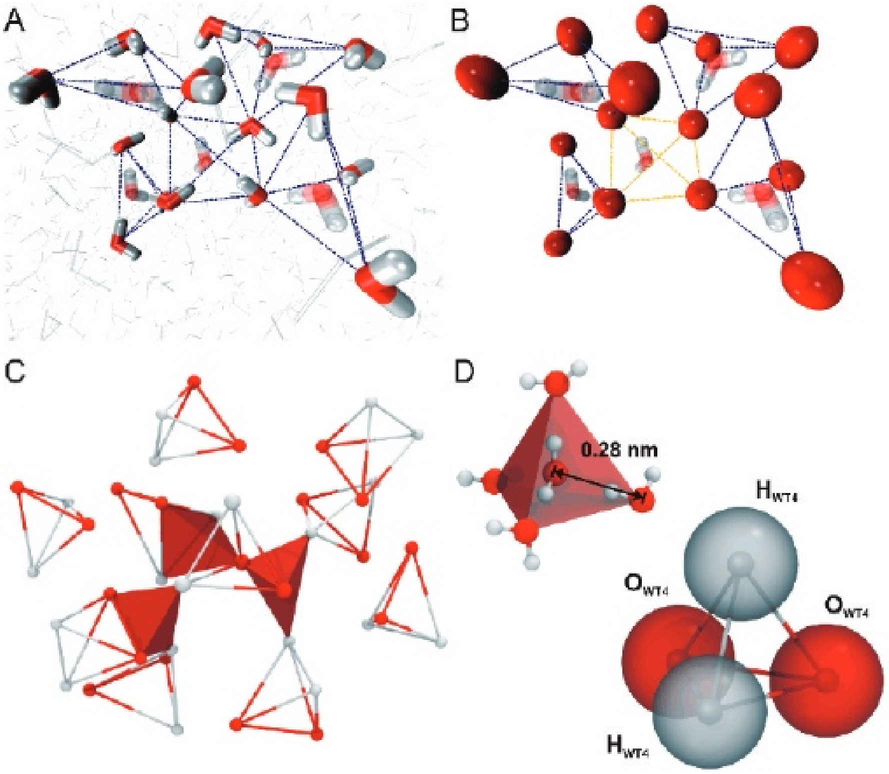


Figure 2.8.: Illustration of the WT4 model used to represent water molecules in the SIRAH force field a) MD snapshot of a typical arrangement of water molecules in bulk. Water molecules tend to cluster together in irregular tetrahedral shapes. b) The positions of oxygen atoms at the corners of the tetrahedrons from panel A are highlighted using red beads. c) The structural organization of WT4 in bulk solution, as captured by a MD snapshot. d) The geometry of the WT4 molecule, with the white and red beads representing H-like and O-like beads with partial charges 0.41a and -0.41e, respectively. Reprinted with permission from [111]. Copyright 2010 American Chemical Society.

DNA. In addition, the WT4 model supports CG/MM hybrid simulations, which provide a powerful tool for studying complex systems.[111] With this approach, regions of interest (i.e. DNA) can be described in full atomistic detail, while solvent can be represented in a CG resolution.

CG Model for Ions

The SIRAH force field is capable of modeling three different ionic species at the CG level, namely NaW+, KW+, and ClW-.[111] This provides a useful tool for simulating systems that involve these ions, such as ion channels or electrolyte solutions. The CG ions are designed to incorporate both ions and water molecules in their first solvation shell, which typically consists of approximately six water molecules. The bead masses are set to the sum of the ion mass and the mass of water molecules, and van der Waals parameters were chosen to match the first minima of the radial distribution function of hydrated ions obtained from experimental data, such as neutron diffraction experiments.

CG Model for Proteins

The SIRAH CG model for proteins and peptides is always used in combination with an explicit solvent of WT4 molecules to address several common limitations of CG force fields.[106, 107] These include the use of a uniform dielectric constant, the lack of long-range interactions, and the need to use constraints to maintain secondary structure. In the SIRAH force field, proteins are mapped onto a CG representation by treating the peptide bonds with a relatively high level of detail, while side chains are modeled at a lower level of structural resolution. This approach allows for a more efficient simulation of large protein systems, while still capturing the essential features of protein structure and function. The backbone representation of the peptide bonds, which retains the positions of the nitrogen(N), α carbon(Cα), and oxygen atoms(O), provides an unbiased description of the conformational space explored by peptides and proteins, without imposing any specific secondary or tertiary structure, as shown on Figure 2.9. The use of partial charges on each bead can approximately account for the formation of hydrogen bond-like interactions.[107, 111] This helps to stabilize the formation of α-helices and β-sheets without the need for ad hoc constraints. In addition, the van der Waals interactions within backbone beads are set to the same values as those in the AMBER99 force field. This ensures that the protein structure achieves the appropriate level of compaction during the formation of α-helices and β-sheets. Assigning dihedral angles between four neighboring beads establishes a direct relationship with all-atom dihedrals.

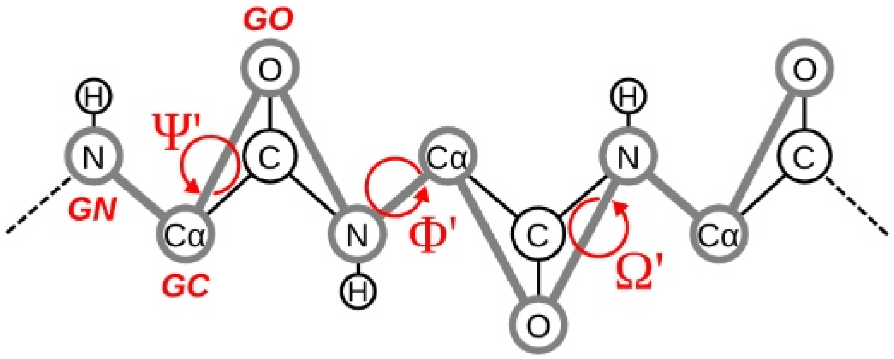
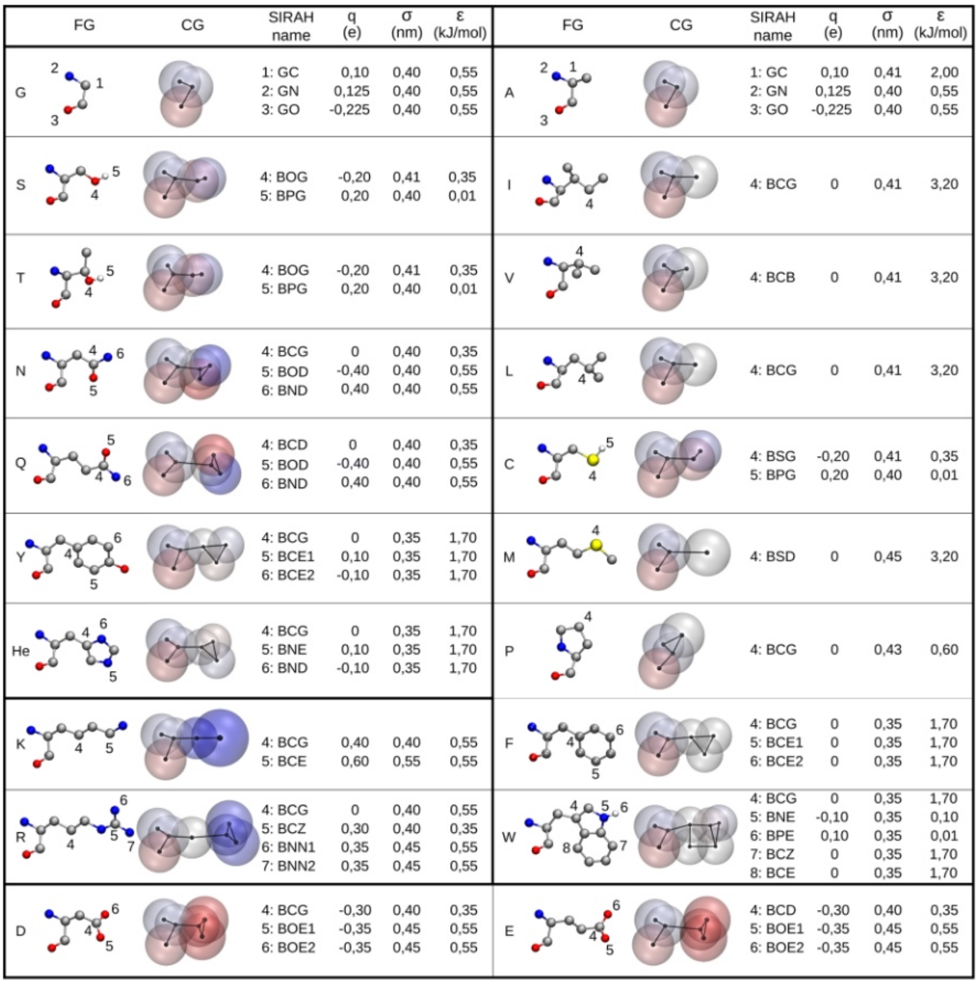


Figure 2.9.: CG mapping scheme of protein backbone. CG beads, colored in red, have been positioned on the atoms of Cα, N, and O, represented by gray circles. The tick lines denote connectivity, while the primary dihedral angles at the CG level have been highlighted. Adapted with permission from [106]. Copyright 2015 American Chemical Society.

The current version of SIRAH includes both neutral and charged termini. Charged termini are created by placing a charge of +1 or -1 on the N- or O-terminal beads. The topology of the CG side chains follows the principle of representing interaction points based on the characteristics of each residue (see Figure 2.10). Hydrophobic amino acid residues, such as Val, Ile, Leu, and Met, are mapped to a single CG bead with zero net charge. Aromatic side chains are mapped to either three (Phe, His, and Tyr) or five beads on a plane (Trp). The beads of polar and charged side chains correspond to charged groups or hydrogen bond acceptors/donors. To ensure the "L" chirality of amino acids, improper dihedral angles are applied to both the backbone and side-chain beads. Furthermore, to enforce the planarity of the tryptophan side chain, two additional improper dihedrals are used. The force field parameters for proteins in the SIRAH model have been updated to the 2.0 version in order to improve the accuracy of the structural descriptors such as such as root mean square deviations, solvent accessible surface, radius of gyration and native contacts.[107] To achieve this, the charges, van der Waals, and dihedral parameters were adjusted in an ad hoc manner. The angular force constants were adjusted to increase side-chain flexibility, and the non-bonded interactions were improved to better capture the hydrophilicity/hydrophobicity of both side-chains and backbone, as well to model cation pi interactios. The use of more terms in Fourier expansions for torsion potentials improved backbone torsion angles and enabled modelling of parallel and anti-parallel β-sheets.[107]



FG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)	FG	SIRAH name	q (e)	σ (nm)	ε (kJ/mol)
G	1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,40 0,40 0,40	0,55 0,55 0,55	A	1: GC 2: GN 3: GO	0,10 0,125 -0,225	0,41 0,40 0,40	2,00 0,55 0,55
S	4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	I	4: BCG	0	0,41	3,20
T	4: BOG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01	V	4: BCB	0	0,41	3,20
N	4: BCG 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	L	4: BCG	0	0,41	3,20
Q	4: BCD 5: BOD 6: BND	0 -0,40 0,40	0,40 0,40 0,40	0,35 0,55 0,55	C	4: BSG 5: BPG	-0,20 0,20	0,41 0,40	0,35 0,01
Y	4: BCG 5: BCE1 6: BCE2	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	M	4: BSD	0	0,45	3,20
He	4: BCG 5: BNE 6: BND	0 0,10 -0,10	0,35 0,35 0,35	1,70 1,70 1,70	P	4: BCG	0	0,43	0,60
K	4: BCG 5: BCE	0,40 0,60	0,40 0,55	0,55 0,55	F	4: BCG 5: BCE1 6: BCE2	0 0 0	0,35 0,35 0,35	1,70 1,70 1,70
R	4: BCG 5: BCZ 6: BNN1 7: BNN2	0 0,30 0,35 0,35	0,40 0,40 0,45 0,45	0,55 0,35 0,55 0,55	W	4: BCG 5: BNE 6: BPE 7: BCZ 8: BCE	0 -0,10 0,10 0 0	0,35 0,35 0,35 0,35 0,35	1,70 0,10 0,01 1,70 1,70
D	4: BCG 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55	E	4: BCD 5: BOE1 6: BOE2	-0,30 -0,35 -0,35	0,40 0,45 0,45	0,35 0,55 0,55

Figure 2.10.: SIRAH CG representation of amino acids. Each amino acid is presented using its one-letter code, all-atom heavy atoms, CG representation, bead names, partial charges, and vdW parameters. Only hydrogen atoms that are used for the CG mapping are shown, and the numbers near the atoms indicate the corresponding CG bead positions. The CG beads are colored based on their charge (negative, red; positive, blue). Adapted with permission from [106]. Copyright 2015 American Chemical Society.

CG Model for DNA

The CG model for DNA uses six beads to represent each nucleotide, while still preserving the important Watson-Crick recognition that occurs in the chemistry of DNA.[108, 109] The CG model can faithfully replicate experimental structures, capture the dynamic breathing of DNA, and accurately depict the various conformational transitions that occur within the molecule. The CG model defines four distinct coarse-grained bases (dax, dtx, dcx, and dgx) that correspond to the all-atom nucleotides.[109] This mapping is illustrated in Figure 2.11 and ensures that the fundamental interactions between nucleotides are maintained, thus preserving the overall "chemical sense" of the system. In the CG model, the phosphate sugar backbone is represented with less detail than the Watson-Crick pairs, as the latter are modeled with a higher level of specificity to accurately capture their interactions. The phosphate groups in the nucleotides are depicted as px beads that occupy the same position as the corresponding phosphorus atoms. Additionally, the bead kx is placed at the position of the C5' atom to establish the 5'-3' direction of each DNA strand. This allows for the formation of the major and minor grooves in the CG model. The kn bead (where kn can be ka, kt, kc, or kg) is located at the position of the C1' atom. To maintain the molecular specificity between the two DNA strands, the superatoms that

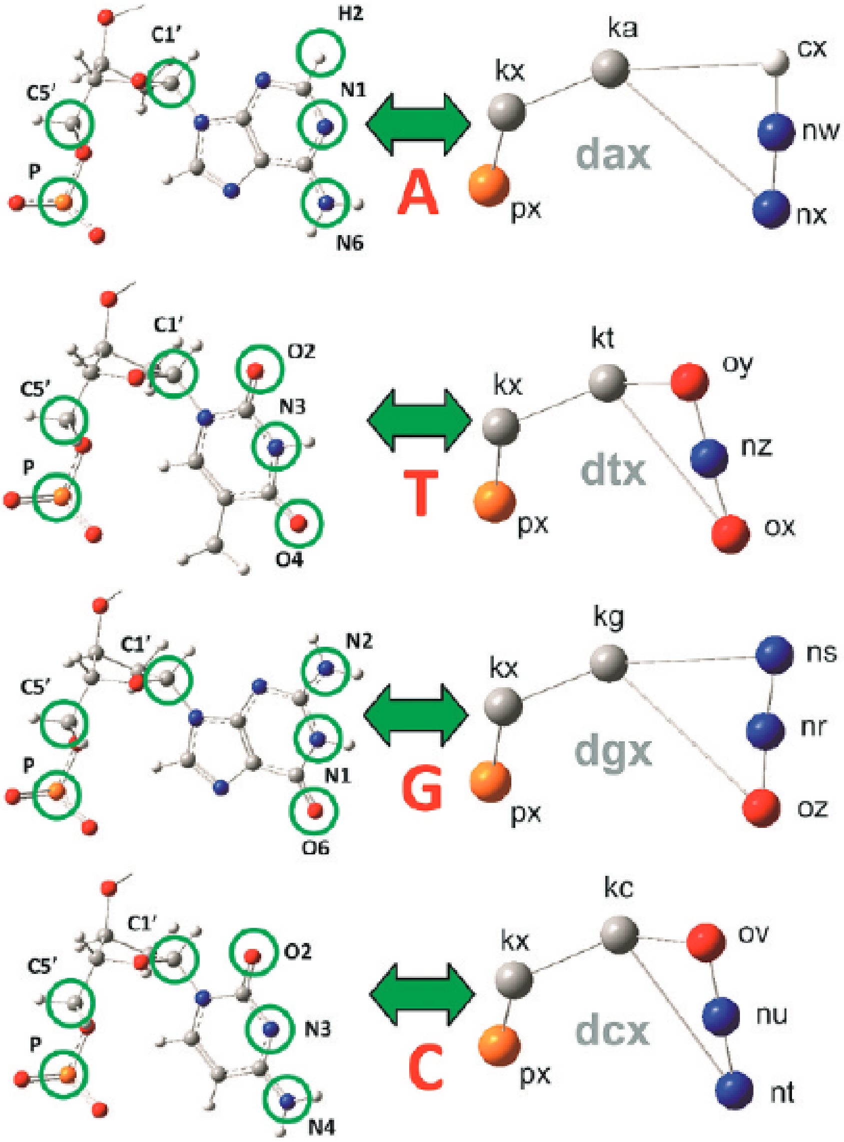


Figure 2.11.: SIRAH CG representation of amino acids. The circular symbols indicate the positions of elements that are retained in the CG representation. Adapted with permission from [109]. Copyright 2010 American Chemical Society.

participate in Watson-Crick interactions in the CG model are placed in the same position as their corresponding atoms.[109] This approach ensures that the all-atom Watson-Crick hydrogen bonds are preserved as two-point electrostatic interactions in the CG model. This model The CG model described in this scheme has been shown to accurately replicate key features of DNA, including solvation spines, electrolyte specificity, and cation-driven narrowing of the minor groove.has been shown to accurately replicate key features of DNA, including solvation spines, electrolyte specificity, and cation-driven narrowing of the minor groove.[108]

2.5. Accelerated Molecular Dynamics Techniques

Simulating rare events is a challenging task that requires specialised techniques. A simple definition of a rare event is a transition from one region of phase space to another, typically separated by a high energy barrier as depicted in Figure 2.12 At equilibrium, one state may be significantly more stable than the other, in which case the less stable state can be referred to as being thermodynamically metastable.[19, 21] The goal of simulating rare events is to study the rate of conversion from one state to another, as well as the possible paths that the system takes in a suitable set of reaction coordinates. Examples of rare events include the conformational changes of biomolecules, chemical reactions, crystallization, and numerous other phenomena across a variety of fields.[19] Simulating rare events is difficult due to the long residence periods in each state, which can exceed the maximum practical length of a simulation. This challenge is particularly pronounced in large, high-dimensional systems containing multiple metastable minima and saddle points. In the field of computational science, calculating free-energy differences is essential because they determine the driving force behind a process. However, rare events can take too long to simulate, making their study infeasible without accelerated MD techniques.[19, 21]

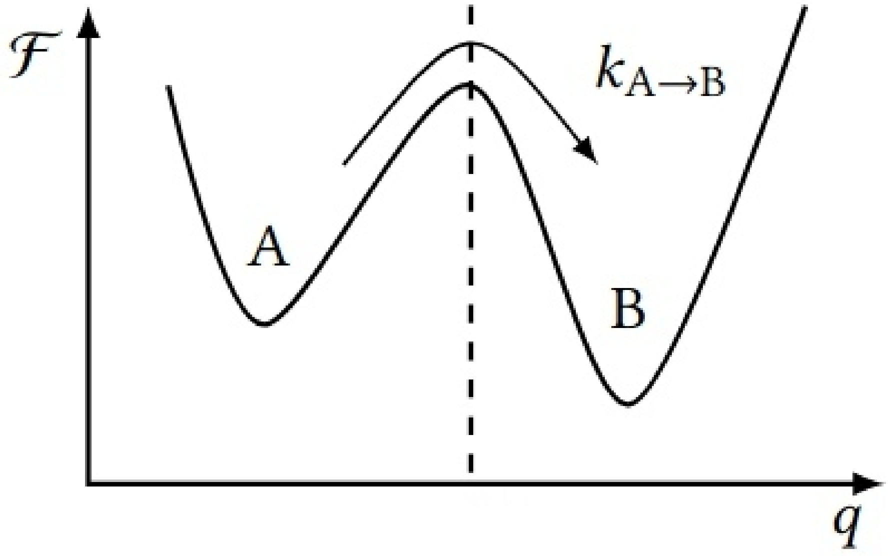


Figure 2.12.: The free energy as a function of a single reaction coordinate, F(q), for a system with two minima labeled A and B. The two minima are separated by a maximum. Adapted with permission from [19]. Copyright 2017 Oxford University Press.

The canonical partition function Q for a system in the NVT ensemble is obtained by integrating over the entire phase space, as shown in the following equation:

$$Q = \int e^{\frac{U(r)}{k_b T}} d^N r \tag{2.26}$$

where $E(r)$ and N denote potential energy and number of degrees of freedom in system.

The Helmholtz free energy A and the canonical partition function Q are related through the equation:

$$A = \frac{1}{k_b T} \ln Q \tag{2.27}$$

The transition between two states can be characterized using a reaction coordinate (ξ), which can take on any order parameter, including changes in the Hamiltonian.[113] Typically, geometric properties such as distance, angle, torsion, or the difference between root mean square deviations from two reference states are used to define ξ.[19, 113] Once the reaction coordinate ξ is defined, the probability distribution of the system along ξ can be calculated by integrating out all degrees of freedom except for ξ:

$$Q(\xi) = \frac{\int \delta[\xi(r) - \xi] e^{\frac{U(r)}{k_b T}} d^N r}{\int e^{\frac{U(r)}{k_b T}} d^N r} \tag{2.28}$$

Here, $Q(\xi)d\xi$ can be interpreted as the probability of finding the system within a small interval of width $d\xi$ centered at ξ. Using the definition of the canonical partition function and the probability distribution along the reaction coordinate, the free energy can be calculated as:

$$A(\xi) = \frac{1}{k_b T} \ln Q(\xi) \tag{2.29}$$

where $A(\xi)$ is basically a PMF.[113] In MD simulations, it is often not feasible to directly calculate equations 2.28 and 2.29. However, according to the ergodic hypothesis, the ensemble average $Q(\xi)$ can be approximated by the time average $P(\xi)$. $P(\xi)$ represents the normalized frequency or histogram of the system being found at specific ξ values during MD runs. However, due to the limited duration of MD simulations, it may not be possible to explore all parts of the reaction coordinate ξ, particularly in regions that correspond to barriers between local minima.[113] Accelerated MD techniques are necessary to efficiently sample rare events and obtain full probability distributions. In order to drive a system over an energy barrier, one can either modify the energy expression to reduce the barrier or restrict the sampling space to the reaction coordinate.[19]

2.5.1. Umbrella Sampling

Umbrella sampling (US), is a biased MD method used to determine free energy along a reaction coordinate.[113, 114] US is a technique used to investigate rare events in computational chemistry by calculating free energy differences along a reaction pathway. It involves applying a bias potential along a reaction coordinate, which can be a one- or multi-dimensional variable that describes the progress of the reaction, to guide a system from one thermodynamic state to another. To simulate intermediate steps along the reaction pathway, multiple windows are created, each of which is associated with a specific value of the reaction coordinate. In each window, a biasing potential is applied to force the system to remain at that particular value of the reaction coordinate. MD simulations are performed at each window to collect statistics on the system's behavior and calculate the free energy associated with each window. The biasing potential utilized in US ensures that the system visits all parts of the reaction coordinate, including regions that would be rarely sampled with conventional MD simulations, such as rare events. The data collected from individual windows in US can be combined to obtain an overall free energy landscape. To obtain an unbiased free energy difference, the effects of the biased potentials should be be removed.[113, 115]

A bias potential $w_i(\xi)$ is introduced in window i to constrain the reaction coordinate ξ at a specific value, as expressed in the following equation:

$$U^b(r) = U^b(r) + w_i(\xi) \tag{2.30}$$

The biased potential energy, $U^b(r)$, and the unbiased potential energy, $U^b(r)$, correspond to the biased and unbiased states, respectively.[113] The bias potential, $w_i(\xi)$, is commonly represented as a quadratic equation, as shown in the following expression:

$$w_i(\xi) = \frac{x_w}{2}(\xi - \xi_i)^2 \tag{2.31}$$

where x_w is the force constant and ξ and ξ_i are the current and reference value of reaction coordinate respectively. Configurations that are far from the reference position i_i will bear a larger weights, and the energy function $U^b(\xi)$ will be inclined towards a specific conformation that is relevant, albeit with a non-Boltzmann distribution.[113] Ideally, the bias potential is chosen in a way that enables uniform sampling across the complete range of the reaction coordinate, ξ. Selecting a bias potential that is either too strong or too weak can result in non-overlapping distributions between the windows. As a result, the optimal bias potential is determined to be $w_{opt} = A(\xi)$, although the function $A(\xi)$ is not known beforehand. This means that selecting the appropriate bias potential requires knowledge of the system's characteristics and properties.

The relation between unbiased and biased probability at window i can be obtained as following:

$$P_i^u(\xi) = P_i^b(\xi)e^{\frac{\omega_i(\xi)}{k_bT}} \left\langle e^{\frac{-\omega_i(\xi)}{k_bT}} \right\rangle \tag{2.32}$$

Here, $P_i^b(\xi)$ is obtained from the of MD simulations of the biased system, while $\omega_i(\xi)$ is given analytically.

The free energy profile at window i can be determined as following:

$$A_i(\xi) = -k_bT\ln(P_i^b(\xi)) - \omega_i(\xi) + F_i \tag{2.33}$$

where $F_i = -k_bT\ln\left\langle e^{\frac{-\omega_i(\xi)}{k_bT}} \right\rangle$.

This derivation is accurate, and the only assumption made is that the sampling conducted in each window is sufficient.[115] The free-energy curves $A_i(\xi)$ from each window can be combined to obtain a global free-energy curve, $A(\xi)$, using techniques for estimating F_i which arise from the introduction of bias.

Several techniques have been proposed to estimate unknown F_i coefficients and determine unbiased probability distribution, but the most commonly used method is weighted histogram analysis (WHAM).[116, 115] The overall distribution is determined by computing a weighted mean of the individual window distributions, as described in the following equation:

$$P^u(\xi) = \sum_i^{windows} p_i(\xi)P_i^u(\xi) \tag{2.34}$$

The weights $p_i(\xi)$ are determined with the aim of minimizing the statistical error of $P^U(\xi)$

The equations for WHAM can be summarized as follows:

$$\begin{aligned} P^u(\xi) &= \frac{\sum_{i=1}^{N_{wind}} n_i(\xi)}{\sum_{i=1}^{N_{wind}} N_i e^{\frac{F_i - \omega_i(\xi)}{k_bT}}} \\ F_i &= k_bT\ln\left\{\sum_{\xi_{bins}} P^u(\xi)e^{\frac{-\omega_i(\xi)}{k_bT}}\right\} \end{aligned} \tag{2.35}$$

Here, the symbol N_{wind} represents the number of simulations, and $n_i(\xi)$ corresponds to the number of counts in the histogram bin that is associated with the value ξ.[113, 116]

The variables $P^u(\xi)$ and F_i are not known and are determined iteratively by equation 2.35 in a self-consistent manner. Once the convergence criterion is met, the most accurate estimation of the unbiased probability distribution $P^u(\xi)$ can be obtained, which in turn can be used to compute the PMF. WHAM allows for the calculation of free energies associated with arbitrary perturbations added to a reference potential in a self-consistent manner, and it can be extended to multiple temperatures or multidimensional reaction coordinates.[116, 113]

3. xDLVO-CG Model: A Tool for Understanding Protein-Protein Interactions

3.1. Introduction

[1] PPIs play a crucial role in determining the solubility, aggregation, precipitation, and crystallization behavior of protein solutions. Understanding PPIs is crucial for many biotechnological processes,[3, 4] but measuring and controlling them experimentally can be challenging. In order to design effective biotechnological processes, it is important to have a comprehensive understanding of the effective PPIs, as well as how environmental conditions such as pH, temperature, and ionic strength can impact the stability and phase behavior of protein solutions.[11, 12] Having this knowledge can help to develop stable protein formulations, prevent unwanted aggregation, and fabricate desired self-assembly constructs while reducing uncontrolled aggregation. Effective PPIs in solution are determined by weak nonspecific interactions, including steric repulsion, van der Waals, hydrophobic interactions, and long- or short-ranged electrostatic repulsion and attraction. The solubility of molecules is determined by their solid-liquid equilibrium, which is influenced by the interactions between the dissolved molecules themselves and also with the solvent.[118] The deviation of osmotic pressure from that of an ideal solution can provide insights into the interactions between dissolved molecules, and these interactions are quantified by virial coefficients (see Equation 2.1). The second virial coefficient, B_{22}, is particularly important in diluted systems as it provides information about the average two-body interactions between protein molecules. The osmotic second virial coefficient, B_{22}, can estimate PPIs in diluted protein solutions, characterising the non-specific protein binding capacity and stickiness.[119] It has been used to semi-quantitatively characterise protein solubility,[120, 121, 122, 123, 124, 125] phase behaviour,[126, 127, 128, 129] and crystallization.[68, 120, 122, 130, 129, 4] Positive B_{22} values indicate repulsion interactions and correspond to good protein solubility, while negative values indicate attractive interactions and protein aggregation.

Several studies have shown that there is a thermodynamic connection between the solubility (S) of a protein and its B_{22} coefficients.[120, 125] With the help of the measured B_{22} coefficient, it is possible to calculate the protein's solubility using the following formula:

$$B_{22} = \frac{-\Delta\mu}{RT}\frac{1}{2M_wS} - \frac{\ln S}{2M_wS} \tag{3.1}$$

where $\Delta\mu$ is defined as the difference between the standard chemical potentials of the protein in two different states. Specifically, it refers to the difference between μ_p^o in solution and μ_p^o in the crystal.

[1] The majority of the content presented in this chapter is based on the work published in Ref. [117], which has been modified and restructured to align with the format of thesis. Author contributions are stated in the 'List of Publications' section at the end of the thesis.

As a result, numerous semi-empirical models have been developed, which utilize adjustable parameters instead of $\Delta\mu$. This is because determining $\Delta\mu$ can be challenging, whereas using adjustable parameters provides greater adaptability in modeling protein solubility.[130, 131, 124] George and Wilson established a quantitative link between the optimal conditions for protein crystallization and the value of B_{22}.[70] They identified a narrow range of B_{22} in which crystallization occurs.[121, 120, 132] Subsequently, several studies have investigated the relationship between protein crystallization and B_{22}, and have found that crystallization usually happens when B_{22} is below 0.0001. Conversely, if B_{22} is excessively negative (meaning it is attractive), it leads to the formation of amorphous precipitates instead of crystals.

The osmotic second virial coefficient can be measured experimentally using various techniques, including membrane osmometry, static light scattering, cloud-point measurements, fluorescence anisotropy, self-interaction chromatography, and sedimentation equilibrium experiments.[127, 133, 134, 135] Despite their usefulness, the experimental techniques for measuring the osmotic second virial coefficient have certain limitations. For example, they can be time-consuming and require large quantities of protein samples, which restricts the range of experimental conditions that can be screened. Additionally, the measured values are often influenced by the experimental setup and cannot reveal details about specific terms that contribute to overall average protein interactions. Consequently, there is a need for theoretical models that can explore a wide range of solution conditions and complement experimental efforts.[136, 137] The B_{22} coefficient can be theoretically calculated using formula 2.3 (as explained in Chapter 2), provided that the PMF is known. The calculation of the B_{22} coefficient has involved various computational methods and representations of protein structure. Monte Carlo simulations and Mayer sampling have been used,[138, 139] with the Mayer sampling[140, 141, 142, 143, 144, 126] as well as MD simulations computing radial distribution functions or PMFs.[145, 146, 147] The counting of all configurations in which proteins interact has also been utilized.[144] These techniques have been applied to protein structures ranging from all-atom to low-resolution, CG models.[148, 149, 145, 147, 146] Apart from those computationally costly techniques, one commonly used theoretical model for calculating B_{22} coefficients is based on the DLVO theory (or xDLVO theory as described in Chapter 2), which represents proteins as ideal spheres.[30] However, while this model can effectively capture the dependence of B_{22} on salt concentrations,[3, **?**] the resulting model parameters are often physically unrealistic and may not be easily transferable to different proteins. Furthermore, proteins experience other types of interactions beyond electrostatic repulsion and van der Waals interactions, which are not accounted for in this model. Therefore, the use of more sophisticated models that account for more realistic representations of protein shape and interactions is necessary to obtain more accurate B_{22} values.

To address this issue, we developed an extended xDLVO-CG model, which includes a CG representation of proteins and an additional ion-protein dispersion interaction term.[117] This model has been tested on four different proteins: lysozyme(LYZ), subtilisin(Subs), bovine serum albumin(BSA), and immunoglobulin(IgG1) and demonstrated semi-quantitative agreement with experimental values, without the need for fitting to experimental B_{22} values.

3.2. Methods

3.2.1. xDLVO-CG Model - Theoretical Framework

The xDLVO-CG model aims to improve upon the spherical protein approximation by incorporating a more accurate representation of protein shapes. This is achieved through the use of CG representations of proteins, which allows for a more realistic modeling of protein shape. The example of CG representation of LYZ protein is shown on Figure 3.1.

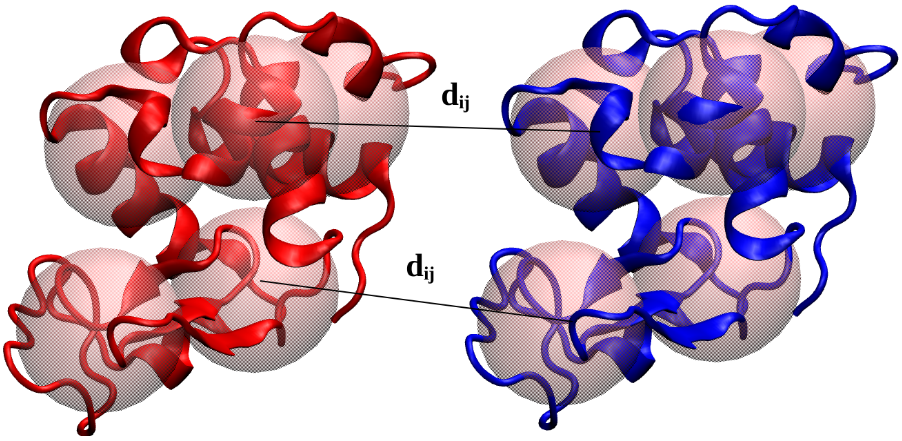


Figure 3.1.: Shape-based coarse-grained model of proteins, shown for lysozyme (PDB code 4nhi) represented by 5 CG beads per protein unit. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry.

Furthermore, the interaction potentials utilized in the model are adapted from the xDLVO model, which is described in greater detail in Chapter 2. To enhance the accuracy further, the model includes an additional potential term, the ion-protein dispersion potential (as explained in Section 3.2.5). This term accounts for the impact of specific ions on protein salting out, as observed in the well-known Hofmeister series. The total interaction potential between two proteins, $W_{22}(r)$, is calculated as the sum of several contributing potential terms, including electrostatic, dispersion (Hamaker or Lennard-Jones), osmotic, and ion-protein interactions, as described by the following equation:

$$W_{22}(r) = \begin{cases} W_{el}(r) + W_{disp}(r) + W_{osm}(r) + W_{i-pr}^{disp}(r) \\ W_{el}(r) + W_{LJ}(r) + W_{osm}(r) + W_{i-pr}^{disp}(r) \end{cases} \tag{3.2}$$

Further details on each of these potential terms are discussed below.

3.2.2. Electrostatic Interactions

Electrostatic interactions between the proteins were calculated using the Debye-Hückel equation,[77] which is represented by the following formula:

$$W_{el}(r) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \frac{Z_i Z_j e^2 \exp\left(\kappa(d_{ij} - r_{ij})\right)}{4\pi\varepsilon_0\varepsilon_r r(1+\frac{\kappa d_{ij}}{4})^2}, \quad r_{ij} > d_{ij} + 2\sigma \tag{3.3}$$

Here, the variable r represents the distance between two proteins' centers of mass. The total number of beads in each protein is denoted by N_1 and N_2, respectively. The initial distance between bead i of the first protein and bead j of the second protein (as per the crystal structure) is represented by d_{ij}, whereas r_{ij} denotes the current distance between beads during protein pulling. The value σ stands for the thickness of the water layer on the protein surface, which is estimated to be approximately 0.1 nm.[150]) Z_i and Z_j denote the charge of CG beads of protein units.

The influence of the low dielectric core, which can lead to short-range image-charge-based repulsion of charged polyelectrolyte chains adsorbed on spherical substrates [151], was not taken into account. For ion-protein electrostatics, the image repulsion is minimal, with only slight contributions noticeable at low salt concentrations, without affecting the B_{22} values. Furthermore, image-charge-based repulsion is negligible at moderate ionic strengths, consistent with previous research in this field [152] [153].

3.2.3. Dispersion Interactions

The dispersion potential $W_{disp}(r)$ describing the attractive van der Waals forces between proteins was calculated using two different approaches. The first approach involved the use of the Hamaker equation (3.4), which describes the van der Waals interaction between two ideal spheres. This equation is derived by summing the London forces over infinitely small volumes between the two spheres, taking into account the polarizability of the molecules and the refractive index of the medium between the spheres [76], and can be calculated according to:

$$W_{disp}(r) = -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \frac{A_H}{12}\frac{1}{N_1 N_2}\left[\frac{d_{ij}}{r_{ij}^2 - d_{ij}^2} + \frac{d_{ij}^2}{r_{ij}^2} + 2\ln(1-\frac{d_{ij}^2}{r_{ij}^2})\right], \qquad r_{ij} > d_{ij} + 2\sigma \tag{3.4}$$

The Hamaker constant, represented by A_H, can be measured through various techniques or estimated using the relation (3.5):

$$A_H = \pi^2 \lambda q_1 q_2, \tag{3.5}$$

It is described by the London-van der Waals constant, λ, and the atom densities of the interacting bodies q_1 and q_2. The London-van der Waals constant is obtained by measuring the polarizabilities and characteristic absorption frequencies of the solute and is related to the refractive indices[154].

Table 3.1 provides a list of the Hamaker constants, denoted as A_H, utilized in this study.

Table 3.1.: The values of the Hamaker constant, A_H, utilized in the calculation of the dispersion potential based on equations (2.7) and (3.4). The values for LYZ, BSA, Subs and IgG1 were obtained from literature sources[155, 125, 156, 157].

Protein	Conditions	A_H [k_BT]
LYZ	All pH and salts	8.0
BSA	All pH and salts	3.0
Subs	NaCl	5.1
Subs	NaSCN	6.8
IgG1	All pH and salts	3.0

In an attempt to find a parameter-free substitute for the Hamaker equation, the Lennard-Jones (LJ) potential was implemented. This approach, labeled xDLVO-CG(LJ), does not require experimentally measured values of A_H and can be expressed as follows:

$$W_{LJ}(r) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.6}$$

The Lennard-Jones parameters of each bead, ε_i and σ_i, were assigned according to Arkhipov et al. as implemented in the shape-based coarse-grained (SBCG) builder in VMD.[158] The interaction strength of each bead was assigned based on the hydrophobic solvent accessible surface area (SASA) for the protein domain represented by a bead, which can be expressed as follows:

$$\varepsilon_i = \varepsilon_{max}\left(\frac{SASA_i^{hydroph}}{SASA_i^{total}}\right)^2 \tag{3.7}$$

Here, ε_{max} is the interaction constant (often taken as 10 kcal/mol), while SASA$i^{hydroph}$ and SASAi^{total} are the hydrophobic and the total SASA of a bead i. For the B_{22} calculations in this Chapter, ε_{max} of 10 kcal/mol and 2 kcal/mol were used for the small protein (LYZ) and other proteins (Subs, BSA, IgG1), respectively.

The LJ radius for each CG bead was determined as the radius of gyration of the corresponding atoms, increased by a constant value of 0.1 nm to simulate the atoms at the edge of the bead. The LJ parameters for each bead pair were obtained using standard combination rules based on the bead's hydrophobic solvent accessible surface area (SASA). The LJ interaction strength between a pair of beads is primarily determined by their SASA, with a stronger interaction between two hydrophobic beads. The LJ parameters between specific bead pairs were calculated using the following combination rules:

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j},\ \sigma_{ij} = 0.5\,(\sigma_i + \sigma_j) \tag{3.8}$$

Here, ε_{ij} represents the interaction strength between beads i and j, while σ_{ij} is the LJ radius between them.

3.2.3.1. The Effect of Hamaker vs. Lennard-Jones Potentials on Calculated B_{22} Values

Several expressions are available to calculate dispersion interactions between proteins based on continuum models. The Hamaker constant, A_H, is commonly used as an adjustable parameter in (x)DLVO models and is fitted to experimental data at different pH and salt conditions. While A_H can be calculated using McLachlan's formulation of Lifshitz's van der Waals forces, which accounts for excess polarizability in a dielectric medium,[159] this approach is less widely used. The resulting B_{22} coefficients obtained using the fitted A_H show good agreement with experiments. However, these adjustments to A$_H$ include error compensations arising from the incorrectly modelled electrostatic, osmotic, and ion-protein dispersion terms. In this study, xDLVO-CG was validated using Hamaker constants previously reported (Table 3.1 without changing their values at different salt and pH conditions (two A_H used only for Subs). The approach showed good correlations between the calculated and experimental B_{22}, as discussed in more details in Section 3.3. However, with the aim was to introduce the protein solubility model using standard potential energy functions without designed fitting to the experimental data points, an attempt was made to replace Hamaker dispersion interactions in the total PMF by the Lennard-Jones interactions , parameterized according to Arkhipov et al.[160], as described in Section 3.2.3. The dispersion potentials for LYZ at pH 7 and Subs at pH 5.5 are shown in Fig.3.2. The strength of the dispersion interactions, introduced by the Hamaker constant, decreases faster than in the case of the LJ interactions. This results in stronger binding of the proteins in the LJ potential model and a small shift of the calculated second virial coefficients towards negative values (see the right panel in Fig.3.2). Overall, the potential depth and the corresponding center of mass (COM) distance of the most attractive interactions between the proteins are similar in both cases, resulting in B_{22} coefficients in the range of the reported experimental data. Analysis of the B_{22} changes as a function of the sampled protein-protein structures indicates higher sensitivity of the PMF with LJ, which arises from the fact that all bead-bead dispersion interactions were parametrized based on the chemical composition of the residues included in a CG bead, and not by uniform values of the Hamaker constant.

