PROTEIN INTERACTIONS, UNFOLDING AND AGGREGATION FROM LOW TO HIGH PROTEIN CONCENTRATIONS VIA COARSE-GRAINED MOLECULAR MODELING AND EXPERIMENTAL CHARACTERIZATION

by

Cesar O. Calero-Rubio

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Fall 2017

PROTEIN INTERACTIONS, UNFOLDING AND AGGREGATION FROM LOW TO HIGH PROTEIN CONCENTRATIONS VIA COARSE-GRAINED MOLECULAR MODELING AND EXPERIMENTAL CHARACTERIZATION

by

Cesar O. Calero-Rubio

Approved: _____
Eric M. Furst, Ph.D.
Chair of the Department of Chemical and Biomolecular Engineering

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Christopher J. Roberts, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Abraham M. Lenhoff, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Eric M. Furst, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Atul Saluja, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

There are many people whose support, both direct and indirect, have contributed to the full achievement of this work. First and foremost, I would like to thank my thesis advisor and mentor, Chris Roberts, for his patience, guidance and full support during the course of this work. I deeply appreciate the time dedicated to all our scientific and not-so-scientific discussions that enhanced my growth as a scientist and a person. Chris constantly challenged me to go beyond what can be seen at first sight and better understand protein behavior from fundamentals. His approach to solving problems will always stay with me during my career.

I would also like to thank the many collaborators and colleagues who shared their labs and knowledge with me. A special thanks to the current and former members of the Drug Product Science and Technology group at Bristol-Myers Squibb for their financial and scientific support during the realization of this work, in particular to Atul Saluja and Erinc Sahin. Thank you for your patience, the many helpful discussions and for teaching me practical industrial perspectives when addressing scientific problems. I also want to thank Kristi Kiick, Xinqiao Jia and Brad Paik for all the support and guidance. I enjoyed meeting with you to discuss peptide behavior and better understand the material properties of solutions. Thanks to Paul Butler and Susana Teixeira at the NIST Center for Neutron Research who were always friendly and helpful to discuss Small Angle Scattering experiments.

I want to thank both current and former member of the Roberts Research Group whom I have shared experiences, and who have assisted me throughout the experimental

and computational issues I encountered. Special thanks to Mahlet Woldeyes, Greg Barnett, Ranendu Ghosh and Marco Blanco, for supporting and training me during my years in the group. Finally, I want to thank my parents Orlando and Elcida, my brother Daniel, my significant other Jackie, and all my friends for their constant support and for providing me with enough distraction from my scientific activities so I could keep moving forward in both my personal and professional lives. All of you have supported me in good and not so good times, and for that I am really grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The diverse behavior of protein solutions can be attributed to the collection of microscopic interactions between solvent, protein, buffer and other cosolute molecules. These interactions can dictate different solution properties such as native and non-native aggregation, phase separation (liquid-liquid separation, crystallization, etc.), opalescence and elevated solution viscosity. These are monitored and controlled during the development and manufacturing of protein solutions for different health care, food, and other industrial applications. Better understanding and control of the interactions among solution molecules can help to optimize the development and manufacturing of protein solutions, leading to a decrease in overall product and process development costs. However, the complex nature of the molecular-scale interactions makes it challenging to identify the key contributions to these interactions and resulting solution behaviors.

Protein-protein interactions, as opposed to protein-solvents and protein-cosolute interactions, are the most studied and arguably better understood solution interactions used during the development of protein-based products. Historically, these interactions have been experimentally characterized at low protein concentrations (c_2) due to several instrumental and theoretical limitations. This has been done using different interaction parameters such as the protein-protein second osmotic virial coefficient (B_{22}), the diffusion interaction parameter (k_D) or their surrogates. These are measured using static and dynamic light scattering (SLS and DLS, respectively) techniques, analytical ultra-centrifugation (AUC) or self-interaction chromatography (SIC), among others. Recent

efforts have focused on developing methodologies to measure interactions at much higher c_2, where SLS coupled with Kirkwood-Buff (KB) solution theory are used to interpret the high-c_2 behavior of protein solutions *via* the protein-protein KB integral (G_{22}). By contrast, simple colloidal (spherical) models have historically been used to capture protein interactions as a function of solution environment and c_2, with larger emphasis on low-c_2 conditions. However, these models might lack enough molecular resolution to capture the c_2-dependent behavior of protein interactions, as well as enough structural definition to model anisotropic protein molecules, such as monoclonal antibodies (mAb).

The viability of using coarse-grained (CG) molecular models to capture low- to high-c_2 protein-protein interactions is examined here for a series of proteins and solutions expanding from globular to mAb proteins. For dilute c_2, several different generic molecular descriptions of a given mAb molecule are used to evaluate differences across protein steric interactions (protein excluded volume effects) and the protein molecular volume using advanced Monte Carlo algorithms. This comparison allows one to find models that can self-consistently capture low- and high-c_2 packing behavior. These models are later used to evaluate the protein solution osmotic compressibility as a function of c_2 using transition matrix Monte Carlo algorithms. The results highlight shortcomings of using spherical models to capture antibody solution interactions, while anisotropic mAb-like models considerably improve consistency between protein packing and molecular volume. For globular proteins, the spherical assumption is expected to accurately represent both protein excluded and molecular volumes, so no additional refinements were performed.

Those CG models that were found to accurately capture the packing behavior of proteins from low to high c_2 are coupled with simple potential of mean force (PMF) models to capture short-ranged non-electrostatic attractions (from van der Waals and solvation effects), screened electrostatic interactions (from the protein anisotropic charge distribution) and highly flexible protein regions, when needed. This is done to minimize the number of model parameters yet capture the main contributions to protein-protein interactions. Experimental measurements of excess Rayleigh scattering were performed as a function of solution pH, buffer identity and concentration, and sucrose content, for a series of NaCl concentrations to estimate B_{22} values and decouple the electrostatic and short-ranged non-electrostatic attractive contributions to B_{22}. The former dominates at low total ionic strength (*TIS*) and the latter at high-*TIS* conditions. The results are used to refine model parameters at low c_2, which are later used to predict high-c_2 excess Rayleigh scattering behavior for similar solution conditions, up to 160 g/L protein concentration. Higher resolution CG models are used to further evaluate strong attractions observed at low *TIS*, caused by highly anisotropic electrostatic attractions. The results indicate excellent agreement between experimental and predicted values when protein interactions are repulsive to weakly attractive ($B_{22}/B_{22,ST} > -3$, with $B_{22,ST}$ representing the excluded volume contribution). Strongly attractive interactions can deviate qualitatively from the experimentally observed behavior. The CG models are also used to obtain additional domain-domain potentials of mean force as a function of c_2, pH, and *TIS* to gain insights into preferential interactions across protein domains.

Added cosolutes, such as sugars, surfactants and other stabilizers, can mediate the interactions between proteins, thus affecting the overall solution properties. These

cosolutes are expected to affect the solvation shells of proteins, which can be cast in the framework of preferential interactions as a mean to better describe the effects of adding cosolutes to protein solutions. However, current preferential interaction frameworks can only be used to interpret experimental protein-cosolute interactions for ternary systems (three-component solutions) while most protein solutions usually contain a minimum of four components (water, protein, buffer and a cosolute). A new derivation is presented here for protein-solvent and protein-cosolute interactions that can be applied to an arbitrary number and concentration of components for dilute c_2. The new framework was used to compare differences of preferential interactions in the absence and presence of buffer, for a series of cosolutes (sucrose, trehalose, sodium phosphate and sodium chloride) and added cosolute concentrations, at fixed solution temperature, pH and pressure. Protein preferential interaction measurements showed that protein-cosolute interactions in the presence (quaternary systems) and absence (ternary systems) of buffer are statistically indistinguishable as the buffer contribution was effectively zero. This would allow to treat quaternary solutions as pseudo-ternary systems, in agreement with the newly derived framework. Additional measurements were performed to identify the effects of adding sucrose to a mAb solution into the protein-protein interactions measured *via* excess Rayleigh scattering, which showed that the addition of sucrose results in increased protein-protein repulsions. Adding sucrose to a mAb solution resulted in preferential solvation/accumulation of sucrose around the protein surface, in good agreement with the orthogonal measurements of protein-protein interactions and computer simulations.

Finally, experimental measurements and molecular scale simulations were performed for a series of Ala-rich peptides to gain insights on the effects of: (i)

modifying side change hydrophobicity and (ii) modifying the chain length into the unfolding and aggregation behavior of peptides. A four-beads-per-amino acid (4bAA) CG model was coupled with Replica Exchange Molecular Dynamics (REMD) to compute unfolding and aggregation transitions and identify intermediate states during the unfolding behavior of five polypeptide sequences with similar chemistry. Circular dichroism (CD) was used to unveil the unfolding and aggregation behavior along with peptide secondary structure of the same sequences as a function of temperature and polypeptide concentration. The results showed a linear increase in thermal stability with chain length, with a decrease in stability with decrease hydrophobicity. This is found to be caused by the increase in solution entropy with the decrease in side-chain hydrophobicity.

Overall, the results in here demonstrate: (i) the effects of solution environment in mediating protein-protein interactions, (ii) how that can be studied within the framework of preferential interactions, (iii) the viability of coupling experimental measurements and computer simulations at low c_2 to predict high-c_2 interactions and better understand the effect of solution formulations at the microscopic level, and (iv) how more detailed CG models can be used to both capture the anisotropic surface charge distribution of proteins as well as the unfolding and aggregation propensities of polypeptides. The approaches can be generalized to other proteins of interest outside those studied here.

Chapter 1

INTRODUCTION

1.1 Motivation

Protein-based drugs are one of the fastest growing class of drug products in the pharmaceutical industry [1]. From this group of pharmaceuticals, monoclonal antibodies (mAbs) are of special interest due to their relatively high solution stability and their high affinity and specificity for targeted antigens [1,2]. During 2013, sales of mAb-based drugs surpassed $75 billion in the global market, mostly due to their flexibility and potency targeting life-threatening and degenerative diseases [1–4]. Similarly, there is increasing interest in developing other proteins for therapeutic use, such as those made with mAb fragments, small proteins, and peptide-based therapies [1,5,6]. Over the past decade the development of protein based drugs has led to enhancing human health by treating rare, intractable and/or life-threatening diseases such as cancer and auto-immune diseases. However, several challenges during the development of these therapies are still present [7–13].

During the development of protein-based therapeutics, a comprehensive group of critical quality attributes (CQAs) are investigated and monitored [10–12,14,15]. This is done to comply with requirements set by various regulatory agencies for product registration [15]. Some of these CQAs include levels and types of protein aggregation, solution viscosity, turbidity, opalescence and phase separation [8,11,16–21]. These CQAs are difficult to predict, especially for highly concentrated protein solutions (upwards of 100 g/L) [8,22,23]. Such high-concentration conditions have been

increasingly targeted for monoclonal antibody (mAb) solutions when developing self-injectable products [8,13]. CQAs are sensitive to changes in solution formulation (such as pH, ion concentration and chemical identity, and the presence of other co-solutes), protein identity (sequence, structure, binding affinity, *etc*.), storage temperature, and interfacial stresses (promoting agitation such as shaking, stirring, *etc*.) [10,12,16,24]. This leads to an exceedingly large experimental space that needs to be addressed during the development stages of any mAb, or for any protein product, to comply with regulations and target markets [7,15,16]. Therefore, it would be advantageous to have physical/mechanistic models that could reduce the experimental space by providing quantitative and qualitative predictions of at least a subset of these properties while at the same time reducing the amount of experimental information required.

So-called "colloidal" or "weak" protein-protein interactions have been shown to correlate, in some cases, with several CQAs such as protein phase separation, opalescence, aggregation rates, and elevated viscosities [17,18,23,25–27]. For instance, Figure 1.1 shows an example of such correlations between protein aggregation rates (in the form of observed rate constants, k_{obs}) and protein-protein interactions (in the form of G_{22}, see below) [23]. Here, a reference was used to compare changes in aggregation rates with the addition of cosolutes to the solution, and a correlation was found between $k_{obs}/k_{obs,ref}$ and G_{22}-$G_{22,ref}$, for $k_{obs,ref}$ and $G_{22,ref}$ representing the values for the reference solutions. This category of protein interactions can be characterized experimentally using light scattering (LS), analytical ultracentrifugation, and/or small-angle neutron and x-ray scattering techniques (SANS and SAXS, respectively) [28–36]. Each of these methods would allow one to measure the second osmotic virial coefficients (referred to as B_{22}) as a function of solution formulation, but this parameter is limited to dilute c_2

(typically below 10 g/L) [23,29]. Recent work has highlighted that changes in protein-protein interactions as a function of protein concentration (c_2) can be quantified using at least a subset of these experimental techniques [23,27,29,37–41]. Static light scattering (SLS) allows one to directly measure protein-protein interactions as a function of c_2 in the form of Kirkwood-Buff (KB) integrals, particularly the protein-protein KB integral, G_{22} [23,29]. From the perspective of industrial laboratories, SLS provides G_{22} and B_{22} values with greater accessibility and/or higher throughput compared to other techniques, especially techniques such as SAXS and SANS [29–31,38,42]. SLS provides an attractive experimental technique to measure low- to high-c_2 interactions and to compare directly with molecular simulations [29,37–39,43].

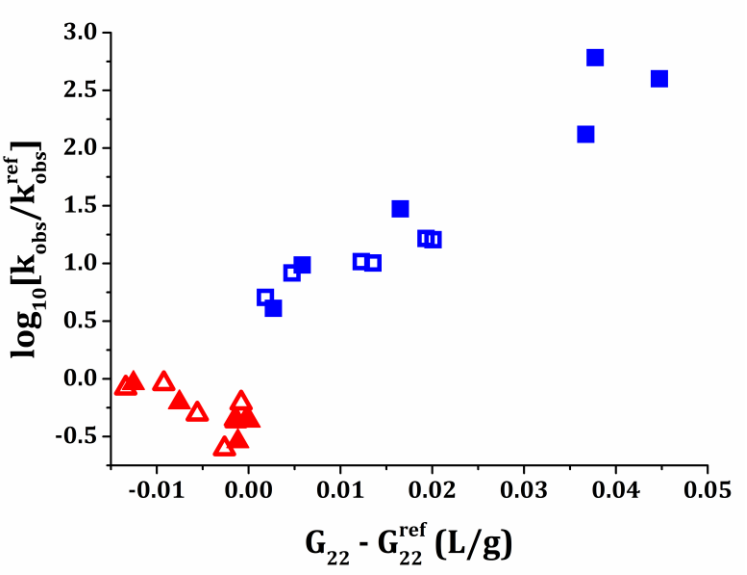

Figure 1.1. Experimental correlation of aggregation rate constants and protein-protein interactions reproduced from reference [23]. A reference was used to compare how changes in G_{22} could correlate with changes in k_{obs}.

Apart from idealized conditions, proteins rarely exist in solutions where water is the only other component [8,10,16]. For almost all biochemical conditions of interest, a minimum of three components are present: water, protein, and a co-solvent or co-solute, such as buffer or salt [8,10,44–49]. Adding co-solutes and/or co-solvents to protein solutions can have significant effects on the overall solution properties and the behavior of the protein molecules, altering protein phase behavior, aggregation rates, conformational stability, and solution viscosity [23,25,46–54]. This has fundamental and practical implications for design, manufacture, and formulation of proteins and other biomolecular solutions [8–10]. However, it is challenging to systematically characterize, quantify, and predict the effects of co-solutes (interchangeably referred as cosolutes and osmolytes in the remainder of this dissertation) on protein properties, and to capture these in terms of measurable protein-cosolute interactions [44,48,49]. All of this is due to the intrinsic complexity of multicomponent solutions.

Timasheff and coworkers [44,52–54] and Schellman and co-workers [48] developed a systematic framework to characterize the effects of added cosolutes on the chemical potential of proteins in solution, in the form of preferential interactions, as opposed to simply “direct interactions”. Preferential interactions are determined by the relative preference of protein molecules to interact with water (the solvent in biochemical applications) versus each of the other molecules in solution [44,48,55]. For ternary solutions, this can lead to two different cases: the first case occurs when the net protein-osmolyte interactions are more favorable (attractive) than the net water-protein interactions, leading to so-called preferential accumulation or preferential binding of osmolyte molecules around the protein surface, while also preferentially excluding water molecules from the solvation layers [44,56,57]. The second case is the opposite,

leading to preferential exclusion of osmolyte molecules from the protein surface. This is illustrated in Figure 1.2, where sucrose, water and a protein molecule are used to represent the dynamic behavior of sucrose interchanging with water molecules in the outer hydration layers characteristic of preferential accumulation of sucrose [44,58]. However, current approaches have only been developed and applied to ternary solutions despite most protein-based products containing a minimum of four components (water, protein, buffer and at least one "stabilizer" or co-solute) [44,48,49,59]. Consequently, evaluating and improving current approaches is necessary to better characterize, understand and develop commercial protein solutions instead of assuming idealized conditions where the collection of all co-solutes is treated as a single "background".

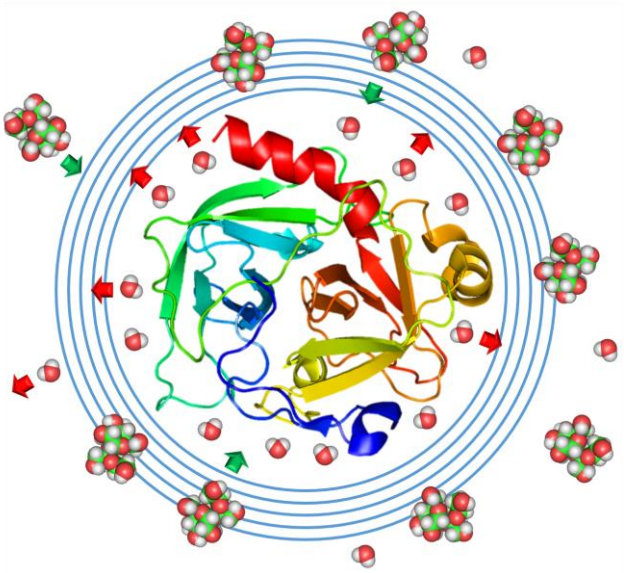

Figure 1.2. Illustrative example of protein solvation/desolvation for the ternary system water-protein-sucrose adapted from reference [60]. The blue circles are meant to represent the hydration layers of water around the protein molecule (PDB: 1EX3) with sucrose and water molecules interchanging on the outer layers to illustrate the preferential accumulation of sucrose around the protein surface.

Except for relatively small proteins and peptides, it is too computationally expensive to simulate most protein systems of interest with all-atom simulations and explicit solvent molecules [61–67]. Although it is possible to simulation large proteins, such as mAbs, with all-atom simulations and explicit solvent, obtaining volume integrals, such as B_{22} and G_{22}, is currently intractable and impractical for industrial applications [63–67]. This is further exacerbated at high c_2, where hundreds to thousands of proteins must be simulated simultaneously [61,68]. Atomistic force fields provide the most accurate method available to identify amino-acid-specific interactions, while coarse-grained (CG) molecular models provide faster computations at the expense of some degree of molecular definition [61,65–72]. Previous work has often used coarse-grained descriptions of inter-particle interactions to describe protein solutions. One example is conditions near crystallization, where coarse-grained (CG) models with implicit solvent have been used to describe protein-protein interactions [71–74]. Similarly, screened-dipole interaction models have been developed and related to thermodynamic properties and phase behavior of proteins in solution [75,76]. Although this simplified description of protein solutions is historically employed as a minimalist approach, the validity of using these low-resolution models for protein solutions has been questioned in some cases. Objections to using CG models arise from the complex nature of protein interactions on how these are influenced by the solution formulation, the protein sequence and structure, and the heterogeneous surface chemistry of proteins. Based in part on such arguments, recent examples raise the question of whether CG models are practically useful for capturing or predicting high-concentration properties of protein solutions [66,77–79]. Thus, it is of interest to address whether CG models are potentially well-suited as a computational framework to circumvent experimental

searches to optimize time and resources, provided these models can accurately predict the experimental behavior of interest.

Several examples of CG models have been previously proposed and studied to capture protein-protein interactions, protein unfolding and, in some cases, the aggregation behavior of peptides and proteins in solution [63,66,80,81]. Spherical models have been historically used to represent protein solutions from the framework of colloidal solutions [71,74,81]. Additional modifications to these spherical models have been achieved by using "patches" with various levels of attractions and repulsions to mimic the natural chemical anisotropicity of protein surfaces [82,83]. Although this protein description is approximate at best, it can provide useful insights into the effects of protein interactions in the behavior of protein solution. Higher resolution models have also been developed to better capture the effects of the protein sequence in the protein-protein interactions [62,63,66]. For instance, a one-bead-per-amino acid (1bAA) model was developed to capture B_{22} changes as a function of the solution ionic environment [66]. This model provides the flexibility of fast computations with sequence resolution that allows it to be used as a protein engineering tool. Similarly, four-bead-per-amino acid (4bAA) models have primarily been designed to capture qualitative structural features of polypeptide unfolding and self-assembly – *e.g.*, the transition between helix and coil configurations for natively helical polypeptides, the existence of local conformations during unfolding, and the formation of tertiary structures for long polypeptides [62,67]. These 4bAA models have also been used to understand the molecular scale interactions that affect the unfolding and self-assembly of Ala-rich peptides in the context of polymer-peptide interactions and amyloid formation. These recent

developments could be used to better interpret or predict experimental measurements of protein solutions.

Despite current advances for computing infrastructure and the development of faster and more accurate experimental techniques, relatively little effort has been devoted to using molecular simulations to efficiently and accurately predict protein-protein interactions of concentrated protein solutions, as well as the thermodynamics of unfolding and aggregation, using CG molecular models. Most efforts have been invested on accurately capturing low-c_2 behavior with highly structurally resolved models [62,66]. Even though qualitative structural features of the unfolding process are reasonably well captured, the quantitative details such as unfolding free energy values, midpoint unfolding temperatures (T_m), and unfolding enthalpy values obtained from molecular simulations (both CG and atomistic) typically do not match those obtained experimentally [62,69,70,84–87]. Furthermore, it is common to use molecular simulations in a "hindsight" manner, where the experimental behavior is already known and the simulations are intended to give molecular-scale insight that is beyond the capabilities of the experiment, or to help confirm or refute hypotheses that were based on interpretation of the experimental data [40,81,84]. Much less work has been devoted to predicting experimental behavior with molecular models, either *a priori* or based on a subset of experimental data that can provide a reference for future predictions.

1.2 Project Goals

This thesis focuses on characterizing "weak" protein-protein, protein-cosolute and protein-solvent interactions, and protein unfolding behavior both experimentally and theoretically, as well as gaining insights into how formulation conditions mediate protein behavior in solution. The specific goals of this thesis are to: (1) develop simple

coarse-grained molecular models that allows one to self-consistently characterize low- to high-concentration protein-protein interactions *via* Monte Carlo simulations; (2) predict high-concentration mAb interactions by coupling low-concentration experimental measurements with coarse-grained molecular scale simulations; (3) use coarse-grained molecular models to capture changes in colloidal stability caused by hydrophobic and electrostatic interactions as a function of pH and protein sequence; (4) examine the validity of these approaches for globular proteins; (5) generalize KB theory for quantifying protein-osmolyte and protein-solvent interactions from high precision density measurements for solutions with more than three components, without limitations on the number and concentration of all components outside of the protein molecules; and (6) evaluate a higher-resolution coarse-grained model to predict unfolding transitions for Ala-rich peptides as a function of peptide sequence and solution temperature. The following subsections provide the background and context needed for the particular examples in this thesis. Additional experimental and computational methodologies are explained in each specific chapter.

1.3 Statistical Mechanical Background of B_{22} and G_{22}

In the remainder of this dissertation, the notation of Casassa and Eisenberg will be employed: component 1 denotes solvent (water), 2 denotes protein along with the stoichiometric counterions needed to maintain electroneutrality, and 3, 4, etc. denote additional solute species [88]. Consequently, c_2 will represent the mass concentration (in g/L or mg/mL) of protein in the solution of interest. Protein-protein interactions can be quantified using a variety of parameters, of which B_{22} is one of the most widely used in the field of protein solution chemistry and statistical thermodynamics [26,36,50,83,89–94]. However, B_{22} is only relevant in the dilute protein limit (*i.e.*, $c_2 \rightarrow$

0 g/L) and is independent of c_2. Thus, quantifying the c_2-dependence of protein-protein interactions requires a different measurement of inter-particle interactions. The protein-protein Kirkwood-Buff integral, G_{22}, is the c_2-dependent analogue of B_{22}, and can be obtained from static light scattering (SLS, see section 1.5) and analytical ultracentrifugation experiments, as well as experiments based on x-ray or neutron scattering (see section 1.6) [29,32,81,95].

The formal definitions of B_{22} and G_{22} are:

$$B_{22} = -\frac{1}{2}\int_r \int_{\Omega_1} \int_{\Omega_2} \left[\exp\left(-\frac{w_{22}(c_2 \to 0, r, \Omega_1\Omega_2)}{k_B T}\right) - 1\right] dr d\Omega_1 d\Omega_2 \quad (1.1)$$

$$G_{22} = \int_r \int_{\Omega_1} \int_{\Omega_2} \left[\exp\left(-\frac{w_{22}(c_2, r, \Omega_1\Omega_2)}{k_B T}\right) - 1\right] dr d\Omega_1 d\Omega_2 \quad (1.2)$$

where w_{22} represents the potential of mean force between two protein molecules. In the examples in this thesis, these are the direct manifestation of "weak" or "colloidal" protein-protein interactions in solution such as van der Waals forces, hydrophobic attractions, and screened electrostatic attractions/repulsions [66,84]. w_{22} is explicitly a function of the center-to-center distance between two protein molecules (r), their relative orientations (Ω_1 & Ω_2), their concentration (c_2) and the solution environment (solvent and other solute concentrations). k_B is Boltzmann's constant and T is the absolute temperature. In equation 1.1, w_{22} is only evaluated in the dilute protein limit, which makes B_{22} a c_2-independent quantity. Conversely, w_{22} depends on c_2 in equation 1.2, as G_{22} is inherently a c_2-dependent quantity. The exception is in the limit of $c_2 \to 0$, when B_{22} and G_{22} are formally equivalent except for the difference in sign and a factor of ½ based on equation 1.1 and 1.2 [29]. Additionally, G_{22} can be related to the protein

isothermal compressibility in an osmotic system, which is also related to the fluctuations in the number of protein molecules in a Grand Canonical ensemble:

$$G_{22} = \frac{V}{\langle N_2 \rangle}\left[k_B T\left(\frac{\partial \langle N_2 \rangle}{\partial \mu_2}\right)_{T,V,\mu_{i\neq 2}} - 1\right] = \frac{V}{\langle N_2 \rangle}\left[\frac{\langle N_2^2 \rangle - \langle N_2 \rangle^2}{\langle N_2 \rangle} - 1\right] \quad (1.3)$$

where V represents the system volume, $\langle N_2 \rangle$ is the average number of proteins within the system volume, $\langle N_2^2 \rangle$-$\langle N_2 \rangle^2$ is the variance in the number of proteins in the system volume, and μ_2 is the protein chemical potential [59]. The derivative of $\langle N_2 \rangle$ with respect to μ_2 in equation 1.3 is equivalent to the osmotic compressibility [59].

1.4 Protein-Water and Protein-Osmolyte Preferential Interactions

Kirkwood-Buff (KB) solution theory is the only comprehensive analytical liquid-state theory that, in principle, allows one to predict all thermodynamic properties of macroscopic systems at a given temperature and solution component average concentrations, based only on molecular-scale properties [59,95]. Proposed by Kirkwood and Buff, this framework defines any macroscopic thermodynamic variable in terms of a set of KB integrals over pair-correlation functions [95]. The KB integral for components *i* and *j* is denoted G_{ij}, and is defined as the volume integral of the average molecular pair-correlation function ($\overline{g}_{ij}(r)$) for component *i* with respect to component *j*, relative to an ideal gas mixture in a grand canonical ensemble:

$$G_{ij} = 4\pi \int \left(\overline{g}_{ij}(r) - 1\right) r^2 dr \quad (1.4)$$

A positive G_{ij} value corresponds to a net attraction between components *i* and *j*, while a negative value corresponds to a net repulsion. $\overline{g}_{ij}(r)$ is a function of T and the concentration of all the components (including those different from *i* and *j*) in an osmotic system, so G_{ij} is an implicit function of T and the bulk concentrations or chemical

potentials of each component. Additionally, $\overline{g}_{ij}(r)$ is equivalent to the orientational-averaged potential of mean force between molecules $\overline{w}_{ij}(r)$:

$$\bar{g}_{ij}(r) = \exp\left[-\frac{\overline{w}_{ij}(r)}{k_B T}\right] \tag{1.5}$$

Consequently, equations 1.2 and 1.4 are equivalent when $i = j = 2$, and when the potential w_{22} is averaged over all possible orientations.

By mathematically inverting the original framework by Kirkwood and Buff, Ben-Naim and others have derived general thermodynamic relationships to calculate G_{ij} values based on changes in measurable thermodynamic properties, such as:

$$\left(\frac{\partial \mu_2}{\partial m_3}\right)_{T,P,m_2\to 0} = c_3(G_{12} - G_{23})\left(\frac{\partial \mu_3}{\partial m_3}\right)_{T,P,m_2\to 0} \tag{1.6}$$

where m_j and μ_j denote the molality and chemical potential of component j, respectively [59]. As noted by the subscript $m_2 \to 0$, equation 1.6 only applies under dilute protein conditions. The difference between the KB integrals (G_{12} - G_{23}) on the right-hand side of equation 1.6 dictates the preferential interactions that are observed experimentally, because the partial derivative on the right-hand side is necessarily positive for an equilibrium system [44,49,59,96]. When (G_{12} - G_{23}) is positive, the water-protein interactions are preferred over the protein-osmolyte interactions (*i.e.*, preferential exclusion of the osmolyte), and *vice versa* when (G_{12} - G_{23}) is negative. This does not specify whether each of G_{12} or G_{23} is positive or negative. The net preferential accumulation or exclusion of co-solutes or water is determined by only the difference between G_{12} and G_{23}, as shown by Timasheff and co-workers, among others [44,48,49].

Protein and cosolute chemical potentials (μ_2 and μ_3) are challenging to measure directly because proteins and most osmolytes are effectively non-volatile and difficult to crystallize from protein solutions. Derivatives of μ_2 and μ_3 versus protein and

osmolyte concentrations cannot typically be measured explicitly but, in some cases, can be inferred from osmotic pressure, vapor pressure and dialysis equilibrium experiments [46,52–54,97–102]. However, these measurements can require substantial amounts of protein, as well as significantly long experimental times to reach equilibrium [101]. An alternative approach to determine G_{ij} values and preferential interactions relies on the relationship between partial specific volumes (PSV) and KB integrals, such as those derived by Ben-Naim for binary and ternary solutions [59,99].

The PSV of component *i* in solution ($\hat{V}_i$) can be directly quantified by evaluating the change in solution density (ρ) as a function of the mass fraction of that component (w_i) at constant *T*, pressure (*P*), and solution compositions [49,59]:

$$\hat{V}_i = \frac{1}{\rho_o} + \left[\frac{d\left(\frac{1}{\rho}\right)}{dw_i}\right]_{T,P,m_{j\neq i}} \tag{1.7}$$

For ternary solutions under dilute protein conditions (*e.g.*, $c_2 < 2$ g/L) , $\hat{V}_2$ can be related to (G_{12} - G_{23}) as follows [49,59]:

$$\hat{V}_2 = \frac{RT\kappa_T}{M_w} - G_{12} + c_3\hat{V}_3(G_{12} - G_{23}) \tag{1.8}$$

Here, κ_T is the isothermal compressibility of the solution, M_w is the molecular weight of the protein, and *R* is the gas constant. For liquid solutions far from the critical point, the leftmost term on the right-hand side of equation 1.8 is sufficiently close to zero to be neglected [29,49]. Therefore, by measuring $\hat{V}_2$ and $\hat{V}_3$ as a function of c_3 (osmolyte concentration in mass/volume units), the type of preferential behavior can be determined experimentally [49,103,104].

However, caution must be taken when using equation 1.8 to analyze $\hat{V}_2$ as a function of osmolyte concentration(s). When $\hat{V}_2$ vs $c_3\hat{V}_3$ follows a non-linear trend,

multiple mathematical combinations of concentration-dependent G_{ij} functions could potentially lead to this behavior, and no information can be conclusively inferred about the difference (G_{12} - G_{23}) [103,104]. Consequently, only the region where $\hat{V}_2$ vs $c_3\hat{V}_3$ shows a linear dependence near $c_3 = 0$ should be used to infer (G_{12} - G_{23}) and identify the type of preferential interactions for a given protein and co-solute by assuming that G_{12} is constant within measurable uncertainties [49,103,104]. Additionally, equation 1.8 can only be used to analyze ternary solutions since an explicit and analogous expression for mixtures with more than three components had not been reported prior to the present work [59]. This is relevant to most protein solutions of practical interest, as they are usually composed of four or more components (*e.g.*, water, protein, buffer and additional osmolytes such as carbohydrates, inorganic salts, and free amino acids). From that perspective, it is important to obtain a generalized expression that allows one to utilize the behavior of $\hat{V}_2$ for multi-component (greater than ternary) solutions. It is also of interest to test how well a pseudo-ternary approximation works if one seeks to effectively ignore the contributions from the buffer, such as is commonly done in many biophysical chemistry contexts.

1.5 Static Light Scattering for Experimental Quantification of Protein-Protein Interactions

Weak protein-protein interactions and the resulting non-ideal thermodynamic properties can be characterized using static laser scattering (SLS) measurements [29,37]. Here, light emitted from a laser source is passed through a protein sample with a detector located at a defined angle (usually 90°) to measure how much light is scattered by the sample from the incoming laser beam [29]. Using an osmotic system coupled with previous theories developed by Lord Rayleigh and Einstein, Blanco *et al.* showed

that the amount of scattered light scattered from any protein solution (in the form of Rayleigh ratios) can be related to the fluctuations in the number of molecules inside the sub-volume created by the laser beam within the protein solution [29]. These contributions are additive, so one could, in principle, subtract any non-protein contribution to the Rayleigh ratio to obtain excess quantities that correspond to specific protein interactions in the solution. By subtracting all contributions besides that of the protein, one would obtain:

$$\frac{R^{ex}}{K} = \frac{R_{protein} - R_{buffer}}{K} = c_2 M_w \left[\frac{M_{w,app}}{M_w} + G_{22} c_2\right]. \quad (1.9)$$

In equation 1.9, R^{ex} corresponds to the protein excess Rayleigh ratio, $R_{protein}$ and R_{buffer} correspond to the Rayleigh ratios of the protein and the buffer solutions, respectively, K is an instrumental constant, and $M_{w,app}$ is the protein apparent molecular mass. All other quantities were defined above. By computing R^{ex} as a function of c_2 (given the concentration of all other components is constant), B_{22} and $M_{w,app}$ can be estimated assuming both quantities are c_2-independent, and B_{22} = -½G_{22}, but this analysis would only apply for dilute solutions [23,29]. Although B_{22} as a function of solution conditions has been shown to correlate, in some cases, with changes in protein solubility, protein solution viscosity, and protein stability, prior work has highlighted that protein-protein interactions and corresponding solution non-idealities can change qualitatively as one moves to high-c_2 conditions [8,22,23,105]. Here, one could use equation 1.9 to estimate G_{22} values if one knew $M_{w,app}$. The canonical assumption $M_w \approx M_{w,app}$ can be used, and previous work has shown this approximation might be accurate for mAb solutions and other large proteins [23,29]. However, one must be aware that $M_{w,app}$ might significantly deviate from M_w, so *a priori* knowledge of $M_{w,app}$ might be required to accurately use equation 1.9 in such cases.

1.6 Protein Interactions from Other Experimental Scattering Techniques

As mentioned in section 1.1, a variety of techniques can be used to measured B_{22} and G_{22}, or surrogates of these. Dynamic light scattering (DLS) can be used to measure collective diffusion coefficients (D_c) as a function of c_2 [23,26,106]. These measurements can used to calculate diffusion interaction parameters (referred as k_D) using the approximation:

$$D_c = D_0(1 + k_D c_2) \tag{1.10}$$

Here, D_0 is equal to D_c for $c_2 \rightarrow 0$ g/L. k_D values are surrogates of $B_{22,}$ where a linear relationship is expected between the two as follows [23]:

$$k_D = 2B_{22} + \alpha_h \tag{1.11}$$

Here, α_h is used to represent the hydrodynamic contribution to k_D and is always positive. Despite this relationship, recent work has highlighted that α_h might change with formulation conditions [23]. Consequently, unless a value for α_h is known, k_D values represent a convolution between thermodynamic (in the form of B_{22}) and hydrodynamic (in the form of α_h) contributions.

Small angle neutron and X-ray scattering (SANS and SAXS, respectively) can be used to measure the structure factor as function of scattering vector (represented as $S(q)$ *vs* q) for any protein solution of interest [31,81,107]. Although there are notable differences in the execution of these techniques (*e.g.*, SANS usually requires hydrogen-deuterium exchange to provide appropriate contrast while SAXS can be destructive for protein molecules for high-intensity radiation exposure), both techniques have been successfully used to measure structure factors [28,31,77,81]. $S(q)$ is related to $g_{22}(r)$:

$$S(q) = 1 + c_2 \int_0^{\infty} \frac{sin(qr)}{qr} [g_{22}(r) - 1] 4\pi r^2 dr \quad (1.12)$$

where q is the scattering vector and c_2 represents the protein concentration [77,81]. For $q \rightarrow 0$, equation 1.12 converges to $1 + c_2 G_{22}$ *via* equation 1.4, and this would be equivalent to the results obtained from SLS. Unfortunately, $S(q \rightarrow 0)$ is experimentally challenging to measure as the regions of detector that correspond to $q = 0$ directly overlap with the signal obtained from the radiation source [31,77,81]. Additionally, the complex and expensive experimental instrumentations required for SANS and SAXS make these techniques inefficient for measuring G_{22} values in comparison to SLS.

1.7 Protein Unfolding and Aggregation

Protein unfolding is the process by which a natively folded protein changes its initial structure (conformation) and forms additional structures that considerably differ from its native conformation [16,31,108,109]. This can be caused by several external stresses such as high or low temperature and pH fluctuations, the addition of destabilizing cosolutes (such as urea and guanidinium hydrochloride), and surface interactions such as shaking and stirring [12,16,21,24,30,110–112]. Protein aggregation is a multi-step process where unfolded or partially unfolded protein molecules are incorporated into more stable protein complexes formed by two or more protein molecules, simply denoted as protein aggregates [16,113] . This can be caused by the same stresses that induce protein unfolding [14,113]. Both processes are interrelated, as unfolding has been identified as a precursor for aggregation [21,24,114]. Experimental techniques such as circular dichroism, differential scanning calorimetry, differential scanning fluorescence and dynamic light scattering, among others can be used to directly measure or infer changes due to protein unfolding and aggregation

[16,30,31,115]. However, those experimental techniques are performed at finite protein concentrations, so it is experimentally challenging to decouple results for unfolding and aggregation as both might be observed simultaneously [23,30,116]. To achieve this, several orthogonal experiments must be performed to gain insights into the independent contributions between protein unfolding, aggregation and interactions for small systems (*e.g.*, polypeptides) [16,23,30,31,58]. Molecular scale simulations can be used to efficiently define both processes, where one can easily identify the dominant protein configurations present during the simulation and how these might change based on intra- and inter-protein interactions [84,87,117,118]. This could help to further understand the relationship between unfolding, aggregation and weak protein interactions in protein solutions, and provides an additional framework to analyze experimental measurements.

1.8 Coarse-Grained Molecular Modeling for Protein Interactions, Unfolding and Aggregation

1.8.1 Coarse-Grained Molecular Models

Although an all-atom description of proteins is ideal to understand the changes in protein solution behavior caused by the mutation of specific amino acids within the protein sequence, by changing the concentration of any component (*e.g.*, protein, counterions, or added co-solutes) or by modifying the solution pH, this level of description often requires the use of explicit solvent, counterion and co-solute molecules [65,67,80]. Due to differences in time scales of the different phenomena that molecules are subject to, all-atom simulations must be performed by considering the fluctuations of smaller molecules (such as water), which would dominate the computational demands of the simulations [118,119]. Unfortunately, most macroscopic behaviors (*e.g.*, protein unfolding, aggregation and weak interactions) occur orders of magnitude

slower than fluctuations of small molecules [118]. This would make most all-atom simulations inefficient or even intractable in comparison to the time needed to perform equivalent experimental measurements. To address this inefficiency, coarse-graining has been increasingly used to model protein solutions as this decreases the computational demands of a computer simulation [62,66,67,118,120]. Likewise, the solvent molecules (including buffer, counter-ions and co-solutes) are parameterized during coarse-graining, leading to implicit solvent simulations [62,66,67,120]. This is of greater need when calculating G_{22} values at high-c_2 conditions, because doing so requires one to include many protein molecules in the same simulation to capture multi-body correlations at elevated concentrations [61,68,121]. This leads naturally to considering a family of CG models, where one must acknowledge that there is no unique choice for the level of coarse-graining with which to treat a given system.

This CG framework solves the inefficiency issues but might lack enough molecular resolution to accurately capture packing behavior at high protein concentrations and amino acid-level perturbations (such as those cause by chemical denaturation, pH titration or point mutations) [40,63,66]. For instance, Figure 1.3 illustrates the large structural and chemical discrepancy between spherical and atomistic descriptions for a mAb molecule. To compensate for this discrepancy, spherical models have been coupled with surface "patches" to represent the complex surface chemistry of a protein molecule to model protein interactions and oligomerization [82,83,122]. Similarly, amino acid-level description CG models, such as the 1bAA model, have been developed to model both weak protein interactions, and to simulate unfolding and aggregation transitions for a variety of proteins [62,63,66,67,84]. This level of molecular resolution compensates for the lack of structural detail in spherical models at

the expense of longer computational times [67,84,120]. However, most work in the literature focuses on the viability of using CG models retrospectively to gain insight into experimental results - *e.g.*, by fitting analytical models to high-c_2 data. Little to no effort has been invested on proving the potential for this type of molecular modeling to provide accurate *in silico* predictions [12,118,123]. Likewise, it is helpful to employ a systematic approach by which to assess a range of different CG models, once one has decided on the key experimental behaviors of interest for the model(s) to predict [63,67]. To achieve this, optimized computer algorithms are also a key to perform comprehensive comparisons across CG models applied to a variety of protein molecules.

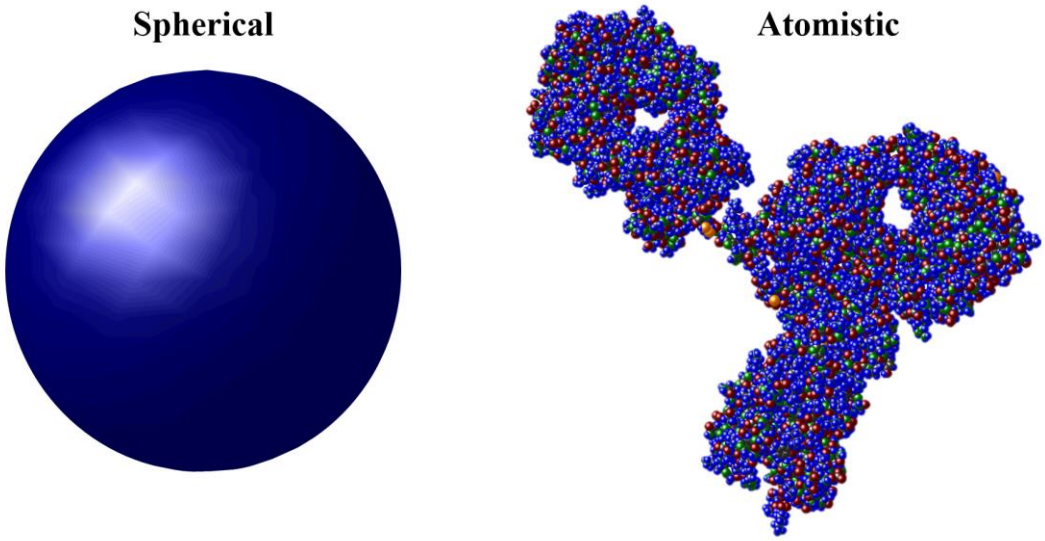

Figure 1.3. Schematic representation of spherical and atomistic descriptions of a mAb molecule. The atomistic description has been color-coded to highlight the complex chemistry of the protein surface.

1.8.2 Computer Algorithms for Predicting Protein-Protein Interactions

Several approaches can be taken to simulate quantities such as B_{22} and G_{22}. For simple spherical models with symmetric interactions, analytical expressions can be numerically integrated using equation 1.1 for low-c_2 conditions, as shown repeatedly in the literature [79,124]. However, more complex computer algorithms are needed to efficiently enumerate asymmetric molecular structures and/or interactions, and for high-c_2 conditions. Monte Carlo (MC) and molecular dynamics (MD) simulations can be used to either compute radial distribution functions that are later integrated such as in equation 1.4, or to directly integrate interaction potentials such as those in equation 1.1 [125–128]. However, previous work has highlighted the short comings of using the first approach (*i.e.*, simulating radial distribution functions to compute properties as shown in equation 1.4) as full convergence ($g(r\rightarrow\infty) = 1$) is required to obtain accurate B_{22} or G_{22} values [126]. In that context, Kofke and coworkers have developed computational algorithms to efficiently compute B_{22} and analogous two-body integrals such as in equation 1.1 [129–133]. This approach is known as the Mayer sampling (MS) algorithm and it corresponds to an umbrella sampling method that allows one to compute B_{22} values based on a known reference (usually a hard-sphere or a Lennard-Jones sphere fluid) [129–133]. This approach has been previously used to simulate protein and protein-like structures at low-c_2 (*i.e.*, only two-body interactions) and shown to provide accurate estimates of B_{22} values comparable to experimentally measured results [63,66].

However, the MS algorithm was never intended to provide results for equation 1.2 (*i.e.*, G_{22}), so additional approaches are needed. This can be achieved by performing simulations that provide the key quantities in equation 1.3. One example is the use of isobaric-isothermal ensembles to simulate values for the isothermal compressibility as a function of c_2, which can be used to compute G_{22} *via* equation 1.3 [134,135]. Likewise,

grand canonical Monte Carlo (GCMC) simulations can be used to obtain the quantitative relation between μ_2 and c_2, which can also be used to compute G_{22} *via* equation 1.3 [68,128,136]. This connects naturally to excess Rayleigh scattering as shown in equations 1.9. Finally, self-consistency across different approaches can be achieved by comparing simulated B_{22} values from MS and G_{22} values from GCMC simulations under dilute c_2, as both quantities are related as c_2 converges towards 0 g/L. Moreover, both quantities can be obtained from SLS experiments.

1.8.3 Computer Algorithms for Modeling Protein Unfolding and Aggregation

To capture both the unfolding and aggregation of proteins in solution, MD simulations are sometimes used to capture fluctuations in secondary, tertiary and quaternary structures of the proteins [62,80,84,137]. To efficiently sample the most thermodynamically relevant configurations, techniques such as thermal annealing are used to avoid getting "trapped" in a single basin of configurations on the energy landscape [128,138]. This relies on temperature steps that allow the simulated system to explore additional configurations at elevated temperatures. Other techniques that rely on the same principle have been proposed to increase sampling efficiency thus reducing computational times. For instance, Replica exchange molecular dynamics (REMD) combines MD and MC approaches to provide efficient sampling to obtain the density of states from the simulation [84,139]. This is achieved by simultaneously simulating replicas of the same system at different temperatures. Low-temperature conditions generate the most stable (minimum energy) configurations while high-temperature conditions prevent the simulation from becoming a small number of stable configurations without exploring the larger energy and conformation space [84,138,139]. REMD relies on simultaneously running several simulations, and this can

considerably increase the computational time. Moreover, REMD can only provide information regarding equilibrium properties since the continuity of the system is broken, so kinetic information is lost in the process [80,138–140]. Regular MD simulations would be needed in this case, but the risk of "simulation jamming" (*i.e.*, when the obtained results are not representative of the entire system but a dominant sub-set) is always present [80,138–140]. Despite its limitations, REMD provides enough information about the transition between folded, unfolded and aggregated protein states, so it becomes useful to identify the presence of folding intermediates that are challenging to characterize experimentally, and that are often thought to be precursor to non-native aggregation.

1.9 Organization of the Dissertation

This dissertation provides a reasonably general framework to quantify weak protein-protein, protein-water and protein-cosolute interactions both computationally and experimentally for multi-component solutions. Additionally, it provides insights into the potential of efficiently using CG models to predict protein-protein weak interactions and unfolding behavior. The remainder of the dissertation is organized as follows.

Chapter 2 presents a computational approach to select a proper CG model to self-consistently capture both the excluded-volume and molecular volume of mAb molecules from low to high protein concentrations. Seven molecular models (six CG models and an all-atom description) are explored to obtain an accurate model to reasonably capture all-atom resolution behavior within acceptable computational times. In particular, that chapter explores the short-comings of using CG models to simulate high-c_2 protein-protein interactions, especially those arising from using spherical

models to model mAb solutions. It also places a bound on effective charge distributions that are physically reasonable for stable protein solutions.

Chapter 3 investigates the viability of using low-c_2 measurements coupled with CG molecular simulations to predict high-c_2 protein-protein interactions arising from electrostatic and non-electrostatic contributions. This is achieved for two mAb molecules by combining models that accurately capture packing constraints (Chapter 2) with simplified interactions across protein molecules. The results highlight the viability of this approach, as well as some short-comings for strongly attractive conditions. Additionally, a framework to compute domain-domain specific interactions is developed based on potentials of mean force for the domains. High-resolution CG models are also used to gain insights into the effects of the local charge distribution of each mAb molecule on their experimental B_{22} values.

Chapter 4 applies the framework developed in Chapter 3 to globular protein solutions. A spherical model with embedded monopole, dipole and non-electrostatic interactions is used to predict high-c_2 interactions from low-c_2 interaction measurements and molecular simulations. The results are discussed from the perspective of short-ranged non-electrostatic and electrostatic interactions, and the effects of pH titration on the electrostatic behavior of the proteins in solution. Additional high-resolution CG simulations are used to illustrate the need for intermediate structural resolution when the charge distribution of the molecule at a given pH is conducive of strong electrostatic interactions.

Chapter 5 reexamines the preferential interaction framework developed by Timasheff and co-workers to experimentally quantify multi-component solutions. This is applied to solutions of a globular protein and a mAb under dilute protein

concentrations (c_2 < 2 g/L) for a series of ionic and non-ionic cosolutes. A general expression for the protein PSV is derived for arbitrarily complex multi-component solutions. Experimental results between ternary and equivalent quaternary solutions are discussed from the perspective of a quasi-ternary solution. The results also indicate non-classical preferential interactions of sugar molecules with both proteins. All-atom simulations are further used to provide a reference framework for evaluating the net-effect of protein-cosolute and protein-water interactions compared to excluded volume contributions.

Chapter 6 focuses on the refinement of a high-resolution CG molecular model to predict unfolding transitions for a series of Ala-rich peptides. The CG model is refined using historical data of a set of Ala-rich peptides and later validated with a new experimental set of peptides. The results are evaluated based on: (1) the effects of single point mutations in the formation of intermediate states and (2) the effects of chain length on the conformational stability of the peptides. Additional simulations are used to evaluate the effects of peptide-peptide interactions in the conformational stability of the peptides to provide better understand the link between protein unfolding and non-native aggregation from an equilibrium thermodynamics stand point. The results indicate the formation of unfolding intermediates for a series of polypeptides with stronger hydrophobic interactions. Finally, the results also show good agreement between experimental and simulated T_m values for some of the tested sequences. This highlight the potential to predict unfolding thermodynamics of polypeptides using tuned CG molecular simulations when aggregation is not present at low temperatures.

Chapter 2

SIMULATING MAB SOLUTIONS FROM LOW TO HIGH PROTEIN CONCENTRATIONS USING COARSE-GRAINED MOLECULAR MODELS WITH DIFFERENT STRUCTURAL RESOLUTIONS

2.1 Introduction

A key challenge in understanding protein-protein interactions from a molecular perspective is the varied length and time scales involved in the behavior of protein molecules in solution [62,66,77,84,141]. This leads to convoluted contributions from mono-molecular events (*e.g.*, unfolding, protein structure fluctuations) and multi-molecular events (*e.g.*, weak protein-protein interactions) in the overall behavior of the proteins in solution. This problem is further enhanced at high protein concentrations (c_2) as multi-body interactions can be affected by the crowding of the protein molecules in solution [81,142–144]. Furthermore, experimentally characterizing highly concentrated protein solutions can be challenging due to the large amount of protein material required, in contrast to low c_2, and current instrumental limits [8,23]. Even with recent advances in instrumentation for measuring viscosity, light scattering, and related methods such as analytical ultracentrifugation, routine measurements of protein-protein interactions at high c_2 remains an outstanding challenge [8]. In some situations, a relatively quick prediction of high-c_2 behavior may be required (for example, due to a change in therapeutic dose or dosing paradigm) with only limited time and resources. This is relevant as current experimental results highlight that protein-protein interactions and their corresponding solution non-idealities can change qualitatively as

one moves to high-c_2 conditions [22,23,43,145]. Therefore, there is an outstanding need for comprehensive models to better understand and predict high-c_2 protein solution behavior while balancing the inherent computational burdens that accompany all-atom simulations (see sections 1.1 and 1.8). In this regard, this chapter illustrates an approach to develop an efficient yet reasonably accurate family of CG molecular-scale models to compute B_{22} and high-c_2 protein-protein interactions of a generic mAb solution that are comparable with experimental results (see sections 1.3 and 1.5). This chapter addresses mAb solutions from a general perspective, with specific cases being treated in Chapter 3. Finally, some of the content in this chapter has been published or included in a peer reviewed journal [146].

2.2 Methods

2.2.1 Coarse-Grained Models

Six coarse-grained (CG) models and an all-atom model were considered in this chapter. Figure 2.1 shows a schematic of five of the CG models and the all-atom model. The other CG model is a single sphere located at the center of mass of the protein as shown in Figure 1.3. The TRIAD, HEXA and DODECA models (nomenclature defined in Figure 2.1 and below) were developed in the present work to resemble the overall shape of a mAb molecule, and used 3 (TRIAD), 6 (HEXA) or 12 (DODECA) beads per protein. Linker segments between the Fab and Fc fragments are shown schematically in Figure 2.1 for the TRIAD, HEXA, and DODECA models. Rigid-body distances, such as distances between the center of mass of each bead, were selected by comparison with existing crystal structures (PBD: IGGY and 1IGT, and reference [147]), and the domain and subdomain centers-of-mass in those protein structures. Figure 2.2 shows illustrative

schematic diagrams of the relevant distances and the locations of each unit for the TRIAD and HEXA models. The two Fab units were treated identically, so for the sake of parity the Fc-Fab distances were treated as equal for both fragments despite small differences in published structures.

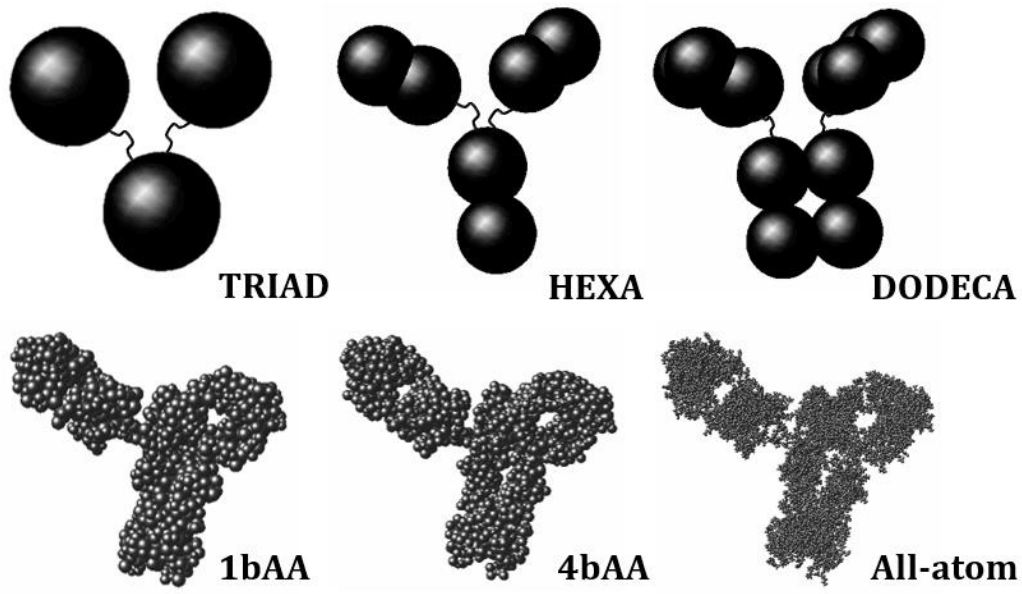

Figure 2.1. Illustrative coarse-grained model depictions. Structures are semi-quantitatively shown to scale to show the level of detail and anisotropy of each model.

The 1bAA (one-bead per amino acid) model is the same as the model by Blanco *et al.*, where each amino acid is treated as a single spherical bead located at the geometrical center of each amino acid [66]. The 4bAA (four-bead per amino acid) model is the same as the model by Bereau and Deserno, except that the protein chain within each domain was not flexible [62]. Similarly, each domain is a rigid-body in the

1bAA model. The 4bAA model treats each amino acid as four independent spherical beads, with one bead each for: the amide group; the alpha carbon; the carbonyl group, and the side chain. The spherical, TRIAD, HEXA and DODECA models are meant to represent the lower limits of molecular detail. Conversely, the 1bAA and 4bAA models represent a more detailed transition towards the all-atom representation, and provide a direct comparison with previous work. Finally, the all-atom model was based on the crystal structure by Padlan [147].

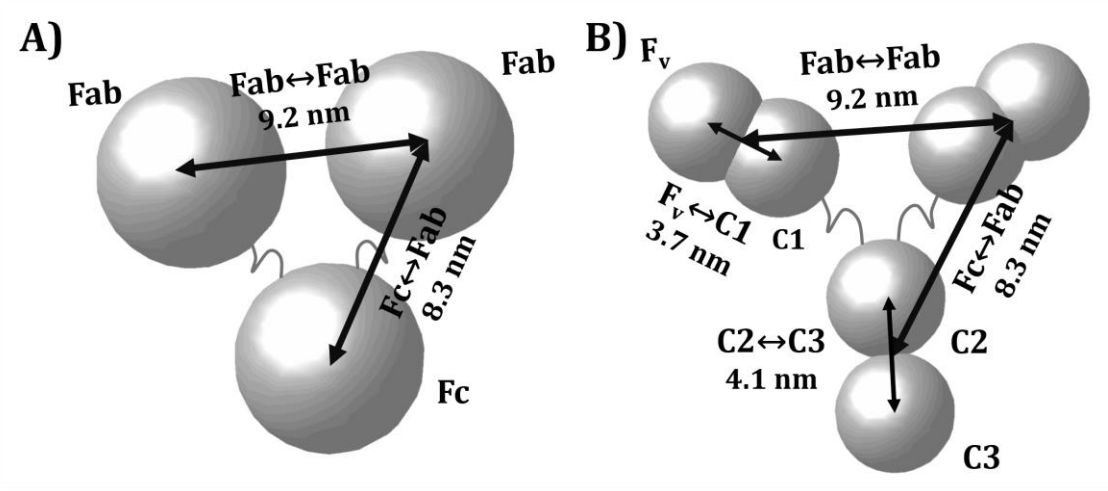

Figure 2.2. Schematic diagrams of the TRIAD (A) and HEXA (B) models, including geometric constraints listed in Table 2.1. The solid connectors between Fab and Fc domains indicate the rigid Fab-Fc linkers listed in Table 2.1.

Table 2.1 lists the values of the distances between domains or subdomains determined from the crystal structure for the TRIAD, HEXA and DODECA models, with spheres centered roughly on the center-of-mass of the corresponding chain(s) in the crystal structures, as those positions were not exactly the same in different crystal

structures. For the DODECA model, the Fab and Fc fragments were each modeled with 4 beads, two for each light or heavy chain subdomain (*e.g.*, V_L, V_H, C_L, C_H1, C_H2, and C_H3). The V_L-V_H, V_L-C_H1, V_H-C_H1 and C_H1-C_H1 distances were the same in each Fab, and the C_H2-C_H3, C_H2-C_H2 and C_H3-C_H3 bead distances were the same for each heavy chain, as shown in Table 2.1 and Figure 2.2 [147]. This was done for the sake of simplicity and as a first attempt to develop a DODECA model. Additional details of how model parameters were selected or refined are given below, including the selection of bead diameters, and choices for the average Fab-Fab spacing.

Table 2.1. Geometric parameters for TRIAD, HEXA, and DODECA models

Model	Fc-Fab (nm)	Fab-Fab (nm)	F_v-C1 (nm)	C2-C3 (nm)	σ_{bead} (nm)
TRIAD	8.3	9.2[†]	-	-	6.1 nm
HEXA	8.3	9.2[†]	3.7	4.1	4.3 nm
DODECA	8.3	9.2[†]	3.5[‡]	4.0[*]	3.5 nm

[†]Distance corresponds to the equilibrium distance (R_{eq}) or rigid-body value
[‡]Equal values for the V_L-V_H, V_L-C_H1, V_H-C_H1 and C_H1-C_H1 distances for the DODECA model
[*]Equal values for the C_H2-C_H3, C_H2-C_H2 and C_H3-C_H3 distances for the DODECA model

2.2.2 Steric Contributions

For calculating the steric contributions to protein-protein interactions at low and high c_2, the steric interactions were modeled with a classical hard-sphere (HS) potential for each bead [148]:

$$u_{ij}^{ST} = \begin{cases} \infty, & if\ r_{ij} \le \sigma_{ij} \\ 0, & otherwise \end{cases} \quad (2.1)$$

u_{ij}^{ST} is the interaction potential for the steric contribution to w_{22} in equations 1.1 and 1.2, r_{ij} represents the distance between the centers of beads i and j, and σ_{ij} is the average diameter between the i-j pair, defined as $\sigma_{ij} = ½(\sigma_{ii}+\sigma_{jj})$. Interactions were treated in the standard manner, as pairwise-additive to account for simultaneous interactions between

multiple proteins at high c_2 [128,148]. Different σ_{ij} values were used for each model. The all-atom model used values proposed in reference [149], including explicit hydrogen atoms when available. For the 4bAA model, diameters were based on prior work: 3.7 Å for the alpha carbon bead; 3.5 Å for the carbonyl group bead; 2.9 Å for the amide group bead; and 5 Å for each side chain bead, with the exception of glycine, which was not assigned a side chain bead [62,84]. The 1bAA model used the values reported by Blanco *et al.* (see Table 3.2) [66].

For the TRIAD, HEXA and DODECA models, each bead diameter (σ_{bead}) in a given model was given the same value, and the following refinement method was used to obtain the particular numerical value of σ_{bead} for a given model. The steric-only value of the second osmotic virial coefficient ($B_{22,ST}$) was first determined for the all-atom model using the Mayer sampling with overlap sampling method (see section 2.2.3). For each of the TRIAD, HEXA, and DODECA models, the value of $B_{22,ST}$ was computed for a series of equally-spaced σ_{bead} values until values were found that bracketed the $B_{22,ST}$ value from the all-atom model. A simple linear interpolation of the σ_{bead} values was used to match the all-atom result. This procedure was iterated, as needed, until $B_{22,ST}$ for the TRIAD, HEXA, or DODECA model and the all-atom model agreed within at least three significant figures. This procedure was unnecessary for the single-sphere model. In that case, σ_{sphere} was simply scaled to match the $B_{22,ST}$ value from the all-atom simulation based on the analytical result for $B_{22,ST}$ of a sphere:

$$\sigma_{sphere} = \left(\frac{3}{2\pi} B_{22,ST}^{all-atom}\right)^{1/3} \tag{2.2}$$

The resulting σ_{bead} and σ_{sphere} values for each of the CG models that were refined here are given in Tables 2.1 and 2.2. Based on this approach, the excluded volume contribution to protein-protein interactions in dilute solution for each of these models

will necessarily be equal to the all-atom model. Work in the literature has shown that the 1bAA and 4bAA models have $B_{22,ST}$ values that are within a few percent of all-atom models for globular proteins, but this has not been extended to monoclonal antibodies, and is a new result reported in this dissertation [63].

Table 2.2. Effective hard-sphere diameter (σ_{eff}) and molecular volume (*via* $B_{12,ST}$) as a function of molecular detail.

Model	# Beads	σ_{eff} (nm)	$v_2 = 2B_{12,ST}$ (mL/g)
Spherical	1	10.45*	2.72
TRIAD	3	10.45*	1.60 ± 0.02
HEXA	6	10.45*	1.27 ± 0.01
DODECA	12	10.45*	1.168 ± 0.004
1bAA	1340	10.61 ± 0.05	1.036 ± 0.002
4bAA	5360	10.46 ± 0.05	0.94 ± 0.01
All-atom	20460	10.45 ± 0.05	0.93 ± 0.01

*Values were matched to the all-atom result.

2.2.3 Mayer Sampling with Overlap Sampling

For any of the models described above or in subsequent sections, B_{22} values were calculated using the Mayer sampling (MS) method employing the overlap sampling algorithm (MSOS) developed by Kofke and coworkers [130–132]. The MS algorithm involves a biased sampling where configurations that do not contribute significantly to B_{22} (*i.e.*, w_{22} values near zero) are avoided. For anisotropic molecules and/or interaction potentials, work elsewhere has illustrated that the location of the reference hard sphere, as well as its size, were relevant to obtain an accurate B_{22} value if a single sphere is used as the reference [63,66,130–132]. Schultz and Kofke developed the MSOS algorithm in part to address this dependence [131,132]. In this work, it was found that although the location of the reference hard sphere was irrelevant as long as it overlaid with some

portion of the protein structure, its diameter (σ_{ref}) still influenced the $B_{22,ST}$ values unless one selected a σ_{ref} value below 7 nm as shown in Figure 2.3. The MSOS approach also provided improved efficiency compared to the MS counterpart. A minimum of 10^8 MC attempts were needed to obtain less than 1% uncertainty in the converged value of $B_{22,ST}$ for models such as 1bAA or 4bAA [63]. Conversely, only 10^5 MC attempts per MSOS run were needed to reach the same level of accuracy as shown in Figure 2.4.

MSOS simulations were performed at constant temperature (298.15 K) with 10^6 MC attempts for both the reference system and the model of interest. Based on the discussion above, a single hard sphere with σ_{ref} = 5 nm located at the center of mass of a given mAb model was used as the reference. Each MC attempt consisted of either a translation or a rotation around the center of mass of the molecule. Translations were performed by selecting a 3D vector from a uniform distribution to translate a randomly selected molecule. Each component of this vector (*i.e.*, the x, y and z components) were drawn from a uniform distribution between *-d* and *d*, for *d* representing the maximum displacement obtained during equilibration (see below) [128]. Rotations were performed using quaternions as explained in reference [150]. The angle and axis of rotation were drawn from a uniform distribution, where the maximum angle of rotation, θ, was set during equilibration (see below). Since the integral in equation 1.1 (*i.e.*, B_{22}) does not depend on the absolute coordinate system but only on the relative distance and orientation between both interacting molecules, the center of the coordinate system was located at the center of mass of one protein molecule, and the other molecule was rotated and translated around the center. The maximum displacement and rotation angle were obtained with a pre-equilibration step consisting of up to 30 cycles of 10^4 MC attempts to obtain an acceptance ratio of 50%. An initial maximum displacement and rotational

angle of d = 1 nm and θ = 50°, respectively, were selected. At the end of each equilibration cycle, the average acceptance ratio of the finishing cycle was compared with the threshold (50% in the present work). d and θ were decreased if the acceptance ratio was below the threshold, or increased otherwise [128,131].