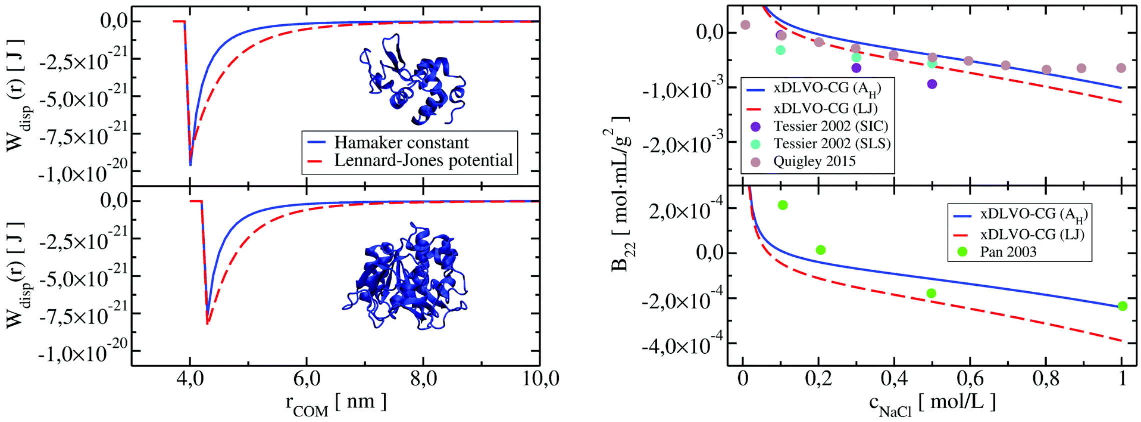


Figure 3.2.: Dispersion potentials for LYZ (upper panel) and Subs (lower panel) represented by the Hamaker constant (in blue), and by the Lennard-Jones potential (in red). The corresponding B_{22} coefficients for LYZ at pH 7 and Subs at pH 5.5 using both potentials are shown on the right panel. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data, labeled as 'Quigley 2015' and 'Pan 2002' were taken from [127, 125].

3.2.4. Osmotic Potential

The attractive interaction between proteins due to ion exclusion from the protein interspace at short distances[79] was computed using equation:

$$W_{osm}(r) = -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{1}{N_1N_2}\frac{4\pi k_B}{3}TD_{ij}^3\rho_3(1-\frac{3r_{ij}}{4D_{ij}}+\frac{r_{ij}^3}{16D_{ij}^3}), \qquad d_{ij}+2\sigma \leq r_{ij} \leq 2D_{ij} \tag{3.9}$$

Here, the radius of a hydrated salt, R_3, is the sum of the hydrated anion and cation radii, and ϱ_3 is the salt density. D_{ij} is defined as the sum of d_{ij}, the distance between two CG beads, and R_3 and σ. The hydrated anion and cation radii values were taken from Marcus et al.[161] (as shown in Table 3.2.5).

3.2.5. Ion-protein Dispersion Interactions

The nature of the salt used has an impact on PPIs, which is reflected in the value of B_{22}.[129] The Hofmeister series describe the propensity of certain salt ions to cause protein precipitation, and is related to the interactions between the ions and the protein[162, 163]. Despite their importance, these interactions are not considered in most solubility models, including (x)DLVO. However, Ninham and his colleagues have shown that incorporating ion-macroion dispersion interactions into the PMF leads to more accurate modeling of salt effects[164, 165, 166]. This effect is significant even for monovalent ions of similar size, and modifies the dependence of B_{22} on ionic strength. To account for these interactions, a new term representing the ion-protein dispersion potential was added to the total PMF in xDLVO-CG. Since the size of cations and anions is much smaller than that of proteins, their interactions with proteins can be approximated as the interaction of a small point charge with an infinite plane:[166, 167]

$$W_{i-pr}^{disp} = -\frac{B_{i-pr}}{r^3} \tag{3.10}$$

where B_{i-pr} is the characteristic ion-protein dispersion coefficient, which depends on the polarizability of both the ion and the protein.[164] The total ion-protein dispersion energy is then obtained by summing the contributions from all individual ion-protein interactions, while accounting for the non-uniform distribution of ions around the protein. The distribution of charges is determined by the Boltzmann distribution, which leads to attraction of ions of opposite charge towards the protein surface and repulsion of ions of the same charge. The number of ions in a spherical shell around the protein is given by

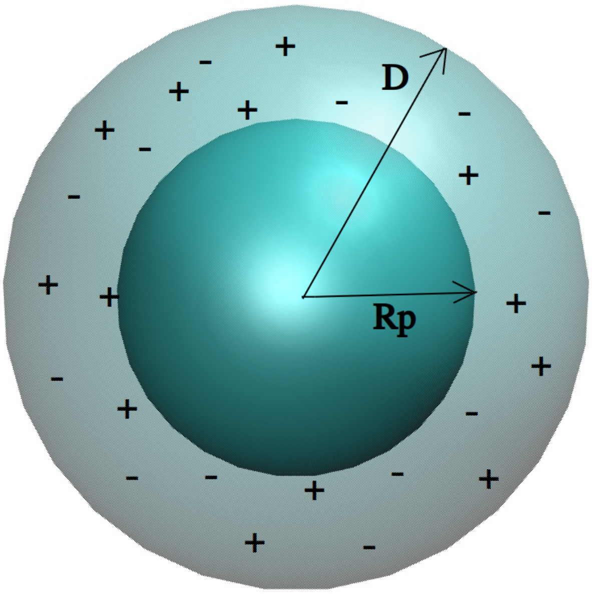


Figure 3.3.: A continuum model of ion-protein dispersion interactions is illustrated. R_p and D denote the protein radius and maximum thickness of a shell when ion-protein interactions are taken into account. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry.

$c_{bulk} \exp\left(-z\phi(r)/r\right) dV$, where $dV = 4\pi r^2 dr$ is the shell volume, z is the charge of an anion or a cation. The electrostatic potential $\phi(r)$ felt by an ion at a distance r from the protein center is approximated by equation (2.8). To obtain the total ion-protein dispersion, the contributions of all ions in the spherical shell are integrated:

$$\begin{aligned} W_{i-pr}^{disp}(r) = -4\pi B_a \int_{R_P+\sigma}^{R_P+D} \frac{c_b \exp\left(-z_a\phi(r)/(k_BT)\right)}{r} dr \\ -4\pi B_c \int_{R_P+\sigma}^{R_P+d} \frac{c_b \exp\left(-z_c\phi(r)/(k_BT)\right)}{r} dr \end{aligned} \tag{3.11}$$

The variables used in the equation are defined as follows: R_p represents the radius of the protein, σ represents the thickness of the water layer surrounding the protein, D represents the maximum thickness of a shell at which ion-protein interactions are considered (as shown in Figure 3.3, and B_a and B_c represent the dispersion coefficients for anions and cations, respectively. The value of D was chosen to be 1 nm, while the anion- and cation-protein dispersion coefficients, B_a and B_c, were obtained from previous studies[166, 164] and are presented in the Table 3.2.5.

3.2.5.1. Impact of Including Ion-protein Dispersion Interactions Term

xDLVO model typically considers the influence of salts on PPIs through osmotic and electrostatic potentials. The electrostatic potentials calculated by Debye-Hückel theory are unable to distinguish between salts of the same valency. On the other hand, the osmotic attraction potential utilizes hydrated salt radii as input, which are only slightly different between different salts. However, since the radii of hydrated monovalent salts are similar, the osmotic term alone cannot explain the observed salt-dependence of the

Table 3.2.: The values of the hydrated salt radii, R_3, utilized in the calculation of the osmotic potential based on equations (2.10) and (3.9), as well as the ion-protein dispersion coefficients, B_{i-pr}, used in the calculation of the ion-protein dispersion potential according to equation (3.11). The values of R_3 and B_{i-pr} were obtained from references [161, 166, 164].

Salt	R_3 [nm]	Ion	$B_{i-pr} \cdot 10^{-50}[J \cdot m^3]$
NaCl	0.442	Na^+	0.454
KCl	0.436	K^+	1.888
NaI	0.464	Cl^-	3.574
NaSCN	0.460	I^-	4.440
		SCN^-	10.000

osmotic second virial coefficients B_{22} of proteins. The Hamaker constant, which describes London-van der Waals interactions between charge-neutral proteins and should not depend on ionic strength,[76, 154] is usually fitted to experimental data to recover the trends according to the Hofmeister series, which reduces predictive power of models. The ion-protein dispersion interaction, which depends on the properties of salt ions and not the protein type, is responsible for promoting attractive PPIs.[168] The impact of including ion-protein dispersion in the xDLVO-CG model can be demonstrated by the resulting B_{22} coefficients for LYZ and Subs, which are shown in Figure 3.4, with and without ion-protein interactions. The omission of the ion-protein dispersion term in the calculation of B_{22} values results in limited differentiation between different salts and fails to replicate the marked differences observed in experimental data. In contrast, the ion-protein dispersion term in the PMF differentiates the salting out efficiency of ions and induces stronger separation between the data obtained using different salts, leading to better correlation with experiments. The model also agrees with the Hofmeister series trend, i.e., $\mathrm{Cl}^- < \mathrm{I}^- < \mathrm{SCN}^-$, where thiocyanate anion is known as a strong salting out agent causing protein precipitation. This is due to the higher ion-protein dispersion coefficients of SCN^-, which is 2.8 times larger than in the case of Cl^-,[164] resulting in stronger ion-protein dispersion interactions. The model is consistent with observations that I^- ions show stronger salting out efficiency of LYZ than Cl^-, indicating the model's agreement with the Hofmeister series trend.[80, 169] These results highlight the importance of accounting for ion-protein dispersion interactions in accurately predicting protein-protein interactions.

3.2.6. Shape Based Coarse Grained Model for Proteins

To account for the anisotropic nature of protein interactions and the variety of configurations that contribute to the average PMF, a shape-based coarse-grained (SBCG) model of proteins was used, which was constructed from the corresponding all-atom structures. To validate the model, four proteins were selected - LYZ, Subs, BSA, and IgG1 - and experimental and theoretically computed B_{22} data were compared. All-atom structures of these proteins (PDB codes 4nhi, 4f5s, 1ndu, and 1mco, respectively) were obtained, which contained all the necessary residues and had at least two identical proteins in the unit cell. Additionally, the effects of pH were incorporated by using the PROPKA method,[170, 171] which was implemented in the PDB2PQR web server,[172, 173] to add hydrogen atoms at the desired pH. The all-atom structures of these proteins (PDB codes 4nhi, 4f5s, 1ndu, and 1mco, respectively) were obtained, which contain all the necessary residues and have at least two identical proteins in the unit cell. A shape-based model with a self-organising neural network topology building algorithm was used to map all-atom structures of proteins to CG representations,[160] as implemented in the VMD program (version 1.9.3).[174] The learning algorithm was initialised with two variables, ϵ and λ, and the starting and final

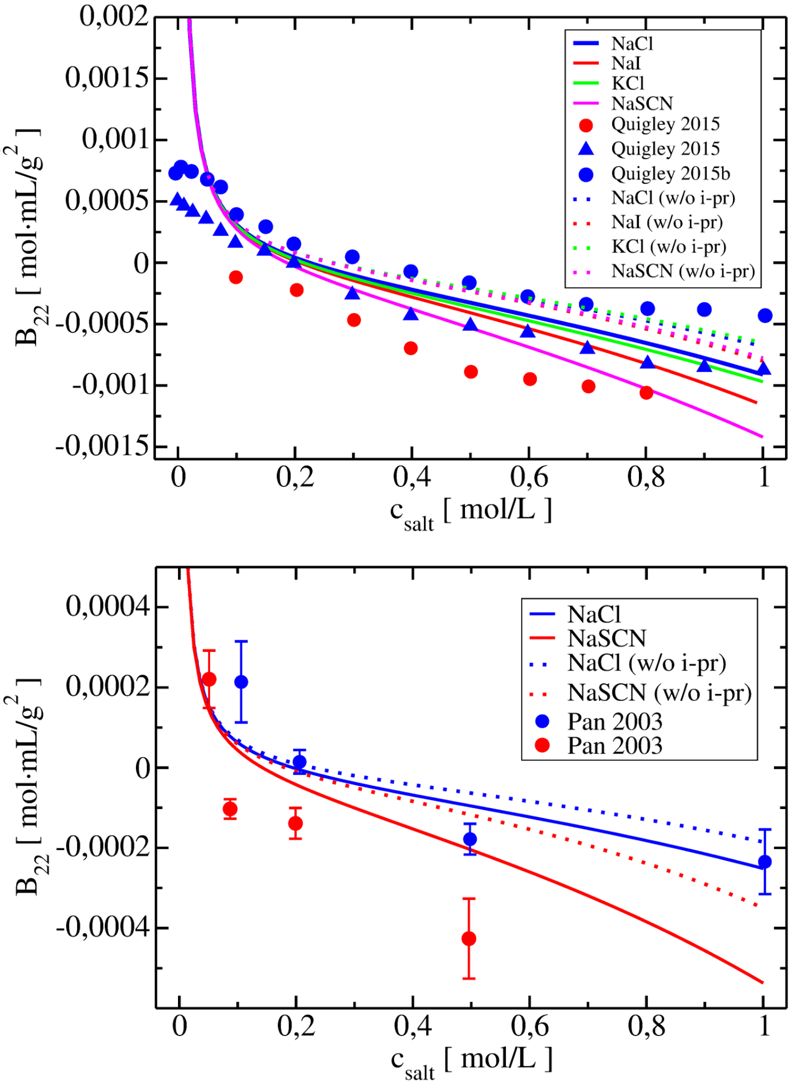


Figure 3.4.: The change in the calculated B_{22} coefficients after adding the ion-protein dispersion interactions term to the potential of mean force of LYZ at pH 4.5 (upper panel) and Subs at pH 5.5 (bottom panel). The B_{22} values calculated using xDLVO-CG are represented with solid curves, while the results from xDLVO-CG without ion-protein interactions are marked with dotted dashed curves. The experimental B_{22} coefficients are denoted with colored circles. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data, labeled as 'Quigley 2015' and 'Quigley 2015b' respectively represent the first moment and maximum peak dynamic light scattering measurements taken from [127]. Data labeled as 'Pan 2003,' were obtained from [125] for each respective pH.

values of ϵ and λ were set to 0.3 and 0.05, and 20 and 0.01, respectively. Each CG bead represented approximately 500 atoms of a given protein, resulting in 5, 10, 20, and 40 CG beads for LYZ, Subs, BSA, and IgG1 (see Figure 3.1 and 3.11), respectively. The CG beads were placed at the COM of the atoms they represented, and their charges were calculated as the sum of the partial charges of the atoms within each bead, determined by PROPKA.

3.2.6.1. Impact of Coarse-Graining Size on Results

Several mapping schemes were evaluated and compared to determine the optimal coarse-graining size for the calculation of B_{22} values. The mapping schemes included a 500-to-1 mapping, a 300-to-1 mapping, a 200-to-1 mapping, and a 150-to-1 mapping, as shown on

Figure 2.3 The results indicate that the choice of coarse-graining size had little impact on the calculated B_{22} values, with the 500-to-1 mapping scheme producing similar results to the 150-to-1 mapping scheme. Moreover, using the 500-to-1 mapping scheme reduced the computational cost of calculating B_{22} by over 80%, from 57.76 minutes to 7.7 minutes for one relative orientation of BSA. Based on these findings, the 500-to-1 mapping scheme was chosen as the optimal coarse-graining size for all proteins.

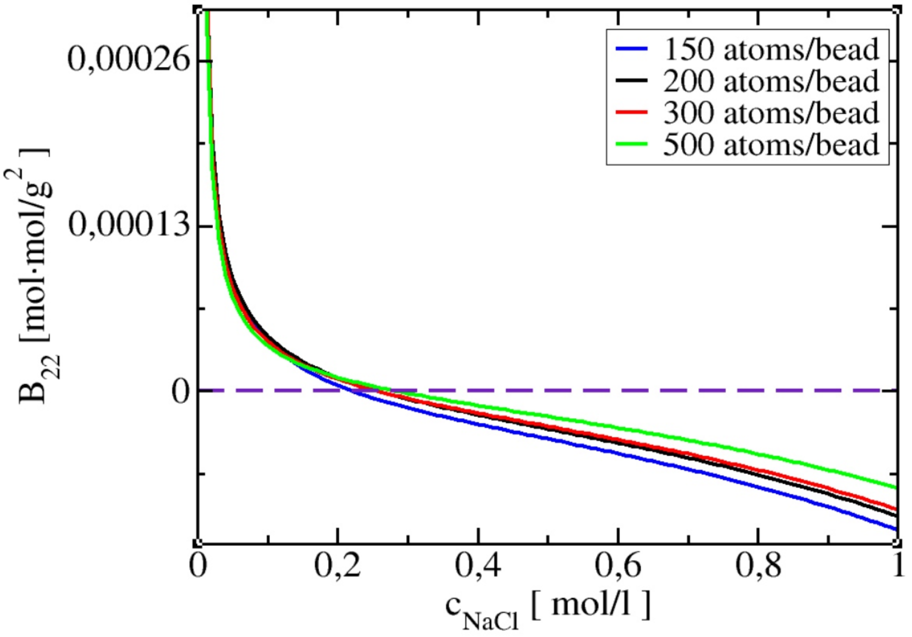


Figure 3.5.: The effect of coarse-graining size: the number of atoms per CG bead, on calculated B_{22} values for BSA at pH 7.4 in NaCl solution. Reprinted with permission from [117]. Copyright 2021 Royal Society of Chemistry.

3.2.7. Protein-protein sampling scheme

Equation 2.3 states that the total B_{22} coefficient is dependent on the protein-protein interactions at different orientations, and therefore requires the sampling of diverse protein configurations to calculate accurately. Two sampling schemes were utilized in this study. The first involves sampling of the PPIs by pulling the protein along a linear trajectory outwards from the crystal structure up to a distance of $R_0 + 30$ nm. The other scheme employs a statistical sampling approach over the configuration space to sample protein pair interactions for relevant relative arrangements. In this approach, the initial crystal structure of the protein is used, with the first protein kept fixed at position (x_1, y_1, z_1), while the second protein is moved uniformly around the first one on the fixed distance r_0. The radial sampling of the second protein is performed by varying the $(\theta,\ \phi,\ r)$ coordinates in the spherical coordinate system with the center at (x_1, y_1, z_1). Ten values of θ, ϕ angles were taken uniformly from $\theta = [0, 2\pi]$ and $v = [-1, 1] \rightarrow \phi = -\arccos(v)$ intervals, which resulted in 83 unique starting configurations (including the protein position in the crystal). In addition, the second protein was subsequently rotated by five different angles $\left[\frac{\pi}{3}, \frac{2\pi}{3}, \pi, \frac{4\pi}{3}, \frac{5\pi}{3}\right]$ around $(x - x_2)$, $(y - y_2)$ or $(z - z_2)$ axis, respectively. These resulted in 16 differently rotated structures, therefore, in the total amount of different starting protein configurations of 83·16(=1328). From these starting configurations, again the protein was pulled along a linear trajectory outwards, up to a distance of R_0+30 nm, along the line connecting the respective center-of-mass. All configurations were checked for possible steric overlap, and overlapping structures were excluded from further calculations. The total PMF is obtained by averaging all angular orientations defined by the starting structures, resulting in the PMF as a function of the COMs of proteins: $W_{22}(r, \Omega_1, \Omega_2) \rightarrow W_{22}(r)$. The second osmotic coefficient as a function of ionic strength is calculated by numerical integration of the averaged PMF at different salt conditions according to equation (2.3).

3.3. Results

The xDLVO-CG model was validated by calculating the osmotic second virial coefficient for four proteins with varying sizes, shapes, and physicochemical properties. The choice of proteins was based primarily on the availability of experimental measurements of B_{22}, although a literature survey revealed significant variations in the experimental B_{22} values for the same protein under similar solution conditions, often due to the use of different measurement techniques. The computed B_{22} results were compared to the experimental B_{22} values from several reports. Additionally, the xDLVO-CG model was compared with other theoretical models for B_{22} calculations, including a spherical xDLVO model that replicated the model reported in Ref. [175] and FMAPB2, an all-atom model for B_{22} calculations in implicit solvent.[147] FMAPB2, which provides higher structural resolution compared to spherical or CG models, was evaluated using a publicly available web-server (http://pipe.rcc.fsu.edu/fmapb2).

3.3.1. Lysozyme

LYZ is a small protein with a globular shape which has been extensively studied in various scientific investigations and has found practical applications as an antibacterial agent. Figure 3.6a and 3.6b depict the calculated B_{22} coefficients for LYZ as a function of ionic strength at two different pH values of 4.5 and 7. The calculated values were compared to the experimentally measured values. [127, 176, 177, 133, 178, 123] The B_{22} values for LYZ are positive at lower ionic strengths, indicating that the effective interactions are repulsive, resulting in stable solutions. However, as the ionic strength increases, the B_{22} values decrease towards negative values, and at approximately 0.23 M NaCl at pH 4.5, the B_{22} value crosses the zero point. This indicates a switch from repulsive to attractive interactions, which may lead to protein aggregation. For LYZ at pH 4.5 and 10 mM NaCl, the electrostatic potential is approximately 5 times stronger than the dispersion potential. This dominance of repulsive interactions at 10 mM NaCl results in a positive B_{22} value, as shown in Figure 3.6. The increase in salt concentration permits opposite charge screening, reducing the strong electrostatic repulsion between the equally charged protein species. This reduction in electrostatic repulsion allows the attractive dispersion interactions to be more pronounced, inducing molecule self-assembly. As seen in Figure 3.6c, the W_{el} value significantly decreases with an increase in NaCl concentration, indicating the reduction in electrostatic repulsion, so the W_{disp} interactions become more prominent. The influence of osmotic and ion-protein interactions is relatively low, but it becomes more significant at higher ionic strengths where protein salting out can occur. Eventually, this leads to a strong attraction between LYZ molecules, as indicated by the increasingly negative PMF shown in Figure 3.6d. Furthermore, the B_{22} values for LYZ are slightly more negative at pH 7 compared to pH 4.5, and the crossing point occurs at around 0.18 M NaCl instead of 0.23 M NaCl. This behavior is attributed to the fact that the protein charge at pH 7 is reduced from +11e to +9e, requiring fewer counterions to screen the electrostatic repulsion between LYZ protein pairs and making the propensity to aggregation more favorable. The data obtained from the xDLVO-CG model correlates well with experimental data for B_{22} values of LYZ, which between themselves show some discrepancies due to differences in the experimental methodology.[127, 176, 177, 133, 178, 123, 179] Interestingly, using a single starting protein-protein configuration from the crystal led to results that were nearly identical to those obtained via the sampling protocol, likely due to the globular and symmetrical shape of LYZ with low charge anisotropy. The xDLVO model accurately

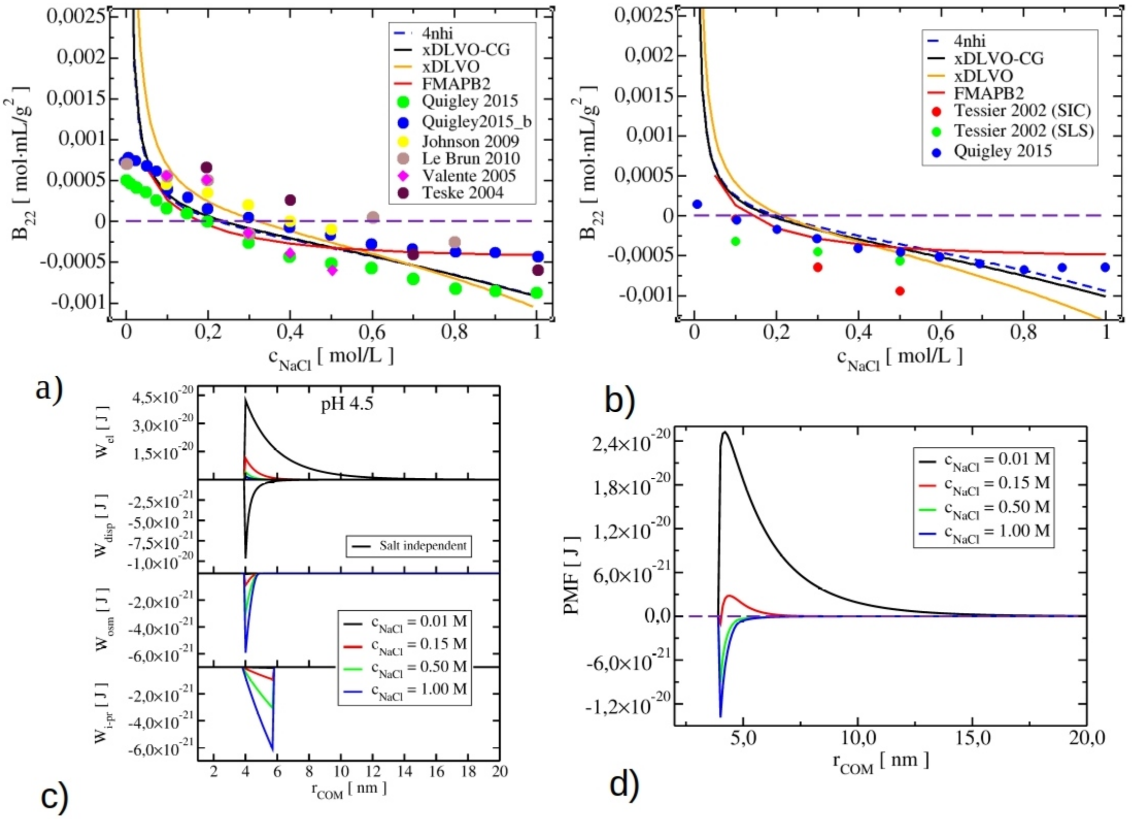


Figure 3.6.: a) The B_{22} coefficients for LYZ at pH 4.5 and b) pH 7 with varying NaCl concentrations computed using xDLVO-CG, xDLVO(spherical model), and FMAPB2 (all-atom representation). The data labeled with a dashed blue curve indicates that the xDLVO-CG calculation was performed for only one relative orientation (based on the crystal structure), otherwise the B_{22} coefficients were calculated for the sampled configurations. The experimental data is represented by colored circles. c) The change in each potential (as given by Equations 3.3-3.11) for LYZ at pH 4.5 as a function of NaCl concentration. d) The total PMF at pH 4.5 as a function of NaCl concentration. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data, labeled as 'Quigley 2015' and 'Quigley 2015b' respectively represent the first moment and maximum peak dynamic light scattering measurements taken from [127]. Data labeled as 'Johnson 2009,' 'Le Brun 2010,' 'Valente 2005,' 'Teske 2004,' and 'Tessier 2002' were taken from [176, 123, 177, 133, 178]

predicts B_{22} values of LYZ owing to its spherical shape, but it tends to overestimate values at higher ionic strengths. Meanwhile, the FMAPB2 model, which employs an all-atom protein representation, provides similar results to the xDLVO-CG model but has a higher computational cost. Calculating B_{22} values using xDLVO-CG and FMAPB2 takes approximately 1.85h (for 100 salt concentrations) and 4 h (for one salt concentration), respectively. The xDLVO-CG model agrees well with experimental data also at other pH values (pH 3, pH 5, and pH 8), as shown in Figure 3.7. However, at pH 3 and 300mM NaCl, both xDLVO and xDLVO-CG models exhibit larger deviations from experimental data, as observed in the left histogram of Figure 3.7. Kalyuzhnyi et al. reported similar observations when they explicitly accounted for all interacting species in the model. Both Kalyuzhnyi et al. and this study reported overestimation of calculated B_{22} at pH values greater than 7 for lysozyme at pH 8, possibly due to ion-specific effects that the model cannot currently account for.[180]

In summary, LYZ protein behavior in solution is highly dependent on the ionic strength and pH values. With an increase in salt concentration, the PPIs become attractive through

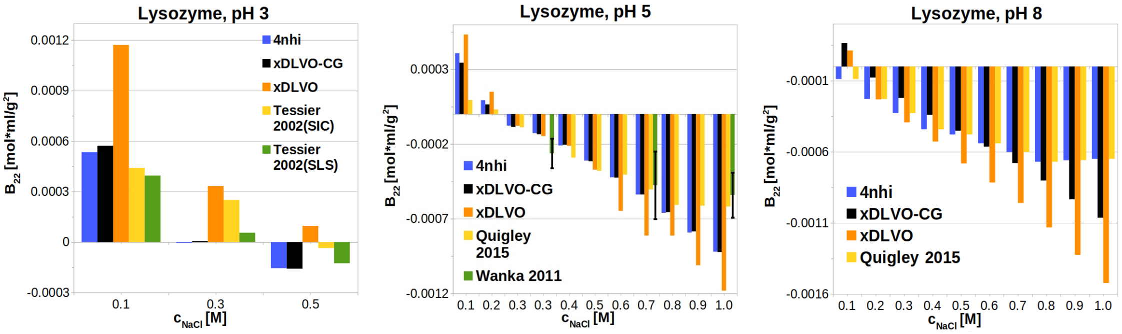


Figure 3.7.: The B_{22} values for LYZ at pH 3, pH 5 and pH 8 computed using xDLVO-CG. Two sampling methods were used: 4nhi, which refers to calculations for proteins from the crystal, and xDLVO-CG with orientational sampling. The results were compared to those obtained using xDLVO with the protein approximated as a sphere. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. Experimental data labeled as 'Quigley 2015,' 'Tessier 2002,' and 'Wanka 2011' were taken from [127, 178, 179], respectively.

dispersion interactions, leading to aggregation. The behavior is more pronounced at higher pH values, as the protein charge is higher, resulting in a stronger screening of electrostatic repulsion.

3.3.2. Validation on Nonspherical Proteins

LYZ is a compact protein that has a spherical shape, making it easy to use xDLVO theory to predict its solubility or likelihood to aggregate. However, predicting the solubility of non-spherical proteins like Subs, BSA, and IgG1 using xDLVO theory is more challenging. Therefore, we validated the accuracy of the xDLVO-CG model for these three proteins. Figure 3.8 shows a simplified representation of the proteins using CG models, where the protein monomers used to sample PPIs are colored blue and red, respectively.

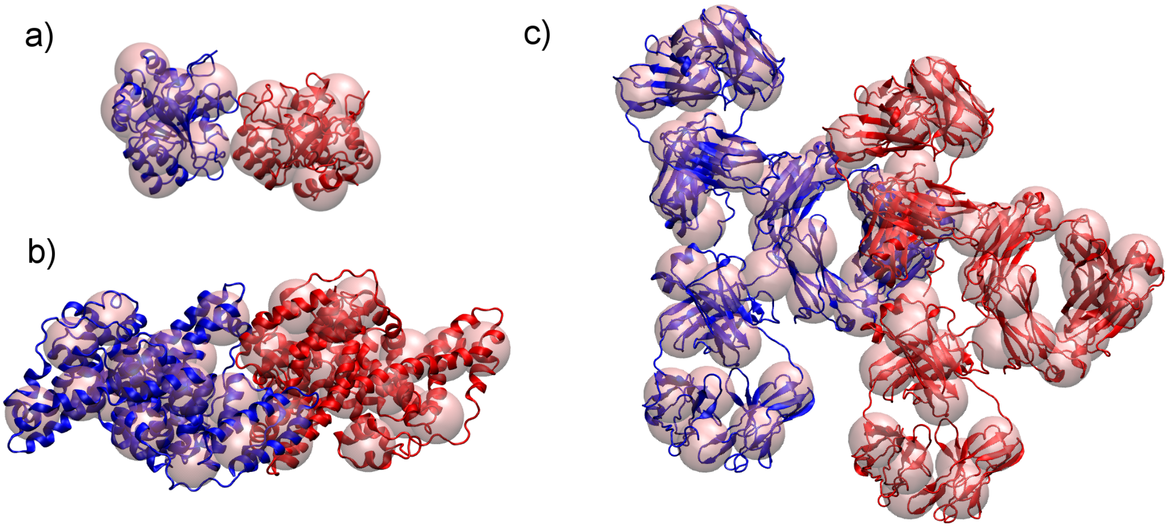


Figure 3.8.: The coarse-grained models of a) subtilisin, b) bovine serum albumin, and c) immunoglobulin type 1 employed in the xDLVO-CG model to calculate the osmotic second virial coefficients. The cartoon representation of the all-atom structure of the proteins is also included to ensure clarity.

3.3.2.1. Subtilisin

Subs is the digesting enzyme used in commercial application as an an engineered variant (called properase). There is a scarcity of experimental data for the second virial coefficients of Subs, with measurements only being done at pH 5.5 for properase variant.[125] Therefore, B_{22} was calculated as a function of ionic strength in the presence of 0-1 M NaCl, as illustrated in Figure 3.9. Significant differences between calculations made with xDLVO and FMAPB2 in comparison to xDLVO-CG were observed for the B_{22} data of Subs at pH 5.5. The B_{22} crossing point, indicating a shift between attractive and repulsive interactions, was shifted to the lower ionic strengths in xDLVO and showed stronger increase of attractive PPIs towards aggregation. This was caused by the representation of the charge distribution by an uniform sphere in xDLVO, whereas the protein has significant anisotropy, which xDLVO-CG better represents. B_{22} from xDLVO, with the experimentally fitted Hamaker constant of 5.1 k_BT, matched the experimental B_{22} at 500 mM NaCl (-1.83·10^{-4} and -1.78·10^{-4} mol·mL/g^2, respectively), but this was the only data point that matched. Another parameter set for xDLVO was used to calculate B_{22} by Pan et al.[125]. The charge of Subs was calculated using the MacroDox program and the Hamaker constant was fitted to achieve nearly quantitative agreement between the experimental measurements and their xDLVO. The protein charge they obtained was higher by 2.7-3.5e (i.e., was +8.7 - +9.42e depending on salt concentration) than that obtained using PROPKA (+6e). Since there is no information about the Subs charge from experimental measurements, the protonation schemes used in either approach cannot be validated. B_{22} calculated using FMAPB2 was shifted far to the negative B_{22} range, deviating far from the experimental B_{22} values. To accurately predict the behavior of Subs at higher pH levels, we conducted calculations based on the protein's properties at pH 7. At this pH, the positive charge of the protein decreased from +6 to +4, which had a significant impact on its aggregation potential (as shown in Figure 3.10). The decrease in positive charge resulted in stronger aggregation, indicating that Subs may be more prone to clumping and precipitation under alkaline conditions. These findings highlight the importance of understanding the pH-dependent behavior of proteins like Subs, particularly in applications where they may be exposed to

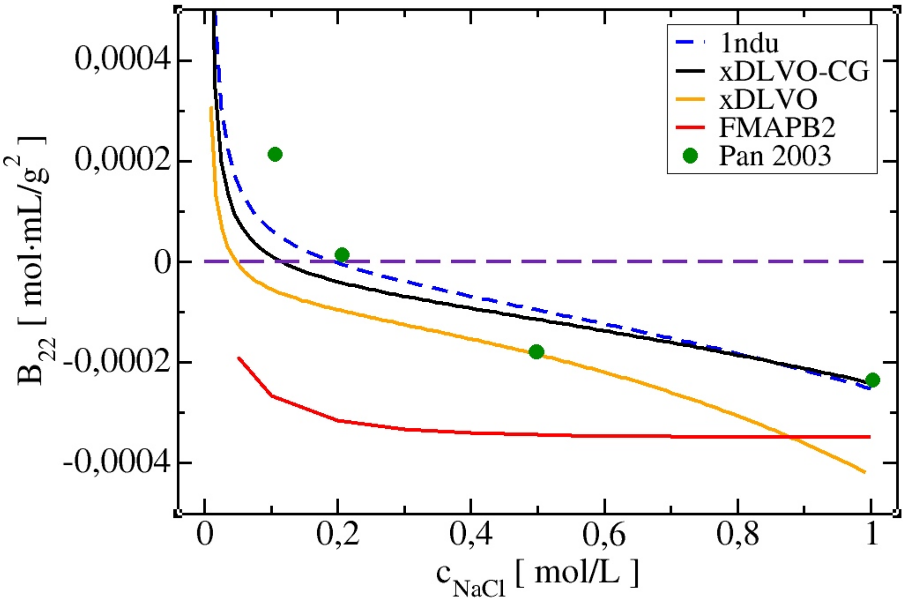


Figure 3.9.: The B_{22} coefficients for Subs at pH 5.5 as a function of NaCl concentration calculated using the xDLVO-CG, xDLVO, and FMAPB2 models. Circles represent experimental data. Reprinted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data, labeled as 'Pan 2003' were obtained from [125].

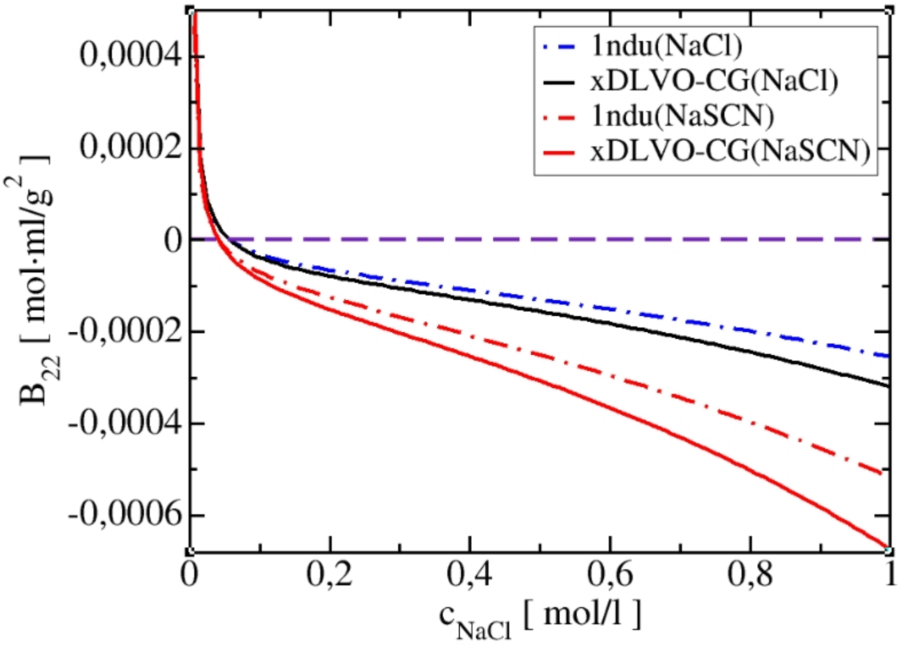


Figure 3.10.: The B_{22} coefficients for Subs at pH NaSCN as a function of NaCl concentration calculated using the xDLVO-CG model. Reprinted with permission from [117]. Copyright 2021 Royal Society of Chemistry.

varying pH environments.

3.3.2.2. Bovine Serum Albumin

The BSA protein belongs to serum albumins that can bind to a variety of substances, including drugs, nutrients, and metals, making them useful in clinical, pharmaceutical, and biochemical applications. The B_{22} calculations of BSA at pH 6.2 and 7.4 were performed and compared with experimental values reported in various references[134, 135, 181, 182], as shown on Figure 3.11. xDLVO-CG reproduces the main trends of the experimental data at lower ionic strength and with the correct trends using sampling of various protein relative orientations. The BSA charge calculated using PROPKA deviates from experimental values, especially at pH 6.2, so H++ protonation was used instead[183], resulting in a slightly better charge estimate. Both protonation schemes underestimate the BSA charge, leading to larger discrepancies between calculated B_{22} and experimental data.[184]) In addition to the discrepancies between calculated and experimental B_{22} values, the experimental data itself is also inconsistent. Different studies have reported different measured values, leading to further uncertainty in the interpretation of the results. This may be due to variations in experimental conditions, such that different experimental techniques or data analysis methods may contribute to the variability in reported values. Therefore, careful consideration and comparison of experimental data from multiple sources is necessary for a comprehensive understanding of PPIs in case of BSA. The calculated B_{22} values using FMAPB2 show a positive shift, similar to the observations for LYZ. However, the B_{22} values obtained using FMAPB2 are consistent with the findings reported by Qin et al. [147]. xDLVO fails to reproduce the B_{22} coefficients at salt concentrations above 100 mM NaCl, even with a better estimate of the total charge. The accuracy of xDLVO results is strongly influenced by the choice of protein radius. For BSA, the experimentally determined radius varies between 3.5-4.1 nm depending on pH and ionic strength[185]. However, using two different radii in xDLVO calculations still failed to capture the experimental trends for B_{22}. The aforementioned results highlight the limitations of xDLVO in accurately predicting

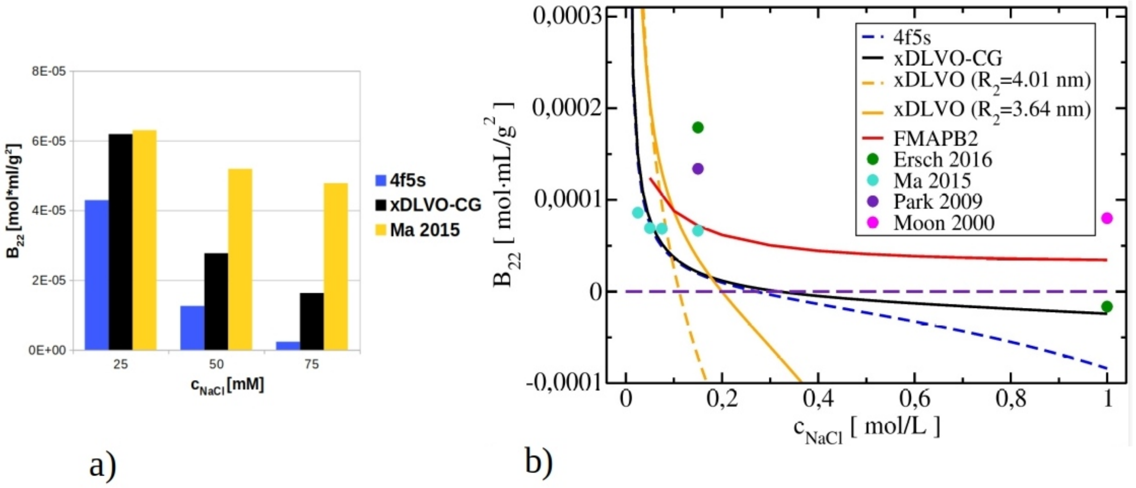


Figure 3.11.: The B_{22} coefficients for BSA at a) pH 6.2 and b) pH 7.4 as a function of NaCl concentration calculated using the xDLVO-CG, xDLVO, and FMAPB2 models. Circles represent experimental data. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data labeled with 'Ma 2015' were taken from [135], while the data labeled with 'Ersch 2016,' 'Park 2009,' and 'Moon 2000' were taken from [134, 181, 182], respectively.

PPIs, as even small variations in the protein radius, along with other parameters, can significantly impact the predicted behavior.

3.3.2.3. Human Immunoglobulin Type I

The complexity of proteins plays a critical role in determining protein-protein interactions and the accuracy of computational approaches under various conditions. The challenges presented by complex proteins are particularly evident in the case of IgG1, which is a member of the monoclonal antibody group and has a unique T-shape. Monoclonal antibodies play a critical role in the immune system's defense by recognizing and binding to specific antigens. The pharmaceutical industry is still facing difficulties in preparing stable formulations of monoclonal antibodies, making the calculation of B_{22} values for IgG1 an essential tool for developing more stable formulations. Several studies have investigated the use of extended DLVO theory to model the second virial coefficients of monoclonal antibodies under different solution conditions and protein concentrations. These studies include those by Calero Rubio et al. [140, 141, 142], Roberts et al. [119], and Singh et al. [186]. In most cases, the models used a direct parametrization from experimental data and various levels of coarse-graining [140]. Figure 3.12 depicts the calculated B_{22} values for IgG1 as a function of NaCl concentration and at three different pH values (pH 5, 5.75, and 6.5), and compared them with experimental values reported in Ref[187]. These values provide insight into the behavior of IgG1 and can aid in the development of more stable formulations. IgG1 is a highly charged protein with charge +56e, +48e, and +27e at pH 5, 5.75, and 6.5, respectively. Therefore, the values of B_{22} are more positive compared to LYZ. The developed xDLVO-CG method was able to reproduce the main experimental trends in the B_{22} values of IgG1, despite representing the large protein using only 40 CG beads that mimicked its basic shape and charge anisotropy (as shown in Figure 3.8c). The CG representation of IgG1 was found to be a critical factor that affected the calculated B_{22} values, as a simpler spherical-based model like xDLVO failed to replicate the experimental trends of non-globular proteins without using parameters fitted from experiments. The all-atom FMAPB2 model yielded similar B_{22} values to xDLVO-CG at lower ionic strengths (as marked in red in Figure C.7), but deviated towards more positive values at higher ionic strengths. Since no B_{22} experiments have been conducted at higher ionic strengths, it is

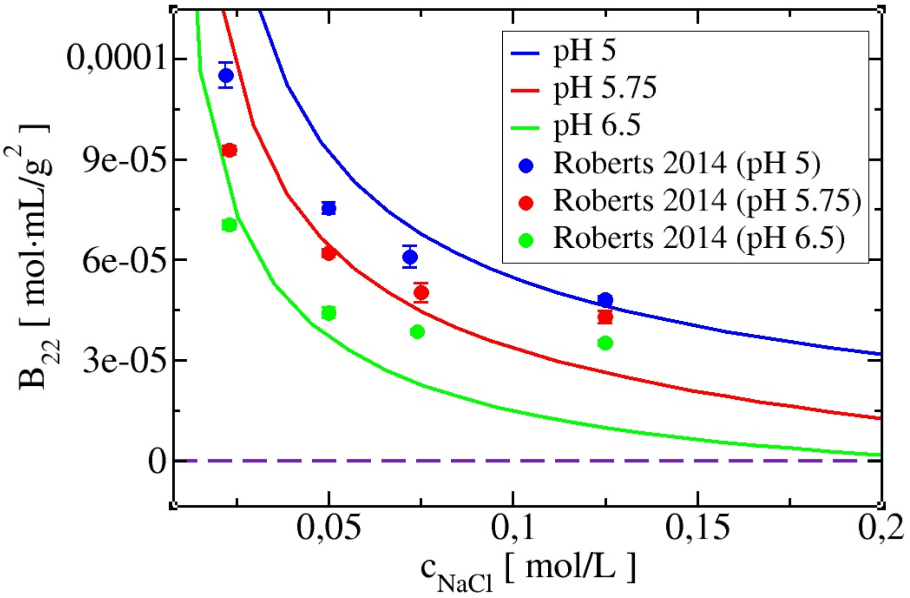


Figure 3.12.: The calculated osmotic second virial coefficients for IgG1 at pH 5, 5.75, and 6.5 using xDLVO-CG compared with experimental data, which is represented by colored circles. Adapted with permission from [117]. Copyright 2021 Royal Society of Chemistry. The experimental data, labeled as 'Roberts 2014 (pH 5)', 'Roberts 2014 (pH 5.75)' and 'Roberts 2014 (pH 6.5)' were taken from [187].

more challenging to compare the models. Furthermore, it was observed that the effect of pH change could be semi-quantitatively modeled with xDLVO-CG (as shown in Figure 3.12), which provides opportunities for further improvements. The calculated values of B_{22} suggest that IgG1 exhibits preferred attractive interactions at higher pH values. This behavior can be attributed to the higher protein charge observed at lower pH, which promotes the solubilization of IgG1 monomers in solution. The variation in the electrostatic potential between IgG1 protein pairs with respect to changes in pH is depicted in Figure 3.13. This figure illustrates how alterations in pH can affect the electrostatic interactions between IgG1 molecules, which can have significant implications for the stability and aggregation propensity of these proteins. The B_{22} values can serve as a valuable tool for understanding the behavior of IgG1 in different solution conditions, which can aid in the development of more stable formulations for this important class of monoclonal antibodies.

3.3.3. Analyzing Intermolecular Interactions in HBV Core Protein Dimers with xDLVO-CG Calculations

Using xDLVO-CG calculations, we analyzed the intermolecular interactions between HBV core protein dimers, which drive the assembly and stability of the dimer of dimers, and indirectly the formation of the whole capsid structures.

The formation of HBV VLPs involves various intermolecular interactions, with electrostatic forces often being the most significant. To gain a better understanding of the PPIs between core protein homodimers (Cp_2) of HBV with different sequence lengths, xDLVO-CG calculations were performed on six different Cp_2 with varying C-ter lengths. The lengths and charges of six different C_p proteins are shown in Table 3.3.

The shortest Cp_2149 dimer was represented by 10 beads in the CG representation, whereas the longest Cp_2183 dimer was represented by 12 beads. Figure 3.14 shows the calculated B_{22} coefficients obtained from the xDLVO-CG calculations performed on the Cp_2 dimers. As far as we know, there are no existing experimental measurements of the second osmotic

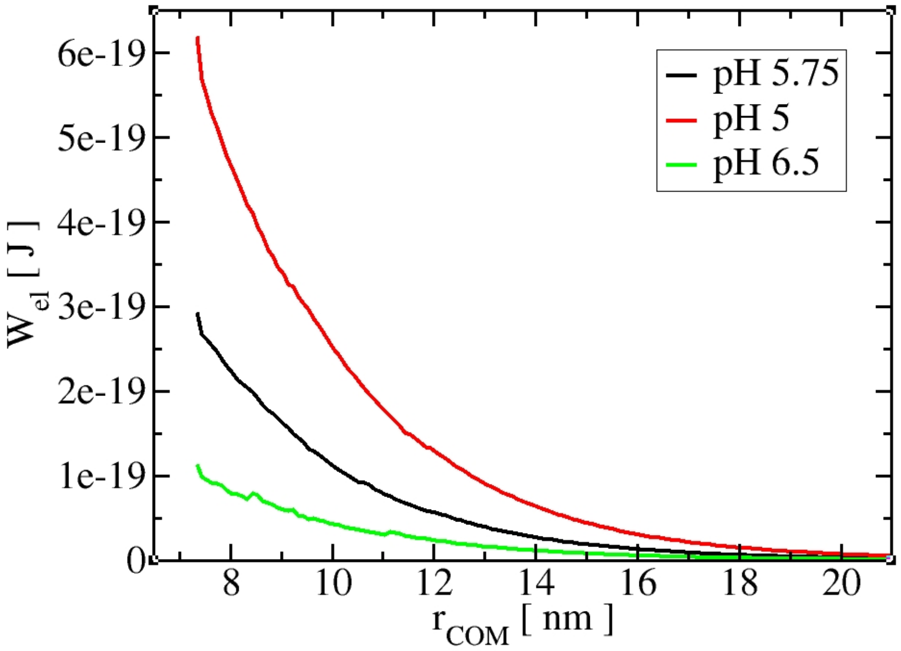


Figure 3.13.: Electrostatic potential between IgG1 proteins at pH 5, pH 5.75 and pH 6.5 and 10 mM NaCl calculated by xDLVO-CG model

Table 3.3.: The charges of core proteins per one Cp monomer, truncated at different positions of the C-ter

C_p, aa	Charges
149	-7
154	-3
157	-2
164	1
167	4
183	8

virial coefficients for Cp_2 core protein dimers that can be used to compare with the values calculated using the xDLVO-CG model. The B_{22} coefficients, calculated using xDLVO-CG model, for Cp_2164, Cp_2157, and Cp_2154 show a fast decrease with the increase of ionic strength due to their low charge states (see charges of Cp monomers in Table 3.3. This indicates that the PPIs between these proteins are more attractive in nature compared to other dimers. It is known that the assembly of Cp_2149, which lacks the flexible C-terminus chain, is highly influenced by ionic strength and only occurs in specific salt conditions. The calculated B_{22} values reveal a preference for attractive PPIs in solutions with NaCl concentrations higher than 0.15 M. Similar observations were made previously for capsid formation from the Cp149 assembly domain. The strongest repulsion interactions, requiring high ionic strength solutions, are observed for Cp_2183, which has a highly positively charged C-ter (+15 charged per single Cp). This allows for strong binding with oppositely charged species, such as nucleic acids, facilitating self-assembly and further capsid or VLP formation. Note that the charges shown in Table 3.3 are calculated for the entire core protein, including the contributions from the assembly domain and C-ter. To investigate the intermolecular interactions between Cp proteins with varying core protein lengths, Figure 3.15 displays the changes in electrostatic potential between the Cp proteins as a function of core protein length. The figure clearly illustrates the correlation between the calculated B_{22} coefficients

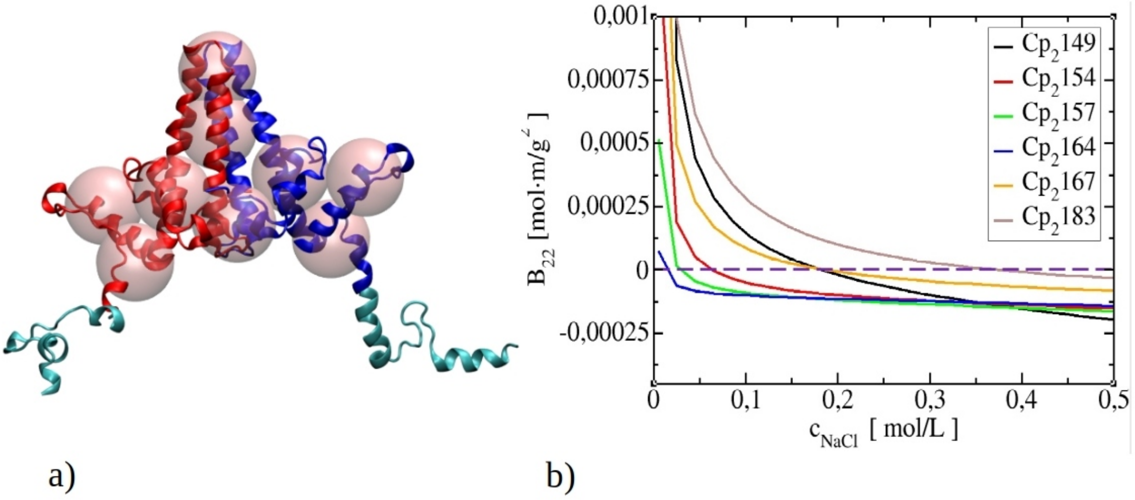


Figure 3.14.: a) Shape-based CG model of Cp$_2$149 dimer superimposed on its all-atom structure (red and blue chains). The flexible C-terminal (residues 150aa to 183aa) of the full length Cp$_2$183 is shown in cyan for illustration purposes. b) B_{22} coefficients calculated for full length HBV core protein dimers (Cp$_2$183) and truncated dimers (Cp$_2$149, Cp$_2$154, Cp$_2$157, Cp$_2$164, Cp$_2$167).

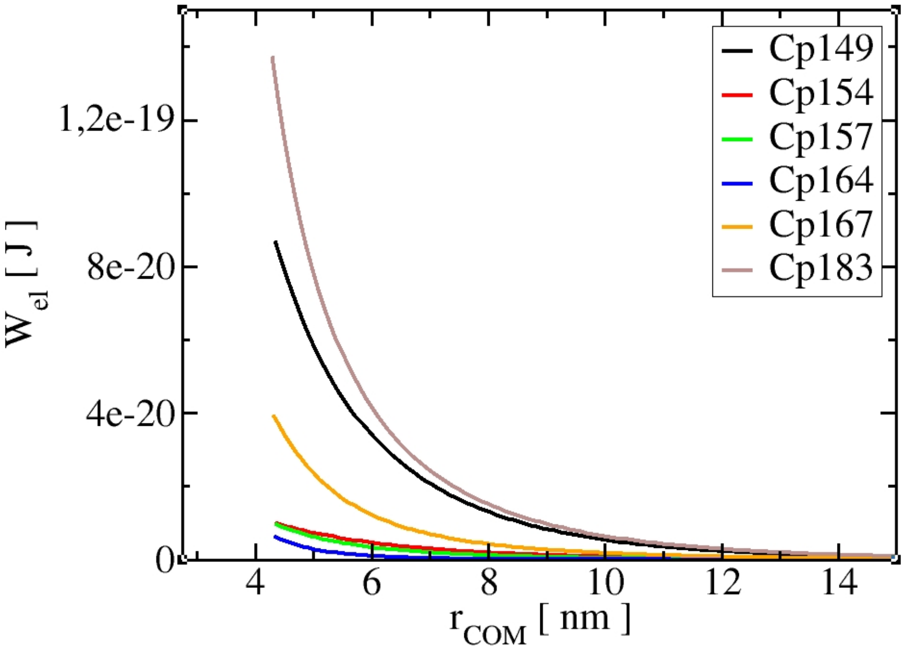


Figure 3.15.: Electrostatic potential between Cp proteins of different core protein length at pH 7 and 10 mM NaCl calculated by xDLVO-CG model

and the extent of electrostatic repulsion among Cp dimers. It shows that Cp dimers with higher repulsion exhibit more positive values of B_{22} coefficients, indicating a strong relationship between these two factors. This observation underscores the importance of considering electrostatic interactions when designing and engineering Cp proteins, as these interactions can significantly impact their stability and functionality. These findings provide insights into the molecular interactions underlying the formation of HBV capsids, which may have implications for the development of new technological processes for HBV VLP processing.

3.4. Conclusion

This Chapter presents our developed xDLVO-CG model that calculates osmotic second virial coefficients (B_{22}) for proteins based on pH, ionic strength, and protein type. The xDLVO-CG model predicts protein solution stability and salt-induced dependencies, and complements experimental measurements of B_{22}, providing insight into protein solubility as a function of colloidal PPIs. The model is derived from the extended DLVO theory and it includes a new term for ion-protein dispersion interactions derived and uses a shape-based coarse-grained representation to account for anisotropic PPIs. This approach reduces or eliminates the need for fitting experimental B_{22} data, which may accelerate investigations of protein processing conditions in pharmaceutical and food industries. The xDLVO-CG model has a unified CG protein mapping scheme, which improves transferability among proteins of different shapes and characteristics, and it does not require special model adjustments for different proteins. The model calculates B_{22} by integrating the PMF over many configurations, including electrostatic, dispersion, osmotic, and ion-protein dispersion interactions. The xDLVO-CG model agrees with experimental measurements of B_{22} coefficients for lysozyme, subtilisin, bovine serum albumin, and human immunoglobulin type 1 at different solution conditions. The model is transferable to larger, irregular, and non-spherical proteins, but due to difficulty to correctly compute the charge of proteins, it may result in deviations from experimental data. However, the model is still more accurate than regular xDLVO models, and with further improvements in electrostatic and solvent effects and better dispersion terms, it has the potential to better predict B_{22} coefficients without the need for time-consuming experiments. The developed computational scheme can also be adjusted for calculating aqueous stability of other colloidal particles, but requires a special CG parameterization scheme. In addition to its application to several benchmark proteins such as lysozyme, subtilisin, bovine serum albumin, and immunoglobulin type 1, the xDLVO-CG model has also been used to compute the osmotic second virial coefficients for hepatitis B virus core protein dimers. This specific application of the xDLVO-CG model to the hepatitis B virus core protein dimers showcases its potential to be applied to a wide range of proteins and provides further validation of its accuracy and reliability in predicting protein stability under different environmental conditions. The results of these calculations can also contribute to a better understanding of the behavior of the hepatitis B virus core protein dimers in solution, which is of significant importance in the development of processing conditions of HBV VLPs.

4. xDLVO-CGhybr Model: Hybrid Method for Protein-Protein Interactions with Poisson Boltzmann and Extended DLVO Theory

4.1. Introduction

[1] Chapter 3 introduced the xDLVO-CG model, a computational approach aimed at assessing protein protein interactions by calculating B_{22} coefficients. This model combines the extended xDLVO theory, which describes the stability of colloidal suspensions, with theCG approach, which simplifies the representation of macromolecules. Chapter 3 provides an extensive elaboration on the significance of PPIs in aqueous solutions for both fundamental science and technological applications. These interactions play a crucial role in determining the stability and solubility of proteins in solution, as well as whether they undergo aggregation or crystallization.[11, 127] The state of proteins in solution is primarily governed by weak, nonspecific interactions that can be influenced by various factors, such as pH, ionic strength, and solvent composition. Understanding these interactions and their dependencies is essential for designing and developing effective strategies to control protein stability and prevent undesirable aggregation or precipitation, which can significantly impact the performance and shelf-life of protein-based products.[189, 7, 190, 123] In this context, B_{22} coefficients serve as a valuable tool for assessing the net effective interactions of proteins in solution. The xDLVO-CG model has demonstrated the ability to compute B_{22} coefficients with reasonable agreement with experimental values, albeit with some discrepancies observed for large, irregularly-shaped proteins. Therefore, there is still room for improvement in the xDLVO-CG model to better model the potential of mean force and achieve a closer agreement with experimental values, particularly for large and irregularly-shaped proteins. The central focus of improving the xDLVO-CG model was to modify the electrostatic potential term to enhance its accuracy in modeling PPIs. Electrostatic interactions play a crucial role in PPIs, as the protonation states of amino acids are influenced by the pH of the solution and the protein environment.[191] Proteins exhibit a wide range of forms and charge anisotropy, which can significantly impact various physicochemical processes. Changing the salt concentration is a common method used to modulate PPIs in solution, as it weakens repulsive electrostatic interactions. However, previous theoretical studies on B_{22} modeling have mostly relied on simplified continuum models based on Debye-Hückel theory, which is an approximation that has not been adequately tested to represent biomolecular electrostatics in context of calculating second osmotic virial coefficients. Some attempts have been made to go beyond Debye-Hückel theory, such

[1] The majority of the content presented in this chapter is based on the work published in Ref. [188], which has been modified and restructured to align with the format of thesis. Author contributions are stated in the perspective journal article.

as creating a modified Poisson-Boltzmann method that incorporates ion-specific effects on spherically represented proteins,[164, 192, 193] using a fast multipole method solved by boundary element method on a residue level coarse grained structure,[194] and proposing the extended Debye-Hückel continuum model for better modeling solvation dynamics.[195] Due to computational constraints, some researchers have used the PB theory to solve electrostatic potential between atomistically represented proteins with the goal of computing B_{22}, but only on a limited number of relative protein orientations.[69] Despite the importance of accurately modeling electrostatic effects in biomolecular systems, insufficient effort has been made in the context of B_{22} computations. While explicit solvent models offer accurate modeling of biomolecular electrostatic effects, they are computationally expensive. Furthermore, using all-atom MD to calculate B_{22} using free energy techniques yields values that are far from experimental values, and LJ interactions need to be weakened by factors of approximately 0.1 to achieve a good match with experiments. The xDLVO-CG model provided reasonably accurate calculations for B_{22} coefficients, but showed some discrepancies for proteins that are large and irregularly-shaped. Although this model is computationally efficient, it may not capture all the molecular-level details that higher resolution models can account for. To address this issue, we introduce the xDLVO-CGhybr model, which improves accuracy by modifying the electrostatic potential term of PMF. This model employs a hybrid approach that combines PB theory and Debye-Hückel theory to accurately calculate the electrostatic contribution to the total interaction potential, regardless of protein size and shape. Moreover, we introduce a CG Lennard-Jones potential that matches reference all-atom potentials for precise prediction of PPIs. We validate our model on six different proteins with varying complexity and shape, and show improved predictions of B_{22} values.

4.2. Methods

4.2.1. xDLVO-CGhybr Model: Interaction Potential Calculation

In the xDLVO-CGhybr model, the interaction potential denoted as $W(r)$ is computed by summing up the electrostatic, dispersion, osmotic, and ion-protein potential terms between pairs of proteins. The resulting equation for the interaction potential is:

$$W_{22}(r) = \begin{cases} W_{el}(r) + W_{disp}(r) + W_{osm}(r) + W_{i-pr}^{disp}(r) \\ W_{el}(r) + W_{LJ}(r) + W_{osm}(r) + W_{i-pr}^{disp}(r) \end{cases} \tag{4.1}$$

The xDLVO-CGhybr model employs the same four interaction potential terms as the previously reported xDLVO-CG model. However, the electrostatic and dispersion potentials, which are the first two terms in the potential function, have undergone modifications in xDLVO-CGhybr. The nature of these modifications will be discussed in detail below. The remaining two potential terms, namely osmotic depletion and ion-protein potential, have remained unaltered and will be briefly elaborated upon. For a more comprehensive explanation of these terms, please refer to Chapter 3. In the xDLVO-CGhybr model, potential terms are computed between proteins that are represented using a CG resolution. However, for electrostatic interactions that occur at short protein separations, full all-atom structures are used. This is in contrast to the xDLVO-CG model, which only utilized CG structures.

4.2.2. Electrostatic Interactions

The xDLVO-CGhybr model utilizes a hybrid approach to determine the electrostatic interaction energy between two proteins. The model applies the Poisson-Boltzmann equation to all-atom protein structures to compute the electrostatic interaction energy for short protein separations up to $R_0 + 2nm$, where R_0 represents the initial center-of-mass distance. Brief overview of PB theory[24, 23, 83, 82] is given in Section 2.3 in Chapter 2. For larger separations, the Debye-Hückel model is used,[77] which is calculated using a CG representation of the protein structures:

$$W_{el}(r) = \begin{cases} E_{PB}(r), & r \leq R_o + 2nm \\ E_{DB}(r), & r > R_0 + 2nm \end{cases} \tag{4.2}$$

This method enables the efficient calculation of interaction potential while retaining essential molecular-level details. When proteins are located in close proximity to each other, the electrostatic interaction energy is obtained by calculating the difference between the total electrostatic free energy of the protein complex and the electrostatic energies of the individual, separated proteins:

$$E_{PB}(r) = G_{complex}(r) - G_{Prot1}(r) - G_{Prot2}(r) \tag{4.3}$$

The term G represents the electrostatic free energy obtained through the use of an iterative solver. For more information on the computational details of these calculations, please refer to Section 4.2.7.

When the distance between proteins exceeds $R_0 + 2nm$, a computationally efficient approach for calculating electrostatic interactions can be achieved through the use of the Debye-Hückel model applied to CG protein structures. The relevant equations are as follows:

$$W_{el}(r) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{Z_i Z_j e^2 \exp\left(\kappa(R_{bi} + R_{bj} - r_{ij})\right)}{4\pi\varepsilon_0\varepsilon_r r(1 + \frac{\kappa(R_{bi}+R_{bj})}{4})^2}, r_{ij} > d_{ij} + 2\sigma \tag{4.4}$$

R_{bi} and R_{bj} are bead radii defined differently in this work and assumed to be equal to the radius of gyration of the atoms that comprise each bead. r is the COM distance between two proteins, N_1 and N_2 are the total number of beads, d_{ij} is the starting distance between bead pairs, r_{ij} is the current bead bead distance during protein translation, σ is the length of the water layer around a protein (0.1 nm), $\varepsilon_r r$ is the relative permittivity, Z_i and Z_j are bead charges, and κ is the inverse Debye length is calculated as follows:

$$\kappa = \sqrt{\frac{2N_A e^2 I}{\varepsilon_0 \varepsilon_r k_B T}} \tag{4.5}$$

where I stands for ionic strength, N_A stands for the Avogadro number, k_B stnds for the Boltzmann constant, and T stands for absolute temperature.

4.2.2.1. Integrating Poisson-Boltzmann and Debye-Hückel Approaches

A hybrid calculation scheme based on PB and Debye-Hückel theory is used to improve the electrostatic part of the potential of mean force. PB calculations are performed on all-atom structures at short COM distances, while Debye-Hückel calculations are performed on CG structures at larger COM distances. The electrostatic binding energy calculated by PB and Debye-Hückel theory for IgG1 and BPTI is shown in Figure 4.1 as a function of COM distances. When the energies obtained by these two methods were compared, it was observed that the largest differences were observed at the initial COM distances, while the energies were relatively similar at larger protein separations. At short COM distances, the energies obtained by the Debye-Hückel model were smaller by a factor of three to five than the values obtained by PB theory. This conclusion applies to all proteins studied where a smooth transition from PB to Debye-Hückel potential was observed. The Debye-Hückel equation represents the analytical solution to the PB equation for the interaction of two homogeneously charged spheres of equal radius and is used as an approximation in other cases. Various approximate solutions of PB equations have been investigated on test molecules or geometric shapes, and their findings show that these models are more accurate at larger intermolecular distances. The larger deviation between these two models at short separations is expected because specific residue-residue interactions can be better described by all-atom protein representation and PB theory. At the protein-protein interface, the effective dielectric constant can shift from the solvent to the protein interior. As a result, these residues effectively interact as if they belong to the same protein within its low dielectric environment, resulting in a higher charge-charge interaction than if they were placed in a solvent medium. These residues become solvated and begin to feel the dielectric environment of the solvent as protein-protein separation increases.[196] The effects of dielectric discontinuity become significant only at separations less than Debye length. The dielectric constant is determined by rolling the solvent probe radius, according to computational methods for numerically solving PB equations. Furthermore, proteins are large molecules whose electrostatic interactions are heavily influenced by their shape and charge distribution. The energy of polarisation due to dielectric interface (second term in Equation 2.17) is highly dependent on protein shape, with partial atomic charges placed near the protein surface contributing the most. These and many other factors can all significantly contribute to the total interaction energy. A hybrid approach was chosen as a compromise between computational speed and accuracy because solving PB equations for protein systems, particularly when calculating second osmotic virial coefficients, is computationally expensive, where electrostatic interactions must be evaluated for a wide range of starting orientations, COM distances, and salt concentrations. It is demonstrated that this method can still provide reasonably accurate modeling of electrostatic effects while also improving predictions of second osmotic coefficients (see Section 4.3). Therefore, it can be concluded that electrostatic interactions are important in determining overall protein interactions, and accurate modelling of electrostatic interactions is critical in determining overall PPIs in solution.

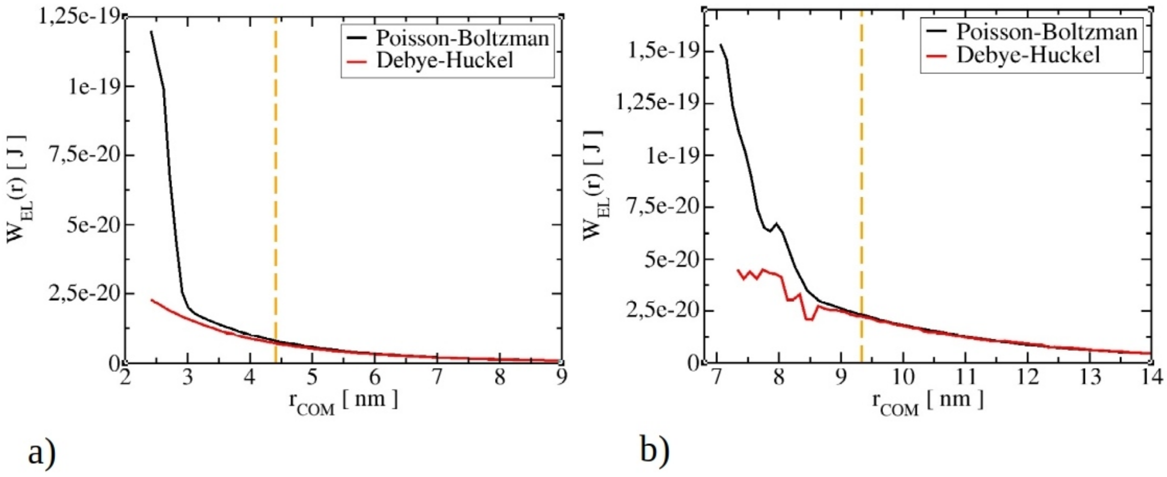


Figure 4.1.: Comparison of calculated electrostatic energy of interactions between a) BPTI and b) IgG1 proteins by solving Poisson-Boltzmann structure on full all-atom structure and by using Debye-Huckel on coarse-grained structures. The vertical dashed orange line indicates COM distance (at Ro+2nm) where electrostatic potential in xDLVO-CGhybr is switched from Poisson Boltzman to Debye-Hückel model.

4.2.3. Dispersion Interactions

In the xDLVO-CGhybr model, the dispersion interactions between proteins are calculated through either the Hamaker potential or the Lennard-Jones potential. The Hamaker potential, which arises from electromagnetic quantum fluctuations,[197, 198, 159] describes the attraction forces between molecules. It is calculated by integrating the London dispersion forces between two homogeneous spheres[76] and is represented by the following equation:

$$W_{Ham}(r) = -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \frac{A_H}{12}\frac{1}{N_1 N_2}\left[\frac{d_{ij}}{r_{ij}^2 - d_{ij}^2} + \frac{d_{ij}^2}{r_{ij}^2} + 2\ln(1 - \frac{d_{ij}^2}{r_{ij}^2})\right], r_{ij} > d_{ij} + 2\sigma \quad (4.6)$$

Furthermore, the dispersion interactions between proteins were also calculated based on the Lennard-Jones potential, represented by following equation:

$$W_{LJ}(r) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], r_{ij} > d_{ij} + 2\sigma \quad (4.7)$$

Here, ε_{ij} and σ_{ij} represent the respective Lennard-Jones parameters for each bead pair. The equations used in this study are identical to those in our previously reported xDLVO-CG model. However, unlike in our previous work where LJ parameters were derived using a simplified method implemented in the coarse-grain builder in VMD, in this study, the parameters were directly parametrized from all-atom LJ potentials (as described in Section 4.2.8.

4.2.3.1. Modeling Dispersion Interactions: Comparing Hamaker and Lennard-Jones Potentials

The interaction between two molecules, as proposed by the Lifshitz theory of van der Waals forces,[197] is rooted in the dipole field created by quantum fluctuations. This results in mutual polarisation between molecules and with the solvent, giving rise to net attractive dispersion interactions. This results in mutual polarization between molecules and the solvent, creating net attractive dispersion interactions. The Lifshitz theory provides a rigorous theoretical framework for calculating Hamaker constants; however, this approach is practically challenging as it necessitates knowledge of full refractive indices spectra.[198, 154] Consequently, Hamaker constants fitted from experiments are more commonly used. In addition, other approaches can be also used, such as scaling the LJ potential, which is known to result in overly attractive interactions.[146]

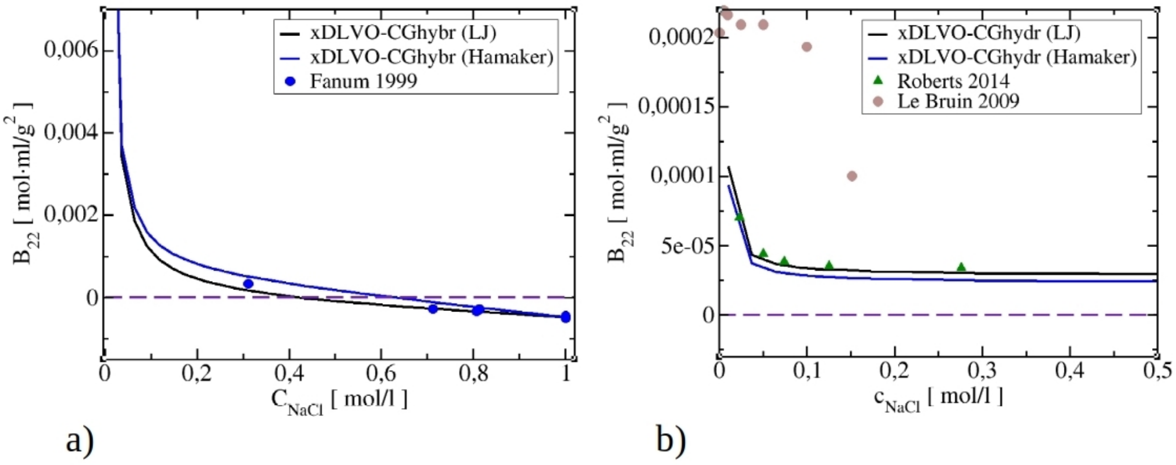


Figure 4.2.: Impact on using Lennard-Jones or Hamaker potential on calculated B_{22} coefficients for a) BPTI at pH 4.9 b) IgG1 at pH 6.5. The experimental data labeled as 'Farnum 1999', 'Roberts 2014' and 'Le Bruin' were taken from [199, 187, 200].

Our approach uses LJ potentials scaled to match the depth of interaction of Hamaker potential, with A_H obtained from literature. While both approaches give similar B_{22} values (as shown on Figure 4.2), Hamaker potential is more convenient to use while Lennard-Jones potential can better model anisotropy of dispersion interactions. As it will be shown in Section 4.3, accurately evaluating electrostatic interactions is crucial for improving the overall predictive power of the model, while the choice between Hamaker or LJ potentials has less impact.

4.2.4. Osmotic Depletion Potential

The osmotic potential is a type of attractive interaction between two proteins that occurs due to the exclusion of ions from the interstitial space between them over short distances. This exclusion causes a local osmotic pressure imbalance, leading to the additional attractive interaction between proteins.[79] The osmotic attraction potential $W_{osm}(r)$ is determined using the following equation:

$$\begin{aligned} W_{osm}(r) = -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{1}{N_1N_2}\frac{4\pi k_B}{3}TD_{ij}^3\rho_3(1-\frac{3r_{ij}}{4D_{ij}}+\frac{r_{ij}^3}{16D_{ij}^3}), \\ d_{ij}+2\sigma \leq r_{ij} \leq 2D_{ij} \end{aligned} \tag{4.8}$$

where R_3 is the mean hydrated radius of the salt (the sum of anion and cation radii), ρ_3 is the salt density, and D_{ij} is defined as $D_{ij} = d_{ij} + R_3 + \sigma$.

4.2.5. Ion-protein Dispersion Interactions

The protein-ion potential describes the total dispersion interaction between the protein and all ions in its vicinity. The protein is represented as an ideal sphere with the charge Z, and ions are distributed non-uniformly around it using the Boltzmann distribution. The total potential is calculated by integrating the contribution of each ion according to:[117]

$$\begin{aligned} W_{i-pr}^{disp}(r) = -4\pi B_a\int_{R_P+\sigma}^{R_P+D}\frac{c_b\exp\left(-z_a\phi(r)/(k_BT)\right)}{r}dr \\ -4\pi B_c\int_{R_P+\sigma}^{R_P+d}\frac{c_b\exp\left(-z_c\phi(r)/(k_BT)\right)}{r}dr \end{aligned} \tag{4.9}$$

In this equation, B_a and B_c are ion-macroion dispersion coefficients that describe the interaction between the protein and anion and cation ions, respectively, as found in the literature. R_p refers to the protein radius, and D represents the thickness of the spherical shell around the protein, which is taken as 1 nm. The charges of the cations and anions are denoted by z_c and z_a, respectively, while c_b represents the bulk salt concentration. Finally, $\phi(r)$ represents the electrostatic potential around a protein sphere that carries a charge Z.

4.2.6. Mapping Proteins into Coarse-Grained Structures and Computational Details of xDLVO-CGhybr Calculations

The all atom structures of proteins were taken from the protein data bank with the codes 1bpi, 3rn3, 2cga, 3nwk, 4f5s and 1mco for bovine trypsin inhibitor (BPTI), ribonuclease A (RbnA), chymotrypsinogen (ChymA), concanavalin A (ConcA), bovine serum albumin (BSA) and human immunoglobulin type I (IgG1) respectively. The chosen PDB structures were checked to see if they contained the missing residues, and if so, the Swiss Model program was used to model them. The PROPKA method (version 3.3)[170, 171] and the PDB2PQR online web server[172, 173] were used to assign the protonation states of protein residues at the desired pH. The protonated all-atom structures were then mapped into the CG representation (with about 500 atoms per bead) using a SBCG model[160] implemented in the VMD program (version 1.9.3).[174] The centre of each bead was placed in the corresponding atoms' COM, and the bead radius was assigned to the radius of gyration and charge as a sum of partial charges. The PMF was calculated by adding interactions between the corresponding bead pairs from each protein pair using Equations 4.1 to 4.9. The PMF and B_{22} were calculated using in-house code, and B_{22} was determined by numerical integration of the PMF over various protein-protein orientations using Equation 2.3. The procedure described in Section 3.2.7 was used to sample protein-protein orientations, except that due to higher computational costs, PMF was determined over fewer protein-protein configurations, i.e. by starting from 83 starting radial positions. For each possible starting configuration, the PMF was determined by translating proteins over a vector connecting the COMs of two protein pairs up to a distance of $R_0 + 2nm$, where R_0 is the initial distance between two proteins.

4.2.7. Poisson-Boltzmann Calculations: Computational Details

PB calculations on all-atom protein structures were performed using an APBS (Adaptive Poisson-Boltzmann Solver).[24, 201] First, all-atom protein structures from the PDB were protonated using the Propka method (as described in Section 4.2.6). The CHARMM force field was used to assign partial charges and Van der Waals radii to the atoms. The linearized finite difference PB equation was solved using APBS. Using this method, the equation is first solved on a coarse grid (with fewer grid points) with large dimensions (grid lengths). The solution is then used to calculate the Dirichlet boundary conditions for a smaller region of interest with a finer grid. Three calculations are required to determine the electrostatic energy of interaction between two proteins: one of the protein-protein complex and one of each of the two proteins separated from each other. The electrostatic interaction energy is calculated as the difference between the complex's electrostatic energy and the electrostatic energies of the separated proteins (See Equation 4.3). The APBS calculations were performed at 20 different monovalent salt concentrations ranging from 10 mM to 1M NaCl. The radii of sodium and chloride were set to 2.0 and 2.23, respectively. One of the proteins was kept fixed in space for each calculation at a specific salt concentration, while another protein was translated along the vector connecting their centers of mass, increasing the distance by one Å in each step. The second protein was moved up to $R_0 + 2nm$ from its starting COM distance R_0, and APBS calculations were performed at each intermediate distance. This allowed to calculate the electrostatic binding energy of protein complexes at various ionic strengths. In each APBS calculation, the number of grid points, the length of the coarse grid and the fine mesh domain lengths were set by the internal APBS script, depending on the size of each studied protein pair. Multiple Debye-Hückel boundary conditions were used, and the molecular surface definition was "smoothed" using 9-point harmonic averaging,[85] with the solvent probe radius set to 1.4 Å and the solvent density set to 10 quadrature points per Å^2. The cubic B-spline discretization was used to map protein charges to the grid. The internal dielectric constant of all proteins studied was set to 4.0, while the external dielectric constant was set to 78.4 (the dielectric constant of water medium). To obtain the electrostatic binding energy of the protein complex for a specific COM distance or ionic strength, six lpbe calculations were required for each separate calculation (two of complex, first protein, or second protein respectively). To ensure proper cancellation of self-solvation energies, all calculations were performed with the same grid spacing.

4.2.8. Lennard Jones Parameters for Coarse-Grained Model

Each bead was assigned its R_{min} parameter ($R_{min} = 2^{1/6}\sigma$), which was chosen to be equal to the radius of gyration of the atoms that make up each bead, and epsilon was fitted to match the all-atom LJ potential. The all-atom LJ potential was computed by translating proteins over vector connections COMs of protein pairs, beginning with five distinct relative orientations. The potentials were calculated using GROMACS[202] tools (gmx energy) and CHARMM36m potential parameters,[203] after assigning protonation states according to Propka and constructing a translation trajectory using GROMACS tools. It was assumed that the all-atom LJ potential was the same for all pH, as a small number of different hydrogens did not have a significant impact on the final potential. Next, the CG Lennard Jones potential was adjusted to match the all-atom potential, with the epsilon parameters variable and the sigma parameters fixed. A least squares algorithm was used for the fitting process. The interaction parameters between different beads were determined using Lorentz Berthelot combining rules. Finally, the depth of the CG LJ potential was scaled to match the depth of the Hamaker dispersion potential, as LJ parameters are parametrized for vacuum and their effective interaction in a solvent medium is smaller in magnitude

4.3. Results and discussions

To validate the model, we conducted calculations on six different proteins with varying sizes and shapes: BPTI, RbnA, ChymA, ConcA, BSA andIgG1, shown on 4.3. The structures and shapes of these proteins are depicted in Figure 4.3, along with their corresponding electrostatic maps. We utilized the Hamaker constant as the only adjustable parameter in our model, which was obtained from the literature or assigned a general value of $5k_bT$ based on the fundamental Lifshitz theory of electrodynamic forces. We compared the results of our calculations to the B_{22} values obtained from previously published experimental data. To facilitate comparison, we also computed B_{22} values using an xDLVO-CG model,[117] an FMAPB2 model,[147] and a standard xDLVO model.[204, ?] The FMAPB2 model uses an all-atom protein representation in combination with an implicit solvent model and is publicly available on a web server (https://pipe.rcc.fsu.edu/fmapb2/). The xDLVO model represents proteins as ideal charged spheres and uses the charge obtained from Propka and the experimentally determined hydrodynamic radius from literature.