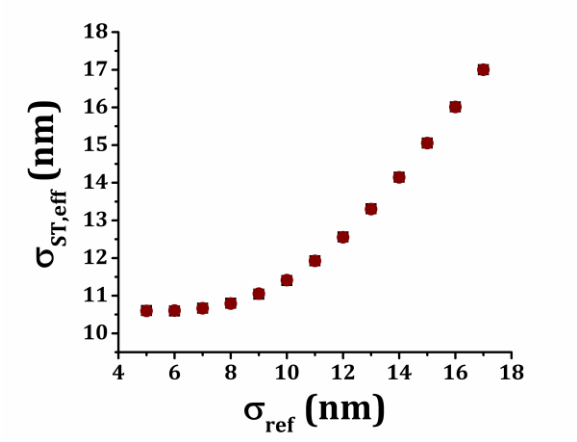

Figure 2.3. Effective hard-sphere diameter as a function of the reference particle diameter. Errors are smaller than the symbols.

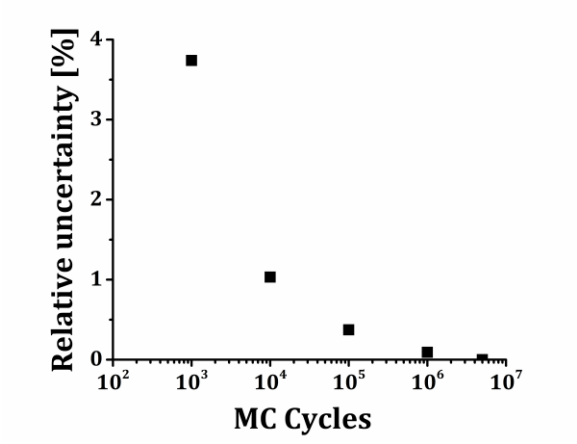

Figure 2.4. Relative uncertainty (uncertainty/average) as a function of the number of MC cycles for $B_{22,ST}$ calculations of the 4bAA model using the MSOS algorithm.

The $B_{22,ST}$ values for the spherical, TRIAD, HEXA and DODECA models were scaled to match the all-atom results (see preceding subsection). As such, $B_{22,ST}$ was not useful to distinguish the models further. The CG models were next compared in terms of the steric contribution to the water-protein second osmotic virial coefficient, $B_{12,ST}$, because that quantity plus the molecular shape are expected to be key features that determine how well proteins "pack" in more concentrated solutions [96,142]. As proteins are strongly solvated by water molecules, short-ranged non-electrostatic protein-protein interactions often involve at least partial overlap between solvation shells, and it is difficult to completely remove the solvation layer except in the case of strong, specific protein-ligand "lock-in-key" binding interactions [44,57,151,152]. Consequently, $B_{12,ST}$ is a reasonable estimate of the effective molecular volume of a protein in water. The mathematical definition of B_{12} is:

$$B_{12} = -\frac{1}{2}\int_r \int_{\Omega_1} \int_{\Omega_2} \left[\exp\left(-\frac{w_{12}\left(c_{1,2} \to 0, r, \Omega_1, \Omega_2\right)}{k_B T} \right) - 1 \right] dr d\Omega_1 d\Omega_2 \qquad (2.3)$$

which shows that the interaction potential is between a water molecule (subscript 1) and a protein molecule (subscript 2). w_{12} represents the potential of mean force between water and protein molecule evaluated at infinite dilution of both components (*e.g.*, only one water and one protein molecule interacting at any given time). This focuses on the excluded volume of a protein with respect to water, without treating the multi-body effects of correlations between hydration "layers" of water near the protein surface. This is in analogy with how surface areas for proteins are often estimated [153]. In equation 2.3, w_{12} is explicitly a function of the center-to-center distance between two protein molecules (r), their relative orientations (Ω_1 & Ω_2) and the solution environment. k_B is Boltzmann's constant and T is the absolute temperature.

The value of $B_{12,ST}$ was determined for a given protein model by using a single "water-sized" hard-sphere with a diameter of 3 Å as component 1. In this case, the same methodology for $B_{22,ST}$ was applied with a slight modification to the reference system. For the protein at the center of the simulation box, the reference was the same as described above. For the moving protein particle, a "water-sized" hard-sphere at the center of mass of the moving protein particle was the natural choice as the reference system. Consequently, these simulations returned values $B_{22,ST}/B_{12,ST}$. $B_{12,ST}$ was calculated by using $B_{22,ST}$ values explained above. Statistical uncertainties for simulated $B_{22,ST}$ and $B_{12,ST}$ values were estimated by performing 5 independent simulations for each model. The standard deviation was used as the estimate of statistical uncertainty, including error propagation. For the spherical, TRIAD, HEXA and DODECA models, a single CPU core was used to execute the MSOS algorithm. Parallel computing (up to ten CPU cores) was used for the 1bAA, 4bAA and all-atom simulations. Consequently, in order to accurately compare CPU times, a calibration curve for the CPU time as a function of the number of cores was obtained from short simulations and all reported values are based on extrapolation to a single-core simulation. Simulations were executed using an Intel® Xeon® E5-2687Wv3 machine (3.1 GHz, 64-bit).

2.2.4 Short-Ranged Non-Electrostatic Attractions, Electrostatic Interactions and Hinge Flexibility

Short-ranged non-electrostatic attractions were modeled with a modified Lennard-Jones potential:

$$u_{ij}^{SR} = 1.3463\ \varepsilon_{SR}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{128} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right] \tag{2.4}$$

while electrostatic interactions were modeled with a modified Yukawa potential:

$$u_{ij}^{EL} = \frac{\alpha^2 q_i q_j}{r_{ij}/\sigma_{ij}} \exp\left[-\kappa\left(r_{ij} - \sigma_{ij}\right)\right] \quad (2.5)$$

u_{ij}^{SR} is the short-ranged non-electrostatic attraction potential, ε_{SR} is the strength of the interaction, u_{ij}^{EL} is the electrostatic potential, q_i and q_j are the valences (also denoted as charges in the remainder of this chapter) of bead *i* and *j* respectively, $1/\kappa$ is the screening length, and α was used as a scaling factor for any q_i to account for the difference between theoretical and effective charges in solution [51,154–158]. α is used in the present chapter instead of ψ_i (see Chapters 3-4) to highlight the use of a different mathematical equation to represent charge-charge interactions in this chapter. ε_{SR} values were kept constant for each bead for a given model and a given round of simulations.

In the case of mAbs, changes in Fab-Fc-Fab angles are more pronounced than fluctuations in Fab-Fc distances [34,107,159–162]. Consequently, the flexibility of the hinge region was imparted *via* a simple harmonic function:

$$u_i^F = k_f\left(R - R_{eq}\right) \quad (2.6)$$

where u_i^F is the flexible potential, k_f represents an effective spring constant, R is the distance between the centers of mass of the two Fab units, and R_{eq} is the equilibrium distance. Equilibrium distances corresponded to representative values of the distances between Fab centers-of-mass found in published mAb crystal structures, which corresponds to an average Fab-Fc-Fab angle of 68°. An additional constraint was also imposed: the distance between the center of mass of each pair of Fab and Fc units (Fc-Fab) was always constant (see Table 2.1). k_f values of 0 (fully flexible), 1, 10 and 100 k_BT/nm^2 were used. With the additional constraint, a k_f value of 0 k_BT/nm^2 allowed a uniform distribution of angles between 44° (Fab-Fab steric overlapping) and 180° while a value of 1 k_BT/nm^2 allowed a Gaussian distribution of angles between 50° and 90°, a

value of 10 k_BT/nm^2 between 60° and 76°, and a value of 100 k_BT/nm^2 between 66° and 70° at T = 298.15 K. This approach is intended to roughly assess the effect of a simple flexible model on the high-c_2 behavior of mAb solutions [107,160–162]. This allows for extreme cases where the Fab-Fc-Fab angle is even greater than those reported in the literature that may approach values as large as 90°, to provide the most conservative estimates of whether hinge flexibility might influence measurable light scattering (osmotic compressibility) values [34,107,147,159–164]. Finally, the potential of mean force was a linear combination of each of the contributions:

$$w_{22} = \sum_{i \neq j} \left(u_{ij}^{ST} + u_{ij}^{SR} + u_{ij}^{EL} \right) + \sum_{i} u_i^F \tag{2.7}$$

where the summations run over indices i and j that denote the beads or atoms on neighboring proteins. Only two protein molecules at a time were used in each MSOS simulation.

2.2.5 Transition Matrix Monte Carlo Simulations and Calculated Excess Rayleigh Scattering

Transition matrix Monte Carlo (TMMC) is a biased MC sampling algorithm where the probability (Π) of the number of protein molecules (N_2), denoted as $\Pi(N_2)$, can be computed by reconstructing a transition probability matrix in a grand-canonical ensemble. TMMC efficiently provides the protein concentration (or number density) as a function of protein chemical potential [68,92,165,166]. That quantity is used here to predict excess Rayleigh scattering behavior by evaluating G_{22} through the simulated osmotic compressibility using histogram reweighting techniques (see equation 1.3) [68,136,165–168]. The distribution $\Pi(N_2)$ can be calculated for an arbitrary choice of protein chemical potential *via*

$$\ln \Pi(N_2|\mu_2) = \ln \Pi(N_2|\mu_o) + \frac{(\mu_2 - \mu_o)}{k_B T} N_2 \quad (2.8)$$

were μ_2 is the desired protein chemical potential and μ_o is the protein chemical potential used to simulate a reference $\Pi(N_2)$. The average values of N_2 (*i.e.*, the protein concentration) and G_{22} then follow from $\Pi(N_2)$ for a given value of μ_2 and applying either form of equation 1.3. The subscript "2" in equation 2.8 indicates that the simulated particles correspond to the protein in an implicit solvent.

The methodology proposed by Errington was used for the TMMC simulations, with an initial uniform distribution for $\Pi(N_2)$ and subsequent updates at the end of each cycle, with each cycle being defined as 10^6 MC attempts [68]. An MC attempt consisted of randomly selecting one of the following for a given protein: a rigid-body rotation, a translation, a vibration (due to the hinge flexibility), or a molecule insertion or deletion. Rotations and translations were performed as explained in section 2.2.3. Molecule insertions and deletions were perform using the methodology explained by Frenkel and Smit [128]. Regular movements (translation, rotations and vibrations) represented 30% of the total movements while deletions and insertions represented the other 70%. Temperature was kept constant at 298.15 K. Preliminary simulations were used to find an adequate reference protein chemical potential value depending on the choice of model parameters and CG model. A box length of up to 120 nm was used and the simulation box was started with an empty system. The final scaled R^{ex}/K and $S_{q=0}$ values were obtained by using equation 1.9 with a M_w value of 146.5 kDa and assuming that $M_{w,app}$ and M_w are equal [23,29]. Statistical uncertainties were estimated by performing 3 independent simulations, and were found to be less than 2% for 10^3 or more MC cycles, therefore all simulations consisted of at least $3x10^3$ MC cycles. Additionally, the CPU time was recorded once the simulation had reached at least 80% flat sampling in

terms of the visited-states histogram, and only the maximum observed CPU time was reported when comparing computational burdens between different models.

2.2.6 Parameter Mapping

To evaluate the sensitivity of the different parameters describing protein-protein interactions at low and high c_2, the following methodology was applied. Initially, the effect of the short-ranged non-electrostatic attraction strength (ε_{SR}) was studied by computing B_{22} at different values of ε_{SR} for the spherical, TRIAD, HEXA and DODECA models. The parameters of the 1bAA, 4bAA and all-atom models were not evaluated as those were set previously at low c_2, and the models are computationally intractable at high c_2. Values of ε_{SR} that returned physically realistic B_{22} values were used in calculations to then evaluate the effect of the charge distribution in each CG model. In this case, B_{22} was computed as a function of two parameters: (1) the charge of the variable region of an IgG1 molecule (Q_{Fab} for TRIADs, Q_{Fv} for HEXAs, and Q_{VH} and Q_{VL} for DODECAs); (2) the ratio of the effective net charge in solution and the theoretical net charge (α), where the net charge is the sum of charges over all beads in the CG model. The theoretical net charge (Z_{thry}) for a given pH value was calculated based on one of the available mAb sequences (PDB: 1IGT), and a target pH value of 5 was selected for illustrative purposes. When calculating charge values for each bead for use in equation 2.5, the charge based on sequence (and pH) was scaled by α, which determines the ratio of the effective charges in solution (*e.g.*, due to ion clustering [154,156,158]) and the theoretical charges obtained from the sequence alone (see below). This approach was used because mAb sequences are highly conserved except for the variable regions. As such, only the charge on the variable region is essentially arbitrary unless one decides to chemically modify the conserved domains. In addition,

the net charge on a protein is a function of not only its amino acid sequence and the pH, but also the ionic environment (*e.g.*, chemical identity and concentration of counterions). The simplest approach to account for the difference between the true net charge in solution (Z_{eff}) and the theoretical net charge (Z_{thry}), is to assume any specific-ion effects are proportional to the local surface charge and therefore $Z_{eff} = \alpha Z_{thry}$. Surface response plots of B_{22} *vs* $Q_{j,thry}$ and α are reported below, where j represents the location (domain) for the charge. Finally, B_{22} values were normalized by $B_{22,ST}$ and model parameters corresponding to selected values of $B_{22}/B_{22,ST}$ were used to evaluate the high-c_2 behavior of a given model.

2.2.7 Theoretical Charge Distribution

Theoretical charges were calculated using the standard Henderson-Hasselbalch equation [169]. The total charge was then computed as the sum of the independent charges. For illustrative purposes, the PDB 1IGT IgG1 molecule was used. For the TRIAD, HEXA and DODECA models, the sequences were partitioned into equal-chain-length units to compute the charge of each bead. For the TRIAD model calculations, the Fab fragments were composed of the upper half of the heavy chain (residues 1-234) and the light chain (residues 1-214), while the Fc fragment was composed of the two lower-halves of the heavy chains (residues 244-474), which excludes the hinge region. For the HEXA model calculations, the F_v domain was composed of the upper half of the light chain (residues 1-107) and the upper quarter of the heavy chain (residues 1-118); the C1 domain was composed of the lower half of the light chain (residues 108-214) and the second quarter of the heavy chain (residues 119-234); the C2 domain was composed of residues 244-357 of each heavy chain; and the C3 domain was composed of the last quarter of each of the heavy chains (residues 358-

474). Finally, for the DODECA model calculations, the heavy chains were portioned into four units (residues 1-118, 119-234, 244-357 and 358-474) and the light chains into two units (residues 1-107 and 108-214) for a total of 12 beads, each with its respective charge. The numerical charge values are shown in Table 2.3 for pH 5 and 7. For all three cases, charges were placed at the geometric center of each bead.

Table 2.3. Theoretical charges for the TRIAD, HEXA and DODECA models at pH 5.0 and 7.0 for an IgG1. The theoretical net charge at pH 5 and pH 7 are 40.8 and 10.2, respectively, for PDB: 1IGT

Model	pH	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
TRIAD†	5.0	12.3	16.2	-	-	-	-
	7.0	4.5	1.2	-	-	-	-
HEXA‡	5.2	2.9	9.4	10.4	5.8	-	-
	7.0	0.1	4.4	2.6	-1.4	-	-
DODECA*	5.0	0.9	3.0	2.0	6.4	5.2	2.9
	7.0	0	0.2	0.1	4.2	1.3	-0.7

†1=Fab, 2=Fc

‡1= F_v, 2=C1, 3=C2, 4= C3

*1=V_L, 2=C_L, 3=V_H, 4=C_H1, 5=C_H2, 6=C_H3

2.3 Steric Contributions to Protein-Protein Interactions *via* B_{12} and B_{22}

From liquid-state theory, it is well established that steric interactions (*i.e.*, "packing constraints") are a key aspect when one considers concentrated systems [59,96,142,144]. Therefore, model comparison and refinements started with considering steric protein-protein interactions at dilute conditions, and then extending to concentrated systems. For dilute solutions, it was computationally tractable to consider all models in Figure 2.1 for refining the models to accurately capture steric repulsions,

while for more concentrated systems a subset of models was considered in order to maintain reasonable computational times.

As noted in the previous sections, the models that were newly developed here (*i.e.*, TRIAD, HEXA and DODECA) each had a single characteristic dimension (σ_{bead}) for their beads or spheres. The previously developed models (*e.g.*, 1bAA, 4bAA and all-atom) had different sets of bead or atom sizes that were set elsewhere [62,66,84]. Considering those previous models first, Table 2.2 lists the value of the effective hard-sphere diameter (σ_{eff}) obtained from equation 2.2 using the simulated $B_{22,ST}$ values as inputs to Figure 2.2 for the 1bAA, 4bAA and all-atom models. The results show that the 4bAA and all-atom σ_{eff} values are equivalent within 0.1%. The 1bAA value is 1.5% larger than the all-atom value, which results in 4.6% higher $B_{22,ST}$ for the 1bAA model when compared with the all-atom model. If needed, this could be addressed easily by slightly reweighting the bead sizes from the original 1bAA model[66]. Those results in Table 2.2 are consistent with previous findings obtained by Grünberger *et al.*, where 1bAA and 4bAA models showed minor differences for a mAb molecule, although the previous work did not include the comparison to an all-atom model for a mAb that is included here [63].

$B_{22,ST}$ was matched to the all-atom value when setting the value of σ_{bead} for the characteristic bead diameter in each CG mAb model that was introduced here. This assured that the different models have the same excluded volume in dilute conditions, and therefore would produce the same reference state for B_{22} for comparison with experimental data. Using the resulting model parameters, the value of $B_{12,ST}$ was calculated as means to estimate the hydrated molecular volume for each model. To a first approximation, the hydrated protein volume is equal to $2B_{12,ST}$. $B_{12,ST}$, along with

the protein geometry, are additional important quantities to assure that the CG proteins will pack at high c_2 in a way that reasonably captures how all-atom structures pack when solvent is present [96,142]. Assuring that $B_{22,ST}$ and $B_{12,ST}$ both mimic the all-atom values for these mAb-like molecule geometries was used as a strategy to form a model framework in which low-c_2 and high-c_2 interactions are captured self-consistently in later calculations that include more than just steric interactions.

Figure 2.5A shows the $B_{12,ST}$ values (main panel) and the corresponding CPU times (inset) as a logarithmic function of the number of beads per model. The results for $B_{12,ST}$ show an asymptotic behavior as the level of coarse-graining or structural detail approaches the atomistic structure. However, the computational burden increases dramatically (power-law behavior) as one increases the level of structural complexity. No significant difference (less than ~ 3 %) was observed between the all-atom, 4bAA and 1bAA models for $B_{12,ST}$. In contrast, the spherical model overestimates $B_{12,ST}$ by almost a factor of 3. As $B_{12,ST}$ is a representation of the effective volume occupied by a single protein in solution, these results imply that spherical models of mAbs that accurately capture low-c_2 behavior in terms of B_{22} will grossly overestimate the protein volume that is relevant at high c_2 where excluded volume "shells" of multiple protein molecules will overlap simultaneously. The agreement between $B_{12,ST}$ for the CG models and the all-atom structure improve greatly as one adopts a mAb-like geometry and increases the number of beads per protein. The results in Figure 2.5A illustrate a compromise between computational costs and molecular-scale structure. If one adopts a HEXA or DODECA representation, $B_{12,ST}$ is within ~20 % of the all-atom value. However, the computational burden for even these dilute calculations is smaller by at least 5 orders of magnitude, compared to the all-atom results.

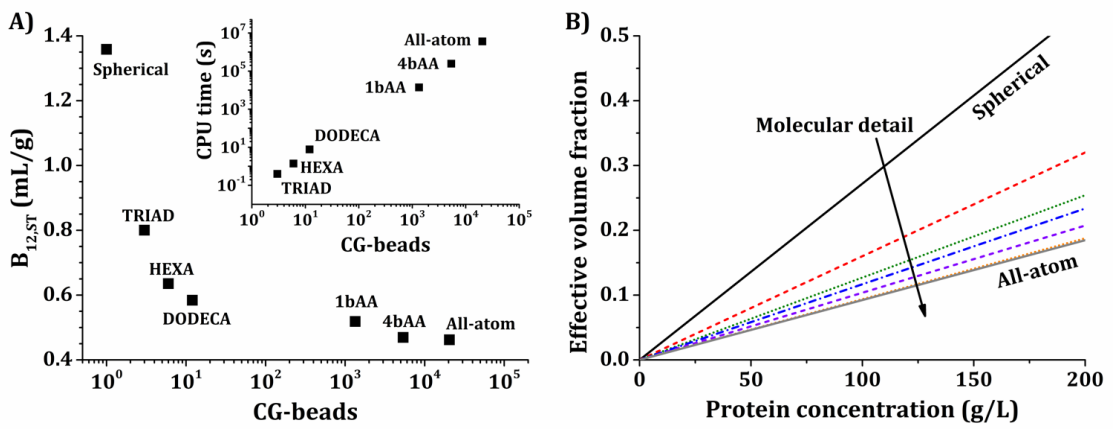

Figure 2.5. **Panel A:** $B_{12,ST}$ and CPU time (inset) as a function of coarse-graining detail for 10^6 MC cycles. **Panel B:** effective volume fraction *vs* protein concentration as a function of coarse-graining molecular detail using $2B_{12,ST}$ as the scaling factor. Labels follow the trend in panel A.

In molecular simulations, it is natural to express concentrations in terms of volume fraction, η, defined as the physical volume occupied by the molecules divided by the volume of the system: $\eta = N_2 v_2 / V$, where N_2 denotes the number of proteins in the system volume (V), and v_2 is the molecular volume of a single protein [128]. The experimentally measured mass concentrations (mass/vol) can be calculated easily from η if one knows v_2 (or an estimate) and the protein molecular weight. In this regard, the values of $2B_{12,ST}$ (set equal to v_2) provide a natural scaling unit for a given model, as they represent an estimate of the physical volume occupied by the hydrated protein (the factor of 2 arises simply by the formal definition of B_{12}, see equation 2.3). Figure 2.5B shows the effect of molecular detail on c_2 (in g/L) calculated from η using $2B_{12,ST}$ as the scaling factor. This shows that using low-resolution CG models in simulations or calculations based on equations of state for spherical models will result in significant

underestimation of the experimental c_2 that corresponds to a given simulated state point, and that error is most pronounced for the spherical model.

The results in Figure 2.5A are only for a simple c_2-independent algorithm like the MSOS, so it could be expected that moving to high-c_2 simulations would result in even greater computational burdens for the higher-resolution structures. As such, unless behaviors that require strong and specific interactions at the amino acid level are essential, the 1bAA, 4bAA, and all-atom models are likely far too computationally burdensome to be practical for moving to high c_2. On the other hand, the spherical and TRIAD models showed such large deviations with respect to the all-atom results that the HEXA and DODECA models may offer an optimum trade-off between computational cost and quantitative accuracy for evaluating the high-c_2 behavior of these models. That will be revisited in Chapter 3.

2.4 Steric Interactions at Elevated Protein Concentrations

High-c_2 behavior was studied by simulating excess Rayleigh scattering profiles based on the osmotic compressibility determined from TMMC simulations for each of the spherical, TRIAD, HEXA and DODECA models. Initially, the TRIAD, HEXA, and DODECA models were simulated as rigid bodies, so no hinge flexibility was included during the sampling (k_f was effectively infinite). Figure 2.6 shows the results of the maximum observed CPU time as a function of molecular detail. Although the CPU time was expected to scale with n^2, where n denotes the number of beads in each model, Figure 2.6 shows a correlation close to $n^{3.4}$. This might be due to the coupling between the increased number of beads ($\sim n^2$) and the more complex energy-landscape or geometric complexity for the steric-only version of the models, because of the higher molecular detail and increased c_2 that results in multi-body interactions. A similar

correlation can be expected for more complicated CG models, as the results showed a consistent scaling when moving from the simple spherical to the DODECA model. Consequently, the use of more complex CG models, such as the 1bAA, 4bAA, or all-atom models, would make this approach computationally inefficient and practically intractable in most cases. Presumably, this scaling can be at least slightly decreased by including optimized sampling methods such as neighbor lists and configurational bias algorithms [61,170]. However, once one adds attractions and long-ranged interactions to the models, the computational burdens for simulating high-c_2 protein solutions will require a large trade-off between structural detail and computational cost.

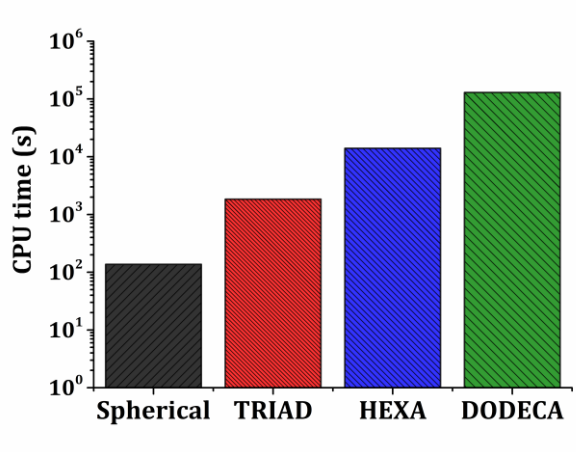

Figure 2.6. CPU time for TMMC simulations as a function of the level of coarse-graining. A flat histogram convergence of 80% was used as a metric for similar convergence between TMMC simulations with different CG models over the same range of protein concentrations.

$S_{q=0}$ as a function of volume fraction (η) was obtained by applying equations 1.3 and 1.9 to the simulated $\Pi(N_2)$ distributions, and is shown in Figure 2.7A from low η to

the highest value for which the TMMC simulations reasonably converged. This represents an unambiguous form of plotting the high-c_2 behavior, as both variables are dimensionless and independent of any reference state. Both forms of equation 1.3 were tested to calculate G_{22}, and no differences between the approaches were found if small steps of chemical potential ($\Delta\mu_2$) were used to compute the isothermal compressibility. Large differences were observed between the spherical model and the DODECA or HEXA model (for example, $S_{q=0}$ is three times larger for spheres than DODECAs or HEXAs at $\eta = 0.1$), with the TRIAD model showing an intermediate behavior between spheres and DODECAs or HEXAs. The results suggest that the packing behavior of spheres and DODECAs or HEXAs differs considerably, while the HEXA and DODECA models are essentially indistinguishable in terms of sterics based on the scattering behavior or static structure factor (Figure 2.7A).

To test whether one can effectively rescale $S_{q=0}$ *vs* η profiles, the dimensionless values from the x-axis of Figure 2.7A were rescaled to dimensional concentration units using the $B_{12,ST}$ values shown in Table 2.2, and the results are shown in Figure 2.7B. Interestingly, all the $S_{q=0}$ results seem to collapse onto similar curves for at least low c_2, once each model's respective $2B_{12,ST}$ is used as the scaling factor. However, one needs to recall that the reported $B_{12,ST}$ values were obtained at a forced constant $B_{22,ST}$ value, so the low-c_2 regime of figure 2.7B necessarily must collapse onto a common curve because the slope of $S_{q=0}$ *vs* c_2 equals $-2B_{22,ST}$ in that limit. As such, the results in Figure 2.7B confirm that the TMMC simulations are internally consistent with the MSOS simulations. Consequently, the different physical behavior observed in figure 2.7A is "corrected" by the inaccurate $B_{12,ST}$ values obtained under the same steric interactions. This indicates that results from spherical models can be numerically adjusted to mimic

behavior obtained from more mAb-like models, but at the expense of an incorrect molecular volume and molecular shape.

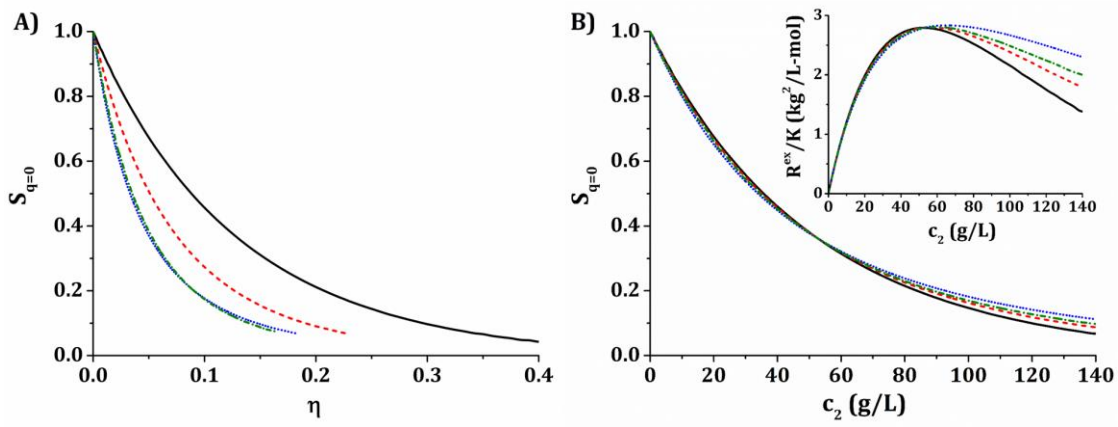

Figure 2.7. Simulated $S_{q=0}$ as a function of volume fraction (η, panel A) or protein concentration (c_2, panel B) after rescaling as described in the main text, for the spherical (solid black), TRIAD (dashed red), HEXA (dotted blue) and DODECA (dash-dotted green) models with steric-only interactions. Inset in panel B corresponds to the excess Raleigh ratio *vs* protein concentration.

To relate back to experimental excess Rayleigh scattering profiles, the inset in Figure 2.7B shows values of R^{ex}/K *vs* c_2. These curves reproduce key qualitative experimental features under repulsive conditions: linear behavior at low c_2, downward curvature at higher c_2, and a turnover (maximum R^{ex}/K) at sufficiently high c_2 [23,105,171]. The quantitative scales of the y- and x-axes also agree with typical experimentally measured values for mAb molecules [23,105,171]. The inset in Figure 2.7B shows that no clear differences can be observed below the turnover concentration, mainly as a result of equal $B_{22,ST}$ and M_w values noted above, and that this turnover occurs around 60 g/L for the four models. However, a 60 g/L concentration of spheres

corresponds to 15% v/v, while the same concentration for DODECAs corresponds to 6.5% v/v as Figures 2.5B and 2.7A show. Additional radial distribution functions of spherical and anisotropic steric-only models showed very different behavior at high c_2 (see reference [23]), indicating large differences in packing behavior as discussed above. This suggests that simulations of mAb molecules based on spherical models might return values that seem to correlate with experimental measurements but lack a reasonable structural behavior of true mAbs in solution, and this will limit their utility as predictive models of high-c_2 behavior. This is also consistent with reports where spherical models have failed to accurately capture the high-c_2 behavior of physicochemical properties and intermolecular interactions of mAb solutions [28,77]. Finally, all of the above shortcomings for steric-only models are expected to be further exacerbated once one includes non-uniform charge distributions and/or short-ranged non-electrostatic attractions. This follows because the packing and location of charges and attractive domains is expected to be important at high c_2, when proteins interact simultaneously with multiple neighboring proteins.

As noted earlier, the computational time of progressing from the HEXA model to the DODECA model was considerable (*i.e.*, scaling with $n^{3.4}$). The same can be expected when moving from the DODECA model to a more structurally detailed CG model. Consequently, if computational efficiency is considered, improvements from the HEXA model to the DODECA model and to more complex models might become unnecessary if only B_{22} and G_{22} or $S_{q=0}$ values are targeted. Based on these considerations, it is proposed that an optimum tradeoff between efficiency and accuracy may be found by further refining CG models that build on a HEXA or DODECA representation, and that reasoning informs most of the examples below.

2.5 Contributions from Hinge Flexibility, Short-Ranged Non-Electrostatic Attractions and Electrostatic Interactions

A flexible hinge is expected to provide easier packing for highly anisotropic models. Consequently, the effect of the hinge flexibility was first evaluated for the predicted scattering and osmotic compressibility behavior with steric-only interactions for the TRIAD and HEXA models. Figure 2.8A shows that a flexible hinge does not significantly affect the osmotic compressibility behavior within the tested c_2 range for the TRIAD model, including the low-c_2 range that is relevant for B_{22} results. Figure 2.8B shows very small differences in the simulated $S_{q=0}$ and R^{ex}/K profiles for the HEXA model as a function of flexibility for simulated c_2 values, and the turnover location is not affected even when the hinge region is so flexible that the Fab domains can move completely unimpeded around the Fc domain (*i.e.*, $k_f = 0$). For both models, these differences became more relevant for values above 140 g/L, suggesting that packing may be significantly affected by the flexible hinge at much higher c_2 but it is indistinguishable up to the threshold concentration used here if one is focused on the osmotic compressibility or related quantities. However, the reader should be aware that this might not hold for small angle scattering measurements where more structural and spatial information can be obtained (see also Chapters 3 and 7) [34,81].

For short-ranged non-electrostatic attractions, the strength of the attractions as well as the range potentially plays a role in the solution behavior of mAb solutions, and this might also be affected by the hinge flexibility. Therefore, the effect of the strength of a short-ranged non-electrostatic attractive potential was tested for the spherical, TRIAD, HEXA and DODECA models. Figure 2.9A shows the results for calculated $B_{22}/B_{22,ST}$ values as a function of the strength of attraction. One observes that $B_{22}/B_{22,ST}$ depends weakly on the strength of the interaction potential (well depth) at low values of

ε_{SR}, but depends strongly on ε_{SR} at higher values. At sufficiently high ε_{SR} values, B_{22} decreases dramatically towards large negative values. Additionally, this downturn in B_{22} values occurs at smaller ε_{SR} values as the number of beads in the model increases. This agrees with previously reported B_{22} values of a range of CG models and is consistent with general statistical-mechanical arguments for short-ranged, anisotropic attractive interactions [63,66,89,92].

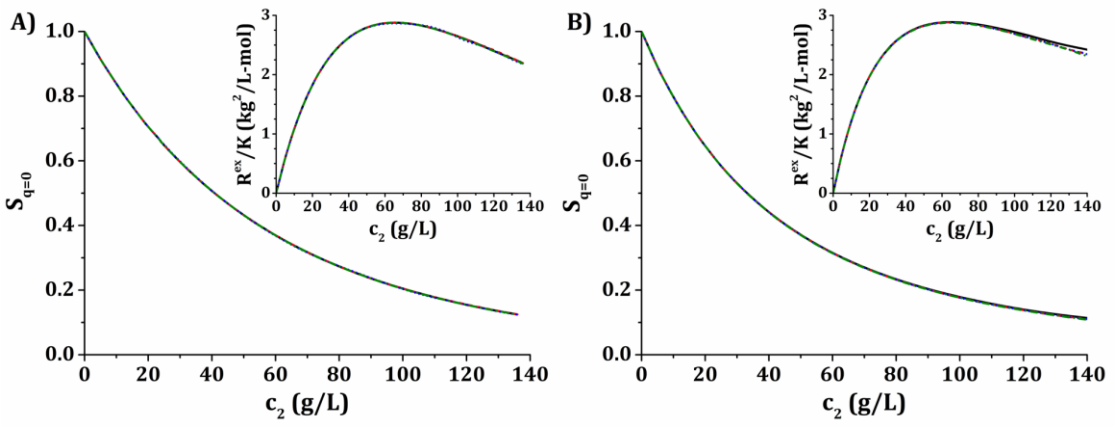

Figure 2.8. Effect of the hinge flexibility on the simulated SLS behavior from low to high c_2 with steric-only interactions for the (A) TRIAD and (B) HEXA models. Results are shown for spring-constant (k_f) values for the Fab-Fab distance that span from infinitely flexible (k_f = 0, dash-dotted green), to increasingly more rigid structures: k_f = 1 (dotted blue), k_f = 10 (dashed red), k_f = 100 (solid black).

Experimental values of $B_{22}/B_{22,ST}$ between the extremes of -10 and 10 have been reported in the literature for a wide range of protein solutions, including mAb solutions, although some methods report values that are only surrogates for B_{22} [17,23,25,29,89,92]. Strongly attractive conditions ($B_{22}/B_{22,ST}$ values approaching -5 to -10) typically result in physically unstable solutions due to irreversible or reversible

protein aggregation for a wide range of proteins [23,50,79,105,172]. It has been observed that $B_{22}/B_{22,ST}$ values converge towards a value between -1 and 0 for mAb solutions at high total ionic strength (~ 300 mM) before highly non-ideal salt effects become important and solutions become unstable at large negative values of B_{22} [23,105,171]. Under those conditions, electrostatic interactions are effectively screened, and attractions are expected to be caused by hydrophobic interactions and van der Waals forces [105]. $B_{22}/B_{22,ST}$ values in the range of -1 to 0 are obtained with the present models for ε_{SR} values between 1.35 k_BT and 2 k_BT for the TRIAD model, 1.2 k_BT and 1.5 k_BT for the HEXA model, and 0.5 k_BT and 0.7 k_BT for the DODECA model.

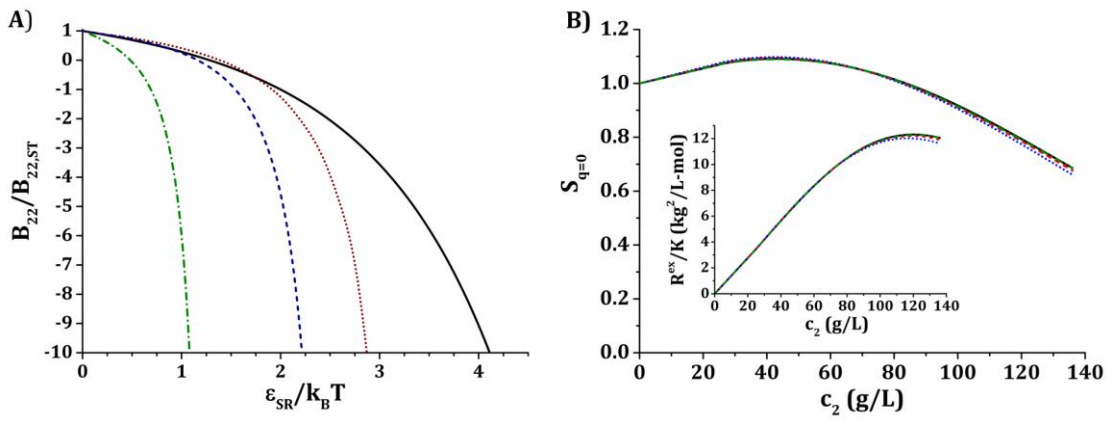

Figure 2.9. **Panel A:** $B_{22}/B_{22,ST}$ as a function of the short-ranged attraction parameter ε_{SR} for spheres (black solid line), TRIADs (red dotted line), HEXAs (blue dashed line) and DODECAs (green dashed-dotted line). **Panel B:** effect of the hinge flexibility as a function of c_2 for the TRIAD model for ε_{SR} = 1.5 k_BT. Lines represent values of k_f = 0 (green), 1 (blue), 10 (red) and 100 (black).

A value of 1.5 k_BT was used to illustrate the effect of hinge flexibility at high c_2 for the TRIAD model with added short-ranged non-electrostatic attractions as shown in

Figure 2.9B. The added attractions produced slightly different $S_{q=0}$ values for a given c_2 than the steric-only behavior in Figure 2.8. The inset to Figure 2.9B shows R^{ex}/K values are slightly but visibly higher at the highest c_2 values in comparison with the steric-only behavior. Interestingly, the flexibility of the hinge caused no significant differences in the osmotic compressibility (*e.g.*, $S_{q=0}$ and R^{ex}/K) values for $c_2 < 140$ g/L, in agreement with the steric-only behavior in Figure 2.8. This suggests that the added flexibility does not considerably affect the overall packing behavior even at those elevated c_2 with short-ranged attractions present. However, the CPU times were much larger for lower k_f (more flexible) models. This motivated simplifying the models to neglect hinge flexibility in what follows below, where additional interactions are included. In the case of the HEXA and more complicated CG models, the hinge flexibility might still show a relevant effect, but the extension of that analysis to other CG models and to more spatially detailed results (such as small angle scattering) was beyond of the scope of the present work and will be discussed in Chapters 3 and 7.

The effect of adding short-ranged non-electrostatic attractions coupled with electrostatic interactions was then evaluated for the TRIAD, HEXA and DODECA models, as those models have a sufficiently small set of parameters that it was practical to perform a reasonably global parameter search. In what follows, the effects of the model parameters for charge-charge interactions are determined primarily for low-c_2 conditions. This is done because most experimental reports provide quantitative scattering results to compare against at only low-c_2 conditions. Effects of electrostatics on high-c_2 behavior are provided only for illustration purposes, as a detailed investigation of electrostatic interactions at high c_2 will be provided in Chapter 3 that builds from the model foundations set here.

In calculations that included electrostatic interactions, the modified Yukawa model was used with charges located at the center of each bead. Since both Fab fragments are chemically identical, the TRIAD model only has two different charge values: one for the Fc and an equal charge for each Fab bead. Similarly, the HEXA model has four charge values following that representation, while the DODECA has six charge values. As a result, a rigid TRIAD model with sterics, short-ranged non-electrostatic attractions and screened electrostatic interactions has five model parameters: ε_{SR}, α, κ, Q_{Fc}, Q_{Fab}. The latter two will depend on the pH and protein sequence and any territorial counterions, while the inverse screening length (*i.e.*, κ) is a function of pH and buffer salt type/concentration [154,156]. A rigid HEXA model adds two additional parameters for a total of seven: ε_{SR}, α, κ, Q_{Fv}, Q_{C1}, Q_{C2} and Q_{C3}, while a rigid DODECA adds two more for a total of nine: ε_{SR}, α, κ, Q_{VL}, Q_{CL}, Q_{VH}, Q_{CH1}, Q_{CH2} and Q_{CH3}. Although additional details can possibly be added to model the short-ranged non-electrostatic attractions and electrostatic contribution, such as different ε_{SR} values or dipoles for different domains, the present models already provide a relatively large parameter space for mapping the global model behavior. Therefore an approach based on experimental observations and general theoretical considerations was used to assess the parameter space in which these models would produce physically realistic behaviors, similar to what was done previously for globular proteins in the absence of electrostatic interactions [63].

It was discussed above how to obtain ε_{SR} values that would return experimentally reasonable observations of $B_{22}/B_{22,ST}$, and a similar approach was taken for the remaining parameters. κ is related to the total ionic strength of the solution (*TIS*) *via* $TIS = 92.42\ \kappa^2$, where *TIS* is given in mM units and κ in nm^{-1}, which assumes any salt

behaves as a 1:1 electrolyte in water at 298.15 K [148]. *TIS* affects electrostatic interactions *via* charge screening, and helps to mediate the balance between short-ranged non-electrostatic attractions and electrostatic interactions. As an example, Figure 2.10 shows predicted values of $B_{22}/B_{22,ST}$ for a model IgG1 at pH 5 using the HEXA model, with the charges shown in Table 2.3 with α = 0.1 and 1 for the purpose of illustration.

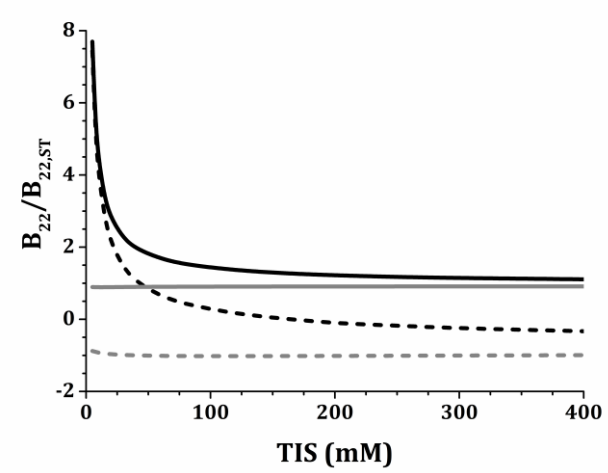

Figure 2.10. $B_{22}/B_{22,ST}$ as a function of *TIS* for the HEXA model with ε_{SR} = 0 (solid line) and 2 k_BT (dashed line), and for charges in Table 2.3 at pH 5 with α = 1 (black) and 0.1 (grey).

As expected, the results show that ε_{SR} is more relevant at high *TIS*, where charges are highly screened. At low *TIS*, the electrostatic contributions are the dominating effect unless the pH is such that the net charges on each domain are relatively small (close to the isoelectric point). Thus, if ε_{SR} and κ are known or assumed, then three degrees of freedom (α, Q_{Fc} and Q_{Fab}) would be left for the TRIAD model, and similarly five degrees of freedom (α, Q_{Fv}, Q_{C1}, Q_{C2} and Q_{C3}) for the HEXA model, and seven (α, Q_{VL}, Q_{CL},

Q_{VH}, Q_{CH1}, Q_{CH2} and Q_{CH3}) for the DODECA model. In this context, one can evaluate the effect on predicted B_{22} of different charge distributions as a method to map out the parameter space that might result in experimentally tractable values. As the variable region of a family of mAb is typically the only engineered or modified portion between two different sequences, only Q_{Fab} for the TRIAD model, Q_{Fv} for the HEXA model, and Q_{VL} and Q_{VH} for the DODECA model were considered as variable parameters in what follows for illustrative results assuming the constant domains were as shown in Table 2.3. Theoretical charges can be obtained from the sequence and pH, and α can be used as a scaling factor to correct for any territorial counterions [154,156]. This would produce B_{22} as a function α and any modified charge (*e.g.*, from point mutations or chemical degradation of charged residues) to study the effect of electrostatic phenomena in the solution behavior.

Figure 2.11 shows an example of a three dimensional surface plot of calculated $B_{22}/B_{22,ST}$ versus a range of realistic model parameters for the TRIAD (panel A), HEXA (panel B) and DODECA (panels C and D) models at low *TIS* where electrostatic interactions are important for determining osmotic virial coefficient values. ε_{SR} was set to: 1.85 k_BT for the TRIAD model; 1 k_BT for the HEXA model; and 0.85 k_BT for the DODECA model. These particular values for ε_{SR} were selected because they resulted in a $B_{22}/B_{22,ST}$ value of -0.5 for each respective model [23]. The theoretical charges were obtained from assuming pH = 5, typical of solution formulations of therapeutic antibodies while avoiding fractional charges on acidic amino acids for the calculations, (see Table 2.3), and using the sequence and structure from the 1IGT PDB [10,23]. α is a scaling factor that represents a deviation from purely mean-field models of screening with no specific-ion effects (*i.e.*, $\alpha = 1$).

Figure 2.10 illustrates that if one increases ε_{SR} for a given charge distribution, this will simply shift the results in Figure 2.11 uniformly to less positive or more negative values of B_{22}. Similarly, increasing *TIS* or decreasing α will bring $B_{22}/B_{22,ST}$ closer to values in Figure 2.9A (*i.e.*, dominated by the short-ranged non-electrostatic attractions), as any charge-charge interaction will be screened out and those contributions to the potential of mean force would be negligible. In Figure 2.11 (all panels), it is notable that the decay towards more negative (attractive) $B_{22}/B_{22,ST}$ values (yellow and white regions) occurs within a rather small window of possible charge distributions that place negative charges on the variable domains, and therefore create strong attractions with the positively charged conserved domains. This is consistent with the results from Figure 2.9A, which show a much faster decay towards lower $B_{22}/B_{22,ST}$ values as the strength of the attractions increases. Consequently, attractive conditions are more sensitive to small perturbations of model parameters, which can lead to phase separation or gelation due to strong attractions. This is in good agreement with experimental observations for protein solutions with strong attractions that result in large increases in viscosity, liquid-liquid phase separation and/or accelerated aggregation rates at elevated c_2 with small changes in solution environment [22,23,43].

It has been discussed above that the transition towards attractive B_{22} values was more sensitive to electrostatic perturbation than the transition towards more repulsive behavior. This is more pronounced for higher values of α and highly anisotropic charge distribution (*e.g.*, the presence of dipoles due to opposite charges within different domains). This value of α is highly dependent on the amount and chemical identity of ions in solution, which highlights the importance of the solution environment and the narrow experimental space under which protein solutions are stable [8,23,25,113,156].

Additionally, Figure 2.11 shows that there is a small window of net charge values (combination of Q_j for the variable domains, as well as α) that would result in physically realistic $B_{22}/B_{22,ST}$ values that allow one to further refine the parameter space of these models for a given pH and sequence. To achieve reasonable values of B_{22}, the values of net charge on the variable regions are either positive or relatively small negative values at this pH unless one considers very small values of α.

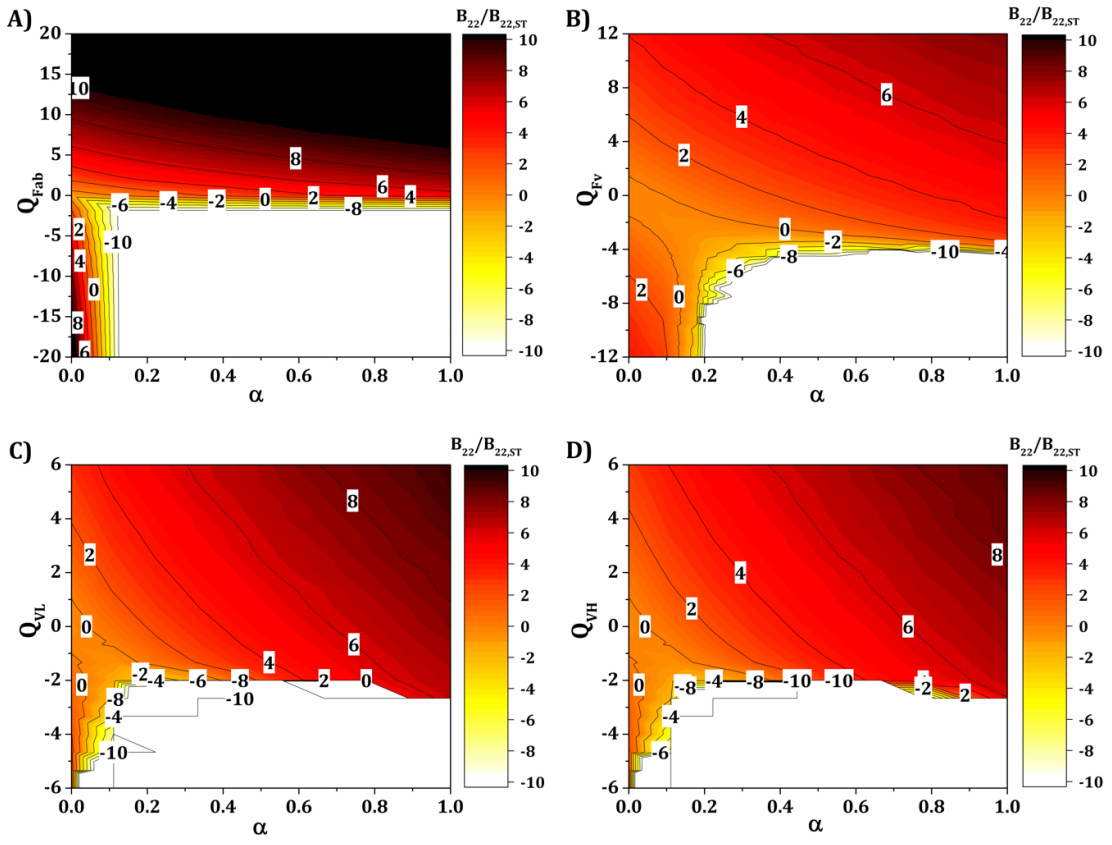

Figure 2.11. Surface response of $B_{22}/B_{22,ST}$ for the TRIAD (panel A), HEXA (panel B) and DODECA (panels C & D) models as a function Q_{Fab}, Q_{Fv}, Q_{VL} or Q_{VH} and α at pH 5 and TIS = 15 mM for a model IgG1 (Table 2.3). Other parameters were set as follows: ε_{SR} = 1.85 k_BT (TRIAD), 1 k_BT (HEXA) and 0.85 k_BT (DODECA) in order to give similar values for B_{22} at high *TIS* across the different models.

Moving to high-c_2 conditions, the structure factor as a function of c_2 was simulated for the TRIAD, HEXA and DODECA models for parameter values that give equal numerical values of $B_{22}/B_{22,ST}$, in order to evaluate if significant differences between the models can be discerned at high c_2 even when starting from the same low-c_2 behavior. At fixed $B_{22}/B_{22,ST}$ values of 0.8 and -0.5, the TRIAD, the HEXA, and the DODECA models each show quantitatively different behaviors at high c_2 as shown in Figure 2.12. Figure 2.12A shows that the HEXA and DODECA models showed similar behavior in the absence of attractions, while the TRIAD model only agrees with the other two up to approximately 50 g/L at an equal $B_{22}/B_{22,ST}$ value. This illustrates that repulsive interactions are less sensitive to the geometric differences in molecular structure. In contrast, Figure 2.12B shows that strongly attractive interactions lead to larger disagreement between the models. This is perhaps not surprising, as close-packed structures depend heavily on the shape of the molecule and correlations between the spatial location of the different domains, and the high-c_2 packing can be expected to change based on the domain-domain interactions. This also highlights a need for development and testing of anisotropic models to simulate high-c_2 behavior of non-globular proteins. Further comparison with experimental behavior is needed to select an optimal model between the HEXA and the DODECA models, particularly under conditions of strongly attractive protein-protein interactions, while factoring in computational burdens. This will be explored in Chapter 3 for two mAb molecules and in Chapter 4 for a globular protein.

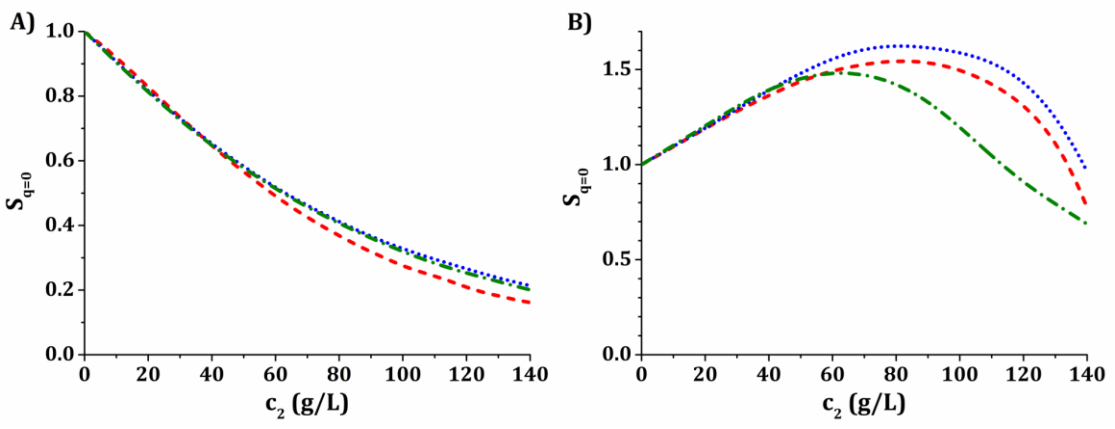

Figure 2.12. Simulated $S_{q=0}$ *vs* c_2 for the TRIAD (dashed red), HEXA (dotted blue) and DODECA (dash-dotted green) models for a constant $B_{22}/B_{22,ST}$ of 0.8 (panel A) and -0.5 (panel B), corresponding to net-repulsive and net-attractive conditions respectively.

2.6 Summary and Conclusions

A comparison of several coarse-grained models with varying molecular detail showed that canonical spherical models overestimate the molecular volume ($2B_{12,ST}$) in comparison to mAb atomistic structures even when matching equal excluded volume ($B_{22,ST}$) contributions at low c_2. Further analysis at high c_2 showed that spherical models lack the physical packing behavior of anisotropic structures such as those of mAbs. A useful balance between numerical accuracy and computational burden was offered by the HEXA (6 beads per protein) and the DODECA (12 beads per protein) models. Analysis of hinge flexibility and short-ranged non-electrostatic attractions showed that the flexibility of the hinge region does not affect the high-c_2 behavior (in terms of the osmotic compressibility) below approximately 140 g/L, therefore rigid models are useful without causing significant additional uncertainty in those conditions for osmotic compressibility calculations. Adding short-ranged hydrophobic and van der Waals attractions primarily affects the solution behavior at high-*TIS* conditions, as expected,

while electrostatic interactions are most relevant at low *TIS*. Analysis of the effect of the charge distribution on $B_{22}/B_{22,ST}$ showed that the presence of highly anisotropic charge distributions leads to unphysically low (negative) $B_{22}/B_{22,ST}$ values, while theoretical charge distributions from the primary sequence and crystal structures result in highly unphysical protein-protein interactions at both low- and high-c_2 conditions. Finally, high-c_2 simulations showed that the level of structural coarse-graining becomes most relevant as interactions move from strongly repulsive to strongly attractive interactions. Combined with the trade-off between structural accuracy and computational burden, this highlights a balance that must be considered when designing CG molecular models for different applications.

Chapter 3

PREDICTIONS OF "WEAK" MAB INTERACTIONS AT HIGH CONCENTRATIONS WITH MOLECULAR SIMULATIONS AND LOW CONCENTRATION EXPERIMENTAL MEASUREMENTS

3.1 Introduction

As described in Chapter 1, reversible ("colloidal" or "weak") protein-protein interactions have been shown to correlate, in some cases, with liquid-liquid phase separation, opalescence, crystallization, aggregation rates, and elevated solution viscosity [17,18,23,25,173]. The contributions to these interactions include steric repulsion, short-ranged non-electrostatic attraction (van der Waals interactions and hydration effects), and electrostatic attraction and repulsion as discussed in Chapter 2. The balance between these forces and their overall contributions to protein interactions depends on the solution conditions and the protein of interest (*e.g.*, sequence and structure) [62,66]. While these same individual contributions are operative at high protein concentrations (c_2), the average distance between the surfaces of adjacent protein molecules is necessarily much smaller than at low c_2 [81,142,144]. This might affect the balance between different forces which will lead to changes in the net protein-protein interactions as c_2 increases [23]. This chapter considers the challenge of using coarse-grained molecular models to predict experimental protein interactions (*via* excess Rayleigh scattering) from low to high c_2. The excess Rayleigh scattering (R^{ex}/K) profiles of two monoclonal antibody protein models provided by Bristol-Myers Squibb (referred as the IgG1 and IgG4) are experimentally determined as a function of pH, total

ionic strength (*TIS*), sucrose content and c_2. Experimental second osmotic virial coefficients (B_{22}) values and protein structural information are used as the only inputs to parameterize coarse-grained models as a function of the strength of the short-ranged non-electrostatic attraction and the effective electrostatic contributions as a function of pH, *TIS* and sucrose content. The parameter tuning is done without prior knowledge of the high-c_2 behavior. The experimental high-c_2 R^{ex}/K results are then predicted using the low-c_2 parameters in transition matrix Monte Carlo simulations. The results are discussed from both qualitative and quantitative perspectives, highlighting strengths and weaknesses of the approach for repulsive and attractive conditions. Additional solution structure measurements are included to better assess the preferential interactions across domains of the same protein, and as a foundation for future model development.

3.2 Materials and Methods

3.2.1 Sample Preparation

Sodium acetate buffer stock solutions were prepared by dissolving glacial acetic acid (Fisher Scientific) in deionized water (MilliQ, Millipore-Sigma) to reach 10 mM acetic acid, and titrated to pH 5.1 ± 0.05 (termed pH 5 below) using a 5 M sodium hydroxide solution (Fisher Scientific). Similarly, 10 mM histidine buffer stock solutions were prepared by dissolving histidine hydrochloride (Sigma) in deionized water and titrating to pH 6.5 ± 0.05 (termed pH 6.5 below). Stock IgG1 and IgG4 solutions were provided by Bristol-Myers Squibb at a starting protein concentration of ~50 g/L. pH 5 and pH 6.5 protein stock solutions were filtered and dialyzed using 10 kDa molecular weight cutoff (MWCO) Spectra/Por dialysis membrane (Spectrum Laboratories,

Rancho Dominguez, CA) with the desired buffer using four 12-hr buffer exchanges at 4 °C to remove any undesired solutes from the original protein solution.

Excipient stock solutions were prepared by dissolving sucrose (HPLC grade, Sigma) and/or NaCl (Fisher Scientific) in 10 mM buffer solutions (acetate for pH 5 and histidine for pH 6.5) to obtain final solutions of 30 w/w % sucrose and/or 1.3 M NaCl. These solutions were titrated to the respective pH with small volumes of a 5 M sodium hydroxide solution. Final protein solutions were prepared gravimetrically by combining (1) protein stock solution, (2) pH-adjusted buffer, (3) pH-adjusted excipient stock solution with matching buffer. The proportions of (1), (2), and (3) were selected to achieve a constant excipient concentration and pH as specified in Table 3.1. This was done for a series of increasing protein concentrations every 0.5 g/L up to a maximum of 10 g/L (for low-c_2 interaction measurements) to ensure dilute protein behavior.

Table 3.1. Summary of formulations for low-c_2 data

Formulation	Additional excipient (concentration range)
pH 5, 10 mM acetate	NaCl (0 - 500 mM)
pH 5, 10 mM acetate + 5% w/w sucrose	
pH 6.5, 10 mM histidine	NaCl (0 - 350 mM)
pH 6.5, 10 mM histidine + 5% w/w sucrose	

For protein solutions above 10 g/L, concentrated protein stock solutions were prepared through membrane centrifugation at ~3200 RCF using 10 kDa MWCO Amicon-Ultra centrifugal tubes (Millipore-Sigma) and two buffer exchange steps. A pH shift was observed as the protein solution was concentrated from ~35 g/L to ~165 g/L, so starting pH values of 4.3 and 5.9 were selected for the dialysis with resulting pH

values of 5.06 ± 0.05 and 6.49 ± 0.05, respectively, for final 165 g/L solutions after centrifugation with two buffer exchange steps. This was performed for both mAb molecules. UV−VIS spectrophotometry (Agilent 8453, Santa Clara, CA) was used to determine the concentration of the protein solutions at 280 nm using an extinction coefficient of 1.54 L g^{-1} cm^{-1} and 1.59 L g^{-1} cm^{-1} for the IgG1 and IgG4, respectively, before and after dilutions from the concentrated stock solutions. Lower concentration protein samples were then prepared by gravimetrically diluting the concentrated protein solution in the desired buffer to obtain c_2 values ranging from 10 to 160 g/L. Less than 0.1% variation between targeted and actual values for the protein and cosolute concentrations was achieved in all cases.

3.2.2 Static Light Scattering (SLS)

Batch SLS experiments were conducted using a Wyatt Technology (Santa Barbara, CA) DAWN HELEOS II instrument with laser wavelength (λ) of 658.9 nm at 25.0 ± 0.1°C. In SLS, the average scattered intensity at 90° can be determined and used to calculate the excess Rayleigh ratio, represented as R^{ex}:

$$R^{ex} = \frac{I_{sample} - I_{buffer}}{I_{toluene} - I_{background}} R_{toluene} * n_{solvent}^{1.983} \tag{3.1}$$

where I is the measured intensity of the sample (I_{sample}), buffer (I_{buffer}), toluene ($I_{toluene}$) and the background radiation ($I_{background}$); $R_{toluene}$ is the Rayleigh ratio of toluene at the measured temperature, and $n_{solvent}$ is the refractive index of the solvent [51]. Measurements of R^{ex} as a function of c_2 can be used to estimate protein-protein interactions in the form of the protein-protein Kirkwood-Buff integral, G_{22}:

$$\frac{R^{ex}}{K} = M_{w,app} c_2 + M_w G_{22} c_2^2 \tag{3.2}$$

where $M_{w,app}$ is the protein apparent molecular weight and M_w is the protein true molecular weight. K is the optical constant and equal to $4\pi^2 n^2 (dn/dc_2)^2 N_A^{-1} \lambda^{-4}$, where n is the solution refractive index, (dn/dc_2) is the change in refractive index of the solution as a function of c_2 (see below) and N_A is Avogadro's number [29]. The zero-q limit for the structure factor ($S_{q=0}$) can be obtained by dividing the right hand side of equation 3.2 by $c_2 M_w$, with the canonical simplification that $M_{w,app} \approx M_w$. In this case, $S_{q=0}$ is equal to $1+ c_2 G_{22}$ and dimensionless. Values of (dn/dc_2) were determined using a J157HA Refractometer (Rudolph Scientific, Hackettstown, NJ) for c_2 values up to 10 g/L for each formulation. A value of 0.203 ± 0.03 mL/g was obtained for buffer-only and NaCl formulations, while 0.220 ± 0.04 mL/g was obtained for all formulations with 5% w/w sucrose, for both pH values and both mAb molecules.

In the limit of dilute protein concentration (*i.e.*, c_2 below approximately 10 g/L and $|c_2 G_{22}| < 0.1$), $G_{22} \approx -2B_{22}$ [23,29]. As B_{22} is independent of c_2, B_{22} values were obtained by fitting experimental excess Rayleigh profiles to equation 3.2 for low-c_2 conditions. Additionally, KB theory and the corresponding analysis is also applicable at higher c_2, so it can be used to quantify protein-protein interactions at high c_2 from SLS data [23,29]. Negative (positive) G_{22} values are equivalent to $S_{q=0}$ values below (above) 1, and corresponds to overall repulsive (attractive) interactions. Correspondingly in dilute solutions, positive (negative) B_{22} values indicate overall repulsion (attraction).

3.2.3 Coarse-Grained mAb Models and Interaction Parameters

3.2.3.1 Low resolution CG models: the HEXA and DODECA models

Two different coarse-grained (CG) molecular models were used to model low-c_2 and predict high-c_2 SLS experimental behavior. These were a subset of the larger

group of different molecular models that were introduced in Chapter 2. These two models can provide an optimal balance between accuracy and computational burden as shown in Chapter 2. Figure 3.1 shows a schematic with geometric constraints of these CG models, referred to as the HEXA and DODECA models in the remainder of this chapter. These models were developed and refined in Chapter 2 to resemble the overall shape of a mAb molecule, and used 6 (HEXA) or 12 (DODECA) beads per protein. Additional model details can be found in Chapter 2.

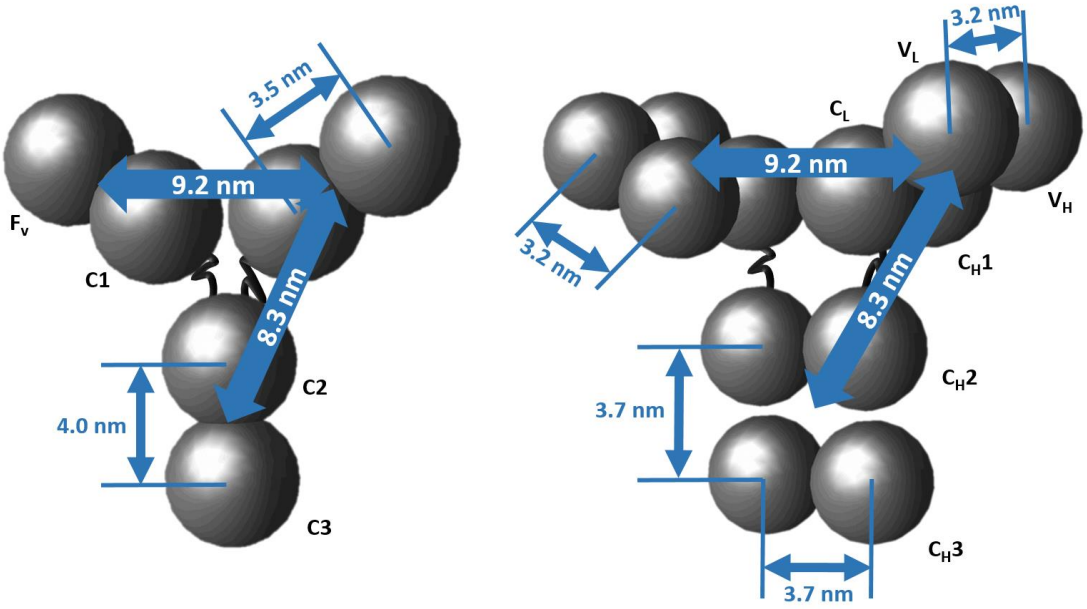

Figure 3.1. Schematic diagrams of the HEXA (left) and DODECA (right) geometries, including refined geometric constraints. The solid-line connectors between Fab and Fc domains indicate a rigid Fab-Fc linker was employed.

A modification to the previously proposed short-ranged non-electrostatic attraction model was made here to achieve an effective attractive range of ~1 nm for

both the HEXA and DODECA models (compared to equation 2.4). This was achieved by modifying the range of attractions for the DODECA model:

$$\frac{u_{SR}(r_{ij})}{k_B T} = \frac{\varepsilon_{SR}}{k_B T}\ c\ \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{128} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{n}\right] \tag{3.3}$$

where ε_{SR} represents the strength of the short-ranged non-electrostatic attractions, n represents the range of the attractions and is equal to 6 for the DODECA model and 10 for the HEXA model, and c is a normalization factor to make the potential energy equal to $-\varepsilon_{SR}$ at its minimum value and is equal to 1.2196 for the DODECA model and 1.3464 for the HEXA model.