4.3.1. Calculation of B_{22} Coefficients for Small and Medium-Sized Proteins

4.3.1.1. Bovine Pancreatic Trypsin Inhibitor

BPTI is a small protein with an ellipsoid shape, consisting of 58 residues with a molecular mass of 6.5 kDa. In physiological conditions, it binds with high affinity to trypsin and other digestive proteases, inhibiting digestive enzymatic activity. Naturally occurring in various plants, such as soybeans, legumes, and grains, BPTI plays a role in a self-defense mechanism. Figure 4.4 shows the calculated B_{22} coefficients for BPTI at pH 4.9, in comparison with xDLVO-CG,[117] FMAPB2,[147] and xDLVO models.[?] The calculated values at low and medium salt concentrations are positive and decrease towards negative with an increase in ionic strength. Theoretical B_{22} data points are crossing zero at approximately 0.42 M, following a similar trend of decrease as the experimental data.[199] At pH 4.9, the protein has a relatively high charge for its small size (+6 according to the Propka method). Moreover, Figure B.2 shows that all but one protein bead are positively charged without a significant protein region bearing negative charge, with only one bead bearing charge -0.06. This uniform local charge distribution contributes to high electrostatic repulsion, requiring an intermediate salt concentration to screen electrostatic interactions and shift protein-protein interactions from repulsive to attractive ones. Calculated values of B_{22} coefficients show nearly quantitative agreement with experimental results of Farnum et al.,[199] who performed static light experiments to measure the experimental B_{22} values. To the best of our knowledge, no other experimental B_{22} measurements were performed on BPTI protein. In comparison, the xDLVO-CG model gives B_{22} values that are shifted negatively compared to the experimental ones and it crosses zero at a lower ionic strength of 0.36 M. B_{22} values calculated with FMAPB2 correlate with xDLVO-CGhybrid results

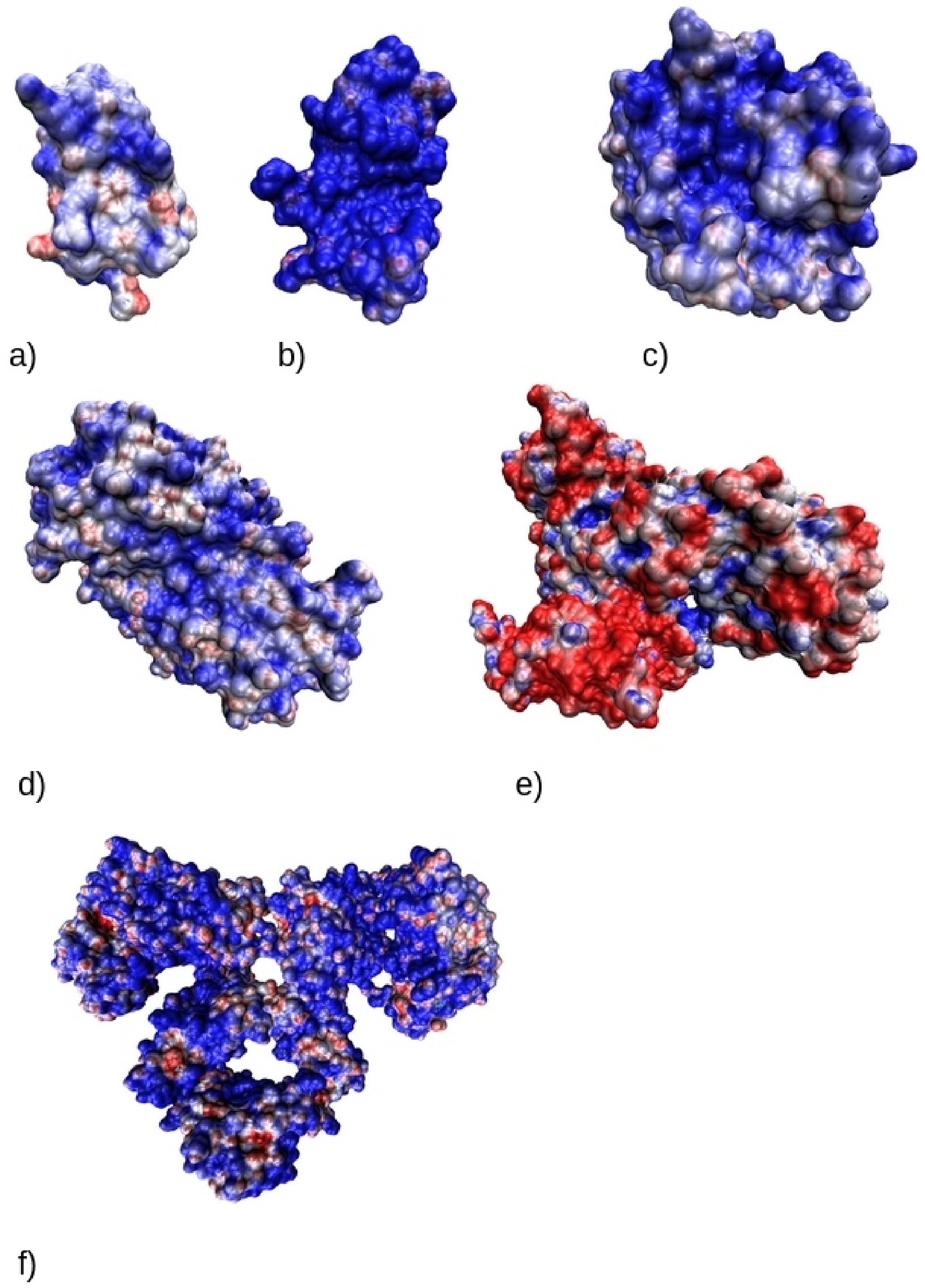


Figure 4.3.: All-atom structures of six proteins are presented alongside their corresponding electrostatic maps. Panel (a) depicts the Bovine trYpsin inhibitor (BPTI) at pH 4.9 with a charge of +6. Panel (b) shows Ribonuclease A (RbnA) at pH 3 with a charge of +16. Panel (c) displays Chimotripsinogen (ChymA) at pH 3 with a charge of +17. Panel (d) illustrates Concanavalin A (ConcA) at pH 4 with a charge of +25. Panel (e) demonstrates Bovine serum albumin (BSA) at pH 7.4 with a charge of -16. Finally, panel (f) presents Human immunoglobulin type 1 (IgG1) at pH 6.5 with a charge of +27. The protein surface is colored according to the electrostatic potential calculated by APBS, with blue marking regions of excess positive charge and red marking regions of excess negative charge. The sizes of the proteins in the illustrations are not to scale, but are depicted for viewer convenience.

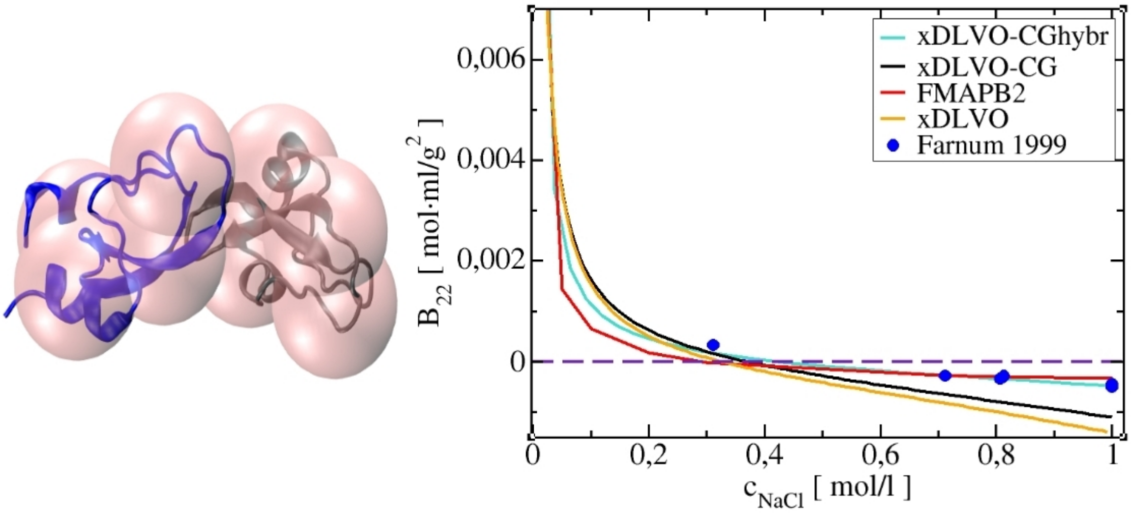


Figure 4.4.: Calculated B_{22} coefficients for BPTI at pH 4.9 as a function of NaCl concentration. The B_{22} values are compared with the values obtained with xDLVO-CG (black), FMAPB2 (red) and xDLVO (orange) models and experimental results (circles). On the left side are shown coarse-grained models of proteins superimposed on an all-atom representation. The experimental data labeled as 'Farnum 1999' were taken from [199].

and match the experimental data at all ionic strengths except at the lowest ionic strength of 0.3 M. At this ionic strength, FMAPB2 model deviates, giving B_{22} values of -0.0000205 $mol \cdot ml/g^2$, while the experimentally determined value is 0.00033199 $mol \cdot ml/g^2$ (and xDLVO-CGhybr gives a value of 0.000331996 $mol \cdot ml/g^2$). Moreover, there are slight differences between calculations made with FMAPB2 and xDLVO-CGhybr at lower ionic strengths below 0.4 M NaCl. FMAPB2 data crosses the 0 point at approximately 0.27 M and gives lower B_{22} values in comparison to xDLVO-CGhybr and xDLVO-CG models. Given the lack of available experimental measurements, it is challenging to judge which model performs better at low salt conditions. The xDLVO model overestimates the B_{22} values and is out of range of experimental values except at the first point. Brownian simulations have also been employed by Mereghetti et al to calculate B_{22} in BPTI solutions.[205]

4.3.1.2. Ribonuclease A

Bovine pancreatic RbnA is a digestive enzyme found in the pancreas that breaks down single stranded RNAs in food. This small protein consists of 124 amino acid residues and has a triangular shape with a molecular mass of 13.7 kDa. Due to its small size and ease of purification, it is frequently used in biochemical research. In fact, several Nobel prizes have been awarded for discoveries in molecular biology that were made with the help of this protein. Figure 4.5 displays the calculated B_{22} coefficients for RbnA at different pH values (pH 3 and 4) and ionic strengths ranging from 50 mM to 1 M NaCl. Experimentally determined B_{22} values at pH 3 show positive values, indicating repulsive protein-protein interactions until a concentration of approximately 0.95 M NaCl.[206] The second osmotic virial coefficients calculated using the xDLVO-CGhybr model follow a similar trend of decrease and predict the repulsive nature of PPIs. The calculated data show good agreement with experimental data at 0.1 M and 1 M NaCl. However, reproducing the non-monotonous experimental data at pH 3 using the CG protein representation is challenging, without considering dynamical changes on an all-atom level. As the ionic strength increases, electrostatic screening diminishes the repulsion between proteins, resulting in negative B_{22} coefficients. At the highest salt concentration (1 M), the experimentally determined B_{22} value is $-2.7110^{-6}mol \cdot ml/g^2$, while the value calculated using the xDLVO-CGhybr model is $-2.72110^{-6} mol \cdot ml/g^2$. Although the calculated B_{22} value is near zero, it is insufficient to indicate the presence of strong attractive PPIs. The B_{22} values calculated at pH 4 are repulsive at almost all salt ranges and agree nearly quantitatively with experimental data, except at the first two points at lower ionic strength. Increasing the ionic strength up to 1 M is not enough to shift PPIs towards attraction, unlike what is observed for many other proteins. The repulsive interactions are caused by electrostatic repulsion, which occurs because the RbnA has a high charge for its size. Figure B.2 indicates that positive charges are evenly distributed throughout the protein, without significant charge anisotropy that could reduce electrostatic repulsion. When comparing the B_{22} values calculated using the xDLVO-CG model, slightly larger values are obtained at low salt concentrations (below approximately 0.3 M). After that point, the values shift towards negative values at higher salt concentrations, leading to larger deviations from experimental data. Similarly, the FMAPB2 model follows a similar trend as the xDLVO-CG calculations until around 0.2 M salt concentration. However, at higher ionic strengths, the FMAPB2 values shift even further towards negative values and deviate further from experimental data. The xDLVO model gives data that falls outside the range of experimental values. It is shifted positively throughout the entire range of salt concentrations, except at the highest ionic strength where it provides a good match. To gain a deeper understanding of the observed trends in changes of the second osmotic virial coefficients, it is essential to examine the potential of mean force in greater detail. Figure 4.6 provides insight into how the PMF changes with the addition of salt. At low salt concentrations, the PMF exhibits strong repulsion between the proteins, primarily due to electrostatic interactions. As the salt concentration increases, the ionic strength of the solution rises, leading to a screening effect that reduces the strength of the repulsive interactions. With further increases in salt concentration, the dispersion interactions between the proteins become more pronounced, resulting in a shift in the dominant interactions from electrostatic to dispersion forces. The changes in the PMF with increasing NaCl concentration reflect a complex interplay between various intermolecular forces that govern the behavior of proteins in solution. Figure 4.6b compares the PMF at a salt concentration of 1 M NaCl obtained using xDLVO-CGhybr and xDLVO-CG models. The results reveal that the PB theory predicts a repulsive peak even at 1 M NaCl,

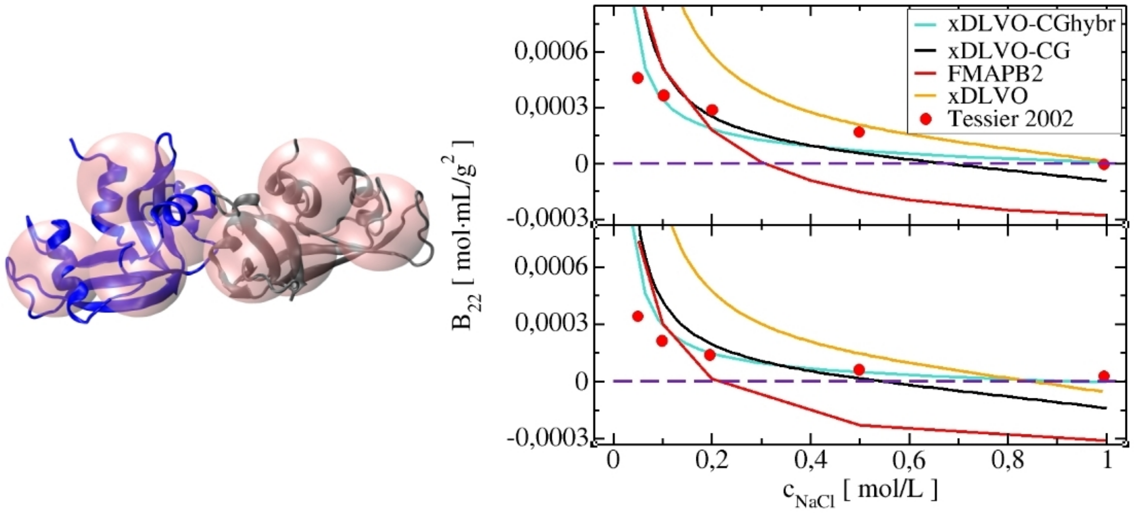


Figure 4.5.: Calculated B_{22} coefficients for RbnA at pH 3 and pH 4 (upper and lower panel) as a function of NaCl concentration. B_{22} values are compared with the values obtained with xDLVO-CG (black), FMAPB2 (red) and xDLVO (orange) models and experimental results (circles). The left side displays coarse-grained models of proteins superimposed on an all-atom representation. The experimental data labeled as 'Tessier 2002' were taken from [206].

but this peak only exists at the first five distances. However, this repulsive peak is not significant enough to cause a major disturbance in the B_{22} value due to the structure of the B_{22} formula (Equation 2.3). By carefully examining the PMF, we can gain a better understanding of the underlying physics of the system and the factors that contribute to changes in the B_{22} values.

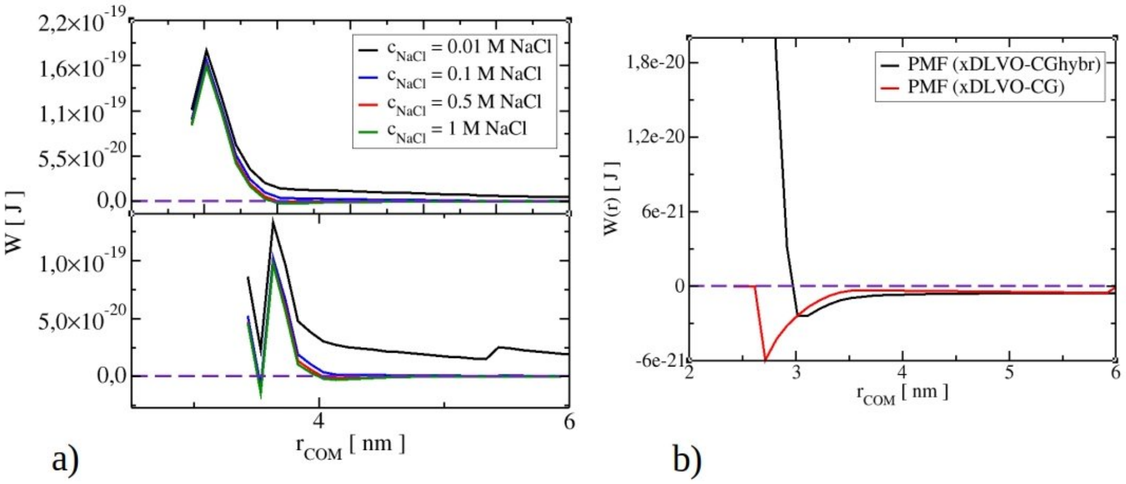


Figure 4.6.: Changes of potential of mean force of BPTI at pH 4.9 (upper panel) and RbnA at pH 3 (lower panel) with increasing NaCl concentration. Strong repulsion between proteins, originated from electrostatic interactions, decreases with an increase of ionic strength. When charge screening is achieved, dispersion interactions become more pronounced. b) Comparison of PMF obtained by xDLVO-CGhybr and xDLVO-CG model

4.3.1.3. Chymotripsynogen

ChymA is a medium-sized globular protein made up of 245 residues. It is a biologically inactive precursor of chymotrypsin, an enzyme that selectively hydrolyzes peptide bonds formed by aromatic residues such as tyrosine, phenylalanine, and tryptophan. Figure 4.7 illustrates the calculated B_{22} values for Chymotrypsinogen A at pH 3 and up to 1 M NaCl salt concentration. Our calculated values almost quantitatively match the experimental data of Tessier and Velev,[178, 207] and semi-quantitatively match other experimental data (of Bajaj and Pjura).[208, 209] In comparison, the data calculated using the xDLVO-CG model are slightly negatively shifted compared to the data obtained by our new model, but the difference slowly fades as the ionic strength increases until it reaches practically the same value at 1 M NaCl. The FMAPB2 calculations show data that are further shifted towards negative values, especially at higher ionic strengths. In contrast, the values calculated by the xDLVO model are slightly positively shifted at low salt concentrations and towards negative at high salt concentrations. Unfortunately, no experimental data were reported for concentrations higher than 0.4 M NaCl, which makes it difficult to compare the performance of these models at higher ionic strengts. Lund et al used Monte Carlo simulations to study the effect of salt concentrations on B_{22} change in ChymA.[210] They utilized a simplified model at the residue level and considered interaction potential based on hard sphere, electrostatics, and Van der Waals contributions. The simulation results were in line with experimental observations, however, the interactions between proteins were found to be slightly more repulsive. Monte Carlo simulations have also been employed by other researchers to calculate B_{22} in ChymA solutions.[69]

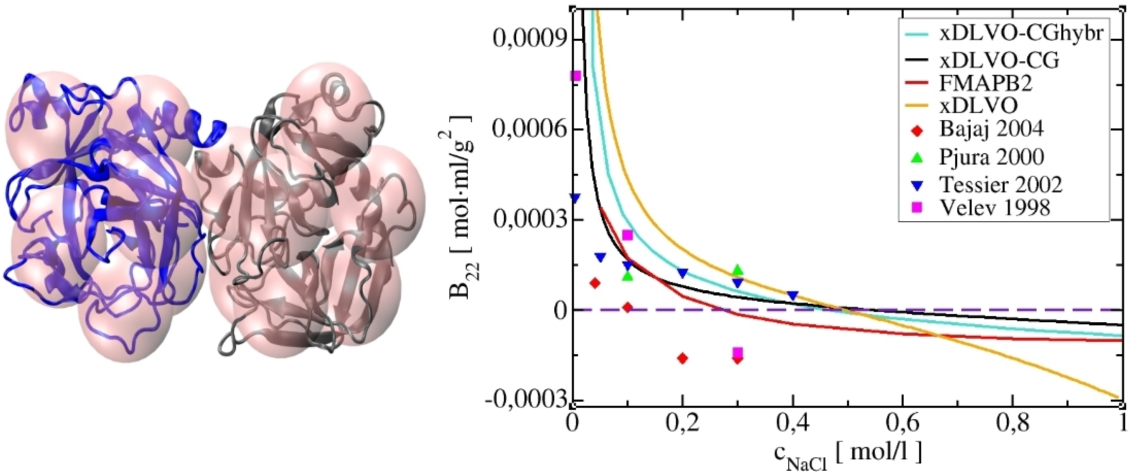


Figure 4.7.: Calculated B_{22} coefficients for ChymA at pH 3 as a function of NaCl concentration. B_{22} values are compared with the values obtained with xDLVO-CG (black), FMAPB2 (red) and xDLVO (orange) models and experimental results (circles). The left side displays coarse-grained models of proteins superimposed on an all-atom representation. The experimental data labeled as 'Bajaj 2004' and 'Pjura 2000' were taken from [209, 208], while 'Tessier 2002' and 'Velev 1998' were taken from [178, 207]

4.3.2. Large proteins

4.3.2.1. Concanavalin A

ConcA is a large protein composed of 237 amino acid residues with a molecular mass of 50 kDa and a planar shape rich in antiparallel beta sheets. It occurs naturally in jack-beans and plays a role in carbohydrate binding. In biochemistry, ConcA is commonly used to characterize sugar-containing molecules and to purify glyco molecules in lectin-affinity chromatography. The protein exhibits a specific dimer-tetramer equilibrium that depends on solution conditions, existing as a homodimer at pH lower than 7 and as a homotetramer at pH higher than 7. We performed B_{22} calculations of ConcA at pH 4 and pH 5 and compared them with experimental values reported in [127], where similar interaction chromatography was used to measure B_{22} values. Since ConcA exists as a dimer at this acidic pH range, calculations were performed between pairs of dimers. The experimental measurements showed that ConcA exhibits attractive PPIs, which cross the zero point at approximately 0.12 M NaCl.[127] The B_{22} coefficients calculated by the FMAPB2 and xDLVO models do not align with the experimental data, and they deviate from experiments significantly. In contrast, xDLVO-CGhybr and xDLVO-CG models can reproduce the general trends of experimental data and predict overall attractive interactions between proteins in solution. However, the agreement between experimentally determined and theoretically determined values was less quantitative in comparison to other proteins, which highlights the limitations of our model. At low ionic strengths, the xDLVO-CGhybr values matched the experimental data well, while at higher ionic strengths, the data was positively shifted compared to the experimentally measured values. At pH 4, the xDLVO-CG values were more positive than both the experimental and xDLVO-CGhybr values. However, at pH 5, the xDLVO-CG model provided a nearly quantitative match with the experimentally measured data and outperformed the xDLVO-CGhybr model. The suboptimal performance of the xDLVO-CGhybr model for the ConcA protein could be attributed to several factors, such as the uncertainty in the correspondence between the protein charges assigned by the Propka method and the actual physical charges. Moreover, the protein's ability to selectively absorb specific ions, which can alter its effective charge, was not accounted for in this model. The ConcA protein is also flexible and can adopt various conformations, especially in the

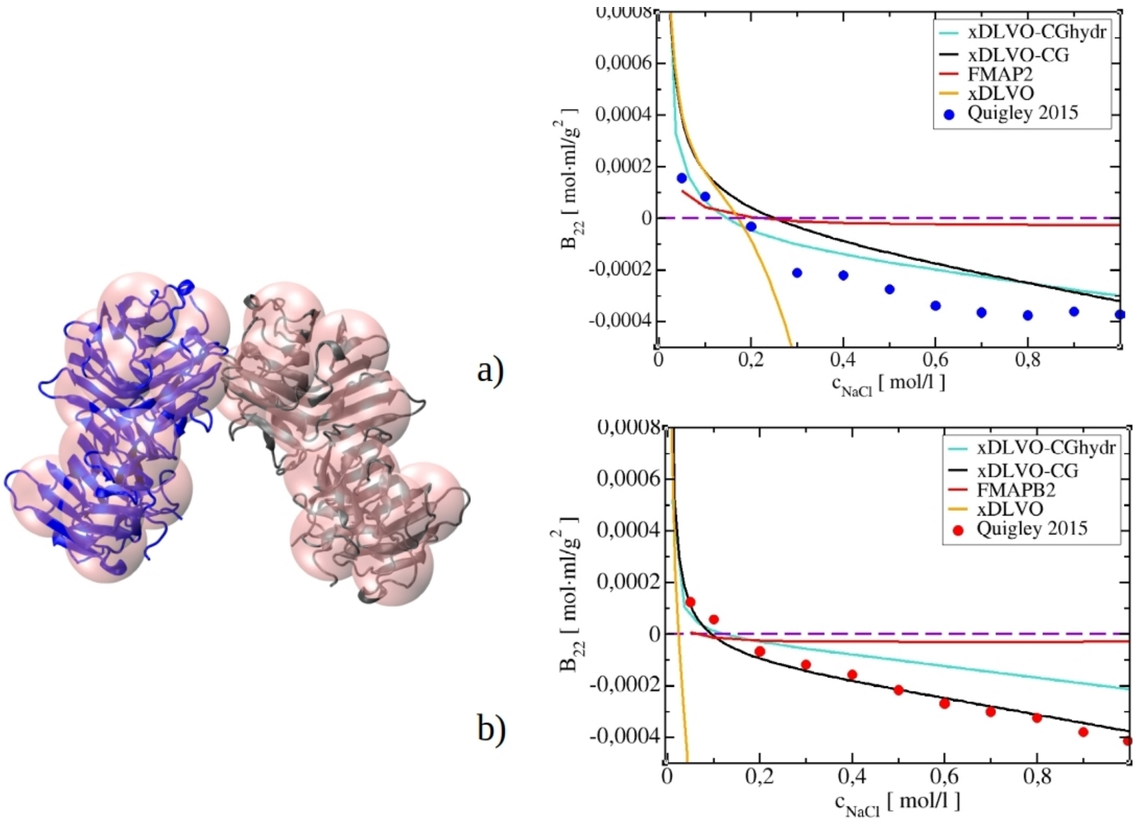


Figure 4.8.: Calculated B_{22} coefficients for ConcA at a) pH 4 and b) pH 5 as a function of NaCl concentration. B_{22} values are compared with the values obtained with xDLVO-CG (black), FMAPB2 (red) and xDLVO (orange) models and experimental results (circles). The left side displays coarse-grained models of proteins superimposed on an all-atom representation. The experimental data, labeled as 'Quigley 2015' for different pH values, were taken from [127].

presence of divalent ions. Furthermore, computational constraints limited our model's evaluation of the PMF over a limited set of 83 relative protein orientations, while the B_{22} value should measure interactions at all possible orientations without favoring individual conformations, which is computationally unfeasible. More advanced methods could be used to address these limitations by assessing the most relevant relative configurations and evaluating a larger set of relative orientations.

4.3.2.2. Bovine Serum Albumin

BSA is a member of the serum albumin family, and is known for its ability to bind ligands, including drugs, nutrients, and metals. With a mass of 65 kDa and an irregular shape, BSA is composed of 583 amino acid residues and a CG structure consisting of 20 beads. In this study, we conducted calculations under conditions of pH 7.4 and NaCl ionic strength up to 1 M, and compared our results with previously reported experimental values.[134, 135, 181] Our calculations, using the xDLVO-CGhybr model, demonstrated improved results compared to previous calculations using the xDLVO-CG model, as shown in Figure 4.9. Our B_{22} values were shifted closer to the experimental values reported by Ma et al., and remained positive up to 1 M NaCl. While our calculated data were closer to FMAPB2 data, FMAPB2 values deviated from the experimentally measured value at 0.2 M, whereas the xDLVO-CGhybr model performed better at capturing experimental trends in the ionic strength range of 0.05 to 0.1 M NaCl. In contrast, at ionic strengths above 0.1 M, the spherical xDLVO model failed to reproduce B_{22} coefficients, with values decreasing abruptly towards the negative range. Our findings highlight the limitations of simple

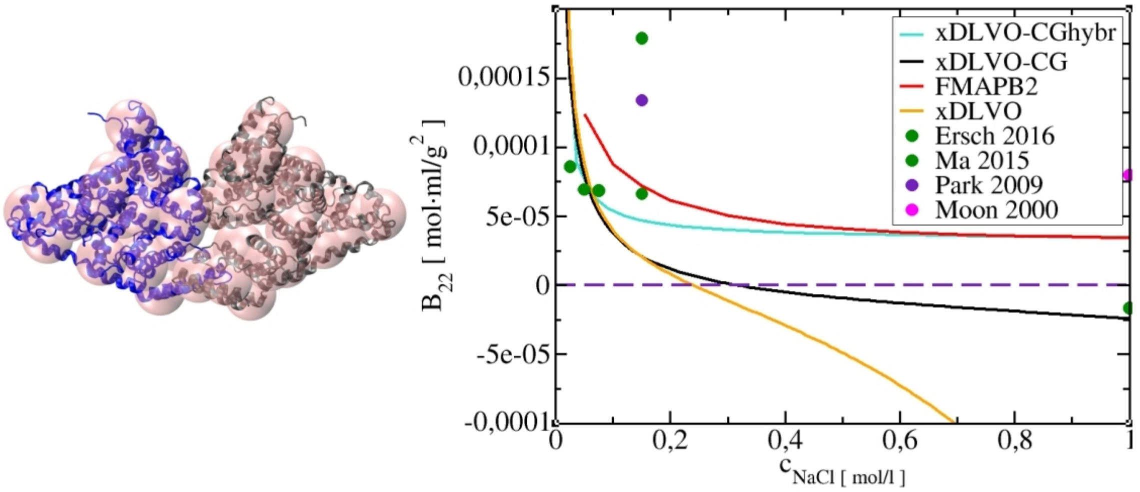


Figure 4.9.: Calculated B_{22} coefficients for BSA at pH 7.4 as a function of NaCl concentration. B_{22} values are compared with the values obtained with xDLVO-CG (black), FMAPB2 (red) and xDLVO (orange) models and experimental results (circles). The left side displays coarse-grained models of proteins superimposed on an all-atom representation. The experimental data labeled with 'Ma 2015' were taken from [135], while the data labeled with 'Ersch 2016,' 'Park 2009,' and 'Moon 2000' were taken from [134, 181, 182], respectively.

spherical models like (x)DLVO in predicting the behavior of complex proteins like BSA, as such models heavily rely on fitted parameters to achieve quantitative agreement with experimental data. In contrast, our results suggest that the xDLVO-CGhybr model is a more suitable approach for accurately predicting the behavior of BSA under both low and high ionic strength conditions. This underscores the importance of developing advanced models that incorporate the structural and dynamic complexities of proteins to better understand their behavior in various environments. Such studies, when combined with experimental measurements, will advance our understanding of the molecular mechanisms underlying the ligand binding capacity of BSA and related proteins, leading to the development of new therapeutic approaches for various diseases.

4.3.2.3. Human Immunoglobulin Type I

IgG1 is a crucial subclass of monoclonal antibodies that play a significant role in the immune system by recognizing and binding to specific antigens. These molecules have extensive biotechnological and pharmaceutical importance and are frequently used in clinical therapies. It is, therefore, essential to develop stable formulations that do not aggregate over time. Several studies have been conducted using DLVO or xDLVO to model second osmotic coefficients of monoclonal antibodies in different solution conditions,[140, 119] as well as to study PPIs at higher protein concentrations.[142] These studies typically involved the use of various levels of coarse graining, and the models are often based on direct parametrization from experimental data. IgG1 has a molecular mass of 13.9 kDa and a characteristic T-shaped structure, composed of 644 residues. Our calculations of IgG1 were performed at pH 6.5 and compared with experimental results reported in the literature, as shown on Figure 4.10. B_{22} data calculated by all models were in closer agreement with experimental values reported by Roberts[187] than by Le Brun,[200] which gave too repulsive values. Our new model showed a clear improvement in modeling second osmotic virial coefficients compared to the xDLVO-CG model, yielding values that were closer to the experimentally determined ones and similar to those obtained using the all-atom FMAPB2 model. In contrast, the simplified xDLVO model yielded data that were completely outside the experimental range, further highlighting its limited predictive power for large and irregular proteins. Our study emphasizes the need for more rigorous theoretical approaches to obtain more quantitative results, particularly for large and complex proteins such as IgG1.

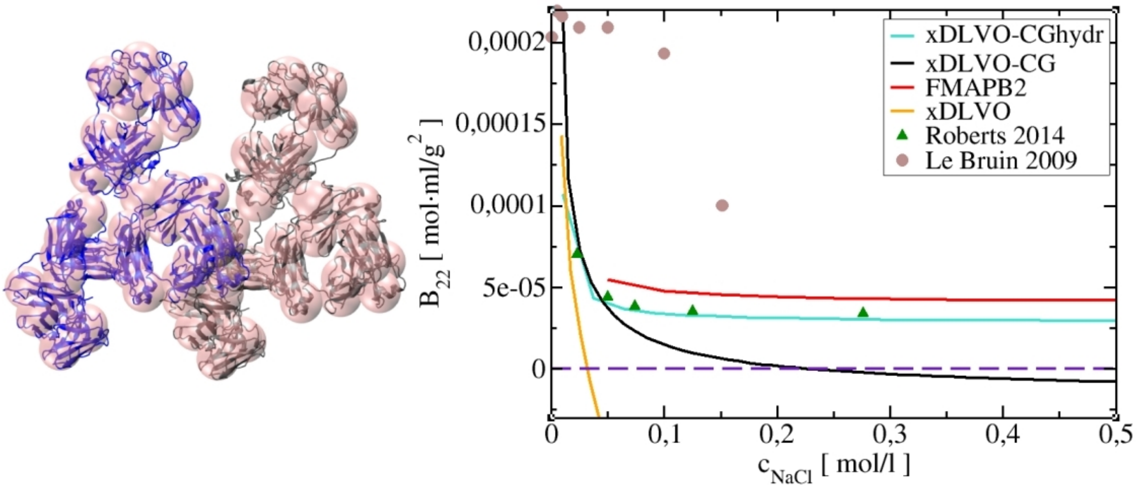


Figure 4.10.: Calculated B_{22} coefficients for IgG1 at pH 6.5 as a function of NaCl concentration. B_{22} values are compared with the values obtained with xDLVO-CG (dashed red), FMAPB2 (dashed orange) and xDLVO (green dots) models and experimental results (circles). The left side displays coarse-grained models of proteins superimposed on an all-atom representation. The experimental data, labeled as 'Roberts 2014' and 'Le Bruin' were taken from [187, 200].

4.4. Conclusion

We have developed a new xDLVO-CGhybr model that represents a significant improvement over our previously reported xDLVO-CG model. The xDLVO-CGhybr model uses a hybrid approach to calculate the electrostatic part of the potential of mean force, which combines Poisson-Boltzmann theory and Debye-Hückel theory to accurately calculate the electrostatic contribution to the total interaction potential for proteins of arbitrary size and shape. Additionally, we introduced a coarse-grained Lennard-Jones potential that was carefully parameterized to match the reference all-atom potentials, allowing for accurate predictions of protein-protein interactions. To validate the accuracy of our model, we tested it on six different proteins with varying levels of complexity and shape, including bovine pancreatic trypsin inhibitor, ribonuclease A, chymotrypsinogen, concanavalin A, bovine serum albumin, and human immunoglobulin type I. Our results showed that the xDLVO-CGhybr model outperformed other theoretical models such as xDLVO and FMAPB2, giving improved predictions of B_{22} values. These results demonstrate the xDLVO-CGhybr model's potential as a reliable tool for studying protein interactions and their behavior in solution, particularly in the context of pharmaceutical and biotechnological applications. However, our xDLVO-CGhybr model does have some limitations, such as assuming rigid protein structures taken from crystal PDB databases and limited protonation schemes. Despite these limitations, our model has been shown to correctly predict the nature of protein-protein interactions in most cases, which is critical for predicting protein stability and solubility in solution. To further enhance the accuracy and versatility of our model, future developments could be pursued, such as using more advanced sampling techniques and incorporating machine learning algorithms to improve the speed and accuracy of interaction potential computations. These potential advances could help expand the scope of our model to predict protein interactions in a wider range of conditions and enable more efficient exploration of parameter space to optimize protein formulations for stability and solubility.

5. Exploring Specific Protein Interactions: Molecular Dynamics Simulations of Beta-Lactoglobulin and Hepatitis B Core Proteins

5.1. Introduction

Chapter 3 and Chapter 4 introduced xDLVO-CG and xDLVO-CGhybr models, respectively, which utilize low-resolution coarse-grained structures to assess nonspecific protein interactions across a wide range of solution conditions. However, accurately modeling specific protein-protein interactions that govern processes such as dimerization or assembly into ordered nanostructures requires higher structural resolution models. This is due to the fact that specific interactions are typically driven by residue-residue interactions and the formation of hydrogen bonds or salt bridges, which necessitate a more detailed representation of the protein structure. Additionally, the use of accelerated MD techniques (as discussed in Chapter 2) can aid in simulating these processes and evaluating the associated free energy changes. This chapter focuses on investigating protein interactions using molecular dynamics simulations, specifically for beta-lactoglobulin (BLG) and the Cp proteins of HBV. To investigate these proteins, different simulation methods were used due to their differing sizes. BLG was simulated using an all-atom MD approach, while coarse-grained simulations were used to simulate the Cp proteins. Given that the focus of this chapter is primarily on MD simulations of HBV core proteins, more detailed information about BLG proteins will be provided in a subsequent section. Through this approach, we aim to gain diverse perspectives and tackle questions that were previously unattainable with the models used in the preceding chapters.

As described in Chapter 1, VLPs based on the HBV core Cp proteins are promising therapeutic agents due to their ability to self-assemble in various systems.[45] These VLPs have been investigated for use in hepatitis B infection treatment and as nucleic acid carriers for gene therapy. HBV capsids are made of 120 dimers of core protein that form an icosahedral structure with T=4 quasi-symmetry.[49] In vitro assembly of Cp proteins into the VLPs is influenced by solution conditions and can be modulated by pH, temperature, or ionic strength. Purification is necessary to remove impurities and obtain pure Cp_2 dimers that can be reassembled with therapeutic nucleic acids. In addition to VLPs made from the naturally occurring Cp protein with 183 aminoacids (Cp183), researchers are investigating Cp proteins with shorter nucleic acid binding regions, with the aim to improve nucleic acid loading efficiency and facilitate the production and processing of HBV-based VLPs.[65, 63] Until now, the most of the experimental and theoretical research thus far has been conducted on truncated Cp proteins with 149 amino acids (Cp149), which lack a nucleic acid binding region and cannot encapsulate nucleic acids. Extensive research has been conducted on truncated Cp149 capsids, with experiments focused on capsid assembly, disassembly, stability, and preventing assembly through mutations or antivirals.[211, 212, 55, 213, 214, 215, 216, 217, 218, 219, 220] Some studies have also investigated the in vitro assembly of empty or RNA-containing HBV capsids using full-length Cp_2183 proteins, with notable differences from truncated Cp_2149 capsids.[221, 222, 66, 47] Full-length Cp_2183 dimers assemble at higher ionic strength (>0,25 M) due to strong electrostatic repulsion among positively charged C-ter domain. Theorethical studies of HBV capsids mainly focused on truncated Cp149 capsids, but some explored full-length capsids via thermodynamic models. Zlotnick et al. created a thermodynamic model for HBV Cp149 capsid assembly predicting assembly kinetics and thermodynamic parameters.[223] Molecular dynamics simulations investigated capsid flexibility, dynamics, and sodium ion distribution in Cp149 capsids, their structural stability with antiviral compounds, and disassembly under mechanical stress.[224, 225, 226, 227, 228, 229, 230, 231, 232] Coarse-grained simulations explored irreversible Cp149 capsid deformation.[160] For Cp183 capsids, thermodynamic models were developed which used low-resolution coarse-grained structures for Cp_2183 dimers and tangentially connected charged spheres for RNA chains.[233, 234, 235, 236] These models could predict the distribution of nucleic acids within the capsid, exposure of the C-terminal domain, and the optimal genome size for encapsidation.

The current understanding of HBV capsid assembly or disassembly is limited by the absence of theoretical studies on the interactions between protein-protein and protein-nucleic acids in HBV capsids and Cp dimers, particularly as the C-ter length changes. Although extensive research has been conducted on HBV capsid assembly, the detailed mechanisms involved in protein-protein and protein-nucleic acid interactions during capsid formation are still unclear. In order to address this gap, we used SIRAH coarse-grained force and umbrella sampling simulations of Cp_2 trimers to investigate the binding energy of Cp_2 dimers with different C-ter lengths (Cp_2183, Cp_2167, Cp_2164, Cp_2157, Cp_2154, and Cp_2149) in an explicit solvent at physiological temperature. Additionally, we simulated Cp_2 trimers in the presence of model DNA to study how nucleic acids stabilize the capsid structure. Our objective is to enhance our comprehension of the primary factors that drive HBV capsid assembly and stability, and to give insights for novel experimental protocols for more efficient production of HBV capsids.