Similarly, a modification to the previously proposed electrostatic models was made to better model experimental data. This was achieved by changing from a Yukawa potential (equation 2.5) to a modified screened-Coulomb potential:

$$\frac{u_{el}(r_{ij})}{k_B T} = \zeta\ \psi_i \psi_j q_i q_j\ \frac{e^{(-\kappa(r_{ij}-\sigma_{ij})}}{\left[\frac{r_{ij}}{\text{Å}}\right]\left[1+\frac{1}{2}(\kappa\sigma_{ij})\right]^2} \tag{3.4}$$

ζ corresponds to the Bjerrum length and is equal to $(4\pi\epsilon\epsilon_o k_B T)^{-1}$, with ϵ representing the solution relative permittivity at a given temperature, ϵ_o is the vacuum permittivity (in units of $q_e^2\ N^{-1}\ m^{-2}$, for q_e representing the elemental charge of an electron), k_B is the Boltzmann constant and T is the absolute temperature. For solutions considered here, ζ was equal to 7.15 Å for any buffer + NaCl formulations at 25 °C, and 7.26 Å for formulations with 5% w/w added sucrose at 25 °C [174]. q_i and q_j are the theoretical valence of domain/fragment/amino acid i and j, respectively, as calculated from the protein sequence (see below), while ψ_i (ψ_j) is used to scale the theoretical charge such that $\psi_i q_i$ ($\psi_j q_j$) is equal to the effective valence in solution, $q_{i,eff}$ ($q_{j,eff}$). σ_{ij} is the average diameter of beads i and j and equal to ½(σ_i + σ_j), where σ_i and σ_j are the diameter of the

*i*th and *j*th bead or (sub)domain, respectively. κ is the Debye screening length based on the *TIS* of the solution, and r_{ij} is the center-to-center distance between the interacting beads *i* and *j*. This modification allows one to better capture electrostatic phenomena as well as it allows for easier comparison with electrophoresis measurements [154,156].

Theoretical valence values (q_i) were calculated using the standard Henderson-Hasselbalch equation [169]. Protein sequences for the IgG1 and IgG4 molecules were provided by Bristol-Myers Squibb, including homology models for better capturing the molecule geometry (see Figure 3.1), which were refined in comparison to Figure 2.2. This sequence was partitioned into equal-chain-length units to compute the charge of each HEXA or DODECA model bead. For the HEXA model calculations, the Fv domain was composed of the upper half of the light chain (residues 1-107) and the upper quarter of the heavy chain (residues 1-118) (*i.e.*, combining both V_H and V_L domains); the C1 domain was composed of the lower half of the light chain (residues 108-214) and the second quarter of the heavy chain (residues 119-234) (*i.e.*, combining both C_H1 and C_L domains); the C2 domain was composed of residues 244-357 of each heavy chain (*i.e.*, both C_H2 domains); and the C3 domain was composed of the last quarter of each of the heavy chains (residues 358-474) (*i.e.*, both C_H3 domains). For the DODECA model calculations, the heavy chains were portioned into four units (residues 1-118 for the V_H, 119-234 for the C_H1, 244-357 for the C_H2, and 358-474 for the C_H3) and the light chains into two units (residues 1-107 for the V_L, and 108-214 for the C_L) for a total of 12 beads, each with its respective net charge. In what follows, the terms valence and charge will be used interchangeably. Examples of the theoretical charge distribution for the DODECA model and the IgG1 are shown in Figure 3.2, with the IgG4 shown in

Figure 3.3. All theoretical charge values are shown in Table 3.2 for pH 5 and 6.5 for both CG models and molecules.

Table 3.2. Theoretical charges at pH 5 and 6.5.

Molecule	Model	q_{VH}	q_{VL}	q_{CL}	q_{CH1}	q_{CH2}	q_{CH3}
IgG1	DODECA, pH 5	1.89	0.72	2.94	6.27	5.07	2.78
	DODECA, pH 6.5	0.29	0.02	0.55	4.53	1.82	-0.20
	HEXA, pH 5	2.61		9.21		10.14	5.56
	HEXA, pH 6.5	0.31		5.08		3.64	-0.40
IgG4	DODECA, pH 5	2.79	2.61	1.94	2.27	1.15	2.89
	DODECA, pH 6.5	2.02	2.02	-0.45	0.53	-1.44	-0.19
	HEXA, pH 5	5.40		4.21		2.30	5.78
	HEXA, pH 6.5	4.04		0.08		-2.88	-0.38

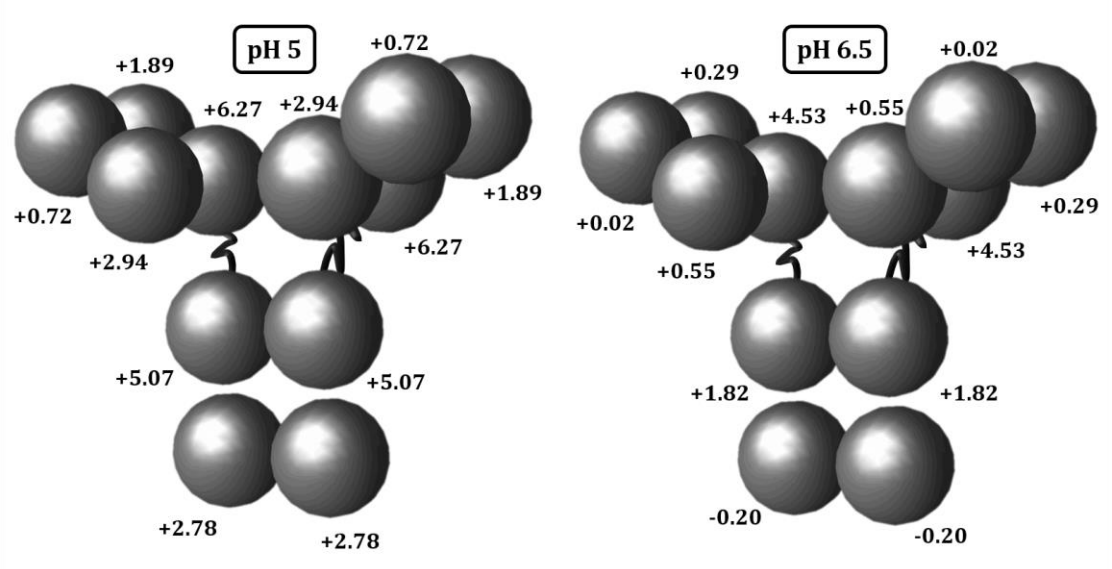

Figure 3.2. Theoretical charge distribution for the DODECA model at pH 5 and 6.5 for the IgG1 molecule.

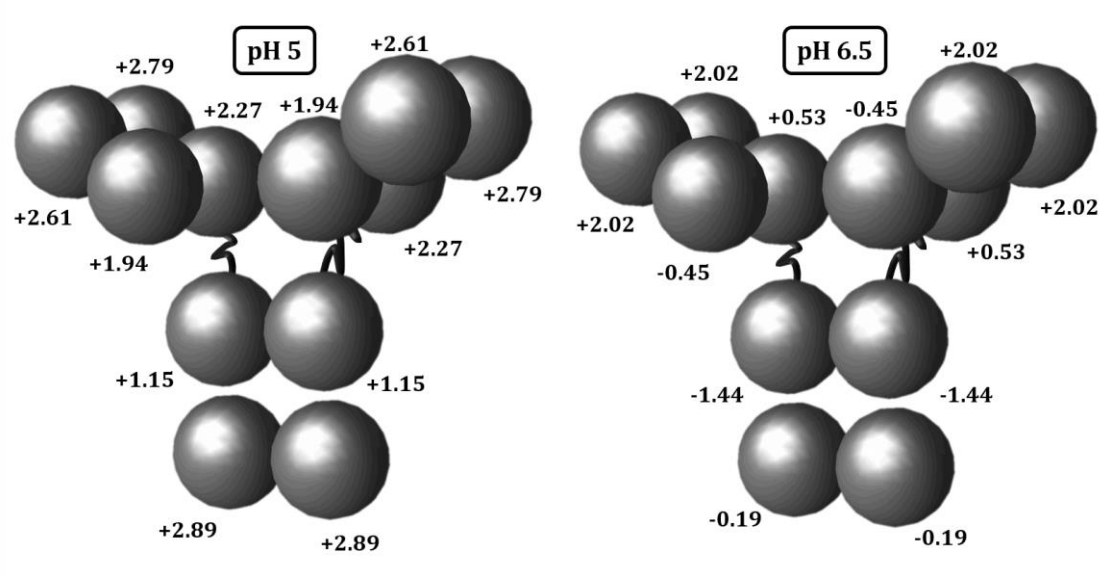

Figure 3.3. Theoretical charge distribution for the DODECA model at pH 5 and 6.5 for the IgG4 molecule.

3.2.3.2 High Resolution CG Model: the 1bAA Model

A structurally higher resolution CG model was used to evaluate interactions between pairs of IgG1 and IgG4 molecules, as a function of the formulation space. A previously developed one-bead-per-amino acid (1bAA) model was used to compute B_{22} values using the Mayer sampling with overlap sampling algorithm (see below). The 1bAA force field proposed in the literature was updated using equation 3.4 to model electrostatic interactions. Short-ranged non-electrostatic attractions were modeled *via*

$$\frac{u_{hp}(r_{ij})}{k_BT} = \begin{cases} \frac{4\varepsilon_{SR}}{k_BT}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} + 1 - \varepsilon_{ij}\right], & if\ \sigma_{ij} < r_{ij} \le r_c \\ \frac{4\varepsilon_{SR}\varepsilon_{ij}}{k_BT}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], & otherwise \end{cases} \tag{3.5}$$

in agreement with work found in the literature [66]. Here, ε_{SR} represents the strength of short-ranged non-electrostatic attractions, $\varepsilon_{ij} = (\varepsilon_i\varepsilon_j)^{1/2}$ is the relative hydrophobicity of

the *i*-*j*th pair of residues with ε_i denoting the specific hydrophobicity of residue *i*, and r_c is the cut-off distance equal to $2^{1/6}\sigma_{ij}$ [62,66]. Additional parameters have been defined elsewhere [66]. Finally, steric-only interactions were modeled using equation 2.1 (*i.e.*, hard-sphere potential). The 1bAA force field allows one to evaluate the combination of the effects of short-ranged non-electrostatic interactions (*e.g.*, hydration effects and van der Waals attractions), electrostatic interactions, and protein excluded volume at the scale of individual amino acids while still being computationally tractable. Each amino acid is represented as a single bead, with different bead sizes and short-ranged non-electrostatic attraction energetics for different amino acid identities, and the charge (q_i) for the *i*th amino acid resides at the center of that bead. Based on nominal pKa and pKb values, at pH 5, all D and E amino acids are approximated as having a charge of -1, while all K, H and R amino acids have a charge of +1. At pH 6.5, only H residues are modeled as deprotonated, changing their charge value from +1 to 0. The values of the specific hydrophobicity scaling parameter (ε_i) and bead sizes (σ_i) are shown in table 3.3.

Table 3.3. ε_i and σ_i parameter values for each of the 20 natural amino acids.

Residue	σ_i (Å)	ε_i
Lys (K)	7.03	0.00
Glu (E)	6.40	0.05
Asp (D)	5.83	0.06
Asn (N)	5.95	0.10
Ser (S)	5.28	0.11
Arg (R)	7.32	0.13
Gln (Q)	6.35	0.13
Pro (P)	5.62	0.14
Thr (T)	5.81	0.16
Gly (G)	4.31	0.17

Residue	σ_i (Å)	ε_i
His (H)	6.29	0.25
Ala (A)	5.02	0.26
Tyr (Y)	7.11	0.49
Cys (C)	4.92	0.54
Trp (W)	6.70	0.64
Val (V)	6.05	0.65
Met (M)	6.32	0.67
Ile (I)	6.36	0.84
Phe (F)	6.95	0.91
Leu (L)	6.55	1.00

3.2.4 Monte Carlo Simulations

3.2.4.1 B_{22} Simulations from Mayer Sampling with Overlap Sampling

The HEXA, DODECA and 1bAA models were used to compute B_{22} for a given pH and *TIS* using the Mayer sampling method employing the overlap sampling algorithm (MSOS) developed by Kofke and coworkers [132]. A similar methodology to the one employed in Chapter 2 was used here: MSOS simulations were performed at 25 °C with 10^7 Monte Carlo (MC) attempts for both the reference system and the model of interest. Each MC attempt consisted of either a translation or a rotation around the center of mass of the first protein molecule using the second molecule as the origin. The maximum displacement and rotation were obtained with a pre-equilibration step of 10^5 MC attempts where these features were adjusted to obtain an acceptance ratio of 50% (see section 2.2.3). The steric-only behavior of the protein was used as a reference, so the simulation directly returned $B_{22}/B_{22,ST}$, where $B_{22,ST}$ represents the steric-only second osmotic virial coefficient (*i.e.*, the value due to only protein excluded volume contributions) as explained in Chapter 2 and below. The following simulations were performed: $B_{22}/B_{22,ST}$ was calculated for ε_{SR} values between 0 and 1.5 k_BT, ψ_i values between 0 and 4, and TIS values between 0 and 520 mM. The obtained $B_{22}/B_{22,ST}$ values were compared to experimental values for further parameter tuning or analyzing the surface response space (see subsections below). Statistical uncertainties were estimated by performing 5 independent simulations for each model and a given solution condition. The standard deviation was used as the estimate of statistical uncertainty, including error propagation.

3.2.4.2 High-c_2 Simulations with Transition Matrix Monte Carlo

Transition matrix Monte Carlo (TMMC) was used to compute R^{ex}/K *vs* c_2 profiles for c_2 values above 10 g/L using the methods described in Chapter 2 and below. The parameters were refined *via* MSOS simulations and compared with low-c_2 experimental B_{22} values as described in the next subsection. The simulations were carried out in a grand-canonical (osmotic) system [68,128,136]. An initially uniform concentration probability distribution was used, which was subsequently reconstructed at the end of each cycle until it converged to the actual probability distribution, with each cycle being defined as 10^6 MC attempts. A MC attempt consisted of one of the following randomly selected moves: a translation, a rotation or a molecule insertion or deletion [128,136]. Translations and rotations represented 30% of all MC attempts, while deletions and insertions represented the remaining 70%. The temperature was held constant at 25 °C. Preliminary simulations were used to find an adequate value of the reference chemical potential, depending on the parameter values (see below). Due to boundary effects, G_{22} was observed to depend on the box size for $c_2 > 30$ g/L and box sizes below 50 nm [128,168]. Consequently, a box length from 60 nm to 180 nm was used, where simulated values of G_{22} were not found to significantly depend on the box length and larger box sizes were used for low-c_2 conditions to decrease the noise on simulated G_{22} values. The simulation box was started with an empty system. G_{22} values were calculated by using histogram reweighting on the c_2 probability distribution *via*

$$c_2 G_{22} = \left(\frac{\langle N_2^2 \rangle - \langle N_2 \rangle^2}{\langle N_2 \rangle} - 1 \right) \tag{3.6}$$

where c_2 represents the average protein concentration in g/L, $\langle N_2 \rangle$ is the average number of protein molecules in the simulation box, and $\langle N_2^2 \rangle - \langle N_2 \rangle^2$ represents the average fluctuations in the number of proteins, all of these for a given choice of protein chemical

potential. This equation is equivalent to a dimensionless version of equation 1.3. Excess Rayleigh (R^{ex}/K *vs* c_2) and $S_{q=0}$ profiles were obtained by inserting simulated values of G_{22} in equation 3.2 with a M_w value of 146.5 kDa and assuming that $M_{w,app}$ and M_w are equal, as is the case for many mAb solutions [23,29,37,175].

3.2.4.3 Higher Order Virial Coefficients *via* MSOS

The HEXA and DODECA models were also used to compute third (B_{222} or $B_{2(3)}$), fourth ($B_{2(4)}$) and fifth ($B_{2(5)}$) virial coefficients for the same model parameters used to compute high-c_2 behavior with the TMMC algorithm. This was achieved by using the MSOS algorithm but in this case applied to three proteins ($B_{2(3)}$), four proteins ($B_{2(4)}$) and five proteins ($B_{2(5)}$) for each respective virial coefficient [131,132,176]. The methodology was similar to the one employed for B_{22} with minor changes for each virial coefficient: MSOS simulations were performed at 25 °C with 10^7-10^{10} Monte Carlo (MC) attempts for both the reference system and the model of interest until achieving convergence for each virial coefficient. On average, this was achieved within 10^8 MC attempts for $B_{2(3)}$, 10^9 for $B_{2(4)}$ and 10^{10} for $B_{2(5)}$. Each MC attempt consisted of either a translation or a rotation around the center of mass of the first, second, third and/or fourth protein molecule using the last molecule as the origin, depending on the number of simulated proteins. The maximum displacement and rotation was obtained with a pre-equilibration step of 10^5 MC attempts, where these features were adjusted to obtain an acceptance ratio of 50% (see section 2.2.3). The steric-only behavior of the protein was also used as a reference, so the simulations directly returned $B_{2(i)}/B_{2(i),ST}$, where $B_{2(i)}$ represents the *i*th osmotic virial coefficient, and $B_{2(i),ST}$ represents its steric-only counterpart (*i.e.*, the value due to only protein excluded volume contributions) [131,148,177–179].

The following simulations were performed: $B_{2(i)}/B_{2(i),ST}$ values were calculated for a subset of [ε_{SR}, ψ_i] values that match those tested with the TMMC algorithm to predict high-c_2 behavior. $B_{2(i),ST}$ values were computed using the same methodology employed in Chapter 2 for $B_{22,ST}$, but applied to the respective higher virial coefficient. In summary, a reference hard sphere in the center of mass of each protein was used in computing $B_{2(i),ST}$ values, which resulted in $B_{2(i),ST}/B_{2(i),HS}$ values, with $B_{2(i),HS}$ representing the *i*th osmotic virial coefficient for a hard sphere fluid. Final $B_{2(i),ST}$ values were obtained by rescaling the simulation results with analytical $B_{2(i),HS}$ results [177]. The resulting $B_{2(i)}/B_{2(i),ST}$ values were then scaled with computed $B_{2(i),ST}$ values (see Chapter 2). Simulated $B_{2(i)}$ values were used to compute the osmotic pressure as shown in equation 3.7, which was later used to compute the necessary derivative to evaluate R^{ex}/K values as shown in equation 3.8. Here, Π_2 represents the osmotic pressure of the protein in solution, R is the gas constant and T is the absolute temperature. Equation 3.7 is obtained by expanding the osmotic pressure of the protein as a function of virial coefficients as is done in the McMillan-Mayer solution theory [59,96,148,177,180]. Likewise, equation 3.8 is obtained by combining equations 3.2, 3.6 (or equation 1.3 in the same matter) and the osmotic compressibility equation, $\kappa_T = (\partial c_2/\partial \Pi_2)_T \cdot c_2^{-1}$.

$$\frac{\Pi_2 M_w}{c_2 RT} = 1 + B_{22}c_2 + B_{2(3)}c_2^2 + B_{2(4)}c_2^3 + B_{2(5)}c_2^4 \tag{3.7}$$

$$\frac{R^{ex}}{K} = M_{w,app}c_2 + M_w c_2 \left[\left(\frac{\partial c_2}{\partial(\Pi_2/RT)}\right)_{T,\mu_{i\neq 2}} - 1\right] \tag{3.8}$$

3.2.4.4 Domain Contact Maps with Radial Distribution Function Simulations

Domain-domain radial distribution functions ($g_{ij}(r)$ *vs* r) were collected in a grand-canonical system using the same DODECA model parameters used to predict

high-c_2 SLS behavior (see below). The same simulation movements used during the TMMC simulations, and discussed in Chapter 2 in detail, were used without biasing the sampling (*e.g.*, grand-canonical simulations [128]). A box length of up to 180 nm was used, and the protein chemical potential was tuned to obtain an average c_2 value of 10, 50, 100 or 150 g/L for a defined set of parameters during the grand-canonical sampling. This allows one to obtain information regarding preferential contacts between domains of the protein molecule. Conditions with only short-ranged non-electrostatic attractions (*i.e.*, $\psi_i = 0$) were used as a comparison to better discern the effect of added charges and packing limitations (intrinsic in the shape of the molecule) in the preferential contacts across protein domains. Histograms for the spatial distribution of protein beads (*i.e.*, how often a bead is observed based on a reference frame) were collected for a uniform grid of 0.2 nm every 10^2 MC attempts for a total of 10^7 MC attempts (*i.e.*, a total of 10^5 entries were recorded for each grid of a histogram), using varying beads (domains) as the reference point [125,126,128]. These histograms were converted to $g_{ij}(r)$ *vs r* values using algorithms explained in reference [128]. Here, the subscripts *i-j* indicate that the radial distribution function is collected based on pairs of domains *i-j* interacting with each other. For instance, $g_{CH3\text{-}VH}(r)$ indicates a radial distribution function of C_H3 domains with respect to V_H domains, averaged over all simulated protein molecules and at a given μ_2 value (or average c_2 value [136]) and parameter set. By using the definition of the potential of mean force [59,128,148], $g_{ij}(r)$ *vs r* were transformed into potentials of mean force, $w_{ij}(r)$, for each possible domain-domain contact *via*

$$\frac{w_{ij}(r)}{k_B T} = -\ln\left[g_{ij}(r)\right] \tag{3.9}$$

The minimum value of such potential of mean force is used to reconstruct domain-domain energy contact surface responses by plotting this minimum value (referred as

$\min(w_{ij})/k_BT$) for each possible domain-domain interaction (*e.g.*, a total of 21 independent contacts for the DODECA model) [148].

3.2.5 Average Relative Deviation (ARD) Calculations and Model Validation

In order to evaluate the effectiveness of the present CG models to model or predict experimental behavior, ARD values were calculated for any given data set *via*:

$$ARD\ (\%) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i^{experimental} - x_i^{predicted}}{x_i^{experimental}}\right| \tag{3.10}$$

where n represents the number of data points and x_i is the experimental or simulated value to be evaluated (*e.g.*, B_{22} *vs TIS*, and R^{ex}/K *vs* c_2 in this work). As the ARD represents the average deviation between the model and the experimental data, a cutoff value between 10% and 20% was used as a criterion for considering a prediction to be quantitatively accurate, as this average deviation can be considered a conservative estimate of the model prediction uncertainty, particularly given typical experimental uncertainties for SLS data.

3.2.6 Tuning Model Parameters from Low-c_2 Data

To predict high-c_2 excess Rayleigh profiles from low-c_2 measurements using the formulated CG models, B_{22} *vs TIS* experimental data were used to tune two model parameters: the strength of short-ranged non-electrostatic attractions ε_{SR} (equation 3.3), and the correction factor to the theoretical charges ψ_i (equation 3.4). Additional model parameters were refined in Chapter 2 and above based only on the geometry of multiple mAbs from their published crystal structures, as well as the homology models for these molecules. Under high-*TIS* conditions, electrostatic interactions are expected to be heavily screened (according to the Debye-Hückel theory), so B_{22} values under these

conditions can be used to set the value of ε_{SR} by combining experimental and simulated data [23,66,76,148,154]. Conversely, low-*TIS* conditions are expected to be dominated by electrostatic interactions, and this can be used to determine an optimized value of ψ_i [23,66]. For simplicity, all ψ_i values were assumed to be equal for all the domains. Consequently, ψ_i will be referred as ψ in the remainder of this work, as an average correction factor for all theoretical charges at a given pH [154,156]. Similarly, ε_{SR} will be used as an averaged short-ranged non-electrostatic attraction strength (for solvation and hydration effects, and van der Waals attractions) and equal for all the molecule domains in the HEXA and DODECA models [57,143].

The following methodology was employed for the parameter tuning exercise: $B_{22}/B_{22,ST}$ *vs TIS* values were simulated using both the HEXA and DODECA models using the MSOS algorithm for a range of [ε_{SR}, ψ] pairs. ARD values were computed for each pair by comparing experimental and simulated $B_{22}/B_{22,ST}$ *vs TIS* results. Experimental $B_{22}/B_{22,ST}$ values between -0.05 and 0.05 were excluded from any ARD calculation to avoid heavy biasing on the final ARD value. Surface plots of ARD *vs* [ε_{SR}, ψ] were constructed, where a funnel-like behavior is expected if there is a unique subset of [ε_{SR}, ψ] pairs that minimizes the ARD results. As there is experimental uncertainty in experimental B_{22} values, the previous exercise would result in a parameter space of [ε_{SR}, ψ] pairs that can accurately mimic the experimental data, as shown below. Consequently, all simulated [ε_{SR}, ψ] pairs that resulted in ARD values below 20% were subsequently used to predict high-c_2 Rayleigh scattering behavior, creating a predicted "envelope" for R^{ex}/K rather than a single curve.

3.3 Steric-Only Behavior as Reference and Equations of State (EoS)

The steric-only behavior can be used as a reference state as this corresponds to the minimum level of interactions any macromolecule would experience in solution (see Chapter 2). In the case of low-c_2 behavior, steric interactions are c_2-independent, resulting in a B_{22} value of 0.01 L/g as calculated in Chapter 2 and termed $B_{22,ST}$ in the remainder of this chapter. This B_{22} value can be used to normalize any B_{22} value across different solution formulations. Consequently, $B_{22}/B_{22,ST}$ values above 1 would be representative of added repulsion (beyond sterics) to the protein, which will be termed as "net-repulsive" in the remainder of this chapter. Likewise, $B_{22}/B_{22,ST}$ values below 1 would be representative of added attractions that overcome the steric-only behavior of the protein, and this will be termed as "net-attractive" in the remainder of this chapter.

For higher c_2 values, it is necessary to develop expressions to compute the c_2-dependent behavior of steric interactions. Chapter 2 showed the results of computing steric-only interactions as a function of c_2 using several CG models, including those used in this chapter. In Chapter 2, grand-canonical MC simulations were carried out to obtain values of c_2 as a function of protein chemical potential (c_2 *vs* μ_2). These results can be further used to compute a steric-only EoS to analytically calculate the high-c_2 behavior due to steric-only interactions [59,148,178,181,182]. This can then be used as a reference state instead of the ideal gas or non-interacting behavior (*i.e.*, $B_{22} = 0$ or $S_{q=0} = 1$). Consequently, two different approaches were used to obtain such an EoS. The first approach was based on the virial expansion as is done in the McMillan-Mayer solution theory (referred to as VE below, as short-hand for virial expansion) [148,180]. VE-EoS provides a simple 4th-order polynomial

$$\frac{\Pi_{2,ST} M_w}{c_2 RT} = A_1 + A_2\eta + A_3\eta^2 + A_4\eta^3 + A_5\eta^4 \tag{3.11}$$

which can easily be used to analytically calculate thermodynamic properties [59,96,148,180]. This approach is expected to strongly deviate from real multi-body behavior as c_2 increases beyond the range of simulated values (*e.g.*, above 160 g/L). The second approach was based on the Carnahan-Starling EoS (referred as the MC-S below, as short-hand for modified Carnahan-Starling), which was previously developed to accurately capture the steric-only behavior of spherical models as a function of volume fraction [148,183]. However, as shown in Chapter 2, spherical models lack the packing resolution to capture high-c_2 mAb behavior, so modifications were performed to the coefficients in this EoS. The MC-S EoS provides a more complicated mathematical function that was inspired by self-consistent statistical mechanical derivations of hard sphere fluids that can still be used to obtain analytical thermodynamic properties and derivatives [148,183]:

$$\frac{\Pi_{2,ST} M_w}{c_2 RT} = \frac{A_1 + A_2\eta + A_3\eta^2 + A_4\eta^3}{(1 + A_5\eta)^3} \tag{3.12}$$

In equations 3.11 and 3.12, $\Pi_{2,ST}$ represents the osmotic pressure of the protein in solution due to only sterics, R is the gas constant and T is the absolute temperature. η corresponds to the protein volume fraction in solution (= $v_2 \cdot c_2$). v_2 is the protein molecular volume and was computed in Chapter 2 using atomistic simulations and found equal to 0.93 mL/g (Table 2.2) for a series of mAb molecules. Both of the proposed analytical steric-only EoS models were fitted to simulated data ($\Pi_{2,ST}$ *vs* η, and κ_T *vs* η) by minimizing the error in both the isothermal compressibility ($\kappa_T = (\partial\eta/\partial\Pi_{2,ST})_T \cdot \eta^{-1}$) and the osmotic pressure ($\Pi_{2,ST}$) as a function of protein volume fraction (η) for values of $\eta < 0.165$ (*i.e.*, $c_2 < 180$ g/L). The resulting parameters obtained from error minimization are shown in Table 3.4. It is worthwhile to point out that

extrapolating to higher volume fractions (or c_2 values) is discouraged as additional parameters might be required to capture even more crowded environments [178,180,184].

Table 3.4. Model parameters for the steric-only EoS

EoS	A_1	A_2	A_3	A_4	A_5
VE	1†	10.551 ± 0.006	62.2 ± 0.2	136 ± 1	468 ± 4
MC-S		7.5177 ± 0.0007	33.84 ± 0.01	-4.60 ± 0.05	-1‡

†Preset to comply with dilute limit behavior (ideal gas EoS)

‡Preset to ensure continuity in the compressibility as well as limiting behavior at $\eta = 1$

3.4 Experimental and Computational "Weak" Protein-Protein Interactions for the IgG1 Molecule

SLS was used to determine excess Rayleigh profiles (R^{ex}/K as a function of c_2) for a series of solution conditions for the IgG1 molecule. At low c_2, these measurements were used to determine B_{22} values as a function of *TIS* by changing the NaCl molarity. Figure 3.4 shows the results of B_{22} *vs TIS* for two series of formulations (buffer + NaCl, and buffer + 5% w/w sucrose + NaCl) and two pH values: 5 (panel A) and 6.5 (panel B), all measured for c_2 < 10 g/L. B_{22} values were normalized using the steric-only behavior from a 3D homology model (0.01 L/g) as a reference state and for easier comparison with the MSOS simulations. $B_{22}/B_{22,ST}$ *vs TIS* profiles differ quantitatively between pH 5 and pH 6.5, and between both sucrose concentrations. At pH 5 and low *TIS*, protein-protein interactions were relatively large and net-repulsive ($B_{22}/B_{22,ST}$ >> 1). Increasing *TIS* by adding NaCl decreases the magnitude of B_{22} until reaching a constant value for *TIS* values above approximately 300 mM. At pH 6.5 and low *TIS*, protein-protein interactions were net-attractive, relative to steric-only interactions

($B_{22}/B_{22,ST}$ < 1). Increasing *TIS* by adding NaCl decreases the magnitude of B_{22} until reaching a constant value for *TIS* values above approximately 300 mM in all situations. B_{22} values at high *TIS* (>300 mM) were the same for both pH values but became less attractive (less negative) with the addition of sucrose, with similar high *TIS* results in the presence of sucrose for both pH values. Conversely, B_{22} values differ significantly across pH at low *TIS* (below 50 mM), where pH 5 resulted in larger (more repulsive) B_{22} values than pH 6.5. The addition of sucrose did not result in statistically distinguishable behavior at low *TIS*.

Insets in Figures 3.4A and 3.4B show the experimental R^{ex}/K *vs* c_2 (high-c_2) results for the formulations presented in Table 3.5 and that correspond to the low-c_2 measurements in the main panels. Additionally, the steric-only behavior for this molecule is shown as a reference, as computed using the VE EoS model (equation 3.11 and Table 3.4). The results in Figure 3.4A (inset, pH 5) show that R^{ex}/K profiles for both buffer-only and sucrose formulations are net-repulsive (R^{ex}/K values below the steric-only behavior), with R^{ex}/K profiles for sucrose below (more repulsive than) those for buffer-only. Adding 100 mM NaCl results in a substantial increase in R^{ex}/K values and brings it above the steric-only behavior. For Figure 3.4B (inset, pH 6.5), the buffer-only formulation overlaps with the steric-only behavior at low c_2, and adding 5% w/w sucrose results in a decrease in the R^{ex}/K profiles (increase in repulsions) while adding 100 mM results in an increase in R^{ex}/K values (increase in attractions). In each formulation (buffer, buffer + sucrose, and buffer + NaCl), R^{ex}/K values at high c_2 at pH 5 are lower in magnitude than those at pH 6.5 for equal c_2 values.

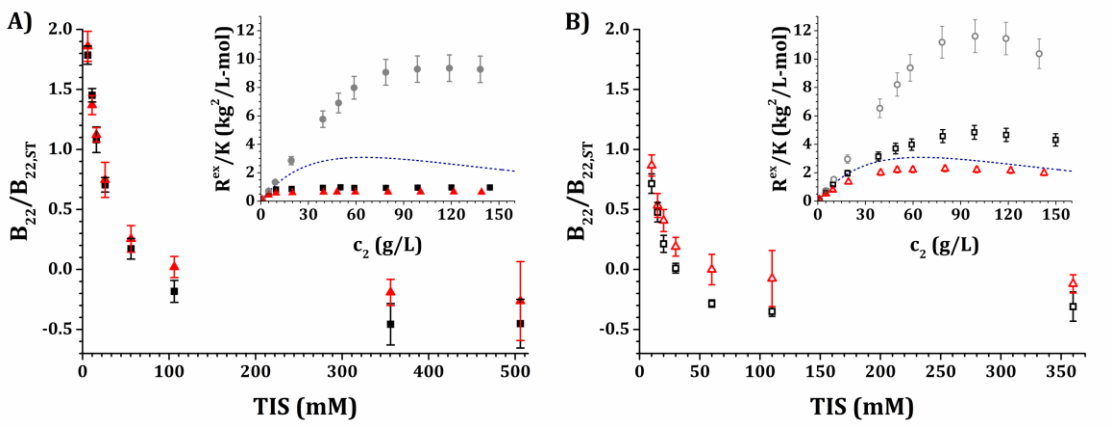

Figure 3.4. **Main panels:** $B_{22}/B_{22,ST}$ values as a function of *TIS* for the IgG1 at pH 5 (panel A) and pH 6.5 (panel B) with added NaCl from 0 to 500 mM. Black symbols represent data with only buffer and added NaCl while red symbols represent the same solutions with 5% w/w added sucrose. **Insets:** high-c_2 data for pH 5 and pH 6.5 for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles). The blue dashed line corresponds to the steric-only behavior calculated using the VE EoS.

Table 3.5. Summary of formulations for high-c_2 SLS data across protein molecules.

Formulation	Short notation
pH 5, 10 mM acetate	pH 5, buffer
pH 5, 10 mM acetate, 5% w/w sucrose	pH 5, sucrose
pH 5, 10 mM acetate, 100 mM NaCl	pH 5, NaCl
pH 6.5, 10 mM histidine	pH 6.5, buffer
pH 6.5, 10 mM histidine, 5% w/w sucrose	pH 6.5, sucrose
pH 6.5, 10 mM histidine, 100 mM NaCl	pH 6.5, NaCl

Protein-protein "weak" interactions are mediated by the solution environment the protein is subject to [23,25,51,156,185]. These interactions have three main contributions: steric or excluded volume effects (repulsive); short-ranged van der Waals interactions and hydration/solvation effects (net attractive or repulsive); and electrostatic interactions (both attractive and repulsive) [66]. Among these, only the

latter should be significantly affected by the solution ionic environment (*via* charge screening) if one neglects ion binding effects [148,154,156,186]. Consequently, the decrease in B_{22} values and plateau behavior that are observed as *TIS* increases in Figure 3.4 can be attributed to the screening of strong charge-charge repulsion with the addition of NaCl, as described by the Debye-Hückel theory [148,154,156]. This agrees with previously published experimental behavior of a number of proteins as a function of ionic environment [23,25,187,188]. Since $B_{22}/B_{22,ST}$ converges towards values less than 1 at high-*TIS* conditions, there should be short-ranged non-electrostatic attraction present in the molecule to overcome the steric repulsion. Additionally, both pH 5 and 6.5 results converged towards equal B_{22} values for TIS values above 300 mM, suggesting that electrostatic contributions are completely screened and the solvation effects and van der Waals attractions present between the molecules are not affected by the differences in buffer chemistry and pH [23,51,66]. Conversely, the difference in interactions at lower *TIS* values suggests different electrostatic behavior with the change in both buffer-type and pH, going from strongly repulsive to mildly attractive (relative to steric-only interactions, $B_{22}/B_{22,ST} = 1$) as pH increases.

It is commonly accepted that, for most proteins, the total effective protein charge approaches zero as the pH of the solution approaches the isoelectric point (pI) of the molecule. Consequently, the strength of electrostatic repulsion (caused by strong charge-charge repulsions) will decrease as the pH approaches the pI of the molecule. The pI of several IgG1 molecules have been reported to lie between 7.5 and 8.5, and the present molecule has a theoretical pI of 7.9. Consequently, the decrease in repulsion with increased pH is expected based on a decrease in the total effective charge of the molecule (mostly due to deprotonated histidine residues at pH 6.5 in comparison to pH

5) and the change in the ion clouding/de-clouding that this might lead to [154,155,158]. This can be observed in Figure 3.2, where the values of the theoretical charges decrease from pH 5 to 6.5, and in some cases (*e.g.*, the C_H3 domain) this can cause a shift in sign. Since short-ranged non-electrostatic attractions are present at both pH values, this decrease in the effective protein charge with increasing pH would decrease the effective electrostatic repulsion (relative to sterics) at low *TIS*, as seen in Figure 3.4.

Collecting B_{22} data as a function of *TIS* allows one to gain insights into two of the main contributions to protein-protein solution interactions: (a) the strength and sign of net electrostatic interactions (observed at low *TIS*) and (b) the strength of short-ranged non-electrostatic attraction (observed at high *TIS*). This is better visualized in Figures 3.5 and 3.6, where data shown in Figure 3.4 were used to tune the model parameters for the HEXA and DODECA models as described above.

Figures 3.5 and 3.6 show a comparison of $B_{22}/B_{22,ST}$ *vs TIS* between experiments and simulations for the HEXA and DODECA models coupled with MSOS simulations, respectively. The experimental data and formulations are the same as those presented in Figure 3.4 and Table 3.1. Here, the shaded areas in the main panels represent the simulated $B_{22}/B_{22,ST}$ *vs TIS* profiles obtained from ARD values below 20% (gray minima in surface plots in the insets). The insets show colored surface plots of ARD as a function of ε_{SR} and ψ values. From those parameter-response surfaces, one can identify a narrow parameter space (values for ε_{SR} and ψ, also referred as [ε_{SR}, ψ] pairs) that accurately captures the low-c_2 experimental behavior with a given CG model. Here, all surface response plots show a funnel-like behavior, where a small subset of [ε_{SR}, ψ] pairs are capable of accurately modeling the experimental data within a 20% ARD. These results showed the capability of the present CG models to quantitatively capture two-particle

behavior as a function of *TIS* and at low c_2. Additionally, Figures 3.5 and 3.6 show that the currently proposed electrostatic interaction model is capable of accurately modeling the B_{22} behavior from low to high TIS as well as the plateau in B_{22} values that occurs at high *TIS* (>300 mM).

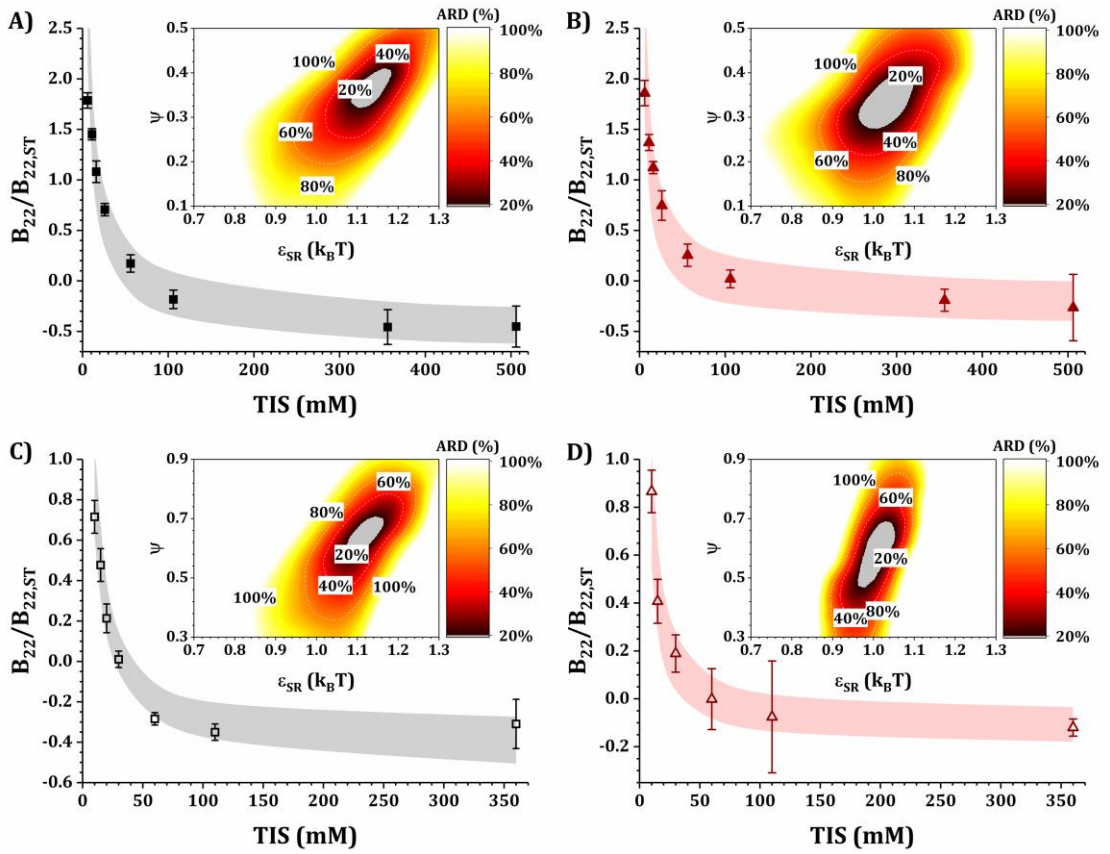

Figure 3.5. Comparison of $B_{22}/B_{22,ST}$ as a function of *TIS* between experimental (symbols) and simulated values (shaded areas) using the HEXA model at pH 5 for buffer (A) and 5% w/w added sucrose (B) conditions and at pH 6.5 for buffer (C) and 5% w/w added sucrose (D) conditions. The insets correspond to surface response of ARD values as a function of ε_{SR} and ψ.

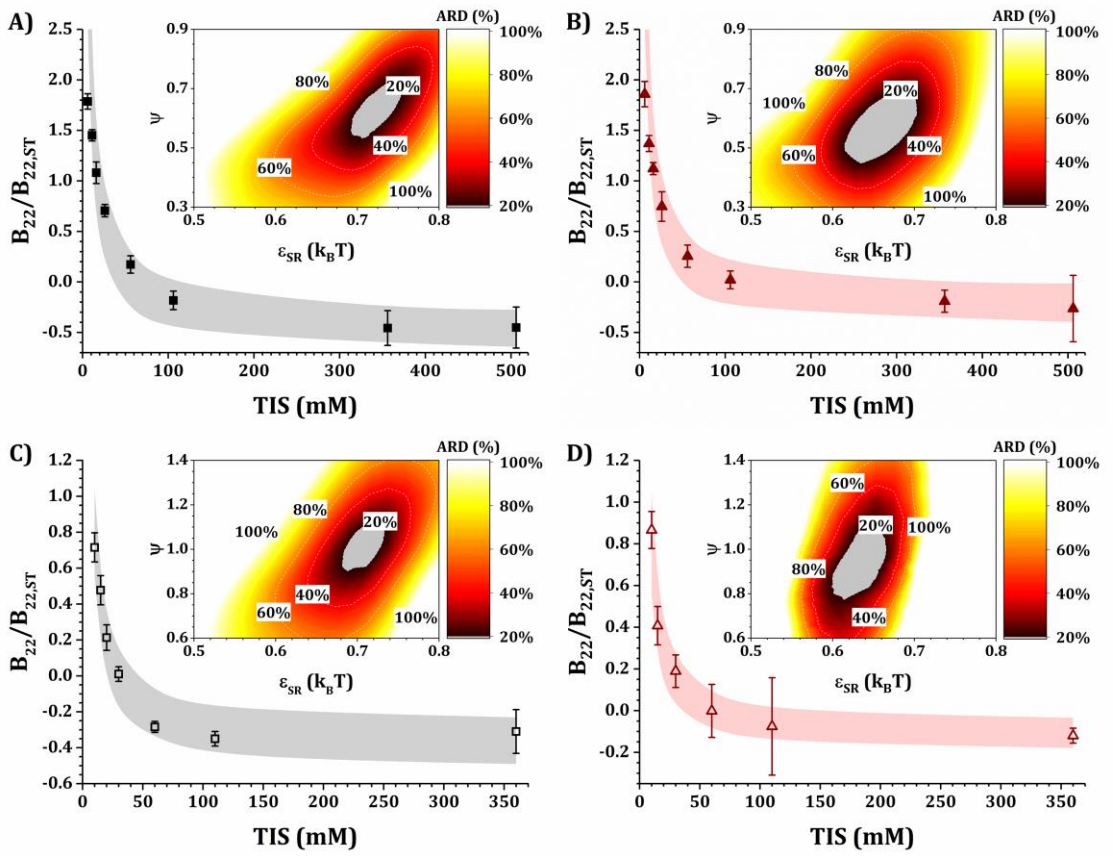

Figure 3.6. Comparison of $B_{22}/B_{22,ST}$ as a function of *TIS* between experimental (symbols) and simulated values (shaded areas) using the DODECA model at pH 5 for buffer (A) and 5% w/w added sucrose (B) conditions and at pH 6.5 for buffer (C) and 5% w/w added sucrose (D) conditions. The insets correspond to surface response of ARD values as a function of ε_{SR} and ψ.

Differences in the values of the parameters within the gray regions are observed when comparing insets in Figures 3.5 and 3.6 across pH (panels A *vs* C, and B *vs* D), added sucrose (panels A *vs* B, and C *vs* D) and model-type (Figures 3.5 *vs* 3.6). By comparing across pH for both Figures 3.5 and 3.6, one observes that the only parameter that is significantly affected is ψ, as it shifts from ~0.35 at pH 5 to ~0.65 at pH 6.5 for the HEXA model, and from ~0.65 at pH 5 to ~1.0 at pH 6.5 for the DODECA model.

This increase in ψ can potentially be explained by a decrease in ion binding due to smaller net charges in the protein molecule (see Figure 3.2 and discussion above) and the possible change in binding affinity of the ions [154,156]. Consequently, the solvation of ions around the protein may change with increasing pH, causing an increase in ψ as the solution charges approach their theoretical value since $\psi \rightarrow 1$ as $q_{i,\text{eff}} \rightarrow q_i$. From the results in Figures 3.5 and 3.6, this change in ψ is only observed across changes in pH, as ψ remains constant when comparing across sucrose content (panels A *vs* B, and C *vs* D). This suggests that the addition of sucrose should only induce a significant non-electrostatic effect to the protein solution behavior. This is also observed in the experimental data, where the values of B_{22} are the same for both buffer-only and buffer + sucrose at low *TIS*, but diverge as *TIS* increases. By comparing panels A and B, and C and D, one can observe that the addition of sucrose correlates with a decrease in the value of ε_{SR}. For the HEXA model, ε_{SR} goes from ~1.1 k_BT for buffer-only to ~1.0 k_BT for buffer + sucrose for both pH values (Figure 3.5). Similarly for the DODECA model, it goes from ~0.72 k_BT to ~0.64 k_BT (Figure 3.6). This decrease in ε_{SR} and increase in B_{22} at high *TIS* with added sucrose suggests changes in the hydration shells of the protein in the form of protein-sucrose interactions, and this will be discussed further below.

Comparing Figures 3.5 and 3.6 also shows that the ε_{SR} values within the gray areas are always lower in magnitude for the DODECA model (0.62 - 0.78 k_BT) than for the HEXA model (1.0 - 1.2 k_BT) for all simulated formulations in Table 3.1. This is due to the decrease in the number of beads/domains by moving from the DODECA to the HEXA model as shown in Chapter 2. Conversely, the magnitude of ψ increases in the DODECA model in comparison to the HEXA model (see numbers above). Although the discussion for ε_{SR} in terms of the differences in the number of domains for HEXAs

vs DODECAs also applies to ψ, changes in the values of the charges also play a relevant role in this case. As expected, charges in the HEXA model are effectively twice the magnitude of those in the DODECA model (see Table 3.2 and Figure 3.2). Charge-charge interactions were modeled *via* equation 3.5, where the electrostatic potential energy is proportional to the product of the charges. Consequently, doubling the value of the charges (by going from DODECAs to HEXAs) would induce an increase in potential energy by a factor of 4, which can compensate and overcome the decrease in the number of simulated domains/beads (from 12 to 6 in this case). Additionally, larger charges are conducive to stronger ion solvation, thus lower ψ values are expected for more coarse-grained (less structurally detailed) models. Consequently, one must be cautious that ε_{SR} and ψ values used in this work and for these CG models are model specific, and likely will differ if one changes the structural resolution of the models (either higher or lower resolution).

It is anticipated that the same qualitative interaction behavior discussed above might apply at higher c_2 values. At pH 5, both buffer and buffer + sucrose conditions R^{ex}/K values are observed to lie below the steric-only behavior (net-repulsive). This agrees with the low-c_2 behavior as B_{22} values were larger than $B_{22,ST}$ ($B_{22}/B_{22,ST}$ ~ 1.8). Nevertheless, buffer + sucrose conditions are observed to be more repulsive (lower $S_{q=0}$ or R^{ex}/K) than buffer-only conditions, and this deviation is more pronounced as c_2 increases. This does not correlate with low-c_2 measurements as both pH conditions resulted in equal B_{22} values within their experimental uncertainties (Figure 3.4A). At pH 6.5, the buffer-only behavior remains net-attractive at high c_2 (above the steric-only curve) but the sucrose conditions are net-repulsive between 10 and 120 g/L, converging towards the steric-only behavior at higher c_2. Once again, these results could not be

predicted from low-c_2 information alone as all measured B_{22} results were net-attractive ($B_{22}/B_{22,ST} < 1$) at pH 6.5, with equal B_{22} results for sucrose and buffer-only formulations at low *TIS* (see discussion above). For all formulations with added NaCl, the R^{ex}/K profiles show net-attractive behavior for both pH values, with stronger attractions at pH 6.5 than at pH 5 and in good agreement with the expectations from their low-c_2 results.

Figures 3.7 and 3.8 show a comparison of the experimental high-c_2 R^{ex}/K *vs* c_2 results. The results from the HEXA and DODECA models are based on the TMMC simulations for the parameter space obtained by fitting low-c_2 data (*cf.*, Figures 3.5 and 3.6). Figure 3.7 corresponds to parameters from Figure 3.5 and the HEXA model. Figure 3.8 corresponds to parameters from Figure 3.6 and the DODECA model. The formulation conditions are the same as shown in Table 3.4. Shaded areas in the main panels in Figure 3.7 and 3.8 represent the confidence intervals of the predicted R^{ex}/K *vs* c_2 profiles using model parameters that resulted in an ARD value below 20% from the low-c_2 parameter tuning (gray regions in insets of Figures 3.5 and 3.6). The symbols in Figure 3.7 and 3.8 represent the same experimental data shown in the insets of Figure 3.4, including 95% confidence intervals as error bars. By visual inspection, the parameters obtained at low-c_2 allow the CG models to be predictive of the high-c_2 behavior within a 20% average deviation from 10 to 150 g/L. The steric-only behavior at high c_2 is also included as a reference in Figures 3.7 and 3.8. None of the predictions in Figures 3.7 and 3.8 utilize experimental data from high c_2 as inputs to the models, but required knowledge of the B_{22} values at each given solution condition.

Although there have been several studies that focus on experimentally correlating low-c_2 measurements with high-c_2 protein physicochemical behavior, results in Figure 3.4 highlight some of the short-comings of these approaches, as interactions

and solution behavior might change as the solution transitions from low to high c_2. As shown in Figures 3.7 and 3.8, the changes in $S_{q=0}$ are not monotonic (not constant G_{22}), which lead to decreased attractions relative to that seen only at low c_2. This is of greater relevance during screening of drug candidates and formulations during early stages of development, where limited access to protein material necessitates measurements at

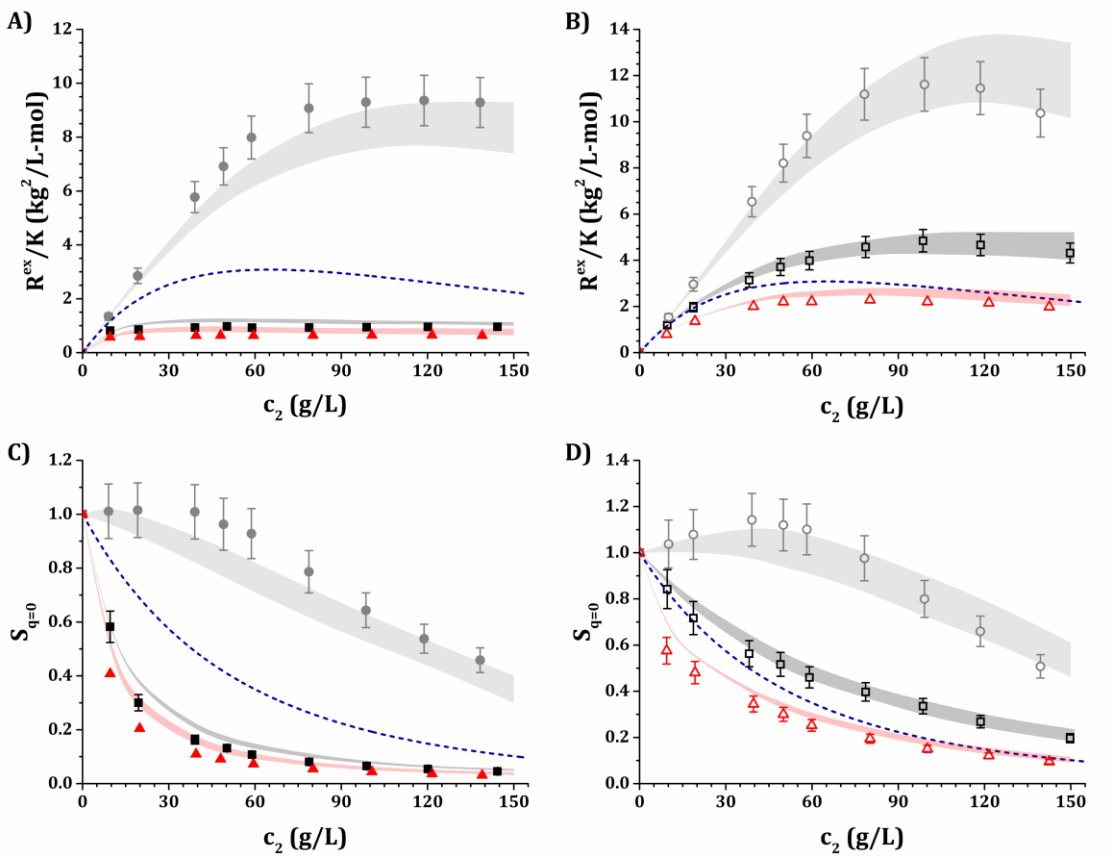

Figure 3.7. High-c_2 predictions of R^{ex}/K and $S_{q=0}$ from low-c_2 parameters with the HEXA model shown in Figure 3.5, for pH 5 (A, C) and pH 6.5 (B, D) and for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles). The symbols represent the experimental while shaded areas represent the model predictions. The blue dashed line represents the steric-only behavior.

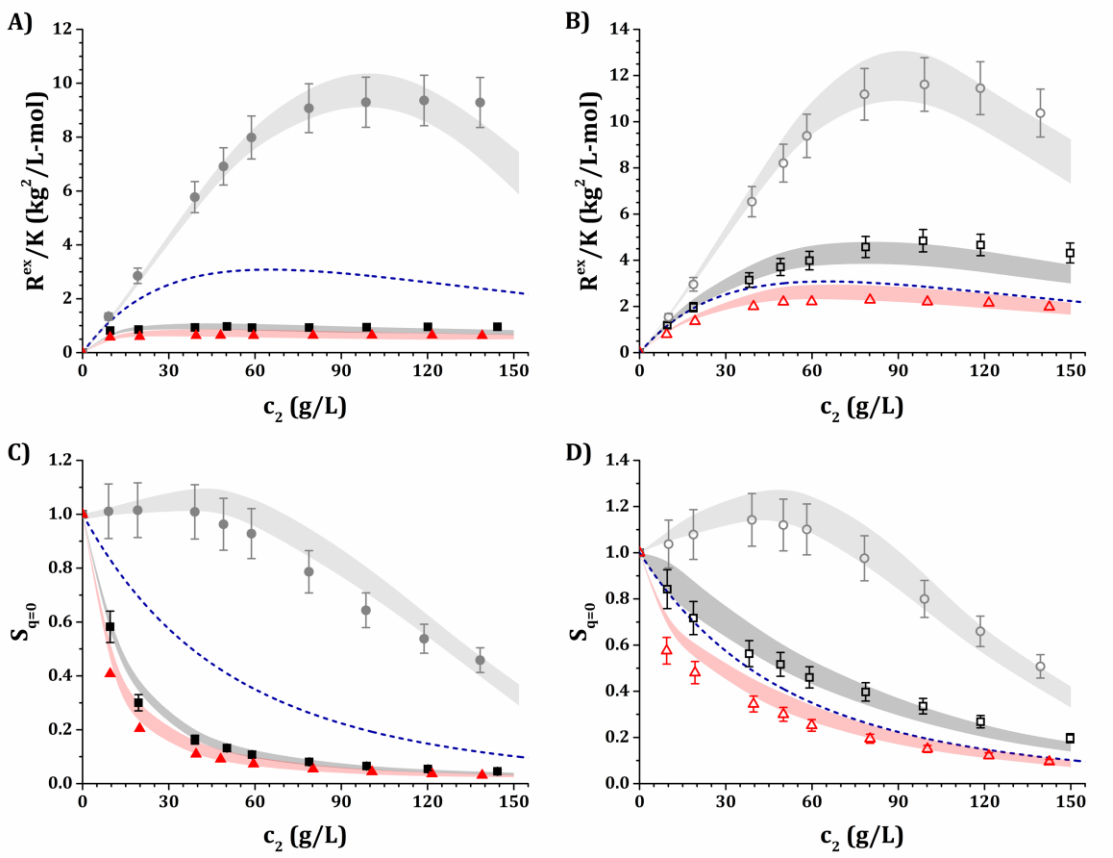

Figure 3.8. High-c_2 predictions of R^{ex}/K and $S_{q=0}$ from low-c_2 parameters with the DODECA model shown in Figure 3.6, for pH 5 (A, C) and pH 6.5 (B, D) and for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles). The symbols represent the experimental while shaded areas represent the model predictions. The blue dashed line represents the steric-only behavior.

low-c_2, dilute solution conditions [8,10]. As the solution is concentrated, the solution behavior is expected to be heavily influenced by the steric contributions based on general arguments from the statistical mechanics of liquids [59,96,142]. Thus, the shape of the molecule is expected to greatly affect the way mAb molecules interact under concentrated conditions (see Chapter 2). The addition of short-ranged interactions (either attractive or repulsive) then mediates preferentially interacting domains, which

might lead to increases in viscosity as reported in previous work [22,25,75,145,171,189]. These two effects (enhanced short-ranged interactions and packing behavior) are reasonably well captured by the HEXA and DODECA model.

The results in Figures 3.5 and 3.6 allow one to obtain a small family of [ε_{SR}, ψ] pairs that can be used to evaluate the predictive capabilities of the HEXA and DODECA models at high-c_2 conditions as shown in Figures 3.7 and 3.8. This approach shows that both models are capable of accurately predicting, not simply regressing, R^{ex}/K *vs* c_2 profiles up to 150 g/L. Small qualitative differences can be observed between the results for the HEXA (Figure 3.7) and DODECA (Figure 3.8) models. While the HEXA model shows smaller deviations at pH 5 than the DODECA model, the opposite is true for pH 6.5. At pH 5, there is a noticeable deviation for values above 120 g/L for the DODECA model where the predicted R^{ex}/K profiles decrease sooner with increasing c_2 than the experimental data (Figure 3.8A). This behavior might be caused by the geometry of the models and the ease of packing of each model. The HEXA model locates all its beads on a single plane while the DODECA model increases the complexity of the model by extending it to two planes (see Figure 3.1). This increase in geometrical complexity potentially adds stronger packing limitations at high c_2 for the DODECA model. Additionally, these two models were simulated by neglecting the flexibility of the hinge region due to limited access to data that can be used to refine such behavior (*e.g.*, SANS or SAXS). The hinge flexibility might correct for this earlier decay in R^{ex}/K *vs* c_2 for the DODECA model by easing the packing constraints of such models as suggested in Chapter 2. However, the addition of a flexible hinge would pose additional computational challenges in terms of convergence or precision of the simulations. That

would be further exacerbated if one permitted full chain flexibility and local unfolding in the simulations [160,164,190].

3.5 Experimental and Computational "Weak" Protein-Protein Interactions for the IgG4 Molecule

Similar to the IgG1 molecule, SLS measurements were performed to determine R^{ex}/K profiles of the IgG4 molecule for similar solution conditions as for the IgG1 molecule (*cf.*, Tables 3.1 and 3.5). Figure 3.9 shows the experimental results of $B_{22}/B_{22,ST}$ *vs TIS* for pH 5 and pH 6.5, all calculated for $c_2 < 5$ g/L. These $B_{22}/B_{22,ST}$ *vs TIS* profiles qualitatively differ between pH 5 and pH 6.5, and between both the IgG1 and the IgG4 molecules. At pH 5, $B_{22}/B_{22,ST}$ values decrease from 1.9 (net-repulsive) to around -0.25 (net-attractive) as *TIS* increases for the IgG4, while the same solution conditions result in values between 1.9 to around -0.5 for the IgG1. Conversely at pH 6.5, the results in Figure 3.9 show opposite behavior for the IgG4 molecule as $B_{22}/B_{22,ST}$ values increase from -9 to around -0.25 with increasing *TIS* for the IgG4, while similar solution conditions result in values between 0.9 (at low *TIS*) to -0.5 (at high *TIS*) for the IgG1. In agreement with the IgG1 molecule, increasing *TIS* by adding NaCl for the IgG4 leads to an eventual plateau in the magnitude of B_{22} for *TIS* values above approximately 300 mM in all situations. Additionally, B_{22} values at *TIS* values above 300 mM were the same for both pH values but became less attractive (less negative) with the addition of sucrose, with similar high *TIS* results in the presence of sucrose for both pH values. This is in good qualitative agreement with results in the previous subsection despite their quantitative differences.

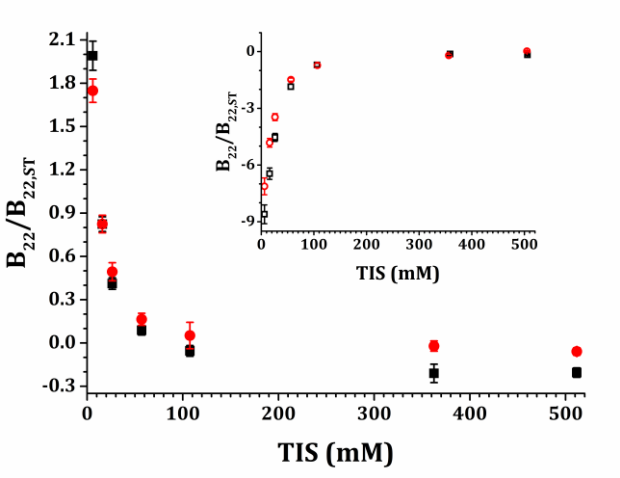

Figure 3.9. $B_{22}/B_{22,ST}$ values as a function of *TIS* for the IgG4 molecule at pH 5 (main panel) and pH 6.5 (inset) with added NaCl from 0 to 500 mM. Black symbols represent data with only buffer and added NaCl while red symbols represent the same solutions with 5% w/w added sucrose.

The results in Figure 3.9 allow one to conclude the following for the IgG4: at pH 5, this molecule experiences strong charge-charge repulsions at low *TIS*, which are screened as *TIS* increases. As B_{22} values are less negative than those of the IgG1 for *TIS* > 300 mM, the IgG4 should be subject to weaker short-ranged non-electrostatic attractions than the IgG1 (at both pH 5 and pH 6.5). However, the IgG4 experiences a different pH dependence than the IgG1 as $B_{22}/B_{22,ST}$ transitions from 1.9 (strong net-repulsive behavior) to -9 (strong net-attractive behavior) upon titration from pH 5 to pH 6.5 for *TIS* ~ 10 mM , while it only decreases from 1.9 to 0.9 for the IgG1 under similar conditions. As discussed in section 3.4 for the IgG1, this decrease in B_{22} can be partially attributed to an overall decrease in effective protein charge as the pH approaches the pI of this molecule (computed as 7.55 for the IgG4 following). If all charges on the protein surface were turned-off, the resulting $B_{22}/B_{22,ST}$ value would be that obtained at *TIS* > 300 mM (around -0.25 for the IgG4). Consequently, an additional electrostatic

phenomenon should be present at pH 6.5 for the IgG4 that is not significant for the IgG1. By analyzing the charge distribution of the molecule (*cf.*, Figures 3.2 and 3.3), one can see that there is a change in the charge sign of the C_H3 domain for both molecules by titrating from pH 5 to 6.5. However, the differences in charge values between the outermost domains (V_H, V_L at the top and C_H2 and C_H3 at the bottom) is more pronounced for the IgG4 molecule than the IgG1 molecule. Consequently, the presence of this strong charge disparity can lead to the presence of strong dipoles, which can be conducive to strong electrostatic attractions as observed at pH 6.5 for the IgG4 [75,76]. Although this might also be present on the IgG1 molecule, the magnitude of these dipole-effects is not as strong as those for the IgG4. Therefore, the net charge-charge repulsion dominates the B_{22} results for the IgG1, leading to positive (yet net-attractive) B_{22} values in contrast to the large negative B_{22} values for the IgG4. Similar dipole dominated behavior can be observed for other protein solutions and it is shown for a globular protein in Chapter 4 [75]. Finally, the addition of sucrose to the IgG4 solutions also led to increases in B_{22} as observed for the IgG1, and mostly dominant at $TIS > 300$ mM. This suggests a decrease in the effective hydrophobicity of the IgG4 as hypothesized above. This will be further explored in Chapter 5.

Figure 3.10 shows the experimental R^{ex}/K *vs* c_2 (high-c_2) results at pH 5 and pH 6.5 for the formulations presented in Table 3.5 and that correspond to the $B_{22}/B_{22,ST}$ measurements in Figure 3.9. Additionally, the steric-only behavior for this molecule is shown as a reference, as computed using the VE EoS model. The results in Figures 3.10A (pH 5) show that R^{ex}/K profiles for both buffer and sucrose formulations are net-repulsive (R^{ex}/K values below the steric-only behavior) at low-c_2 conditions (below 80 g/L), but transition to net-attractive (R^{ex}/K values above the steric-only behavior) at

much higher c_2 values (> 80g/L), with R^{ex}/K profiles for sucrose below (more repulsive than) those for buffer-only under all c_2 values. The addition of 100 mM NaCl results in a substantial increase in R^{ex}/K values above the steric-only behavior across the whole concentration range, and greater than the R^{ex}/K results of the IgG1. In contrast, the results in Figure 3.10B (pH 6.5) show that all R^{ex}/K profiles are net-attractive. The buffer and sucrose formulations show a steep increase in R^{ex}/K values, to the extent that solutions above 40 g/L could not be characterized as the scattering signal sharply increases above the limits of detection on available instrumentation. This rapid increase in R^{ex}/K values with c_2 does correlate with the measured B_{22} values, and can be expected from a strongly attractive system. The addition of 100 mM NaCl results in a large decrease in R^{ex}/K values, but larger than the results of the IgG1. This further suggests that the strong attractions are caused by strong electrostatic (and attractive) interactions as short-ranged non-electrostatic attractions are weaker for the IgG4 than the IgG1. This is in good agreement with the B_{22} results for the IgG1 and IgG4. Both results at pH 5 and 6.5 agree with the low-c_2 (B_{22}) measurements, as buffer conditions were always more attractive than sucrose conditions, and increasing *TIS* resulted in a decrease (increase) in repulsions (attractions) at pH 5, with the inverse *TIS* dependence at pH 6.5.