5.2. Methods

5.2.1. System Setup and Molecular Dynamics Simulations

The all-atom structure of Cp protein was obtained from the PDB database under the code 6htx, and any missing atoms or residues were reconstructed using the Swiss-Model program. The protein structure was modelled up to residue 183 based on the 6htx crystal structure of Cp. To set the conformation of the flexible C-ter domain, which is not resolved in any of the available PDB structures, the Cp dimer structure obtained by Ingemar et al. was used.[221] They used Bayesian inference to obtain ensembles of conformations that match experimental SAXS spectra measured in diluted conditions where Cp dimers exist in an unassembled state. The crystallographic symmetry operations given in the PDB were then applied to the Cp_2183 dimer to obtain the structure of the whole capsid. To investigate the effect of changing the length of the C-ter domain, truncated trimers of dimers were obtained by in silico cutting the protein chains at desired positions. The comparison of the total number of atoms, beads and charge of each Cp_2 dimer is shown in Table 5.1. The MD simulations were conducted using the GROMACS program,[202, 112] and the SIRAH coarse-grained force field was used.[106] The reconstructed proteins were protonated at pH 7 using the PROPKA method[170, 171] via the PDB2PQR online web server.[172, 173] The fully protonated all-atom structures were then mapped onto coarse-grained structures using SIRAH tools.[237] Example of CG structure of Cp_2149 dimers obtained by SIRAH

Table 5.1.: The total number of atoms, beads and charges of Cp_2 dimers, truncated at different positions of the C-terminus

C_p, aa	Number of atoms	Number of beads	Charges
149	4722	1456	-14
154	4928	1518	-6
. 157	5026	1550	-4
164	5276	1628	2
167	5428	1670	8
183	5952	1856	16

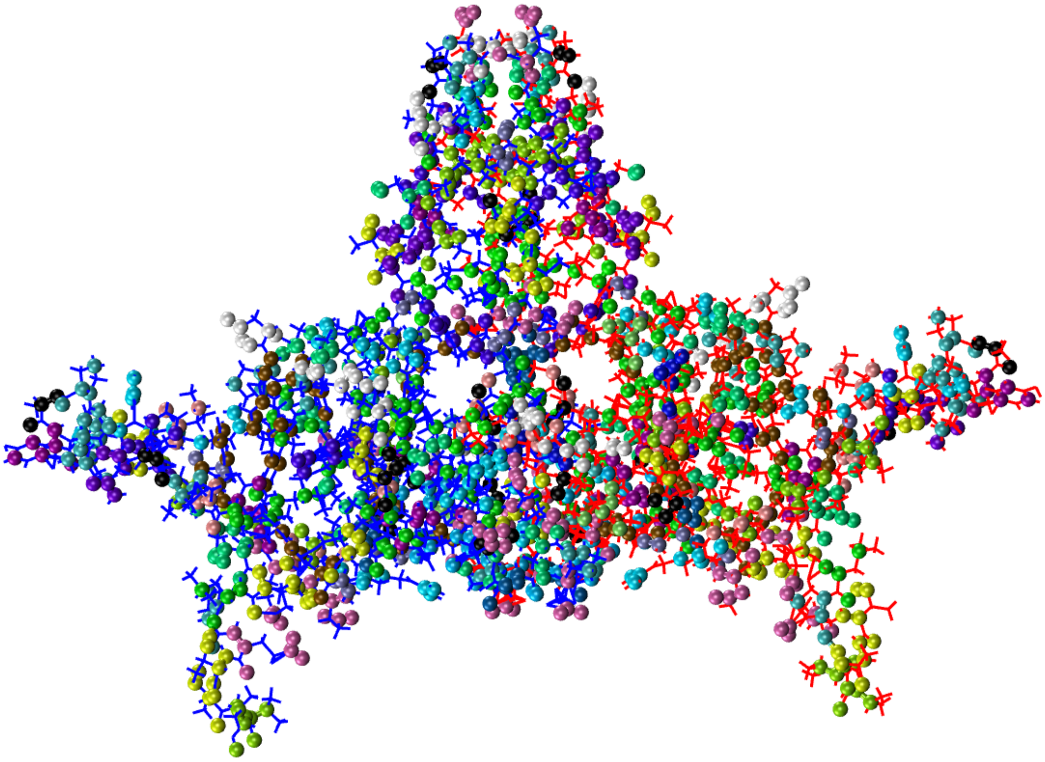


Figure 5.1.: Example of CG structure of $Cp_2$149 dimers obtained by SIRAH tools. For clarity, SIRAH CG structure is shown sumperimposed on all-attom structure (depicted by lines representarion).

tools is shown on Figure 5.1. The coarse-grained proteins were placed in a rectangular box and solvated using pre-equilibrated WT4 molecules from the SIRAH force field. The WT4 molecules placed within 0.3 nm from the protein were removed to allow for relaxation of protein side chains during minimization while resolving local water gradients in the equilibration stage. Na and Cl ions were added to the system to achieve a concentration of 37.5 mM, including electro-neutralising counterions. The system underwent a two-step energy minimization, first by the steepest descent algorithm was employed for 5000 steps, with protein backbone atoms harmonically restrained with 1000 kJ/mol·nm, and then by of an unrestrained minimization for another 5000 steps. Equilibration was performed under the NVT ensemble at 300 K using the V-rescale thermostat.[95] For the first 5 ns, all protein CG beads were harmonically restrained, and in the subsequent 25 ns, only backbone beads of proteins were weakly restrained, while solvent molecules and protein side chains were allowed to move and relax. The final 25 ns of unrestrained production runs were performed under NPT conditions, at 300 K and 1 bar pressure, using the V-rescale thermostat and

Parrinello-Rahman barostat with isotropic pressure coupling.[94] A timestep of 20 fs was used in all simulations, and the neighbour lists were updated every 10 steps. Electrostatic interactions were calculated using Particle Mesh Ewald,[238] with a direct cutoff of 1.2 nm and grid spacing of 0.2 nm, and van der Waals interactions were calculated with a cutoff of 1.2 nm.

5.2.2. Umbrella Sampling Simulations

US simulations were employed to simulate trimers of dimers of six different lengths at C-terminus in order to estimate protein-protein interactions and investigate the effects of nucleic acid interactions on trimer disassembly. The COM distance between the first two Cp_2 dimers and the third dissociated Cp_2 dimer was selected as the reaction coordinate. To create the initial configurations for the US windows, pulling simulations were performed using the last snapshot of a production run. These pulling simulations involved moving the Cp_2 dimer away from the trimer of dimers along the vector that connects the COM distances of the two groups, as depicted in Figure 5.2. During the pulling simulations, the first two dimers were harmonically restrained, and a force was applied to separate the third dimer. The pulling force used was 500 kJ/mol·nm, and the rate of pulling was 0.001 nm/ps. The US windows were asymmetrically distributed, with a spacing of 0.03 nm used for the first 35 umbrella windows and 0.08 nm used for the remaining windows. A total of 104 US windows were generated, spanning a distance range from R_0 to roughly $R_0 + 4nm$, where R_0 represents initial COM distance. The system was equilibrated for 3.5 ns using the NPT ensemble at 300 K and 1 bar in each window, followed by a 100 ns MD run with a V-rescale thermostat and Parrinello–Rahman barostat under the NPT conditions.[94] In each window, a bias harmonic potential of 1500 kJ/mol·nm was applied. The WHAM was utilized to calculate the potential of mean force by removing the effect of applied bias.[115] 5.1.

5.2.3. Molecular Docking and MD Simulations of Cp_2 Trimer-DNA Complexes

In order to investigate the effect of nucleic acids on protein interactions within the Cp_2 trimer, simulations were conducted with a DNA molecule (PDB code 1hw2) attached to the trimer. The SIRAH force field and the same computational setup as previously described were employed. It should be noted that RNA molecules could not be used in these simulations due to the lack of force field parameters in SIRAH. The DNA molecule was docked onto the trimer of dimers using the HDOCK web server (available at http://hdock.phys.hust.edu.cn/).[239] The HDOCK method utilizes a hybrid docking strategy that automatically incorporates binding interface information into traditional global docking. This is achieved through the use of an iterative knowledge-based scoring function.[240, 241] The top 100 docked structures with the best docking score, representing the best binding affinity, were simulated for 50 ns using MD. The binding energy between the DNA and the Cp_2 trimers was calculated using the Molecular Mechanics Generalized Born Surface method (MM-GBSA) method[242] and analysed with the gmmpbsa analysis tool.[243] The binding energy between the Cp_2 trimer and DNA was determined using a single trajectory approach by subtracting the total molecular mechanics and solvation energy of the complex from the total energy of the individual components. The best five structures were used for umbrella sampling simulations, where the Cp_2 dimer was separated from the trimer with the DNA molecule attached in the vicinity to determine its impact on trimer disassembly. The DNA molecule was not part of any pulling group and was not restrained during the pulling and umbrella sampling simulations.

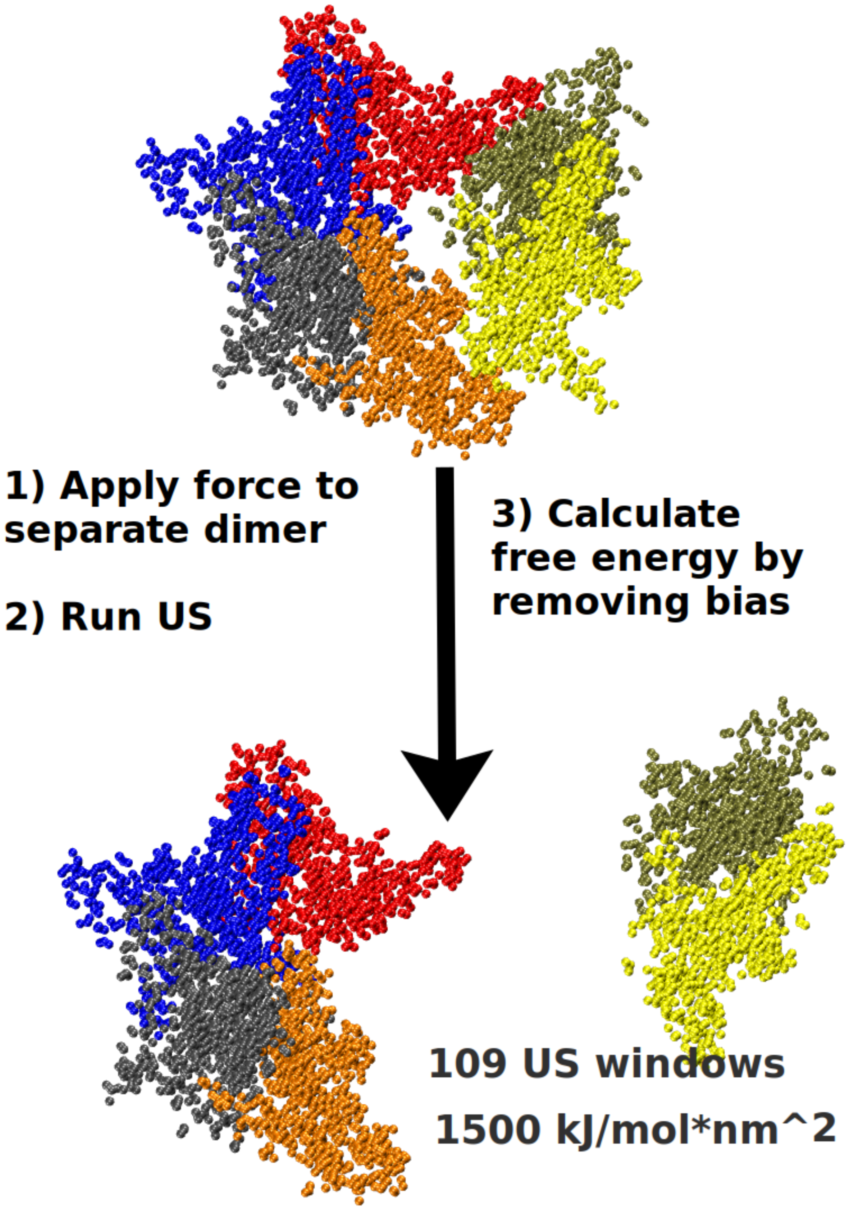

1) Apply force to separate dimer
2) Run US
3) Calculate free energy by removing bias
109 US windows
1500 kJ/mol*nm^2

Figure 5.2.: Scheme of employed umbrella sampling simulations

5.2.4. Molecular Dynamics Simulations of Beta-lactoglobulin

US simulationa were performed between BLG monomers using the CHARMM36m force field[203] and SPC water model[244] in GROMACS (version 2019.2).[202] The BLG monomers were aligned to the x-axis with periodic boundary conditions in a rectangular box measuring 18.0 x 9.5 x 9.5 nm^3. The BLG protein structures were protonated at pH 3 and pH 7 according to the PROPKA method and PDB2PQR webserver. Sodium and chloride ions were added to achieve 10 mM and 100 mM salt concentrations. The systems were first minimized using the steepest descent algorithm with position restraints on the protein heavy atoms for 30,000 steps. Equilibration was then performed under NVT and NPT ensembles at 300 K and for 400 ps each using the Berendsen thermostat.[93] The Berendsen weak coupling method was used to maintain pressure isotropically at 1.0 bar. The simulations were performed with a timestep of 2 fs and short-range nonbonded interactions were cut off at 1.2 nm. The particle mesh Ewald (PME) algorithm was used to evaluate full electrostatic interactions beyond 1.2 nm.[238] Pulling simulations were performed along the x-axis for 1.2 ns with a spring constant of 1000 kJ (mol nm2)$^{-1}$ and a pull rate of 0.005 nm/ps, as shown on Figure 5.3. For BLG simulation at pH of 3 and a salt concentration of 100 mM sodium chloride, the spring constant was 1500 kJ (mol nm2)$^{-1}$. 5.1. The starting configurations for the US windows were taken from the snapshots of pulling trajectories. A sampling window spacing of 0.0625 nm was used for COM distances shorter than 4.1 nm and 0.125 nm for distances longer than 4.2 nm, resulting in an asymmetric distribution. A total of 61 US windows were generated, and each window underwent equilibration using NPT ensemble at 300 K and 1 bar for 400 ps, followed by a 20 ns MD run with NPT ensemble using the Nose-Hoover thermostat[245] and Parrinello-Rahman barostat.[94] Analysis was performed using the WHAM.[115]

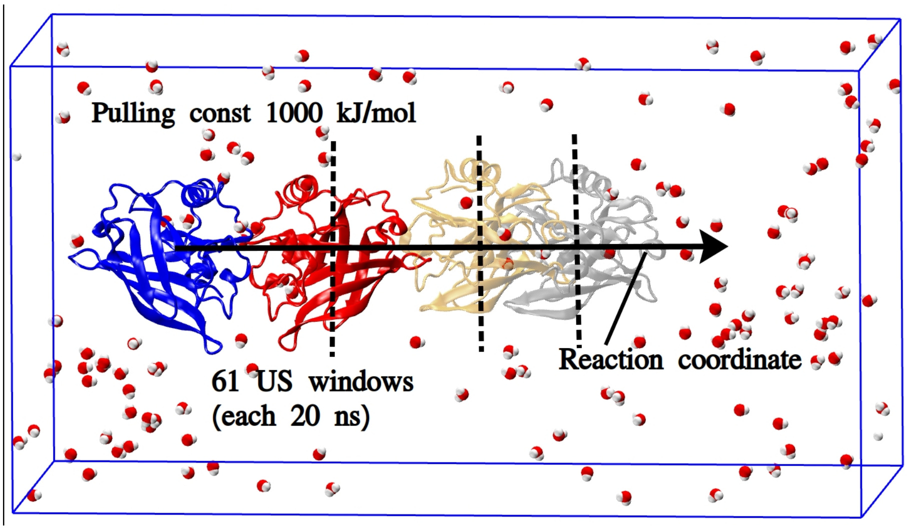


Figure 5.3.: Scheme of employed umbrella sampling simulations for Beta-Lactoglobulin.

5.3. Results and Discussion

5.3.1. The Interplay of Specific and Nonspecific Interactions of Beta-Lactoglobulin

[1] Proteins are composed of diverse amino acids that give rise to both nonspecific (e.g. electrostatic and dispersion) and specific interactions (e.g. hydrogen bonding), which play a crucial role in determining the self-assembly and dispersity of protein systems. These interactions can lead to the formation of larger protein aggregates and are influenced by factors such as pH and salt concentration. BLG is a small globular protein found in milk, is an example of a protein whose self-assembly is influenced by both nonspecific and specific interactions. BLG exists as a monomer or dimer in diluted solutions depending on pH and ionic strength.[247] At pH values of pH < 3 and pH > 8, BLG monomers are dominant, while at intermediate pH values, dimers and higher oligomers occur. The pH-dependent behavior of BLG is due to the charge states of many amino acids, which are influenced by pH and ionic strength. The key role of electrostatic interactions in BLG's protein-protein interactions is due to the pH-dependent charge states of many amino acids in the protein sequence(Figure 5.4a). However, the dimer binding site of BLG is stabilized by several hydrogen bonds.[248, 249] Amino acids at the binding interface become positively charged at low pH, which prevents dimer formation unless higher salt concentration is added to neutralize the charges. Thus, at intermediate pH values, an interplay of PPIs of various characters occurs, and BLG exists in solution as a mixture of monomers and dimers, with ratio of species dependent on pH and salt concentration. The B_{22} coefficients and binding

[1] The content presented in this section is based on the work published in Ref. [246], which has been modified and restructured to align with the format of thesis.

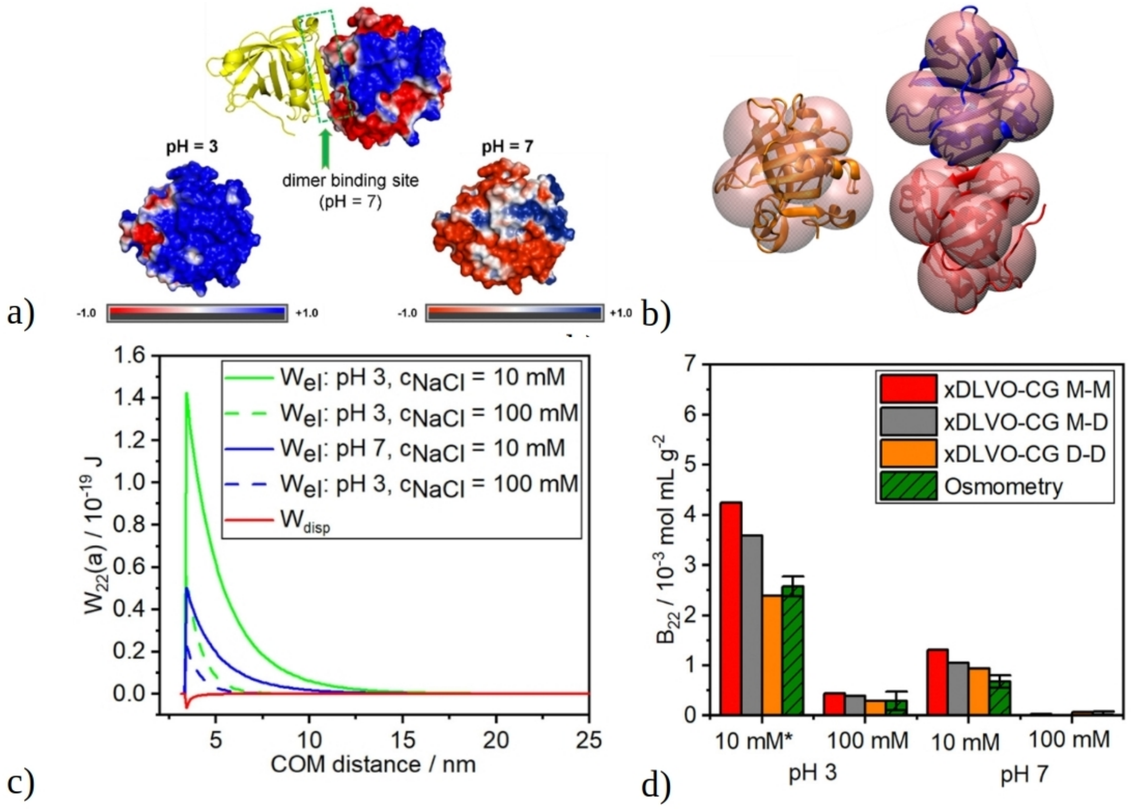


Figure 5.4.: a) Graphical representation of the charge distribution of the BLG monomer and dimer as a function of the solution pH. b) Coarse-grained representation of BLG monomer (on the left) and dimer (on the right) used to calculate B_{22} coefficient with xDLVO-CG. c) Change of electrostatic potential between BLG monomers at different pH and NaCl concentration d) The osmotic second virial coefficient, B_{22}, at different pH 3 and pH 7 and 10 mM and 100 mM Nacl, calculated for monomer–monomer (M–M), monomer–dimer (M–D) and dimer–dimer (D–D) pairs. The calculated values were compared with membrane osmometry measurements (green). Adapted with permission from [246]. Copyright 2022 Royal Society of Chemistry.

energies for dynamically interacting systems were determined by combining xDLVO-CG calculations with umbrella sampling MD simulations. The binding energy of BLG in its dimer state was estimated as a function of pH and ionic strength, and the calculated values were compared with experimental data obtained Uttinger et al.[246] The monomer-dimer equilibrium constant was determined using analytical ultracentrifugation (AUC), which allowed the exact ratio of monomers vs dimers to be determined. This information was used to interpret membrane osmometry experiments that gave the averaged B_{22} of the solution. The results, presented in Figure 5.4d, show that the calculated B_{22} values for BLG at pH 3 and pH 7 and at NaCl concentrations of 10 mM and 100 mM using xDLVO-CG are consistent with membrane osmometry measurements. The B_{22} values suggest that electrostatic repulsion interactions dominate due to the total charge of BLG monomers of +18 and -8 at pH 3 and pH 7, respectively, resulting in more positive B_{22} values, as shown on Figure 5.4c. The pH dependence of B_{22} is noticeable, and an increase in ionic strength promotes self-assembly towards the dimeric state. However, B_{22} changes slightly differently for the monomer–monomer, monomer–dimer, and dimer–dimer cases, with more pronounced differences at pH 3. Nonetheless, the B_{22} values at the measured salt concentrations do not differ significantly, indicating an equilibrium state. The results of US calculations which simulated dimer dissociation are presented in Figure 5.5, and show that the stability of BLG dimers is strongly influenced by the pH and salt concentration of the solution. At pH

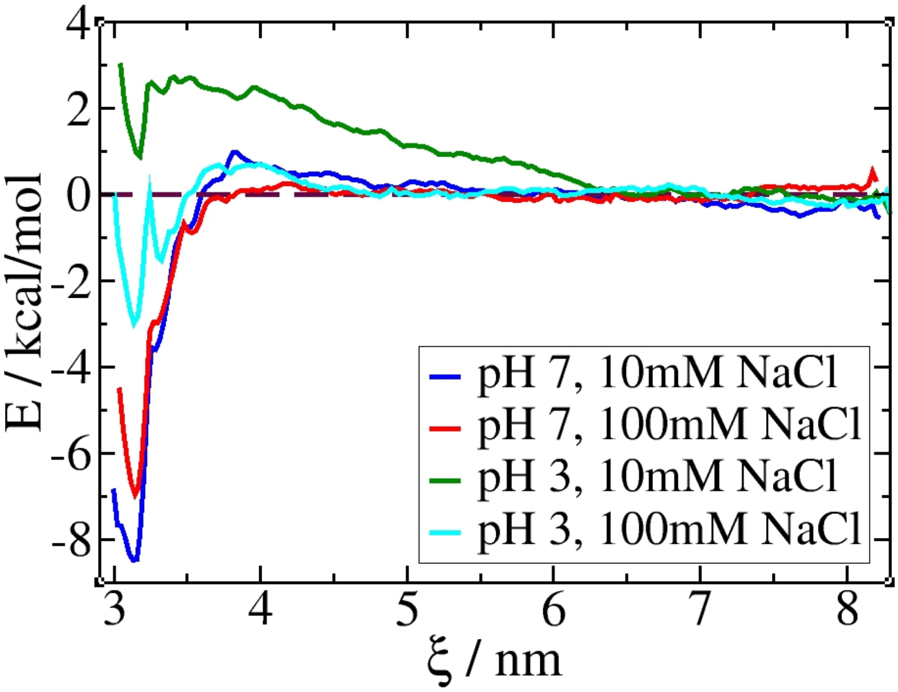


Figure 5.5.: Free energy of BLG dimerization at pH 3 and pH 7 and for 10 mM and 100 mM NaCl calculated umbrella sampling simulations at 300 K and atmospheric pressure. Adapted with permission from [246]. Copyright 2022 Royal Society of Chemistry.

7 and NaCl concentrations of 10 mM and 100 mM, the binding energy of the BLG dimer is -8.3 kcal/mol and -6.9 kcal/mol, respectively, indicating that the dimers are relatively stable under these conditions. These results are consistent with the experimental data reported.[246] In contrast, at pH 3 the BLG dimers are less stable, with a binding energy of -4.2 kcal/mol at 100 mM NaCl. The equilibrium between monomers and dimers is shifted towards monomers compared to pH 7. At pH 3 and 10 mM NaCl, the energy of binding is repulsive (+1.2 kcal/mol), which means that dimers are thermodynamically unstable at these conditions. These results highlight the importance of pH and salt concentration in determining the stability of BLG dimers, and provide insight into the mechanisms underlying their formation and dissociation in different solution conditions.

5.3.2. Assessing the Stability of Hepatitis B Capsid Fragments using Coarse-Grained Simulations and Free Energy Methods

The stability of HBV based VLPs is influenced by a variety of interactions, with electrostatic forces often playing a dominant role. To investigate the PPIs between core protein homodimers (Cp_2) of VLPs with different C-ter length and explain their self-assembly propensity versus interactions with nucleic acids, we employed umbrella sampling simulations. For computational efficiency, we chose to simulate a smaller capsid fragment, a trimer of Cp_2 dimers (as illustrated in Figure 5.2). Experimental studies have identified the trimer of dimers as an important intermediate nucleate formed during capsid assembly or disassembly reactions.[217, 223] Using the umbrella sampling simulations, we calculated the free energy of binding in a complex, specifically the separation of a Cp_2 dimer from a trimer of dimers (see Figure 5.2), with different core protein lengths considered. The addition of residues at the C-terminus domain from Cp149 to Cp183 (as shown in Figure 5.1) changes the total protein charge (see Table 5.1). For instance, the $Cp_2$149 dimer has a charge of -14e, while the $Cp_2$183 dimer has a charge of +16e. This shift from negative to positive charge is primarily due to the addition of positively charged arginine residues, which are abundantly present in the flexible 150-183aa C-ter located inside the capsid.

5.3.2.1. Steered Molecular Dynamics Simulations

In order to investigate the rare event of dimer separation, steered MD simulations were performed by applying an external force to dissociate the dimer from the trimer of dimers. Conventional MD simulations would require prohibitively long simulation times to observe such an event, which motivated the use of steered MD. The force response during the pulling of the dimer from the trimers was measured, and the results revealed that $Cp_2$164 and $Cp_2$183 dimers required the highest force to dissociate, while truncated $Cp_2$149 and $Cp_2$154 dimers required the lowest force. These findings suggest that the length of the C-terminal domain affects the stability of the Cp_2 dimer within the trimer of dimers, with longer C-terminal domains potentially resulting in stronger dimer interactions due to additional stabilizing interactions between the C-terminal domains and the assembly competent core. In addition to calculating the force response, the pulling simulations provide insights into the molecular-level mechanism of dimer dissociation. The force gradually built up until it reached a maximum value at approximately 2.5 ns, at which point the force persisted at its maximum value before gradually decreasing. This behavior corresponds to the beginning of dimer dissociation, where fewer residues are in contact, and less force is needed to dissociate the remaining residues. Once the dimer is fully separated from the trimer, less force is required to pull it away, although some force is still necessary. Figure 5.6b illustrates a snapshot of the $Cp_2$183 system at the end of the steered MD simulation, after it has fully separated from the trimer. By understanding the factors that influence the stability of Cp_2 dimers within the trimer of dimers, this knowledge could enable researchers to engineer VLPs with specific properties that are desirable for drug delivery and vaccine development. The ratio of forces obtained during steered MD simulations can provide insight into the relative stability of different structures, as more stable structures typically require greater force to separate. However, direct comparison of force response is complicated by the fact that each Cp_2 dimer has a different mass, with heavier dimers requiring more force to separate. Furthermore, the presence of flexible residues on the C-terminus side may cause partial unfolding during dimer separation, requiring additional force and further complicating direct force comparison. The next section of this chapter will present the results of US simulations to calculate the binding energy in a quantitative manner, using the snapshots obtained from the pulling trajectories as starting configurations for a series of US windows.

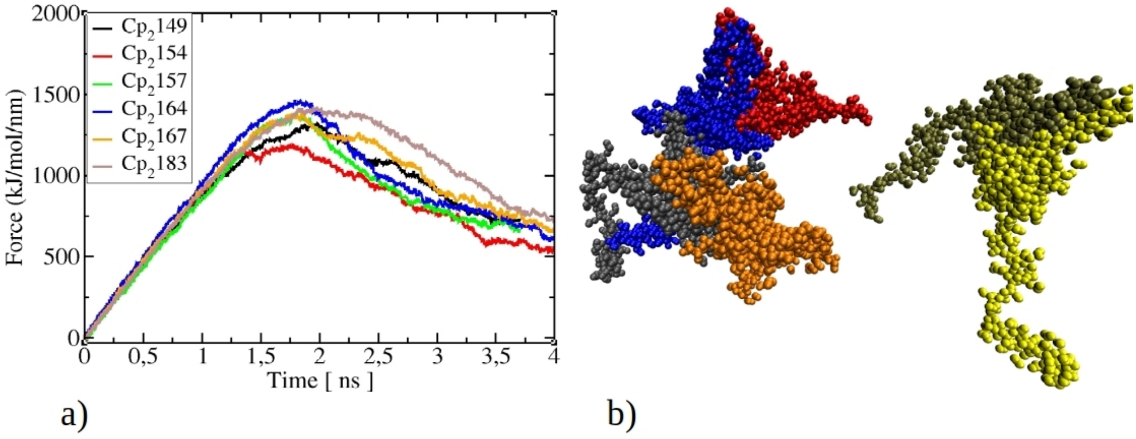


Figure 5.6.: a) The force response during pulling Cp_2 dimer from from trimer of dimers as a function of core protein length b) Illustration of last snapshot from steered MD upon separating $Cp_2$183 dimer from from trimer of dimers

5.3.2.2. Free Energy of Binding from Umbrella Sampling

The results of US simulations shown in Figure 5.7 indicate that the assembly of Cp_2 trimers of different lengths is thermodynamically favorable. The PMF obtained by umbrella sampling revealed a negative free energy difference between the assembled and dissociated states for all six proteins, indicating that assembly is energetically favorable. Interestingly, even though the $Cp_2$149 trimer has a relatively high charge (-14), the protein-protein interactions were found to be only modestly attractive. This suggests that the efficient VLP formation of these proteins is achieved through the delicate screening of local charge repulsions at relatively low ionic strengths. To assess the binding energy of Cp149 obtained from our simulations, we compared it to the values reported in the literature. Unfortunately, there is no data available for proteins of other lengths. Ceres et al. estimated the effective subunit-subunit energy to be -3.1 kcal/mol at 0.15 M NaCl, increasing up to -4 kcal/mol at 0.7 M NaCl by using a thermodynamic model to analyze the capsid assembly kinetics obtained by MALLS-SEC (Multi-angle laser light scattering analysis of SEC) measurements.[59] In a similar study, Chevreuil et al. estimated the effective subunit-subunit interaction to be -6kbT (-3.56 kcal/mol) by fitting a theoretical model to the capsid melting curve.[212] Other experimental studies also reported binding energies ranging from -4.15 kcal/mol at 50 mM NaCl to -5.1 kcal/mol at 500 mM NaCl.[216] It is worth noting that the experimental values for subunit-subunit interactions refer to the effective interaction between subunits averaged over 240 possible subunit contacts within the capsid. In contrast, in our MD simulations, the Cp_2 dimer has contacts with two other dimers within the trimer of dimers, so the effective subunit-subunit contact is equal to half of the binding energy obtained by US (i.e., -3.9 kcal/mol). Our simulation value is slightly more attractive than the experimentally reported values because our simulations were performed at 37.5 mM NaCl, while the reported experiments were conducted at 0.15 to 0.7 M NaCl. However, it should be mentioned that we have not simulated the whole capsid, and the experimentally reported values were obtained by thermodynamic models that do not take into account the full atomic structure. It is also worth noting that the experimentally observed change in subunit-subunit energy with increasing NaCl concentration (from 0.05 M to 0.7 M NaCl) is relatively low. Overall, our simulation results indicate a favorable thermodynamic assembly of $Cp_2$149 trimers and shed light on the importance of screening local charge repulsions for efficient VLP formation. Subunit-subunit interactions are fundamental to the assembly of viral capsids. Due to the limited experimental data on subunit-subunit interaction energies, computational simulations can provide valuable insights into these interactions. The Cp149 proteins are the only ones for which experimental values of subunit-subunit interaction energies are available. However, simulations have shown that positively charged arginines at the C-terminus can significantly perturb the PPI between Cp_2 dimers. This effect is not uniform, and $Cp_2$154 and $Cp_2$183 trimers have lower dimer-dimer interactions compared to $Cp_2$149, leading to weaker attractive forces that are not sufficient to induce efficient self-assembly into VLPs. The $Cp_2$183 trimer, for instance, has the lowest binding energy of all six Cp_2 trimers. However, the free energy minima is still negative, indicating weakly attractive overall PPIs. Higher ionic strengths (>0.25 M NaCl) are needed to screen electrostatic interactions and initiate efficient VLP assembly, either without or with nucleic acids, giving empty or nucleic acids encapsulated capsids. Over 90% full-length Cp183 capsids form empty under in vivo conditions, and low solubility for $Cp_2$183 dimers were reported.

In contrast, the other three Cp_2 dimers with intermediate core protein lengths interact more strongly within the trimer, stabilizing it and increasing the final trimer dissociation energy. $Cp_2$157 trimers are more attractive than $Cp_2$154 trimers, with a free energy minima of -7.4 kcal/mol. The stabilizing effect is strongest for the trimer formed from $Cp_2$164 proteins, with a free energy minima of -10.5 kcal/mol. Previous theoretical investigations

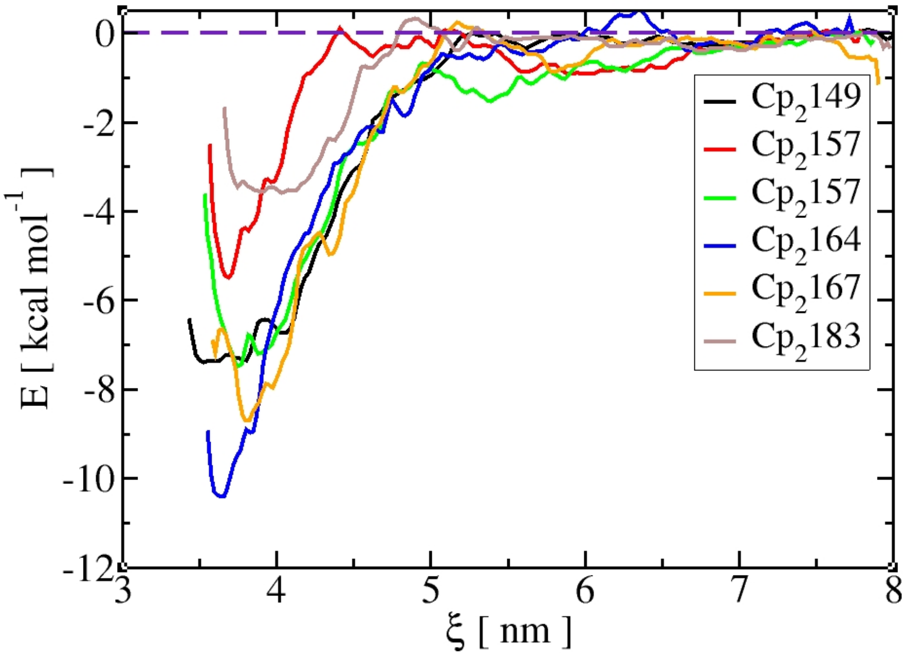


Figure 5.7.: Free energy change of trimer dissociation obtained by US simulations at 300 K and atmospheric pressure. US was performed on SIRAH trimers of dimers of 6 different lengths of core protein (Cp149, Cp154, Cp157, Cp164, Cp167, Cp183)

have suggested that capsids with intermediate lengths, such as Cp164, may be more stable than Cp167 and Cp183 capsids. However, experimental works have reported solubility issues with Cp_2 dimers with intermediate lengths, indicating that processing these proteins may be cumbersome. Weak subunit-subunit interactions promote efficient capsid growth by enabling fast dimer exchange, which prevents kinetic traps and corrects misassembled structures. Even small alterations in binding energy can have a significant impact on VLP formation. Core protein allosteric modulator components that bind into $Cp_2$149 dimer interfaces and slightly increase the energy of interaction (from -3.1 to -4.4 kcal/mol at 0.15 M NaCl) can prevent the assembly or enhance it into aberrant structures.[250, 251] Therefore, it is essential to understand the properties of subunit-subunit interactions to optimize viral capsid assembly for therapeutic purposes.

5.3.2.3. Analyzing Enthalpy Terms From Molecular Dynamics Trajectories

To gain a deeper understanding of protein-protein interactions and the factors influencing the interaction between Cp_2 dimers, we performed a thorough analysis of the dynamical change of enthalpy terms during pulling and US simulations. Our results, shown in Figure 5.8a, indicate that the magnitude of Lennard-Jones interactions becomes slightly more attractive as the core protein length increases, with Cp164 and Cp183 proteins exhibiting more attractive interactions at the initial COM distances. This is likely due to the longer protein chains, which may establish more contacts and contribute to a higher magnitude of Lennard-Jones interactions. However, the difference between Lennard-Jones interactions is not substantial, and the trends may slightly vary across different runs. In Figure 5.8b, we show the change in Coulomb interactions between the Cp_2 dimer and the rest of the trimer during the US simulations for different lengths of the C-ter domain. Electrostatic interactions are one of the most critical driving forces during the assembly and disassembly of capsids, and in this case, they are significantly modulated by the core protein length. Cp183 and Cp149 exhibit the largest Coulomb electrostatic repulsion in the gas phase due to their total charge among all six Cp_2 trimers (-42 and +48, respectively). At the shortest COM distances, $Cp_2$183 exhibits a repulsive energy of approximately 3200 kcal/mol, while

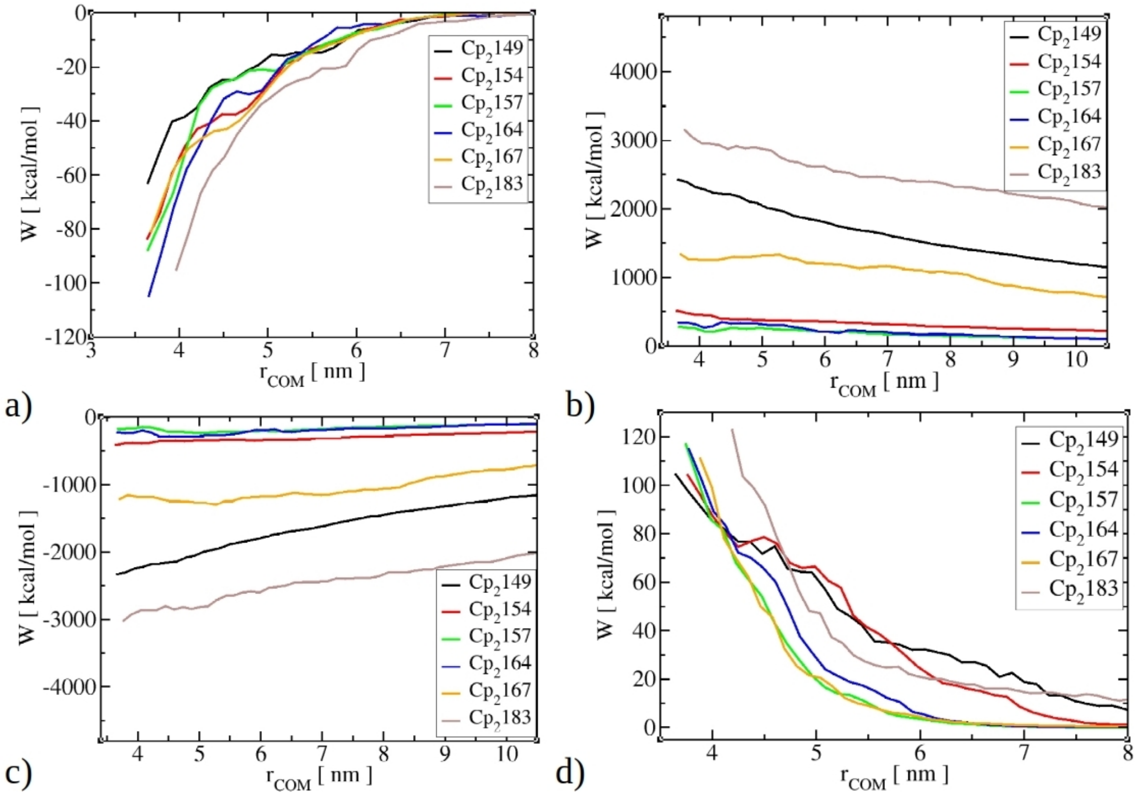


Figure 5.8.: Dynamical change of enthalpic interactions between Cp_2 dimer and Cp_2 dimers of dimers during US runs of a) Lennard Jones interactions b) Coulomb energy (calculated by subtracting the total Coulomb energy of complex from energies of single components) c) Solvation energy (calculated on all-atom structures backmapped from CG trajectories) d) Total electrostatic contribution of binding energy (obtained by subtracting solvation energies from Coulomb energies)

Cp_2149 exhibits a lower energy of approximately 2500 kcal/mol. Compared to full-length Cp183 proteins, Cp_2167 dimers exhibit approximately three times smaller electrostatic repulsion in the range of 1000 kcal/mol. Coulomb electrostatic repulsion is significantly reduced for Cp_2157, Cp_2154, and Cp_2167 trimers with intermediate lengths of the C-ter due to their lower total charges. Among them, the Cp_2164 trimer exhibits the lowest energy of Coulomb interaction within a range of 250 kcal/mol at the shortest COM distances and 100 kcal/mol at the largest COM distances (approximately 8 nm) when the Cp_2 dimer is fully separated from the starting trimer structure. We also calculated the contribution of solvation energy to binding by backmapping coarse-grained MD trajectories to all-atom trajectories, and solvation energies were computed by the MM-GBSA method, as shown in Figure 5.8c. Due to the large protein structure and high computational cost, solvation energies were calculated using the Born model. Our results indicate that solvation energies are more attractive for Cp_2 dimers with larger total charges and larger Coulomb repulsive interactions, such as Cp_2149, Cp_2167, and Cp_2183. However, even though the solvation energy can be quite attractive, the overall contribution of electrostatic obtaining energy (obtained by subtracting solvation energies from Coulomb energies) is unfavorable for all Cp_2 dimers, as shown in Figure 5.8d. Here, Cp149, Cp154, and Cp183 proteins exhibit more repulsive electrostatic binding energy compared to the other three dimers, which correlates with the trends obtained with umbrella sampling. It is worth noting that there might be uncertainty in determining solvation energy due to the use of the Born solvation model and applying it to backmapped all-atom trajectories. However, the general conclusions remain valid.