Combining the results at low and high c_2 for the IgG4 shows a partial disagreement in comparison to the IgG1. Similar B_{22} values were obtained for equivalent formulations at pH 5 for both molecules. Net-repulsive behavior was observed across the whole c_2 range at pH 5 without added NaCl for the IgG1, while B_{22} was less attractive for the IgG4 than the IgG1 at high *TIS* conditions. Based solely on B_{22} results, one might conclude that the IgG4 is more colloidally stable (*i.e.*, less prone to strong attractions) than the IgG1. However, this is not the case at high c_2 as the IgG4

experiences stronger attractions than the IgG1 (see discussion above), highlighting the need to measure the interaction behavior of each molecule at the c_2 of interest and avoiding the assumption that low-c_2 interactions are directly predictive of high-c_2 behavior because this inherently neglects the high entropy contributions that arise from packing constraints (see Chapter 2) and added interactions.

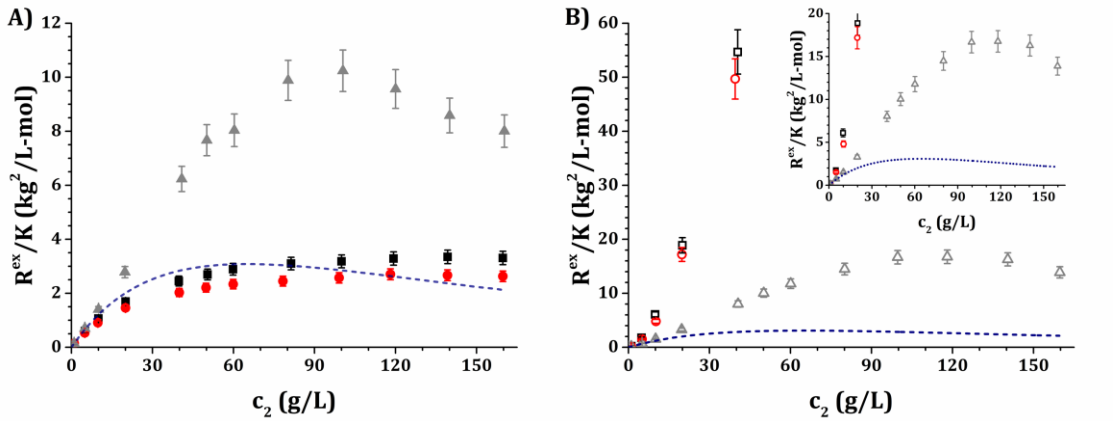

Figure 3.10. R^{ex}/K values as a function of c_2 for the IgG4 molecule at pH 5 (A) and pH 6.5 (B) for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles).. Black symbols represent data with only buffer and added NaCl while red symbols represent the same solutions with 5% w/w added sucrose. The blue dashed-line represents the sterics-only behavior from the VE-EoS.

Similar to section 3.4, $B_{22}/B_{22,ST}$ *vs TIS* values were compared between experiments and simulations using the HEXA and DODECA models coupled with MSOS simulations. This allows one to reconstruct ARD *vs* [ε_{SR}, ψ] surface responses to obtain parameters that model the low-c_2 SLS results. This was first done at pH 5 and the results are shown in Figure 3.11 for the HEXA and DODECA models, where the experimental data and formulations are the same as those presented in Figure 3.9 and

Table 3.1. The shaded areas represent the simulated $B_{22}/B_{22,ST}$ *vs TIS* profiles obtained from ARD values below 20%. These parameters were later used to predict high-c_2 behavior at pH 5 and the formulations in Table 3.5. Figure 3.12 shows a comparison of the experimental and predicted high-c_2 R^{ex}/K results as a function of c_2 using the HEXA and DODECA models. Those are based on the TMMC simulations for the parameter space obtained by fitting low-c_2 data as explained above. As before, the shaded areas represent the model predictions with the symbols representing the experimental data. R^{ex}/K and the quantity ($S_{q=0}$ - $S_{q=0,ST}$) are plotted as a function of c_2, with $S_{q=0,ST}$ representing the steric-only zero-q structure factor calculated using the VE-EoS.

The results in Figure 3.5, 3.6 and 3.11 show that the selected models (HEXA and DODECA) with the selected force fields (Figure 3.1, equations 3.3 and 3.4, and those in Chapter 2) are capable of modeling B_{22} (low-c_2) behavior for mAb molecules and conditions from net-repulsive to mildly net-attractive ($B_{22}/B_{22,ST} > -1$). Similarly, Figures 3.7, 3.8 and 3.12 show that the resulting models can capture the high-c_2 behavior of both molecules. However, the current approach (*i.e.*, assuming low-c_2 parameters are constant and predictive of high-c_2 conditions) is not always accurate across molecules, solution formulations and/or models. Under some conditions, the DODECA and HEXA models accurately predict the high-c_2 behavior (*e.g.*, IgG1, both pH 5 and 6.5, NaCl). In some cases, the HEXA model performed better predictions than the DODECA models (*e.g.*, IgG1, pH 6.5, buffer), and *vice versa* (*e.g.*, IgG4, pH 5, NaCl). Conversely, none of the models were capable of quantitatively or semi-quantitatively predicting high-c_2 from only low-c_2 parameters for certain conditions, especially for the IgG4 (*e.g.*, IgG4, pH 5, buffer). This highlights the need for additional refinements with high-c_2 information as there are additional phenomena

that can be altered by increasing c_2 (*e.g.*, a change in the counter-ions and solvation layers of the protein as the average protein-protein distance is reduced at high c_2).

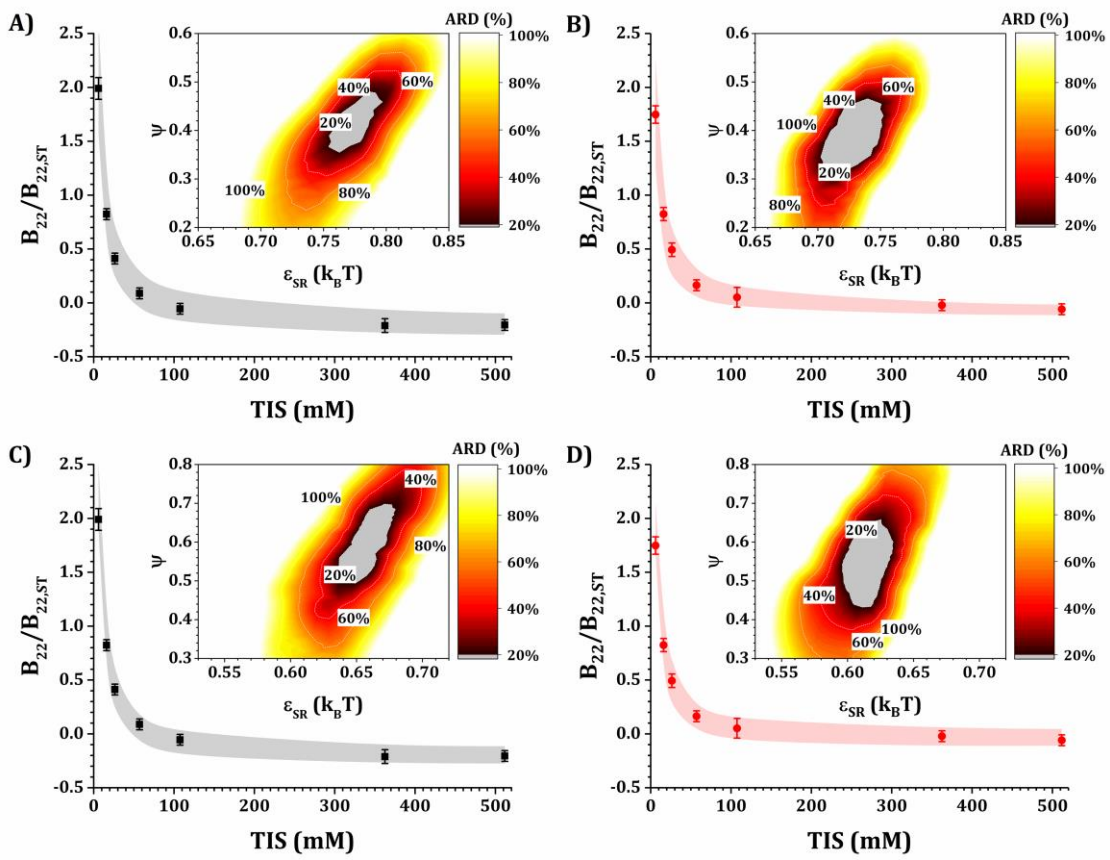

Figure 3.11. Comparison of $B_{22}/B_{22,ST}$ as a function of *TIS* between experimental (symbols) and simulated values (shaded areas) using the HEXA and DODECA models at pH 5 for buffer (A: HEXA, B: DODECA) and 5% w/w added sucrose (C: HEXA, D: DODECA) conditions. The insets correspond to surface response of ARD values as a function of ε_{SR} and ψ.

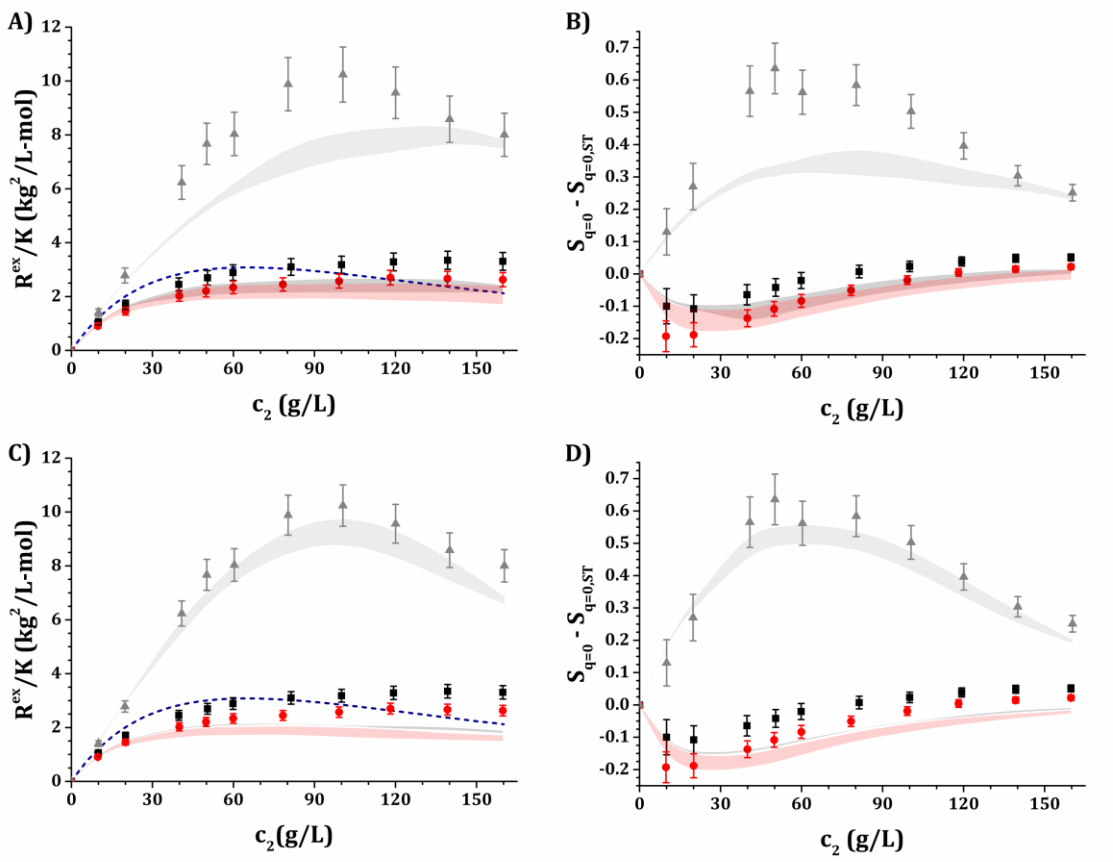

Figure 3.12. High-c_2 predictions of R^{ex}/K and ($S_{q=0}$ - $S_{q=0,ST}$) from low-c_2 parameters with the HEXA (panels A and B) and DODECA (panels C and D) model shown in Figure 3.11, and for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles).

In spite of the previous results, there are conditions where the current approach does not apply for both low and high c_2. This can be seen in Figure 3.13, where $B_{22}/B_{22,ST}$ *vs TIS* values, and R^{ex}/K and ($S_{q=0}$ - $S_{q=0,ST}$) *vs* c_2 at pH 6.5 were modeled using the DODECA model. Comparable results were obtained for conditions with the HEXA model, so those have been excluded to avoid an excessive number of figures. In Figures 3.13A and 3.13B, one can observe that the DODECA model is incapable of quantitatively capturing the experimentally measured $B_{22}/B_{22,ST}$ *vs TIS* behavior, and no

parameter set was found to result in less than 40% ARD for both panels. Consequently, only those parameters that minimized the ARD and fit the measured $B_{22}/B_{22,ST}$ values at either 0 mM NaCl or 100 mM were used to construct Figures 3.13C and 3.13D. Here, one can see that the model is incapable of predicting the high-c_2 behavior for both buffer and sucrose conditions. On the other hand, the model does qualitatively predict the results at 100 mM NaCl, with large deviations as observed for some conditions at pH 5. It is worth pointing out that these conditions are highly attractive, with attractions being caused by electrostatic interactions. Similarly, protein charges are physically located on the surface of the protein, and not in the center as assumed in the present models. This can also be seen by the unphysically high ψ values in the inset of Figure 3.13A and 3.13B as values of ψ are expected to lie between 0 and 1 as this parameter is expected to represent the change in effective charge due to ion binding. Strong net-attractive behavior is highly dependent on molecular orientation, proximity, packing and available volume, so low resolution models, such as the HEXA and DODECA models, are incapable of capturing the highly complex nature of strong protein attractions. This can be addressed with high resolution CG model, which will be explored in section 3.8.

The magnitude of the interactions observed for the IgG4 at pH 6.5 and low *TIS* conditions can be conducive to phase separation, high solution viscosity and other physicochemical instabilities [17,19,23,25,145,173,191]. Consequently, these conditions are avoided during the development of protein based solutions. Despite the model lacking enough resolution to accurately predict the high-c_2 behavior, it was capable of modelling strong electrostatic attractions at low *TIS* and low c_2 (see Figure 3.13, panels A and B). Consequently, the HEXA and DODECA model can still be useful

to interpret low-c_2 data for strongly attractive conditions similar to the approaches explored in section 3.7.

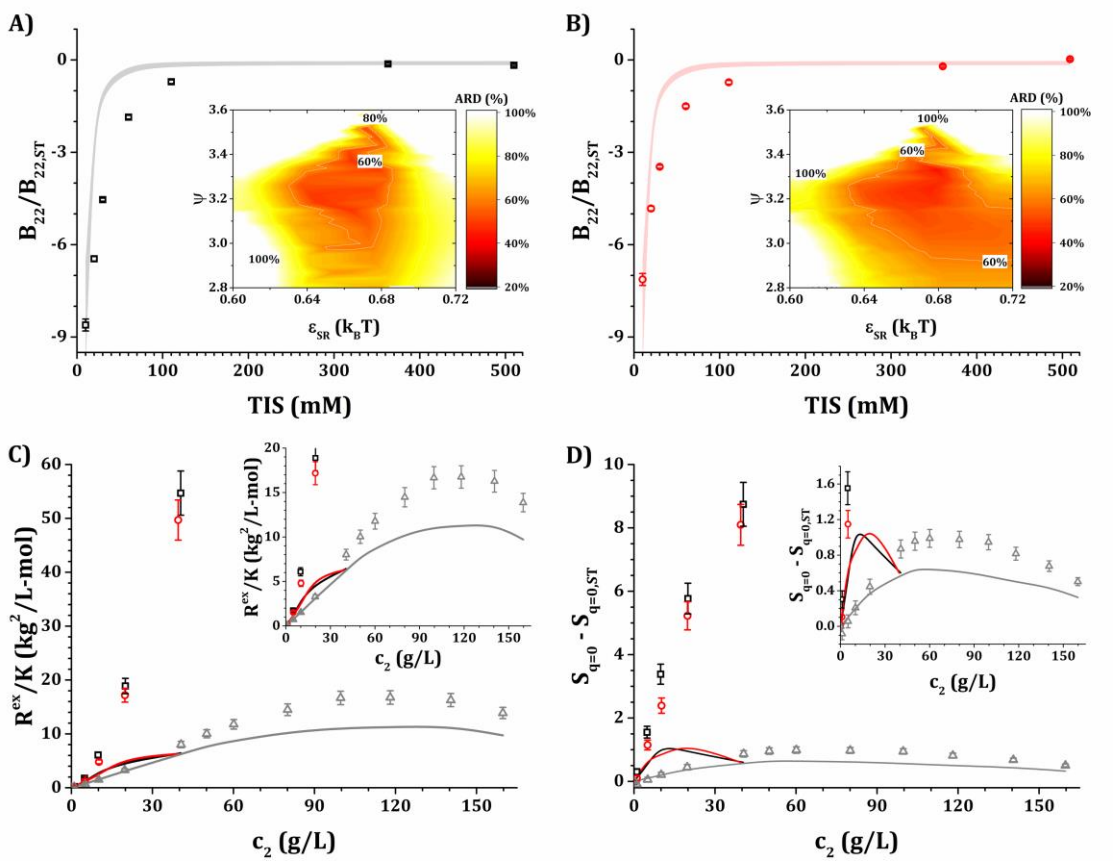

Figure 3.13. Panels A and B: Comparison of $B_{22}/B_{22,ST}$ *vs TIS* between experimental (symbols) and simulated (shaded areas) values using the DODECA model at pH 6.5 for buffer (A) and sucrose (B) conditions, with their respective ARD surface responses as a function of ε_{SR} and ψ in the insets. **Panels C and D:** High-c_2 predictions of R^{ex}/K and ($S_{q=0}$ - $S_{q=0,ST}$) from low-c_2 parameters with the DODECA model at pH 6.5 and for buffer-only (black squares), 5% w/w sucrose (red triangles) and 100 mM NaCl (gray circles).

3.6 TMMC vs MSOS Simulations for Predicting High-c_2 Interactions

The computational approach used to simulate high-c_2 protein interactions in sections 3.4 and 3.5 relied on simulating hundreds to thousands of molecules in an open box, and allowing for equilibration [68,170]. Although this is done to capture both the energy and entropy contributions to the solution as a function of c_2 to better assess the effects of increasing c_2, this approach might not be ideal if one is more interested in optimizing the time needed to obtain a predicted data set than the overall accuracy of the prediction (within a reasonable value). In this situation, the approach taken by the McMillan-Mayer solution theory becomes of interest, where a generic EoS can be created based on a polynomial expansion (*i.e.*, the virial expansion) where the only parameters needed are the so-called osmotic virial coefficients (see equation 3.7). These virial coefficients can be computed using the MSOS algorithm as explained in section 3.2.4.3. Consequently, this subsection explores the viability of using such an approach to predict high-c_2 behavior, by using the accurate results from the TMMC simulations as the targeted result.

Figure 3.14 shows four illustrative cases for all the parameter sets obtained in previous subsections for the HEXA and DODECA models. Here, $S_{q=0}$ *vs* c_2 profiles were computed using the TMMC algorithm or the MSOS algorithm coupled with truncated versions of equation 3.7 (depending on the number of virial coefficients used) in conjunction with equation 3.8, for the HEXA and DODECA models. Figure 3.14 illustrates that the MSOS approach (*i.e.*, using equations 3.7 and 3.8) can replicate the TMMC simulation if one uses up to the fourth (4^{th}) virial coefficient and, in some situations, the fifth (5^{th}) virial coefficient for conditions that exhibit net-repulsion (Figures 3.14A and 3.14C) to weak net-attraction (Figures 3.14B) up to 150 g/L of protein concentration. However, adding more virial coefficients is needed to

quantitatively capture conditions that exhibit moderate to strong net-attraction (Figure 3.14D). This is also in agreement with previous analyses of the short-comings of using the virial EoS to capture solution behavior of concentrated systems [148,176,180].

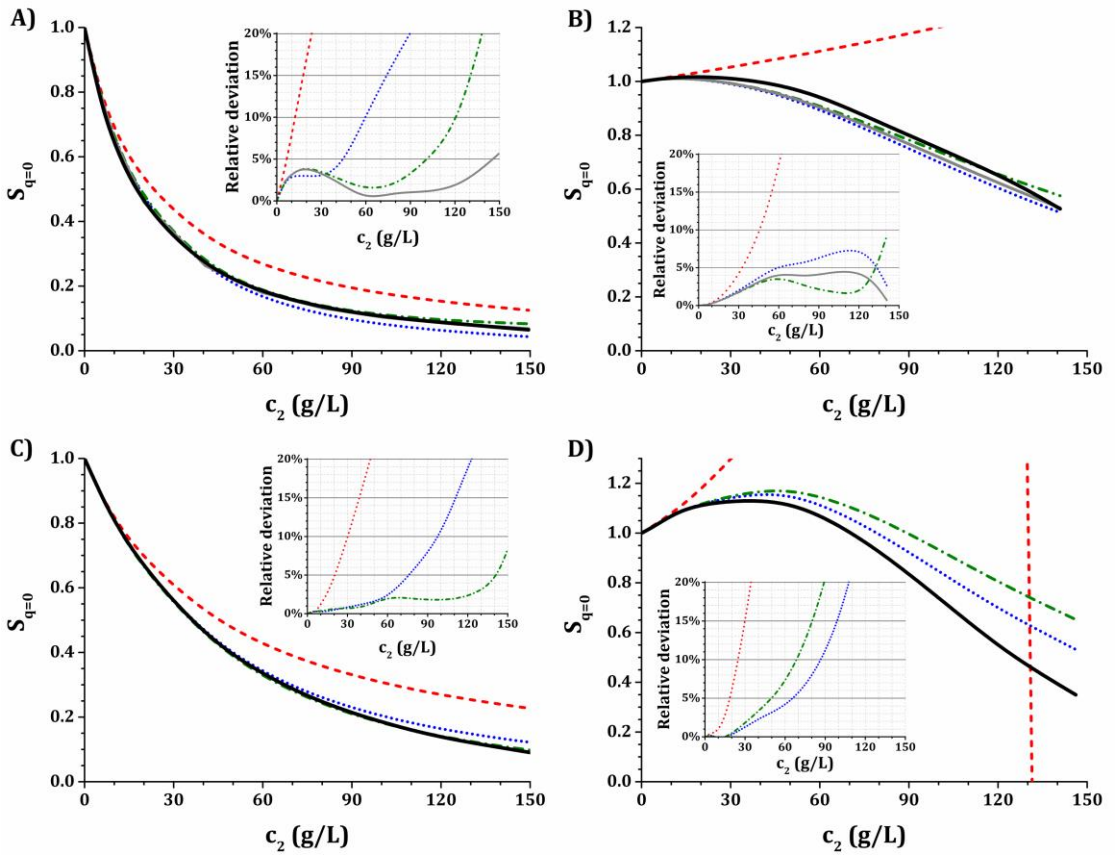

Figure 3.14. Comparison of the MSOS and TMMC approaches for case studies using the HEXA (A & B) and DODECA (C & D) models. Black solid lines represent the TMMC results, while red dashed, blue dotted, green dash-dotted and gray solid lines represent the MSOS results with up to the 2nd, 3rd, 4th and 5th virial coefficient, respectively. Insets correspond to the relative deviation as a function of c_2 for each model using the TMMC results as the reference.

Table 3.6 shows the measured $B_{22}/B_{22,ST}$ values for each of the formulations in Table 3.5 and for both mAb molecules, and whether or not the MSOS approach equals the TMMC approach up to 150 g/L within 10% relative deviation. Here, it can be seen that most conditions could be modeled with the MSOS algorithm, mostly excluding conditions with the IgG4 at pH 6.5. This suggest that the presence of strong electrostatic attractions require higher structural definition (see discussion above and in Chapter 2). Table 3.6 also suggests a preliminary $B_{22}/B_{22,ST}$ threshold to determine if the MSOS approach is accurate. The MSOS approach was accurate for the IgG1 at pH 5 with 100 mM NaCl, but inaccurate for the IgG1 at pH 6.5 with 100 mM in comparison with the TMMC approach. Consequently, the present results would suggest a value of $B_{22}/B_{22,ST} > -0.2$ as a limit at which the MSOS approach up to the 5^{th} virial coefficient is equivalent to the TMMC approach for $c_2 < 150$ g/L. However, additional molecules and solution conditions should be explored before a definite threshold is set.

Unfortunately, it is challenging to compare the TMMC and MSOS approaches based on practical use since additional factors, such as sampling optimization and convergence, are difficult to predict for different c_2 values. High-c_2 conditions require larger numbers of simulated molecules in the TMMC approach. This increases both the time to convergence and the total simulation time. Consequently, selecting one approach over the other must be based on the requirements of the user. Using the MSOS algorithm requires running a minimum of 4 or 5 molecular simulations (one for each virial coefficient), while the TMMC algorithm only requires a single but more comprehensive simulation. Consequently, if the goal is to obtain a single c_2 prediction, it would be wiser to run a single TMMC simulation without the need to corroborate whether the MSOS approach is accurate. On the other hand, if a series of concentrations is desired, it would

be wiser to use the MSOS algorithm when possible (*e.g.*, $B_{22}/B_{22,ST} > -0.2$) as this provides an analytical EoS with further flexibility. Finally, one needs to consider that the MSOS approach will not be able to provide additional information about the solution structure at very high c_2, which is discussed in the following subsection.

Table 3.6. Viability of using the MSOS approach for previously discussed formulations and mAb molecule

Molecule	Formulation	$B_{22}/B_{22,ST}$	MSOS ~ TMMC?
IgG1	pH 5, buffer	1.79 ± 0.08	Yes
	pH 5, sucrose	1.86 ± 0.13	Yes
	pH 5, NaCl	-0.18 ± 0.09	Yes
	pH 6.5, buffer	0.72 ± 0.08	Yes
	pH 6.5, sucrose	0.87 ± 0.09	Yes
	pH 6.5, NaCl	-0.35 ± 0.05	No
IgG4	pH 5, buffer	2.0 ± 0.1	Yes
	pH 5, sucrose	1.75 ± 0.08	Yes
	pH 5, NaCl	-0.06 ± 0.04	Yes
	pH 6.5, buffer	-8.6 ± 0.5	No
	pH 6.5, sucrose	-7.1 ± 0.4	No
	pH 6.5, NaCl	-0.71 ± 0.09	No

3.7 Domain-Domain Contact Maps *via* $g_{ij}(r)$ from Molecular Simulations

Beyond modeling and predicting low- and high-c_2 interaction parameters (*e.g.*, B_{22} and G_{22} values), there is additional interest in obtaining information regarding what pair(s) of contacts are responsible for strong molecular attractions or repulsions. This information could be further used to better engineer any protein of interest [188,192–194]. This can be achieved for any mAb molecule using domain-domain $g_{ij}(r)$ simulations (see section 3.2.4.4). Figure 3.15 shows two examples of $g_{ij}(r)$ values as a

function of c_2, for C_H3-C_H3 and V_H-V_H domain-pairs. These results can be transformed into potentials of mean force and used to compute the minimum free energy upon contact (within 1 nm) for each pair of domain-domain interactions for the DODECA model using equation 3.9. Figure 3.16 shows the results for a purely attractive system (*i.e.,* $\psi = 0$) for c_2 = 10, 50, 100 and 150 g/L. Likewise, Figures 3.17 and 3.18, and Figures 3.19 and 3.20 show the result for pH 5 and pH 6.5, respectively, for conditions with only buffer (3.17 and 3.19) and with 100 mM added NaCl (3.18 and 3.20) applied to the IgG1 molecule.

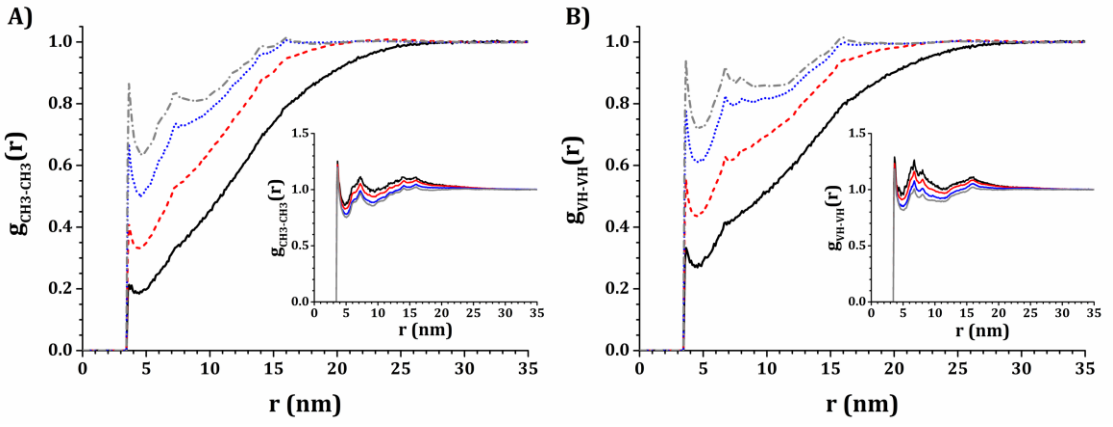

Figure 3.15. Example of $g_{ij}(r)$ results for C_H3-C_H3 (panel A) and V_H-V_H (panel B) pairs at 10 g/L (black), 50 g/L (red), 100 g/L (blue) and 150 g/L (gray). Insets correspond to the same conditions in main panel but with $\psi = 0$.

As can be seen in Figure 3.15, there is a notable change in $g_{ij}(r)$ as a function of c_2. In the main panels of Figure 3.15, the likelihood of observing close contact (within 1 nm) between domains increases as a function of concentration for repulsive conditions for both the V_H and C_H3 domains. This is likely driven by the reduced space among

proteins at high c_2 since $g_{ij}(r)$ represents the likelihood of having a given mAb domain within a defined distance r, averaged over all other domains and molecules in solution [148]. However, the likelihood of close contacts decreases as a function of c_2 when only short-ranged non-electrostatic attractions are present (insets in Figure 3.15). This might suggest a limitation in the free-packing of proteins at high-c_2 conditions (*e.g.*, crowding effects [142]), leading to a decrease in the effective attractions among protein molecules due to packing limitations as discussed in Chapter 2. This can be better seen by analyzing Figures 3.16-3.20.

Figure 3.16 shows that under net-attractive conditions, there is a high likelihood of observing favorable (negative energy) contacts involving the outermost external beads (C_H3, V_H and V_L domains) such us C_H3-V_H, V_H-C_H2 and V_L-C_H1 contact-pairs, among others. This is highlighted by red filled areas in Figure 3.16. However, this likelihood decreases with increasing c_2 despite the molecules only experiencing steric repulsions and short-ranged non-electrostatic attractions, in agreement with the observations in Figure 3.15 (insets). Moreover, the interactions between external (C_H3, V_L and V_H) and internal (C_H2, C_H1 and C_L) beads transition from favorable at 10 g/L (red colored in panel A), to unfavorable at 150 g/L (blue colored in panel D). Similarly, all internal beads (C_H2, C_H1 and C_L domains) experience weak favorable-contacts at low c_2 among themselves (*e.g.*, C_H2-C_H1, C_H1-C_L contacts, *etc.*), which transition to unfavorable (positive energy) contacts at higher c_2. These results for the potential of mean force upon contact for a purely attractive case suggests that there is a significant entropy penalty for internal beads to interact due to the packing limitations inherent in the elongated shape of a mAb molecule. Additionally, this entropy penalty increases at higher c_2 values.

The addition of like charges to all the protein domains might lead to the case represented in Figure 3.17. For strongly repulsive conditions, there is very low likelihood of observing close contact among protein beads at low c_2 as seen in Figure 3.17A. Increasing c_2 leads to a decrease in the contact free-energy (more favorable interactions) across external beads as seen in Figure 3.17D. Similarly, the likelihood of these contacts follows the charge distribution shown in Figure 3.2. The charge of the V_L domain is lower than that of the V_H domain, which is lower than that of the C_H3 domain. This can be seen in Figure 3.17, where the molecules have a tendency to first orient their V_L and V_H domains into contact as c_2 increases (panels A *vs* B) before including C_H3 domains (panels B *vs* C). Additionally, a lower interaction free-energy can be observed involving V_L domains than any other inter-domain interactions (C & D) at high c_2, in agreement with this domain having the smallest charge. This leads to a preference for V_L domains to interact with themselves. This is of relevance in cases where V_L domains are more aggregation prone, so preventing their likelihood to interact closely might decrease the aggregation rates of a protein as a whole [113,192]. Similarly, this can be the case for the C_H3 and V_H domains as these are external domains. Due to their higher charge value and the entropy penalty mentioned above, only unfavorable contacts between internal beads is observed across the whole c_2 range in Figure 3.17.

Figure 3.18 shows the effect of partially screening the charge-charge repulsive interactions with increasing *TIS*. Here, it can be observed that interactions between external beads is favorable across the whole c_2 range, with interactions involving internal beads only being relevant at low c_2, in agreement with the results in Figure 3.16. This is in part due to the entropy penalty mentioned above and the higher charge values

of the internal beads, which lead to less favorable contacts in comparison to external beads as also observed in Figure 3.17.

Figures 3.19 and 3.20 show that a change in the charge distribution can change the preferential interaction map. Figure 3.2 suggests that the C_H3 domain transitions from positive to negative charge from Figure 3.17 to Figure 3.19, which leads to preferentially attractive interactions with all the other domains in the molecule. This can be seen in all four panels in Figure 3.19, where the interaction energy between any domain and C_H3 domain is always lower than for the same domain interaction with a non-C_H3 domain. This behavior was observed for the V_L domain in Figure 3.17 due to the change in charge values, and this highlights how changes in the charge distribution (*e.g.*, by mutations, chemical degradation or pH titration) can cause a shift in preferential interactions across protein domains [23,111,192,195]. As observed in Figures 3.16-3.18, only external domain contacts are favorable at high c_2 for Figure 3.19, once again highlighting the entropy penalty that arises from the packing constraints experienced by these molecules. Finally, Figure 3.20 shows that screening charge-charge interactions leads to a map that closely resembles Figure 3.16, with small modifications caused by the charge values as mentioned above. This is expected as any contact map should converge to Figure 3.16 (or its equivalent depending on the value of ε_{SR}) as charges are heavily screened and only short-ranged non-electrostatic attractions dominate the interactions across domains.

Interestingly, the contacts between internal beads (*i.e.*, C_H2, C_H1 and C_L domains) were observed to be unfavorable across the five studied conditions, including cases that are net-attractive (Figures 3.16, 3.18 and 3.20). These results were obtained for the condition where all beads have the same short-ranged non-electrostatic attraction

strength. Consequently, a highly energetic contribution must be present between internal beads to overcome this entropic penalty if one experimentally observes preferential contact across internal domains. Additionally, if internal contacts were experimentally present at high c_2, there would be a strong energetic penalty for disrupting such strong domain-domain interactions. The contribution of these strong internal interactions coupled with the shape of these protein molecules might be responsible for the formation of stable protein networks, protein phase separation and high solution viscosity as has been observed and hypothesized for some mAb solutions in the literature [40,75,78,124,171,172,196–199]. Interestingly, the results presented here suggest that it would be possible to identify favorable contacts between beads at as low as 10 g/L of protein concentration, and that this behavior might be predictive of the high c_2 behavior. Although these five figures are not expected to represent the whole experimental space for all mAb proteins, the results presented in this chapter are expected to highlight how minor changes in the chemical identity of some protein domains might affect the overall solution behavior of a mAb solution.

The results from Figures 3.16-3.20 are shown as cases of study and potential uses for the HEXA and DODECA models beyond zero-q limit (SLS) data. Consequently, additional experimental data with higher structural resolution might be needed to better understand domain-domain preferential interaction. One possibility is using SAXS and SANS experiments [28,34,200]. However, additional refinement of the models is required for this to be the case. This is out of the scope of this dissertation, and potential uses and examples are discussed in Chapter 7.

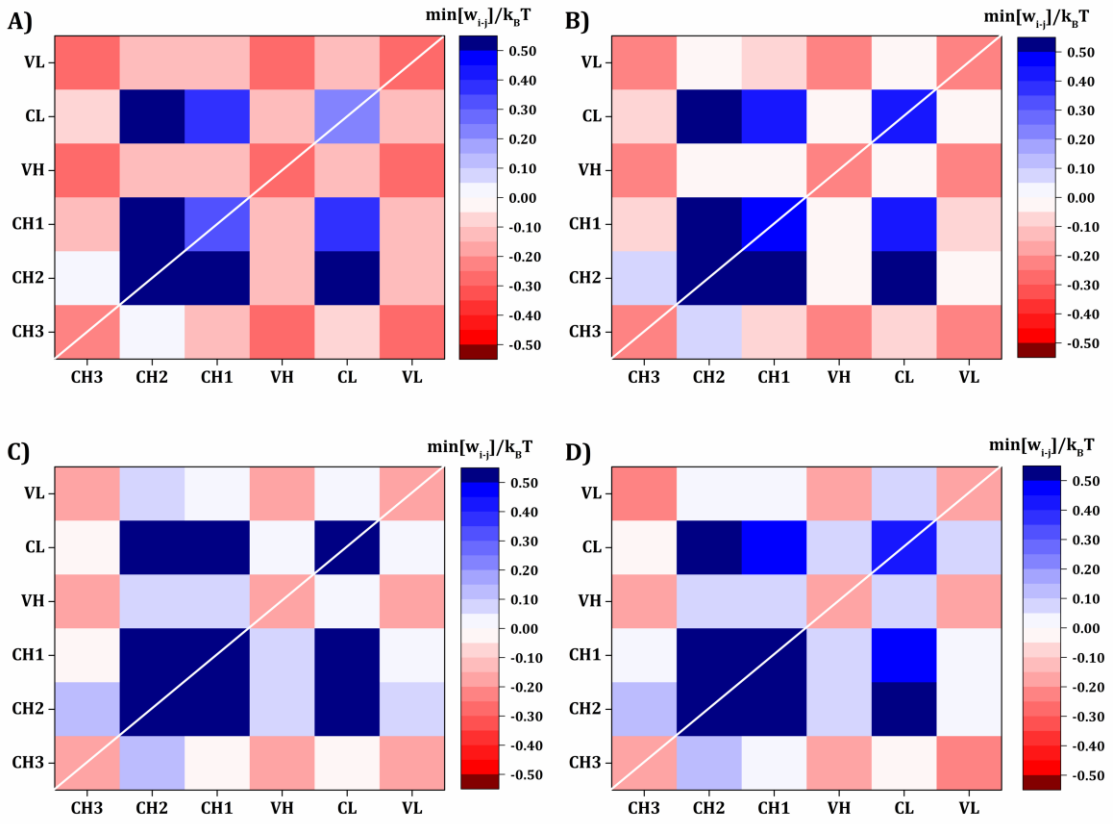

Figure 3.16. Domain-domain contact map for attractive conditions (ε_{SR} = 0.72 k_BT, ψ = 0) and for c_2 equal to 10 (A), 50 (B), 100 (C) and 150 (D) g/L.

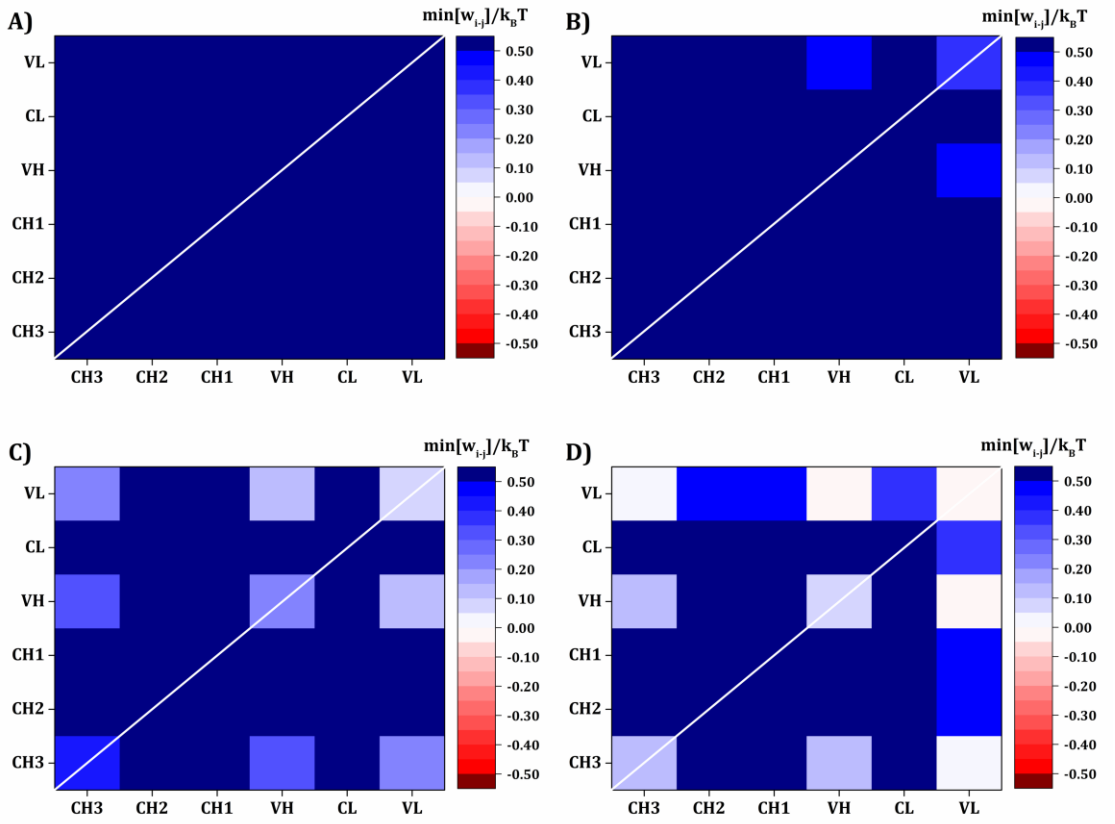

Figure 3.17. Domain-domain contact map for a condition that fits the IgG1, pH 5, buffer conditions (ε_{SR} = 0.72 k_BT, ψ = 0.48, *TIS* = 6 mM) and for c_2 equal to 10 (A), 50 (B), 100 (C) and 150 (D) g/L.

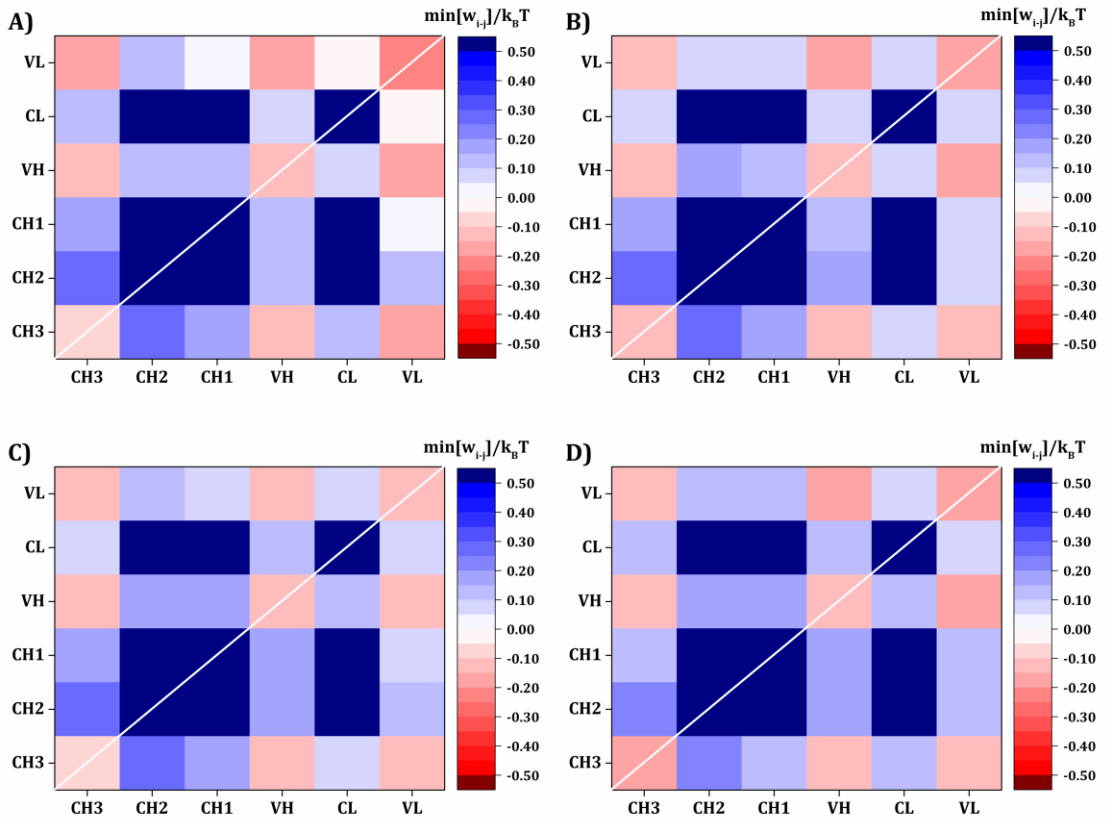

Figure 3.18. Domain-domain contact map for a condition that fits the IgG1, pH 5, 100 mM NaCl conditions (ε_{SR} = 0.72 k_BT, ψ = 0.48, *TIS* = 106 mM) and for c_2 equal to 10 (A), 50 (B), 100 (C) and 150 (D) g/L.

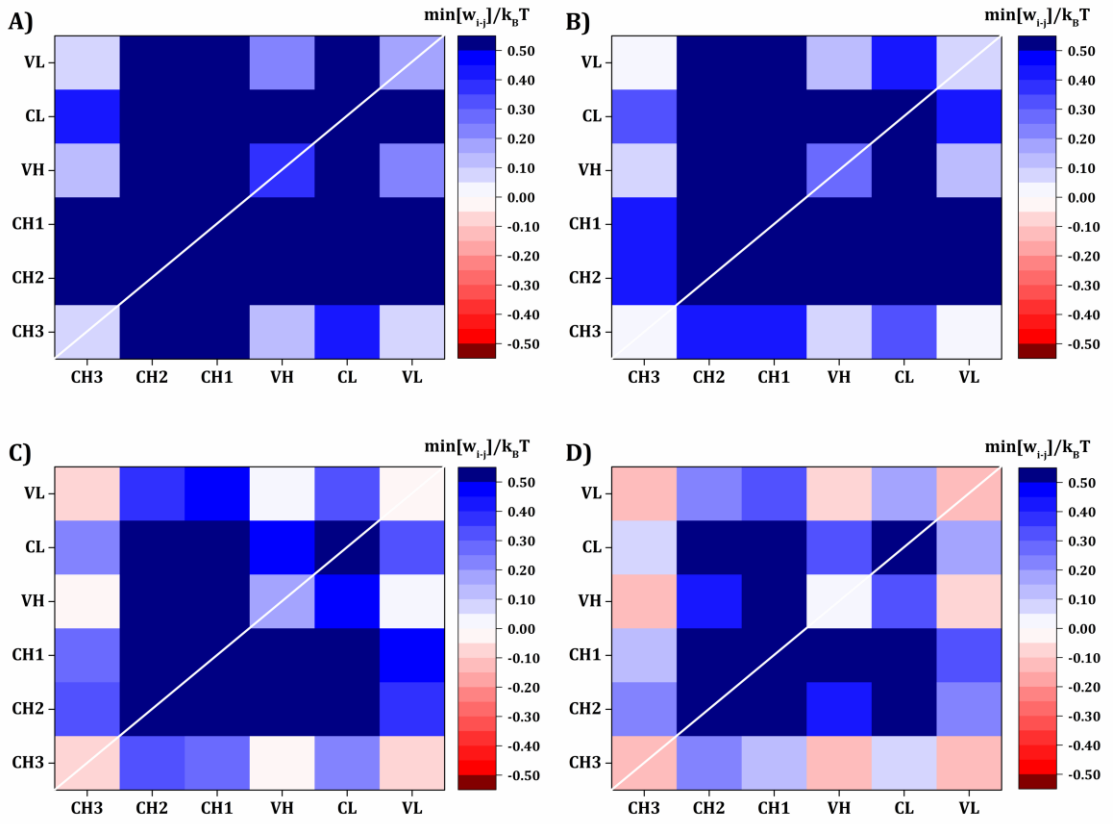

Figure 3.19. Domain-domain contact map for a condition that fits the IgG1, pH 6.5, buffer conditions (ε_{SR} = 0.72 k_BT, ψ = 0.9, *TIS* = 10 mM) and for c_2 equal to 10 (A), 50 (B), 100 (C) and 150 (D) g/L.

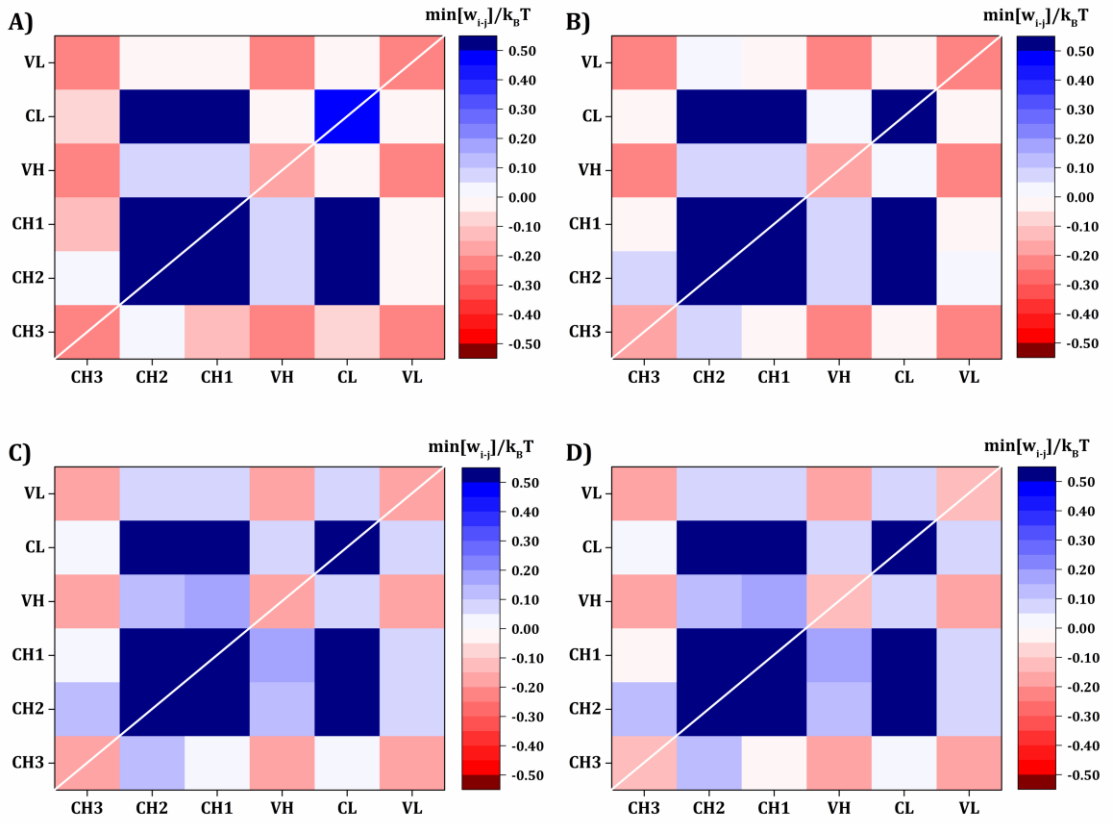

Figure 3.20. Domain-domain contact map for a condition that fits the IgG1, pH 6.5, 100 mM NaCl condition (ε_{SR} = 0.72 k_BT, ψ = 0.9, *TIS* = 110 mM) and for c_2 equal to 10 (A), 50 (B), 100 (C) and 150 (D) g/L.

3.8 Capturing B_{22} Behavior with an Amino Acid Resolution CG Model

To assess whether or not the anisotropic surface distribution of charged residues can predict the transition from repulsive to strongly attractive electrostatic interactions from pH 5 to pH 6.5 for the IgG4 molecule, $B_{22}/B_{22,ST}$ response surfaces were calculated by computing $B_{22}/B_{22,ST}$ as a function of ε_{SR}, ψ and *TIS* using the 1bAA model as explained in section 3.2.3.2. The value of ε_{SR} was selected to assure that accurate $B_{22}/B_{22,ST}$ values would be obtained at high *TIS* conditions for both mAb molecules. Figure 3.21 shows the results of evaluating $B_{22}/B_{22,ST}$ as a function of ε_{SR} for a system

with no electrostatic contributions (*i.e.,* $\psi = 0$). Figure 3.21 shows that the IgG4 molecule is more sensitive to values of ε_{SR} than the IgG1 molecule as the same value of ε_{SR} provides lower (more negative) $B_{22}/B_{22,ST}$ values for the IgG4 than the IgG1. Since the experimental results in Figures 3.4 and 3.9 combined would suggest that the B_{22} values for the IgG4 molecule are higher (less attractive) than the IgG1 molecule at high-*TIS* conditions (above 300 mM), the results in figure 3.21 suggest that the value for ε_{SR} to properly model the experimental $B_{22}/B_{22,ST}$ results is lower for the IgG4 molecule (ε_{SR} = 0.44 k_BT) than for the IgG1 molecule (ε_{SR} = 0.50 k_BT). Similarly, the results in figure 3.21 would suggest that the IgG4 molecule is more sensitive to perturbations in ε_{SR} than the IgG1 molecule. These results also suggest than the 1bAA model might can capture the non-electrostatic tendencies of protein molecules. Additionally, the discrepancies between low and high c_2 combined with the current results might suggest that non-ideal cosolute-protein interactions are present in the protein solutions that makes the IgG1 behave more attractively than the IgG4 at pH 5 and 6.5 at low-c_2 and high-*TIS* conditions but change with increasing c_2. A framework to analyze these interactions will be explored in Chapter 5.

After finding a value of ε_{SR} that can model the non-electrostatic contributions to $B_{22}/B_{22,ST}$, response surfaces can be constructed of $B_{22}/B_{22,ST}$ as a function of ψ and *TIS*. These response surfaces are expected to show one of three limiting cases, depending on the degree of anisotropy of the surface charge distribution: (1) monopole dominated behavior, such that $B_{22}/B_{22,ST}$ is large and positive at low *TIS*, and decreases monotonically with increasing *TIS* (*e.g.*, Figure 3.4); (2) multipole dominated behavior, such that $B_{22}/B_{22,ST}$ is large and negative at low *TIS*, and increases monotonically with increasing *TIS* (*e.g.*, inset in Figure 3.9); (3) $B_{22}/B_{22,ST}$ shows a transition between

monopole- and multipole-dominated regions as ψ increases. This behavior could also be observed as a function of *TIS* at a fixed ψ value. However, the experimental measurements had previously set the value for *TIS*, so the discussion below is developed from the perspective of ψ as a degree of freedom.

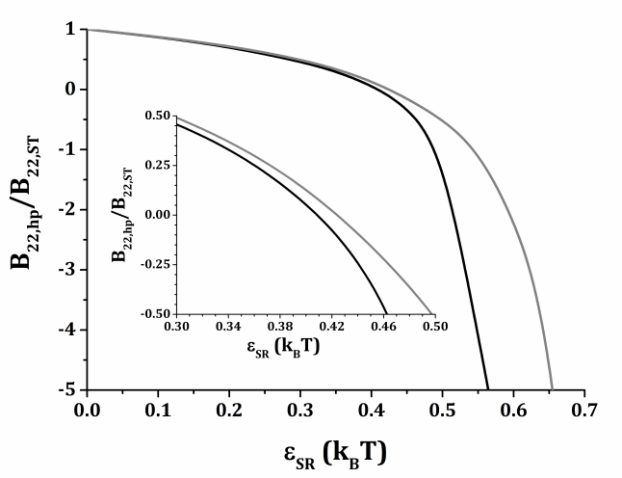

Figure 3.21. B_{22} as a function of ε_{SR} using the 1bAA CG model the IgG1 (gray) and the IgG4 (black) molecules.

The shape of the response surface should depend on the solution pH as well as the protein sequence and structure. Because these are CG models using implicit solvent, the effective values for ψ and ε_{SR} can be modified experimentally by changing the properties of the solution – *e.g.*, by adding additional excipients that mediate protein-protein interactions, as well as by specific-ion effects that lead to preferential exclusion or accumulation of ions near the protein surface (see Chapter 5). Figures 3.22 and 3.23 show the response surfaces for pH 5 and pH 6.5, respectively, for the IgG1 and IgG4 using atomistic homology models provided by BMS. The two figures show that the

IgG1 and IgG4 exhibit a case of monopole dominated behavior (case 1 above) at pH 5, and this behavior continues for the IgG1 at pH 6.5. However, the IgG4 experiences a multipole-dominated behavior at pH 6.5 (case 2, above). These results are in quantitative or semi-quantitative agreement with measured $B_{22}/B_{22,ST}$ values (Figures 3.4 and 3.9), and provides further evidence that the charge distribution of both molecules are dominating their interaction behavior. Finally, these results suggest that this type of response surface has the potential to be used more generally as a tool to assess how anisotropic surface-charge distributions affect protein-protein interactions without arbitrary definitions of charge "patches" or other geometric measures of anisotropic interactions that are difficult to generalize, and allows one to better infer the effect of point mutations in the overall colloidal stability of the molecule [111,153,192]. However, this model is currently limited to low-c_2 simulations as it would be computationally intractable to apply the TMMC approach to this molecule with current computational infrastructures.

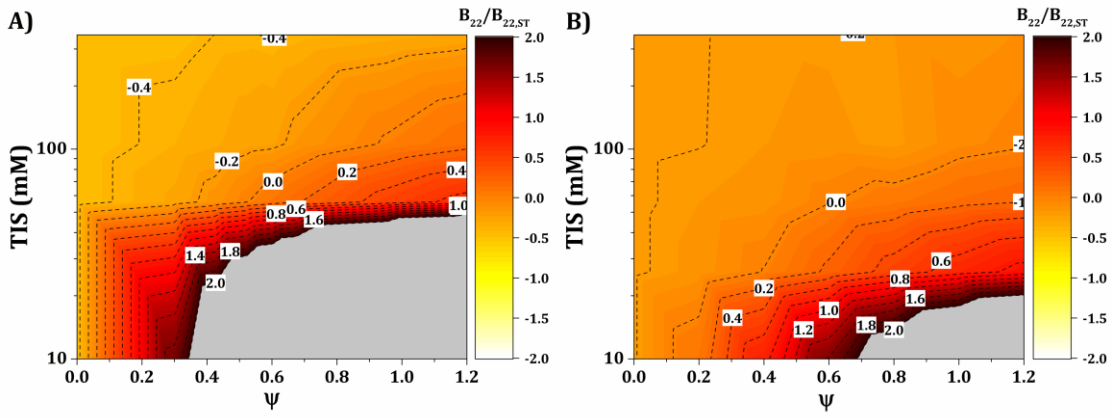

Figure 3.22. $B_{22}/B_{22,ST}$ response surface of the 1bAA CG model as a function of ψ and *TIS* for the IgG1 (panel A: ε_{SR} = 0.5 k_BT) and the IgG4 (panel B: ε_{SR} = 0.44 k_BT) at pH 5.

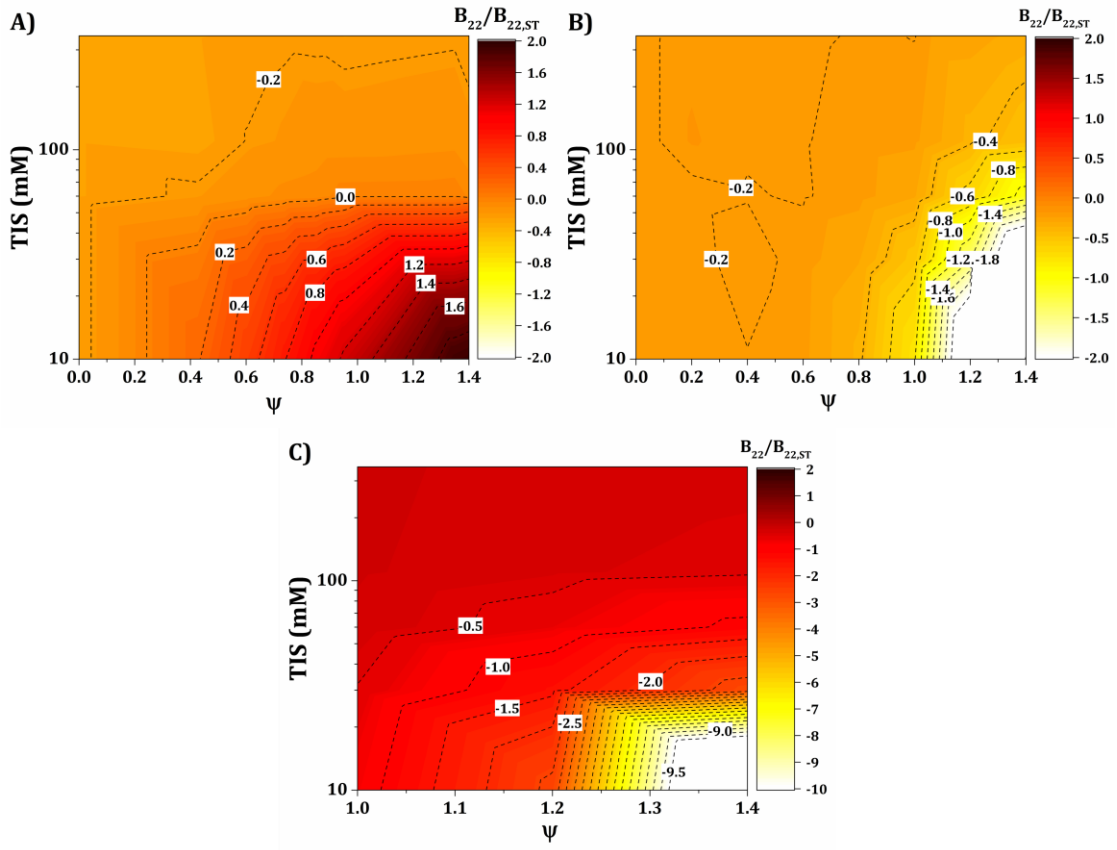

Figure 3.23. $B_{22}/B_{22,ST}$ response surface of the 1bAA CG model as a function of ψ and *TIS* for the IgG1 (panel A: ε_{SR} = 0.5 k_BT) and the IgG4 (panels B and C: ε_{SR} = 0.44 k_BT) at pH 6.5. Panel C is a zoomed-in version of Panel B with a different $B_{22}/B_{22,ST}$ range for easier visualization.

3.9 Summary and Conclusions

Static light scattering was used to quantify "weak" protein-protein interactions of two mAb molecules as a function of c_2 for a range of pH and *TIS* values, and sucrose concentration. These included conditions that resulted in both net-repulsive and net-attractive protein interactions at low *TIS*, and at low- to high-c_2 conditions. Two coarse-grained molecular models were tested to evaluate their potential to predict excess Rayleigh profiles and zero-q structure factors at high c_2. Additional domain-domain

preferential contact maps were built based on simulations of radial distribution functions and potentials of mean force. Low-c_2 results showed that the IgG1and IgG4 molecules exhibits net-repulsive behavior at low *TIS* and pH 5, which transitions to net-attractive behavior as the solution *TIS* increases. At pH 6.5, the IgG1 antibody showed weakly net-attractive behavior from low to high *TIS*. Conversely, the IgG4 showed strong net-attractive behavior at low *TIS* caused by strong electrostatic attractions. At high TIS (>300 mM), statistically equal values were found at both pH conditions, with weaker attractions observed for the IgG4 than the IgG1 in this *TIS* regime. For all measured pH and *TIS* conditions, the addition of 5% w/w sucrose to the solution induced weaker net-attractions with increasing *TIS*. This behavior was also observed at high c_2, where formulations with 5% w/w sucrose always resulted in more repulsive behavior. For conditions without sucrose present and at high c_2, buffer-only formulations shifted from net-repulsive behavior (relative to steric-only interactions) at pH 5 to net-attractive at pH 6.5, while formulations with 100 mM NaCl resulted in net-attractive behavior relative to steric contributions at both pH values and for both molecules. Domain-domain preferential contact maps showed the added potential of these low-resolution CG models to provide additional domain-based information that can be used to better engineer proteins with good colloidal stability. Finally, the 1bAA model was used to identify if anisotropic charge distributions were responsible for strong attractions for the IgG4 at pH 6.5. The results obtained from the 1bAA model quantitatively and qualitatively agree with those measured experimentally, and demonstrated that the IgG4 molecule transitions from repulsive electrostatic interactions at pH 5 to attractive electrostatic interactions at pH 6.5, solely based on the molecule sequence and homology model.

In terms of model predictions from low to high c_2, the quantitative differences were not statistically significant for net-repulsive to weakly net-attractive conditions, and therefore both models could be used to accurately predict high-c_2 behavior depending on the requirements of the user (*e.g.*, computational burden and molecular features). However, the HEXA and DODECA models failed to predict high-c_2 interactions based only on low-c_2 parameters for strongly attractive conditions (*e.g.*, IgG4 at pH 6.5). An additional approach using the MSOS algorithm to compute osmotic virial coefficients was tested and found to mimic the results from the TMMC algorithm for conditions where $B_{22}/B_{22,ST} > -0.2$. The simulations results showed that both CG models, the HEXA and DODECA models, were able to quantitatively or semi-quantitatively predict the experimental data based solely on parameters obtained by combining B_{22} *vs TIS* experimental and simulated data collected at low c_2.

Chapter 4

PREDICTIONS OF PROTEIN-PROTEIN INTERACTIONS OF A GLOBULAR PROTEIN AT HIGH CONCENTRATIONS WITH MOLECULAR SIMULATIONS AND LOW CONCENTRATION EXPERIMENTAL MEASUREMENTS

4.1 Introduction

Chapters 2 and 3 discussed the potential of using CG models to predict high-c_2 "weak" protein-protein interactions for mAb solutions using solely low-c_2 experimental information and structural information from available crystal structures. This chapter further examines the challenge of using CG models to predict experimental protein interactions (*via* excess Rayleigh scattering) from low to high c_2 for a globular protein. The excess Rayleigh scattering (R^{ex}/K) profiles of a model protein, α-chymotrypsinogen A (aCgn), are experimentally determined as a function of pH, *TIS* and c_2. Experimental B_{22} values are used to parameterize potential of mean force (PMF) models as a function of *TIS* and pH, in terms of the strength of short-ranged non-electrostatic attractions, the effective net charge, and the effective dipole moment of aCgn at a given pH. The model parameterization is done without knowledge of the high-c_2 behavior, as done in Chapter 3. The experimental high-c_2 R^{ex}/K results are then predicted using the low-c_2 parameters for the same PMF models in transition matrix Monte Carlo simulations. The results are discussed from both qualitative and quantitative perspectives, highlighting strengths and weaknesses of the approach for repulsive and attractive conditions. Finally, a higher-resolution coarse-grained (CG) model is used to rationalize and potentially predict the qualitative change from repulsive

to attractive electrostatic interactions, without the need for defining or assuming the existence of relevant charged “patches” or other geometric measures of anisotropic surface charge distributions. The experimental data presented in this chapter were adapted from a previously published work and the all the results have been published in a peer-reviewed journal [201].

4.2 Material and Methods

4.2.1 Buffer and Protein Solutions Preparation

Sodium phosphate buffer stock solutions were prepared by dissolving sodium phosphate monobasic anhydrous (Fisher Scientific, Fair Lawn, NJ) in deionized water (MilliQ, Millipore-Sigma) to reach 5 mM sodium phosphate, and titrated to pH 7.0 ± 0.05 (termed pH 7 below) using a 5 M sodium hydroxide solution (Fisher Scientific). Stock sodium acetate buffer (40 mM) was prepared by dissolving anhydrous sodium acetate (Fisher Scientific) and glacial acetic acid (Fisher Scientific) in distilled, deionized water and titrating to pH 5.0 ± 0.05 (termed pH 5 below). Stock salt buffer solutions for both pH 5 and pH 7 were prepared using the same procedures as above with the gravimetric addition of 10, 50, 100, or 300 mM NaCl (Fisher Scientific). All buffer solutions were filtered and stored at 4 °C prior to use.

Bulk protein solutions were prepared from 5X crystallized lyophilized aCgn (Worthington Biochemical, Lakewood, NJ) dissolved in stock buffer solution (pH 7) to a protein concentration of ~15 g/L. A stock solution of 35 g/L phenylmethlylsulfonyl fluoride (PMSF) was prepared by dissolving PMSF (Fluka Chemical, Ronkonkoma, NY; Sigma-Aldrich, St. Louis, MO) in 100% anhydrous ethanol (Decon Labs, King of Prussia, PA). The bulk aCgn solution (40 mL at ~15 g/l) was treated incrementally (in

250 µL aliquots) with 1 mL of 35 g/L PMSF (10x PMSF mole excess) in order to deactivate potential proteolytically active residual proteases in the commercial material [45]. The resulting protein solutions were then filtered and dialyzed using 10 kDa molecular-weight-cutoff (MWCO) dialysis membrane (Spectr/Por, Spectrum Laboratories, Rancho Dominguez, CA) in the desired buffer with four 12-hr buffer exchanges at 4°C to remove any residual salt impurities from the commercial protein material.

Concentrated protein stock solutions (>100 g/L) were prepared by membrane centrifugation at ~3200 RCF using 10 kDa MWCO Amicon-Ultra centrifugal tubes (Millipore). UV-VIS spectrophotometry (Agilent 8453, Santa Clara, CA) was used to determine the protein concentration using a wavelength of 280 nm with an extinction coefficient of 1.97 L g^{-1} cm^{-1} [202]. Lower-c_2 protein samples were then prepared by gravimetrically diluting the concentrated protein solution in the desired buffer to obtain c_2 values ranging from 1 to 100 g/L.

4.2.2 Static Light Scattering Measurements

Static light scattering (SLS) experiments were conducted using a Wyatt Technology (Santa Barbara, CA) DAWN HELEOS II instrument with laser wavelength of 658.9 nm at a temperature of 25°C. The same methodology employed in Chapter 3 was used here (section 3.2.2). In summary, G_{22} values were approximated as $-2B_{22}$ in the limit of dilute protein concentrations (*i.e.*, $c_2 < 10$ g/L). A negative (positive) G_{22} value is equivalent to a value of $S_{q=0}$ below (above) 1, corresponding to net-repulsive (net-attractive) interactions relative to an ideal gas mixture. Correspondingly in dilute solutions, positive (negative) B_{22} indicates net repulsion (attraction) relative to the ideal case of two non-interacting point particles.

4.2.3 Transition Matrix Monte Carlo Simulations and Simulated R^{ex}/K Values

Transition matrix Monte Carlo (TMMC) was used to compute excess Rayleigh ratios (R^{ex}) and zero-q limit structure factors ($S_{q=0}$) as a function of protein concentration. Details of the methodology are the same as in section 3.2.4.2. An initial uniform distribution was used for $\Pi(N_2)$, and the updated distribution was subsequently reconstructed at the end of each cycle until it converged to the equilibrium distribution, with each cycle being defined as 10^6 MC attempts. A MC attempt consisted of one of the following randomly selected moves: a translation, a rotation (for CG models with a dipole, see below) or a molecule insertion or deletion. Regular movements (translations and rotations) represented the initial 30% of the MC attempts, while deletions and insertions represented the other 70%. The temperature was kept constant at 298.15 K. Preliminary simulations were used to find an adequate value of the reference chemical potential, depending on the choice of the interaction model (see section 3.2.4.2 and below). A box length of up to 180 nm was used and the simulation box was started with an empty system. The final scaled R^{ex}/K values were obtained by using equation 1.3 with a M_w value of 25.7 kDa and assuming that $M_{w,app} \approx M_w$. This might lead to an initial uncertainty as high as 5% depending on the measured $M_{w,app}$ at low-c_2 (see Chapters 2 and 3, and results below) [23,29,81].

4.2.4 Interaction Potential Models

TMMC simulations were carried out using PMF models as pairwise additive force fields, as is typically done in colloidal simulations and as in Chapters 2 and 3 [27,61,76,77]. These were based on a classical spherical model that included a square-well potential (u_{sw}) to model steric repulsion and short-ranged non-electrostatic attraction (*e.g.*, van der Waals, hydration and excluded volume effects) [148]:

$$u_{sw}(r_{ij}) = \begin{cases} \infty, & r_{ij} \leq \sigma \\ -\varepsilon_{sw}, & \sigma < r_{ij} < \lambda_{sw}\sigma \\ 0, & otherwise \end{cases} \quad (4.1)$$

and a screened electrostatic potential (u_{el}) decomposed into three contributions: monopole-monopole interactions (u_{qq}), monopole-dipole interactions ($u_{q\mu}$) and dipole-dipole interactions ($u_{\mu\mu}$) [76]

$$u_{el} = u_{qq} + u_{q\mu} + u_{\mu\mu} \quad (4.2)$$

The electrostatic interactions in equation 4.2 were modeled using the equations shown by Bratko *et al.* (equations 1-7 in reference [76]) and reproduced in Appendix A. In equation 4.1, the parameters of interest are the well-width ($\lambda_{sw}\sigma$), the well-depth or strength of short-ranged non-electrostatic attractions (ε_{sw}), and the effective hard-sphere diameter of the protein (σ). The value of σ was selected to provide an accurate value of the steric contribution to B_{22} ($B_{22,ST}$) so as to match the value reported for an all-atom structure of aCgn (see Chapter 2). The resulting value of σ was 4.65 nm based on results published by Grünberger *et al*. [63]. λ_{sw} was set to 1.1, giving an effective range of non-electrostatic attractions equal to 4.65 Å. ε_{sw} was used as a fitting/tuning parameter as done in Chapter 3 with ε_{SR} (the different subscript was used to highlight the different non-electrostatic potential used in this chapter compare to Chapter 2 and 3). The parameters that dictate the contributions from u_{el} are the effective net charge (Q_{eff}) and the effective dipole moment of the protein (μ_{eff}) in solution (see Appendix A). The values of Q_{eff} and μ_{eff} were determined using the procedure below.

The formal definition of B_{22} is given in equation 1.1, where $w_{22}(c_2 \rightarrow 0, r, \Omega_1, \Omega_2)$ is the PMF between a pair of proteins before any averaging of the orientation-dependent

contributions [59,76,148]. It is evaluated in the dilute limit (denoted by $c_2 \rightarrow 0$), such that multi-protein interactions are irrelevant [131,148]:

$$w_{22}(c_2 \rightarrow 0, r_{ij}, \Omega_1, \Omega_2) = u_{sw}(r_{ij}) + u_{el}(r_{ij}, \Omega_1, \Omega_2) \tag{4.3}$$

This PMF is a function of the inter-protein distance (r_{ij}), and two orientational degrees of freedom (Ω_1, Ω_2). However, these two orientational degrees of freedom can be pre-averaged by computing an orientation-averaged Boltzmann factor of the interaction potential, as shown in equation 1.5 [76]. In this sense, equation 1.1 can be reduced to a one-dimensional integral with respect to r_{ij}:

$$B_{22} = -\frac{1}{2}\int_0^{\infty}\left[\exp\left(-\frac{\overline{w}_{22}(c_2 \rightarrow 0, r_{ij})}{k_B T}\right) - 1\right] dr \tag{4.4}$$

This allows one to compute B_{22} more efficiently using numerical integration to solve the integral in equation 4.4. Consequently, an orientation-averaged dipole equation was used to compute B_{22} values [76]. The electrostatic interactions were decoupled into an orientation-independent interaction (monopole-monopole) and an orientation-averaged interaction (monopole-dipole and dipole-dipole interactions) using the mathematical models proposed by Bratko *et al.* (equations 9 to 22 in reference [76] and reproduced in equations A.8 to A.12 in Appendix A). The orientational-averaged electrostatic model involves the same adjustable parameters: Q_{eff} and μ_{eff}. B_{22} was subsequently obtained by numerical integration of equation 4.4 using Matlab™ (Mathworks Inc., Natick, MA).

4.2.5 B_{22} from a High-Resolution CG Model Coupled with Mayer Sampling with Overlap Sampling

A structurally high-resolution CG model was used to evaluate whether the anisotropic distribution of charges would be expected to lead to multipole dominated

behavior (see Chapter 3). The previously developed one-bead-per-amino acid (1bAA) model was used to compute B_{22} using the Mayer sampling method employing the overlap sampling algorithm (MSOS), as in Chapter 3. The updated 1bAA force field proposed in Chapter 3 was used here. This is written in simplified notation and shown in equations 3.4 and 3.5, with additional parameters in Table 3.3. In summary, the 1bAA model treats each amino acid as a single bead, and the charge (q_i or q_j) for a given amino acid resides at the center of that bead [66]. Charges reside in the center of the side-chain bead for charged amino acids. Based on nominal pKa values, at pH 5, all D and E amino acids are approximated as having a charge of -1, while all H, K, and R amino acids have a charge of +1. For pH 7, the same charge states apply for D, E, K, and R amino acids, while H is neutral.

The same methodology employed in Chapter 3, section 3.2.4.1, was used here: MSOS simulations were performed at constant temperature (298.15 K) with 10^7 MC attempts for both the reference system and the interaction model of interest. Each MC attempt consisted of either a translation or a rotation around the center of mass of the molecule. The maximum displacement and rotation for a given step was obtained with a pre-equilibration step of 10^5 MC attempts to obtain an acceptance ratio of 50%. The steric-only behavior was used as the reference, so the simulation directly returned $B_{22}/B_{22,ST}$, and no subsequent rescaling was needed. The strength of short-ranged non-electrostatic attractions (*i.e.*, van der Waals and hydrophobic), ε_{SR}, was first evaluated to find a value that fits experimentally measured values for *TIS* > 300 mM. This parameter was held constant for the following simulations: $B_{22}/B_{22,ST}$ was calculated for ψ_i values between 0 and 2 in increments of 0.2, and *TIS* values of 10, 60, 110 and 300 mM (see equation 3.4 and Chapter 3 for more information). For simplicity, all ψ_i values

were treated equal, so the term ψ is used instead in the remainder of this chapter to represent an average charge correction for all side chains. Simulated $B_{22}/B_{22,ST}$ values were used to build surface response plots for further analysis as in Chapter 3.

4.2.6 Average Relative Deviation (ARD) Calculations and Model Validation.

To evaluate the effectiveness of colloidal models to fit or predict experimental weak interactions, the average relative deviation (ARD) was calculated for any given data set as shown in equation 3.10. As the value of ARD is a measure of the average deviation between the model and the experimental data, a cutoff value between 5% and 25% was used below as the criterion for considering a prediction to be quantitatively accurate, as this average deviation can be considered a conservative estimate of the model prediction uncertainty.

4.3 Interactions at Dilute Protein Concentrations

SLS was used to determine excess Rayleigh scattering (R^{ex}/K) as a function of c_2 for aCgn solutions at a range of solution conditions. Figure 4.1 shows experimental R^{ex}/K *vs* c_2 as a function of pH and NaCl concentration for $c_2 < 10$ g/L reproduced from reference [201]. The R^{ex}/K *vs* c_2 profiles differ qualitatively between pH 5 (panel A) and pH 7 (panel B). Increasing *TIS* by adding NaCl decreases the upward curvature in R^{ex}/K *vs* c_2 at pH 7, while it decreases downward curvature at pH 5. Upward (downward) curvature for R^{ex}/K *vs* c_2 in Figure 4.1 indicates net-attractive (net-repulsive) interactions. The high-*TIS* conditions at pH 5, and all the pH 7 conditions, show net-attraction. Conversely, pH 5 conditions show repulsive interactions at low *TIS*. This indicates that screened electrostatic interactions are present at both pH conditions.

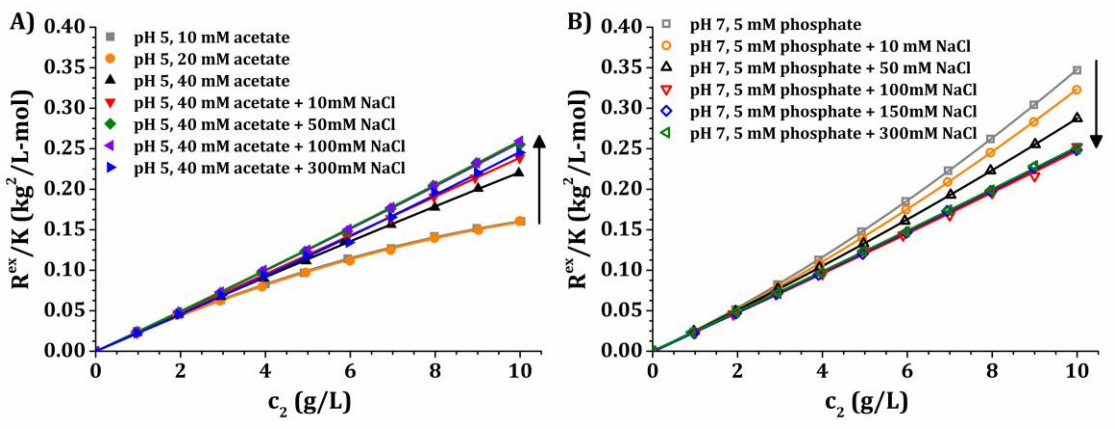

Figure 4.1. Excess Rayleigh scattering as a function of protein concentration for sodium acetate at pH 5 (panel A) and sodium phosphate at pH 7 (panel B), with increasing *TIS* as indicated by the arrows. Lines represent the fits to equation 1.9.

The data in Figure 4.1 were used to regress values of $M_{w,app}$ and G_{22} from equation 1.9 for aCgn at each of the solution conditions. $M_{w,app}$ was not found to be statistically different from M_w of aCgn (25.7 kDa) in all buffer conditions (Figure 4.2A). That is, the normalized apparent molecular weights ($M_{w,app}/M_w$) were approximately equal to 1, indicating that no measurable aggregation or solvent-solute non-idealities were present. The G_{22} values were used to calculate $B_{22}/B_{22,ST}$, using the relationship $B_{22} = -G_{22}/2$. A value of $B_{22,ST} = 4.9$ mL/g was used as per Grünberger *et al.* [63], and was treated as independent of solution conditions. On this scale, values of $B_{22}/B_{22,ST}$ larger (less) than 1 indicate net-repulsive (net-attractive) interactions beyond steric repulsion. As in the earlier chapters, the terms net-repulsive and net-attractive will be used relative to the steric-only value (*i.e.*, $B_{22}/B_{22,ST} = 1$) rather than the value for an ideal gas mixture (*i.e.*, $B_{22}/B_{22,ST} = 0$) [23,59,148].

The resulting values of $B_{22}/B_{22,ST}$ are plotted in Figure 4.2B as a function of *TIS*. The values in Figure 4.2B illustrate that the nature of protein-protein interactions is qualitatively different at pH 5 and 7. At low *TIS*, $B_{22}/B_{22,ST}$ decreases with increasing pH, consistent with decreasing the net protein charge as pH values approach the pI of the protein. The theoretical net-charge of aCgn is +5 and +9 at pH 7 and pH 5, respectively (as calculated using the amino acid sequence for aCgn [75] in PROPKA [203]). At pH 7, $B_{22}/B_{22,ST}$ was lower than 1 for all *TIS*, indicating net-attractive interactions. Additionally, the $B_{22}/B_{22,ST}$ values increase (become less negative) with increasing *TIS*. This indicates that attractive electrostatic interactions contribute strongly to the interactions, and this is likely due to dipole or higher multipole interactions that presumably overcome monopole-monopole repulsions [154]. Conversely, at pH 5 the interactions are strongly net-repulsive ($B_{22}/B_{22,ST} >> 1$) at low *TIS*, and this changes to slightly net-attractive interactions with increasing *TIS*. This trend is consistent with canonical behavior for screened monopole-monopole interactions between charged spheres that have short-ranged non-electrostatic attractions as in Chapter 3. Overall, the interactions at low c_2 and low *TIS* for aCgn are dominated by screened electrostatic interactions at pH 5 and 7, with primarily monopole-monopole repulsions at pH 5, and dipole or multipole contributions dominating closer to the isoelectric point (pI) at pH 7. At high *TIS*, both pH 5 and pH 7 conditions result in quantitatively similar values of B_{22} and show weakly net-attractive behavior ($B_{22}/B_{22,ST}$ near -1).

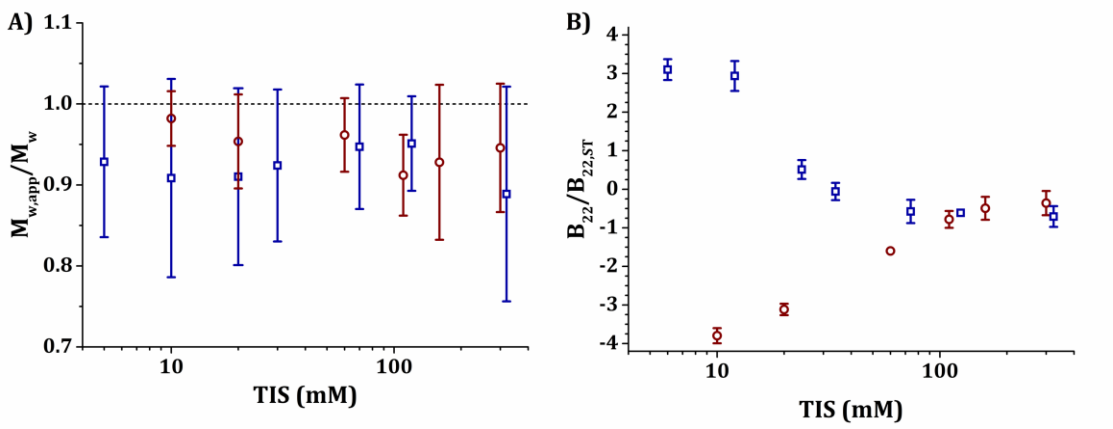

Figure 4.2. Normalized apparent protein molecular weights ($M_{w,app}/M_w$, panel A) and second osmotic virial coefficient ($B_{22}/B_{22,ST}$, panel B) at pH 7 (red circles) and pH 5 (blue rectangles) as a function of *TIS*. Error bars represent 95% confidence levels from fitting equation 1.9.

4.4 Modeling Weak Protein-Protein Interactions at Low c_2

The orientation-averaged interaction models derived by Bratko *et al.* (see Appendix A) were used to regress model parameters based on the measured B_{22} values as a function of *TIS* (Figure 4.2B) [76]. The data for pH 5 indicated predominantly screened monopole-monopole interactions at low *TIS*. Consequently, the $B_{22}/B_{22,ST}$ data at pH 5 were fit to an electrostatic model that only included repulsive electrostatic interactions (*i.e.*, $\mu_{eff} = 0$) along with short-ranged non-electrostatic attractions and steric repulsions. Based on the strongly attractive electrostatic interactions evident in Figure 4.2B at pH 7, $B_{22}/B_{22,ST}$ *vs TIS* for those conditions were fit to the screened electrostatic model that included monopole-monopole, monopole-dipole and dipole-dipole interactions. This was done as a first-order approximation to capture the experimental results at both pH values. As noted in the Methods section, the monopole model has two fitted parameters (ε_{sw} and Q_{eff}), and the monopole + dipole model has three fitted parameters (ε_{sw}, Q_{eff} and μ_{eff}).

Figure 4.3A shows the best-fit (solid line) for the PMF models to the pH 5 data, with confidence intervals from a 10% ARD shown as dashed lines. The inset shows the response surface of ARD from the model of $B_{22}/B_{22,ST}$ *vs TIS*, relative to the experimental results, as a function of ε_{sw} and Q_{eff}. Figure 4.3B shows the analogous results for pH 7 with confidence intervals from a 10% ARD as dashed lines, and with the inset showing the response surface of ARD as a function of Q_{eff} and μ_{eff}. For the pH 7 response surface, the number of parameters was reduced from three to two by using the $B_{22}/B_{22,ST}$ values at pH 5 and pH 7 at the highest *TIS* value to obtain a common value of ε_{sw} = 1.7 k_BT so as to capture the plateau value of $B_{22}/B_{22,ST}$ *vs TIS*.

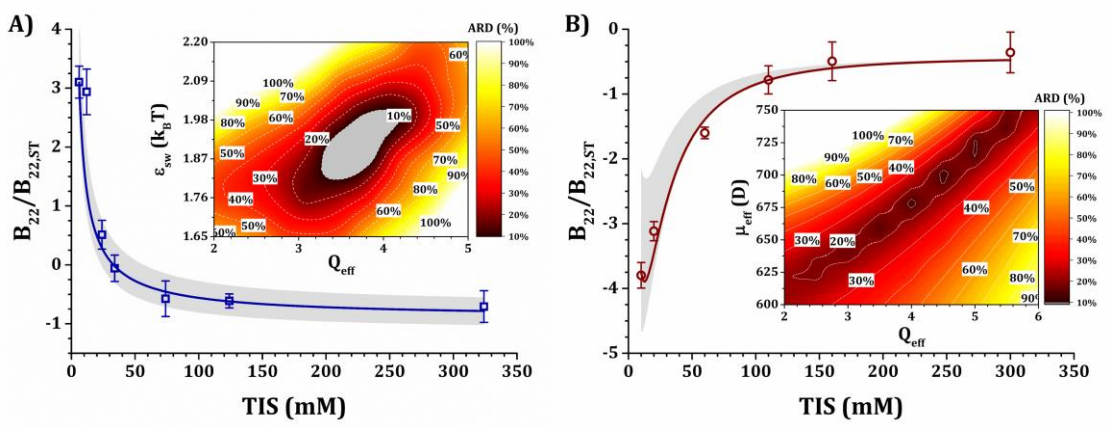

Figure 4.3. Parameter optimization of the PMF models (solid lines) with the experimental results from Figure 4.2B (symbols), as a function of *TIS*. **Panel A:** pH 5 with minimum ARD line shown at ε_{sw} = 1.87 k_BT, Q_{eff} = 3.4 and dashed lines indicating range of values obtained from 10% ARD calculations (gray area in inset). **Panel B:** pH 7 with ε_{sw} = 1.65 k_BT, Q_{eff} = 4.2, μ_{eff} = 693 D and dashed lines indicating 10% ARD range from inset. **Insets:** surface plots of ARD for computed and experimental B_{22} values across the range of *TIS*, as a function of ε_{sw} and Q_{eff} (pH 5, panel A) and Q_{eff} and μ_{eff} (pH 7, panel B).

The results in Figure 4.3 show that the simple spherical description of aCgn is able to qualitatively and quantitatively capture the experimental $B_{22}/B_{22,ST}$ *vs TIS* behavior. Additionally, the value for the dipole moment (μ_{eff}) of aCgn at pH 7 (693 D) calculated from these model fits is semi-quantitatively similar to values previously reported by Velev *et al.* (518 D) as well as calculations done using the approach described by Felder *et al.* (553 D) [41,204]. The next subsection addresses the question of whether this interaction model can predict high-c_2 behavior if it is used to extrapolate to high-c_2 conditions for aCgn, similar to Chapter 3.

4.5 TMMC and Simulated Excess Rayleigh Scattering at High c_2 and pH 5

TMMC was used to test whether the PMF models that were fit at low c_2 could be predictive of high-c_2 behavior, specifically R^{ex}/K *vs* c_2 beyond the dilute limit (c_2 > 10 g/L). Experimental R^{ex}/K values at high c_2 were measured in increments of 10 g/L up to c_2 = 100 g/L at *TIS* = 20, 30, 70 and 120 mM for pH 5; and 10, 20, 60 and 110 mM for pH 7. Using the parameter values derived from only the experimental results in the dilute limit (*i.e.*, using only $B_{22}/B_{22,ST}$ values as in section 3.4 and 3.5), TMMC simulations were performed at the same *TIS* and pH values listed above. To assess the sensitivity of the predictions from TMMC to the choice of model parameters, the simulations were performed across the small parameter space shown in the insets in Figure 4.3 within an ARD cut-off range of 20% around the global minimum for the best-fit parameters from the low-c_2 data. In what follows, the ARD values based on all of the R^{ex}/K data (*i.e.*, all c_2 and *TIS* values) for a given pH and choice of model parameter values are denoted as ARD_{All}. That is done to determine if a single set of model parameters can predict all the R^{ex}/K results from low to high c_2 and *TIS* for a given pH.

An alternative approach would be to optimize model parameters at a given *TIS* and pH, and that will also be explored below.

The ARD_{All} response surface for the results at pH 5 are shown in Figure 4.4A, with the experimental data spanning from low to high c_2 and *TIS*. Comparing the ARD response surface from low c_2 (Figure 4.3A, inset) with the ARD_{All} response surface for high c_2 (Figure 4.4A), the small subset of parameters obtained at low c_2 can reasonably predict the high-c_2 behavior at pH 5 over the whole *TIS*- range. A small subset of parameter values produces an overall ARD below 5%, showing that this CG description could quantitatively predict high-c_2 behavior based on a training set at low c_2 for net-repulsive conditions and weakly net-attractive conditions. Figure 4.4B shows the measured and predicted R^{ex}/K values by using parameters within the minimum ARD_{All} range from Figure 4.4A. There are small but systematic deviations between the predicted and experimental R^{ex}/K values with increasing c_2 for *TIS* = 30 mM conditions (red circles), but, otherwise, the model quantitatively captures the experimental data.

An alternative is to refine separate parameter sets for each *TIS*, in order to acknowledge that the effective charge in the CG model is a lumped parameter and could change with added NaCl based on preferential salt-protein interactions [49,154,155]. Figure 4.4C shows the analogue to Figure 4.4B, but for case where the model parameters for each *TIS* value are optimized separately, based on the ARD profiles for the low-c_2 data for that *TIS* value (see Appendix A for individual-*TIS* ARD calculations). Figure 4.4C shows that the simulated results quantitatively match the experimental results, with ARD values as low as 2%, for each of the *TIS* conditions. The agreement between model

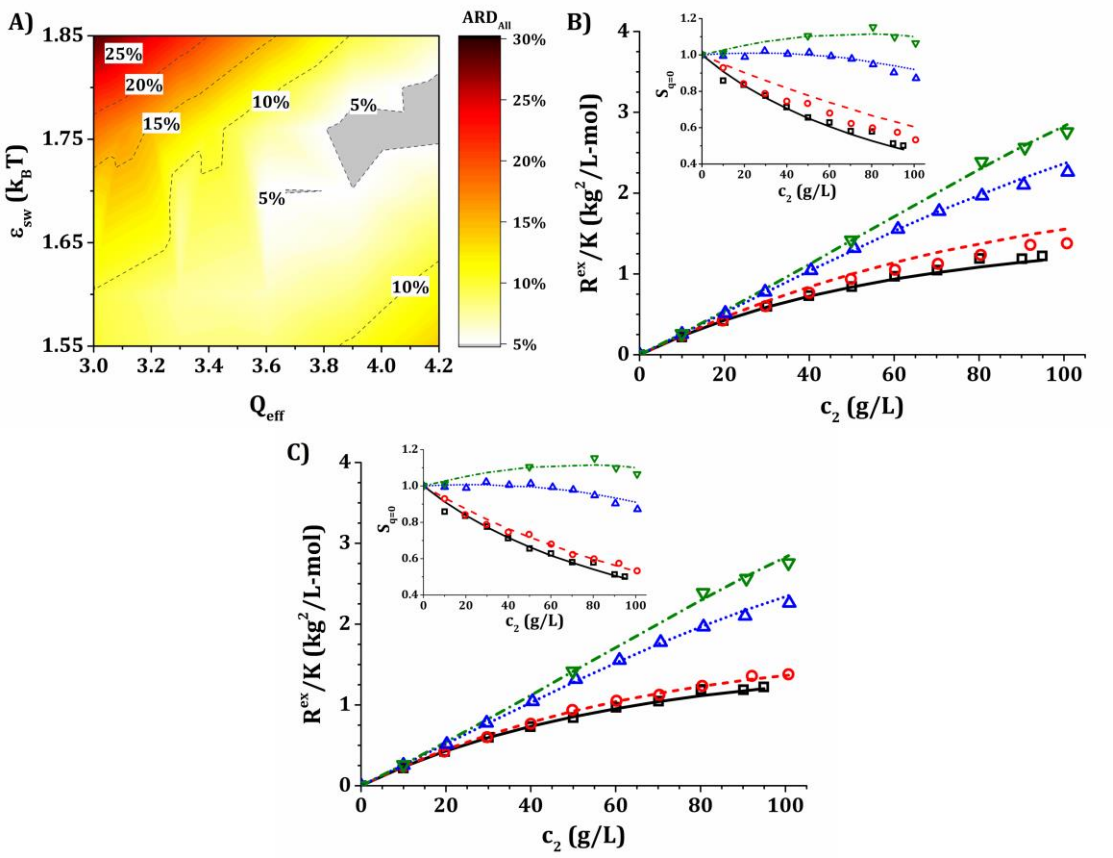

Figure 4.4. Comparison of experimental R^{ex}/K *vs* c_2 profiles at high c_2 with predictions from TMMC simulations at pH 5. **Panel A:** contour plot of ARD between experimental and predicted R^{ex}/K *vs* c_2 values. The gray area corresponds to ARD values below 5%. **Panel B:** overlay of experimental results (symbols) and predictions from simulation (lines) for the parameter values in panel A that minimized the overall ARD (ε_{sw} = 1.8 k_BT, Q_{eff} = 4.2*)*. **Panel C:** analogue to panel B but using the individual ARD in Appendix A. Insets for panels B and C are the corresponding $S_{q=0}$ *vs* c_2 transformation of the main panels. Colors represent *TIS* = 20 mM (black), 30 mM (red), 70 mM (blue) and 120 mM (green).

predictions and experimental results is slightly improved in Figure 4.4C when compared to Figure 4.4B, but at the cost of needing multiple sets of model parameters. Overall, the results in Figure 4.4 indicate that net-repulsive and weakly attractive interactions

from low to high c_2 can be captured accurately by a spherical CG model, in agreement with the results obtained in Chapter 3 for the case of two mAb proteins.

4.6 TMMC and Simulated Excess Rayleigh Scattering at High c_2 and pH 7

The same methodology was applied to pH 7 conditions, and the ARD_{All} response surface is shown in Figure 4.5A. In this case, there are two distinct regions for the predictions: a liquid-liquid (L-L) split regime at high μ_{eff}/Q_{eff} values, and a single-phase regime for low μ_{eff}/Q_{eff} values. The phase-separated regime is due to strong dipole-dipole and dipole-monopole attractions that overcome the monopole-monopole repulsion, as expected based on previous work that focused on the solution behavior of dipolar molecules [75,76]. Within the single-phase regime, a range of low ARD_{All} values was obtained, but no values below 20% were observed when all the *TIS* data sets at pH 7 were used. From inspection of individual-ARD plots (see Appendix A), the lowest-*TIS* condition (10 mM) resulted in considerably higher ARD values and skewed the parameter sets. Consequently, an additional ARD response surface was obtained by excluding the *TIS* = 10 mM (*i.e.*, buffer-only) condition, and that is shown in Figure 4.5B. The same two types of regimes (single phase and L-L split) are observed, but a portion of the ARD response surface clearly shows the desired low (< 10% ARD) behavior towards the lowest μ_{eff}/Q_{eff} values (bottom right corner). Figure 4.5C shows the comparison between predicted and experimental R^{ex}/K profiles for the parameter ranges in the low-ARD region of Figure 4.5B, and this excludes a prediction for the buffer-only condition. Figure 4.5C is analogous to Figure 4.5D, but using the parameter range based on individual *TIS* fittings as done above. The results in Figures 4.5C and 4.5D show that refining model parameters for individual *TIS* conditions is not required for *TIS* values greater than 10 mM. However, predictions from low-c_2 conditions are

not accurate at high-c_2 for strongly attractive conditions (*e.g.*, buffer-only), even if the parameters are regressed only for those low-*TIS* conditions.

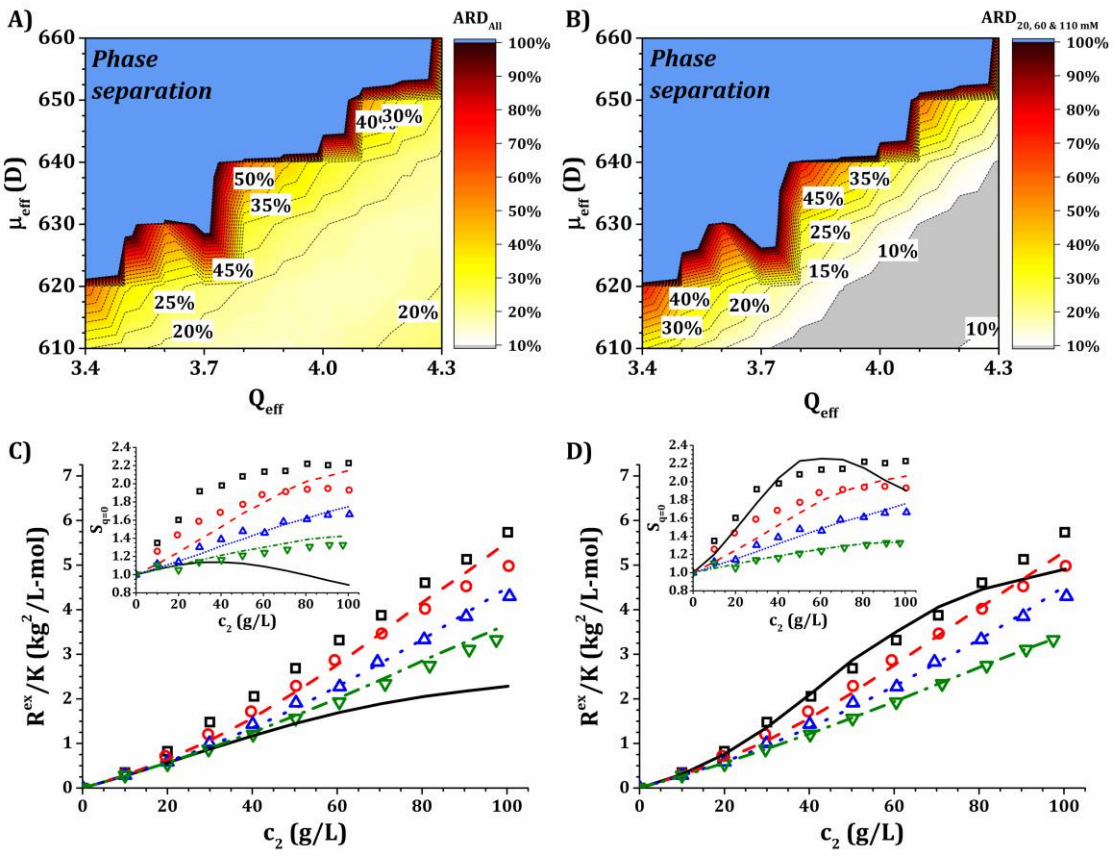

Figure 4.5. Comparison of experimental R^{ex}/K profiles at high c_2 and predictions from TMMC simulations at pH 7. **Panel A:** contour plot of ARD between experimental and predicted R^{ex}/K *vs* c_2 values. **Panel B:** contour plot of ARD between experimental and predicted R^{ex}/K *vs* c_2, excluding the lowest *TIS* (10 mM). **Panel C:** overlay of experimental results (symbols) and predictions from simulation (lines) for the parameter values in panel A that minimized the overall ARD (ε_{sw} = 1.7 k_BT, Q_{eff} = 4.2, μ_{eff} = 630 D). **Panel D:** Analogue to panel C but using the individual ARD plots in Appendix A. Insets for panels B and C are the corresponding $S_{q=0}$ *vs* c_2 transformation of the main panels. Colors represent *TIS* = 10 mM (black), 20 mM (red), 60 mM (blue) and 110 mM (green).

Interestingly all parameter sets that provided good agreement between experimental and simulated R^{ex}/K profiles in Figures 4.4 and 4.5 correspond to ARD regions below 20% for the $B_{22}/B_{22,ST}$ analysis (Figure 4.3). Therefore, with the exception of the lowest-*TIS* conditions at pH 7, the parameters obtained from just low-c_2 analysis (*i.e.*, $B_{22}/B_{22,ST}$ *vs TIS*) provided quantitatively or semi-quantitatively accurate predictions of the high-c_2 behavior. However, for very low *TIS* and strongly attractive electrostatic interactions at pH 7, there is poor agreement between the predictions and the experimental R^{ex}/K data. This failure suggests that simplified CG interaction models lack some of the key physics to properly describe attractive electrostatic interactions at higher c_2, and presumably that includes either accounting for higher terms in a multipole expansion, or working with a structurally more detailed CG models [66,76,82,154].

There is an alternative interpretation of the results in Figure 4.5: predictions for electrostatic attractions at high c_2 are sensitive to the value of *TIS* at low-*TIS* conditions. This would follow because the ion screening-length affects monopole-monopole repulsion, monopole-dipole and dipole-dipole attractions differently – each type of interaction decays as r^{-1}, r^{-2} and r^{-3}, respectively, with r denoting the intermolecular separation (see Appendix A). For low-*TIS* conditions, the balance between repulsion and attraction can be delicate, and strongly dependent on c_2. Therefore, slight changes in *TIS* can tilt the balance towards a monopole-dominated behavior or a dipole-dominated behavior as c_2 increases. Appendix A illustrates this qualitatively for the interaction models in this work, and shows that strongly attractive interactions at high-c_2 and low-*TIS* conditions (below ~20 mM) may be expected to be difficult to predict quantitatively using simple spherical CG models. These concerns notwithstanding, the

results in Figure 4.5 suggest that useful predictions about qualitative effects may be possible (*e.g.*, very strong multi-body attractions) using minimalistic CG models.

The limitations at strongly attractive conditions are further illustrated in Figure 4.6, which shows the maximum *TIS* value at which a given set of μ_{eff} and Q_{eff} values would result in L-L phase separation at pH 7 based on TMMC simulations, with the remaining model parameters fixed as described before. Figure 4.6 shows that high values of μ_{eff}/Q_{eff} can result in phase separation with *TIS* as low as 50 mM. Increased net-charge (Q_{eff}) requires a higher minimum dipole moment (μ_{eff}) to observe L-L separation, illustrating the sensitive balance between repulsive and attractive electrostatic interactions.

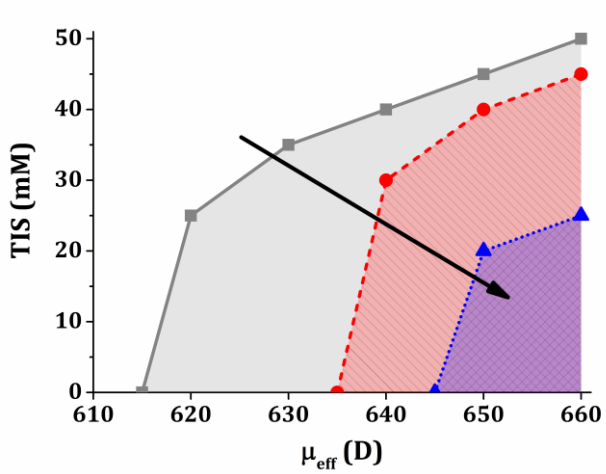

Figure 4.6. Predicted phase separation at pH 7 as a function of dipole moment and TIS with increasing net-charge values indicated by the arrow *(Q_{eff} = 3.4, 3.8 and 4.2)*. The colored area represents the conditions under which liquid-liquid phase separation was observed in the simulations (low *TIS* and higher μ_{eff} values with increasing Q_{eff}).

As mentioned above, predictions might be improved at high-c_2 conditions if one included higher-order terms in the multipole expansion. However, that would require more adjustable parameters and the accuracy of the model is still likely to break down once c_2 reaches large enough values to be relevant to dense liquid phases (>> 100 g/L) because the detailed and complex surface-charge distribution would be expected to play a stronger role in producing localized multipole(s) on the protein surface [82,141]. These results for the spherical model resemble those observed experimentally for the IgG4 in Chapter 3, and suggest the need to use higher molecularly detailed CG models to capture conditions where the charge distribution can lead to strong attractive electrostatic interactions. This is explored for aCgn in what follows.

4.7 Higher Molecular Resolution CG Simulations

As done in section 3.8, protein-protein interaction response surfaces were calculated by computing $B_{22}/B_{22,ST}$ as a function of *TIS* and the ratio between solution and theoretical charges, ψ. For this, a value for ε_{SR} was first needed. Figure 4.7 shows the response of $B_{22}/B_{22,ST}$ by perturbing ε_{SR} and is qualitatively similar to the IgG1 and IgG4 results shown in Figure 3.21 (*i.e.*, $B_{22}/B_{22,ST}$ decays rapidly towards negative values as ε_{SR} increases) [63,82]. Based on the experimental $B_{22}/B_{22,ST}$ results at *TIS* > 300 mM (Figure 4.2B), ε_{SR} = 0.36 k_BT was used to reconstruct response surfaces to show one of three limiting cases, depending on the degree of anisotropy of $B_{22}/B_{22,ST}$ *vs* [*TIS*, ψ] . These simulations are expected to show (1) monopole dominated behavior, (2) multipole dominated behavior or (3) $B_{22}/B_{22,ST}$ shows a transition between monopole- and multipole-dominated regions as ψ increases (see section 3.8 for more information). This behavior can also be observed as a function of *TIS*. However, experiments are usually performed for a set *TIS* value, so the discussion below is

developed from the perspective of ψ as a degree of freedom as described above and in section 3.8.

Figure 4.8 shows the response surfaces for pH 5 (panel A) and pH 7 (panel B) using the 1bAA model and the published crystal structure of aCgn (PDB: 1EX3). The two figures show that aCgn exhibits a transition from monopole-dominated to multipole-dominated behavior as a function of pH (case 3, above). However, the change in pH shifts the ψ value at which this transition is located. For pH 5, the transition is observed at a very high value of the effective electrostatic interactions ($\psi > 2$), while it occurs at a lower value ($\psi > 1.8$) for pH 7. These results are consistent with the existence of strong multipole interactions for pH 7 for aCgn, but not for pH 5. Figure 4.9 shows the comparison of experimental $B_{22}/B_{22,ST}$ *vs TIS* with the values predicted by the 1bAA model at each pH if one selects the value of ψ to minimize ARD ($\psi \sim 0.67$ for pH 5, $\psi \sim 1.5$ for pH 7). The value of ψ is related to the average ratio of the effective solution charge to the theoretical value of each charged residue due to territorial ion binding. Physically, a ψ value different from 1 can be expected if there is non-ideal ion accumulation that causes greater than ideal charge screening ($\psi < 1$, as in pH 5) or multivalent ion (de)clouding and binding which can cause an increase in the effective charge ($\psi > 1$, as in pH 7) [154,158]. An alternative reason is simply that the 1bAA model groups entire amino acids into single beads, and the bead diameter is different than that of the atom center that corresponds to the real charge site for a given amino acid. Therefore, caution should be used to not over-interpret the physical meaning of $\psi > 1$ with CG molecular models.

A minimum for $B_{22}/B_{22,ST}$ at low to intermediate *TIS* values for pH 7 is characteristic of strong electrostatic attractions caused by multipole interactions which,

in theory, can be overcome by very long-ranged monopole repulsions at lower *TIS*. However, the experimental and simulated data show qualitative deviations at the lowest *TIS* values. This is possibly due to the inherent limitations of treating ion-screening with a mean-field description at such low *TIS* [154,158]. The results in Figures 4.8 and 4.9 suggest that this level of CG structural model can accurately capture the anisotropic charge distribution of proteins in solution at low c_2, but the model accuracy is lost at low *TIS* (below on the order of 50 mM) and the parameter set may depend on pH in agreement with the results obtained in section 3.8. These results highlight the importance of experimental "training sets" to make even the higher-structural-resolution CG models more effective, and the potential to use the 1bAA model for smaller proteins outside mAbs, as well as the limitations for strongly attractive electrostatic interactions. Further refinements of the model will be required to better capture the strong electrostatic attractions observed at pH 7 for aCgn and pH 6.5 for the IgG4 (Chapter 3), and this will be further discussed in Chapter 7.

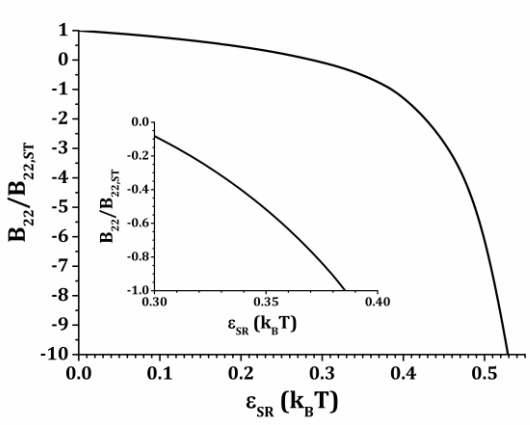

Figure 4.7. $B_{22}/B_{22,ST}$ behavior for the 1bAA CG molecular model as a function of ε_{SR}.

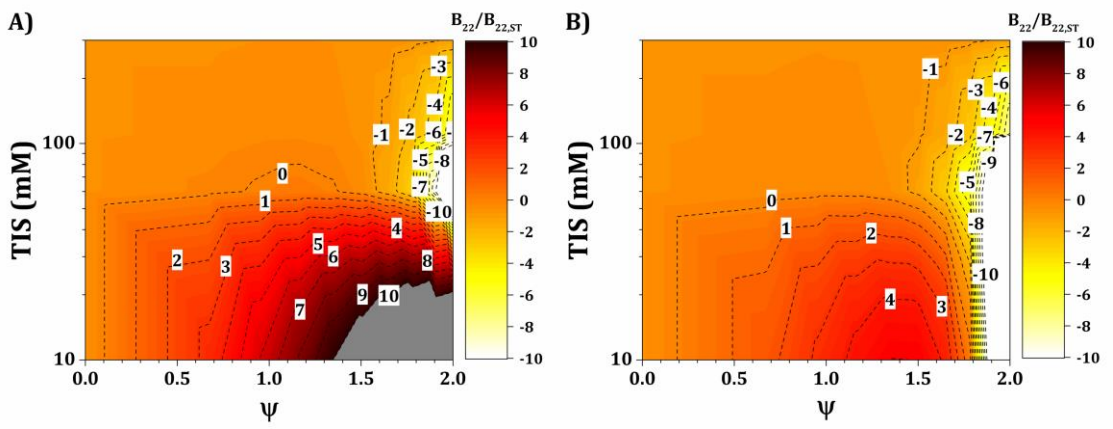

Figure 4.8. $B_{22}/B_{22,ST}$ response surfaces for the 1bAA CG molecular model as a function of *TIS* and ψ for pH 5 (panel A) and pH 7 (panel B) at ε_{SR} = 0.36 k_BT. The gray areas correspond to $B_{22}/B_{22,ST}$ > 10, while white areas correspond to $B_{22}/B_{22,ST}$ < -10.

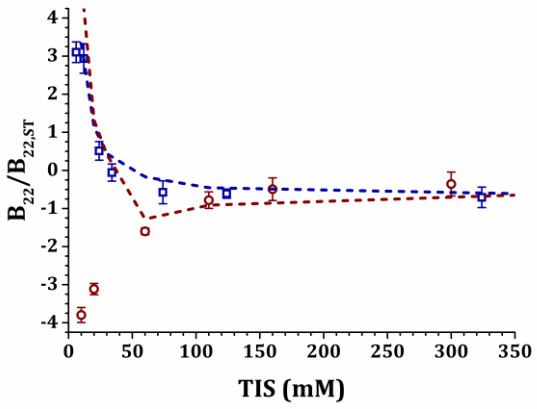

Figure 4.9. Experimental $B_{22}/B_{22,ST}$ *vs TIS* with best fit parameter sets from the 1bAA CG model. Blue squares (pH 5, ψ = 0.67) and red circles (pH 7, ψ = 1.5) represent the experimental data while dashed lines represent the simulated values. The simulated values overlap with the experimental data except at the low-*TIS* conditions for pH 7.

4.8 Summary and Conclusions

Static light scattering (SLS) was used to quantify net protein-protein interactions of aCgn at a range of pH, *TIS*, and c_2 that produce net-repulsive or net-attractive electrostatic interactions at low *TIS*. A spherical CG model was tested for its ability to capture the net protein interactions from low to high c_2 for both net-repulsive and net-attractive behavior of a globular-protein solution. The results show that colloidal models can quantitatively capture the data if a combination of screened monopole, screened dipole model, and short-ranged non-electrostatic attractions (*via* a square-well interaction model) is used. Additionally, the results presented in the chapter show that the low-c_2 model parameters were quantitatively predictive of the interactions at high c_2 if the net protein interactions are repulsive or slightly attractive compared to steric-only interactions. However, for strongly attractive conditions, where the effect of charge anisotropy is dominant, the low-c_2 experimental data coupled with spherical CG models were only able to qualitatively or semi-quantitatively predict the high-c_2 behavior. Independent of the spherical models, the approach used in Chapter 3 for identifying anisotropic charge distributions causing attractions was validated in this chapter, but challenges still exist for strongly attractive electrostatic conditions at $TIS < 50$ mM. Finally, the results discussed in this chapter are in good agreement with the results presented in Chapter 3.

Chapter 5

PREFERENTIAL INTERACTIONS FOR MULTI-COMPONENT SOLUTIONS USING INVERSE KIRKWOOD-BUFF SOLUTION THEORY

5.1 Introduction

As mentioned in section 1.4, with the exception of idealized conditions, proteins rarely exist in solutions where water is the only other component [8,10,16]. A minimum of three components are usually present in aqueous protein solutions: water, protein, and buffer molecules or other ions [8,10]. The addition of buffer allows one to better control the pH of the solution. However, this is usually not enough to stabilize protein solution, so the addition of co-solutes and/or co-solvents can significantly stabilize the overall protein solution, such as controlling protein phase behavior, reducing aggregation, enhancing conformational stability, and lowering solution viscosity [23,25,44,49,189,205–209]. This has fundamental and practical implications for design, manufacture, and formulation of proteins and other biomolecular solutions. Unfortunately, current frameworks that focus on quantifying protein-cosolute preferential interactions are only applicable for ternary solutions, so there is a basic need for more comprehensive theories that would allow one to study solutions with an unlimited number of components.

This chapter will focus on the development of a generalized expression for the protein partial specific volume, $\hat{V}_2$, in terms of KB integrals for any number of solution components, starting from the framework developed by Ben-Naim [44,49,59]. $\hat{V}_2$ is accessible experimentally *via* high-precision density measurements. This expression is

used to compare the experimental results for solutions of α-chymotrypsinogen A (aCgn), for both ternary (water-protein-osmolyte or water-protein-buffer) solutions and quaternary (water-protein-osmolyte-buffer) solutions. This is done to demonstrate the use of the new generalized expression to ternary solutions (in agreement with previous results [49]) as well as quaternary or multi-component solutions. Additionally, $\hat{V}_2$ results for the IgG1 molecule discussed in Chapter 3 are used to gain insights about the molecular origins of the increased protein-protein repulsion in the presence of sucrose at both pH 5 and 6.5.

The remainder of this chapter is organized as follows. The materials and methods section includes experimental and computational methods, as well as a short summary of inverse KB solution theory and its relation to $\hat{V}_2$ in order to set a context for the following derivations. A derivation is presented for a generalized expression that relates $\hat{V}_2$ to only KB integrals and measurable quantities such as partial specific volumes ($\hat{V}_i$) and concentrations (c_i) for the case of the protein component at infinite dilution (*i.e.*, $c_2 \to 0$) and an arbitrary number and concentration of cosolutes. Additional mathematical details are included in Appendix B. The resulting general expression is then used to analyze experimental density data for ternary and quaternary aCgn solutions in terms of KB integrals and preferential interactions, including some conditions from Chapter 4. Additional measurements of quaternary solutions of one of the antibody molecules from Chapter 3 are presented towards the end of this chapter to further explain experimental results shown in Chapter 3. The results are also discussed within the context of different physical contributions to each KB integral, using infinite-dilution, steric-only contributions from atomistic models as a reference state. Some of the content in this chapter has been published or included in peer reviewed journals [60].

5.2 Materials and Methods

5.2.1 Summary of Inverse KB Solution Theory and Formal Relationships Between $\widehat{V}_i$ and KB Integrals

This section begins with a short summary of the results from Ben-Naim and others [59,97,210]. Equation 1.4 shows the definition of an arbitrary KB integral from component *i* and *j* given by Kirkwood and Buff [95,210]. Unless the radial distribution function is known (*e.g.*, from molecular simulations and an assumed intermolecular potential function for simple geometries), it would be more practical to determine G_{ij} values from experimental data. The relevant G_{ij} values are those in a grand-canonical ensemble, where molecules are allowed to diffuse in and out of the system volume as discussed in previous chapters [59,95]. In this case, G_{ij} can be defined as a function of the fluctuations in the number of molecules of each component (N_i) in the system:

$$G_{ij} = V\left(\frac{\langle N_i N_j\rangle - \langle N_i\rangle\langle N_j\rangle}{\langle N_i\rangle\langle N_j\rangle} - \frac{\delta_{ij}}{\langle N_i\rangle}\right) \tag{5.1}$$

where $\langle N_i\rangle$ represents the ensemble average for N_i, and $\langle N_i\, N_j\rangle$ is the ensemble-average covariance of components *i* and *j* (*i.e.*, how much fluctuations for N_i and N_j are correlated). *V* is the system volume and δ_{ij} is the Kronecker delta (*i.e.*, 1 for $i = j$ and 0 for $i \neq j$). This is the generalized expression from which equation 1.3 is obtained from [59]. In what follows, the number concentration (molecules per unit volume) of component *i* will be expressed as $\rho_i = \langle N_i\rangle/V$.

Similarly, an expression for the change in $\langle N_i\rangle$ as a function of the chemical potential of any other component (μ_j) can be obtained from standard statistical thermodynamic fluctuation theory [59]:

$$k_B T\left(\frac{\partial\langle N_i\rangle}{\partial\mu_j}\right)_{T,V,\mu_{k\neq j}} = \langle N_i N_j\rangle - \langle N_i\rangle\langle N_j\rangle, \tag{5.2}$$

It should be noted that the partial derivative is at constant chemical potential values of all components except for *j*. Also, the subscripts *i* and *j* can be interchanged for each side in equation 5.2. Combining equations 5.1 and 5.2, and rearranging, results in:

$$b_{(n)}^{ij} = \frac{k_B T}{V}\left(\frac{\partial\langle N_i\rangle}{\partial\mu_j}\right)_{T,V,\mu_{k\neq j}} = k_B T\left(\frac{\partial\rho_i}{\partial\mu_j}\right)_{T,V,\mu_{k\neq j}} = \rho_i\rho_j G_{ij} + \rho_i\delta_{ij} \tag{5.3}$$

where $b_{(n)}^{ij}$ is the (*i*,*j*) component of the *n*-dimensional $B_{(n)}$ matrix, with *n* denoting the number of components of the solution. This matrix represents how changes in the chemical potential of a given component induce changes in the average concentrations of all other components in the solution in a grand-canonical ensemble, and this arises from the interactions between all of the components (in a multi-body fashion). The matrix $B_{(n)}$ must be symmetric, as interchanging indices in equations 5.1 to 5.3 results in equivalent expressions. Therefore, $G_{ij} = G_{ji}$ and $b_{(n)}^{ij} = b_{(n)}^{ji}$, as is assumed for symmetric fluids [59]. Ben-Naim further derived a general definition for the partial molar volume ($\overline{V}_\alpha$) of any component α in an *n*-component system by using the cofactors of $B_{(n)}$ (equations 5.4 to 5.6 and in Appendix B):

$$\overline{V}_\alpha = \frac{\beta_\alpha}{\eta}, \tag{5.4}$$

$$\beta_\alpha = \sum_i^n \rho_i c_{(n)}^{i\alpha}, \tag{5.5}$$

$$\eta = \sum_i^n\sum_j^n \rho_i\rho_j c_{(n)}^{ij}, \tag{5.6}$$

Here, a slightly different nomenclature is introduced to aid in the subsequent derivations. In equations 5.4 to 5.6, $c_{(n)}^{ij}$ is the *i*-*j*th cofactor of the *n*-dimensional $B_{(n)}$ matrix and is also the (*i*,*j*) component of the corresponding cofactor matrix denoted as

$C_{(n)}$. Equation 5.7 is an additional expression that will be used below as part of the derivation of a general expression for $\overline{V}_\alpha$ in terms of the set of G_{ij} rather than the cofactors of $B_{(n)}$. $|B_{(n)}|$ represents the determinant of the matrix $B_{(n)}$.

$$k_B T \kappa_T = \frac{\left|B_{(n)}\right|}{\eta} \tag{5.7}$$

5.2.2 General Expression for $\overline{V}_\alpha$ in Terms of KB Integrals

Throughout this section, $B_{(n\text{-}1|k)}$ will represent a (n-1)-dimensional matrix that is derived from the previous $B_{(n)}$ matrix by deleting component k (deleting the k-th row and column) and further rearranging the matrix as is done in the calculation of cofactors (see Appendix B). Similarly, $c^{ij}_{(n-1|k)}$ represents the i-jth cofactor of the new $B_{(n\text{-}1|k)}$ matrix and the i-jth component of the cofactor matrix $C_{(n\text{-}1|k)}$. Equation 5.8 is obtained by solving equation 5.4 as shown in Appendix B:

$$\overline{V}_\alpha = \frac{\rho_\alpha \left[\left|B_{(n-1|\alpha)}\right| - \sum_{j\neq\alpha}^{n-1}\left(\rho_j G_{j\alpha} \sum_{i\neq\alpha}^{n-1} \rho_i c^{ij}_{(n-1|\alpha)}\right)\right]}{\rho_k \left[\sum_{i\neq k}^{n-1}\sum_{j\neq k}^{n-1} \rho_i\rho_j c^{ij}_{(n-1|k)} + \rho_k\left(G_{kk}\sum_{i\neq k}^{n-1}\sum_{j\neq k}^{n-1}\rho_i\rho_j c^{ij}_{(n-1|k)} + f_{cross}\right)\right]} \tag{5.8}$$

To proceed, the equality $k = \alpha$ will be used as there are not restrictions on the values that k and α can take (see Appendix B). Doing so allows one to eliminate both prefactors (ρ_α and ρ_k) from equation 5.8 as they will cancel each other. In addition, the concentration of component α will be assumed to be sufficiently low that α can be treated as being infinitely dilute (*i.e.*, $\rho_\alpha \rightarrow 0$). Under this assumption, the second term in the denominator of equation 5.8 will be negligible. Physically, this means that α-α interactions do not contribute significantly to $\overline{V}_\alpha$, so the dominant contributions to η (equation 5.6) come from the remaining (n-1) components. This assumption causes equation 5.8 to simplify to equation 5.9.

$$\overline{V}_\alpha = \left|B_{(n-1|\alpha)}\right| \Big/ \left(\sum_{i\neq\alpha}^{n-1} \sum_{j\neq\alpha}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|\alpha)} \right) - \left(\sum_{j\neq\alpha}^{n-1} \rho_j G_{j\alpha} \sum_{i\neq\alpha}^{n-1} \rho_i c^{ij}_{(n-1|\alpha)} \right) \Big/ \left(\sum_{i\neq\alpha}^{n-1} \sum_{j\neq\alpha}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|\alpha)} \right) \quad (5.9)$$

Equation 5.9 shows two different contributions to $\overline{V}_\alpha$: the first ratio on the right-hand side is that for a (*n*-1)-component solution (*i.e.*, by completely deleting component α from the solution) while the second ratio is effectively a mathematical expansion from that initial solution by adding the individual contributions arising from all pairs of α-*i* interactions, for any $i \neq \alpha$. This is more easily seen when equation 5.9 is combined with equations 5.4 and 5.7. The first term on the right-hand side of equation 5.9 is equivalent to the isothermal compressibility of the (*n*-1)-component mixture, while the second term can be rearranged and written in terms of only KB integrals, partial molar volumes, and component molar densities, resulting in equation 5.10:

$$\overline{V}_\alpha = k_B T \kappa_T - \sum_{j\neq\alpha}^{n-1} \rho_j G_{j\alpha} \overline{V}_j, \quad (5.10)$$

This equation applies under infinitely dilute conditions for component α. It does not impose any restriction on the concentrations of any of the other (*n*-1) components. Therefore, equation 5.10 can be used for protein solutions with any concentration of added cosolutes if the concentration of protein can be considered sufficiently dilute to neglect contributions to $\overline{V}_2$ from protein-protein interactions. Empirically, this typically corresponds to c_2 on the order of a few g/L or less (see Chapters 3 and 4, and results below) [23,29,49].

To obtain an expression in term of preferential interactions, the identity $\sum_j \rho_j \overline{V}_j = 1 \;\rightarrow\; \rho_k \overline{V}_k = 1 - \sum_{j \neq k} \rho_j \overline{V}_j$ can be combined with equation 5.10 to give equation 5.11, where component *k* is taken as the solvent (*e.g.*, water). If the system of interest is an aqueous solution with protein at low c_2, then $k = 1$ and $\alpha = 2$, and one obtains equation 5.12. Here, the identity $\bar{V}_i = \hat{V}_i M_{w,i}$ is used, where the overbar denotes a partial molar volume (units of volume/mole) and the caret denotes a partial specific volume (units of volume/mass). Similarly, $c_i = \rho_i M_{w,i}$, so $\rho_j \overline{V}_j = c_j \hat{V}_j$. The G_{ij} values in equation 5.11 have volume/mole units, but volume/mass units in equation 5.12.

$$\overline{V}_\alpha = k_B T \kappa_T - G_{k\alpha} + \sum_{\substack{j \neq \alpha \\ j \neq k}}^{n-2} \rho_j \overline{V}_j \left(G_{k\alpha} - G_{j\alpha} \right) \tag{5.11}$$

$$\hat{V}_2 = \frac{RT\kappa_T}{M_{w,2}} - G_{12} + \sum_{j \neq 1,2}^{n-2} c_j \hat{V}_j \left(G_{12} - G_{2j} \right) \tag{5.12}$$

The only assumption in the derivation above was the condition of infinite dilution for one of the components (α in equation 5.11 or *2* in equation 5.12). Therefore, equations 5.11 and 5.12 can be applied to solutions containing an arbitrary number of components, and over any physically realizable set of concentrations for components other than α or 2, respectively. For a ternary solution, equation 5.12 reduces to equation 1.8, and for a binary solution (*e.g.*, protein in pure water) it reduces to the first two terms on the right-hand side of equations 5.11 and 5.12, in agreement with the exact derivations for two- and three-component systems [49,59].

5.2.3 Experimental Determination of $\hat{V}_2$ Values

Experimental data for $\hat{V}_2$ values *vs* cosolute concentrations (*e.g.*, c_3, c_4, etc.) were obtained to illustrate the use of equation 5.12 for assessing protein-water and protein-

cosolute interactions. aCgn at pH 7 and the IgG1 molecule used in Chapter 3 at pH 5 and 6.5 are studied in this chapter as protein model systems. There were two main sets of solution conditions for aCgn. The first set consisted of ternary solutions formed by water, aCgn and a given cosolute chosen from: sucrose, trehalose, sodium phosphate and sodium chloride. The second set consisted of quaternary solutions formed by adding either sucrose, trehalose, or NaCl to ternary solutions of aCgn in 5 mM sodium phosphate aqueous buffer. For the IgG1 molecule, there were two sets of quaternary solutions containing water, protein, sucrose (as varying osmolyte) and a buffer molecule (10 mM acetate for pH 5 and 10 mM histidine for pH 6.5).

The same buffer stock solution presented in Chapters 3 and 4 (sodium phosphate buffer for aCgn, and acetate and histidine buffers for IgG1) were used for the experimental results in this chapter, which are summarized in what follows. Stock solutions with a range of buffer concentrations were prepared by dissolving sodium phosphate monobasic anhydrous (Fisher Scientific), glacial acetic acid (Fisher Scientific) or histidine hydrochloride (Sigma) in deionized water (MilliQ, Millipore-Sigma), and subsequently titrated to the respective pH (7.0 ± 0.05 for phosphate buffers, 6.5 ± 0.05 for histidine buffers and 5.1 ± 0.05 for acetate buffers) using a 5 M sodium hydroxide solution (Fisher Scientific). All buffer solutions were filtered with 0.22 μm filters (Millipore) and stored at 4 °C prior to use.

For a given set of solutions containing aCgn, the following procedure was used: aCgn powder (Worthington Biochemical Corp.) was dissolved into a 5 mM phosphate buffer solution at pH 7.0 ± 0.05 to an approximate protein concentration (c_2) of 15 g/L. A phenylmethlylsulfonyl fluoride (PMSF) solution was prepared by dissolving PMSF (Fluka) in 100% anhydrous ethanol (Sigma-Aldrich). To deactivate residual serine

proteases that are anecdotally present in commercial sources of aCgn, 1 mL of a 35 g/L PMSF solution (in 100 µL aliquots) was used for each gram of aCgn in solution [45]. The deactivated protease(s) precipitated readily and was later removed by centrifugation. The remaining aCgn solution was triple dialyzed against the desired buffer solution (see below) using 10 kDa molecular weight cutoff (MWCO) Spectr/Por dialysis membrane (Spectrum Laboratories, Rancho Dominguez, CA) at 4 °C to remove any residual salt impurities from the commercial material, as well as residual ethanol from the PMSF treatment.

IgG1 protein solutions were prepared as follows. A stock IgG1 solution was provided by Bristol-Myers Squibb at a starting protein concentration of ~50 g/L. pH 5 and 6.5 protein stock solutions were filtered and dialyzed using 10 kDa molecular weight cutoff (MWCO) Spectra/Por dialysis membrane (Spectrum Laboratories, Rancho Dominguez, CA) in the desired buffer (10 mM acetate for pH 5 and 10 mM histidine for pH 6.5) with four 12-hr buffer exchanges at 4°C to remove any undesired solutes from the original protein solution.

For ternary solutions with sodium phosphate as the osmolyte (aCgn only), the protein solution was dialyzed against selected sodium phosphate concentrations (5, 20 and 30 mM, as needed) at pH 7.0 ± 0.05. For all other ternary solutions, the protein solution was dialyzed against deionized water and then titrated to pH 7.0 ± 0.05 using a 50 mM sodium hydroxide solution (in 10 µL aliquots). For all aCgn quaternary solutions, the protein solution was dialyzed against a 5 mM phosphate buffer solution at pH 7.0 ± 0.05. The resulting protein stock solutions were filtered (Millipore, 0.22 µm) to eliminate any residual insoluble PMSF as well as any other contaminant.

Osmolyte stock solutions were prepared by dissolving sucrose (HPLC grade, Sigma), D-(+)-trehalose (Fisher Scientific), NaCl (Fisher Scientific) or sodium phosphate monobasic anhydrous (Fisher Scientific) in deionized water (for ternary solutions) or buffer solutions (for quaternary solutions) to obtain final solutions of 30% w/w sucrose, 30% w/w D-(+)-trehalose (hereafter referred to simply as trehalose) or 1 M NaCl. These solutions were titrated to their respective pH with small volumes of a 1 M sodium hydroxide solution. Final protein solutions were prepared gravimetrically by combining (1) protein-water or protein-buffer stock solution, (2) pH-adjusted water or buffer, (3) cosolute-water or cosolute-buffer stock solution. The proportions of (1), (2), and (3) were selected to achieve a constant cosolute molality for a series of increasing protein concentrations, up to a maximum c_2 of 1.5 g/L to ensure infinitely dilute protein behavior (see Chapters 3 and 4). Final osmolyte and protein concentrations were later calculated and corrected with measured density values (see below). Less than 0.1% variation between targeted and actual values for the protein and cosolute concentrations was achieved in all cases.