5.3.2.4. Mapping Residue Contacts During Disassembly Process

To gain insight into the disassembly process of the trimer of Cp_2 dimers, the US trajectories were analyzed by focusing on the time evolution of residue-residue contacts between Cp_2 dimers during the simulated disassembly process. We identified which residues from different Cp_2 dimers were in close contact in the starting crystal structure and monitored the breaking of contacts during dimer separation from the trimer of Cp_2 dimers, as well as tracked how long specific residue-residue contacts between residues were retained during the simulation. Residue pairs with the largest number of counts were considered hotspot interactions, as they represent the residues that dissociate the most slowly and are therefore crucial for overall capsid stability. These contacts are broken last during the dissociation process, indicating their importance. Figures 5.9 and 5.10 show the contact maps of the probability of maintaining contacts between individual residues during Cp_2 dimer dissociation for all six trimers of Cp_2 dimers with different core protein lengths. Although the individual maps differ in the frequency of specific residue-residue contacts, they all share similar features. The first ten residues that maintain contacts for the longest time are mostly uncharged and hydrophobic, located within dimer-dimer interfaces in alpha helices through which respective dimers come into close contact and interact. This finding is consistent with previous studies that concluded that capsid binding is governed by hydrophobic forces contributed by burial of apolar residues placed in subunit-subunit interfaces. Figures 5.11 and 5.12 illustrates the position of residues that maintain a high number of contacts in blue. These residues are mostly located at the end of helical regions through which dimers establish binding, close to the hinge region (141aa to 149aa) that serves as a linker between the assembly domain and the nucleic acid C-ter binding domain. The inset of pictures in Figures 5.11 and 5.12 for Cp_2149 dimers and Cp_2183 dimers, respectively, show some of the most occurring individual residues. In addition to non-charged residues, some

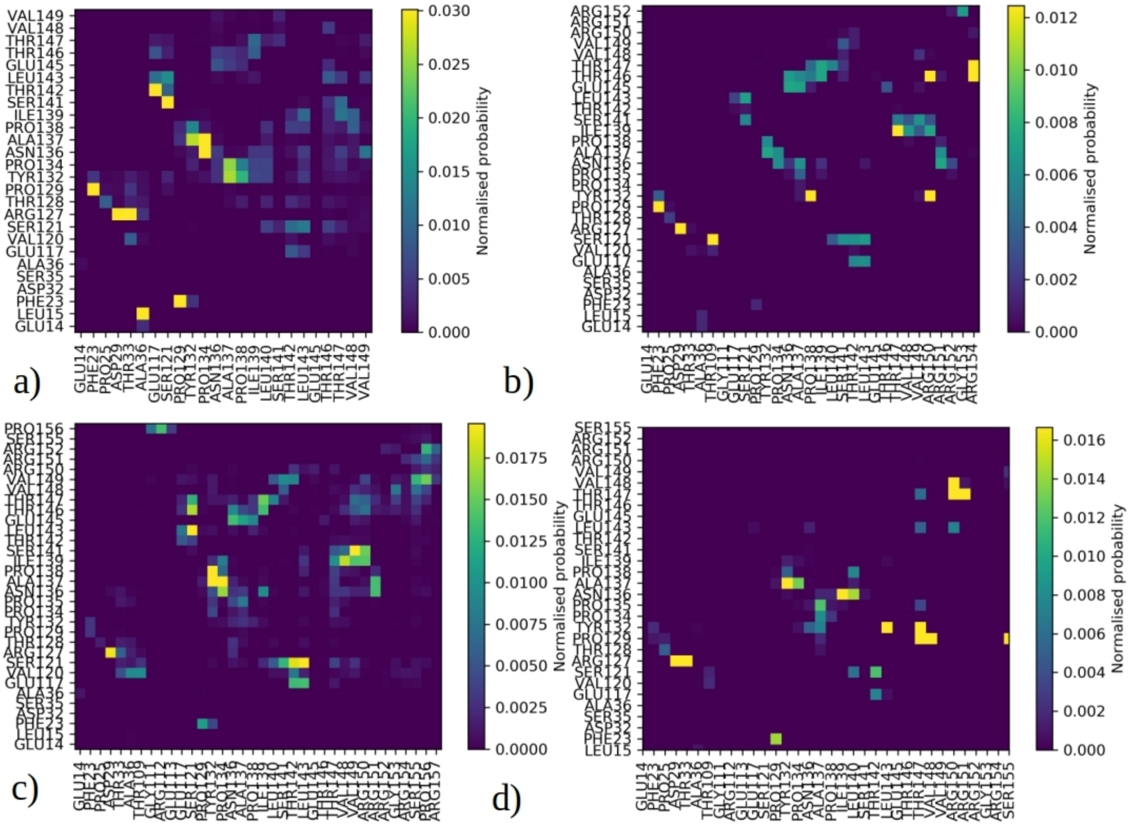


Figure 5.9.: Frequency of residue contacts between Cp_2 dimers during trimer of Cp_2 dissociation for a) Cp_2149 b) Cp_2154 c) Cp_2157 d) Cp_2164 dimers

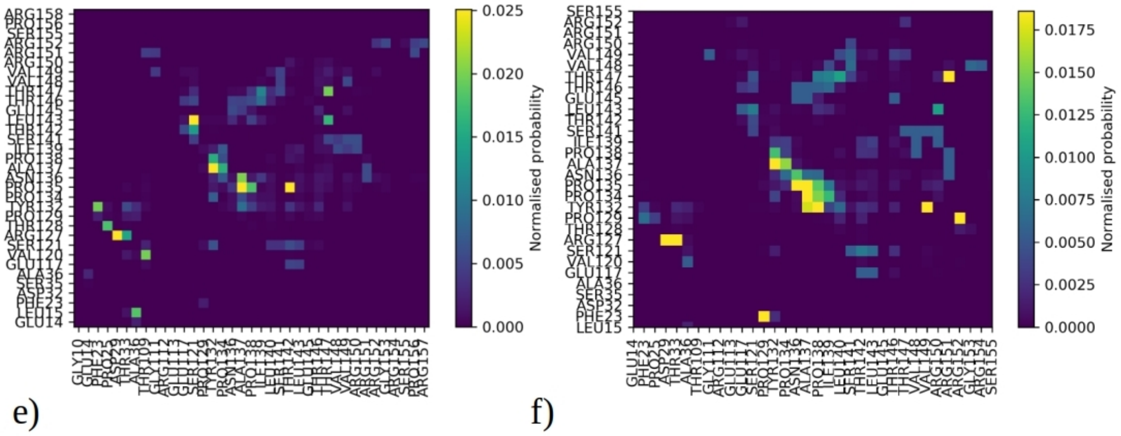


Figure 5.10.: Probability of maintaining residue-residue contacts during Cp_2 dimer dissociation from trimers of dimers for a) $Cp_2$167 and b) $Cp_2$183

charged residues establish contacts, such as positively charged ARG129 and negatively charged ASP29. This residue-residue pair is maintained for a long time for all six trimers

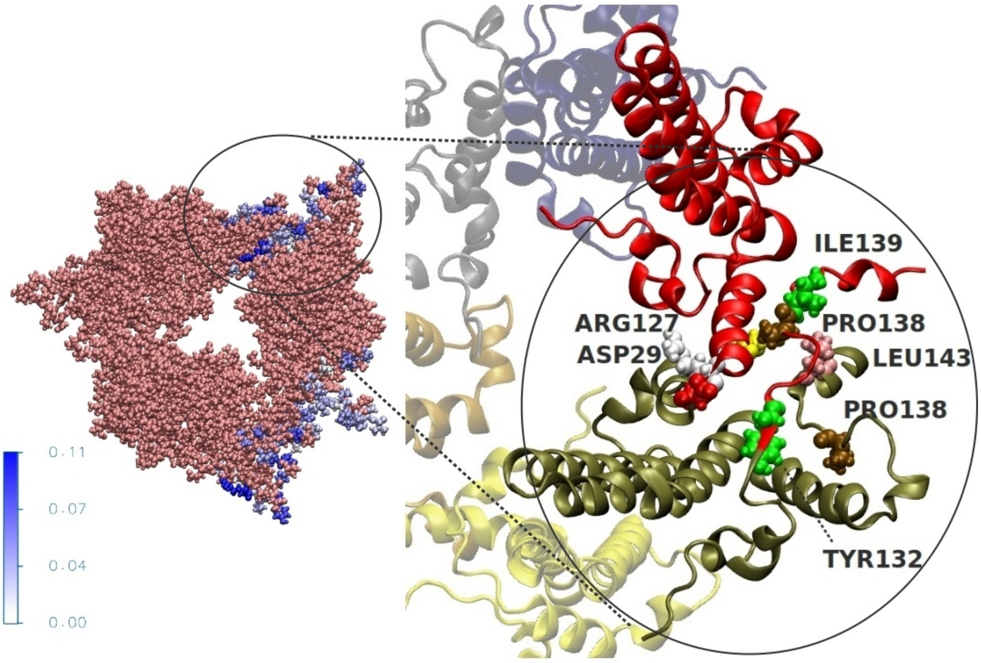


Figure 5.11.: Illustration of residues which maintain contact between dimers with highest probability for $Cp_2$149. On the left side, the whole trimer of dimers is depicted, with residues of high contact probability colored in blue. On the right side, relevant parts are magnified to provide more details.

of Cp_2 dimers, with $Cp_2$183 having the second-highest number of counts. The electrostatic attractive interaction between these two residues further contributes to dimer-dimer binding additional stabilization apart from hydrophobic residues. For Cp_2 dimers containing the nucleic acid binding domain, some of them also establish contacts with residues of other dimers. Residues 149 to 156, for example, might participate in maintaining contacts with residues of other dimers, such as ARG151 and ARG152 with THR147. This probably brings additional stabilization to the dimer-dimer binding energy because arginines are positively charged and the Cp149 core is negatively charged. Several residues have high contacts in common, including SER 121, THR128, TYR132, PRO134, PRO129, ILE139, and THR147. Experimental studies have shown that some of these residues, such as

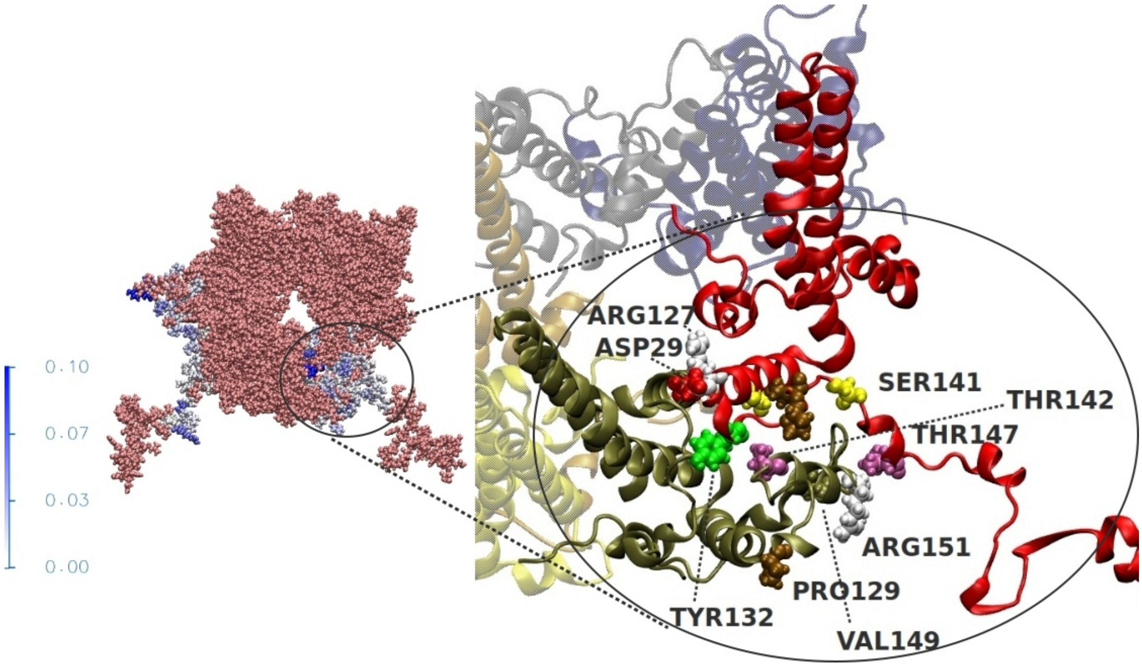


Figure 5.12.: Illustration of residues which maintain contact between dimers with highest probability for Cp_2183. On the left side, the whole trimer of dimers is depicted, with residues of high contact probability colored in blue. On the right side, relevant parts are magnified to provide more details.

TYR132, are important for capsid assembly. For example, the mutation of TYR132 to ALA in Cp_2 dimers causes a deficiency in capsid assembly, although it can undergo co-assembly if mixed with wild-type proteins.[220, 252, 253] Most of the experimental mutation studies were performed on truncated C149 capsids; however, these results suggest that the hotspot residues governing assembly are not significantly influenced by the addition of the nucleic binding region at the C-terminal side.

5.3.2.5. Characterizing Binding Energies in Cp_2 Trimer-DNA Complexes using Docking and MM-PBSA Method

The fundamental biological role of viruses is to encapsulate and safeguard genetic material within capsids. However, if HBV based VLPs are to be utilized as therapeutic delivery agents, it is crucial to incorporate therapeutic nucleic acids into capsids. Therefore, we conducted a study to investigate how the stability of trimers of Cp_2 dimers is affected by the attachment of nucleic acids. The stability of HBV capsids plays a critical role in assembly, disassembly, purification steps, as well as in their ability to deliver and release cargo effectively at the intended site in the organism. The shortest Cp_2149 dimers do not possess a nucleic acid binding region, whereas other dimers such as Cp_2154, C_2157, Cp_2164, Cp_2167, and Cp_2183 are capable of binding nucleic acids to varying degrees, with Cp_2183 possessing the full C-terminus length necessary for efficient nucleic acid binding. We performed MD simulations of Cp_2 trimer-DNA complexes to investigate the affinity of Cp for nucleic acid binding as a function of core protein length. Even though HBV-based VLPs primarily bind RNA molecules, we chose DNA since currently, the SIRAH force field has parameters only for DNA molecules. However, several experimental studies have demonstrated that HBV-based VLPs can bind negatively charged polyelectrolytes, such as RNAs, DNAs, and negatively charged polymers, nonspecifically. First, we docked DNA to the trimer of Cp_2 (Figure 5.14 displays the top ten docked poses of the trimer-DNA complexes), and then we used the top 100 docked complexes with the highest docking scores for 50 ns MD simulations to enable the trimer-DNA structure to relax at a finite temperature and for the

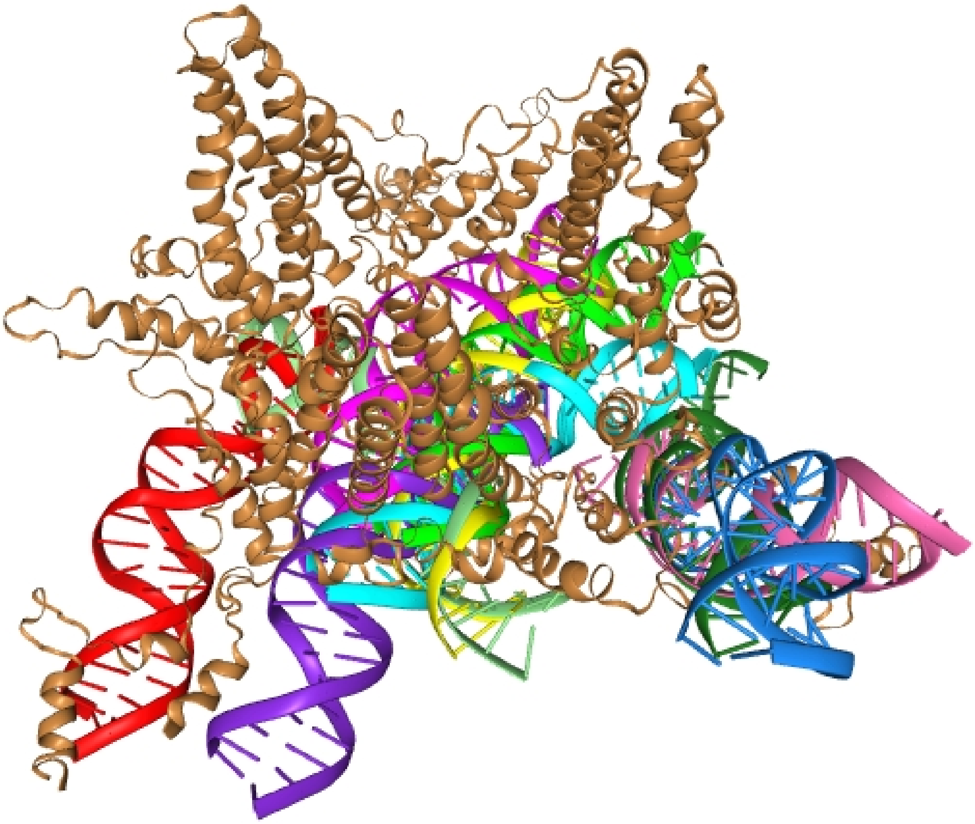


Figure 5.13.: The top ten docked poses of the Cp_2183 trimer-DNA complexes. The proteins are depicted in a new cartoon representation, colored in brown, while the DNA molecule is shown in ten different positions and colored with different colors for illustrative purposes.

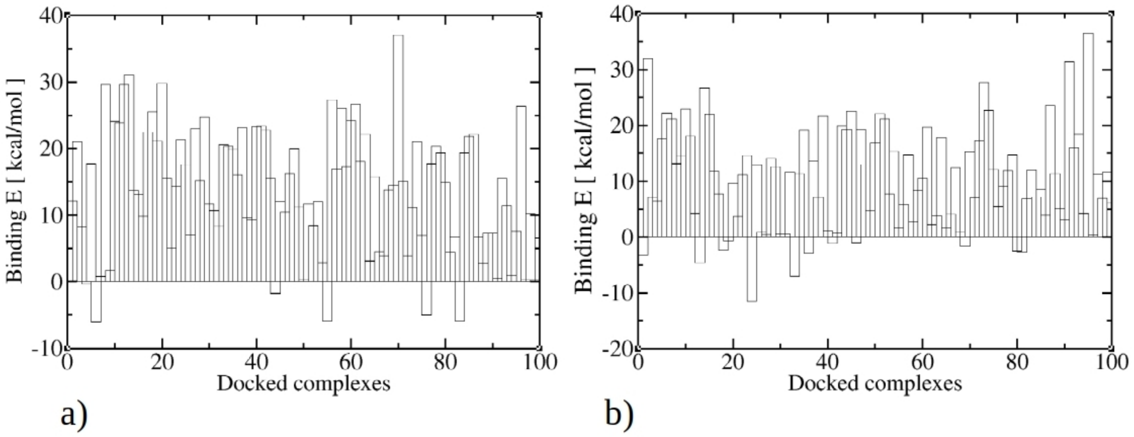


Figure 5.14.: The calculated binding energy of trimer-DNA for 100 MD simulations starting from docked DNA structures for a) Cp_2154 b) Cp_2157 trimers of dimers

DNA to adjust its binding position within the trimer of Cp_2 dimers. We calculated the Cp_2 trimer-DNA binding energy and its dependency on different C-terminus lengths using the MM-GBSA method and a single trajectory approach. Figures 5.14, 5.15a and 5.16 show the calculated binding energies for each of the 100 MD simulations for each C-terminus length. The results show that, with the addition of the arginine-rich C-terminus region, the affinity towards nucleic acids significantly increases. Specifically, Figure 5.14a illustrates that, for Cp_2154 dimers, most docked trimer-DNA complexes exhibit a repulsive binding

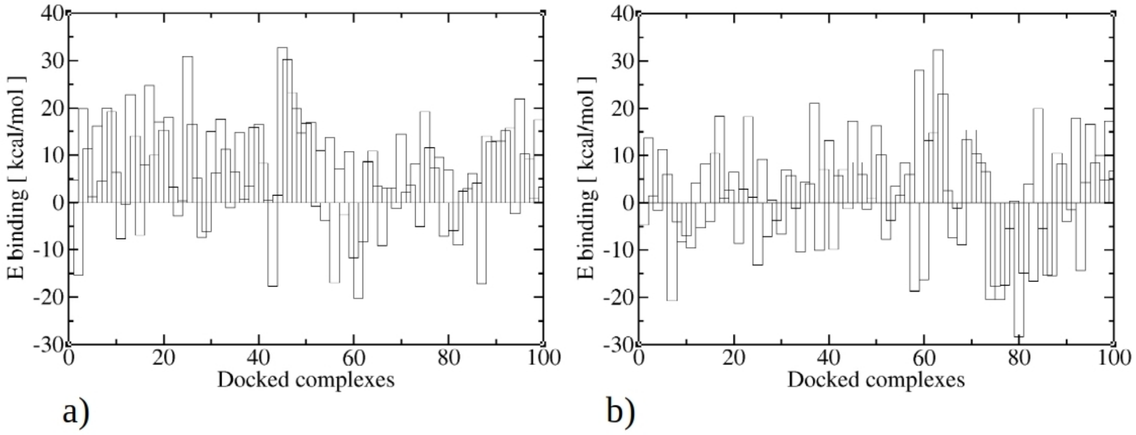


Figure 5.15.: The calculated binding energy of trimer-DNA for 100 MD simulations starting from docked DNA structures for a) Cp_2164 b) Cp_2167 trimers of dimers

energy ranging from +1 to +36 kcal/mol, with around half of the complexes falling within the range of +20 kcal/mol. For Cp_2157 trimer-DNA complexes, compared to Cp_2154, there is a higher proportion of complexes in the range of +10 kcal/mol, with around a quarter of complexes being in the +10 kcal/mol range and another quarter in the +20 kcal/mol range, and only twelve complexes exhibit attractive energy within the range of -3 to -10 kcal/mol (as shown in Figure 5.14b). With the addition of more positively charged residues to the C-terminus domain, an increasing proportion of trimer-DNA complexes exhibit an attractive binding energy. For instance, Figure 5.15a reveals that the Cp_2164 trimer-DNA complex has 24 complexes exhibiting an attractive binding energy ranging from -0.5 to -20 kcal/mol. The trend of a higher proportion of trimer-DNA complexes exhibiting attractive binding energy continued with Cp_2167 trimer-DNA complexes, as 41 complexes exhibited an attractive binding energy within the range of -1 to -28 kcal/mol (Figure 5.15b). For the Cp_2183 trimer-DNA complex, only 31 trimer-DNA complexes exhibited repulsive energy, while the majority of complexes exhibited attractive energy within the range of -1 kcal/mol to -37 kcal/mol, as shown on Figure 5.16a. To further support this trend, Figure Figure 5.16b displays the change of binding energy over simulation time for the most attractive trimer-DNA complex for each of the trimers of Cp_2 dimers with varying C-terminus length. Overall, these findings indicate that the overall binding energy is heavily dependent on the

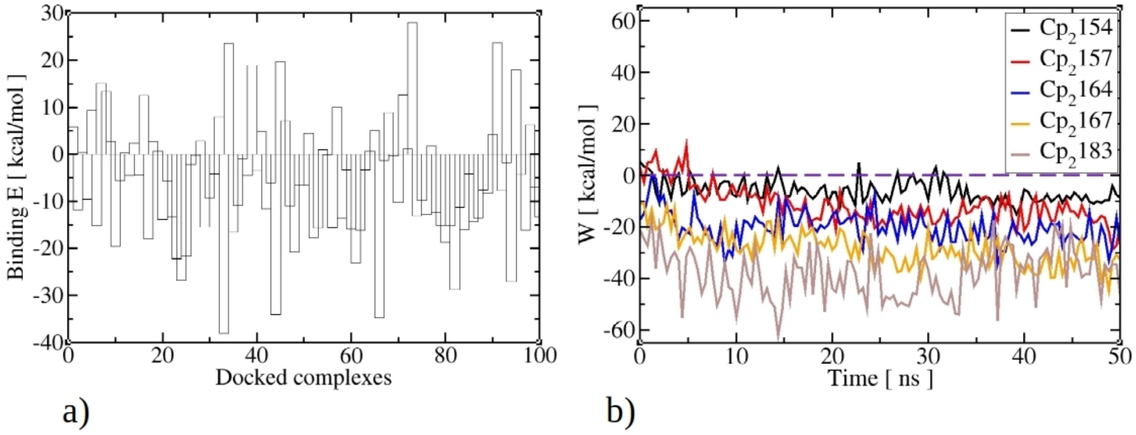


Figure 5.16.: The calculated binding energy of trimer-DNA for 100 MD simulations starting from docked DNA structures for a) Cp_2183 b) Change of binding energy throughout simulation time of the most attractive trimer-DNA complex for each of the trimers of Cp_2 dimers with different length of C-ter

DNA binding pose, and some complexes with shorter C-termini can exhibit more attractive binding energies compared to those with longer C-termini. For instance, the 43rd docked complex of Cp_2164 displays an energy of -18 kcal/mol, while nearly half of the Cp_2183 trimer-DNA complexes have energies within the range of -10 kcal/mol. This highlights the importance of considering the DNA binding pose and the role of the C-terminus in determining the stability of the trimer-DNA complexes.

5.3.2.6. Impact of DNA Binding on Stability of Trimers of Cp_2 Dimers

In this study, we investigated the impact of a bound DNA molecule on the stability of trimers of Cp_2 dimers. To accomplish this, we conducted US simulations in a manner similar like in previous section, but with the DNA attached in the vicinity. We selected five trimer-DNA complexes for each C-terminus length, which exhibited the most attractive binding energy according to the MM-GBSA method, as the starting structures for pulling simulations. The goal of the pulling simulations was to dissociate the Cp_2 dimer from the trimer of dimers and generate starting structures for US simulations. The US simulations aimed to determine the effect of DNA on the energy required to separate the Cp_2 dimers from trimers of dimers. Figures 5.17, 5.18 and 5.19 illustrate the calculated free energy change of Cp_2 dissociation from the corresponding trimers in the presence of DNA. To evaluate the impact of DNA on the stability of trimers of Cp_2 dimers, we performed an analysis specifically for Cp_2 dimers that contained the nucleic acid binding regions Cp_2154, Cp_2157, Cp_2164, Cp_2167, and Cp_2183. As a point of comparison, we included the free energy of Cp_2 dissociation for pure Cp_2 trimers without DNA attached in the vicinity, which was previously presented in a previous Section 5.3.2.2 and is depicted by the dashed line. The results show that DNA stabilizes Cp_2 trimers, resulting in a more attractive (negative) free energy of dissociation. Furthermore, the DNA binding pose also affects the binding energy between Cp_2 dimers in the trimer. As a result, the average binding energy of each of the five different DNA binding poses is depicted in Figure 5.19b. It should be noted that our simulations which produced more attractive binding energies were mostly due to direct contact between the separated dimer and the DNA molecule. While we selected the dimer that was farthest from the DNA for separation, in some cases direct contact could not be avoided. This suggests that dimers have a higher propensity to dissociate in regions with lower nucleic acid density, which is relevant since nucleic acids are not evenly distributed around each dimer in real capsids. The degree of stabilization varies based on the length of the binding region. For example, in the case of trimers of Cp_2154

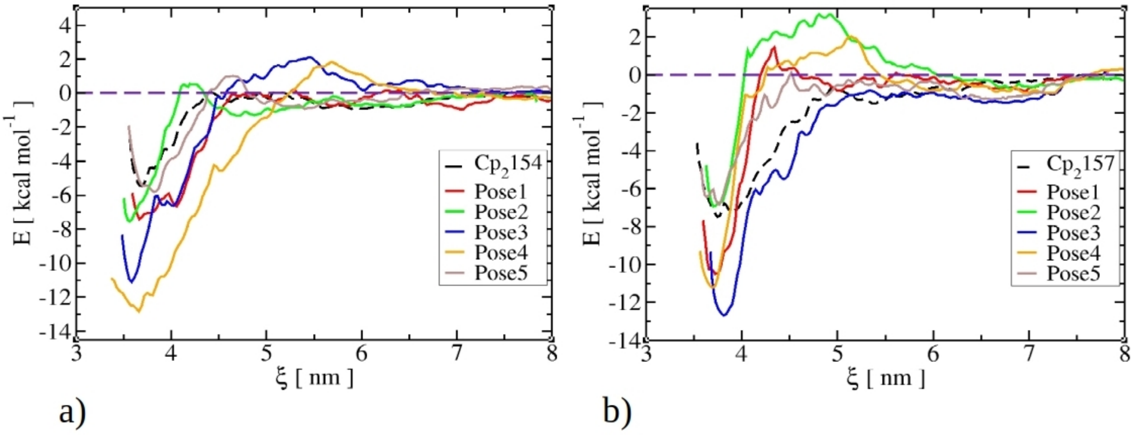


Figure 5.17.: Free energy of Cp_2 dimer binding obtained by US when Cp_2 dimer is dissociated from trimer in presence of DNA. The calculated free energy of separating Cp_2 from Cp_2 trimer by using umbrella sampling technique for a) Cp_2154 b) Cp_2157 where DNA is attached to the trimer at different position.

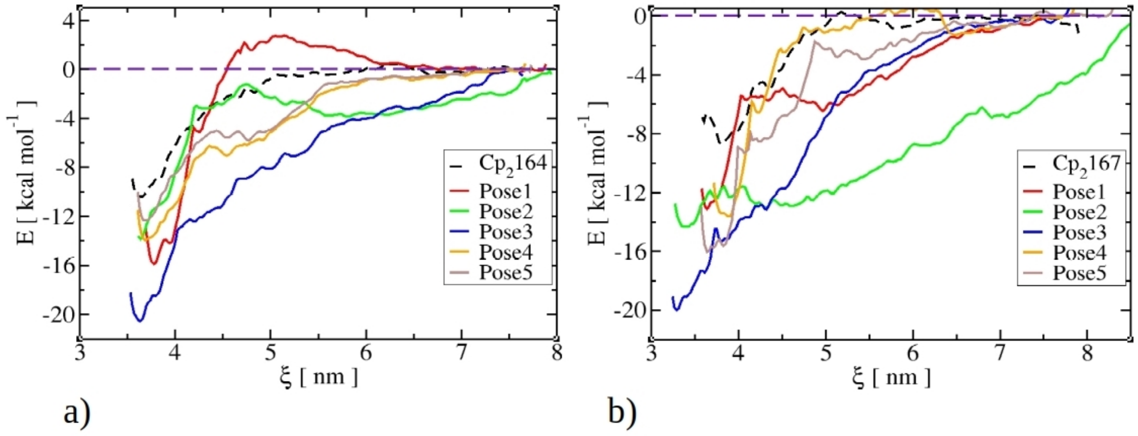


Figure 5.18.: Free energy of Cp_2 dimer binding obtained by US when Cp_2 dimer is dissociated from trimer in presence of DNA. The calculated free energy of separating Cp_2 from Cp_2 trimer by using US simulationa for a) $Cp_2$164 b) $Cp_2$167 where DNA is attached to the trimer at different position.

dimers, the presence of DNA results in a shift of the binding energy to values ranging from -6 kcal/mol to -13 kcal/mol, depending on the DNA binding pose. The average binding energy over five US simulations with different DNA binding poses is -9 kcal/mol, which is 1.65 times more attractive than the energy in the absence of DNA. Similarly, for trimers of $Cp_2$157 dimers, the average binding energy in the presence of DNA is approximately -9.2 kcal/mol. Here, the stabilization factor is 1.22, as the binding energy in the absence of DNA is slightly more attractive compared to Cp154 proteins. Notably, the average binding energy of $Cp_2$154 and $Cp_2$157 dimers in the presence of DNA is 1.4 kcal/mol more attractive than that of trimers of $Cp_2$149, which lack nucleic acid binding regions. This is indirectly supported by capsid disassembly experiments conducted by Valentic et al.,[63] where Cp154 and Cp157 capsids exhibited slightly lower dimer yields compared to Cp149 capsids, suggesting that Cp154 and Cp157 capsids are more stable. The stabilization effect is particularly strong for trimers with longer nucleic acid binding regions, such as $Cp_2$164, $Cp_2$167, and $C_2$2183. These proteins have more positively charged residues at the C-terminus, which allows them to establish strong interactions with DNA and stabilize

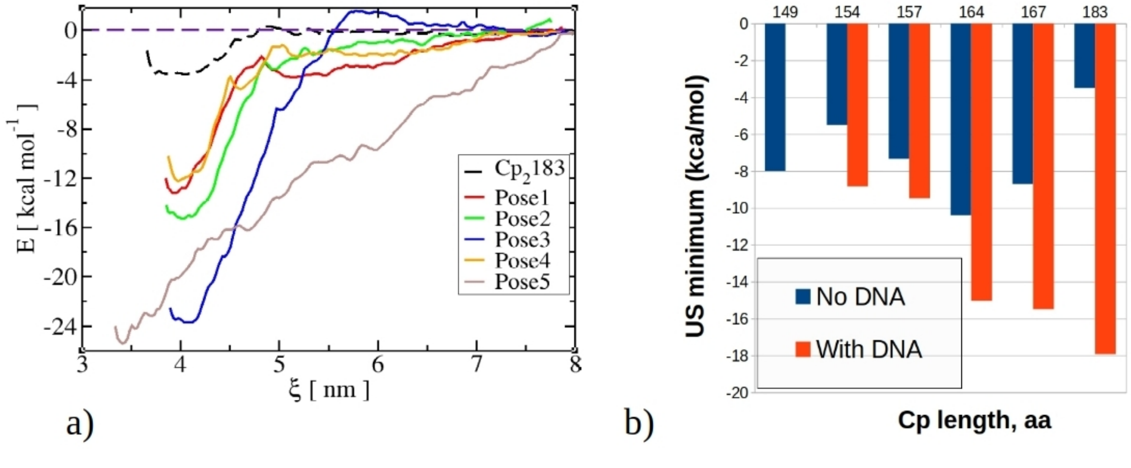


Figure 5.19.: a) Free energy of $Cp_2$183 dimer binding obtained by US when $Cp_2$183 dimer is dissociated from trimer in presence of DNA b) Summary of US simulations without (in blue) and with (in red) bonded DNA to the trimer. Results for Cp_2 trimer-DNA complex are average of five different US simulations, in which DNA was attached to the trimer at different positions.

the system. For example, Cp_2164 and Cp_2167 dimers exhibit binding energies of -10.4 kcal/mol and -8.7 kcal/mol in the absence of DNA, while the average binding energy in the presence of DNA is -15 kcal/mol and -16 kcal/mol, respectively. As a result, the trimers are stabilized by a factor of 1.49 and 1.84, respectively. The strongest stabilization effect is observed for trimers of Cp_2183 dimers, where the attractive energy is 4.9 times stronger compared to the case when DNA is not present. The efficiency of VLP (dis)assembly depends on the complex interplay between core proteins and nucleic acids. The Cp_2183 trimer displays moderately attractive dimer-dimer interactions and a strong affinity for negatively charged DNA or RNA, resulting in the formation of an exceptionally stable capsid. Consequently, the processing of these VLPs can be challenging. Similar effects may also be observed in the case of Cp_2164 and Cp_2167 capsids. Several experimental studies have found that Cp183 capsids, which encapsulate nucleic acids, exhibit higher stability compared to truncated and empty Cp149 capsids. Valentic et al. reported that Cp164 and Cp167 capsids had lower disassembly yields compared to all other capsids, indicating that they were the most stable, while Cp183 dimers showed yields within the range of Cp154 and Cp157 capsids. Furthermore, Ma et al. demonstrated through a theoretical model of capsid thermodynamics that Cp164 capsids were more stable than Cp183 capsids. However, our calculations showed that Cp183 dimers had the highest binding energy in the presence of DNA compared to other dimers, contrary to previous findings suggesting that Cp164 should be more stable. It should be noted that our simulations did not capture the full effects of the capsid environment due to the absence of the entire capsid structure. Sominskaya et al. measured that capsids with different core protein lengths encapsulate varying amounts of nucleic acids,[65] which was also supported by theoretical calculations based on a capsid thermodynamic model, which showed that capsids are most stable with an optimal amount of nucleic acids, while smaller or larger amounts can destabilize the capsid. Taken together, these studies highlight the complex interplay between capsid proteins and nucleic acids in determining the stability of viral capsids. While some capsids have been shown to be more stable than others, the optimal amount of nucleic acids and the effects of the capsid environment on stability are still being explored. Furthermore, our calculations suggest that capsids with shorter nucleic acid binding regions may be more optimal for efficient VLP production. These capsids can still bind some nucleic acids without being over-stabilized, which may impede the ease of assembly and release of nucleic acid cargo during VLP processing and handling.

5.4. Conclusion

In this chapter, a molecular dynamics study was conducted on beta-lactoglobulin proteins and fragments of hepatitis B capsids, i.e. trimers of Cp_2 dimers as representative structure of whole capsid. The study employed the accelerated molecular dynamics technique (umbrella sampling) to investigate the free energy of beta-lactoglobulin dimerization and the free energy of dissociation of trimers of Cp_2 dimers of hepatitis B core proteins. To increase computational efficiency, a coarse-grained SIRAH force field was utilized, which retains the positions of backbone atoms while preserving high structural details of protein structure. The results of the study showed a delicate dependence of the monomer-dimer equilibrium of beta-lactoglobulin on the pH and ionic strength of the solution. In the study of the core proteins of hepatitis B capsids, different variants of the protein were examined, beginning with naturally occurring Cp proteins that had the full length of the nucleic acid binding region, and then variants with gradually truncated the residues at the C-terminus until they reached a variant that lacked a nucleic acid binding region. In addition, the study also investigated the stabilization of trimers of Cp_2 dimers through the attachment of DNA to core proteins. This serves to attenuate the electrostatic repulsion among positively charged C-termini, which are responsible for binding nucleic acids. The results of the umbrella sampling simulations showed that the free energies were highly dependent on the length of the C-terminus side of the core protein, which is further influenced by the attachment of DNA molecules serving as an additional stabilizing factor for the trimers of dimers. Compared to the Cp149 proteins lacking a nucleic acid binding domain, Cp183 with the full length of the C-terminus and Cp154 which contain only a small part of the C-terminus, exhibit less attractive binding energy in the absence of nucleic acids, indicating destabilization. In contrast, proteins with intermediate lengths of C-terminus, such as Cp157, Cp164, and Cp167, were found to be stabilized. This is because the total charge of the proteins decreases as one goes from the fully truncated Cp149 proteins with a charge of -7 to the Cp183 proteins with a charge of +8. When DNA was attached to trimers of dimers, the binding energy became more attractive with the increase in the length of the nucleic acid binding region. This effect was most dramatic for Cp183, where the binding energy became five times more attractive, transitioning from slightly attractive to overly attractive energy. The study highlights the importance of the delicate balance between protein-protein and protein-nucleic acid interactions in the stability of VLPs. The results indicate that structures that are too stable can be difficult to process experimentally, while structures with insufficient binding energy may not form easily.

6. Summary

Proteins are fundamental macromolecules with a vast array of functions and characteristics, making them suitable for various biopharmaceutical applications, such as targeted drug delivery and vaccines. Virus-like particles are a highly promising protein-based therapeutic with strong immunogenicity, therapeutic nucleic acid encapsulation, and potential use in gene therapy. In this thesis, we modeled protein-protein interactions using coarse-grained protein structures and different levels of theory, ranging from computationally affordable continuum models to nearly atomistic, high-resolution coarse-grained models combined with accelerated molecular dynamics. Our calculations of second osmotic virial coefficients provided insight into the effective interaction between protein solutions, a critical factor in ensuring the overall stability of protein solutions and preventing unwanted aggregation during biotechnological processing. Additionally, we explored the dimerization and stability of proteins aggregates using accelerated molecular dynamics techniques.

Chapter 3 introduced a newly developed xDLVO-CG model that predicts protein solution stability and salt-induced dependencies. Specifically, xDLVO-CG is used for calculating second osmotic virial coefficients of proteins in solution. The model uses a shape-based coarse-grained representation to account for anisotropic protein-protein interactions, which reduces or eliminates the need for fitting experimental data. The model has been validated with experimental data for several benchmark proteins and can be applied to a wide range of proteins. The model has also been used to compute osmotic second virial coefficients for hepatitis B virus core protein dimers, demonstrating its potential for predicting protein stability under different conditions.

Chapter 4 presented the xDLVO-CGhybr model, which is an improvement over the xDLVO-CG model presented in the previous chapter. The xDLVO-CGhybr model uses a hybrid approach to calculate the electrostatic part of the potential of mean force, combining Poisson-Boltzmann theory applied on all-atom structures of proteins at short separation and Debye-Hückel theory applied on coarse-grained structures at larger interprotein separations for accurate calculation. The model shows improved predictions of second osmotic virial coefficient values compared to other models available in literature.

xDLVO-CGhybr and its predecessor xDLVO-CG are already valuable tools for calculating second osmotic virial coefficients and studying protein interactions, but there is still much potential for further developments to enhance their accuracy and versatility. One possible direction is to improve the methods for fast and easy calculations of Hamaker constants by Lifshitz theory. Additionally, machine learning approaches could be applied to calculate electrostatic energies more quickly and accurately. In cases where protein conformational flexibility is important, potential terms from the model could be used in Monte Carlo simulations to sample the most relevant protein-protein orientations. This approach can help to better capture the dynamic nature of protein interactions and improve the accuracy of the predictions. Overall, these potential developments have the potential to enhance the accuracy and versatility of xDLVO-CGhybr and xDLVO-CG models, which in turn can broaden their impact and expand their scope of applicability in pharmaceutical and biotechnological applications.

In Chapter 5, molecular dynamics simulations investigated the free energy of beta-lactoglobulin dimerization and the dissociation of trimers of dimers of hepatitis B core proteins. The study revealed the delicate dependence of beta-lactoglobulin's monomer-dimer equilibrium on pH and ionic strength, and the complex interplay between protein-protein and protein-nucleic acid interactions in virus-like capsid stability. The study showed that free energies of trimers of dimers dissociation were highly dependent on the length of the C-terminus, with intermediate proteins exhibiting the most stability. Attaching DNA to trimers of dimers increased binding energy, especially for full-length Cp183 proteins, increasing five times compared to the absence of nucleic acids. The simulations highlighted the importance of balancing protein-protein and protein-nucleic acid interactions in capsid stability. These findings have implications for the development of biopharmaceuticals and gene therapies, where overly stable or insufficiently stable structures may be problematic. The study's findings underscore the potential of molecular dynamics simulations in studying protein interactions and the behavior of virus-like capsids in different conditions, providing valuable insights into the development of biopharmaceuticals and gene therapies.

Future studies could investigate the effect of different factors such as temperature or solvent type on protein interactions, contributing to a more comprehensive understanding of the behaviour of core HBV proteins in solutions. Future work can involve simulating whole HBV capsids that encapsulate different amounts of nucleic acids. The combination of these simulations with experiments can significantly enhance the understanding of capsid stabilities and aid experimentalists in designing appropriate systems. Additionally, computational studies could explore the introduction of mutations or epitope sequences, as well as covalent modifications of core proteins, to identify optimal virus-like particle candidates. Finally, studying the selectivity of capsids to target specific cells in the body for releasing therapeutic cargo, and understanding the triggering of capsid disassembly in vivo conditions to release the cargo are both essential for successful therapeutic applications.

Protein research is crucial for developing novel therapies and materials for various applications. In addition, Proteins and peptides can also serve as building blocks for new materials, energy sources, and devices. However, efficient technological processes must be developed to realise the full potential of proteins. A crucial step in the development of any protein product is to ensure that protein solutions are stable and capable of long-term storage. The world today is changing at a rapid pace, with a rapidly increasing population, leading to new potential threats to human health, as exemplified by the recent COVID pandemic. Consequently, the need for novel therapeutic approaches that can effectively address these challenges becomes crucial. Virus-like particles hold great promise as future vaccines and gene therapies. However, to turn this promise into reality, a collaborative effort between experimental and computational researchers is essential.

7. Zusammenfassung

Proteine sind grundlegende Makromoleküle mit einer Vielzahl von Funktionen und Eigenschaften, die sie für verschiedene biopharmazeutische Anwendungen, wie die gezielte Abgabe von Medikamenten und Impfstoffen, geeignet machen. Virusähnliche Partikel sind ein vielversprechendes Therapeutikum auf Proteinbasis mit starker Immunogenität, therapeutischer Nukleinsäureverkapselung und potenziellem Einsatz in der Gentherapie. In dieser Arbeit modellierten wir Protein-Protein-Wechselwirkungen unter Verwendung grobkörniger Proteinstrukturen und verschiedener Theorieniveaus, die von rechnerisch erschwinglichen Kontinuumsmodellen bis zu nahezu atomistischen, hochauflösenden grobkörnigen Modellen in Kombination mit beschleunigter Molekulardynamik reichen. Unsere Berechnungen der zweiten osmotischen Virialkoeffizienten gaben Aufschluss über die effektive Wechselwirkung zwischen Proteinlösungen, ein entscheidender Faktor für die Gewährleistung der Gesamtstabilität von Proteinlösungen und die Verhinderung unerwünschter Aggregation während der biotechnologischen Verarbeitung. Darüber hinaus untersuchten wir die Dimerisierung und Stabilität von Proteinaggregaten mit Hilfe beschleunigter Molekulardynamiktechniken.

In Kapitel 3 wurde ein neu entwickeltes xDLVO-CG-Modell vorgestellt, das die Stabilität von Proteinlösungen und salzinduzierte Abhängigkeiten vorhersagt. Konkret wird xDLVO-CG zur Berechnung der zweiten osmotischen Virialkoeffizienten von Proteinen in Lösung verwendet. Das Modell verwendet eine formbasierte grobkörnige Darstellung, um anisotrope Protein-Protein-Wechselwirkungen zu berücksichtigen, wodurch die Notwendigkeit der Anpassung experimenteller Daten reduziert oder beseitigt wird. Das Modell wurde mit experimentellen Daten für mehrere Referenzproteine validiert und kann auf eine breite Palette von Proteinen angewendet werden. Das Modell wurde auch zur Berechnung der zweiten osmotischen Virialkoeffizienten für Hepatitis-B-Virus-Kernproteindimere verwendet, was sein Potenzial für die Vorhersage der Proteinstabilität unter verschiedenen Bedingungen unter Beweis stellt.

In Kapitel 4 wurde das xDLVO-CGhybr-Modell vorgestellt, das eine Verbesserung gegenüber dem im vorherigen Kapitel vorgestellten xDLVO-CG-Modell darstellt. Das xDLVO-CGhybr-Modell verwendet einen hybriden Ansatz zur Berechnung des elektrostatischen Teils des Interaktionspotentials, indem es die Poisson-Boltzmann-Theorie, die auf atomistische Strukturen von Proteinen bei kurzen Abständen angewandt wird, und die Debye-Hückel-Theorie, die auf grobkörnige Strukturen bei größeren Abständen zwischen den Proteinen angewandt wird, zur genauen Berechnung kombiniert. Das Modell zeigt eine verbesserte Vorhersage der Werte des zweiten osmotischen Virialkoeffizienten im Vergleich zu anderen in der Literatur verfügbaren Modellen.

xDLVO-CGhybr und sein Vorgänger xDLVO-CG sind bereits wertvolle Werkzeuge für die Berechnung von zweiten osmotischen Virialkoeffizienten und die Untersuchung von Proteinwechselwirkungen, aber es gibt noch viel Potenzial für weitere Entwicklungen, um ihre Genauigkeit und Vielseitigkeit zu verbessern. Eine mögliche Richtung ist die Verbesserung der Methoden zur schnellen und einfachen Berechnung der Hamaker-Konstanten mit Hilfe der Lifshitz-Theorie. Außerdem könnten Ansätze des maschinellen Lernens angewandt werden, um elektrostatische Energien schneller und genauer zu berechnen. In Fällen, in denen die Flexibilität der Proteinkonformation von Bedeutung ist, könnten Potenzialterme aus dem Modell in Monte-Carlo-Simulationen verwendet werden, um die wichtigsten Protein-Protein-Orientierungen zu ermitteln. Dieser Ansatz kann dazu beitragen, die dynamische Natur der Proteinwechselwirkungen besser zu erfassen und die Genauigkeit der Vorhersagen zu verbessern. Insgesamt haben diese möglichen Entwicklungen das Potenzial, die Genauigkeit und Vielseitigkeit von xDLVO-CGhybr- und xDLVO-CG-Modellen zu verbessern, was wiederum ihre Wirkung und ihren Anwendungsbereich in pharmazeutischen und biotechnologischen Anwendungen erweitern kann.

In Kapitel 5 wurden die freie Energie der Beta-Lactoglobulin-Dimerisierung und die Dissoziation von Trimeren von Dimeren der Hepatitis-B-Kernproteine anhand Molekulardynamiksimulationen untersucht. Die Studie zeigte die empfindliche Abhängigkeit des Monomer-Dimer-Gleichgewichts von Beta-Lactoglobulin vom pH-Wert und der Ionenstärke sowie das komplexe Zusammenspiel von Protein-Protein- und Protein-Nukleinsäure-Wechselwirkungen bei der virusartigen Kapsidstabilität. Die Studie zeigte, dass die Differenz der freien Energie der Dissoziation von Trimeren von Dimeren stark von der Länge des C-Terminus abhängen, wobei Zwischenproteine die größte Stabilität aufweisen. Das Anhängen von DNA an Trimere von Dimeren erhöhte die Bindungsenergie, insbesondere für Cp183-Proteine in voller Länge, und zwar um das Fünffache im Vergleich zur Abwesenheit von Nukleinsäuren. Die Simulationen verdeutlichen, wie wichtig das Gleichgewicht zwischen Protein-Protein- und Protein-Nukleinsäure-Interaktionen für die Stabilität des Kapsids ist. Diese Ergebnisse haben Auswirkungen auf die Entwicklung von Biopharmazeutika und Gentherapien, bei denen zu stabile oder unzureichend stabile Strukturen problematisch sein können. Die Ergebnisse der Studie unterstreichen das Potenzial von Molekulardynamiksimulationen bei der Untersuchung von Proteininteraktionen und des Verhaltens von virusähnlichen Kapsiden unter verschiedenen Bedingungen und liefern wertvolle Erkenntnisse für die Entwicklung von Biopharmazeutika und Gentherapien.

Künftige Studien könnten die Auswirkungen verschiedener Faktoren wie Temperatur oder Lösungsmitteltyp auf die Proteininteraktionen untersuchen und so zu einem umfassenderen Verständnis des Verhaltens von HBV-Kernproteinen in Lösungen beitragen. Zukünftige Arbeiten könnten die Simulation ganzer HBV-Kapsiden beinhalten, die unterschiedliche Mengen an Nukleinsäuren einkapseln. Die Kombination dieser Simulationen mit Experimenten kann das Verständnis der Kapsidstabilität erheblich verbessern und Experimentatoren bei der Entwicklung geeigneter Systeme unterstützen. Darüber hinaus könnten Computerstudien die Einführung von Mutationen oder Epitopsequenzen sowie kovalente Modifikationen von Kernproteinen untersuchen, um optimale Kandidaten für virusähnliche Partikel zu identifizieren. Schließlich sind die Untersuchung der Selektivität von Kapsiden für die Freisetzung von therapeutischen Molekülen auf bestimmte Zellen im Körper und das Verständnis der Auslösung des Kapsidabbaus unter in vivo-Bedingungen für die Freisetzung der Molekülen von wesentlicher Bedeutung für erfolgreiche therapeutische Anwendungen.

Die Proteinforschung ist von entscheidender Bedeutung für die Entwicklung neuer Therapien und Materialien für verschiedene Anwendungen. Das Verständnis der Proteinmechanismen ist angesichts der wachsenden Bevölkerung und neu auftretender Krankheiten besonders wichtig. Proteine und Peptide können nicht nur als therapeutische Wirkstoffe, sondern auch als Bausteine für neue Materialien, Energiequellen und Geräte dienen. Um das volle Potenzial von Proteinen auszuschöpfen, müssen jedoch effiziente technologische Verfahren entwickelt werden. Um diese Ziele zu erreichen, sind weitere Fortschritte aus theoretischer und experimenteller Sicht erforderlich. Die in dieser Arbeit vorgestellten Arbeiten sowie mögliche zukünftige Entwicklungen sind ein bemerkenswerter Schritt in Richtung dieser Ziele. Durch die Erforschung der Eigenschaften und des Verhaltens von Proteinen können die Forscher ihr volles Potenzial zum Nutzen der Gesellschaft ausschöpfen.

8. Acknowledgments

I would like to express my heartfelt gratitude to Professor Wolfgang Wenzel for providing me with the opportunity to carry out this research project in his group. I am also deeply grateful to Professor Wenzel and Dr. Mariana Kozlowska for their exceptional scientific guidance and unwavering support throughout my PhD studies. My sincere thanks also go to Professor Jürgen Hubbuch for his valuable input and advice as my second supervisor.

I would like to extend my gratitude to Dr. Maximilian Uttinger and Vanessa Lautenbach from the group of Professor Wolfgang Peukert at the University of Erlangen for their fruitful collaboration and support.

I am indebted to my fellow group colleagues for their professional help, valuable insights, and companionship, especially David, Helmy, Anna, Niklas, and Meysam. Their friendship and support have been invaluable.

Last but not least, I would like to express my deepest appreciation to my family and friends for their unconditional love, support, and encouragement throughout this journey. I am also grateful to all the people I have met and worked with along the way, whose contributions have made this experience truly rewarding.

9. List of Publications

[1] 1. Srdjan Pusara, Peyman Yamin, Wolfgang Wenzel, Marjan Krstić, and Mariana Kozlowska. A coarse-grained xDLVO model for colloidal protein–protein interactions. 23(22):12780–12794, 2021.

Author contributions: conceptualization: P. Y., W. W. and M. K.; methodology: S. P., W. W., M. Kr. and M. K.; software, M. Kr., S. P.; validation: S. P. and M. K.; formal analysis: S. P. and M. K.; investigation: S. P.; resources: W. W.; data curation: S. P.; writing – original draft preparation: S. P. and M. K.; writing – review and editing: S. P., W. W., M. Kr. and M. K.; visualization: S. P. and M. K.; supervision: W. W. and M. K.; project administration: M. K.; funding acquisition: W. W.

2. M. J. Uttinger, C. S. Hundschell, V. Lautenbach, S. Pusara, S. Bäther, T. R. Heyn, J. K. Keppler, W. Wenzel, J. Walter, M. Kozlowska, A. M. Wagemans, and W. Peukert. Determination of specific and non-specific protein–protein interactions for beta-lactoglobulin by analytical ultracentrifugation and membrane osmometry experiments. 18(35):6739–6756, 2022.

3. Srdjan Pusara, Wolfgang Wenzel, and Mariana Kozlowska. Accurate calculation of second osmotic virial coefficients of proteins using mixed poisson-boltzmann and extended dlvo theory. Molecular Systems Design & Engineering, 2023

4. Srdjan Pusara, Wolfgang Wenzel and Mariana Kozlowska. Exploring Hepatitis B Capsid Fragment Stability with Coarse-Grained Molecular Dynamics and Free Energy Calculations , In preparation

5. Srdjan Pusara, Angela Valentic, Wolfgang Wenzel, Mariana Kozlowska, Jürgen Hubbuch. Changes in structure, dispersity and phase behavior of proteins: The virus-like particles in the presence of nucleic acids(Book Chapter), Springer, Submitted

[1] The author contributions for other articles are detailed in the respective journal articles.

Appendix

A. List of Parameters in Equations

- A_H: Hamaker constant
- M_W: Molar mass of protein
- R_p: Protein radius
- Z: Protein charge
- T: Absolute temperature
- ϵ_r: Relative permitivity
- κ: Debye length
- σ: Water layer around protein
- I: Ionic strength
- R_3: Salt radius
- ρ_3: Salt density
- ϵ_{ij}: Epsilon parameter of Lennard-Jones potential between two coarse-grained beads
- σ_{ij}: Sigma parameter of Lennard-Jones potential between two coarse-grained beads

B. Coarse-Grained Structures and Their Charge Distribution of Proteins Used in xDLVO-CG and xDLVO-CGhybr Calculations

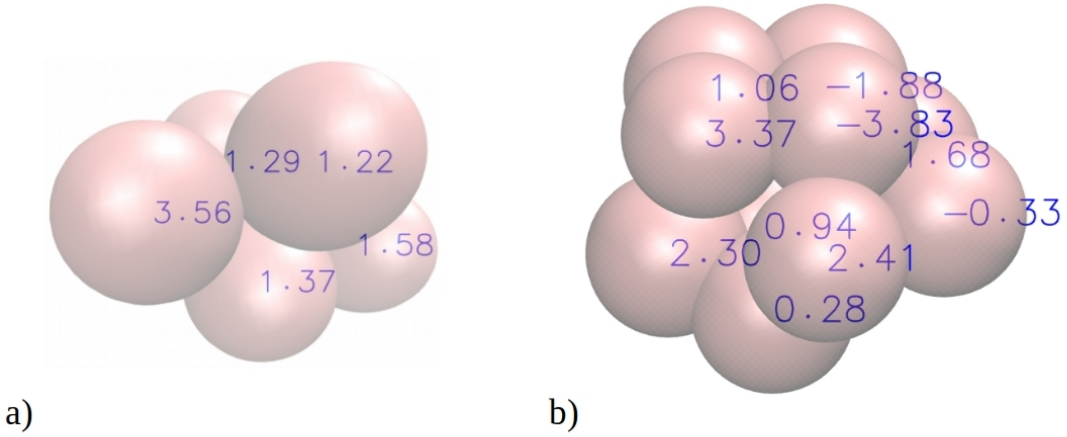


Figure B.1.: Charge distribution over coarse grained beads for a) LYZ at pH 7 and b) Subs at pH 5.5

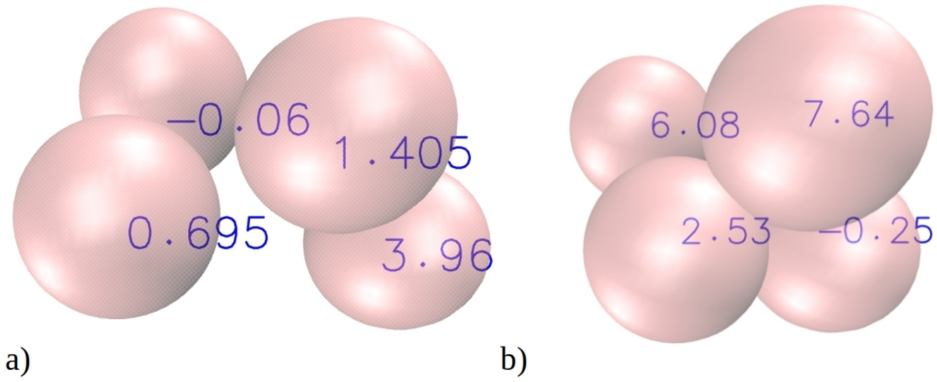


Figure B.2.: Charge distribution over CG beads for a) BPTI at pH 4.9 and b) RbnA at pH 3

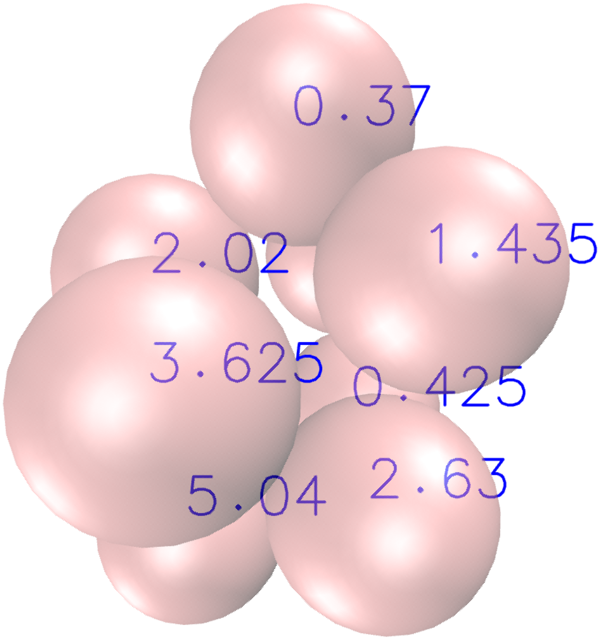


Figure B.3.: Charge distribution over coarse grained beads for ChymA at pH 3

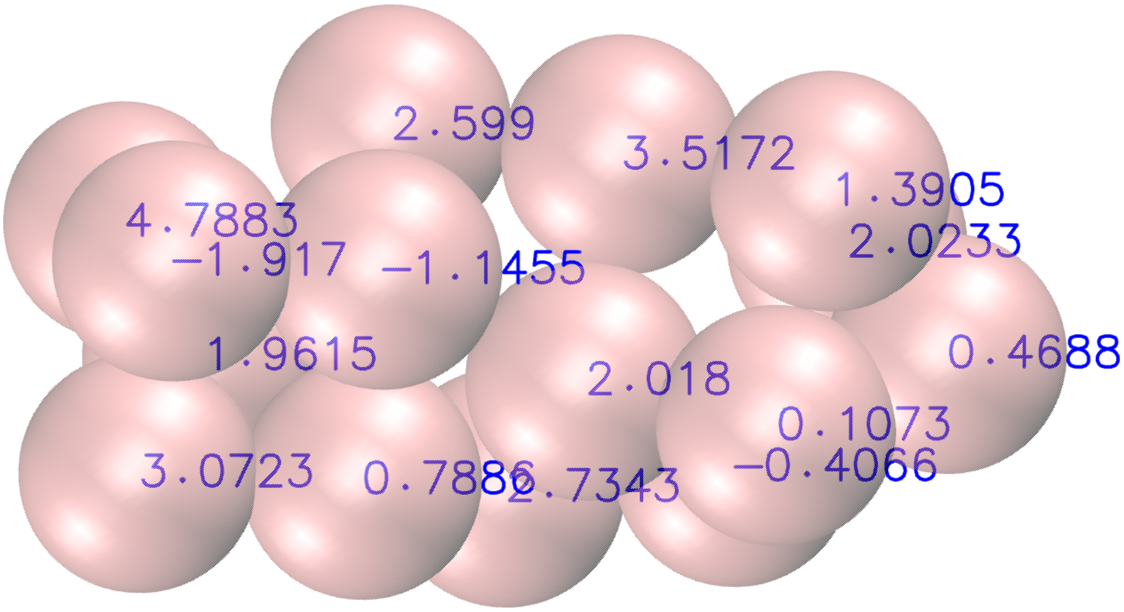


Figure B.4.: Charge distribution over coarse grained beads for ConcA at pH 4

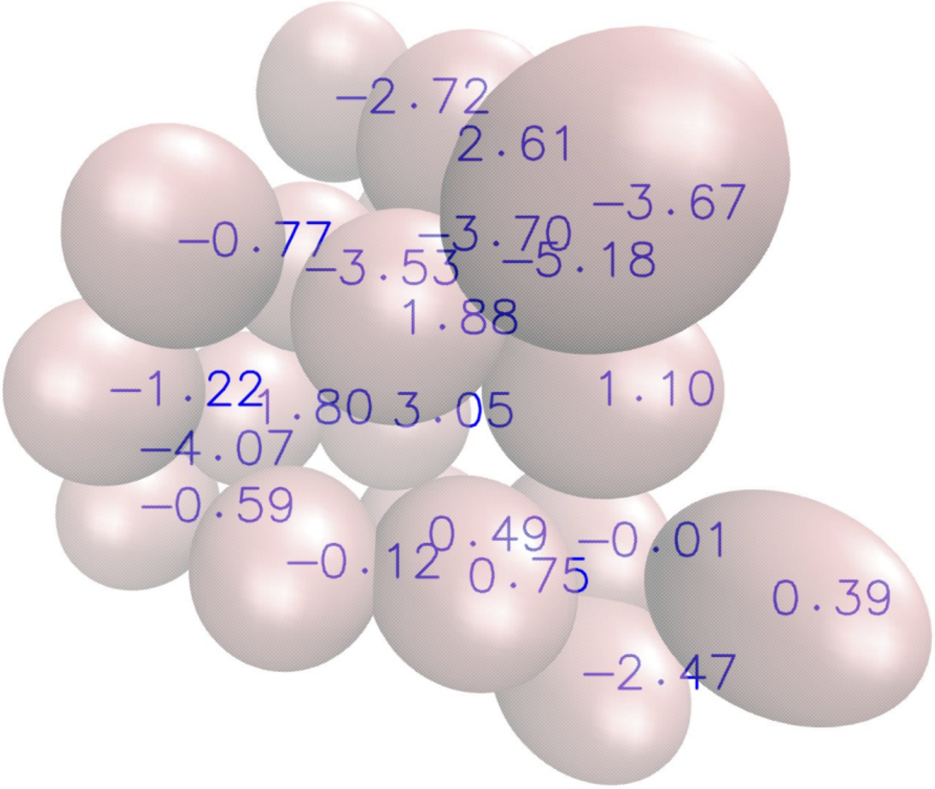


Figure B.5.: Charge distribution over coarse grained beads for BSA at pH 7.4

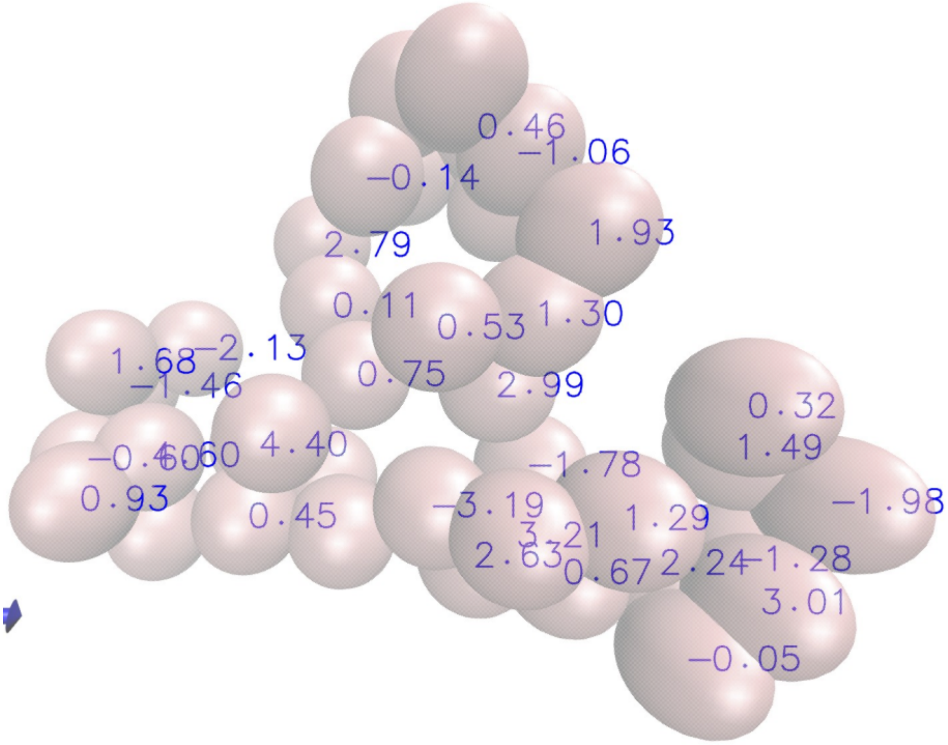


Figure B.6.: Charge distribution over coarse grained beads for IgG1 at pH 6.5

C. Calculated Second Osmotic Virial Coefficients of Proteins

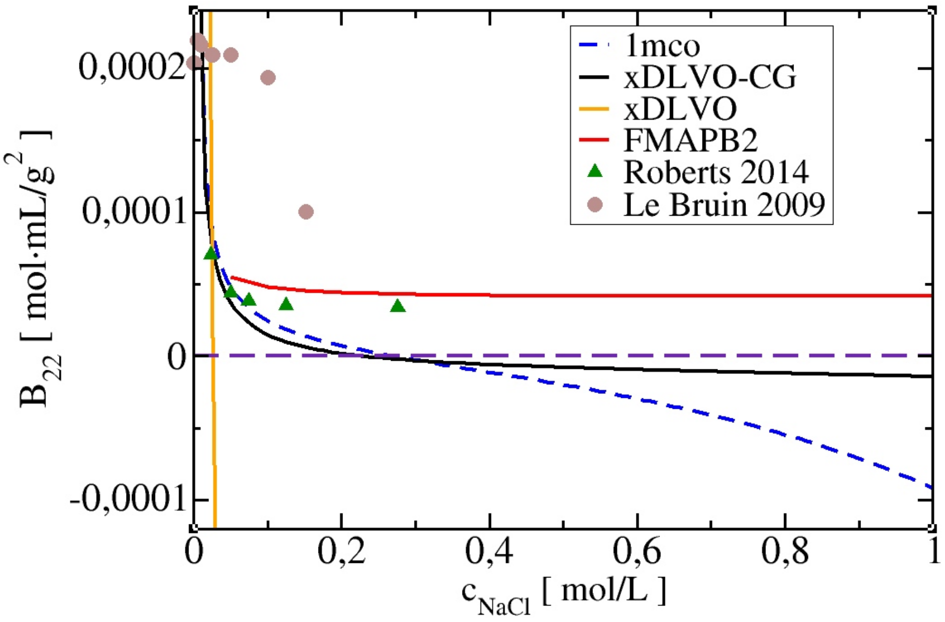


Figure C.7.: Second osmotic virial coefficients for IgG1 at pH 6.5 as a function of NaCl concentration calculated using xDLVO-CG, xDLVO and FMAPB2 models in comparison with experimental data. The experimental data, labeled as 'Roberts 2014' and 'Le Bruin' were taken from [187, 200].

D. Overlap of Histograms from Umbrella Sampling Simulations of Cp Proteins

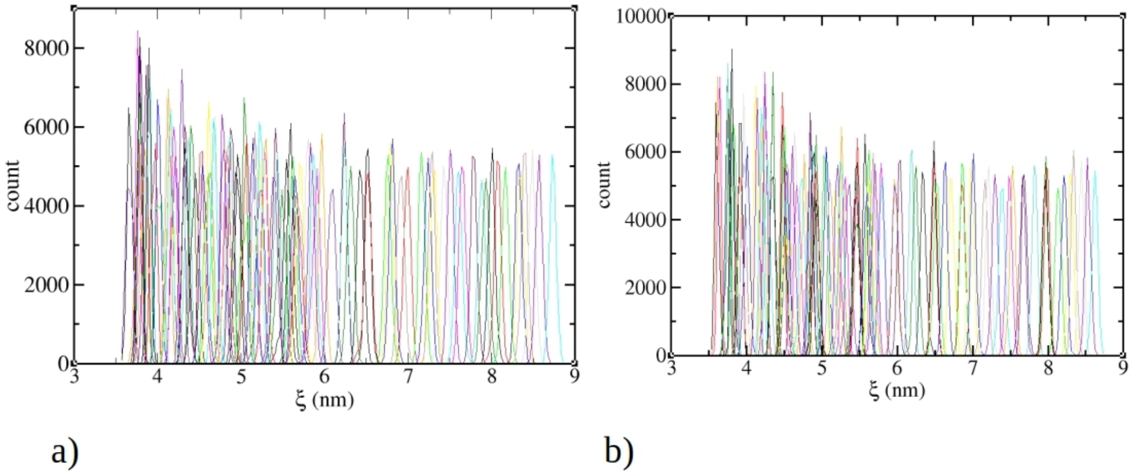


Figure D.8.: Overlap of histograms from umbrella sampling simulations for disassembly of trimers of a) Cp_2149 and b) Cp_2154 dimers

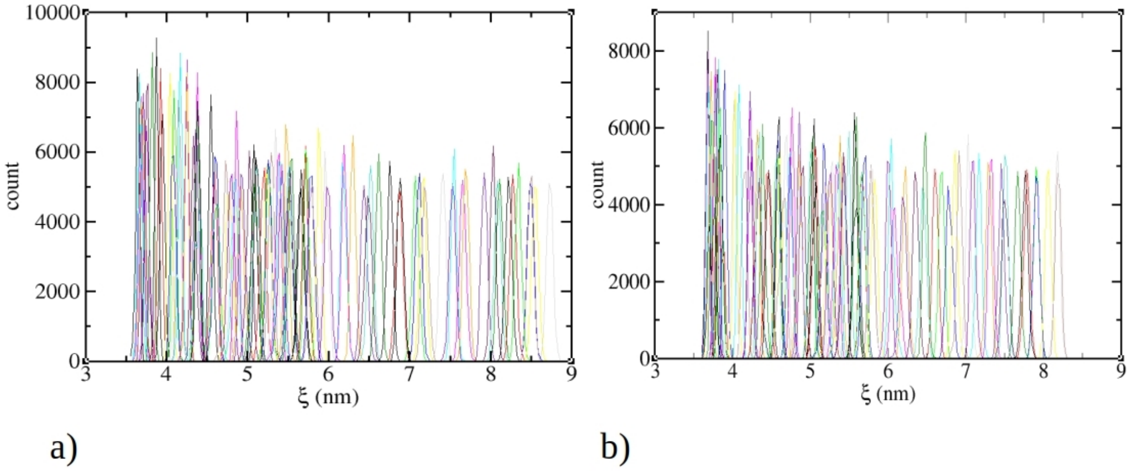


Figure D.9.: Overlap of histograms from umbrella sampling simulations for disassembly of trimers of a) Cp_2157 and b) Cp_2164 dimers

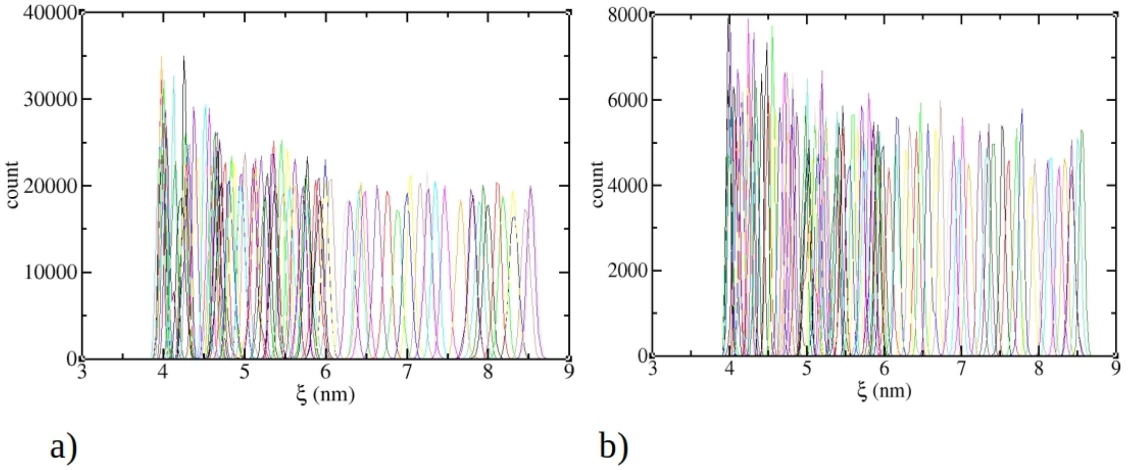


Figure D.10.: Overlap of histograms from umbrella sampling simulations for disassembly of trimers of a) Cp_2167 and b) Cp_2183 dimers

Bibliography

[1] Haian Fu, Romesh R. Subramanian, and Shane C. Masters. 14-3-3 proteins: Structure, function, and regulation. *Annual Review of Pharmacology and Toxicology*, 40(1):617–647, 2000.

[2] Hong Jiang, Xiaoyu Zhang, Xiao Chen, Pornpun Aramsangtienchai, Zhen Tong, and Hening Lin. Protein lipidation: Occurrence, mechanisms, biological functions, and enabling technologies. *Chemical Reviews*, 118(3):919–988, 2018.

[3] Thomas L. Moore, Laura Rodriguez-Lorenzo, Vera Hirsch, Sandor Balog, Dominic Urban, Corinne Jud, Barbara Rothen-Rutishauser, Marco Lattuada, and Alke Petri-Fink. Nanoparticle colloidal stability in cell culture media and impact on cellular interactions. *Chemical Society Reviews*, 44(17):6287–6305, 2015.

[4] Natalie Rakel, Katharina Christin Bauer, Lara Galm, and Juergen Hubbuch. From osmotic second virial coefficient (B22) to phase behavior of a monoclonal antibody. *Biotechnology Progress*, 31(2):438–451, 2015.

[5] Shannon A. Marshall, Greg A. Lazar, Arthur J. Chirino, and John R. Desjarlais. Rational design and engineering of therapeutic proteins. *Drug Discovery Today*, 8(5):212–221, 2003.

[6] Dimiter S. Dimitrov. *Therapeutic Proteins*, pages 1–26. Humana Press, Totowa, NJ, 2012.

[7] Mary E. M. Cromwell, Eric Hilario, and Fred Jacobson. Protein aggregation and bioprocessing. *The AAPS Journal*, 8(3):E572–E579, 2006.

[8] Pascal Braun and Anne-Claude Gingras. History of protein–protein interactions: From egg-white to complex networks. *PROTEOMICS*, 12(10):1478–1498, 2012.

[9] Eric Dickinson. Proteins at interfaces and in emulsions: Stability, rheology, and interactions. *J. Chem. Soc., Faraday Trans.*, 94(12):1657–1669, 1998.

[10] Taco Nicolai. Gelation of food protein-protein mixtures. *Advances in colloid and interface science*, 270:147–164, 2019.

[11] S. D. Durbin and G. Feher. Protein crystallization. *Annual Review of Physical Chemistry*, 47(1):171–204, 1996.

[12] Ren-Bin Zhou, Hui-Ling Cao, Chen-Yan Zhang, and Da-Chuan Yin. A review on recent advances for nucleants and nucleation in protein crystallization. *CrystEngComm*, 19(8):1143–1155, 2017.

[13] Jan A. Miernyk and Jay J. Thelen. Biochemical approaches for discovering protein–protein interactions. *The Plant Journal*, 53(4):597–609, 2008.

[14] Irene M.A. Nooren and Janet M. Thornton. Diversity of protein–protein interactions. *The EMBO Journal*, 22(14):3486–3492, 2003.

[15] Tord Berggård, Sara Linse, and Peter James. Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS*, 7(16):2833–2842, 2007.

[16] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part i. experimental techniques and databases. *PLOS Computational Biology*, 3(3):1–8, 03 2007.

[17] Lukasz Salwinski and David Eisenberg. Computational methods of analysis of protein–protein interactions. *Current Opinion in Structural Biology*, 13(3):377–382, 2003.

[18] Stewart A. Adcock and J. Andrew McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5):1589–1615, 2006.

[19] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids.* 2017.

[20] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[21] Pratyush Tiwary and Axel van de Walle. *A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics*, pages 195–221. Springer International Publishing, Cham, 2016.

[22] Michael Feig and Charles L Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current Opinion in Structural Biology*, 14(2):217–224, 2004.

[23] Nathan A. Baker. Poisson–boltzmann methods for biomolecular electrostatics. In *Methods in Enzymology*, volume 383, pages 94–118. Elsevier, 2004.

[24] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

[25] Paweł Grochowski and Joanna Trylska. Continuum molecular electrostatics, salt effects, and counterion binding—a review of the poisson–boltzmann theory and its modifications. *Biopolymers*, 89(2):93–113, 2008.

[26] Gene Lamm. *The Poisson–Boltzmann Equation*, chapter 4, pages 147–365. John Wiley Sons, Ltd, 2003.

[27] Alexey V. Onufriev and David A. Case. Generalized born implicit solvent models for biomolecules. *Annual Review of Biophysics*, 48(1):275–296, 2019.

[28] Donald Bashford and David A. Case. Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51(1):129–152, 2000.

[29] Gary A. Huber and J. Andrew McCammon. Brownian dynamics simulations of biological molecules. *Trends in Chemistry*, 1(8):727–738, 2019.

[30] Overbreek J. Tm G. Verwey, E J W. Theory of the Stability of Lyophobic Colloids. *Elsevier*, page 6, 1948.

[31] J. Lyklema, H.P. van Leeuwen, and M. Minor. Dlvo-theory, a dynamic reinterpretation. *Advances in Colloid and Interface Science*, 83(1):33–69, 1999.

[32] B.W. Ninham. On progress in forces since the dlvo theory. *Advances in Colloid and Interface Science*, 83(1):1–17, 1999.

[33] M. Boström, D. R. M. Williams, and B. W. Ninham. Specific ion effects: Why dlvo theory fails for biology and colloid systems. *Phys. Rev. Lett.*, 87(16):168103, 2001.

[34] Ciying Qian, Xinlin Liu, Qin Xu, Zhiping Wang, Jie Chen, Tingting Li, Qingbing Zheng, Hai Yu, Ying Gu, Shaowei Li, and Ningshao Xia. Recent progress on the versatility of virus-like particles. *Vaccines*, 8(1):139, 2020.

[35] Marcus J. Rohovie, Maya Nagasawa, and James R. Swartz. Virus-like particles: Next-generation nanoparticles for targeted therapeutic delivery: ROHOVIE et al. *Bioengineering & Translational Medicine*, 2(1):43–57, 2017.

[36] Andris Zeltins. Construction and characterization of virus-like particles: A review. *Molecular Biotechnology*, 53(1):92–107, 2013.

[37] Robert L Garcea and Lutz Gissmann. Virus-like particles as vaccines and vessels for the delivery of small molecules. *Current Opinion in Biotechnology*, 15(6):513–517, 2004.

[38] Florence Boisgérault, Gabriel Morón, and Claude Leclerc. Virus-like particles: a new family of delivery systems. *Expert Review of Vaccines*, 1(1):101–109, 2002.

[39] Stanislaw J. Kaczmarczyk, Kalavathy Sitaraman, Howard A. Young, Stephen H. Hughes, and Deb K. Chatterjee. Protein delivery using engineered virus-like particles. *Proceedings of the National Academy of Sciences*, 108(41):16998–17003, 2011.

[40] Brett D. Hill, Andrew Zak, Eshita Khera, and Fei Wen. Engineering virus-like particles for antigen and drug delivery. *Current Protein & Peptide Science*, 19(1), 2017.

[41] John C. Burnett, John J. Rossi, and Katrin Tiemann. Current progress of sirna/shrna therapeutics in clinical trials. *Biotechnology Journal*, 6(9):1130–1146, 2011.

[42] Shao-Hua Chen and Getu Zhaori. Potential clinical applications of sirna technique: benefits and limitations. *European Journal of Clinical Investigation*, 41(2):221–232, 2011.

[43] Rahul R. Nikam and Kiran R. Gore. Journey of sirna: Clinical developments and targeted delivery. *Nucleic Acid Therapeutics*, 28(4):209–224, 2018.

[44] Vivek K. Sharma, Raman K. Sharma, and Sunil K. Singh. Antisense oligonucleotides: modifications and clinical trials. *Med. Chem. Commun.*, 5(10):1454–1471, 2014.

[45] P. Pumpens and E. Grens. The true story and advantages of the famous hepatitis b virus core particles: Outlook 2016. *Molecular Biology*, 50(4):489–509, 2016.

[46] Christian Trépo, Henry L Y Chan, and Anna Lok. Hepatitis b virus infection. *The Lancet*, 384(9959):2053–2063, 2014.

[47] J. Zachary Porterfield, Mary Savari Dhason, Daniel D. Loeb, Michael Nassal, Stephen J. Stray, and Adam Zlotnick. Full-length hepatitis b virus core protein packages viral and heterologous RNA with similarly high levels of cooperativity. *Journal of Virology*, 84(14):7174–7184, 2010.

[48] Balasubramanian Venkatakrishnan and Adam Zlotnick. The structural biology of hepatitis b virus: Form and function. *Annual Review of Virology*, 3(1):429–451, 2016.

[49] Adam Zlotnick, Balasubramanian Venkatakrishnan, Zhenning Tan, Eric Lewellyn, William Turner, and Samson Francis. Core protein: A pleiotropic keystone in the hbv lifecycle. *Antiviral research*, 121:82–93, 2015.

[50] A. Zlotnick, N. Cheng, S. J. Stahl, J. F. Conway, A. C. Steven, and P. T. Wingfield. Localization of the c terminus of the assembly domain of hepatitis b virus capsid protein: Implications for morphogenesis and organization of encapsidated RNA. *Proceedings of the National Academy of Sciences*, 94(18):9556–9561, 1997.

[51] M Seifer and D N Standring. A protease-sensitive hinge linking the two domains of the hepatitis b virus core protein is exposed on the viral capsid surface. *Journal of Virology*, 68(9):5548–5555, 1994.

[52] M Nassal. The arginine-rich domain of the hepatitis b virus core protein is required for pregenome encapsidation and productive viral positive-strand DNA synthesis but not for virus assembly. *Journal of Virology*, 66(7):4107–4116, 1992.

[53] Ahmed Diab, Adrien Foca, Fabien Zoulim, David Durantel, and Ourania Andrisani. The diverse functions of the hepatitis b core/capsid protein (HBc) in the viral life cycle: Implications for the development of HBc-targeting antivirals. *Antiviral Research*, 149:211–220, 2018.

[54] Zachary D. Harms, Lisa Selzer, Adam Zlotnick, and Stephen C. Jacobson. Monitoring assembly of virus capsids with nanofluidic devices. *ACS Nano*, 9(9):9087–9096, 2015.

[55] Charlotte Uetrecht, Norman R. Watts, Stephen J. Stahl, Paul T. Wingfield, Alasdair C. Steven, and Albert J. R. Heck. Subunit exchange rates in hepatitis b virus capsids are geometry- and temperature-dependent. *Physical Chemistry Chemical Physics*, 12(41):13368, 2010.

[56] Willem K. Kegel and Paul van der Schoot. Competing hydrophobic and screened-coulomb interactions in hepatitis b virus capsid assembly. *Biophysical Journal*, 86(6):3905–3913, 2004.

[57] Adam Zlotnick. Are weak protein–protein interactions the general rule in capsid assembly? *Virology*, 315(2):269–274, 2003.

[58] Corinne A. Lutomski, Nicholas A. Lyktey, Zhongchao Zhao, Elizabeth E. Pierson, Adam Zlotnick, and Martin F. Jarrold. Hepatitis b virus capsid completion occurs through error correction. *Journal of the American Chemical Society*, 139(46):16932–16938, 2017.

[59] Pablo Ceres and Adam Zlotnick. Weak proteinprotein interactions are sufficient to drive assembly of hepatitis b virus capsids. *Biochemistry*, 41(39):11525–11531, 2002.

[60] Yao Zhang, Yongdong Liu, Bingyang Zhang, Shuang Yin, Xiunan Li, Dawei Zhao, Weiying Wang, Jingxiu Bi, and Zhiguo Su. In vitro preparation of uniform and nucleic acid free hepatitis b core particles through an optimized disassembly-purification-reassembly process. *Protein Expression and Purification*, 178:105747, 2021.

[61] Nils Hillebrandt, Philipp Vormittag, Annabelle Dietrich, Christina H. Wegner, and Jürgen Hubbuch. Process development for cross-flow diafiltration-based vlp disassembly: A novel high-throughput screening approach. *Biotechnology and Bioengineering*, 118(10):3926–3940, 2021.

[62] Ivars Petrovskis, Ilva Lieknina, Andris Dislers, Juris Jansons, Janis Bogans, Inara Akopjana, Jelena Zakova, and Irina Sominskaya. Production of the hbc protein from different hbv genotypes in e. coli. use of reassociated hbc vlps for packaging of ss- and dsrna. *Microorganisms*, 9(2), 2021.