The density of each protein solution was measured using a DMA 4500 density meter (Anton-Paar, Ashland, VA) and a DDM 2911 Plus density meter (Rudolph Scientific, Hackettstown, NJ). Both instruments were used for comparison, and no quantitative differences were observed if consistent calibrating solutions and conditions were used. All measurements were done at 25.00 ± 0.02 °C and ambient pressure. $\hat{V}_2$ values were determined from density measurements as a function of protein weight fraction using equation 1.7, as previously described and illustrated in Figure 5.1 [49]. Linear regression was used to obtain the intercept and the slope as needed in equation

1.7. A 95% confidence interval for $\hat{V}_2$ was obtained from the corresponding t-value and standard error of the slope and the intercept with error propagation [211].

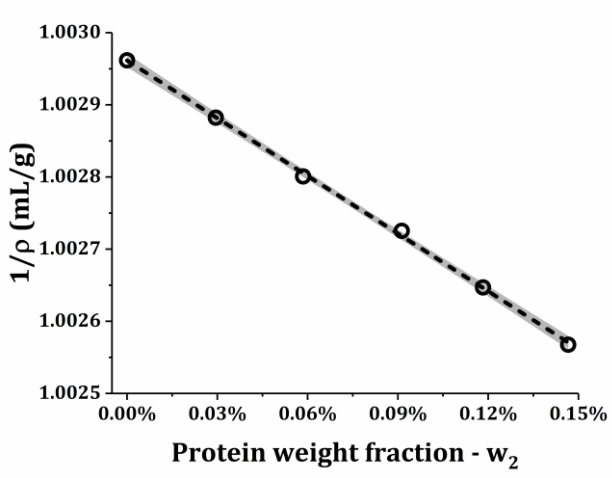

Figure 5.1. Inverse density as a function of protein weight fraction for water-aCgn solutions at 25 °C and pH 7. $\hat{V}_2$ in pure water is determined from the slope and intercept as shown in equation 1.7. The experimental error bars are smaller than the symbols. The dashed line represents the linear fit while the narrow surrounding gray area represents the 95% confidence level of the linear fit.

All data reported in the present work are at atmospheric pressure, 25 °C, and pH 5, 6.5 or 7 depending on the protein solution (see above). Figure 5.1 shows an illustrative plot of experimental inverse density ($1/\rho$) values as a function of protein weight fraction (w_2) for the simplest case of binary mixtures of aCgn and water. The data sets of inverse solution density *vs* protein weight fraction all showed qualitatively similar, linear behavior such as that illustrated in Figure 5.1. This was observed for all binary, ternary, and quaternary solutions discussed in subsequent sections. The lack of curvature in all data sets for $1/\rho$ *vs* w_2 indicated that the c_2 range was sufficiently low to assure that

protein-protein interactions could be neglected and the $c_2 \rightarrow 0$ limit was maintained for subsequent data analysis [23,49]. This is also consistent with the SLS data presented in Chapters 3 and 4 for quantifying the magnitude of protein-protein interactions for aCgn and IgG1 under equivalent solution formulations. The slope and intercept from the linear fit to a given data set of $1/\rho$ *vs* w_2 with fixed cosolute concentration(s) were then used with equation 1.7 to determine $\hat{V}_2$ for a given solution condition. The PSV values for cosolutes were determined in a comparable manner from linear fits for $1/\rho$ *vs* the cosolute weight fraction with all other concentrations held fixed.

5.2.4 Molecular Scale Simulations for Steric-Only Interactions at Infinite Dilution

The experimentally measured PSV values provide a quantitative assessment of the relative interactions between aCgn or IgG1 and water (*via* G_{12}) and cosolutes (*via* G_{23}). The values are, by definition, based on a reference state of an ideal gas mixture [59,95]. With that in mind, it is useful to consider what an ideal steric contribution to G_{2j} (for any $j \neq 2$) would be, and use that as an alternative reference state when comparing experimental G_{2j} values. One unambiguous option is the value of G_{2j} under infinite-dilution conditions of components 2 and j for steric-only interactions, termed $G^{\infty}_{2j,ST}$ in what follows. $G^{\infty}_{2j,ST}$ denotes the 2-body infinite dilution KB integral between components 2 and j, based on steric-only interactions as already discussed in Chapter 2. Therefore, $G^{\infty}_{2j,ST}$ values were calculated using the Mayer sampling with overlap sampling (MSOS) algorithm with the methodology already described in Chapter 2 for $B_{12,\mathrm{ST}}$ ($G^{\infty}_{2j,ST} = -2B_{2j,\mathrm{ST}}$). This methodology was applied to aCgn using the experimental crystal structure (PDB: 1EX3) to provide the three-dimensional protein structure, and to an atomistic homology model provided by BMS for the IgG1 molecule. Briefly, the

MSOS algorithm allows one to compute cluster integrals to obtain virial coefficient values. In the case of two-body integrals, these are equivalent to KB integrals at infinite dilution of both components. In the present examples, $G_{2j,ST}^{\infty}$ was computed by using an all-atom description of aCgn and IgG1 molecules and accounting for only steric interactions of each independent atom (see Chapter 2 for more details). Although the protein is treated in an all-atom fashion, the water or cosolute molecule is treated as a simple sphere with diameter σ_{exc}. In the case of water, σ_{exc} was taken as 3 Å, while for sucrose and trehalose, σ_{exc} was approximated as lying between 7 and 10 Å for the discussion below [97,212,213]. Sodium and chloride ions are estimated as 2.3 Å and 3.3 Å, respectively, but were treated as a single species with an average size of 2.8 Å for the discussion below [214].

5.3 Ternary aCgn Solutions: Water (1) + aCgn (2) + Cosolute (3)

Protein-water and protein-cosolute interactions in ternary solutions were evaluated for aqueous solutions of aCgn with different choices of cosolute: sodium phosphate, sodium chloride, trehalose and sucrose. The effect of adding sodium phosphate to a binary solution of water and aCgn was first evaluated for sodium phosphate molarities that are in the typical range used for buffering protein solutions (0 to 30 mM) at neutral pH. The PSV of sodium phosphate ($\hat{V}_3$) is reported in Table 5.1. This value was independent of buffer concentration in this range and is in good agreement with previous reports [215]. $\hat{V}_2$ values as a function of $c_3\hat{V}_3$ for sodium phosphate as the cosolute are shown in Figure 5.2. Based on equation 1.8 (or equivalently the ternary version of equation 5.12), the change in $\hat{V}_2$ as a function of buffer concentration ($\hat{V}_2$ *vs* $c_3\hat{V}_3$) can be related to the preferential interactions (G_{12}-G_{23}) between the osmolyte and protein molecules (relative to water-protein interactions) by

examining the linear region of $\hat{V}_2$ *vs* $c_3\hat{V}_3$. The results in Figure 5.2 show that $\hat{V}_2$ is effectively independent of buffer concentration over the range of concentrations that were tested, and this indicates negligible preferential interactions at this pH. A linear fit of the data gives a slope (G_{12}-G_{23}) that is not statistically different from zero (Table 5.2). Note that a zero slope in Figure 5.2 or equation 5.12 does not require ideal (non-interacting) behavior but instead indicates effectively equal contributions from protein-water interactions (*via* G_{12}) and protein-cosolute interactions (*via* G_{23}).

Table 5.1. PSV values with 95% confidence intervals for aCgn and for each of the cosolutes in water or in 5 mM sodium phosphate aqueous buffer

Component i	Concentration range	$\hat{V}_i$ in water (mL/g)	$\hat{V}_i$ in aqueous buffer (mL/g)
Sodium phosphate	0 – 30 mM	0.052 ± 0.002	N.D.
Sodium chloride	0 – 500 mM	0.307 ± 0.002	0.299 ± 0.006
Trehalose	0 – 24% w/w	0.653 ± 0.001	0.647 ± 0.002
Sucrose	0 – 24% w/w	0.623 ± 0.001	0.620 ± 0.002
aCgn	0 – 2 g/L	0.733 ± 0.009	0.733 ± 0.005
IgG1	0 – 2 g/L	0.706 ± 0.003	0.708 ± 0.002

Table 5.2. Values with 95% confidence intervals of G_{12} and (G_{12}-G_{23}) from linear fits to $\hat{V}_2$ *vs* $c_3\hat{V}_3$ for aCgn in water (ternary system), or in aqueous 5 mM sodium phosphate buffer (quaternary system) for each cosolute.

Solute	-G_{12} ternary (mL/g)	-G_{12} quaternary (mL/g)	(G_{12}-G_{23}) ternary (mL/g)	(G_{12}-G_{23}) quaternary (mL/g)
Sodium phosphate	0.735 ± 0.002	N.D.	21 ± 26	N.D.
Sodium chloride	0.730 ± 0.003	0.726± 0.005	-2.71 ± 0.79	-2.39 ± 0.81
Trehalose	0.722 ± 0.008	0.731 ± 0.009	-0.56 ± 0.06	-0.55 ± 0.10
Sucrose	0.725 ± 0.012	0.720 ± 0.009	-0.64 ± 0.09	-0.62 ± 0.10

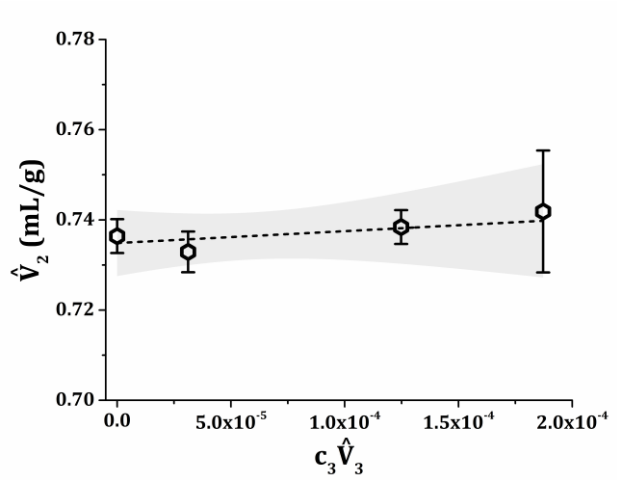

Figure 5.2. aCgn $\hat{V}_2$ values as a function of sodium phosphate concentration for the ternary water–aCgn–sodium phosphate systems at 25 °C and pH 7. The dashed line represents the linear fit while the surrounding gray area represents the 95% confidence level of the linear fit.

Sodium chloride, trehalose and sucrose were also tested as cosolutes to evaluate their preferential interactions in ternary water-aCgn-cosolute mixtures. Sucrose and trehalose were each evaluated between 0 and 24% w/w, while NaCl was evaluated between 0 and 500 mM. $\hat{V}_3$ values for sucrose, trehalose, and NaCl are shown in Table 5.1. These values were independent of cosolute concentration and are in excellent agreement with reports found in the literature. $\hat{V}_2$ values as a function of $c_3\hat{V}_3$ for each cosolute are shown in Figures 5.3A (NaCl), 5.3B (trehalose) and 5.3C (sucrose). For all three cosolutes, the results show that $\hat{V}_2$ decreases linearly with increasing cosolute concentration for all but the highest $c_3\hat{V}_3$ values, characteristic of negative (G_{12}-G_{23}) values and preferential accumulation of each cosolute around the protein surface. In contrast to the results for sodium phosphate, the slopes were statistically different than zero in each case (Table 5.2).

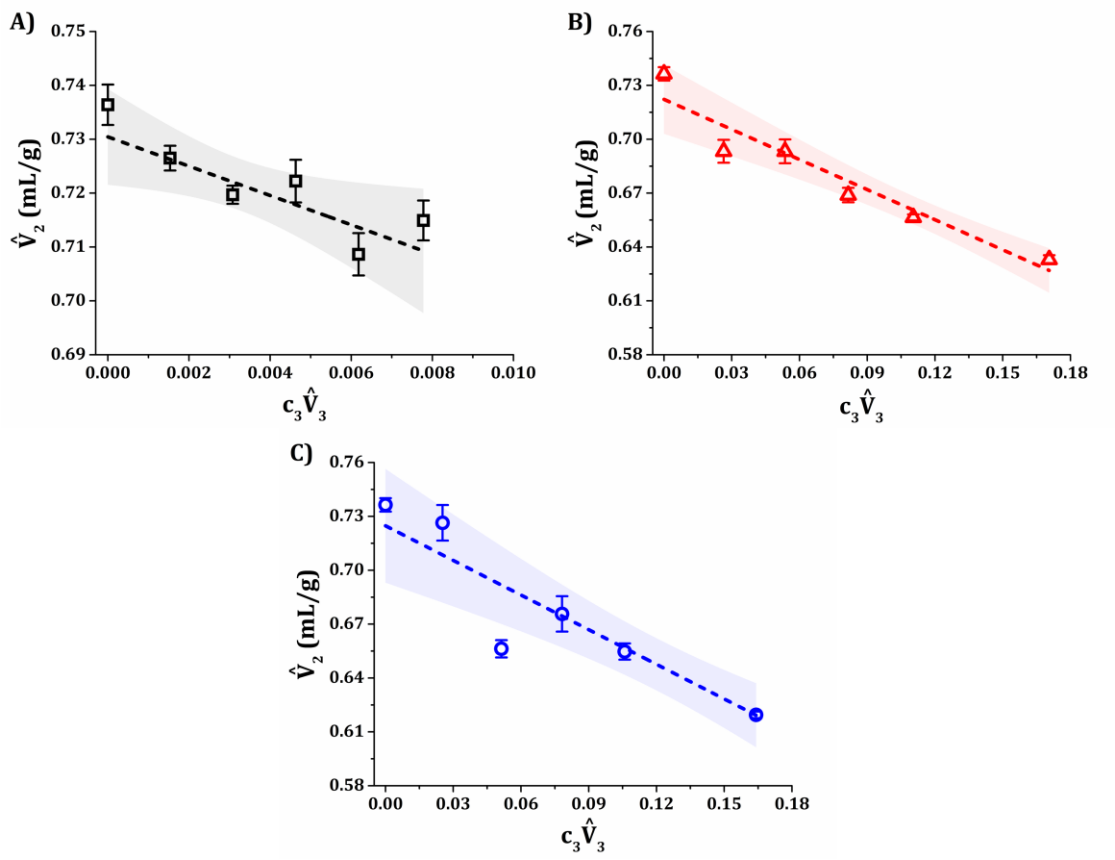

Figure 5.3. aCgn $\hat{V}_2$ values as function of cosolute concentration at 25 °C and pH 7 for ternary solutions of water, aCgn, and NaCl (panel A), trehalose (panel B) or sucrose (panel C). The dashed lines represent linear fits while the surrounding color shaded areas represent the 95% confidence level of the fits.

5.4 Quaternary aCgn Solutions: Water (1) + aCgn (2) + Cosolute (3) + Buffer (4)

The next step was to evaluate whether the presence of low-concentration buffer salts impact the preferential interactions when compared to the buffer-free ternary solutions. This was done by quantifying the changes in $\hat{V}_2$ with increasing sodium chloride, trehalose, or sucrose concentrations for aCgn in aqueous phosphate buffered solutions at a fixed buffer concentration (5 mM sodium phosphate). The same

methodology described above was used, with the single difference that every solution included 5 mM sodium phosphate. The measured $\hat{V}_3$ values for the ternary water-buffer-cosolute solutions are shown in Table 5.1. All these values are somewhat lower than those without the addition of buffer, albeit barely outside of the statistical confidence intervals in each case. This might suggest a small effect of adding sodium phosphate to water-cosolute solutions, but these are much smaller than the magnitude of the protein preferential interactions discussed below and in the previous section.

$\hat{V}_2$ *vs* $c_3\hat{V}_3$ values in phosphate buffered solutions with different cosolutes (component 3) are shown in Figures 5.4A (NaCl), 5.4B (trehalose) and 5.4C (sucrose). Similar to the case for the corresponding ternary solutions, the buffered solutions of aCgn with each of these cosolutes exhibit a negative slope for $\hat{V}_2$ *vs* $c_3\hat{V}_3$. Inspection of Figures 5.3 and 5.4 shows that the results with phosphate buffer are minimally or negligibly different from those for the ternary solutions without buffer. The values for fitted intercepts (-G_{12}) and slopes (G_{12}-G_{23}) are equal for ternary cases (Figure 5.3) and quaternary cases (Figure 5.4) within their statistical uncertainties (Table 5.2 and overlaid results in Figure 5.4).

The fact that phosphate buffer had little or no effect on the net protein-water and protein-cosolute interactions based on density measurements is consistent with equation 5.12, given that $c_4\hat{V}_4$ was constant and small for each case tested here, and the preferential interactions of sodium phosphate with aCgn in ternary protein-water-buffer solutions were statistically no different from zero (see discussion above, Figure 5.2 and Tables 5.1 and 5.2). This was only tested for phosphate buffer molarities up to 30 mM, therefore it is not clear if it would hold to much higher sodium phosphate concentrations.

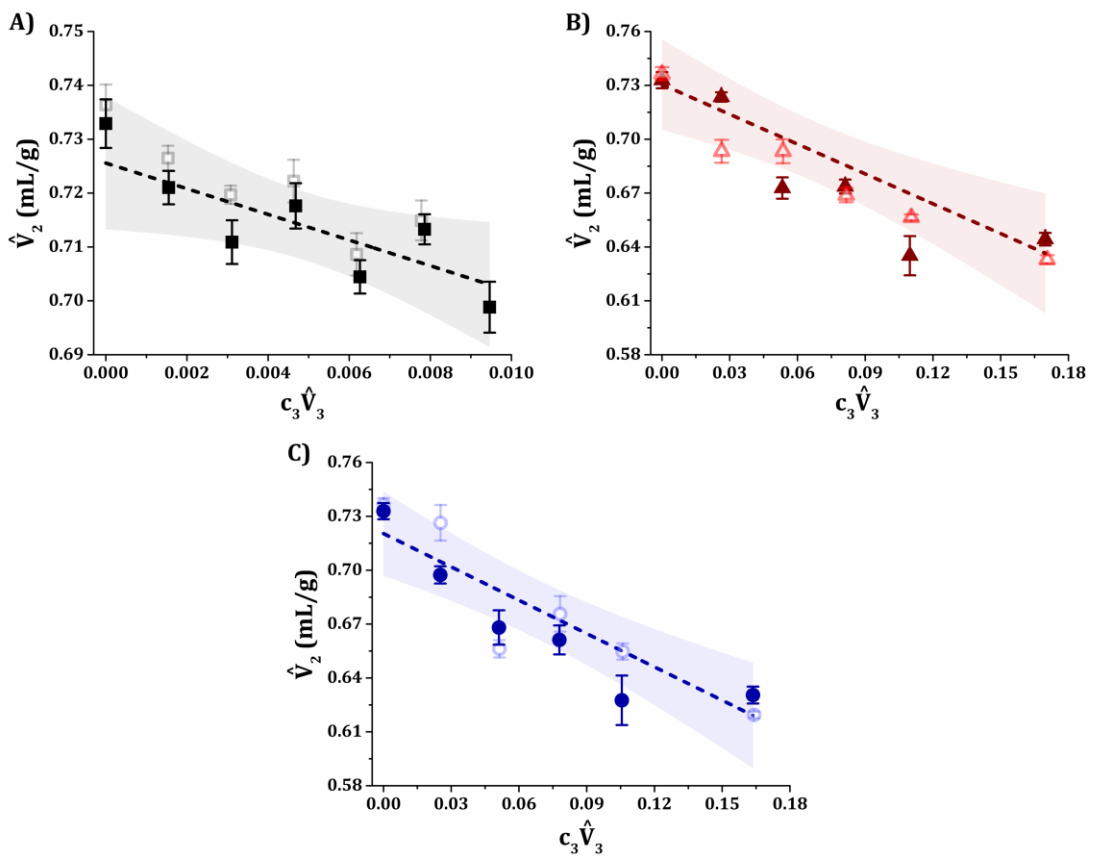

Figure 5.4. aCgn $\hat{V}_2$ values as a function of one of the cosolute concentrations at 25 °C and pH 7 for quaternary solutions of water, aCgn, 5 mM phosphate buffer, and NaCl (panel A), trehalose (panel B) or sucrose (panel C). The dashed lines represent linear fits while the surrounding color shaded areas represent the 95% confidence level of the fits. For comparison, closed symbols represent the quaternary solution data while the open symbols represent the ternary solution data from Figure 5.3.

Based on results presented in the literature [49], it is expected to see differences in the preferential behavior of an osmolyte as a function of pH, even for those that are inherently uncharged (such as sugars and non-ionic surfactants). In general, it is important to not assume a given behavior (*e.g.*, preferential accumulation or exclusion of cosolutes) will hold for a different protein or solvent environment. Instead this should

be verified experimentally using techniques such as those illustrated here or based on complementary techniques illustrated in the literature [44,49,52,101,102].

The results in Figures 5.3A and 5.4A indicate that NaCl is preferentially accumulated for aCgn. This behavior is similar to that seen for other proteins that have favorable salt-protein interactions that lead to “weak binding”, “preferential binding”, or “territorial binding” of counterions [44,45,154,156]. In the context of the Hofmeister series, salts that are preferentially accumulated at protein surfaces are termed chaotropes, while those that are preferentially excluded are termed kosmotropes [45,156,216]. A common assumption is that NaCl lies near the middle of the Hofmeister series for most proteins, and is neither preferentially accumulated nor preferentially excluded to a substantial extent. For aCgn at pH 7, it is apparent that NaCl behaves more like a chaotrope. Although the theoretical overall valence of aCgn is +5 at pH 7.0 (see Chapter 4), there are many charged acidic and basic side chains at neutral pH. Chapter 4 shows that electrostatic protein-protein interactions are net attractive under these solution conditions, consistent with a highly anisotropic surface charge distribution. This could potentially be used to rationalize the preferential accumulation of Na^+ and Cl^- ions around the protein surface.

Preferential accumulation of hydrophilic, uncharged cosolutes such as sucrose and trehalose is unusual when compared with canonical expectations with other proteins [44,46]. Adding sucrose or other disaccharides has historically promoted protein flocculation, phase separation, and/or increased free energies for protein unfolding [44,46]. Based on colloidal theories, those types of behavior have been attributed to unfavorable steric interactions between sugars and proteins (*i.e.*, excluded volume effects), relative to the interactions between water and the protein surface [44,97,217].

Therefore, the canonical expectation is that sugars such as trehalose and sucrose will be preferentially excluded from the surface of proteins in aqueous solution. The results in Figures 5.3 and 5.4 indicate the opposite behavior for aCgn, suggesting steric repulsions between the cosolutes and aCgn are overcome by significant attractive interactions. While unusual, preferential accumulation of sugars such as sorbitol has been reported in some cases. Experimental results for G_{12} and G_{23} do not, per se, explain why a given cosolute is preferentially accumulated or excluded relative to water. However, they provide a potentially useful test for models and theoretical treatments of protein-cosolute preferential interactions. It should also be emphasized that preferential behavior can change based on solution variables such as pH, which can change the surface charge distribution and accumulation of counterions [49,154,156].

5.5 Quaternary IgG1 Solutions: Effects of Sucrose in Protein-Protein Interactions

As mentioned above, sucrose has been historically categorized as a preferentially excluded excipient [44,46,217]. This preferential exclusion from the protein surface causes a reduction in the available free-volume for the protein molecules. This induces stronger attraction between proteins driven by the steric repulsion between protein and sucrose molecules (in the form of depletion forces). However, results presented in Figures 3.4 - 3.8 in Chapter 3 (SLS results from low to high c_2) cannot be explained with this commonly accepted sucrose-protein preferential interaction. Stronger protein-protein repulsions were observed for IgG1 and IgG4 solutions with sucrose than those where the sucrose was absent. This increase in repulsions was manifest in smaller magnitudes of ε_{SR} obtained from parameter tuning for formulation with sucrose compared to those without sucrose (Chapter 3).

In this case, PSV values were measured for similar solution conditions as shown in Chapter 3 for the IgG1 molecule. These results are shown in Figure 5.5A for pH 5 (10 mM acetate + sucrose) and 5.5B for pH 6.5 (10 mM histidine + sucrose) and the linear fit slopes (G_{12}-G_{23}) and intercepts (-G_{12}) are shown in Table 5.3. These results (negative slopes at both pH 5 and pH 6.5 in Figure 5.5 and Table 5.3) suggest preferential accumulation or solvation of sucrose around the protein surface, in agreement with the previously reported results for aCgn and historical results where sucrose was also found to solvate proteins [49].

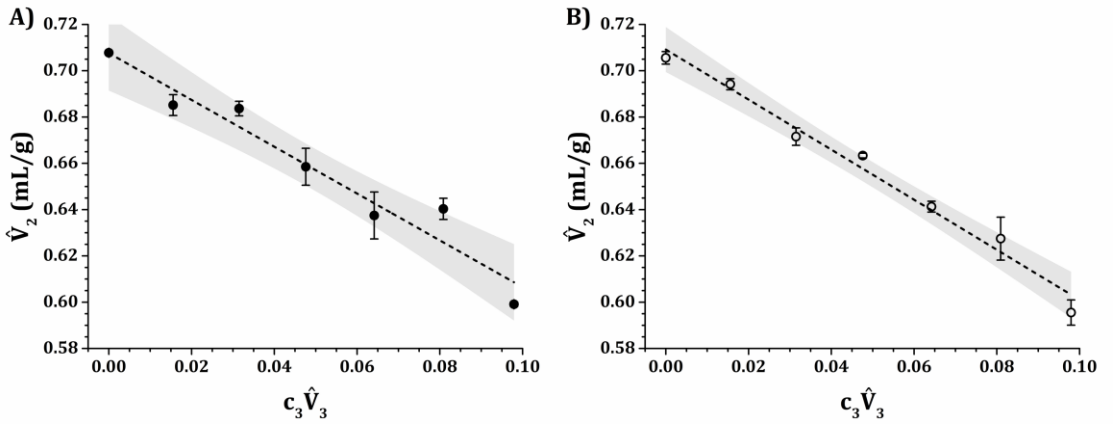

Figure 5.5. IgG1 $\hat{V}_2$ values as a function of sucrose concentrations at 25 °C for quaternary solutions of water, IgG1, sucrose and 10 mM acetate (pH 5, panel A) or histidine (pH 6.5, panel B) buffer. The dashed lines represent the linear fits. Shaded areas represent the 95% confidence level of the fits.

By combining the results from Figures 3.4-3.8 in Chapter 3, one could hypothesize how the addition of sucrose induces stronger repulsion between protein molecules. The accumulation of sucrose around the protein surface (*i.e.*, protein solvation by sucrose) might be affecting the way IgG1 molecules interact through two

different mechanisms. First, the solvation by sucrose around the protein surface displaces water molecules from the hydration layers (protein dehydration), decreasing the gain in free-energy upon protein-protein close contact [57,143,152,218]. Second, sucrose has a larger molecular diameter than water (see below), and a protein solvated by sucrose might experience an increase in its effective excluded volume, increasing the strength of apparent steric repulsion between proteins.

Table 5.3. $-G_{12}$ and (G_{12}-G_{23}) values with 95% confidence intervals from linear fits to $\hat{V}_2$ *vs* $c_3\hat{V}_3$ for the IgG1 at pH 5 and pH 6.5 for quaternary solutions with varying sucrose concentration.

Formulation	$-G_{12}$ (mL/g)	(G_{12}-G_{23}) (mL/g)
10 mM acetate, pH 5	0.710 ± 0.004	-1.02 ± 0.08
10 mM histidine, pH 6.5	0.709 ± 0.004	-1.08 ± 0.07

Both contributions are expected to be present upon addition of sucrose, but the current experimental data do not allow one to resolve which mechanism might be dominating the observed solution behavior. Independent of which mechanism contributes more to the obtained results, it can be concluded the sucrose might be acting as a "coating" agent around the protein in solution, enhancing its stability by decreasing the strength of short-ranged non-electrostatic attractions (*via* hydrophobic effects) between protein molecules [57,207,218]. This is in quantitative agreement with studies of sugars as stabilizing excipients in freeze-drying and spray-drying experiments, and with previous results of sugars acting as preferentially solvating molecules [49,219,220]. Additional SLS results shown in Chapter 3 for the IgG4 molecule in the presence of sucrose show equivalent results to those for the IgG1 molecule (increased

repulsion). This suggests that similar effects might be observed for a variety of molecules and might not be exclusive to aCgn and the IgG1 in the present work.

5.6 Competing Contributions to Preferential Interactions: Implications from Steric-Only Models

The results in Figures 5.2-5.5 and Tables 5.1 to 5.3 provide assessments of the relative protein-water and protein-cosolute interactions *via* G_{12} and G_{23}, respectively, for aCgn and the IgG1 molecule. However, any G_{2j} value (for any $j \neq 2$) only provides the net result of a combination of different contributions to the protein-water and protein-cosolute interactions, using an ideal gas mixture (*i.e.*, non-interacting mixture) as the reference state. As a result, all the G_{2j} values have negative values, indicating net repulsion in all cases. This is a consequence, at least in part, of the large magnitude of steric repulsion because the protein is much larger than the size of water or any of the cosolutes. For any real system, a more natural reference state would be one with steric-only interactions. In this context, comparing experimental G_{2j} values with calculated $G_{2j,ST}^{\infty}$ values (see section 5.2.4 above and in Chapter 2) can provide a useful context for further analyses, where the subscript ST denotes only steric contributions to G_{2j}, while the superscript ∞ indicates infinite dilution of both protein and component j (*i.e.*, only 2-body interactions).

Based on that reasoning, molecular simulations were performed to calculate $G_{2j,ST}^{\infty}$ based on the crystal structure of aCgn and the IgG1, as a function of the hard sphere diameter σ_{exc} for a simple spherical component j to represent a cosolute molecule. The resulting $G_{2j,ST}^{\infty}$ *vs* σ_{exc} values are shown in Figures 5.6A (aCgn) and 5.6B (IgG1). The values for $G_{2j,ST}^{\infty}$ for reasonable values of σ_{exc} for water (3 Å) and the different cosolutes (10 Å) in the present work are all large and negative as shown in Table 5.4.

The value of $G_{2j,ST}^{\infty}$ for aCgn and a water-sized sphere is approximately 42% larger (more negative) than experimental G_{12} values such as those in Tables 5.2. This is also observed with the IgG1 molecule, where $G_{21,ST}^{\infty}$ values were 64% higher than G_{12} values. $G_{2j,ST}^{\infty}$ can also be interpreted as the volume that the center of a given molecule j (*e.g.*, sucrose, trehalose, water, etc.) cannot access due to the volume displaced by the protein. Consequently, the fact that experimental G_{12} values (*i.e.*, $-\hat{V}_2$ for binary water-protein solutions) are less negative than the steric-only estimate is not surprising, as attractive interactions add positive contributions to KB integrals (equation 5.12), and water is expected to have strong favorable hydrogen bonding and van der Waals interactions with the surface of hydrophilic proteins such as aCgn.

The results in Figure 5.6 were fit to a linear function for $\sigma_{\text{exc}} \leq 8$ Å, as results in this region are linear. The intercept (or $G_{2j,ST}^{\infty}$ as $\sigma_{\text{exc}} \to 0$) results of this fits were then computed and presented in Table 5.4. These values can be interpreted as the true molecular volume, or the physical space occupied by the molecule in vacuum. Interestingly, it is common in some fields for experimental values of $\hat{V}_2$ in pure water to be used as the molecular volume or solvent-excluded volume for a given protein. The values in Tables 5.2 and 5.4 clearly show that such an assumption might be reasonable for aCgn if one accepts up to 7% deviation. However, this is not the case for the IgG1 molecule, as molecular volumes were higher than $\hat{V}_2$ values measured in ternary solutions at both pH 5 and pH 6.5 and deviations above 13% can be observed by comparing the values in Tables 5.2 and 5.4. Similarly, it would be more accurate to use the protein volume including, at least, the first hydration layer, as proteins are always expected to be hydrated as the disruption of the very first hydration layer around the

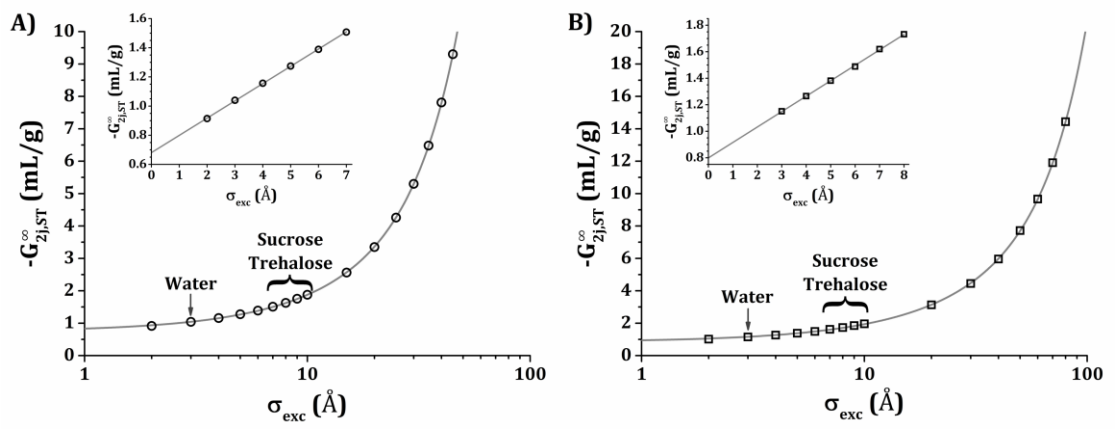

Figure 5.6. $-G^{\infty}_{2j,ST}$ values as a function of the hard sphere diameter of component *j*, using an aCgn crystal structure (PDB: 1EX3, panel A) and a homology model for the IgG1 (panel B) coupled with the MSOS algorithm. The black solid line in the main panels represents a cubic fit of the simulated data. The insets illustrate the linear region for $-G^{\infty}_{2j,ST}$ *vs* σ_{exc} use to extract $-G^{\infty}_{2j,ST}$ for $\sigma_{exc} \rightarrow 0$ Å (Table 5.4).

Table 5.4. Values with 95% confidence intervals of $-G^{\infty}_{2j,ST}$ computed using the MSOS algorithm. Values for $\sigma_{exc} \rightarrow 0$ were extrapolated using the intercept of a linear fit as an estimate for $-G^{\infty}_{2j,ST}$ (insets in figure 5.6).

Protein	$-G^{\infty}_{2j,ST}$		
	$\sigma_{exc} \rightarrow 0$	σ_{exc} = 3 Å	σ_{exc} = 10 Å
aCgn	0.690 ± 0.002	1.04 ± 0.01	1.88 ± 0.02
IgG1	0.80 ± 0.01	1.151 ± 0.003	1.96 ± 0.03

protein would result in a large energetic penalty (see Chapter 2) [57]. Table 5.4 also provides the value of $-G^{\infty}_{2j,ST}$ for the hydrated proteins, where discrepancies with $\hat{V}_2$ for both cases can be observed. Consequently, assumptions of $\hat{V}_2$ being equivalent to the molecular volume (either in vacuum or solvated) are fortuitous and only seem to be useful for aCgn. In general, this assumption is anticipated to be greatly in error for other

aqueous protein solutions because experimental $\hat{V}_2$ values have large contributions from attractions between water molecules and protein molecules. Those attractions can cause experimental $\hat{V}_2$ values to greatly underestimate the excluded volume and molecular volume of proteins in aqueous solution as in the case of the IgG1 molecule. Additionally, the -$G^{\infty}_{2j,ST}$ value for the IgG1 differ from those obtained in Chapter 2, as -$G^{\infty}_{2j,ST}$ (there termed $2B_{12,\mathrm{ST}}$) was equal to 0.924 mL/g in Table 2.2. These differences were found to arise from the use of a different crystal structure with explicit hydrogen atoms in the present calculation in comparison to that in Chapter 2. This addition of explicit hydrogen atoms is expected to provide, on average, a 1.2 Å layer around the protein, which accounts for most of the deviations found in the calculations.

Those limitations of existing assumptions notwithstanding, the results in Figures 5.3 to 5.6, combined with those in Tables 5.2 to 5.4 and the general expressions in equation 5.12, illustrate that experimental $\hat{V}_2$ values in binary and higher-order mixtures should be interpreted in terms of a balance of steric interactions that provide large negative contributions to G_{2j}, and other interactions that can provide either positive (attractive) or negative (repulsive) contributions to G_{2j} values. Hydration of protein surfaces is an obvious example of attractive interactions between proteins and water molecules. In the present case of aCgn and the IgG1, the fact that G_{23} values were less negative than G_{21} values shows that significant attractions can also occur between proteins and sugars (Tables 5.2 and 5.4). It is notable in this context that $G^{\infty}_{23,ST}$ for sucrose and trehalose (*e.g.*, the protein-sugar excluded volume) is necessarily more negative than $G^{\infty}_{12,ST}$ for water (Table 5.4) because the cosolutes have much larger molecular diameters than that of water, so ($G^{\infty}_{12,ST}$-$G^{\infty}_{23,ST}$) values are large and positive. Therefore, the fact that experimental (G_{12}-G_{23}) values are large and negative indicates

that the strength of net protein-cosolute attractions are that much larger than the net protein-water interactions. In other words, the attractions between sugar and protein molecules must be much stronger than those between water and protein molecules to overcome the intrinsic preferential exclusion originated by the difference in their molecular size (sucrose and trehalose being physically larger than water). Physically, this may arise due to a combination of sugar-protein hydrogen bonds and van der Waals interactions, and highlights that treating uncharged cosolutes simply as sterically excluded objects can be a large oversimplification. This is also observed in Figures 5.4-5.6, and Table 5.3 for the IgG1 molecule, where stronger protein-sucrose interactions are observed since (G_{12}-G_{23}) values are larger in magnitude than those from aCgn (Tables 5.2 and 5.3). Finally, the results presented in this dissertation are in excellent agreement with previously reported protein systems where reported $\hat{V}_2$ values are smaller in magnitude than $G_{12,ST}^{\infty}$ values [49].

5.7 Summary and Conclusions

A generalized expression for $\hat{V}_2$ in terms of Kirkwood-Buff integrals for an arbitrary number of cosolutes was derived based on a framework developed by Ben-Naim. This new expression was applied to protein solutions at infinite dilution ($c_2 \rightarrow 0$) and used to evaluate ternary (water-protein-osmolyte) and quaternary (water-protein-cosolute-buffer) solutions. Interactions between aCgn, water, and added cosolutes were quantified in terms of KB integrals at pH 7 and 25 °C. Sodium phosphate as an osmolyte showed no significant preferential interactions below 30 mM and dilute protein conditions. On the other hand, sodium chloride, sucrose, and trehalose showed preferential accumulation under dilute protein conditions in both ternary (no buffer) and quaternary (5 mM sodium phosphate buffer) solutions. No significant quantitative

differences were found between ternary and quaternary solutions for sodium chloride, sucrose, and trehalose, in agreement with the derived expression for multicomponent solutions and the measured aCgn-sodium phosphate interactions. Similarly, the interactions between the IgG1 molecule and sucrose were quantified using the same approach under the same solution conditions used in Chapter 3. These results combined with those in Chapter 3 suggest preferential accumulation (solvation) of sucrose around the protein surface of the IgG1 molecule. In this context, sucrose was found to act as a coating agent around the protein surface decreasing the effective hydration of the protein, and this led to stronger repulsion between protein molecules observed at both low and high c_2. Calculations of infinite-dilution steric interactions between aCgn and the IgG1 molecule with water-sized or sugar-sized species highlighted the presence of strong attraction between water or cosolutes and the protein molecules. The analyses in this chapter suggest reevaluation of the use of $\hat{V}_2$ as direct estimate of molecular volume for protein solution modeling and simulations. Independent of the model results, the experimental results for protein solutions with sucrose and trehalose contrast with commonly used preferential-interaction models that assume preferentially excluded behavior between proteins and sugar molecules, and highlights that protein-cosolute and protein-water interactions should be measured more broadly to aid in the development of improved understanding and modeling of preferential interactions and protein solution behavior.

Chapter 6

PREDICTING THE UNFOLDING TRANSITIONS OF POLYPEPTIDE SOLUTIONS WITH COARSE-GRAINED MODELING

6.1 Introduction

In previous chapters, a series of different CG molecular models was used to simulate protein interactions and solution behavior. However, rigid models were primarily considered in Chapters 3-5 based on the discussion in Chapter 2. In contrast to those simulations, this chapter will focus on fully flexible polypeptide chains, where the flexibility of the backbone is needed to properly capture the structural configurations of these chains in solution. Therefore, a more structurally detailed CG model than those used in Chapters 3-5 with added flexibility will be considered in this chapter.

Polypeptide self-assembly and aggregation can be used as a kinetically controlled and tunable process to form new structures based on various peptide sequences [31,84,221–223]. In each case, a molecular-scale description of the process(es) is needed if one wishes to design or predict the behavior and relative stability of key intermediate species – *e.g.*, as a function of peptide sequence and solution environment [123,224–228]. This is experimentally challenging, as few experimental techniques allow for the identification of the role of each specific residue in the unfolding and aggregation behavior of a defined sequence [16,31]. Additionally, such techniques are relatively low throughput, expensive, and/or have large sample material requirements. This poses challenges for testing and design of a range of protein

sequences and solution environments, and helps to motivate development of modeling approaches to aid in those efforts.

This chapter focuses on refining the previously proposed 4bAA protein model to more accurately quantify unfolding thermodynamics for a series of helical Ala-rich polypeptides as a function of chain length. The model was used with replica exchange molecular dynamics to make *a priori* predictions for the unfolding thermodynamics and pathways for a set of new Ala-rich peptides that were then experimentally synthesized and characterized with circular dichroism spectroscopy, for comparison to model predictions. The particular choices of peptides were based on previous Ala-rich sequences and future applications that focus on control of peptide-peptide interactions in multiblock peptide-polymer conjugates to manipulate assembly [84]. The results in this chapter have been published in a peer-reviewed journal [229].

6.2 Material and Methods

6.2.1 Four-Bead-per-Amino Acid (4bAA) Coarse-Grained Model

To predict and model the unfolding and self-association of peptides, along with the thermodynamics of the steps involved in those processes, an implicit-solvent CG molecular model was used. It was a modified version of the 4bAA force field proposed by Bereau and Deserno that was extended previously to include long-ranged screened electrostatic interactions [62,84]. Each amino acid is represented as the combination of four spherical beads as follows: one for the amide group (N), one for the alpha carbon (C_α), one for the carbonyl group (C') and one for the side chain group (C_β), as shown in Figure 6.1. The first three beads correspond to the peptide backbone and are able to interact *via* steric interactions, bond stretching and bending, and hydrogen bonding (H-

bonding). The last bead represents the side chain, with the exception of glycine where no fourth bead is included. The side-chain bead is used to capture the specificity of interactions between side chains, as well as the charge and relative hydrophobicity of each residue. Interactions between beads include local and non-local effects, as follows.

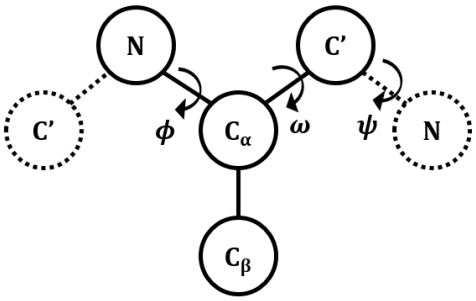

Figure 6.1. Schematic of the 4bAA CG model based on reference [62].

Local interactions correspond to bond distances (2-body interactions), bond angles (3-body interactions), and torsion and improper angles (4-body interactions) due to the planarity of the peptide bond. Local interactions exist only between beads that are covalently bonded to each other. Non-local interactions account for steric repulsion, hydrophobic attraction, hydrogen bonding, and electrostatic interactions that occur between beads that are not covalently linked to each other. Previously, the interaction parameters (other than electrostatics) were parameterized against NMR and crystallographic data in order to capture the secondary and tertiary structures of different polypeptides in their folded state(s) [62]. The electrostatic interactions were parameterized separately based on experimental light scattering data to give accurate values of osmotic second virial coefficients for globular proteins as a function of ionic

strength [66,84]. This extended Bereau-Deserno (EBD) coarse-grained model treats solvent (water + buffer + salts/solutes) implicitly in order to make the computations tractable for the range of different sequences and solution conditions of interest here and for future work. Effects of different salts are only captured in a mean-field manner, by accounting for deviations from the Debye-Hückel theory for monovalent ions in aqueous solution [66,76,84,148].

The force field for the EBD model is given by the linear combination of each contribution to the interactions as shown in equation 6.1, where w_{total} is the total potential energy (strictly, the potential of mean force) for a given configuration of molecule(s) in the simulation. u_{ij}^{bond} and u_{ijk}^{angle} correspond to the bond-length and bond-angle interactions between two and three contiguously bonded beads, respectively. u_{ijkl}^{tors} and u_{ijkl}^{imp} correspond to the torsion and improper angles interactions due to the backbone constraints. Those terms restrict the possible secondary structures of the peptide through the torsion angles ϕ, ψ and ω (Figure 6.1) and the stereoisomer constraints of an amino acid (*i.e.*, L- or D- side chain) [62,84]. The last four terms in equation 6.1 corresponds to the steric ($u_{ij}^{sterics}$), hydrophobic (u_{ij}^{bp}), H-bonding (u_{ij}^{bb}), and electrostatic (u_{ij}^{elec}) contributions to the potential energy. The *i*, *j*, *k* and *l* subscripts on the summations indicate that the summations are over all *i*-*j* pairs, *i*-*j*-*k* triplets or *i*-*j*-*k*-*l* quartets of beads in each corresponding case. All the beads are subject to steric repulsions, while only the amide and carbonyl groups can form hydrogen bonds, and the side chain beads contribute to hydrophobic and electrostatic interactions. More details about the origin of the CG model and parameters can be found in previous work [62,84]. As described below, the H-bonding parameter was refined as part of the present

work, to more accurately capture experimental unfolding thermodynamics of helical polypeptides. Additional model descriptions can be found in references [62,84].

$$w_{total} = \sum_{i<j} u_{ij}^{bond} + \sum_{i<j<k} u_{ijk}^{angle} + \sum_{i<j<k<l} \left(u_{ijkl}^{tors} + u_{ijkl}^{imp} \right) + \sum_{i<j} u_{ij}^{sterics} + \sum_{i<j} u_{ij}^{hp} + \sum_{i<j} u_{ij}^{hb} + \sum_{i<j} u_{ij}^{elec} \quad (6.1)$$

6.2.2 Molecular Dynamics Simulations

The conformational stability, intra-peptide and inter-peptide interactions of each sequence were evaluated by performing replica-exchange molecular dynamics (REMD) simulations at constant volume for a fixed number of peptides, coupled to a Nosé-Hoover thermostat to generate a correct canonical distribution for each replica [84,128,139]. REMD is a suitable method to accurately calculate the ensemble-averaged properties at defined temperature intervals, as it helps to prevent the simulation from becoming locked in local minima of the energy landscape at relatively low temperatures [139]. REMD can be coupled with weighted histogram analysis methods (WHAM) to determine the density of states of the system and then calculate all thermodynamic state functions and ensemble averaged structural properties over the range of simulated temperatures [84,128,230–233].

All REMD results presented here were simulated in a cubic box including a single peptide or two peptides with the same sequence. The single-peptide simulations were used to assess changes in the conformational stability of the sequences in the ideal dilute regime (no peptide-peptide interactions) where a box length (L) of 18 nm was used to ensure the box size was larger than the size of the completely extended peptide. Two-peptide simulations were implemented to evaluate the initial effect of peptide-

peptide interactions on the conformational stability and likely self-assembly behavior of selected peptide sequences. In this case, L was set to make the effective or local concentration of the peptide solution equal to 1 mM. However, the term "concentration" in this context is more properly understood as a measurement of "confinement" or "crowding" for the two peptides within the simulation box, and should not be confused with the experimental definition of c_2 used in previous chapters. Experimentally, high peptide concentrations cause higher probabilities of inter-peptide interactions, a decrease in accessible volume ("crowding effects"), as well as higher probabilities of three or more peptides interacting simultaneously (see Chapters 2-4). The former two are captured reasonably *via* confinement of two peptides within a small simulation box [142,144]. However, the lattermost effect cannot be captured in simulations unless multiple (more than 2) peptides are simulated simultaneously, and this was not done in this chapter in the interest of computational time limitations. Chapters 2-4 illustrate true high-concentration simulations using less structurally resolved models.

Solution pH was held constant at 7.4, as this is relevant to biotechnology applications and allows one to treat all Lys (K) and Glu (E) residues as charged (+1 and -1 respectively). The total ionic strength, *TIS*, was kept constant at 20 mM to avoid complete screening of the electrostatic interactions while maintaining experimentally realistic *TIS* conditions. A set of replicas in REMD were distributed between 220 K and 400 K with the total number of replicas adjusted to each simulated peptide(s) and solution conditions to assure a replica swap acceptance ratio between 30% and 50%. An integration time step of $\delta_t = 0.0035\tau$ (~1 fs for $\tau = 0.3$ ps) was used and swaps between replicas were attempted every 1τ for replica-exchange steps. The initial configuration of each replica was chosen randomly with the peptides in a helical configuration. Each

simulation employed two thermal equilibrium periods of $5x10^4$ τ each, using standard molecular dynamics and REMD, respectively. A sampling period of $1x10^6$ τ was performed using REMD, where the molecular configurations and energy values of each replica were stored every 1τ for subsequent analysis.

Heat capacity values were reconstructed by using WHAM to evaluate the density of states to compute the variance in the energy of the system [84,234]. Mid-points of unfolding temperatures (T_m) and thermodynamic properties (*e.g.*, enthalpies, entropies and free energies of unfolding) were calculated by using a two-state transition model fitted to the simulated heat capacity values [235]. Ensemble-averaged helix contents were calculated by computing the number of residues in a helical configuration and then averaging over the weighted ensemble of configurations during the simulation. For a residue *i* to be defined as helical, it had to comply with the following characteristics: (a) it must form a hydrogen bond with its *i+4th* or *i-4th* residue, (b) its torsion angle φ lies between -150° and -30°, and (c) its torsion angle ψ lies between -90° and 10° [62]. Other order parameters were also calculated from the simulation: the CONGENEAL score [236], using a perfect helical peptide as a reference; the radius of gyration (R_g) [84]; the number of side-chain contacts, calculated as the number of side-chain beads with energy magnitude $\geq 0.49\ k_BT$ (hydrophobic or hydrophilic attractions); the number of charge-charge contacts; the number of C_α-C_α contacts; the total number of hydrogen bonds and the C_α-C_α mean square distance (RMSD). Given the nature of REMD, and the use of an implicit solvent model, it was not possible to reliably infer kinetics or time scales for transitions such as unfolding from these calculations.

Principal component analysis (PCA) was employed to assess the multivariate nature of protein unfolding and solution behavior [202]. All the computed order

parameters were combined and subjected to a PCA treatment: the covariance matrix of the normalized order parameters was calculated for each replica. The eigenvalues and eigenvectors of each covariance matrix were subsequently calculated, and all the normalized order parameters were subjected to a vector projection using the eigenvectors obtained from the replica closest to the mid-point of unfolding in order to study the same projection (direction of change) for all the replicas. Finally, a histogram analysis was employed to evaluate correlations between principal components (PCs). The obtained PCs were organized descendingly by eigenvalues (*e.g.*, PC1's eigenvalue > PC2's eigenvalue) as the magnitude of their corresponding eigenvalues are representative of the amount of information captured by each respective PC. This allows one to select only the PCs with highest eigenvalues for further analysis, and it was found that more than 80% of the information was contained in PC1 and PC2 together in all cases. Therefore, a combined PC1-PC2 analysis is the focus in the following subsections. Inclusion of the other PCs beyond PC1 and PC2 did not alter any of the conclusions drawn below. Additionally, cartoons of snapshot structures were added for T values around the T_m for better readability. Additional information regarding the methodologies of the 4bAA model, REMD, WHAM and PCA can be found in Blanco's and Bereau-Deserno's works, among others [62,84].

6.2.3 Peptide Synthesis and Purification

Lyophilized peptide powders were provided by Bradford Paik, and his description of the materials and methods from reference [229] have been adapted here for the sake of completeness. All materials were purchased from Fischer Scientific (Pittsburgh, PA) except where otherwise indicated. Peptides were synthesized on a Rink Amide Resin (ChemPep, Wellington, FL). Specifically, the sequences listed in Table

6.1 (shorthand notations used in Table 6.1: AQEK, FAQEK, and AQK18) were synthesized with a PS3 peptide synthesizer (Protein Technologies, Tucson, AZ). Longer sequences (shorthand notation in Table 6.1: AQK27 and AQK35) were synthesized with a Focus XC peptide synthesizer (AAPTec Inc, Louisville, KY). Fmoc-alanine, Fmoc-lysine(boc), Fmoc-glutamic acid (t-butyl), Fmoc-glutamine(trt), and Fmoc-phenylalanine were all purchased from ChemPep. The N-terminus of each peptide was acetylated, and peptides were cleaved in 95% trifluoroacetic acid (TFA), 2.5% H_2O, and 2.5% triisopropylsilane (Sigma-Aldrich, St. Louis, MO). TFA was mostly evaporated, and peptides were then precipitated twice into cold ethyl ether. Samples were redissolved in water, frozen in liquid nitrogen, and lyophilized. Dried samples were then reconstituted in water and purified by preparative-scale reverse-phase high-performance liquid chromatography (RP-HPLC) using a Waters Xbridge BEH130 Prep C-18 column. The mobile-phase comprised gradients of degassed, deionized water with 0.1% TFA and acetonitrile with 0.1% TFA, at a flow rate of 21 mL/min. Peptide was detected by UV absorbance at 214 nm, and fractions were collected and lyophilized. Molecular weights of the purified peptides were verified by electrospray ionization mass spectroscopy (ESI-MS).

Table 6.1. Synthesized peptide sequences and short-hand notations

Sequence	Short-hand notation	Molecular weight (kDa)
AAQEAAAAQKAAAAQEAAA	AQEK	2.04
AAQEFAAAQKAAAFQEAAA	FAQEK	2.19
K(AAAQ)4K	AQK18	1.95
K(AAAQ)3K(AAAQ)3K	AQK27	2.92
K(AAAQ)4K(AAAQ)4K	AQK35	3.75

6.2.4 Peptide Solutions and Circular Dichroism (CD) Spectroscopy Experiments

The experimental characterization of the peptide solutions was performed by Bradford Paik, and his description of the methods and results have been adapted from reference [229] for the sake of completeness and to provide experimental comparison to computer simulations. Experimental characterization of the average secondary structure of peptide samples was conducted *via* CD spectroscopy on a Jasco 810 CD spectropolarimeter (Jasco Inc, Easton, MD, USA). Peptides were dissolved in 10 mM potassium phosphate at pH 7.4 with a final peptide concentration of 0.125 g/L. *TIS* was adjusted for select sample preparation by addition of 500 mM potassium chloride or potassium fluoride stock solutions. Samples were briefly sonicated to aid in the dissolution of the lyophilized peptides, and CD spectra were recorded using a quartz cell with 1 mm optical path length. Samples for full wavelength scans at various temperatures were cooled for three minutes at 0 °C prior to the start of the experiment. Scans were recorded from 0 °C to 80 °C, at 10 °C increments, with a step-and-hold heating rate of 1 °C/min between temperatures. Samples underwent subsequent cooling to 0 °C at the same increments and cooling rate between isothermal hold steps. For each wavelength scan, the scanning rate was 50 nm/min, with a response time of 4 s. Wavelengths from 195 nm to 250 nm were recorded at increments of 0.5 nm.

Measurement of peptide unfolding was also conducted by recording the mean-residue ellipticity (MRE) values at 222 nm ($[\Theta]_{MRE,222}$) every 0.5 °C, from 0 °C to 80 °C, while the temperature was increased at a constant rate of 1 °C/min. Samples were subsequently cooled back to 0 °C at 1 °C/min while recording $[\Theta]_{MRE,222}$. In some cases, peptide solutions at higher peptide concentrations (100 μM or 1 mM) were prepared as above, and then were incubated at 60 °C for two weeks to observe whether slow conformational transition(s) or changes in aggregation state occurred. Full wavelength

scans of these incubated samples were performed as described above, but at a single temperature of 60 °C.

6.3 Tuning the EBD 4bAA Model for Unfolding Thermodynamics

As noted in Chapter 1, a challenge for both atomistic and CG molecular models is to accurately produce the thermodynamics of unfolding for polypeptides and proteins [65,69,85]. Consequently, a series of single-peptide simulations using different model parameters were performed for three previously studied polypeptide sequences, in order to test and possibly refine the original model parameters to assure the original or modified model can capture the unfolding thermodynamics at least semi-quantitatively. The polypeptide sequences were taken from Scholtz *et al*. [237]. The generic formula $Y(AEAAKA)_nF$ was used and values of n = 3, 4 and 5 were considered to assess the effects of chain-length on helix stability, as was done experimentally by Scholtz *et al*. [237]. Table 6.2 provides short-hand notations and sequences for each of the peptides considered in this work. A solution pH and *TIS* values of 7.4 and 20 mM, respectively, were used to match the reported experimental conditions. Although no T_m values were explicitly reported in reference [237], T_m values were extracted from the data by identifying the temperature at which the second derivative of the mean residue ellipticity at 222 nm was equal to zero.

To tune the EBD 4bAA model, only parameters affecting non-local forces were modified so as to maintain the prior structural agreement of folded structures with NMR and crystallographic measurements [62]. The parameters that characterize the strength of hydrophobic attractions (ε_{HP}) and hydrogen bonds (ε_{HB}) were subject to a simple perturbation analysis to assess which of those exerts the strongest effect on the T_m value. The originally formulated values by Bereau and Deserno were perturbed ±10% around

the previously reported parameter values [62]. Figure 6.2A shows an example of the simulation results and analysis for the YAF3 sequence from Scholtz *et al*., where the constant volume heat capacity (c_v) is plotted as a function of temperature (T) and the T_m is identified as the T value at which the heat capacity reaches a maximum. From Figure 6.2A, it is clear that ε_{HB} is the most significant parameter affecting the T_m. This can be anticipated, as unfolding transitions for helical peptides necessarily require breakage of back-bone hydrogen bonds.

Table 6.2. Peptide sequences use for tuning and short-hand notations

Sequence	Short-hand notation	Molecular weight (kDa)
Y(AEAAKA)3F	YAF3	2.30
Y(AEAAKA)4F	YAF4	2.95
Y(AEAAKA)5F	YAF5	3.59

Based on those results, simulations at values of ε_{HB} = 4.5, 5.0 and 5.5 k_BT were performed for the YAF3, YAF4 and YAF5 sequences to find the optimum ε_{HB} value that allowed experimental and predicted T_m values to align quantitatively. Note that because this is an implicit-solvent model, all H-bonding energy values are inherently relative to water-peptide H-bonding energies. Figure 6.2B shows the simulated T_m values as a function of the ε_{HB} parameter while Figure 6.2C shows the comparison between experimental and simulated T_m for the same ε_{HB} values. A simple linear interpolation was used to find the optimal ε_{HB} value that matches both experimental and simulated T_m values. The resulting hydrogen bond strength was ε_{HB} = 5.09 k_BT, which was used for subsequent simulations. For reference, this is approximately 85% of the value used in previous work [62,84].

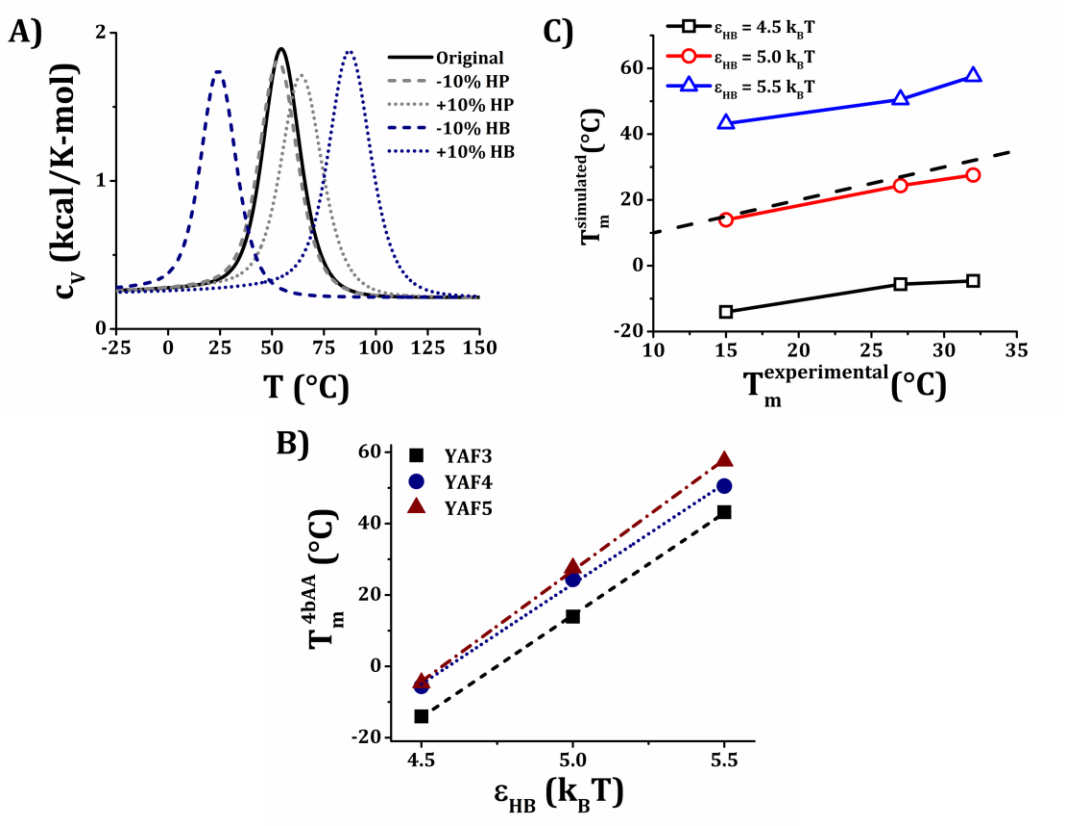

Figure 6.2. **Panel A:** perturbation analysis of the hydrophobic (ε_{HP}) and H-bonding (ε_{HB}) parameters for the YAF3 sequence at pH 7.4 and *TIS* = 20 mM. **Panel B:** mid-point of unfolding obtained from the 4bAA model as a function of the H-bonding parameter for the sequences in Table 6.2. Straight lines represent a linear interpolation between the data points. **Panel C:** comparison between simulated and experimental results of the T_m for three ε_{HP} values. The dashed line represents a 1:1 match ($y = x$).

6.4 Simulating Thermal Unfolding for Ala-Rich Peptides Using the Tuned EBD 4bAA Model

Following the tuning of the model, the sequence AQEK was initially used. For additional sequences, Phe residues were substituted for Ala residues at positions 5 and 15 (A5F and A15F mutations) to provide hydrophobicity, and pi-stacking capability, a hallmark of amyloid fibril formation. This substitution yielded the sequence FAQEK

(Table 6.1). A series of variants with different sequence lengths were synthesized using AAAQ repeats with terminal lysine residues to provide solubility, and a central lysine was used for the two longest sequences. As an additional consideration, glutamic acid residues were eliminated from the sequences to provide a uniform charge. The resulting series of peptide sequences are summarized in Table 6.1, and were selected to evaluate the capabilities of the tuned CG model to capture the effect of (i) selective point mutations for residues with different hydrophobicity and (ii) modifications in the length of the polypeptide. AQEK and FAQEK sequences were selected based on the former argument, while the AQK18, AQK27 and AQK35 sequences were selected for the latter argument. Additionally, this particular set of sequences allows evaluation of the sensitivity of the tuned computational approach for thermodynamic properties that compare to those reported from prior work [62,84].

To validate the thermodynamic tuning and better understand the conformational stability of the proposed sequences, REMD simulations were carried out at pH 7.4 and *TIS* = 20 mM for each of the five sequences listed in Table 6.1 using one peptide in the simulation box for a given sequence (referred as single-peptide simulations in what follows). A natural output from REMD simulations with WHAM is the polypeptide heat capacity (c_v) as a function of temperature (T). Using $c_v(T)$ profiles makes no assumptions regarding a two-state or multi-state model for the process of unfolding, and allows one to easily characterize T_m and assess the cooperativity of the process *via* the location of the peak position and sharpness of the transition [84,111,235,238]. One can subsequently analyze configurations from selected temperatures along the $c_v(T)$ profile to deduce the structural changes that occur during thermal unfolding [235].

Illustrative results are given in Figure 6.3A, where $c_v(T)$ is shown for each of the five peptide sequences in Table 6.1. Inspection of Figure 6.3A shows that each of the five sequences displays reasonably two-state unfolding behavior, based on the observation of a single, relatively sharp and symmetric peak in each case. Configurations at temperatures significantly lower than T_m correspond to the predominantly folded states, and while those significantly above T_m correspond to unfolded states. This can be visualized from a structural perspective by plotting the average helix content as a function of temperature (Figure 6.3B), where the transition from folded to unfolded states is essentially a sigmoidal function, as expected for an idealized two-state unfolding transition [235]. Moreover, T_m values from the $c_v(T)$ plots correspond to a 50% change in the average helix content from the folded to the unfolded state. For a two-state transition, T_m is also expected to correspond to the temperature at which the inflection point occurs in Figure 6.3B for a given sequence, and this agrees with the standard analysis of circular dichroism data (see section 6.7) [237].

6.5 Unfolding Thermodynamics as a Function of Peptide Sequence from REMD Simulations

From inspection of Figure 6.3, one can observe that the length of the peptide sequences considerably affects the thermal stability of the peptide. The T_m increases greatly as the chain-length increases as shown in Figure 6.4A, and as expected based on prior experimental results [237]. However, an unexpected result is the noticeable difference in T_m that was shown when comparing the AQEK, FAQEK and AQK18 sequences. The AQEK sequence showed a considerably lower T_m than the FAQEK and AQK18 sequences. This was initially assessed by evaluating the differences in Gibbs energy (ΔG_{un}, Figure 6.4B), enthalpy (ΔH_{un}, Figure 6.4C) and entropy (ΔS_{un}, Figure

6.4D) between the folded and unfolded configurations as a function of temperature obtained by fitting a two-state transition model to the simulated $c_v(T)$ results (Figure 6.3A), as is done in the case of experimental results from differential scanning calorimetry (DSC) experiments [235]. These thermodynamic properties correlate with the length of the sequence (longer chains resulted in higher values of T_m, ΔG_{un}, ΔH_{un}, and ΔS_{un}) so a much higher entropy change upon unfolding occurs as the chain-length increases, and this is compensated by a much higher enthalpy change upon unfolding. Overall, this leads to a net increase in the Gibbs energy of unfolding and T_m as chain-length increases. Despite having a higher enthalpy of unfolding, the AQEK sequence showed a much lower T_m than the FAQEK and AQK18 sequences (Figure 6.4C). Figure 6.4 indicates that these differences in T_m are mainly caused by a considerable difference in the entropy of unfolding. However, the energetic behavior also correlates with the trends in (all but the AQEK) T_m values (Figures 6.4A and 6.4C), so a balance between energetic and entropic behavior was considered next.

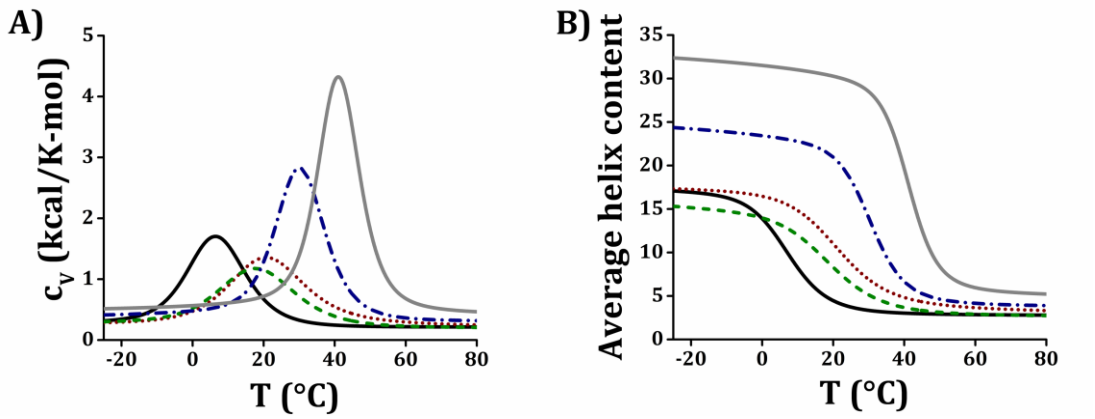

Figure 6.3. Heat capacity (panel A) and average helix content (panel B) as a function of temperature for single-peptide simulations. Lines represent the AQEK (black solid), FAQEK (red dotted), AQK18 (green dashed), AQK27 (blue dashed-dotted) and AQK35 (grey solid) at pH 7.4 and *TIS* = 20 mM.