[63] Angela Valentic, Jakob Müller, and Jürgen Hubbuch. Effects of different lengths of a nucleic acid binding region and bound nucleic acids on the phase behavior and purification process of HBcAg virus-like particles. *Frontiers in Bioengineering and Biotechnology*, 10:929243, 2022.

[64] J. Zachary Porterfield and Adam Zlotnick. A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance. *Virology*, 407(2):281–288, 2010.

[65] Irina Sominskaya, Dace Skrastina, Ivars Petrovskis, Andris Dishlers, Ieva Berza, Maria Mihailova, Juris Jansons, Inara Akopjana, Irina Stahovska, Dzidra Dreilina, Velta Ose, and Paul Pumpens. A VLP library of c-terminally truncated hepatitis b core proteins: Correlation of RNA encapsidation with a th1/th2 switch in the immune responses of mice. *PLoS ONE*, 8(9):e75938, 2013.

[66] Margaret Newman, Pong Kian Chua, Fan-Mei Tang, Pei-Yi Su, and Chiaho Shih. Testing an electrostatic interaction hypothesis of hepatitis b virus capsid stability by using an in vitro capsid disassembly/reassembly system. *Journal of Virology*, 83(20):10616–10626, 2009.

[67] Shuyu Liu, Jian He, Chiaho Shih, Kunpeng Li, Aguang Dai, Z. Hong Zhou, and Jingqiang Zhang. Structural comparisons of hepatitis b core antigen particles with different c-terminal lengths. *Virus Research*, 149(2):241–244, 2010.

[68] B.L Neal, D Asthagiri, O.D Velev, A.M Lenhoff, and E.W Kaler. Why is the osmotic second virial coefficient related to protein crystallization? *Journal of Crystal Growth*, 196(2-4):377–387, 1999.

[69] B.L. Neal, D. Asthagiri, and A.M. Lenhoff. Molecular Origins of Osmotic Second Virial Coefficients of Proteins. *Biophysical Journal*, 75(5):2469–2477, 1998.

[70] A. George and W. W. Wilson. Predicting protein crystallization from a dilute solution property. *Acta Crystallographica Section D Biological Crystallography*, 50(4):361–365, 1994.

[71] Donald A. McQuarrie. *Statistical Mechanics*. Harper's chemistry series. Harper Collins, New York, 1976.

[72] Dominik Horinek. *DLVO Theory*, pages 343–346. Springer New York, New York, NY, 2014.

[73] Marcel Herhut, Christoph Brandenbusch, and Gabriele Sadowski. Inclusion of mPRISM potential for polymer-induced protein interactions enables modeling of second osmotic virial coefficients in aqueous polymer-salt solutions. *Biotechnology Journal*, 11(1):146–154, 2016.

[74] Marcel Herhut, Christoph Brandenbusch, and Gabriele Sadowski. Non-monotonic course of protein solubility in aqueous polymer-salt solutions can be modeled using the sol-mxDLVO model. *Biotechnology Journal*, 11(2):282–289, 2016.

[75] R A Curtis, J M Prausnitz, and H W Blanch. Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes. *BIOTECHNOLOGY AND BIOENGINEERING*, 57(1):11, 1998.

[76] H.C. Hamaker. The London—van der Waals attraction between spherical particles. *Physica*, 4(10):1058–1072, 1937.

[77] E. Hückel. Zur Theorie der Elektrolyte. In *Ergebnisse der exakten naturwissenschaften*, volume 3, pages 199–276. Springer Berlin Heidelberg, Berlin, Heidelberg, 1924.

[78] TG Mason. Osmotically driven shape-dependent colloidal separations. *Physical Review E*, 66(6):060402, 2002.

[79] Sho Asakura and Fumio Oosawa. Interaction between particles suspended in solutions of macromolecules. *Journal of Polymer Science*, 33(126):183–192, 1958.

[80] Franz Hofmeister. Zur Lehre von der Wirkung der Salze. *Archiv für experimentelle Pathologie und Pharmakologie*, 24(4):247–260, 1888.

[81] Feng Dong, Brett Olsen, and Nathan A. Baker. Computational methods for biomolecular electrostatics. In *Biophysical Tools for Biologists, Volume One: In Vitro Techniques*, volume 84 of *Methods in Cell Biology*, pages 843–870. Academic Press, 2008.

[82] Nathan A Baker. Improving implicit solvent simulations: a poisson-centric view. *Current Opinion in Structural Biology*, 15(2):137–143, 2005.

[83] Pengyu Ren, Jaehun Chun, Dennis G. Thomas, Michael J. Schnieders, Marcelo Marucho, Jiajing Zhang, and Nathan A. Baker. Biomolecular electrostatics and solvation: a computational perspective. *Quarterly Reviews of Biophysics*, 45(4):427–491, 2012.

[84] Nathan A. Baker. *Biomolecular Applications of Poisson–Boltzmann Methods*, chapter 5, pages 349–379. John Wiley Sons, Ltd, 2005.

[85] Stephen T. Kottmann. Harmonic averaging of smooth permittivity functions in finite-difference poisson–boltzmann electrostatics. *Theoretical Chemistry Accounts*, 119(5):421–427, 2008.

[86] Dan Ben-Yaakov, David Andelman, Rudi Podgornik, and Daniel Harries. Ion-specific hydration effects: Extending the poisson-boltzmann theory. *Current Opinion in Colloid Interface Science*, 16(6):542–550, 2011.

[87] Vojko Vlachy. Ionic effects beyond poisson-boltzmann theory. *Annual Review of Physical Chemistry*, 50(1):145–165, 1999.

[88] Li Xiao, Changhao Wang, and Ray Luo. Recent progress in adapting poisson–boltzmann methods to molecular simulations. *Journal of Theoretical and Computational Chemistry*, 13(03):1430001, 2014.

[89] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[90] I. P. Omelyan, I. M. Mryglod, and R. Folk. Optimized verlet-like algorithms for molecular dynamics simulations. *Phys. Rev. E*, 65(5):056706, 2002.

[91] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.

[92] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.

[93] 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. *The Journal of Chemical Physics*, 81(8):3684–3690, 1994.

[94] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.

[95] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.

[96] González, M.A. Force fields and molecular dynamics simulations. *JDN*, 12:169–200, 2011.

[97] Sereina Riniker. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *Journal of Chemical Information and Modeling*, 58(3):565–578, 2018.

[98] Valentina Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144–150, 2005.

[99] Arieh Warshel and Michael Levitt. Folding and stability of helical proteins: Carp myogen. *Journal of Molecular Biology*, 106(2):421–437, 1976.

[100] Michael Levitt and Shneior Lifson. Refinement of protein conformations using a macromolecular energy minimization procedure. *Journal of Molecular Biology*, 46(2):269–279, 1969.

[101] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2):227–249, 1976.

[102] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[103] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[104] Adam Liwo, Jaroslaw Pillardy, Cezary Czaplewski, Jooyoung Lee, Daniel R. Ripoll, Malgorzata Groth, Sylwia Rodziewicz-Motowidlo, Rajmund Kamierkiewicz, Ryszard J. Wawak, Stanislaw Oldziej, and Harold A. Scheraga. UNRES: a united-residue force field for energy-based prediction of protein structure—orgin and significance of multibody terms. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 193–200. ACM, 2000.

[105] Parimal Kar, Srinivasa Murthy Gopal, Yi-Ming Cheng, Alexander Predeus, and Michael Feig. Primo: A transferable coarse-grained force field for proteins. *Journal of Chemical Theory and Computation*, 9(8):3769–3788, 2013.

[106] Leonardo Darré, Matías Rodrigo Machado, Astrid Febe Brandner, Humberto Carlos González, Sebastián Ferreira, and Sergio Pantano. SIRAH: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *Journal of Chemical Theory and Computation*, 11(2):723–739, 2015.

[107] Matías R. Machado, Exequiel E. Barrera, Florencia Klein, Martín Sóñora, Steffano Silva, and Sergio Pantano. The sirah 2.0 force field: Altius, fortius, citius. *Journal of Chemical Theory and Computation*, 15(4):2719–2733, 2019.

[108] Matias R. Machado and Sergio Pantano. Exploring laci–dna dynamics by multiscale simulations using the sirah force field. *Journal of Chemical Theory and Computation*, 11(10):5012–5023, 2015.

[109] Pablo D. Dans, Ari Zeida, Matías R. Machado, and Sergio Pantano. A coarse grained model for atomic-detailed dna simulations with explicit electrostatics. *Journal of Chemical Theory and Computation*, 6(5):1711–1725, 2010.

[110] Exequiel E. Barrera, Matías R. Machado, and Sergio Pantano. Fat sirah: Coarse-grained phospholipids to explore membrane–protein dynamics. *Journal of Chemical Theory and Computation*, 15(10):5674–5688, 2019.

[111] Leonardo Darré, Matías R. Machado, Pablo D. Dans, Fernando E. Herrera, and Sergio Pantano. Another coarse grain model for aqueous solvation: WAT FOUR? *Journal of Chemical Theory and Computation*, 6(12):3793–3807, 2010.

[112] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[113] Johannes Kästner. Umbrella sampling. *WIREs Computational Molecular Science*, 1(6):932–942, 2011.

[114] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[115] Jochen S. Hub, Bert L. de Groot, and David van der Spoel. g_wham—a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of Chemical Theory and Computation*, 6(12):3713–3720, 2010.

[116] Marc Souaille and Benoıt Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Computer Physics Communications*, 135(1):40–57, 2001.

[117] Srdjan Pusara, Peyman Yamin, Wolfgang Wenzel, Marjan Krstić, and Mariana Kozlowska. A coarse-grained xDLVO model for colloidal protein–protein interactions. *Physical Chemistry Chemical Physics*, 23(22):12780–12794, 2021.

[118] Mariana Kozlowska, Pawel Rodziewicz, Tillmann Utesch, Maria Andrea Mroginski, and Anna Kaczmarek-Kedziera. Solvation of diclofenac in water from atomistic molecular dynamics simulations – interplay between solute–solute and solute–solvent interactions. *Physical Chemistry Chemical Physics*, 20(13):8629–8639, 2018.

[119] D. Roberts, R. Keeling, M. Tracka, C. F. van der Walle, S. Uddin, J. Warwicker, and R. Curtis. The Role of Electrostatics in Protein–Protein Interactions of a Monoclonal Antibody. *Molecular Pharmaceutics*, 11(7):2475–2489, 2014.

[120] B. Guo, S. Kao, H. McDonald, A. Asanov, L.L. Combs, and W. William Wilson. Correlation of second virial coefficients and solubilities useful in protein crystal growth. *Journal of Crystal Growth*, 196(2):424–433, 1999.

[121] C. Haas, J. Drenth, and W. William Wilson. Relation between the Solubility of Proteins in Aqueous Solutions and the Second Virial Coefficient of the Solution. *The Journal of Physical Chemistry B*, 103(14):2808–2811, April 1999.

[122] S. Ruppert, S. I. Sandler, and A. M. Lenhoff. Correlation between the Osmotic Second Virial Coefficient and the Solubility of Proteins. *Biotechnology Progress*, 17(1):182–187, 2001.

[123] Virginie Le Brun, Wolfgang Friess, Stefan Bassarab, and Patrick Garidel. Correlation of protein-protein interactions as assessed by affinity chromatography with colloidal protein stability: A case study with lysozyme. *Pharmaceutical development and technology*, 15(4):421–430, 2010.

[124] L. F. M. Franco and P. de A. Pessôa Filho. On the relationship between the solubility of proteins and the osmotic second virial coefficient. *Brazilian Journal of Chemical Engineering*, 30(1):95–104, 2013.

[125] X. Pan and C. E. Glatz. Solvent Effects on the Second Virial Coefficient of Subtilisin and Solubility. *Crystal Growth & Design*, 3(2):203–207, 2003.

[126] Mahlet A. Woldeyes, Cesar Calero-Rubio, Eric M. Furst, and Christopher J. Roberts. Predicting Protein Interactions of Concentrated Globular Protein Solutions Using Colloidal Models. *The Journal of Physical Chemistry B*, 121(18):4756–4767, 2017.

[127] A Quigley and DR Williams. The second virial coefficient as a predictor of protein aggregation propensity: a self-interaction chromatography study. *European Journal of Pharmaceutics and Biopharmaceutics*, 96:282–290, 2015.

[128] Erinc Sahin, Adeola O. Grillo, Melissa D. Perkins, and Christopher J. Roberts. Comparative Effects of pH and Ionic Strength on Protein–Protein Interactions, Unfolding, and Aggregation for IgG1 Antibodies. *Journal of Pharmaceutical Sciences*, 99(12):4830–4848, 2010.

[129] André C. Dumetz, Ann M. Snellinger-O'Brien, Eric W. Kaler, and Abraham M. Lenhoff. Patterns of protein–protein interactions in salt solutions and implications for protein crystallization. *Protein Science*, 16(9):1867–1877, 2007.

[130] Chirag M. Mehta, Edward T. White, and James D. Litster. Correlation of second virial coefficient with solubility for proteins in salt solutions. *Biotechnology Progress*, 28(1):163–170, January 2012.

[131] Amith D. Naik and Sunil S. Bhagwat. Optimization of an Artificial Neural Network for Modeling Protein Solubility. *Journal of Chemical & Engineering Data*, 50(2):460–467, 2005.

[132] Kristen Demoruelle, Bin Guo, Shangming Kao, Heather M. McDonald, Dragan B. Nikic, Steven C. Holman, and W. William Wilson. Correlation between the osmotic second virial coefficient and solubility for equine serum albumin and ovalbumin. *Acta Crystallographica Section D Biological Crystallography*, 58(10):1544–1548, 2002.

[133] Christopher A. Teske, Harvey W. Blanch, and John M. Prausnitz. Measurement of LysozymeLysozyme Interactions with Quantitative Affinity Chromatography. *The Journal of Physical Chemistry B*, 108(22):7437–7444, 2004.

[134] Carsten Ersch, Lennart L.C. Meijvogel, Erik van der Linden, Anneke Martin, and Paul Venema. Interactions in protein mixtures. part i: Second virial coefficients from osmometry. *Food Hydrocolloids*, 52:982–990, 2016.

[135] Yingfang Ma, Diana M. Acosta, Jon R. Whitney, Rudolf Podgornik, Nicole F. Steinmetz, Roger H. French, and V. Adrian Parsegian. Determination of the second virial coefficient of bovine serum albumin under varying pH and ionic strength by composition-gradient multi-angle static light scattering. *Journal of Biological Physics*, 41(1):85–97, 2015.

[136] X. X. Li, X. D. Xu, Y. Y. Dan, and M. L. Zhang. The factors during protein crystallization: A review. *Crystallography Reports*, 53(7):1261–1266, 2008.

[137] Eli Ruckenstein and Ivan L. Shulgin. Effect of salts and organic additives on the solubility of proteins in aqueous solutions. *Advances in Colloid and Interface Science*, 123:97–103, 2006.

[138] Weimin Li, Bjorn A Persson, Maxim Morin, Manja A Behrens, Mikael Lund, and Malin Zackrisson Oskolkova. Charge-induced patchy attractions between proteins. *The Journal of Physical Chemistry B*, 119(2):503–508, 2015.

[139] Leigh J. Quang, Stanley I. Sandler, and Abraham M. Lenhoff. Anisotropic Contributions to Protein–Protein Interactions. *Journal of Chemical Theory and Computation*, 10(2):835–845, 2014.

[140] Cesar Calero-Rubio, Atul Saluja, and Christopher J. Roberts. Coarse-Grained Antibody Models for “Weak” Protein–Protein Interactions from Low to High Concentrations. *The Journal of Physical Chemistry B*, 120(27):6592–6605, 2016.

[141] Cesar Calero-Rubio, Ranendu Ghosh, Atul Saluja, and Christopher J. Roberts. Predicting Protein-Protein Interactions of Concentrated Antibody Solutions Using Dilute Solution Data and Coarse-Grained Molecular Models. *Journal of Pharmaceutical Sciences*, 107(5):1269–1281, 2018.

[142] Cesar Calero-Rubio, Atul Saluja, Erinc Sahin, and Christopher J Roberts. Predicting High-Concentration Interactions of Monoclonal Antibody Solutions: Comparison of Theoretical Approaches for Strongly Attractive Versus Repulsive Conditions. *J. Phys. Chem. B*, 123(27):5709–5720, 2019.

[143] Christopher J. O'Brien, Cesar Calero-Rubio, Vladimir I. Razinkov, Anne S. Robinson, and Christopher J. Roberts. Biophysical characterization and molecular simulation of electrostatically driven self-association of a single-chain antibody. *Protein Science*, 27(7):1275–1285, 2018.

[144] Alfredo Jost Lopez, Patrick K. Quoika, Max Linke, Gerhard Hummer, and Jürgen Köfinger. Quantifying Protein–Protein Interactions in Molecular Simulations. *The Journal of Physical Chemistry B*, 124(23):4673–4685, 2020.

[145] Marco A. Blanco, Erinc Sahin, Anne S. Robinson, and Christopher J. Roberts. Coarse-Grained Model for Colloidal Protein Interactions, B_{22} , and Protein Cluster Formation. *The Journal of Physical Chemistry B*, 117(50):16013–16028, 2013.

[146] Austin C. Stark, Casey T. Andrews, and Adrian H. Elcock. Toward Optimized Potential Functions for Protein–Protein Interactions in Aqueous Solutions: Osmotic Second Virial Coefficient Calculations Using the MARTINI Coarse-Grained Force Field. *Journal of Chemical Theory and Computation*, 9(9):4176–4185, 2013.

[147] Sanbo Qin and Huan-Xiang Zhou. Calculation of Second Virial Coefficients of Atomistic Proteins Using Fast Fourier Transform. *The Journal of Physical Chemistry B*, 123(39):8203–8215, 2019.

[148] O.D. Velev, E.W. Kaler, and A.M. Lenhoff. Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen. *Biophysical Journal*, 75(6):2682–2697, 1998.

[149] Adrian H. Elcock and J. Andrew McCammon. Calculation of Weak Protein-Protein Interactions: The pH Dependence of the Second Virial Coefficient. *Biophysical Journal*, 80(2):613–625, 2001.

[150] Yannis Georgalis, Athina Zouni, Wolfram Eberstein, and Wolfram Saenger. Formation dynamics of protein precrystallization fractal clusters. *Journal of Crystal Growth*, 126(2):245–260, 1993.

[151] A. G. Cherstvy and R. G. Winkler. Polyelectrolyte Adsorption onto Oppositely Charged Interfaces: Image-Charge Repulsion and Surface Curvature. *The Journal of Physical Chemistry B*, 116(32):9838–9845, 2012.

[152] K.A Karraker and C.J Radke. Disjoining pressures, zeta potentials and surface tensions of aqueous non-ionic surfactant/electrolyte solutions: theory and comparison to experiment. *Advances in Colloid and Interface Science*, 96(1):231–264, 2002.

[153] M. Boström, V. Deniz, G.V. Franks, and B.W. Ninham. Extended dlvo theory: Electrostatic and non-electrostatic forces in oxide suspensions. *Advances in Colloid and Interface Science*, 123:5–15, 2006.

[154] S. Nir and M. Andersen. Van der waals interactions between cell surfaces. *The Journal of Membrane Biology*, 31(1):1–18, 1977.

[155] C.M. Roth, B.L. Neal, and A.M. Lenhoff. Van der Waals interactions involving proteins. *Biophysical Journal*, 70(2):977–987, 1996.

[156] Martin Muschol and Franz Rosenberger. Interactions in undersaturated and supersaturated lysozyme solutions: Static and dynamic light scattering results. *The Journal of Chemical Physics*, 103(24):10424–10432, 1995.

[157] Gang Wang, Zsigmond Varga, Jennifer Hofmann, Isidro E Zarraga, and James W Swan. Structure and relaxation in solutions of monoclonal antibodies. *The Journal of Physical Chemistry B*, 122(11):2867–2880, 2018.

[158] Anton Arkhipov, Ying Yin, and Klaus Schulten. Four-Scale Description of Membrane Sculpting by BAR Domains. *Biophysical Journal*, 95(6):2806–2821, 2008.

[159] Srinivasan Damodaran. On the Molecular Mechanism of Stabilization of Proteins by Cosolvents: Role of Lifshitz Electrodynamic Forces. *Langmuir*, 28(25):9475–9486, 2012.

[160] Anton Arkhipov, Peter L. Freddolino, and Klaus Schulten. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure*, 14(12):1767–1777, 2006.

[161] Yizhak Marcus. A simple empirical model describing the thermodynamics of hydration of ions of widely varying charges, sizes, and shapes. *Biophysical Chemistry*, 51(2):111–127, 1994.

[162] Yanjie Zhang and Paul S Cremer. Interactions between macromolecules and ions: the hofmeister series. *Current Opinion in Chemical Biology*, 10(6):658–663, 2006.

[163] Halil I. Okur, Jana Hladílková, Kelvin B. Rembert, Younhee Cho, Jan Heyda, Joachim Dzubiella, Paul S. Cremer, and Pavel Jungwirth. Beyond the Hofmeister Series: Ion-Specific Effects on Proteins and Their Biological Functions. *The Journal of Physical Chemistry B*, 121(9):1997–2014, 2017.

[164] Mathias Boström, Frederico W. Tavares, Barry W. Ninham, and John M. Prausnitz. Effect of salt identity on the phase diagram for a globular protein in aqueous electrolyte solution. *The Journal of Physical Chemistry B*, 110(48):24757–24760, 2006.

[165] Mathias Boström, David R. M. Williams, and Barry W. Ninham. Surface Tension of Electrolytes: Specific Ion Effects Explained by Dispersion Forces. *Langmuir*, 17(15):4475–4478, 2001.

[166] Livia A. Moreira, Mathias Boström, Barry W. Ninham, Evaristo C. Biscaia, and Frederico W. Tavares. Effect of the ion-protein dispersion interactions on the protein-surface and protein-protein interactions. *Journal of the Brazilian Chemical Society*, 18(1):223–230, 2007.

[167] Andrea Salis and Barry W Ninham. Models and mechanisms of hofmeister effects in electrolyte solutions, and colloid and protein systems revisited. *Chemical Society Reviews*, 43(21):7358–7377, 2014.

[168] Drew F. Parsons, Mathias Boström, Pierandrea Lo Nostro, and Barry W. Ninham. Hofmeister effects: interplay of hydration, nonelectrostatic potentials, and ion size. *Physical Chemistry Chemical Physics*, 13(27):12352, 2011.

[169] W. Kunz, J. Henle, and B. W. Ninham. 'Zur Lehre von der Wirkung der Salze' (about the science of the effect of salts): Franz Hofmeister's historical papers. *Current Opinion in Colloid & Interface Science*, 9(1):19–37, 2004.

[170] Chresten R. Søndergaard, Mats H. M. Olsson, Michał Rostkowski, and Jan H. Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pka values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, 2011.

[171] Hui Li, Andrew D. Robertson, and Jan H. Jensen. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics*, 61(4):704–721, 2005.

[172] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(2):W522–W525, 2007.

[173] Todd J. Dolinsky, Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32(2):W665–W667, 2004.

[174] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[175] Marcel Herhut, Christoph Brandenbusch, and Gabriele Sadowski. Modeling and prediction of protein solubility using the second osmotic virial coefficient. *Fluid Phase Equilibria*, 422:32–42, 2016.

[176] David H. Johnson, Arun Parupudi, W. William Wilson, and Lawrence J. DeLucas. High-Throughput Self-Interaction Chromatography: Applications in Protein Formulation Prediction. *Pharmaceutical Research*, 26(2):296–305, 2009.

[177] Joseph J. Valente, Beth G. Fryksdale, Douglas A. Dale, Alfred L. Gaertner, and Charles S. Henry. Screening for physical stability of a Pseudomonas amylase using self-interaction chromatography. *Analytical Biochemistry*, 357(1):35–42, 2006.

[178] Peter M. Tessier, Abraham M. Lenhoff, and Stanley I. Sandler. Rapid Measurement of Protein Osmotic Second Virial Coefficients by Self-Interaction Chromatography. *Biophysical Journal*, 82(3):1620–1631, 2002.

[179] J. Wanka and W. Peukert. Optimized Production of Protein Crystals: From 1D Crystallization Slot towards 2D Supersaturation B22 Diagram. *Chemical Engineering & Technology*, 34(4):510–516, 2011.

[180] Yuriy V. Kalyuzhnyi and Vojko Vlachy. Explicit-water theory for the salt-specific effects and Hofmeister series in protein solutions. *The Journal of Chemical Physics*, 144(21):215101, 2016.

[181] YoonKook Park and Ginkyu Choi. Effects of pH, salt type, and ionic strength on the second virial coefficients of aqueous bovine serum albumin solutions. *Korean Journal of Chemical Engineering*, 26(1):193–198, 2009.

[182] Y. U. Moon, R. A. Curtis, C. O. Anderson, H. W. Blanch, and J. M. Prausnitz. Protein—protein interactions in aqueous ammonium sulfate solutions. lysozyme and bovine serum albumin (bsa). *Journal of Solution Chemistry*, 29(8):699–718, 2000.

[183] Ramu Anandakrishnan, Boris Aguilar, and Alexey V. Onufriev. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*, 40(W1):W537–W541, 05 2012.

[184] Vincent L Vilker, Clark K Colton, and Kenneth A Smith. The osmotic pressure of concentrated protein solutions: Effect of concentration and ph in saline solutions of bovine serum albumin. *Journal of Colloid and Interface Science*, 79(2):548–566, 1981.

[185] Oleksii V. Khorolskyi, Nikolay P. Malomuzh, Oleksii V. Khorolskyi, and Nikolay P. Malomuzh. Macromolecular sizes of serum albumins in its aqueous solutions. *AIMS Biophysics*, 7(4):219–235, 2020.

[186] Priyanka Singh, Aisling Roche, Christopher F. van der Walle, Shahid Uddin, Jiali Du, Jim Warwicker, Alain Pluen, and Robin Curtis. Determination of Protein–Protein Interactions in a Mixture of Two Monoclonal Antibodies. *Molecular Pharmaceutics*, 16(12):4775–4786, 2019.

[187] S. Bassarab S. Mühlau Roberts V. Le Brun, W. Friess, European Journal of Pharmaceutics P. Garidel, 75 16–25. D. Biopharmaceutics, 2010, R. Keeling, M. Tracka, C. F. van der Walle, S. Uddin, J. Warwicker, and R. Curtis. Specific Ion and Buffer Effects on Protein–Protein Interactions of a Monoclonal Antibody. *Molecular Pharmaceutics*, 12(1):179–193, 2015.

[188] Srdjan Pusara, Wolfgang Wenzel, and Mariana Kozlowska. Accurate calculation of second osmotic virial coefficients of proteins using mixed poisson-boltzmann and extended dlvo theory. *Molecular Systems Design & Engineering*, 2023.

[189] D Leckband and S Sivasankar. Forces controlling protein interactions: theory and experiment. *Colloids and Surfaces B: Biointerfaces*, 14(1):83–97, 1999.

[190] A.C.A. Roque, C.R. Lowe, and M.A. Taipa. Antibodies and genetically engineered related molecules: Production and purification. *Biotechnology Progress*, 20(3):639–654, 2004.

[191] Andrea Salis, Mathias Boström, Luca Medda, Francesca Cugia, Brajesh Barse, Drew F. Parsons, Barry W. Ninham, and Maura Monduzzi. Measurements and theoretical interpretation of points of zero charge/potential of bsa protein. *Langmuir*, 27(18):11597–11604, 2011.

[192] M. Boström, F. W. Tavares, D. Bratko, and B. W. Ninham. Specific Ion Effects in Solutions of Globular Proteins: Comparison between Analytical Models and Simulation. *The Journal of Physical Chemistry B*, 109(51):24489–24494, 2005.

[193] Eduardo R. A. Lima, Evaristo C. Biscaia, Mathias Boström, Frederico W. Tavares, and John M. Prausnitz. Osmotic second virial coefficients and phase diagrams for aqueous proteins from a much-improved poissonboltzmann equation. *The Journal of Physical Chemistry C*, 111(43):16055–16059, 2007.

[194] Bongkeun Kim and Xueyu Song. Calculations of the second virial coefficients of protein solutions with an extended fast multipole method. *Phys. Rev. E*, 83(1):011915, 2011.

[195] Xueyu Song. Solvation dynamics in ionic fluids: An extended debye–hückel dielectric continuum model. *The Journal of Chemical Physics*, 131(4):044503, 2009.

[196] Francesco Pizzitutti, Massimo Marchi, and Daniel Borgis. Coarse-graining the accessible surface and the electrostatics of proteins for proteinprotein interactions. *Journal of Chemical Theory and Computation*, 3(5):1867–1876, 2007.

[197] Evgenni Mikhailovich Lifshitz, M Hamermesh, et al. The theory of molecular attractive forces between solids. In *Perspectives in theoretical physics*, pages 329–349. Elsevier, 1992.

[198] Srinivasan Damodaran. Electrodynamic pressure modulation of protein stability in cosolvents. *Biochemistry*, 52(46):8363–8373, 2013.

[199] Michael Farnum and Charles Zukoski. Effect of glycerol on the interactions and solubility of bovine pancreatic trypsin inhibitor. *Biophysical journal*, 76(5):2716–2726, 1999.

[200] Virginie Le Brun, Wolfgang Friess, Stefan Bassarab, Silke Mühlau, and Patrick Garidel. A critical evaluation of self-interaction chromatography as a predictive tool for the assessment of protein–protein interactions in protein formulation development: A case study of a therapeutic monoclonal antibody. *European Journal of Pharmaceutics and Biopharmaceutics*, 75(1):16–25, 2010.

[201] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the apbs biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.

[202] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[203] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 2017.

[204] Daniel E. Kuehner, Harvey W. Blanch, and John M. Prausnitz. Salt-induced protein precipitation: Phase equilibria from an equation of state. *Fluid Phase Equilibria*, 116(1):140–147, 1996.

[205] Paolo Mereghetti, Razif R. Gabdoulline, and Rebecca C. Wade. Brownian dynamics simulation of protein solutions: Structural and dynamical properties. *Biophysical Journal*, 99(11):3782–3791, 2010.

[206] Peter M. Tessier, Harvey R. Johnson, Rajesh Pazhianur, Bryan W. Berger, Jessica L. Prentice, Brian J. Bahnson, Stanley I. Sandler, and Abraham M. Lenhoff. Predictive crystallization of ribonuclease a via rapid screening of osmotic second virial coefficients. *Proteins: Structure, Function, and Bioinformatics*, 50(2):303–311, 2002.

[207] O.D. Velev, E.W. Kaler, and A.M. Lenhoff. Protein interactions in solution characterized by light and neutron scattering: Comparison of lysozyme and chymotrypsinogen. *Biophysical Journal*, 75(6):2682–2697, 1998.

[208] P.E. Pjura, A.M. Lenhoff, S.A. Leonard, and A.G. Gittis. Protein crystallization by design: chymotrypsinogen without precipitants 1 1edited by i. a. wilson. *Journal of Molecular Biology*, 300(2):235–239, 2000.

[209] Harminder Bajaj, Vikas K. Sharma, and Devendra S. Kalonia. Determination of second virial coefficient of proteins using a dual-detector cell for simultaneous measurement of scattered light intensity and concentration in SEC-HPLC. *Biophysical Journal*, 87(6):4048–4055, 2004.

[210] Mikael Lund and Bo Jönsson. A mesoscopic model for protein-protein interactions in solution. *Biophysical Journal*, 85(5):2940–2947, 2003.

[211] Jinsheng Zhou, Adam Zlotnick, and Stephen C. Jacobson. Disassembly of single virus capsids monitored in real time with multicycle resistive-pulse sensing. *Analytical Chemistry*, 94(2):985–992, 2022.

[212] Maelenn Chevreuil, Lauriane Lecoq, Shishan Wang, Laetitia Gargowitsch, Naïma Nhiri, Eric Jacquet, Thomas Zinn, Sonia Fieulaine, Stéphane Bressanelli, and Guillaume Tresset. Nonsymmetrical dynamics of the HBV capsid assembly and disassembly evidenced by their transient species. *The Journal of Physical Chemistry B*, 124(45):9987–9995, 2020.

[213] Lisa Selzer, Sarah P. Katen, and Adam Zlotnick. The hepatitis b virus core protein intradimer interface modulates capsid assembly and stability. *Biochemistry*, 53(34):5496–5504, 2014.

[214] Lisa Selzer and Adam Zlotnick. Assembly and release of hepatitis b virus. *Cold Spring Harbor perspectives in medicine*, 5(12), 2015.

[215] Corinne A. Lutomski, Nicholas A. Lyktey, Elizabeth E. Pierson, Zhongchao Zhao, Adam Zlotnick, and Martin F. Jarrold. Multiple pathways in capsid assembly. *Journal of the American Chemical Society*, 140(17):5784–5790, 2018.

[216] Roi Asor, Lisa Selzer, Christopher John Schlicksup, Zhongchao Zhao, Adam Zlotnick, and Uri Raviv. Assembly reactions of hepatitis b capsid protein into capsid nanoparticles follow a narrow path through a complex reaction landscape. *ACS Nano*, 13(7):7610–7626, 2019.

[217] Roi Asor, Christopher John Schlicksup, Zhongchao Zhao, Adam Zlotnick, and Uri Raviv. Rapidly forming early intermediate structures dictate the pathway of capsid assembly. *Journal of the American Chemical Society*, 142(17):7868–7882, 2020.

[218] Pablo Ceres, Stephen J. Stray, and Adam Zlotnick. Hepatitis b virus capsid assembly is enhanced by naturally occurring mutation f97l. *Journal of Virology*, 78(17):9538–9543, 2004.

[219] Zhongchao Zhao, Joseph Che-Yen Wang, Carolina Pérez Segura, Jodi A. Hadden-Perilla, and Adam Zlotnick. The integrity of the intradimer interface of the hepatitis b virus capsid protein dimer regulates capsid self-assembly. *ACS Chemical Biology*, 15(12):3124–3132, 2020.

[220] Angela Patterson, Zhongchao Zhao, Elizabeth Waymire, Adam Zlotnick, and Brian Bothner. Dynamics of hepatitis b virus capsid protein dimer regulate assembly through an allosteric network. *ACS Chemical Biology*, 15(8):2273–2280, 2020.

[221] Ryan C. Oliver, Wojciech Potrzebowski, Seyed Morteza Najibi, Martin Nors Pedersen, Lise Arleth, Najet Mahmoudi, and Ingemar André. Assembly of capsids from hepatitis b virus core protein progresses through highly populated intermediates in the presence and absence of RNA. *ACS Nano*, 14(8):10226–10238, 2020.

[222] Pong Kian Chua, Fan-Mei Tang, Jyuan-Yuan Huang, Ching-Shu Suen, and Chiaho Shih. Testing the balanced electrostatic interaction hypothesis of hepatitis b virus DNA synthesis by using an *In Vivo* charge rebalance approach. *Journal of Virology*, 84(5):2340–2351, 2010.

[223] Adam Zlotnick, Jennifer M. Johnson, Paul W. Wingfield, Stephen J. Stahl, and Dan Endres. A theoretical model successfully identifies features of hepatitis b virus capsid assembly. *Biochemistry*, 38(44):14644–14652, 1999.

[224] Jodi A Hadden, Juan R Perilla, Christopher John Schlicksup, Balasubramanian Venkatakrishnan, Adam Zlotnick, and Klaus Schulten. All-atom molecular dynamics

of the hbv capsid reveals insights into biological function and cryo-em resolution limits. *Elife*, 7:e32478, 2018.

[225] Carolina Pérez-Segura, Boon Chong Goh, and Jodi A. Hadden-Perilla. All-atom MD simulations of the HBV capsid complexed with AT130 reveal secondary and tertiary structural changes and mechanisms of allostery. *Viruses*, 13(4):564, 2021.

[226] Jieying Zang, Min Liu, Huan Liu, and Lina Ding. A molecular simulation study of hepatitis b virus core protein and the nuclear protein allosteric modulators of phthalazinone derivatives. *Physical Chemistry Chemical Physics*, 24(38):23209–23225, 2022.

[227] Huihui Liu, Susumu Okazaki, and Wataru Shinoda. Heteroaryldihydropyrimidines alter capsid assembly by adjusting the binding affinity and pattern of the hepatitis b virus core protein. *Journal of Chemical Information and Modeling*, 59(12):5104–5110, 2019.

[228] Jing Tu, Jiao Jiao Li, Zhi Jie Shan, and Hong Lin Zhai. Exploring the binding mechanism of heteroaryldihydropyrimidines and hepatitis b virus capsid combined 3d-QSAR and molecular dynamics. *Antiviral Research*, 137:151–164, 2017.

[229] Kazushi Fujimoto, Motohiro Fukai, Ryo Urano, Wataru Shinoda, Tetsuya Ishikawa, Katsumi Omagari, Yasuhito Tanaka, Atsushi Nakagawa, and Susumu Okazaki. Free energy profile of permeation of entecavir through hepatitis b virus capsid studied by molecular dynamics calculation. *Pure and Applied Chemistry*, 92(10):1585–1594, 2020.

[230] Go Watanabe, Shunsuke Sato, Mitsuo Iwadate, Hideaki Umeyama, Michiyo Hayakawa, Yoshiki Murakami, and Shigetaka Yoneda. Molecular dynamics simulations to determine the structure and dynamics of hepatitis b virus capsid bound to a novel anti-viral drug. *CHEMICAL & PHARMACEUTICAL BULLETIN*, 64(9):1393–1396, 2016.

[231] Anna Pavlova, Leda Bassit, Bryan D. Cox, Maksym Korablyov, Christophe Chipot, Dharmeshkumar Patel, Diane L. Lynch, Franck Amblard, Raymond F. Schinazi, and James C. Gumbart. The mechanism of action of hepatitis b virus capsid assembly modulators can be predicted from binding to early assembly intermediates. *Journal of Medicinal Chemistry*, 65(6):4854–4864, 2022.

[232] Zhaleh Ghaemi, Martin Gruebele, and Emad Tajkhorshid. Molecular mechanism of capsid disassembly in hepatitis b virus. *Proceedings of the National Academy of Sciences*, 118(36):e2102530118, 2021.

[233] Jehoon Kim and Jianzhong Wu. A molecular thermodynamic model for the stability of hepatitis b capsids. *The Journal of Chemical Physics*, 140(23):235101, 2014.

[234] Jehoon Kim and Jianzhong Wu. A thermodynamic model for genome packaging in hepatitis b virus. *Biophysical Journal*, 109(8):1689–1697, 2015.

[235] Dong Meng, Rex P Hjelm, Jianming Hu, and Jianzhong Wu. A theoretical model for the dynamic structure of hepatitis b nucleocapsid. *Biophysical Journal*, 101(10):2476–2484, 2011.

[236] Tao Jiang, Zhen-Gang Wang, and Jianzhong Wu. Electrostatic regulation of genome packaging in human hepatitis b virus. *Biophysical Journal*, 96(8):3065–3073, 2009.

[237] Matías R. Machado and Sergio Pantano. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics*, 32(10):1568–1570, 2016.

[238] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An N log(N) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[239] Yumeng Yan, Di Zhang, Pei Zhou, Botong Li, and Sheng-You Huang. Hdock: a web server for protein–protein and protein–dna/rna docking based on a hybrid strategy. *Nucleic acids research*, 45(W1):W365–W373, 2017.

[240] Sheng-You Huang and Xiaoqin Zou. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 72(2):557–579, 2008.

[241] Yumeng Yan, Zeyu Wen, Xinxiang Wang, and Sheng-You Huang. Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking: Docking tool for predicting protein-protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 85(3):497–512, 2017.

[242] Nadine Homeyer and Holger Gohlke. Free energy calculations by the molecular mechanics poissonboltzmann surface area method. *Molecular Informatics*, 31(2):114–122, 2012.

[243] Rashmi Kumari, Rajendra Kumar, Open Source Drug Discovery Consortium, and Andrew Lynn. *g_mmpbsa* —a GROMACS tool for high-throughput MM-PBSA calculations. *Journal of Chemical Information and Modeling*, 54(7):1951–1962, 2014.

[244] Herman JC Berendsen, James PM Postma, Wilfred F van Gunsteren, and Jan Hermans. Interaction models for water in relation to protein hydration. pages 331–342. Springer, 1981.

[245] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 1984.

[246] M. J. Uttinger, C. S. Hundschell, V. Lautenbach, S. Pusara, S. Bäther, T. R. Heyn, J. K. Keppler, W. Wenzel, J. Walter, M. Kozlowska, A. M. Wagemans, and W. Peukert. Determination of specific and non-specific protein–protein interactions for beta-lactoglobulin by analytical ultracentrifugation and membrane osmometry experiments. *Soft Matter*, 18(35):6739–6756, 2022.

[247] Michael Gottschalk, Hanna Nilsson, Helena Roos, and Bertil Halle. Protein self-association in solution: The bovine -lactoglobulin dimer and octamer. *Protein Science*, 12(11):2404–2411, 2009.

[248] Kazumasa Sakurai, Motohisa Oobatake, and Yuji Goto. Salt-dependent monomer-dimer equilibrium of bovine -lactoglobulin at pH 3. *Protein Science*, 10(11):2325–2335, 2008.

[249] Sanaullah Khan, Richard Ipsen, Kristoffer Almdal, Birte Svensson, and Pernille Harris. Revealing the dimeric crystal and solution structure of -lactoglobulin at pH 4 and its pH and salt dependent monomer–dimer equilibrium. *Biomacromolecules*, 19(7):2905–2912, 2018.

[250] Christopher John Schlicksup, Patrick Laughlin, Steven Dunkelbarger, Joseph Che-Yen Wang, and Adam Zlotnick. Local stabilization of subunit–subunit contacts causes global destabilization of hepatitis b virus capsids. *ACS Chemical Biology*, 15(6):1708–1717, 2020.

[251] Zhenning Tan, Karolyn Pionek, Nuruddin Unchwaniwala, Megan L. Maguire, Daniel D. Loeb, and Adam Zlotnick. The interface between hepatitis b virus capsid proteins affects self-assembly, pregenomic RNA packaging, and reverse transcription. *Journal of Virology*, 89(6):3275–3284, 2015.

[252] Klaus Klumpp, Angela M Lam, Christine Lukacs, Robert Vogel, Suping Ren, Christine Espiritu, Ruth Baydo, Kateri Atkins, Jan Abendroth, Guochun Liao, et al. High-resolution crystal structure of a hepatitis b virus replication inhibitor bound to the viral core protein. *Proceedings of the National Academy of Sciences*, 112(49):15196–15201, 2015.

[253] Christina R Bourne, Sarah P Katen, Matthew R Fulz, Charles Packianathan, and Adam Zlotnick. A mutant hepatitis b virus core protein mimics inhibitors of icosahedral capsid self-assembly. *Biochemistry*, 48(8):1736–1742, 2009.