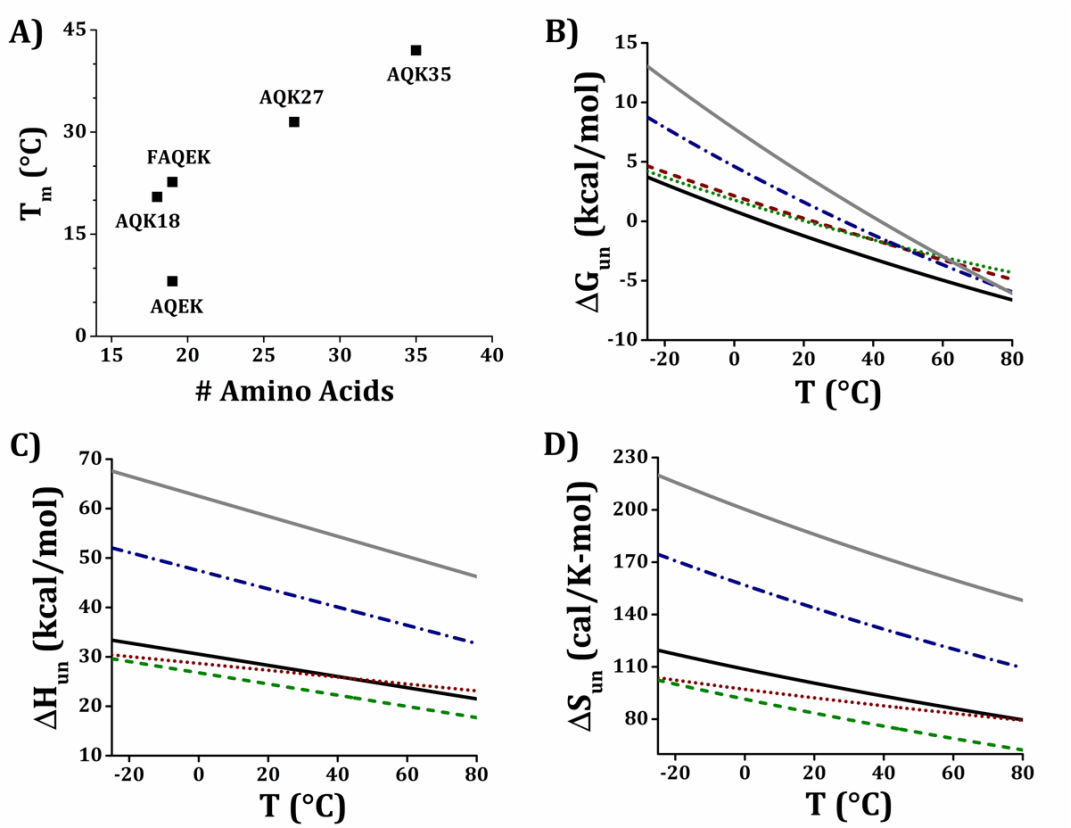

Figure 6.4. **Panel A:** mid-point of unfolding temperature from fitting the simulation results to a 2-state model, as a function of the number of amino acids in the sequence. **Panels B-D:** Gibbs energy (panel B), enthalpy (panel C), and entropy (panel D) of unfolding as a function of temperature obtained from a standard two-state transition model. Line types are as in Figure 6.3.

This was assessed by evaluating the molecular events involved in the unfolding transitions of these five sequences. Changes in potential energy and non-local contributions within the polypeptide chain were evaluated to further understand the molecular events that lead to the results in Figures 6.3-6.4. It is useful to point out that the local contributions (*i.e.*, average bond lengths and angles, and torsional and improper angles) are the same for all the sequences, so differences in the simulated thermodynamic properties should arise mainly from non-local contributions (*i.e.*,

sterics, H-bonding, side chain hydrophobic and electrostatic interactions). Although the total potential energy is used to compute the heat capacity, it does not show any features that conclusively explain significant changes in T_m (Figure 6.5A). The H-bonding energy (Figure 6.5B) reflects a similar trend observed from Figure 6.4A, suggesting that the increased number of H-bonding interactions as chain-length increases is mainly responsible for increases in T_m and ΔH_{un}, in accordance with standard arguments [235,237,239].

However, this does not sufficiently explain the differences between AQEK, FAQEK and AQK18. The first two sequences showed equal H-bonding baselines (Figure 6.5B) while the third one showed weaker H-bonding energy as expected for a shorter peptide, and that would seem to contradict the results in Figure 6.4A. In that regard, the contribution from hydrophobic interactions between side chains (Figure 6.5C) also shows a correlation with changes in T_m. Stronger hydrophobic interactions lead to higher T_m values, as expected from prior results [240–242]. However, caution is needed before concluding that experimental T_m values should scale with larger hydrophobic interactions between side chains. Stronger hydrophobic interactions of exposed side chains in the folded or unfolded states are expected to lead to stronger polypeptide self- association, which may be unavoidable at finite concentrations needed for experimental measurements of unfolding. The latter typically leads to lower T_m values [225,240,243]. Finally, the electrostatic contributions (Figure 6.5D) do not show any correlation with observed T_m values, and the energy values are two orders of magnitude smaller than the other two contributions. Therefore, those contributions are considered effectively negligible in terms of the thermodynamics of the unfolding process of these peptides. Consequently, the results support the view that a balance

between the chain length (and number of H-bonding contacts) and side-chain hydrophobicity dominates the conformational stability of these peptides. While the former increases both H-bonding and hydrophobic energies for chemically similar sequences, the latter can alter the position of the T_m by a few degrees without modifications of the length of the chain.

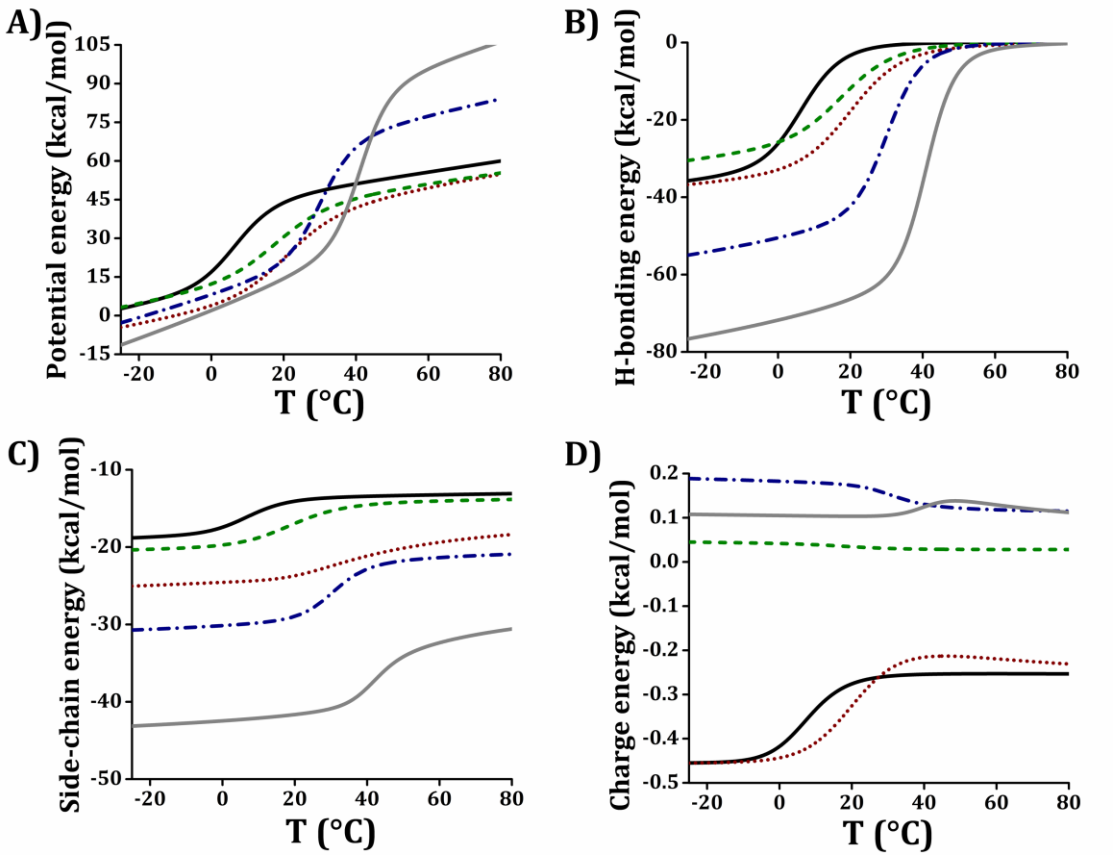

Figure 6.5. Total potential energy (panel A), and its contributions from hydrogen bonding (panel B), hydrophobic attractions (panel C) and electrostatic interactions (panel D) as a function of temperature obtained from WHAM. The total potential energy includes local energies due to bond fluctuations. Line types are the same as in Figure 6.3.

The above points notwithstanding, by analyzing the shape of the curves in Figure 6.5, an interesting behavior is notable for the FAQEK and AQK35 sequences in comparison to the other three sequences that could be useful in explaining the deviations from the simple length-dependent scaling in Figure 6.4A. First, all the sequences appeared to follow a two-state transition (Figures 6.3 and 6.5A-B), which is consistent with the experimental results for those sequences shown and discussed below. Nevertheless, Figures 6.5C-D show different behavior. The hydrophobic contribution shows sigmoidal behavior for the AQEK, AQK18 and AQK27 sequences but it exhibits slightly different behavior for the FAQEK and AQK35 sequences. The FAQEK sequence shows a subtle increase in hydrophobic energy without a defined upper base line as temperature rises, despite substantial changes and defined base lines in H-bonding energy and average helix content at the same temperatures.

Together, these indicate that the unfolding transition might not be representative of an idealized two-state transition. That is supported by Figure 6.5D, where a maximum in the electrostatic contribution is observed for FAQEK above T_m, so the sigmoidal behavior is lost. Such a loss is not observed for AQEK despite its similar sequence and identical location of charged residues. This maximum suggests that FAQEK is subject to a hydrophobic collapse, which allows the equally charged residues (Glu) to approach closer while unfolding. Consequently, the addition of two Phe residues changes the unfolding events compared to AQEK, and this increases the T_m by considerably decreasing the entropy of unfolding. Similarly, AQK35 shows a maximum in the charge energy, which suggests that the charge-charge distances (Lys-Lys distances) decreases as the peptide unfolds. This can only be explained if there is a hydrophobic collapse after unfolding in an equivalent way as that of FAQEK, which causes Lys residues to

similarly become closer during unfolding. Conversely, AQK18 and AQK27 shows a similar behavior to AQEK as the electrostatic energy decreases during unfolding following a sinusoidal transition. This suggests that Lys residues are overall moved further apart during unfolding. Thus, this analysis suggests the existence of intermediate states during unfolding for the FAQEK and AQK35 while validating the former assumption of an idealized two-state transition for the AQEK, AQK18 and AQK27. Therefore, an additional approach was taken to elucidate the proposed hypotheses.

6.6 Principal Component Analysis (PCA) to Obtain Unfolding Intermediates

PCA was performed with the order parameters obtained from the simulations to evaluate the molecular/structural basis for changes in unfolding thermodynamics across the different sequences. In Figures 6.6-6.12, the normalized probability (Π) of observing principal component 1 (PC 1) and principal component 2 (PC 2) is plotted (log scale) in a contour plot as a function of PC 1 and PC 2 (*i.e.*, $\log_{10} \Pi$ (PC 1, PC 2) vs [PC 1, PC 2]) for the sequences in Table 6.1. For two-state unfolding transitions, only two well-identified and well-populated states or regions should be observed in this type of plot. This was also done for the potential energy, by replacing either PC1 or PC2.

The results for AQEK and FAQEK show markedly different behaviors in Figure 6.6, where panels A-C and D-E show, respectively, representative probability surfaces for AQEK and FAQEK, as a function of temperature. The probability surfaces are for $T << T_m$ (panels A and D), $T \sim T_m$ (panels B and E) and $T >> T_m$ (panels C and F). Inspection of panels A to C shows that AQEK follows a reasonably ideal two-state transition, as only two main regions are observed. On the other hand, FAQEK shows intermediate states, as four regions are observed in Figure 6.6E and two regions are observed in Figure 6.6F.

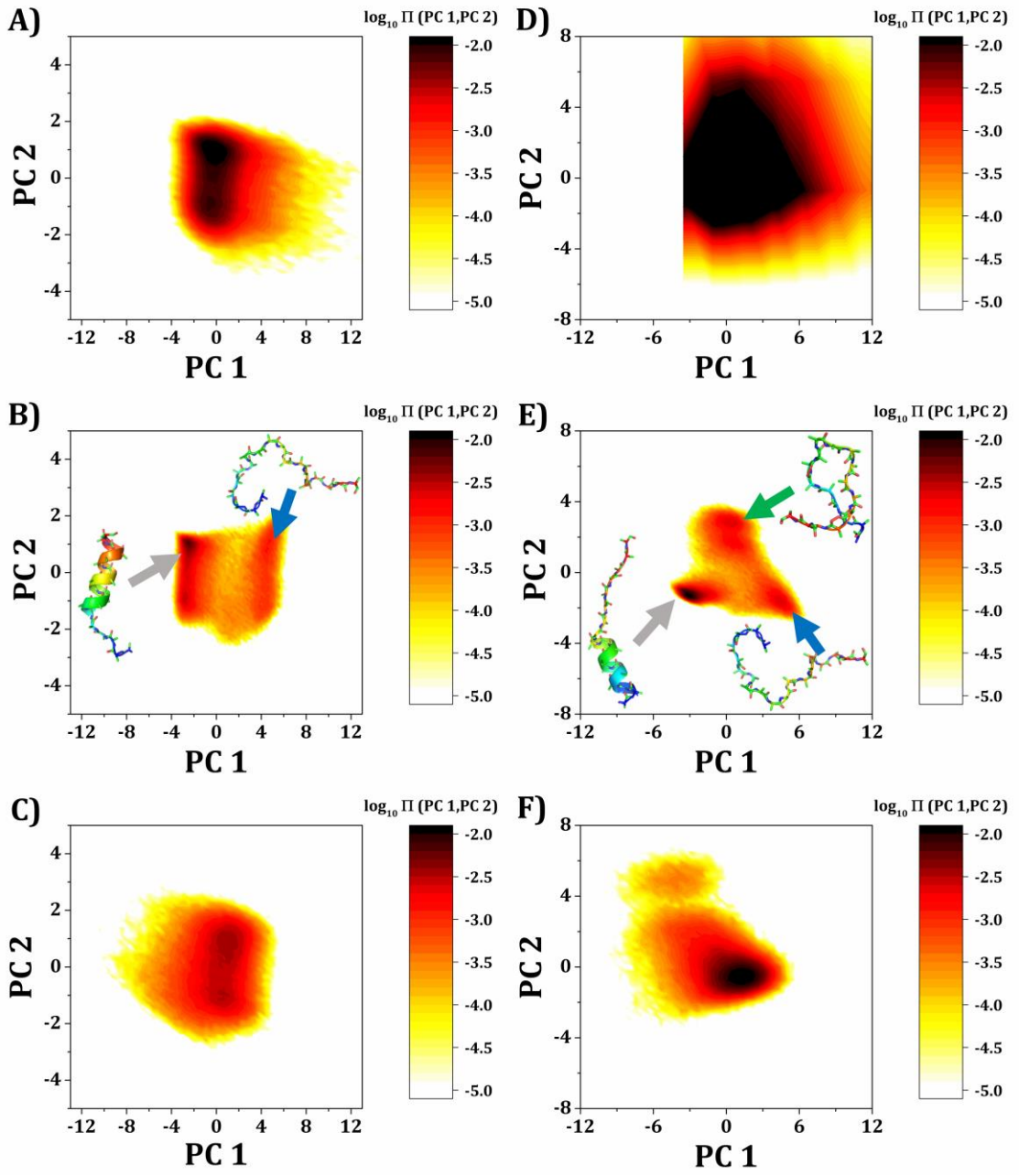

Figure 6.6. Surface plots based on PCA of AQEK at -68 °C (A), 6 °C (B) and 117 °C (C), compared to FAQEK at -53 °C (D), 24 °C (E) and 127 °C (F) with snapshot structural cartoons around the T_m for better readability. For reference, the T_m values for AQEK and FAQEK are 7 °C and 21 °C respectively. Surface plots represent the normalized histogram for the probability of observing values of PC 1 and PC 2 in a log scale.

Structures obtained from the MD simulation show that the intermediate states are represented by a loop formed around the Lys residue for FAQEK during unfolding, leading to a configuration resembling a small molten globule at temperatures near the T_m (see Figures 6.6E and 6.7). These configurations are promoted by the lower energy that both Phe residues allow the peptide to obtain during collapse. This collapse causes both Glu residues to come closer than in the folded and fully-unfolded, expanded states, which is responsible for the maximum observed in Figure 6.5D for the red dashed curve. Additionally, two intermediate configurations were observed: where the Phe residues interact with one another and where Phe residues interact with any nearby Ala residue. Those intermediate configurations are responsible for the decrease in the entropy of unfolding by not allowing the peptide to fully explore other unfolded configurations. These findings agree with experimentally observed events of protein unfolding in other systems, where molten globules are usually observed before the protein fully unfolds or refolds [240,244]. The inclusion of more hydrophobic residues caused an increase in T_m by limiting the number of configurations that the unfolded state could populate, rather than simply lowering the energy of the folded state. While some of the loss of entropy in the unfolded state is compensated by the favorable (lower energy) interactions within the unfolded state, the net result is that the unfolded state(s) are destabilized enough compared to folded states that a higher T is needed to achieve complete unfolding in the simulations for the FAQEK, compared to the AQEK.

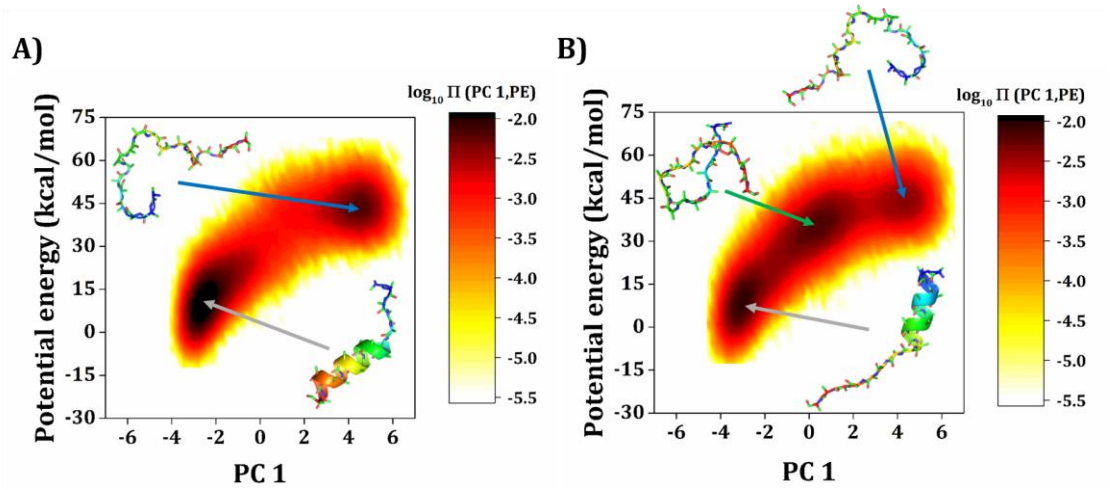

Figure 6.7. Further PCA analysis of the potential energy with snapshot structures of the AQEK at 6 °C (A) and the FAQEK at 24 °C (**B**). Higher (lower) energy states correspond to unfolded (folded) configurations.

Turning to the series of peptides with common sequence and different lengths, PCA was also carried out for the AQK18, AQK27 and AQK35 simulations as a function of temperature. AQK18 shows a clear two-state transition like that observed for the AQEK (Figures 6.6-6.8). Using the T_m of AQEK as a reference, Figure 6.5 and the discussion above indicate that the higher T_m for the AQK18 results from: (i) decreased enthalpy of unfolding that results from the breaking of fewer hydrogen bonds as well as from higher hydrophobicity in the sequence (Figures 6.4C, 6.5B-C), (ii) a significantly reduced entropy of unfolding due to the shorter sequence (fewer configurations), and (iii) less extended structures in the unfolded state due to the sequence higher hydrophobicity in comparison with the AQEK.

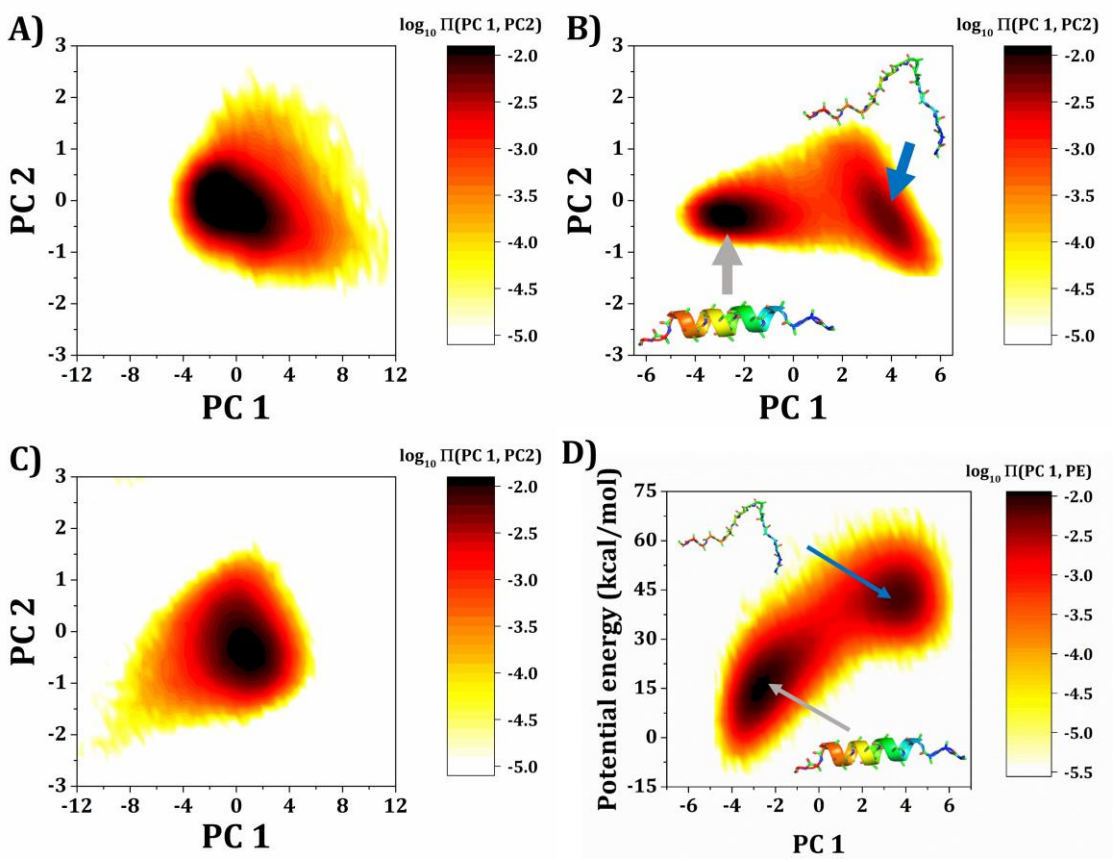

Figure 6.8. PCA of the AQK18 at -69 °C (A), 18 °C (B & D) and 112 °C (C) with snapshot structural cartoons around the T_m (=18 °C). Higher (lower) energy states correspond to unfolded (folded) configurations.

For the AQK27, there is a more apparent two-state transition like that observed for AQEK (Figures 6.9-6.10). However, a reasonably well populated "intermediate" was observed. This arises from the higher likelihood of collapse of the longer chain in the unfolded or partly unfolded state(s), in comparison to the shorter sequences, as well as the addition of a charge residue (Lys) in the center of the sequence. This also arises from an increased number of stabilizing contacts in (partly) collapsed states for the longer sequence. The partly collapsed configurations resemble those in the unfolding

intermediates observed in analysis of the FAQEK behavior, although the intermediates for the AQK27 are less stable due to the weaker hydrophobic interactions (Figure 6.9).

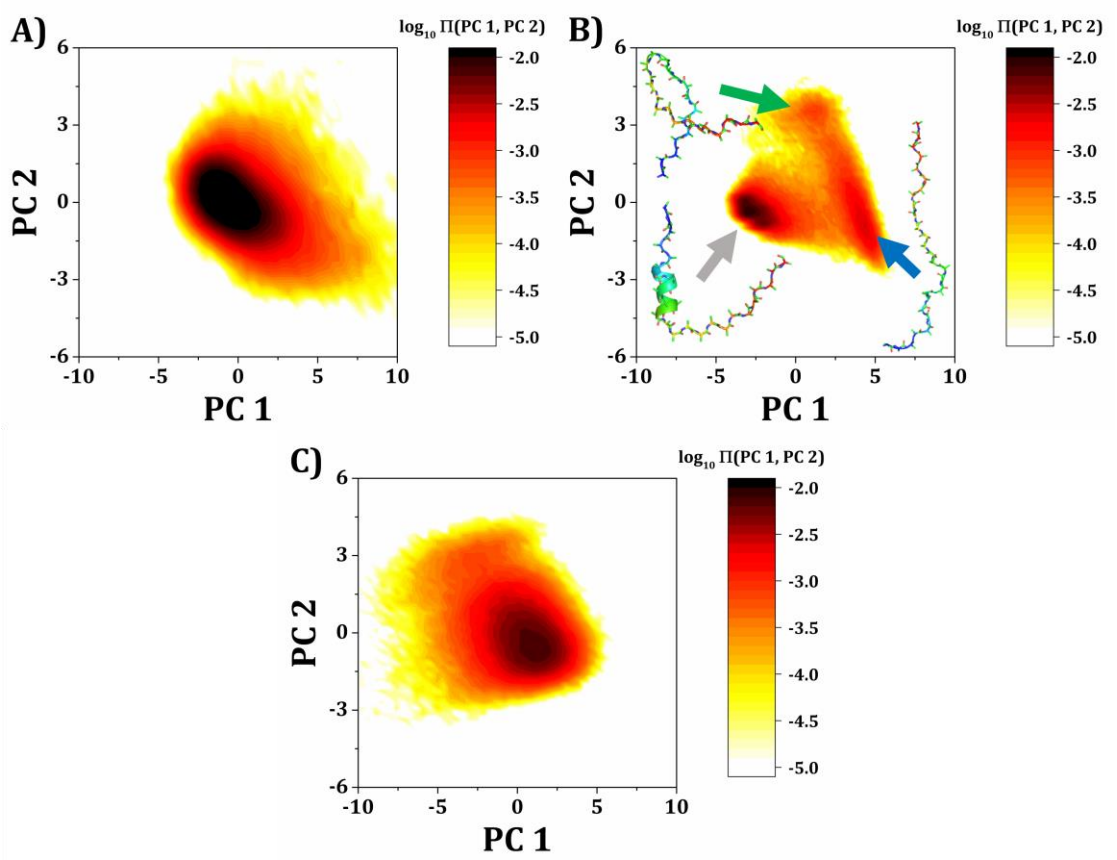

Figure 6.9. PCA of the AQK27 at -48 °C (A), 30 °C (B) and 102 °C (C) with snapshot structural cartoons around the T_m (=32 °C).

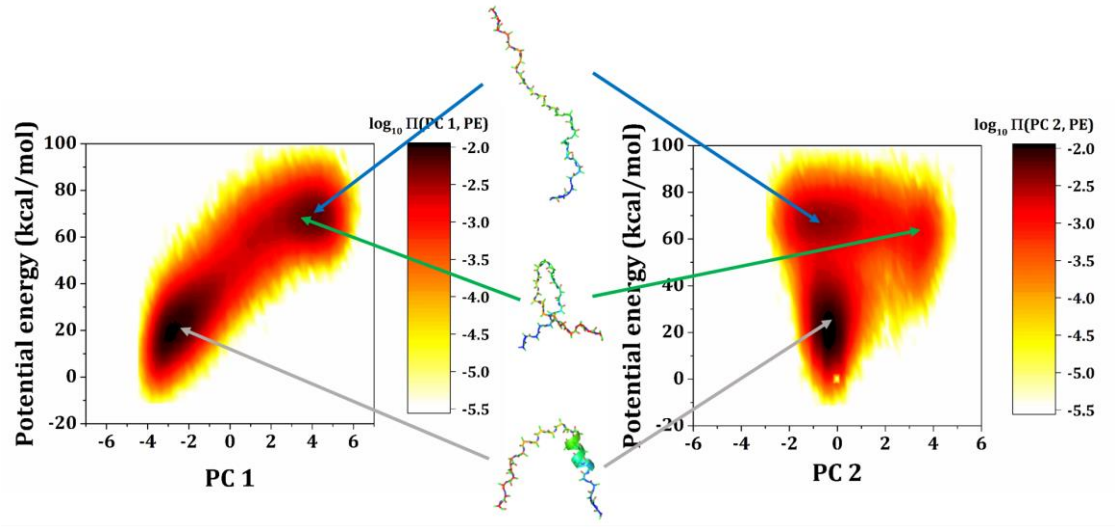

Figure 6.10. PCA with snapshot structures of the AQK27 sequence at 30 °C. Higher (lower) energy states correspond to unfolded (folded) configurations. Energy states in between correspond to stable intermediates.

Similarly, the AQK35 sequence shows a series of intermediate states resulting from chain collapse (Figures 6.11-6.12). These observations are consistent with other experimental and computational results in which molten globules have been observed during refolding [65,244]. In this context, the AQEK sequence is effectively an outlier in Figure 6.4A, as this peptide does not show such behavior and its unfolding represents an idealized two-state transition.

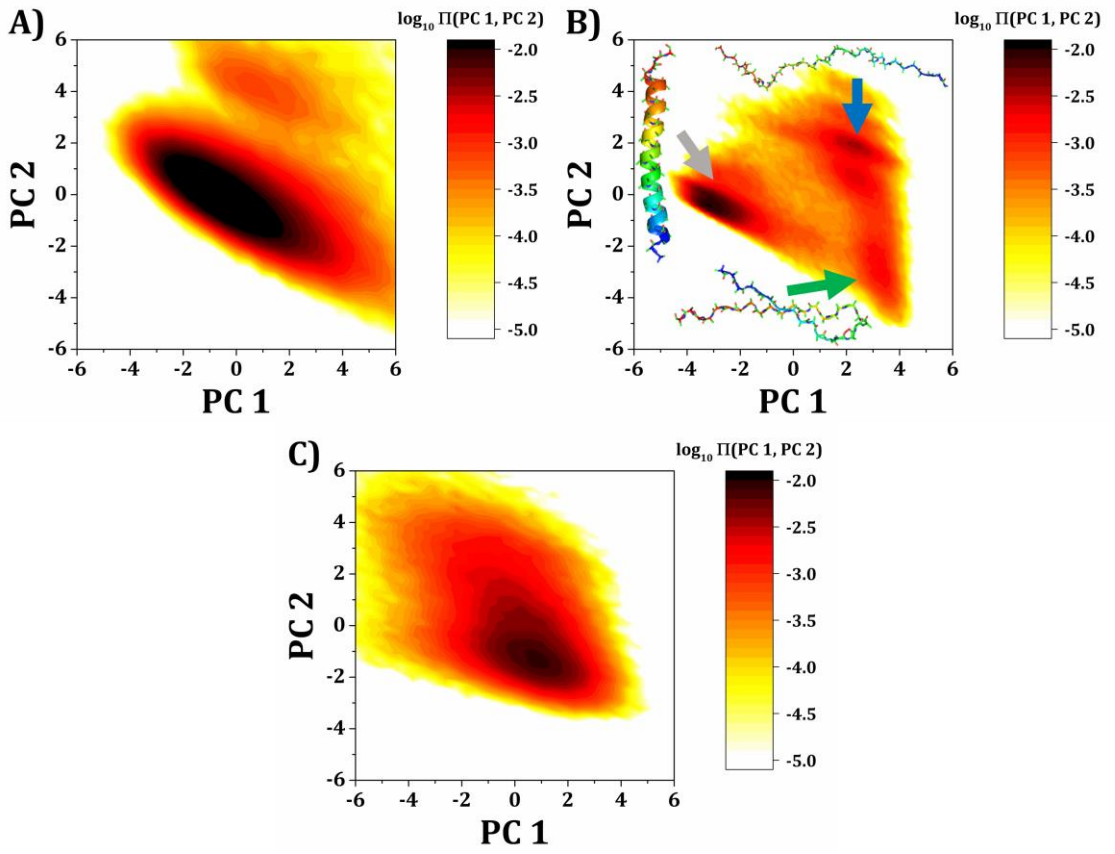

Figure 6.11. PCA of the AQK35 at -28 °C (A), 42 °C (B) and 107 °C (C) with snapshot structural cartoons around the T_m (= 45 °C).

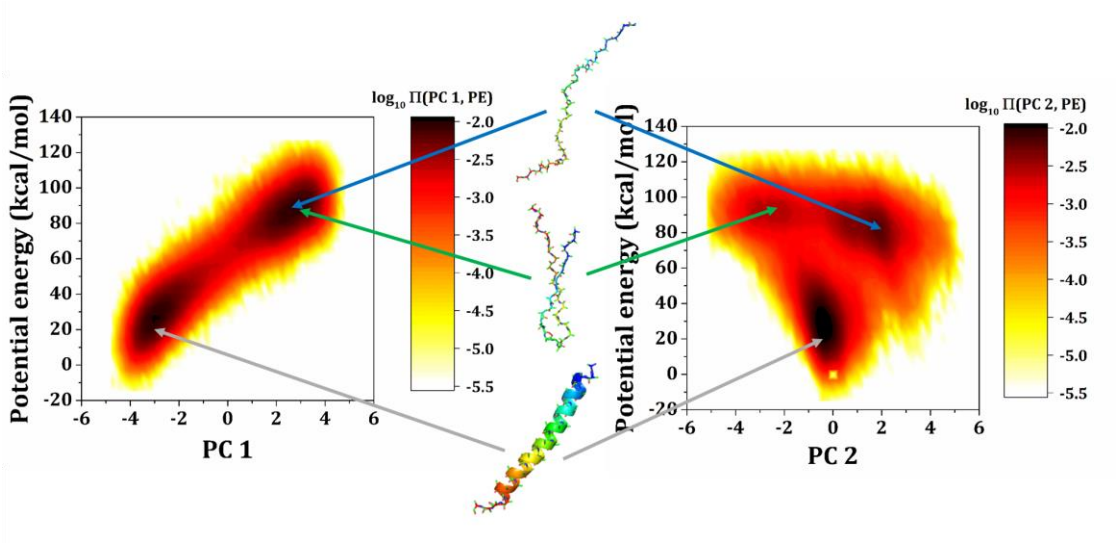

Figure 6.12. PCA with snapshot structures of the AQK35 sequence at 42 °C. Higher (lower) energy states correspond to unfolded (folded) configurations. Energy states in between correspond to stable intermediates.

6.7 Experimental Measurements of Peptide Unfolding and Validation of Predicted Behavior

To experimentally test the simulation results presented and discussed above, the same set of Ala-rich peptides in Table 6.1 were experimentally characterized. The experimental data shown in this section have been adapted from reference [229]. The helical content and unfolding of each of the peptides were characterized *via* CD spectroscopy, heating the peptide solutions from 0 °C to 80 °C at a concentration of 0.125 g/L. Representative full wavelength spectra for the peptides are shown in Figures 6.13-6.15. At low *T*, all peptide sequences showed spectra with characteristic α-helical features, with the minima at 208 nm and 222 nm. For sequences AQK18, AQK27, AQK35 a clear isodichroic point was observed, indicating a two-state transition from α-helix to random coil (Figures 6.14-6.15). Alternatively, an isodichroic point was not clearly observed for the AQEK and FAQEK sequences, signifying the presence of

unordered states as the peptides unfold, in partial agreement with the discussion above (Figure 6.13). The $[\Theta]_{MRE}$ value at 222 nm at 0° C was used to calculate the fractional helicity using a previously reported method that is based on idealized, long helices (Table 6.3) [245].

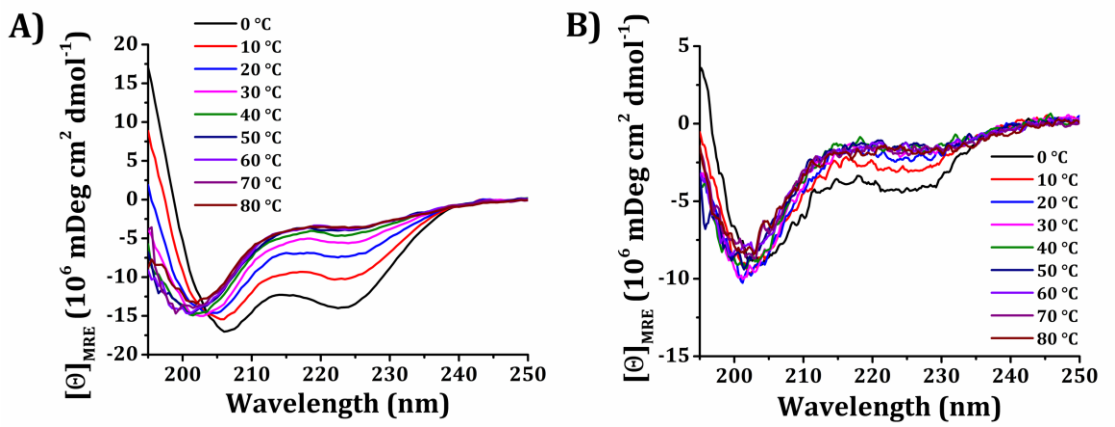

Figure 6.13. Full wavelength spectra of AQEK (A) and FAQEK (B) during heating at 0.125 g/L in 10 mM phosphate buffer (pH 7.4). Peptide samples were heated from 0 °C to 80 °C at 10 °C increments.

Table 6.3. Mean-residue ellipticity ($[\Theta]_{MRE, 222}$) and percent helicity (%-helicity) of the five studied sequences obtained from CD measurements at 0 °C.

Sequence	$[\Theta]_{MRE, 222}$	%-helicity
AAQEAAAAQKAAAAQEAAA (AQEK)	-12412	23
AAQEFAAAQKAAAFQEAAA (FAQEK)	-4462	8
$K(A_3Q)_4K$ (AQK18)	-19198	37
$K(A_3Q)_3K(A_3Q)_3K$ (AQK27)	-23561	43
$K(A_3Q)_4K(A_3Q)_4K$ (AQK35)	-27625	49

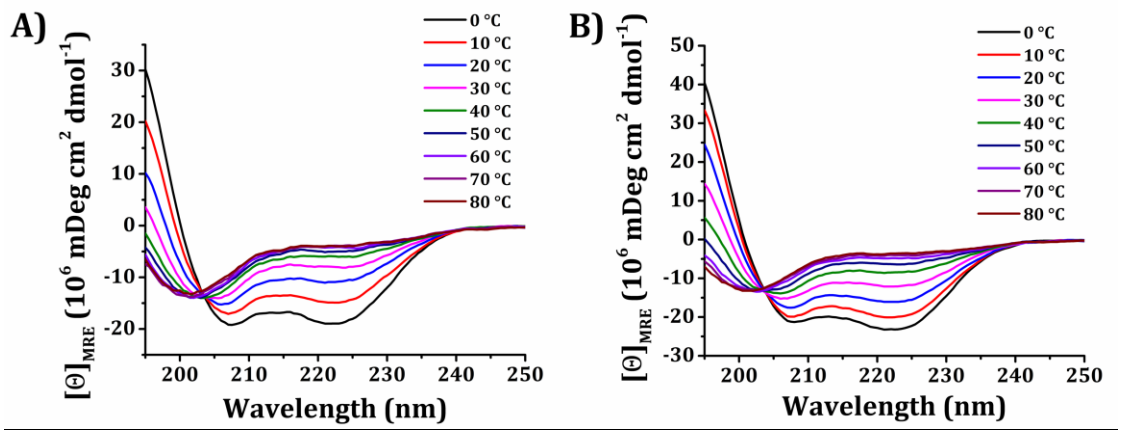

Figure 6.14. Full wavelength spectra of the AQK18 (**A**) and AQK27 (**B**) during heating at 0.125 g/L in 10 mM phosphate buffer (pH 7.4). Peptide samples were heated from 0 °C to 80 °C at 10 °C increments.

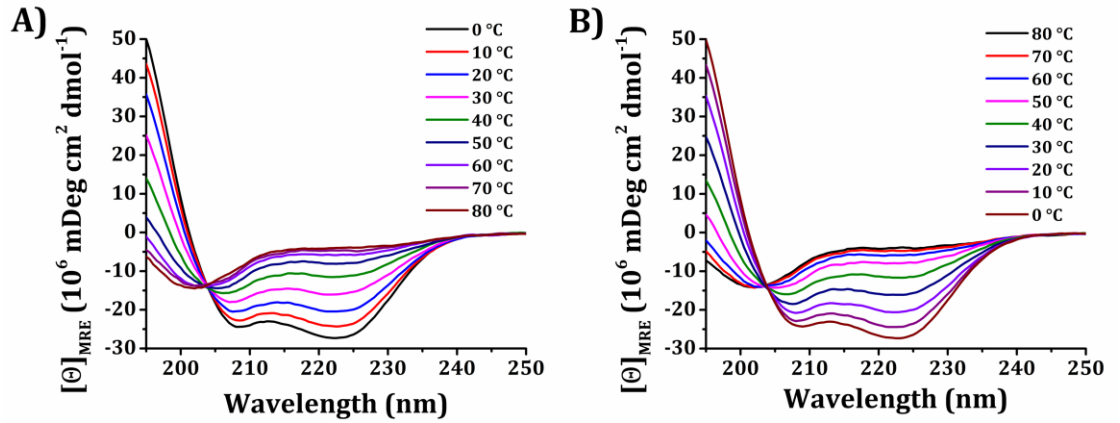

Figure 6.15. Full wavelength spectra of the AQK35 during heating (**A**) and subsequent cooling (**B**) at 0.125 g/L in 10 mM phosphate buffer (pH 7.4). Peptide samples were heated from 0 °C to 80 °C, and cooled back to 0 °C at 10 °C increments.

Samples were estimated to have average helical contents at 0 °C, relative to an ideal helix, of 12%, 23%, 35%, 40%, and 48% for FAQEK, AQEK, AQK18, AQK27 and AQK35, respectively. While the helical content increased reliably with peptide

length in both the simulations (above) and experiments, the experimental helical content values are not equivalent to the simulated average helix contents owing to necessary differences in how the values are determined. Helical content determined experimentally from the CD data is based on comparison to a hypothetical perfect helix, while simulated average helix contents are based on measured torsion angles and hydrogen bonds within the simulation. The helicity estimates from CD, however, are in agreement with values reported previously for other short, Ala-rich peptides of similar length [237,239].

The FAQEK sequence showed a lower percent helicity in comparison to AQEK and AQK18. This differs from the simulated peptide model, where the lowest helical content was observed for the AQEK sequence. A previous experimental study reported a similar trend, where inclusion of Phe residues in a short, Ala-rich peptide was observed to lower the helical content of a peptide compared to identical sequences lacking a Phe residue [246]. This behavior is not observed in the single-peptide simulations, suggesting that inter-peptide interactions can play a role in the stability and unfolding of the FAQEK and other sequences with highly hydrophobic residues. For the other two peptides (AQEK and AQK18), the differences observed in the simulations were borne out experimentally, with AQEK having lower percent helicity than AQK18 despite their similar lengths. As discussed above, this is caused by the stronger hydrophobic contacts within the AQK18 resulting in lower enthalpies and entropy of unfolding. Since the entropy decreases more than the enthalpy, the Gibbs energy of unfolding increases in comparison to the AQEK sequence.

The stability of each of these peptides was also characterized experimentally to assess whether they are prone to aggregation and β-sheet formation. Peptide solutions

were heated over a series of temperature values, and absolute intensity values of the $[\Theta]_{MRE}$ value at 222 nm and 208 nm decreased sigmoidally with increasing *T*, indicating unfolding of the peptide (Figure 6.16). $[\Theta]_{MRE,222}$ values were monitored upon heating from 0 °C to 80 °C, as well as upon subsequent cooling back to 0 °C, to analyze the reversibility of unfolding. At pH 7.4, all peptides recovered their original spectra and $[\Theta]_{MRE,222}$ upon cooling to 0 °C, indicating the conformational transitions are reasonably reversible on the timescales of the measurements. Quantitatively accurate midpoint unfolding temperatures could not be reliably measured, as the pre-transition baseline for each peptide solution was not accessible at temperatures above freezing. This indicated that at least a fraction of the peptides is significantly unfolded at 0 °C, and the values of T_m from these experiments are thus only treated as "apparent T_m" values. That notwithstanding, the inflection points in Figure 6.16 show a shift towards higher temperatures with increasing peptide chain length, which is also observed in the simulations. Such an increase in the experimental apparent T_m was also observed previously for $(AEAAKA)_n$ sequences, which demonstrated reversible unfolding, consistent with the results from the simulations here [237,239].

The unfolding curves of AQEK, FAQEK, and AQK18 were similar to the unfolding curves of sequences with shorter chain lengths (14 and 20 residues), in that they exhibited only a portion of the transition region, no flat/linear pre-transition region above 0 °C, and showed a clear post-transition region at higher temperatures. AQK27 and AQK35 also had similar unfolding curves compared to the longer $(AEAAKA)_n$ repeats (26, 32, 38, and 50 residues), with only a portion of the pre-transition region (upper baseline) being observable. Together, the trends obtained from the experimental

characterization of AQEK, AQK18, AQK27 and AQK35 agree with the results obtained from the molecular simulations.

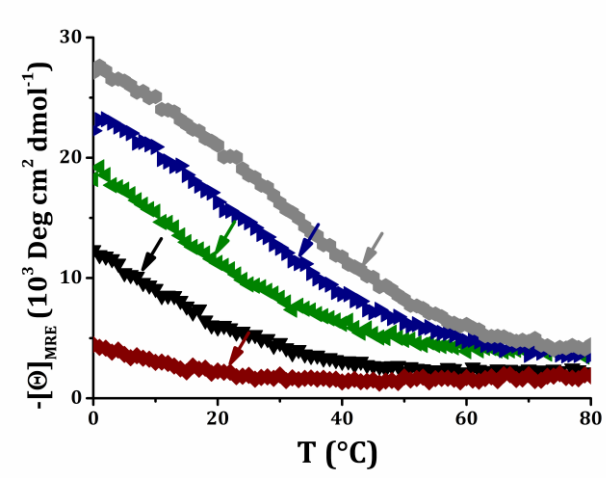

Figure 6.16. Full melting curves at 0.125 g/L in 10 mM phosphate buffer. $[\Theta]_{MRE}$ values at 222 nm observed while samples were heated from 0 °C to 80 °C. Symbols represent the AQEK (black), FAQEK (red), AQK18 (green), AQK27 (blue) and AQK35 (grey) while colored arrows point to the simulated T_m values from Figure 6.4A.

6.8 Peptide-Peptide Interactions Mediating Unfolding Behavior

However, a conspicuous discrepancy was apparent in comparison to the results from simulations summarized above for the FAQEK sequence: the experimental T_m value is lower than those of any of the other sequences. This discrepancy was hypothesized to be a result of inter-peptide interactions in the experimentally probed unfolding process, which can cause a decrease in apparent T_m values when the unfolded peptides interact and aggregate. Therefore, preliminary simulations were performed where two peptides were present in the simulation box, and the box size was selected to provide an effective peptide concentration of 1 mM. These simulations were used as an

initial approach to gain insights into the discrepancies between observed experiments and single-peptide simulations. Average helix content and heat capacity values were calculated as a function of temperature in a manner analogous to the single-peptide and are presented in Figure 6.17.

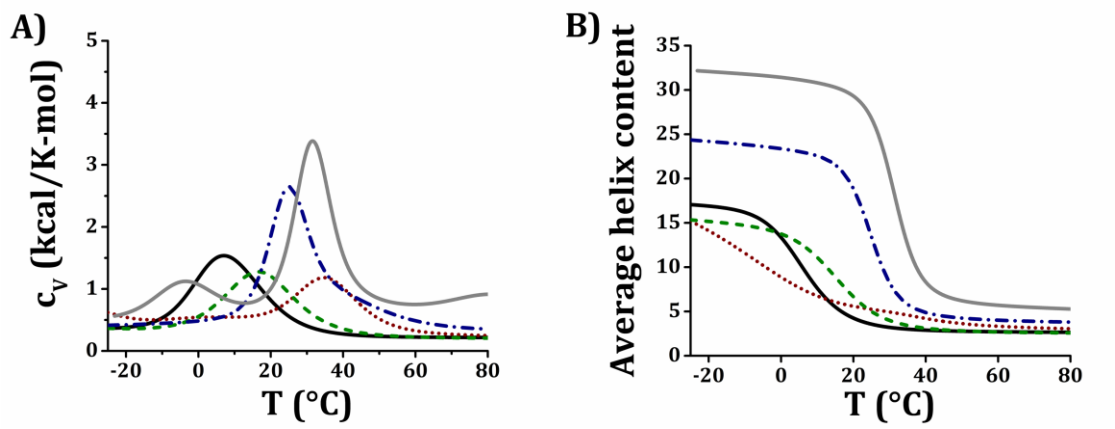

Figure 6.17. Heat capacity (panel A) and average helix content (panel B) as a function of temperature for two-peptide simulations. Line types are as in Figure 6.3.

Comparison of the results in Figure 6.17 to those in Figure 6.3 clearly show a shift in T_m for the FAQEK in the two-peptide simulation, while the T_m values for the other four sequences are not significantly affected by the presence of a second peptide. The results are in better agreement with experimental CD profiles (Figure 6.16) and quantitative helical content values discussed above, and support the conclusion that the experimental unfolding behavior of the FAQEK peptides is considerably affected by inter-peptide interactions that cause non-ideal behaviors not considered in the idealized data analysis of DSC and CD experiments.

Finally, interesting behaviors were observed during the two-peptide simulations, which is reflected in the $c_v(T)$ plots (Figure 6.17A). Additional peaks and poorly defined unfolded baselines at higher temperatures were obtained for some sequences, in contrast to the single-peptide results (Figure 6.3A). For all the sequences, it is expected that T_m values from $c_v(T)$ profiles correspond to almost 50% change in the average helix content. However, that value was observed below 0 °C for the FAQEK sequence. Consequently, the peak observed at 35 °C in Figure 6.17A does not correspond to the unfolding and T_m of individual FAQEK molecules. The AQK35 sequence shows a similar peak above 80 °C, and the AQK27 sequence shows a poorly defined baseline at temperatures just above the T_m. Inspection of representative configurations from the simulations as a function of T indicates that these secondary peaks represent the breakage of weak peptide-peptide complexes that had formed during unfolding, and these resemble the initial steps of nucleation of peptide aggregation [247].

To further elucidate the previous observations, experimental samples were made at 1 mM and 0.1 mM peptide at pH 7.4. Samples were incubated at 60 °C for one week and aggregation was qualitatively examined by solution turbidity. Initial experimental results showed aggregates in FAQEK and AQK35 samples, which partially validates this hypothesis and suggests that inter-peptide interaction and self-association of the FAQEK, AQK27 and AQK35 sequences may considerably affect the unfolding characteristics of these peptides. This highlights that when modeling idealized unfolding transitions, the models should be complemented with simulations that permit inter-peptide (or inter-protein) interactions. However, given that additional experimental data with higher-level structural resolution are not yet available for the current systems to validate the simulations, it seems unreasonable to extend the interpretation of the multi-

peptide effects on unfolding thermodynamics beyond these qualitative conclusions. Additional implications will be discussed in Chapter 7.

Finally, it should be noted that most of the simulations were carried out before the experimental data on the peptides in Table 6.1 were available, and that the tuned EBD 4bAA model was tuned against a different set of peptides (Table 6.2) to allow it to yield predictions that were qualitatively and quantitatively similar to the experimental results for the peptides in Table 6.1. This suggests that the tuned EBD 4bAA model is not limited to the helical peptides here, and can be used as an effective tool to computationally screen peptide sequences and unfolding thermodynamics for future applications regarding peptide and protein stability, as well as inter-peptide interactions.

6.9 Summary and Conclusions

An implicit-solvent CG molecular model was successfully tuned to capture unfolding thermodynamics of a series of Ala-rich peptides. This model was based on a former model and refined for a set of published peptide sequences. It was further used to provide insight into the unfolding events of a series of new Ala-rich peptides. Initial single-peptide simulations (idealized dilute limit) for AQEK and FAQEK (Table 6.1) revealed that the inclusion of Phe residues disrupted the idealized two-state transition exhibited by the AQEK sequence, allowing the formation of stable intermediate states resulting in a decreased entropy of unfolding and a higher T_m. Simulations for the AQK18, AQK27 and AQK35 sequences showed that increases in chain-length have a significant impact on the enthalpy of unfolding by increases in hydrogen bonding, which leads to increases in T_m as chain-length increases as long as the basic chemistry is held constant. Additionally, sequences with charged residues in the middle of the chain showed higher likelihood of unstable intermediate states during unfolding, promoting

intermediate collapsed states. CD experiments later showed good agreement between simulated and experimental apparent T_m values for four of the five sequences that were tested. The FAQEK sequence showed deviations that were hypothesized to be attributed to the presence of aggregates within the experimental solutions, suggesting the incorporation of inter-peptide interactions in the analysis of the unfolding behavior. This was corroborated qualitatively by two-peptide simulations that showed that the interaction between peptides can cause a dramatic change in T_m, and this may be responsible for the low apparent T_m values observed for the FAQEK sequence. This supports the initial hypothesis of aggregation behavior affecting the unfolding thermodynamics of this sequence and encourages the development of computational tools that incorporate both unfolding and aggregation for future work.

Chapter 7

SUMMARY AND FUTURE WORK

7.1 Summary

This work focused on exploring the viability of combining experimental training sets and CG molecular modeling to predict protein and peptide interactions and unfolding from low to high protein/peptide concentrations and as a function of solution conditions, such as pH, TIS and sucrose concentrations. This dissertation studied the effects of molecular shape in packing at high protein concentrations for mAb solutions, explored the viability of predicting high-concentration weak protein-protein interactions for mAb and globular protein solutions from low-concentration measurements coupled with CG molecular modeling, reexamined and extended the previous framework to estimate protein-cosolute and protein-water interactions for multicomponent solutions *via* high-precision density measurements, and demonstrated the potential of simplified CG models to predict unfolding thermodynamics of short polypeptide sequences. A combination of experimental techniques, such as SLS and CD, and computational algorithms, such as MSOS and TMMC, were utilized in the realization of this dissertation. The approaches tested in this dissertation were applied to different protein/peptide solutions, ranging from short peptides (~2 kDa) to mAbs (~150 kDa), proving the flexibility of the tools to be extended to any protein solution of interest.

Several CG models with varying molecular detail were used to evaluate the effect of molecular shape on protein molecular volume and packing behavior. In the case of mAb solutions, the canonical spherical model was found to overestimate the

molecular volume (*via* $B_{12,ST}$ simulations) in comparison to atomistic structures, even when the excluded volume ($B_{22,ST}$) contributions were matched at low-c_2 conditions. This would limit the accuracy of spherical models to quantitatively predict high-c_2 mAb behavior. Computer simulations at higher c_2 demonstrated that spherical models lack the physical packing behavior that arises from the anisotropic structure proper of mAbs. This can be extended to other elongated or non-globular proteins. For mAb solutions, a practical balance between accuracy and computational time was considered when selecting a proper CG model. Only high-resolution models (1bAA and 4bAA) were found to mimic the atomistic behavior, but their computational time would be intractable at high c_2. On the other hand, low resolution models (*e.g.*, spherical models) were found to underestimate protein packing despite their fast computation. Consequently, models such as the HEXA (6 beads per protein) and the DODECA (12 beads per protein) models were found to provide acceptable accuracy in comparison to the atomistic behavior while retaining tractable computational times for high-c_2 simulations.

The effects of the flexibility of the hinge region on mAbs were considered at high c_2. Simulations performed for flexible *vs* rigid molecules showed that the flexibility of the hinge region does not affect the high-c_2 behavior (in terms of the osmotic compressibility) below ~140 g/L, thus rigid models are useful without causing significant additional uncertainty in those conditions for osmotic compressibility simulations. However, it could be expected that a flexible hinge would provide additional short-ranged configurations that a rigid model would not capture, so it would need to be considered for comparing against experimental measurements sensitive to these molecular scale events (*e.g.*, SANS and SAXS).

Additional interactions beyond sterics were considered, including short-ranged hydrophobic and van der Waals attractions *via* a modified Lennard-Jones potential, and change-charge interactions *via* a modified Yukawa potential. It was found that short-ranged non-electrostatic attractions primarily affected the solution behavior when charges are screened (*i.e.*, high-*TIS* conditions), while electrostatic interactions are most relevant at low *TIS*, both as expected from the Debye-Hückel theory. Analysis of the effect of the charge distribution on low-c_2 interactions *via* B_{22} maps showed that the presence of highly anisotropic charge distributions leads to unphysically negative B_{22} values, while theoretical charge distributions from the primary sequence and crystal structures result in highly unphysical protein-protein interactions at both low- and high-c_2 conditions. This can have practical implications in the design of mAb sequences when considering the colloidal stability of the molecule. Finally, high-c_2 simulations showed that the level of structural coarse-graining becomes most relevant as interactions move from strongly repulsive to strongly attractive interactions. Combined with the trade-off between structural accuracy and computational burden, this highlights a balance that must be considered when designing CG molecular models for different applications.

Static light scattering experiments were performed to quantify "weak" protein-protein interactions of IgG1, IgG4 and aCgn protein solutions as a function of c_2, pH, *TIS* and sucrose concentration. These included conditions that resulted in both net-repulsive (*e.g.*, IgG1 and aCgn at pH 5) and net-attractive protein interactions (*e.g.*, aCgn at pH 7 and IgG4 at pH 6.5) at low *TIS*, and at low- to high-c_2 conditions. Three low-resolution CG models were used to evaluate the potential to predict excess Rayleigh profiles and zero-q structure factors at high c_2 based on low-c_2 training sets. Two models (the HEXA and DODECA) were considered for the mAb proteins and one model (a

spherical model) for all aCgn solutions. Contributions from sterics, short-ranged non-electrostatic attractions and electrostatic interactions were included for each of the studied models.

For the mAb solutions, B_{22} (low-c_2) results showed that the IgG1 and IgG4 exhibit net-repulsive behavior at low *TIS* and pH 5, which transitions to net-attractive behavior as *TIS* increases. This suggested strong charge-charge repulsions at pH 5 for both protein molecules. At pH 6.5, these antibodies showed net-attractive behavior from low to high *TIS*. For the IgG1 molecule, these resulted from lower effective charge values resulting in weaker electrostatic repulsions and weakly net-attractive conditions at low *TIS* (*e.g.*, $B_{22}/B_{22,ST} \sim 1$). Conversely, the IgG4 showed strong electrostatic attraction resulting from a disparity in charges between the Fab and Fc regions as low-*TIS* conditions resulted in more attractive behavior than at high *TIS*. For both molecules at all measured pH and *TIS* conditions, the addition of 5% w/w sucrose to the protein solutions induced weaker (stronger) net-attractions (net-repulsions) with increasing *TIS*. In terms of model predictions from low to high c_2, the quantitative differences were not statistically significant at pH 5. However, the models were able to predict the behavior for the IgG1 at pH 6.5 (weak net-attraction) while failed to quantitatively capture the behavior for the IgG4 at the same pH (strong net-attractions). Therefore, both models could be used to accurately predict high-c_2 interactions based solely on low-c_2 experimental data and structural information under net-repulsive to slightly net-attractive conditions depending on the requirements of the user (*e.g.*, computational burden and molecular features) for strongly repulsive to weakly attractive conditions.

For aCgn solutions, a canonical spherical model was used to capture both low- and high-c_2 weak protein-protein interactions. Experimental results showed that aCgn

behaves qualitatively similarly to both mAb molecules at pH 5 (*e.g.*, it transitions from net-repulsive to net-attractive behavior as *TIS* increases at pH 5), and it resembles the IgG4, pH 6.5 behavior at pH 7 (*e.g.*, it transitions from strongly net-attractive to weakly net-attractive behavior as *TIS* increases at pH 7). The simulations showed that canonical spherical models are capable of quantitatively capturing the data if a combination of screened monopole, screened dipole model, short-ranged non-electrostatic attractions, and steric repulsions is used. Additionally, low-c_2 model parameters were quantitatively predictive of the interactions at high c_2 if the net interactions are repulsive or slightly attractive compared to steric-only interactions, in agreement with the results for the mAb molecules. However, for strongly attractive conditions, where the effect of charge anisotropy is dominant, the spherical CG models were only able to qualitatively or semi-quantitatively predict the high-c_2 behavior.

For all protein solutions, a high-resolution 1bAA CG model was used to identify if strong net-attractions could be caused by an anisotropic charge distribution for both the IgG4 at pH 6.5 and aCgn at pH 7. It was found that this CG model was accurate on predicting the strong electrostatic attractions exhibited by the IgG4 molecule while provided insights in the differences between pH 5 and pH 7 for aCgn. However, challenges still exist for strongly attractive electrostatic conditions at *TIS* < 50 mM as predictions suggest net-repulsion for aCgn at pH 7, which was not observed experimentally.

To complement the results obtained in Chapters 3 and 4, a generalized expression to obtain preferential interactions for multicomponent solutions was derived in Chapter 5. This new expression was applied to protein solutions at infinite dilution ($c_2 \rightarrow 0$) and used to evaluate ternary (water-protein-buffer/cosolute) and quaternary

(water-protein-cosolute-buffer) solutions for aCgn and the IgG1. For aCgn at pH 7 and 25 °C, sodium phosphate as an osmolyte showed no significant preferential interactions below 30 mM and dilute protein conditions. Conversely, sodium chloride, sucrose, and trehalose showed preferential accumulation under dilute protein conditions in both ternary (no buffer) and quaternary (5 mM sodium phosphate buffer) solutions. By comparing the results for ternary and quaternary solutions, it was found that the presence of 5 mM phosphate buffer in the solutions does not alter the preferential interactions between water, protein and the added cosolute, in good agreement with the derived expression for multicomponent solutions and the measured aCgn-buffer interactions.

For the IgG1 molecule at both pH 5 and 6.5 and 25 °C, experimental protein-sucrose interactions suggest preferential accumulation (solvation) by sucrose around the protein surface. By combining the results from SLS measurements and computer simulations in Chapter 3, it was concluded that sucrose acts as a coating agent around the protein surface decreasing the effective hydration of the protein, and this led to stronger (weaker) repulsions (attractions) between protein molecules observed at both low and high c_2.

Additional molecular simulations of infinite-dilution steric interactions between aCgn and the IgG1 molecule with a water-sized or sugar-sized species highlighted the presence of strong attraction between water or cosolutes and the protein molecules. This also highlighted the misuse of $\hat{V}_2$ as a direct estimate of molecular volume for protein solution modeling and simulations. The results in Chapter 5 highlight the need to evaluate protein-cosolute interactions on an individual basis as observed results might not apply to other protein solutions. Fortunately, the experimental and computational

frameworks developed in this thesis can be easily applied to any protein solutions of interest.

Finally, a 4bAA CG molecular model was successfully tuned to capture unfolding thermodynamics of a series of Ala-rich peptides. Initial single-peptide simulations for AQEK and FAQEK sequences revealed that the inclusion of Phe residues disrupted the idealized two-state transition exhibited by the AQEK sequence, allowing the formation of stable intermediate states resulting in a decreased entropy of unfolding and a higher T_m. Simulations for the AQK18, AQK27 and AQK35 sequences showed that increases in chain-length have a significant impact in the enthalpy of unfolding by increases in hydrogen bonding energies, which leads to increases in T_m as chain-length increases as long as the basic chemistry is held constant. Experimental characterization of the five peptide sequences showed good agreement between simulated and experimental apparent T_m values for four of the five sequences that were tested. The FAQEK sequence showed deviations that were hypothesized to be attributed to the presence of aggregates within the experimental solutions, suggesting the incorporation of inter-peptide interactions in the analysis of the unfolding behavior. This was partially corroborated qualitatively by two-peptide simulations that showed that the interaction between peptides can cause a dramatic change in T_m, and this may be responsible for the low apparent T_m values observed for the FAQEK sequence.

7.2 Future Work

This dissertation provides initial frameworks to predict or measure protein-protein interactions from low to high c_2, and protein-cosolute interactions and unfolding transitions under dilute protein conditions. The methodologies provided in this

document act as a starting point for many future applications of protein solution behavior, which are discussed below.

7.2.1 *q*-Dependent Structure Factors: The Effect of a Flexible Hinge Region for Highly Attractive Conditions

The analyses in this thesis only focused on the zero-*q* limit of the structure factor measured *via* static light scattering. The zero-*q* limit of the structure factor is expected to represent the behavior of a point particle in solution. Thus, unless strong attractions strongly dependent on protein-protein orientation and proximity are present, the results in Chapter 2 (*e.g.*, no effects from the flexibility of the hinge) should stand. However, this might not be the case for the *q*-dependent structure factors as measured using small angle X-ray and neutron scattering techniques (SAXS and SANS, respectively), which have been shown to be sensitive to the flexible regions of mAbs and other proteins.

Examples of SAXS and SANS measurements for the IgG1 molecule are shown in Figures 7.1 and 7.2 and show differences as a function of protein concentration and solution formulation beyond the zero-*q* limit. Simulations that self-consistently capture the zero-*q* limit (as in Chapters 3-4) as well as the *q*-dependent structure factors (as those in Figures 7.1-7.2) could further provide additional information regarding the protein flexibility and domain-domain interactions as was preliminary shown in Chapter 3. However, further refinements in current algorithms would be required to better optimize the sampling of flexible molecules as adding the flexibility will expand the energy space of the system, adding computational challenges that need to be overcome to achieve convergence within practical computational time frames. Additionally, the MSOS algorithm used in Chapters 2-4 was never designed to be used with flexible molecules, so additional time would need to be invested in developing or reexamining

frameworks that incorporate highly flexible region of proteins into the simulations. These combined with SAXS/SANS measurements would provide more comprehensive molecular models that can be potentially predictive of protein-protein interactions from low to high c_2 and the dominant contributions arising from each domain.

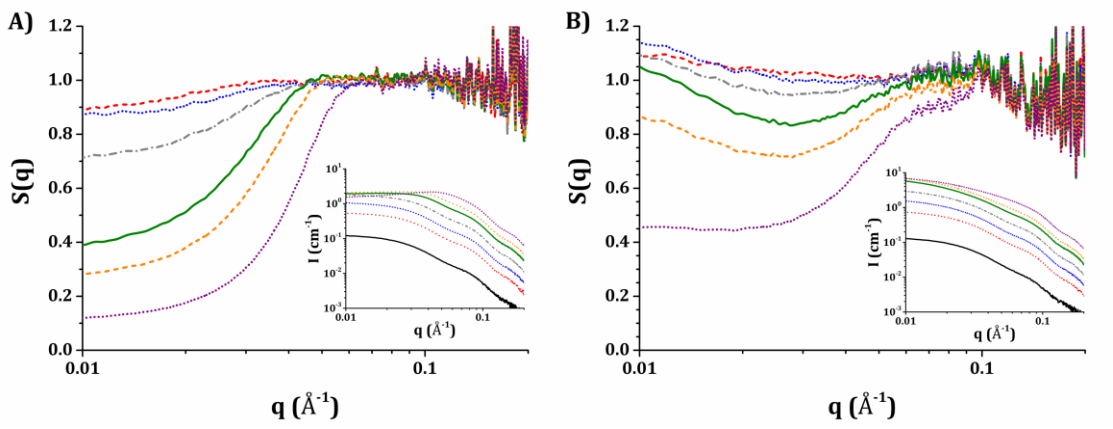

Figure 7.1. *S*(*q*) vs *q* measurements *via* SAXS for the IgG1 molecule at pH 5 for buffer (A) and 100 mM NaCl (B) conditions as a function of c_2: 1 (solid black), 5 (dashed red), 10 (dotted blue), 20 (dash-dotted gray), 40 (solid green), 60 (dashed orange) and 120 (dotted purple) g/L. Insets correspond to the same data in *I*(*q*) vs *q* form.

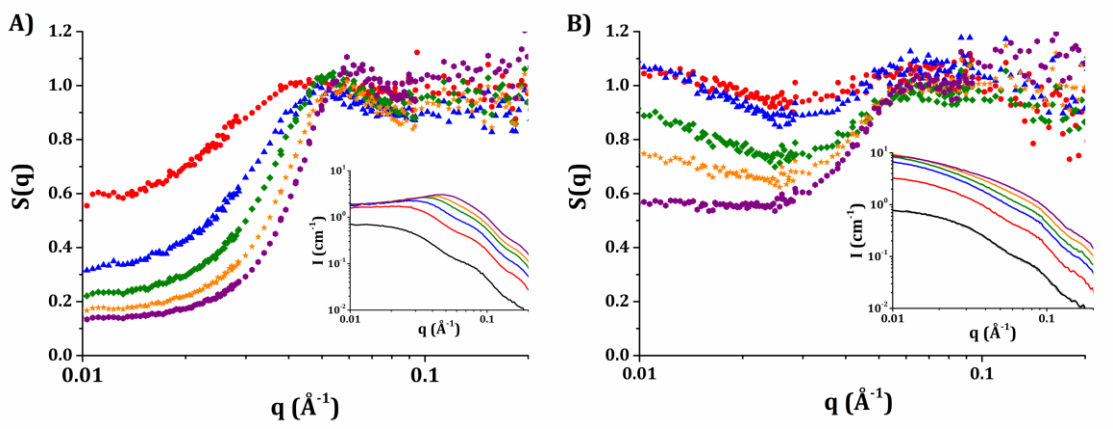

Figure 7.2. *S*(*q*) vs *q* measurements *via* SANS for the IgG1 molecule at pH 5 for buffer (A) and 100 mM NaCl (B) conditions as a function of c_2: 5 (black and gray), 20 (red circles), 40 (blue triangles), 60 (green diamonds), 80 (orange stars) and 100 (purple hexagons) g/L. Insets correspond to the same data in *I*(*q*) vs *q* form.

7.2.2 CG Modeling for *in-silico* Predictions of Colloidal Stability of Protein Solutions

In-silico predictions of protein solution properties are of large interest during the early stages of development of protein-based drugs as limited amounts of protein material are available. The approach and results at the end of Chapters 3-4 provided insights into the charge distribution having an impact in net-attractive interactions at low *TIS* for the IgG4 and aCgn. The results for the mAb molecules showed excellent agreement with all the experimentally measured B_{22} values from low to high *TIS*. Only the sequence and a crystal structure (or homology model) was required to perform such simulations without any prior knowledge of the experimental data. Consequently, the results obtained in Chapter 3 and 4 might suggest that this approach could be used to predict the formulation space for any protein of interest without any knowledge of experimental behavior. Similar approaches that rely on the protein structure and predefined data bases already exist [153,192]. However, most of them rely on

calculations based on a single protein, while the approach in Chapter 3 (*i.e.*, using MSOS to map B_{22} as a function of model parameters) intrinsically rely on the interactions between two or more proteins. Additionally, this approach might provide insights into the amino acid contacts responsible for the most attractive interactions, and the nature of these interactions (electrostatic or non-electrostatic) which can be used in the design and development of protein sequences for different industrial applications.

Although practical predictions of high-c_2 behavior with this approach is not encouraged due to current limitations in computational infrastructures, the approach of using higher order virial coefficients as explored in Chapter 3 could be applied to the 1bAA model. This could further reveal how multi-protein interactions might mediated the preferential interactions between domains/amino acids. Additionally, this approach is not limited to the 1bAA model as done in Chapters 3 and 4, but it could be extended to the 4bAA model discussed in Chapters 2 and 6, or more structurally complex CG models. However, a balance between computational time and accuracy needs to be considered in selecting the model, and this is very likely to depend on the protein of interest as larger proteins would demand larger computational times than smaller protein would with the same level of molecular resolution.

7.2.3 Predictions of Protein Crystallization and Phase Stability

The phase stability of protein solutions is relevant during the development and manufacturing of protein-based drugs. Crystallization can be used as a separation technique to either purify or characterize protein solutions *via* affinity precipitation, crystallography and NMR experiments. Additionally, liquid-liquid phase separation has been observed in protein solutions, which is undesired during the fill and finish steps of protein manufacturing. Both are expected to arise from strong protein-protein attractions

that makes it favorable for the protein to be present in a highly packed configuration (*e.g.*, a crystal, an amorphous solid or a highly dense liquid phase). Most of these properties are currently investigated experimentally. Predictions are currently achieved heuristically using values of interactions parameters at dilute limits (such as B_{22} or surrogates) as predictors of phase stability. Although this has been proven to provide, in some cases, accurate predictions of phase stability, the provided thresholds are rather arbitrary and set to conditions that are extremely attractive, under which phase stability is already compromised at very low c_2.

Interestingly, the TMMC algorithm was initially developed by Prof. Errington at the University of Buffalo as a technique to simulate the phase equilibrium of liquids as it can provide the chemical potentials of two phases at equilibrium. Consequently, the approach described in Chapters 2-4 can be used to evaluate the phase stability of protein solutions, as done in Chapter 4. However, it needs to be taken into account that crystallization is very dependent on the orientation of the proteins in the crystal, so high resolutions models would be needed to accurately predict crystallization. This is not the case for liquid-liquid splits as the TMMC approach can provide both the concentrations of the dilute and dense phases. In the worst-case situation, this approach could provide the maximum c_2 under which the protein solution is present in a single phase, or whether the current formulation(s) might be conducive of phase separation as the c_2 increases. Further research would be needed to identify the viability of predicting both phases, especially the dense phase, but the potential for these techniques to achieve so is latent.

7.2.4 Modeling Aggregation from Low to High c_2 for Peptide and Protein Solutions

Protein aggregation needs to be mitigated during development and manufacturing of protein products. However, many solution conditions, ranging from pH to buffer type and concentration, storage temperature and the addition of cosolutes, needs to be considered during the analysis of aggregation propensity of protein solutions. The results obtained in this thesis highlight the potential of using computer simulations with small experimental training sets to predict protein-protein interactions and unfolding thermodynamics with good accuracy and practical computational times. Protein aggregation arises from a combination of protein interactions and unfolding. Hence, current approaches could be directly extended to evaluate the effects of multibody effects in protein unfolding and further aggregation. This was first addressed in Chapter 6 by simultaneously simulating two interacting peptides, but this does not account for other effects arising from multi-body interactions. Although the approach in Chapter 6 could be extended to simulating hundreds of proteins or peptides as done in Chapters 2-4, the resulting simulation times would make this approach impractical. However, current advances in Discrete Molecular Dynamics and similar algorithms have been shown to reduce computational times by two (2) orders of magnitude in comparison to canonical MD algorithms (as those used in Chapter 6) [85]. Results from experimental CD and DSC measurements coupled with aggregation kinetic values (such as those obtained using parallel temperature initial rates, PTIR, or simultaneous multiple sample light scattering, SMSLS, techniques [248]) can be used as inputs to train the molecular models. Although these new approaches would come with additional short-comings (such as not predicting the right unfolding energies), the likelihood of gaining additional insights into the connection between multi-protein interactions, protein

unfolding, and aggregation might be of large interests to those studying protein or polypeptide solutions.

7.2.5 Protein Mixtures: Protein-Protein Cross-Interactions from Low to High c_2

Current developments in the biopharma industry have geared towards the design of protein complexes or mixtures of different proteins to enhance the potency or specificity of the drug. Although this dissertation focused on single-protein solutions, most of the techniques developed or used here can be applied to protein mixtures (*e.g.*, those with at least two species of proteins) with minor alterations or assumptions. First, the framework developed by Blanco *et al.* in the analysis of SLS data can be extended to multi-component solutions. This would give rise to co-protein interactions (*e.g.*, protein A *vs* protein B interactions, which could be termed $B_{2A\text{-}B}$) that would require a large experimental data set to evaluate. This arises from the need to evaluate the c_2-dependence of $B_{2A\text{-}A}$, $B_{2B\text{-}B}$ and $B_{2A\text{-}B}$, so experiments will scale with n^x, where n is the number of concentrations and x is the number of protein species in the solution. B_{2AB} values could also be measured as a function of the concentration of species A or B using the framework in Chapter 5 by treating the concentrated protein as a cosolute. Additionally, computer simulations can be performed for any mixture using frameworks based on those developed in Chapters 2-4. By considering that the addition of mixtures should not considerably increase computational times, an approach combining limited SLS and density measurements with simulations can provide the needed parametrization to predict the effect of mixing two or more protein species, significantly reducing the amount of experimental measurements required to obtain similar qualitative results.

REFERENCES

[1] R.S. Aggarwal, What's fueling the biotech engine-2012 to 2013., Nat. Biotechnol. 32 (2014) 32–9.

[2] D.M. Ecker, S.D. Jones, H.L. Levine, The therapeutic monoclonal antibody market, MAbs. 7 (2015) 9–14.

[3] T.T. Hansel, H. Kropshofer, T. Singer, J.A. Mitchell, A.J.T. George, The safety and side effects of monoclonal antibodies, Nat. Rev. Drug Discov. 9 (2010) 325–338.

[4] V. Pillay, H.K. Gan, A.M. Scott, Antibodies in oncology, N. Biotechnol. 28 (2011) 518–529.

[5] K.D. Miller, J. Weaver-Feldhaus, S.A. Gray, R.W. Siegel, M.J. Feldhaus, Production, purification, and characterization of human scFv antibodies expressed in Saccharomyces cerevisiae, Pichia pastoris, and Escherichia coli., Protein Expr. Purif. 42 (2005) 255–67.

[6] R.E. Bird, K.D. Hardman, J.W. Jacobson, S. Johnson, B.M. Kaufman, S.M. Lee, T. Lee, S.H. Pope, G.S. Riordan, M. Whitlow, Single-chain antigen-binding proteins., Science. 242 (1988) 423–6.

[7] W. Wang, S.K. Singh, N. Li, M.R. Toler, K.R. King, S. Nema, Immunogenicity of protein aggregates - Concerns and realities, Int. J. Pharm. 431 (2012) 1–11.

[8] S.J. Shire, Z. Shahrokh, J. Liu, Challenges in the development of high protein concentration formulations., J. Pharm. Sci. 93 (2004) 1390–402.

[9] M. Vázquez-Rey, D.A. Lang, Aggregates in monoclonal antibody manufacturing processes, Biotechnol. Bioeng. 108 (2011) 1494–1508.

[10] S. Uchiyama, Liquid formulation for antibody drugs, Biochim. Biophys. Acta - Proteins Proteomics. 1844 (2014) 2041–2052.

[11] D. Awotwe-Otoo, C. Agarabi, G.K. Wu, E. Casey, E. Read, S. Lute, K.A. Brorson, M.A. Khan, R.B. Shah, Quality by design: Impact of formulation variables and their interactions on quality attributes of a lyophilized monoclonal

antibody, Int. J. Pharm. 438 (2012) 167–175.

[12] C.J. Roberts, T.K. Das, E. Sahin, Predicting solution aggregation rates for therapeutic proteins: approaches and challenges., Int. J. Pharm. 418 (2011) 318–33.

[13] S. Mitragotri, P.A. Burke, R. Langer, Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies., Nat. Rev. Drug Discov. 13 (2014) 655–72.

[14] C.J. Roberts, Protein aggregation and its impact on product quality, Curr. Opin. Biotechnol. 30 (2014) 211–217.

[15] A.S. Rosenberg, Effects of protein aggregates: An immunologic perspective, AAPS J. 8 (2006) E501–E507.

[16] W. Wang, C.J. Roberts, Aggregation of Therapeutic Proteins, John Wiley & Sons, Inc., 2010.

[17] A.S. Raut, D.S. Kalonia, Opalescence in Monoclonal Antibody Solutions and Its Correlation with Intermolecular Interactions in Dilute and Concentrated Solutions, J. Pharm. Sci. 104 (2015) 1263–1274.

[18] M. Muschol, F. Rosenberger, Liquid–liquid phase separation in supersaturated lysozyme solutions and associated precipitate formation/crystallization, J. Chem. Phys. 107 (1997) 1953–1962.

[19] A.S. Raut, D.S. Kalonia, Pharmaceutical Perspective on Opalescence and Liquid-Liquid Phase Separation in Protein Solutions, Mol. Pharm. 13 (2016) 1431–1444.

[20] E. Rosenberg, S. Hepbildikler, W. Kuhne, G. Winter, Ultrafiltration concentration of monoclonal antibody solutions: Development of an optimized method minimizing aggregation, J. Memb. Sci. 342 (2009) 50–59.

[21] T. Menzen, W. Friess, Temperature-ramped studies on the aggregation, unfolding, and interaction of a therapeutic monoclonal antibody, J. Pharm. Sci. 103 (2014) 445–455.

[22] M.S. Neergaard, D.S. Kalonia, H. Parshad, A.D. Nielsen, E.H. Møller, M. van de Weert, Viscosity of high concentration protein formulations of monoclonal antibodies of the IgG1 and IgG4 subclass - prediction of viscosity through protein-protein interaction measurements., Eur. J. Pharm. Sci. 49 (2013) 400–10.

[23] R. Ghosh, C. Calero-Rubio, A. Saluja, C.J. Roberts, Relating Protein-Protein

Interactions and Aggregation Rates From Low to High Concentrations., J. Pharm. Sci. 105 (2016) 1086–96.

[24] T. Perevozchikova, H. Nanda, D.P. Nesta, C.J. Roberts, Protein Adsorption, Desorption, and Aggregation Mediated by Solid-Liquid Interfaces, J. Pharm. Sci. 104 (2015) 1946–1959.

[25] B.D. Connolly, C. Petry, S. Yadav, B. Demeule, N. Ciaccio, J.M.R. Moore, S.J. Shire, Y.R. Gokarn, Weak interactions govern the viscosity of concentrated antibody solutions: high-throughput analysis using the diffusion interaction parameter., Biophys. J. 103 (2012) 69–78.

[26] A. Quigley, D.R. Williams, The second virial coefficient as a predictor of protein aggregation propensity: A self-interaction chromatography study, Eur. J. Pharm. Biopharm. 96 (2015) 282–290.

[27] D. Roberts, R. Keeling, M. Tracka, C.F. Van Der Walle, S. Uddin, J. Warwicker, R. Curtis, The role of electrostatics in protein-protein interactions of a monoclonal antibody, Mol. Pharm. 11 (2014) 2475–2489.

[28] E.J. Yearley, I.E. Zarraga, S.J. Shire, T.M. Scherer, Y. Gokarn, N.J. Wagner, Y. Liu, Small-angle neutron scattering characterization of monoclonal antibody conformations and interactions at high concentrations., Biophys. J. 105 (2013) 720–31.

[29] M.A. Blanco, E. Sahin, Y. Li, C.J. Roberts, Reexamining protein-protein and protein-solvent interactions from Kirkwood-Buff analysis of light scattering in multi-component solutions, J. Chem. Phys. 134 (2011) 225103.

[30] E. Sahin, A.O. Grillo, M.D. Perkins, C.J. Roberts, Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies, J. Pharm. Sci. 99 (2010) 4830–4848.

[31] R.M. Murphy, A.M. Tsai, Misbehaving proteins: Protein (mis)folding, aggregation, and stability, Springer, 2006.

[32] P. Schuck, Analytical Ultracentrifugation as a Tool for Studying Protein Interactions., Biophys. Rev. 5 (2013) 159–171.

[33] D.J. Scott, P. Schuck, A brief introduction to the analytival ultracentrifugation of proteins for beginners, in: D.J. Scott, S.E. Harding, A.J. Rowe (Eds.), Anal. Ultracentrifugation, Royal Society of Chemistry, Cambridge, 2005: pp. 7–25.

[34] N.J. Clark, H. Zhang, S. Krueger, H.J. Lee, R.R. Ketchem, B. Kerwin, S.R.

Kanapuram, M.J. Treuheit, A. McAuley, J.E. Curtis, Small-angle neutron scattering study of a monoclonal antibody using free-energy constraints, J. Phys. Chem. B. 117 (2013) 14029–14038.

[35] J.D. Mills, M. Ben-Nun, K. Rollin, M.W.J. Bromley, J. Li, R.J. Hinde, C.L. Winstead, J.A. Sheehy, J.A. Boatz, P.W. Langhoff, Atomic Spectral Methods for Ab Initio Molecular Electronic Energy Surfaces: Transitioning From Small-Molecule to Biomolecular-Suitable Approaches, J. Phys. Chem. B. 120 (2016) 8321–8337.

[36] D.J. Winzor, M. Deszczynski, S.E. Harding, P.R. Wills, Nonequivalence of second virial coefficients from sedimentation equilibrium and static light scattering studies of protein solutions, Biophys. Chem. 128 (2007) 46–55.

[37] A.P. Minton, Recent applications of light scattering measurement in the biological and biopharmaceutical sciences, Anal. Biochem. 501 (2016) 4–22.

[38] M. Hofmann, M. Winzer, C. Weber, H. Gieseler, Prediction of Protein Aggregation in High Concentration Protein Solutions Utilizing Protein-Protein Interactions Determined by Low Volume Static Light Scattering, J. Pharm. Sci. 105 (2016) 1819–1828.

[39] A.P. Minton, Static light scattering from concentrated protein solutions, I: General theory for protein mixtures and application to self-associating proteins., Biophys. J. 93 (2007) 1321–8.

[40] P.D. Godfrin, I.E. Zarraga, J. Zarzar, L. Porcar, P. Falus, N.J. Wagner, Y. Liu, Effect of Hierarchical Cluster Formation on the Viscosity of Concentrated Monoclonal Antibody Formulations Studied by Neutron Scattering, J. Phys. Chem. B. 120 (2016) 278–291.

[41] O.D. Velev, E.W. Kaler, a M. Lenhoff, Protein interactions in solution characterized by light and neutron scattering: comparison of lysozyme and chymotrypsinogen., Biophys. J. 75 (1998) 2682–97.

[42] R. Chaudhuri, Y. Cheng, C.R. Middaugh, D.B. Volkin, High-Throughput Biophysical Analysis of Protein Therapeutics to Examine Interrelationships Between Aggregate Formation and Conformational Stability, AAPS J. 16 (2014) 48–64.

[43] M.S. Neergaard, A.D. Nielsen, H. Parshad, M. Van De Weert, Stability of monoclonal antibodies at high-concentration: Head-to-head comparison of the IgG1 and IgG4 subclass, J. Pharm. Sci. 103 (2014) 115–127.

[44] S.N. Timasheff, Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components., Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 9721–6.

[45] C.P. Schneider, D. Shukla, B.L. Trout, Arginine and the hofmeister series: The role of ion-ion interactions in protein aggregation suppression, J. Phys. Chem. B. 115 (2011) 7447–7458.

[46] B.S. Kendrick, B.S. Chang, T. Arakawa, B. Peterson, T.W. Randolph, M.C. Manning, J.F. Carpenter, Preferential exclusion of sucrose from recombinant interleukin-1 receptor antagonist: role in restricted conformational mobility and compaction of native state., Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 11917–11922.

[47] J. Zhang, D.D. Banks, F. He, M.J. Treuheit, G.W. Becker, Effects of sucrose and benzyl alcohol on GCSF conformational dynamics revealed by hydrogen deuterium exchange mass spectrometry, J. Pharm. Sci. 104 (2015) 1592–1600.

[48] J.A. Schellman, Protein Stability in Mixed Solvents: A Balance of Contact Interaction and Excluded Volume, Biophys. J. 85 (2003) 108–125.

[49] G.V. Barnett, V.I. Razinkov, B.A. Kerwin, S. Blake, W. Qi, R.A. Curtis, C.J. Roberts, Osmolyte Effects on Monoclonal Antibody Stability and Concentration-Dependent Protein Interactions with Water and Common Osmolytes, J. Phys. Chem. B. 120 (2016) 3318–3330.

[50] R.A. Lewus, P.A. Darcy, A.M. Lenhoff, S.I. Sandler, Interactions and phase behavior of a monoclonal antibody, Biotechnol. Prog. 27 (2011) 280–289.

[51] G.V. Barnett, V.I. Razinkov, B.A. Kerwin, T.M. Laue, A.H. Woodka, P.D. Butler, T. Perevozchikova, C.J. Roberts, Specific-Ion Effects on the Aggregation Mechanisms and Protein-Protein Interactions for Anti-streptavidin Immunoglobulin Gamma-1., J. Phys. Chem. B. 119 (2015) 5793–804.

[52] K. Gekko, S.N. Timasheff, Mechanism of Protein Stabilization by Glycerol: Preferential Hydration in Glycerol-Water Mixtures, Biochemistry. 20 (1981) 4667–4676.

[53] G. Xie, S.N. Timasheff, Temperature dependence of the preferential interactions of ribonuclease A in aqueous co-solvent systems: thermodynamic analysis., Protein Sci. 6 (1997) 222–232.

[54] G. Xie, S.N. Timasheff, Mechanism of the stabilization of ribonuclease a by sorbitol: Preferential hydration is greater for the denatured than for the native

protein, Protein Sci. 6 (1997) 211–221.

[55] A. Ben-Naim, A.M. Navarro, J.M. Leal, A Kirkwood-Buff analysis of local properties of solutions., Phys. Chem. Chem. Phys. 10 (2008) 2451–60.

[56] A. Ben-Naim, Solvent Effects on Protein Association and Protein Folding, Biopolymers. 29 (1990) 567–596.

[57] M.E. Paulaitis, L.R. Pratt, Hydration theory for molecular biophysics., Adv. Protein Chem. 62 (2002) 283–310.

[58] G.V. Barnett, W. Qi, S. Amin, E.N. Lewis, V.I. Razinkov, B.A. Kerwin, Y. Liu, C.J. Roberts, Structural Changes and Aggregation Mechanisms for Anti-Streptavidin IgG1 at Elevated Concentration, J. Phys. Chem. B. 119 (2015) 15150–15163.

[59] A. Ben-Naim, Statistical Thermodynamics for Chemists and Biochemists, Plenum Press, 1992.

[60] C. Calero-Rubio, C. Strab, G.V. Barnett, C.J. Roberts, Protein Partial Molar Volumes in Multi-Component Solutions From the Perspective of Inverse Kirkwood-Buff Theory, J. Phys. Chem. B. 121 (2017) 5897–5907.

[61] V. Shen, J. Cheung, J. Errington, T. Truskett, Coarse-grained strategy for modeling protein stability in concentrated solutions. II: phase behavior, Biophys. J. 90 (2006) 1949–1960.

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

[63] A. Grünberger, P.K. Lai, M.A. Blanco, C.J. Roberts, Coarse-grained modeling of protein second osmotic virial coefficients: Sterics and short-ranged attractions, J. Phys. Chem. B. 117 (2013) 763–770.

[64] W.G. Noid, Perspective: Coarse-grained models for biomolecular systems., J. Chem. Phys. 139 (2013) 90901.

[65] T. Zhang, P.H. Nguyen, J. Nasica-Labouze, Y. Mu, P. Derreumaux, Folding Atomistic Proteins in Explicit Solvent Using Simulated Tempering, J. Phys. Chem. B. 119 (2015) 6941–6951.

[66] M.A. Blanco, E. Sahin, A.S. Robinson, C.J. Roberts, Coarse-Grained Model for Colloidal Protein Interactions, B22, and Protein Cluster Formation, J. Phys. Chem. B. 117 (2013) 16013–16028.

[67] M.S. Shell, Systematic coarse-graining of potential energy landscapes and dynamics in liquids., J. Chem. Phys. 137 (2012) 84503.

[68] J. Errington, Direct calculation of liquid–vapor phase equilibria from transition matrix Monte Carlo simulation, J. Chem. Phys. 118 (2003) 9915–9925.

[69] T.T. Foley, M.S. Shell, W.G. Noid, The impact of resolution upon entropy and information in coarse-grained models., J. Chem. Phys. 143 (2015) 243104.

[70] N.J.H. Dunn, W.G. Noid, Bottom-up coarse-grained models that accurately describe the structure, pressure, and compressibility of molecular liquids, J. Chem. Phys. 143 (2015) 243148.

[71] V. Tozzini, Minimalist models for proteins: a comparative analysis, Q. Rev. Biophys. 43 (2010) 333–371.

[72] M. Baaden, S.J. Marrink, Coarse-grain modelling of protein-protein interactions., Curr. Opin. Struct. Biol. 23 (2013) 878–86.

[73] R. Piazza, Interactions in protein solutions near crystallisation: a colloid physics approach, J. Cryst. Growth. 196 (1999) 415–423.

[74] R. Piazza, V. Peyre, V. Degiorgio, "Sticky hard spheres" model of proteins near crystallization: A test based on the osmotic compressibility of lysozyme solutions, Phys. Rev. E. 58 (1998) R2733–R2736.

[75] S.N. Singh, S. Yadav, S.J. Shire, D.S. Kalonia, Dipole-dipole interaction in antibody solutions: correlation with viscosity behavior at high concentration, Pharm. Res. 31 (2014) 2549–2558.

[76] D. Bratko, A. Striolo, J.Z. Wu, H.W. Blanch, J.M. Prausnitz, Orientation-averaged pair potentials between dipolar proteins or colloids, J. Phys. Chem. B. 106 (2002) 2714–2720.

[77] P.D. Godfrin, R. Castañeda-Priego, Y. Liu, N.J. Wagner, Intermediate range order and structure in colloidal dispersions with competing interactions., J. Chem. Phys. 139 (2013) 154904.

[78] E.J. Yearley, P.D. Godfrin, T. Perevozchikova, H. Zhang, P. Falus, L. Porcar, M. Nagao, J.E. Curtis, P. Gawande, R. Taing, I.E. Zarraga, N.J. Wagner, Y. Liu, Observation of small cluster formation in concentrated monoclonal antibody solutions and its implications to solution viscosity., Biophys. J. 106 (2014) 1763–70.

[79] P.D. Godfrin, N. Valadez-Perez, R. Castaneda-Priego, N. Wagner, Y. Liu, Generalized phase behavior of cluster formation in colloidal dispersions with competing interactions, Soft Matter. 10 (2014) 5061–71.

[80] S. Piana, K. Lindorff-Larsen, D.E. Shaw, Protein folding kinetics and thermodynamics from atomistic simulation., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 17845–50.

[81] M.A. Blanco, T. Perevozchikova, V. Martorana, M. Manno, C.J. Roberts, Protein-protein interactions in dilute to concentrated solutions: α-Chymotrypsinogen in acidic conditions, J. Phys. Chem. B. 118 (2014) 5817–5831.

[82] C.J. Roberts, M.A. Blanco, Role of anisotropic interactions for proteins and patchy nanoparticles, J. Phys. Chem. B. 118 (2014) 12599–12611.

[83] G. Foffi, F. Sciortino, On the possibility of extending the Noro-Frenkel generalized law of correspondent states to nonisotropic patchy interactions., J. Phys. Chem. B. 111 (2007) 9702–5.

[84] B.A. Paik, M.A. Blanco, X. Jia, C.J. Roberts, K.L. Kiick, Aggregation of poly(acrylic acid)-containing elastin-mimetic copolymers, Soft Matter. 11 (2015) 1839–1850.

[85] M. Cheon, I. Chang, C.K. Hall, Extending the PRIME model for protein aggregation to all 20 amino acids., Proteins. 78 (2010) 2950–60.

[86] N.J.H. Dunn, W.G. Noid, Bottom-up coarse-grained models with predictive accuracy and transferability for both structural and thermodynamic properties of heptane-toluene mixtures, J. Chem. Phys. 144 (2016) 204124.

[87] S.-H. Chong, S. Ham, Protein Folding Thermodynamics: A New Computational Approach, J. Phys. Chem. B. 118 (2014) 5017–5025.

[88] E.F. Casassa, H. Eisenberg, THERMODYNAMIC ANALYSIS OF MULTICOMPONENT SOLUTIONS., Adv. Protein Chem. 19 (1964) 287–395.

[89] B.L. Neal, D. Asthagiri, A.M. Lenhoff, Molecular Origins of Osmotic Second Virial Coefficients of Proteins, Biophys. J. 75 (1998) 2469–2477.

[90] P.M. Tessier, A.M. Lenhoff, S.I. Sandler, Rapid measurement of protein osmotic second virial coefficients by self-interaction chromatography., Biophys. J. 82 (2002) 1620–31.

[91] P.M. Tessier, S.I. Sandler, A.M. Lenhoff, Direct measurement of protein osmotic second virial cross coefficients by cross-interaction chromatography, Protein Sci. 13 (2004) 1379–1390.

[92] D.W. Siderius, W.P. Krekelberg, C.J. Roberts, V.K. Shen, Osmotic virial coefficients for model protein and colloidal solutions: Importance of ensemble constraints in the analysis of light scattering data, J. Chem. Phys. 136 (2012) 175102.

[93] N. Rakel, K.C. Bauer, L. Galm, J. Hubbuch, From osmotic second virial coefficient (B22) to phase behavior of a monoclonal antibody, Biotechnol. Prog. 31 (2015) 438–451.

[94] T. Young, C. Roberts, Structure and thermodynamics of colloidal protein cluster formation: Comparison of square-well and simple dipolar models, J. Chem. Phys. 131 (2009) 125104.

[95] J.G. Kirkwood, F.P. Buff, The Statistical Mechanical Theory of Solutions. I, J. Chem. Phys. 19 (1951) 774.

[96] D.A. McQuarrie, Statistical mechanics, University Science Books, 2000.

[97] J.M. Schurr, D.P. Rangel, S.R. Aragon, A Contribution to the Theory of Preferential Interaction Coefficients, Biophys. J. 89 (2005) 2258–2276.

[98] P.E. Smith, Equilibrium dialysis data and the relationships between preferential interaction parameters for biological systems in terms of kirkwood-buff integrals, J. Phys. Chem. B. 110 (2006) 2862–2868.

[99] P.E. Smith, Chemical potential derivatives and preferential interaction parameters in biological systems from Kirkwood-Buff theory., Biophys. J. 91 (2006) 849–56.

[100] S. Shimizu, Estimation of excess solvation numbers of water and cosolvents from preferential interaction and volumetric experiments, J. Chem. Phys. 120 (2004) 4989–4990.

[101] E.S. Courtenay, M.W. Capp, C.F. Anderson, M.T. Record Jr., Vapor Pressure Osmometry Studies of Osmolyte - Protein Interactions : Implications for the Action of Osmoprotectants in Vivo and for the Interpretation of " Osmotic Stress " Experiments in Vitro, Biochemistry. 39 (2000) 4455–4471.

[102] D.B. Knowles, I.A. Shkel, N.M. Phan, M. Sternke, E. Lingeman, X. Cheng, L. Cheng, K. O'Connor, M.T. Record, Chemical Interactions of Polyethylene

Glycols (PEGs) and Glycerol with Protein Functional Groups: Applications to Effects of PEG and Glycerol on Protein Processes, Biochemistry. 54 (2015) 3528–3542.

[103] G.V. Barnett, V.I. Razinkov, B.A. Kerwin, S. Blake, W. Qi, R.A. Curtis, C.J. Roberts, Reply to Comment on " Osmolyte Effects on Monoclonal Antibody Stability and Concentration-Dependent Protein Interactions with Water and Common Osmolytes " Reply to Comment on " Osmolyte Effects on Monoclonal Antibody Stability and Concentration- Dependent, J. Phys. Chem. B. 120 (2016) 11333–11334.

[104] J. Rösgen, M. Auton, Comment on "Osmolyte Effects on Monoclonal Antibody Stability and Concentration-Dependent Protein Interactions with Water and Common Osmolytes," J. Phys. Chem. B. 120 (2016) 11331–11332.

[105] D. Arzenšek, D. Kuzman, R. Podgornik, Hofmeister Effects in Monoclonal Antibody Solution Interactions, J. Phys. Chem. B. 119 (2015) 10375–10389.

[106] A. Saluja, R.M. Fesinmeyer, S. Hogan, D.N. Brems, Y.R. Gokarn, Diffusion and sedimentation interaction parameters for measuring the second virial coefficient and their utility as predictors of protein aggregation., Biophys. J. 99 (2010) 2657–65.

[107] M.M. Castellanos, A. McAuley, J.E. Curtis, Investigating Structure and Dynamics of Proteins in Amorphous Phases Using Neutron Scattering, Comput. Struct. Biotechnol. J. 15 (2017) 117–130.

[108] C.J. Roberts, Non-native protein aggregation kinetics, Biotechnol. Bioeng. 98 (2007) 927–938.

[109] M.E.M. Cromwell, E. Hilario, F. Jacobson, Protein aggregation and bioprocessing, AAPS J. 8 (2006) E572–E579.

[110] S. Telikepalli, H.E. Shinogle, P.S. Thapa, J.H. Kim, M. Deshpande, V. Jawa, C.R. Middaugh, L.O. Narhi, M.K. Joubert, D.B. Volkin, Physical Characterization and In Vitro Biological Impact of Highly Aggregated Antibodies Separated into Size-Enriched Populations by Fluorescence-Activated Cell Sorting., J. Pharm. Sci. 104 (2015) 1575–91.

[111] E. Sahin, J.L. Jordan, M.L. Spatara, A. Naranjo, J.A. Costanzo, W.F. Weiss IV, A.S. Robinson, E.J. Fernandez, C.J. Roberts, Computational design and biophysical characterization of aggregation-resistant point mutations for γD crystallin illustrate a balance of conformational stability and intrinsic aggregation propensity, Biochemistry. 50 (2011) 628–639.

[112] E. Sahin, W.F. Weiss IV, A.M. Kroetsch, K.R. King, R.K. Kessler, T.K. Das, C.J. Roberts, Aggregation and pH-Temperature Phase Behavior for Aggregates of an IgG2 Antibody, J. Pharm. Sci. 101 (2012) 1678–1687.

[113] C.J. Roberts, Therapeutic protein aggregation: Mechanisms, design, and control, Trends Biotechnol. 32 (2014) 372–380.

[114] H. Wu, R. Kroe-Barrett, S. Singh, A.S. Robinson, C.J. Roberts, Competing aggregation pathways for monoclonal antibodies., FEBS Lett. 588 (2014) 936–41.

[115] Y. Li, W.F. Weiss, C.J. Roberts, Characterization of high-molecular-weight nonnative aggregates and aggregation kinetics by size exclusion chromatography with inline multi-angle laser light scattering., J. Pharm. Sci. 98 (2009) 3997–4016.

[116] G. Thiagarajan, A. Semple, J.K. James, J.K. Cheung, M. Shameem, A comparison of biophysical characterization techniques in predicting monoclonal antibody stability, MAbs. 8 (2016) 1088–1097.

[117] A.R. Fersht, V. Daggett, Protein Folding and Unfolding at Atomic Resolution, Cell. 108 (2002) 573–582.

[118] T.J. Lane, D. Shukla, K.A. Beauchamp, V.S. Pande, To milliseconds and beyond: challenges in the simulation of protein folding., Curr. Opin. Struct. Biol. 23 (2013) 58–65.

[119] X.C. Yan, J. Tirado-Rives, W.L. Jorgensen, Hydration Properties and Solvent Effects for All-Atom Solutes in Polarizable Coarse-Grained Water, J. Phys. Chem. B. 120 (2016) 8102–8114.

[120] M.S. Shell, The relative entropy is fundamental to multiscale and inverse thermodynamic problems., J. Chem. Phys. 129 (2008) 144108.

[121] M. Lapelosa, T.W. Patapoff, I.E. Zarraga, Molecular simulations of the pairwise interaction of monoclonal antibodies, J. Phys. Chem. B. 118 (2014) 13132–13141.

[122] J.K. Cheung, V.K. Shen, J.R. Errington, T.M. Truskett, Coarse-Grained Strategy for Modeling Protein Stability in Concentrated Solutions. III: Directional Protein Interactions, Biophys. J. 92 (2007) 4316–4324.

[123] P.L. Freddolino, C.B. Harrison, Y. Liu, K. Schulten, Challenges in protein folding simulations: Timescale, representation, and analysis., Nat. Phys. 6 (2010)

751–758.

[124] Y. Liu, W.-R. Chen, S.-H. Chen, Cluster formation in two-Yukawa fluids., J. Chem. Phys. 122 (2005) 44507.

[125] D.M. Heyes, Molecular dynamics at constant pressure and temperature, Chem. Phys. 82 (1983) 285–301.

[126] K. Koga, Osmotic second virial coefficient of methane in water, J. Phys. Chem. B. 117 (2013) 12619–12624.

[127] D. Mercadante, S. Milles, G. Fuertes, D.I. Svergun, E.A. Lemke, F. Gräter, Kirkwood-Buff Approach Rescues Overcollapse of a Disordered Protein in Canonical Protein Force Fields, J. Phys. Chem. B. 119 (2015) 7975–7984.

[128] D. Frenkel, B. Smit, Understanding molecular simulation, Second Edi, 2002.

[129] K.R.S. Shaul, A.J. Schultz, A. Perera, D.A. Kofke, Integral-equation theories and Mayer-sampling Monte Carlo: a tandem approach for computing virial coefficients of simple fluids, Mol. Phys. 109 (2011) 2395–2406.

[130] A.J. Schultz, N.S. Barlow, V. Chaudhary, D. a. Kofke, Mayer Sampling Monte Carlo calculation of virial coefficients on graphics processors, Mol. Phys. 111 (2012) 1–9.

[131] J.K. Singh, D.A. Kofke, Mayer sampling: Calculation of cluster integrals using free-energy perturbation methods, Phys. Rev. Lett. 92 (2004) 220601–1.

[132] A.J. Schultz, D.A. Kofke, Virial coefficients of model alkanes, J. Chem. Phys. 133 (2010) 104101.

[133] A.J. Schultz, D.A. Kofke, Sixth, seventh and eighth virial coefficients of the Lennard-Jones model, Mol. Phys. 107 (2009) 2309–2318.

[134] K.S. Rane, J.R. Errington, Using Monte Carlo simulation to compute liquid-vapor saturation properties of ionic liquids., J. Phys. Chem. B. 117 (2013) 8018–30.

[135] H.C. Andersen, Molecular dynamics simulations at constant pressure and/or temperature, J. Chem. Phys. 72 (1980) 2384.

[136] D. Siderius, V. Shen, Use of the grand canonical transition-matrix Monte Carlo method to model gas adsorption in porous materials, J. Phys. Chem. C. 117 (2013) 5861–5872.

[137] J.C. Flores-Canales, M. Kurnikova, Targeting Electrostatic Interactions in Accelerated Molecular Dynamics with Application to Protein Partial Unfolding, J. Chem. Theory Comput. (2015) 150519074007002.

[138] T. Zang, L. Yu, C. Zhang, J. Ma, Parallel continuous simulated tempering and its applications in large-scale molecular simulations., J. Chem. Phys. 141 (2014) 44113.

[139] Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding, Chem. Phys. Lett. (1999) 141–151.

[140] D. Frenkel, G. Mooij, B. Smit, Novel scheme to study structural and thermal properties of continuously deformable molecules, J. Phys. Condens. …. 4 (1992) 3053–3076.

[141] W. Li, B.A. Persson, M. Morin, M.A. Behrens, M. Lund, M. Zackrisson Oskolkova, Charge-induced patchy attractions between proteins, J. Phys. Chem. B. 119 (2015) 503–508.

[142] S. Zimmerman, A. Minton, Macromolecular crowding: biochemical, biophysical, and physiological consequences, Annu. Rev. Biophys. (1993) 27–65.

[143] M.H. Priya, S. Merchant, D. Asthagiri, M.E. Paulaitis, Quasi-chemical theory of cosolvent hydrophobic preferential interactions., J. Phys. Chem. B. 116 (2012) 6506–13.

[144] J.K. Cheung, T.M. Truskett, Coarse-Grained Strategy for Modeling Protein Stability in Concentrated Solutions, Biophys. J. 89 (2005) 2372–2384.

[145] D.S. Tomar, S. Kumar, S.K. Singh, S. Goswami, L. Li, Molecular basis of high viscosity in concentrated antibody solutions: Strategies for high concentration drug product development, MAbs. 8 (2016) 216–228.

[146] C. Calero-Rubio, A. Saluja, C.J. Roberts, Coarse-Grained Antibody Models for "weak" Protein-Protein Interactions from Low to High Concentrations, J. Phys. Chem. B. 120 (2016) 6592–6605.

[147] E.A. Padlan, Anatomy of the antibody molecule., Mol. Immunol. 31 (1994) 169–217.

[148] S.I. Sandler, An Introduction to Applied Statistical Thermodynamics, John Wiley & Sons, Inc., 2010.

[149] W.D. Cornell, P. Cieplak, C.I. Bayly, K.M. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Foz, J.W. Caldwell, P.A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids and Organic Molecules, J. Am. Chem. Soc. 117 (1995) 5179.

[150] C. Karney, Quaternions in molecular modeling, J. Mol. Graph. Model. 25 (2007) 595–604.

[151] S.J. Perkins, Protein volumes and hydration effects, Eur. J. Biochem. 157 (1986) 169–180.

[152] J.A. Schellman, Temperature, stability, and the hydrophobic interaction., Biophys. J. 73 (1997) 2960–4.

[153] V. Voynov, N. Chennamsetty, V. Kayser, B. Helk, B.L. Trout, Predictive tools for stabilization of therapeutic proteins, MAbs. 1 (2009) 580–582.

[154] T. Laue, B. Demeler, A postreductionist framework for protein biochemistry, Nat. Chem. Biol. 7 (2011) 331–334.

[155] D. Roberts, R. Keeling, M. Tracka, C.F. van der Walle, S. Uddin, J. Warwicker, R. Curtis, Specific Ion and Buffer Effects on Protein–Protein Interactions of a Monoclonal Antibody, Mol. Pharm. 12 (2015) 179–193.

[156] Y.R. Gokarn, R.M. Fesinmeyer, A. Saluja, V. Razinkov, S.F. Chase, T.M. Laue, D.N. Brems, Effective charge measurements reveal selective and preferential accumulation of anions, but not cations, at the protein surface in dilute salt solutions, Protein Sci. 20 (2011) 580–587.

[157] A. Arslanargin, A. Powers, T.L. Beck, S.W. Rick, Models of Ion Solvation Thermodynamics in Ethylene Carbonate and Propylene Carbonate, J. Phys. Chem. B. 120 (2016) 1497–1508.

[158] F. Zhang, M.W.A. Skoda, R.M.J. Jacobs, S. Zorn, R.A. Martin, C.M. Martin, G.F. Clark, S. Weggler, A. Hildebrandt, O. Kohlbacher, F. Schreiber, Reentrant condensation of proteins in solution induced by multivalent counterions, Phys. Rev. Lett. 101 (2008) 148101.

[159] J. Charles A Janeway, P. Travers, M. Walport, M.J. Shlomchik, The structure of a typical antibody molecule, Garland Science, 2001.

[160] L. Bongini, D. Fanelli, F. Piazza, P. De Los Rios, S. Sandin, U. Skoglund, Freezing immunoglobulins to see them move., Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 6466–6471.

[161] M. Galanti, D. Fanelli, F. Piazza, Conformation-controlled binding kinetics of antibodies, Sci. Rep. 6 (2016) 18976.

[162] C. De Michele, P. De Los Rios, G. Foffi, F. Piazza, Simulation and Theory of Antibody Binding to Crowded Antigen-Covered Surfaces, PLoS Comput. Biol. 12 (2016) 1–17.

[163] L.J. Harris, S.B. Larson, K.W. Hasel, A. McPherson, Refined structure of an intact IgG2a monoclonal antibody, Biochemistry. 36 (1997) 1581–1597.

[164] L.J. Harris, E. Skaletsky, A. McPherson, Crystallographic structure of an intact IgG1 monoclonal antibody., J. Mol. Biol. 275 (1998) 861–872.

[165] J. Wang, R.H. Swendsen, Transition Matrix Monte Carlo Method, J. Stat. Phys. 106 (2002) 245–285.

[166] M. Fitzgerald, R.R. Picard, R.N. Silver, Canonical transition probabilities for adaptive Metropolis simulation, EPL (Europhysics Lett. 46 (1999) 282–287.

[167] J.R. Errington, Prewetting transitions for a model argon on solid carbon dioxide system., Langmuir. 20 (2004) 3798–804.

[168] V. Shen, J. Errington, Metastability and instability in the Lennard-Jones fluid investigated by transition-matrix Monte Carlo, J. Phys. Chem. B. 108 (2004) 19595–19606.

[169] H.N. Po, N.M. Senozan, The Henderson-Hasselbalch Equation: Its History and Limitations, J. Chem. Educ. 78 (2001) 1499–1503.

[170] J.K. Singh, J.R. Errington, Calculation of phase coexistence properties and surface tensions of n-alkanes with grand-canonical transition-matrix monte carlo simulation and finite-size scaling., J. Phys. Chem. B. 110 (2006) 1369–76.

[171] W.G. Lilyestrom, S. Yadav, S.J. Shire, T.M. Scherer, Monoclonal antibody self-association, cluster formation, and rheology at high concentrations., J. Phys. Chem. B. 117 (2013) 6373–84.

[172] Y. Liu, L. Porcar, J. Chen, W.R. Chen, P. Falus, A. Faraone, E. Fratini, K. Hong, P. Baglioni, Lysozyme protein solution with an intermediate range order structure, J. Phys. Chem. B. 115 (2011) 7238–7247.

[173] A. George, W.W. Wilson, Predicting protein crystallization from a dilute solution property., Acta Crystallogr. D. Biol. Crystallogr. 50 (1994) 361–5.

[174] C.G. Malmberg, A.A. Maryott, Dielectric constants of aqueous solutions of

dextrose and sucrose, J. Res. Natl. Bur. Stand. (1934). 45 (1950) 299.

[175] E. Sahin, C.J. Roberts, Size-exclusion chromatography with multi-angle light scattering for elucidating protein aggregation mechanisms, Methods Mol. Biol. 899 (2012) 403–423.

[176] T.B. Tan, A.J. Schultz, D.A. Kofke, Virial coefficients, equation of state, and solid–fluid coexistence for the soft sphere model, Mol. Phys. 109 (2011) 123–132.

[177] H. Gould, J. Tobochnik, D.S. Lemons, Statistical and Thermal Physics with Computer Applications, Princeton University Press, 2011.

[178] C. Zhang, B.M. Pettitt, Computation of high-order virial coefficients in high-dimensional hard-sphere fluids by Mayer sampling, Mol. Phys. 112 (2014) 1427–1447.

[179] F.H. Ree, Seventh Virial Coefficients for Hard Spheres and Hard Disks, J. Chem. Phys. 46 (1967) 4181.

[180] S. Vafaei, B. Tomberli, C.G. Gray, McMillan-Mayer theory of solutions revisited: Simplifications and extensions, J. Chem. Phys. 141 (2014) 154501.

[181] K.R.S. Shaul, A.J. Schultz, D.A. Kofke, Mayer-sampling Monte Carlo calculations of uniquely flexible contributions to virial coefficients, J. Chem. Phys. 135 (2011) 124101.

[182] J. de Boer, Molecular distribution and equation of state of gases, Reports Prog. Phys. 12 (1949) 314.

[183] N.F. Carnahan, K.E. Starling, Equation of State for Nonattracting Rigid Spheres, J. Chem. Phys. 51 (1969) 635–636.

[184] A.J. Schultz, D.A. Kofke, Fifth to eleventh virial coefficients of hard spheres, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 90 (2014) 23301.

[185] H. Wu, K. Truncali, J. Ritchie, R. Kroe-Barrett, S. Singh, A.S. Robinson, C.J. Roberts, Weak protein interactions and pH- and temperature-dependent aggregation of human Fc1, MAbs. 7 (2015) 1072–1083.

[186] V. Kumar, N. Dixit, L.L. Zhou, W. Fraunhofer, Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations., Int. J. Pharm. 421 (2011) 82–93.

[187] N. Kim, R.L. Remmele, D. Liu, V.I. Razinkov, E.J. Fernandez, C.J. Roberts, Aggregation of anti-streptavidin immunoglobulin gamma-1 involves Fab unfolding and competing growth pathways mediated by pH and salt concentration, Biophys. Chem. 172 (2013) 26–36.

[188] L. Li, S. Kumar, P.M. Buck, C. Burns, J. Lavoie, S.K. Singh, N.W. Warne, P. Nichols, N. Luksha, D. Boardman, Concentration dependent viscosity of monoclonal antibody solutions: Explaining experimental behavior in terms of molecular properties, Pharm. Res. 31 (2014) 3161–3178.

[189] J.J. Hung, A.U. Borwankar, B.J. Dear, T.M. Truskett, K.P. Johnston, High concentration tangential flow ultrafiltration of stable monoclonal antibody solutions with low viscosities, J. Memb. Sci. 508 (2016) 113–126.

[190] M.M. Castellanos, N.J. Clark, M.C. Watson, S. Krueger, A. McAuley, J.E. Curtis, Role of Molecular Flexibility and Colloidal Descriptions of Proteins in Crowded Environments from Small-Angle Scattering, J. Phys. Chem. B. 120 (2016) 12511–12518.

[191] Y. Wang, R.F. Latypov, A. Lomakin, J.A. Meyer, B.A. Kerwin, S. Vunnum, G.B. Benedek, Quantitative Evaluation of Colloidal Stability of Antibody Solutions using PEG-Induced Liquid–Liquid Phase Separation, Mol. Pharm. 11 (2014) 1391–1402.

[192] C.J. O'Brien, M.A. Blanco, J.A. Costanzo, M. Enterline, E.J. Fernandez, A.S. Robinson, C.J. Roberts, Modulating non-native aggregation and electrostatic protein-protein interactions with computationally designed single-point mutations, Protein Eng. Des. Sel. 29 (2016) 231–243.

[193] J. Warwicker, S. Charonis, R.A. Curtis, Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design, Mol. Pharm. 11 (2014) 294–303.

[194] S.B. Geng, J.K. Cheung, C. Narasimhan, M. Shameem, P.M. Tessier, Improving Monoclonal Antibody Selection and Engineering using Measurements of Colloidal Protein Interactions., J. Pharm. Sci. 103 (2014) 3356–63.

[195] B.M. Baynes, B.L. Trout, Rational design of solution additives for the prevention of protein aggregation., Biophys. J. 87 (2004) 1631–9.

[196] J.A. Pathak, R.R. Sologuren, R. Narwal, Do Clustering Monoclonal Antibody Solutions Really Have a Concentration Dependence of Viscosity?, Biophys. J. 104 (2013) 913–923.

[197] T.M. Young, C.J. Roberts, A quasichemical approach for protein-cluster free energies in dilute solution, J. Chem. Phys. 127 (2007) 165101.

[198] A. Chaudhri, I.E. Zarraga, S. Yadav, T.W. Patapoff, S.J. Shire, G.A. Voth, The role of amino acid sequence in the self-association of therapeutic monoclonal antibodies: Insights from coarse-grained modeling, J. Phys. Chem. B. 117 (2013) 1269–1279.

[199] J. Arora, Y. Hu, R. Esfandiary, H.A. Sathish, S.M. Bishop, S.B. Joshi, C.R. Middaugh, D.B. Volkin, D.D. Weis, Charge-mediated Fab-Fc interactions in an IgG1 antibody induce reversible self-association, cluster formation, and elevated viscosity, MAbs. 8 (2016) 1561–1574.

[200] G.V. Barnett, W. Qi, S. Amin, E. Neil Lewis, C.J. Roberts, Aggregate structure, morphology and the effect of aggregation mechanisms on viscosity at elevated protein concentrations, Biophys. Chem. 207 (2015) 21–29.

[201] M.A. Woldeyes, C. Calero-Rubio, E.M. Furst, C.J. Roberts, Predicting Protein Interactions of Concentrated Globular Protein Solutions Using Colloidal Models, J. Phys. Chem. B. 121 (2017) 4756–4767.

[202] Y.I. Li, B.A. Ogunnaike, C.J. Roberts, Multi-Variate Approach to Global Protein Aggregation Behavior and Kinetics : Effects of pH , NaCl , and Temperature for a -Chymotrypsinogen A, 99 (2010) 645–662.

[203] M. Rostkowski, M.H. Olsson, C.R. Søndergaard, J.H. Jensen, Graphical analysis of pH-dependent properties of proteins predicted using PROPKA, BMC Struct. Biol. 11 (2011) 6.

[204] C.E. Felder, J. Prilusky, I. Silman, J.L. Sussman, A server and database for dipole moments of proteins, Nucleic Acids Res. 35 (2007) 512–521.

[205] L. Sapir, D. Harries, Macromolecular Stabilization by Excluded Cosolutes: Mean Field Theory of Crowded Solutions, J. Chem. Theory Comput. 11 (2015) 3478–3490.

[206] L. Nicoud, M. Sozo, P. Arosio, A. Yates, E. Norrant, M. Morbidelli, Role of cosolutes in the aggregation kinetics of monoclonal antibodies, J. Phys. Chem. B. 118 (2014) 11921–11930.

[207] M.H. Priya, H.S. Ashbaugh, M.E. Paulaitis, Cosolvent preferential molecular interactions in aqueous solutions., J. Phys. Chem. B. 115 (2011) 13633–42.

[208] A.U. Borwankar, B.J. Dear, A. Twu, J.J. Hung, A.K. Dinin, B.K. Wilson, J. Yue,

J.A. Maynard, T.M. Truskett, K.P. Johnston, Viscosity Reduction of a Concentrated Monoclonal Antibody with Arginine-HCl and Arginine-Glutamate, Ind. Eng. Chem. Res. 55 (2016) 11225–11234.

[209] B.J. Dear, J.J. Hung, T.M. Truskett, K.P. Johnston, Contrasting the Influence of Cationic Amino Acids on the Viscosity and Stability of a Highly Concentrated Monoclonal Antibody, Pharm. Res. 34 (2017) 193–207.

[210] K.E. Newman, Kirkwood-Buff solution theory: derivation and applications, Chem. Soc. Rev. 23 (1994) 31–40.

[211] B.A. Ogunnaike, Random phenomena : fundamentals of probability and statistics for engineers, CRC Press, 2010.

[212] N.C. Ekdawi-Sever, P.B. Conrad, J.J. de Pablo, Molecular Simulation of Sucrose Solutions near the Glass Transition Temperature, J. Phys. Chem. A. 105 (2001) 734–742.

[213] A. Lerbret, P. Bordat, F. Affouard, M. Descamps, F. Migliardo, How homogeneous are the trehalose, maltose, and sucrose water solutions? An insight from molecular dynamics simulations, J. Phys. Chem. B. 109 (2005) 11046–11057.

[214] R.D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Crystallogr. Sect. A. 32 (1976) 751–767.

[215] H. Bianchi, P.R. Tremaine, Thermodynamics of aqueous phosphate solutions: Apparent molar heat capacities and volumes of the sodium and tetramethylammonium salts at 25°C, J. Solution Chem. 24 (1995) 439–463.

[216] R. Majumdar, P. Manikwar, J.M. Hickey, H.S. Samra, H.A. Sathish, S.M. Bishop, C.R. Middaugh, D.B. Volkin, D.D. Weis, Effects of salts from the Hofmeister series on the conformational stability, aggregation propensity, and local flexibility of an IgG1 monoclonal antibody, Biochemistry. 52 (2013) 3376–3389.

[217] C. Ebel, H. Eisenberg, R. Ghirlando, Probing protein-sugar interactions., Biophys. J. 78 (2000) 385–93.

[218] A. Ben-Naim, Solvent-induced interactions: Hydrophobic and hydrophilic phenomena, J. Chem. Phys. 90 (1989) 7412.

[219] D. Awotwe-Otoo, C. Agarabi, E.K. Read, S. Lute, K.A. Brorson, M.A. Khan,

R.B. Shah, Impact of controlled ice nucleation on process performance and quality attributes of a lyophilized monoclonal antibody., Int. J. Pharm. 450 (2013) 70–8.

[220] N. Jovanović, A. Bouchard, G.W. Hofland, G.J. Witkamp, D.J.A. Crommelin, W. Jiskoot, Distinct effects of sucrose and trehalose on protein stability during supercritical fluid drying and freeze-drying, Eur. J. Pharm. Sci. 27 (2006) 336–345.

[221] G.A. Hudalla, T. Sun, J.Z. Gasiorowski, H. Han, Y.F. Tian, A.S. Chong, J.H. Collier, Gradated assembly of multiple proteins into supramolecular nanomaterials., Nat. Mater. 13 (2014) 829–836.

[222] E. Sahin, K.L. Kiick, Macromolecule-induced assembly of coiled-coils in alternating multiblock polymers., Biomacromolecules. 10 (2009) 2740–9.

[223] A. Top, K.L. Kiick, C.J. Roberts, Modulation of self-association and subsequent fibril formation in an alanine-rich helical polypeptide., Biomacromolecules. 9 (2008) 1595–603.

[224] A. Ben-Naim, Theoretical aspects of self-assembly of proteins: a Kirkwood-Buff-theory approach., J. Chem. Phys. 138 (2013) 224906.

[225] D.S. Tomar, V. Weber, B.M. Pettitt, D. Asthagiri, Importance of Hydrophilic Hydration and Intramolecular Interactions in the Thermodynamics of Helix-Coil Transition and Helix-Helix Assembly in a Deca-Alanine Peptide., J. Phys. Chem. B. 120 (2016) 69–76.

[226] R.S. Farmer, L.M. Argust, J.D. Sharp, K.L. Kiick, Conformational Properties of Helical Protein Polymers with Varying Densities of Chemically Reactive Groups., Macromolecules. 39 (2006) 162–170.

[227] R.S. Farmer, A. Top, L.M. Argust, S. Liu, K.L. Kiick, Evaluation of conformation and association behavior of multivalent alanine-rich polypeptides., Pharm. Res. 25 (2008) 700–8.

[228] R.S. Farmer, K.L. Kiick, Conformational behavior of chemically reactive alanine-rich repetitive protein polymers., Biomacromolecules. 6 (2005) 1531–9.

[229] C. Calero-Rubio, B. Paik, X. Jia, K.L. Kiick, C.J. Roberts, Predicting unfolding thermodynamics and stable intermediates for alanine-rich helical peptides with the aid of coarse-grained molecular simulation, Biophys. Chem. 217 (2016) 8–19.

[230] J. Yin, D. Landau, Massively parallel Wang–Landau sampling on multiple GPUs, Comput. Phys. Commun. 183 (2012) 1568–1573.

[231] F. Wang, D. Landau, Efficient, multiple-range random walk algorithm to calculate the density of states, Phys. Rev. Lett. 86 (2001) 2050–2053.

[232] A. Ferrenberg, R. Swendsen, Optimized monte carlo data analysis, Phys. Rev. Lett. 6 (1989) 1195–1198.

[233] A. Ferrenberg, R. Swendsen, New Monte Carlo technique for studying phase transitions, Phys. Rev. Lett. 61 (1988) 2635–2638.

[234] A. Kadoura, S. Sun, A. Salama, Accelerating Monte Carlo molecular simulations by reweighting and reconstructing Markov chains: Extrapolation of canonical ensemble averages and second derivatives to different temperature and density conditions, J. Comput. Phys. 270 (2014) 70–85.

[235] J.M. Andrews, C.J. Roberts, Non-native aggregation of α-chymotrypsinogen occurs through nucleation and growth with competing nucleus sizes and negative activation energies, Biochemistry. 46 (2007) 7558–7571.

[236] D.P. Yee, K.A. Dill, Families and the structural relatedness among globular proteins., Protein Sci. 2 (1993) 884–99.

[237] J.M. Scholtz, H.H. Qian, E.J. York, J.M. Stewart, R.L. Baldwin, Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water., Biopolymers. 31 (1991) 1463–70.

[238] P.L. Privalov, N.N. Khechinashvili, A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study., J. Mol. Biol. 86 (1974) 665–84.

[239] J.M. Scholtz, S. Marqusee, R.L. Baldwin, E.J. York, J.M. Stewart, M. Santoro, D.W. Bolen, Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water., Proc. Natl. Acad. Sci. 88 (1991) 2854–2858.

[240] D. Stigter, D.O. Alonso, K.A. Dill, Protein stability: electrostatics and compact denatured states., Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 4176–80.

[241] K. a Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem., Annu. Rev. Biophys. 37 (2008) 289–316.

[242] E.I. Lin, M.S. Shell, Can peptide folding simulations provide predictive

information for aggregation propensity?, J. Phys. Chem. B. 114 (2010) 11899–908.

[243] C.M. Dobson, Principles of protein folding, misfolding and aggregation., Semin. Cell Dev. Biol. 15 (2004) 3–16.

[244] R.R. Goluguri, J.B. Udgaonkar, Rise of the helix from a collapsed globule during the folding of monellin., Biochemistry. 54 (2015) 5356–5365.

[245] R.J. Kennedy, K.Y. Tsang, D.S. Kemp, Consistent helicities from CD and template t/c data for N-templated polyalanines: Progress toward resolution of the alanine helicity problem, J. Am. Chem. Soc. 124 (2002) 934–944.

[246] A. Chakrabartty, T. Kortemme, R.L. Baldwin, Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions., Protein Sci. 3 (1994) 843–852.

[247] J.M. Andrews, W.F. Weiss IV, C.J. Roberts, Nucleation, growth, and activation energies for seeded and unseeded aggregation of α-chymotrypsinogen A, Biochemistry. 47 (2008) 2397–2403.

[248] G. V. Barnett, M. Drenski, V. Razinkov, W.F. Reed, C.J. Roberts, Identifying protein aggregation mechanisms and quantifying aggregation rates from combined monomer depletion and continuous scattering, Anal. Biochem. 511 (2016) 80–91.

Appendix A

ADDITIONAL INFORMATION FOR ACGN CG MODELING

A.1 Electrostatic Potential of Mean Force models

Electrostatic interaction were modeled using a dipole model reproduced from Bratko *et al*. The electrostatic interaction model was divided into three contributions: monopole-monopole (u_{qq}), monopole-dipole ($u_{q\mu}$), and dipole-dipole ($u_{\mu\mu}$) interactions as shown in equations A.1 to A.7:

$$u_{qq}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon\epsilon_o r_{ij}} S_0(r_{ij}, \kappa) \tag{A.1}$$

$$u_{q\mu}(r_{ij}, \theta_j) = \frac{q_i \mu_j \cos(\theta_j)}{4\pi\epsilon\epsilon_o r_{ij}^2} S_1(r_{ij}, \kappa) \tag{A.2}$$

$$u_{\mu\mu}(r_{ij}, \theta_i, \theta_j, \phi) = \frac{\mu_i \mu_j \left[2S_2(r_{ij}, \kappa)\cos(\theta_i)\cos(\theta_j) - S_3(r_{ij}, \kappa)\sin(\theta_i)\sin(\theta_j)\cos(\phi)\right]}{4\pi\epsilon\epsilon_o r_{ij}^3} \tag{A.3}$$

$$S_0(r_{ij}, \kappa) = \frac{\exp\left[-\kappa(r_{ij} - \sigma)\right]}{\left(1 + \frac{\kappa\sigma}{2}\right)^2} \tag{A.4}$$

$$S_1(r_{ij}, \kappa) = \frac{3\exp\left[-\kappa(r_{ij} - \sigma)\right](1 + \kappa r_{ij})}{\left(1 + \frac{\kappa\sigma}{2}\right)\left[2 + \kappa\sigma + \frac{(\kappa\sigma)^2}{4} + \left(1 + \frac{\kappa\sigma}{2}\right)\frac{\epsilon_p}{\epsilon}\right]} \tag{A.5}$$

$$S_2(r_{ij}, \kappa) = \frac{9\exp\left[-\kappa(r_{ij} - \sigma)\right]\left[1 + \kappa r_{ij} + \frac{1}{2}(\kappa r_{ij})^2\right]}{\left[2 + \kappa\sigma + \frac{(\kappa\sigma)^2}{4} + \left(1 + \frac{\kappa\sigma}{2}\right)\frac{\epsilon_p}{\epsilon}\right]^2} \tag{A.6}$$

$$S_3(r_{ij},\kappa) = \frac{9\exp\left[-\kappa\left(r_{ij}-\sigma\right)\right]\left[1+\kappa r_{ij}\right]}{\left[2+\kappa\sigma+\frac{(\kappa\sigma)^2}{4}+\left(1+\frac{\kappa\sigma}{2}\right)\frac{\epsilon_p}{\epsilon}\right]^2} \tag{A.7}$$

Bratko *et al.* derived a simplified version for both equations A.2 and A.3 by orientationally averaging the monopole-dipole and dipole-dipole interactions as shown in equaitons 9 to 22 of reference [76], and reproduced here in equations A.8 to A.12 as shown below:

$$u_{el}(r_{ij}) = u_{qq}(r_{ij}) - \bar{u}_{q,\mu}(r_{ij}) \tag{A.8}$$

$$\frac{\bar{u}_{q,\mu}(r_{ij})}{\mathrm{k_B T}} = \ln[4+4\cosh(\alpha_3)+\exp(2\alpha_2)+8\cosh(\alpha_1)+\exp(-2\alpha_2)\cos(2\alpha_1)] + 2\ln\left[\frac{\alpha_1^{-1}\sinh(\alpha_1)}{2+\cosh(\alpha_1)}\right] - \ln[2] \tag{A.9}$$

$$\alpha_1 = \frac{\beta q_i \mu_j}{4\pi\epsilon\epsilon_o r_{ij}^2} S_1(r_{ij},\kappa) \tag{A.10}$$

$$\alpha_2 = \frac{\beta \mu_i \mu_j}{4\pi\epsilon\epsilon_o r_{ij}^3} S_2(r_{ij},\kappa) \tag{A.11}$$

$$\alpha_3 = \frac{\beta \mu_i \mu_j}{4\pi\epsilon\epsilon_o r_{ij}^3} S_3(r_{ij},\kappa) \tag{A.12}$$

In equations A.1 to A.12, q_i and q_j represent the charges of particles *i* and *j*, respectively. Similarly, μ_i and μ_j represents the dipole moments of particles *i* and *j*. ϵ and ϵ_p are the relative permittivity of the medium and protein/particle, respectively, while ϵ_o represents the permittivity of vacuum. The values used in Chapter 4 for ϵ, ϵ_p and ϵ_o were 80, 4, and 8.85 x 10^{-12} C^2 N^{-1} m^{-2}, respectively. r_{ij} is the center-to-center of mass distance

between particles *i* and *j*, and θ_i, θ_j and ϕ represent the relative angles between the same pair of particles as shown in Figure A.1. k_B is Boltzmann's constant, T is the absolute temperature and $\beta = (k_B T)^{-1}$. $1/\kappa$ is the screening length and is related to the total ionic strength (*TIS* in mM) by the equation $\kappa = 0.10435 \cdot (TIS/\text{mM})^{0.5}$ at 298.15 K. $\bar{u}_{q,\mu}$ is the orientational-average monopole-dipole and dipole-dipole interactions ($u_{q\mu}$ and $u_{\mu\mu}$). In Chapter 4, the terms Q_{eff} and μ_{eff} were used as replacements for q_i and q_j, and μ_i and μ_j, respectively, for all spherical (colloidal) models since all simulated particles were given the same charge and dipole moment, so there was no distinction between the properties of particle *i* and *j*.

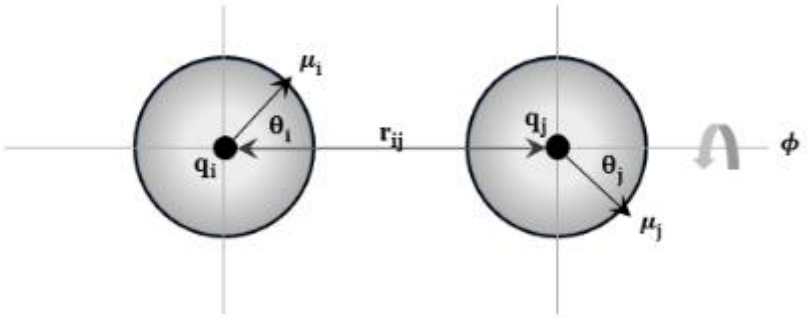

Figure A.1. Representation of interacting particle (adopted from Bratko *et al.*).

A.2 ARD Surface Response Plots for Individual *TIS* Values for pH 5 and 7

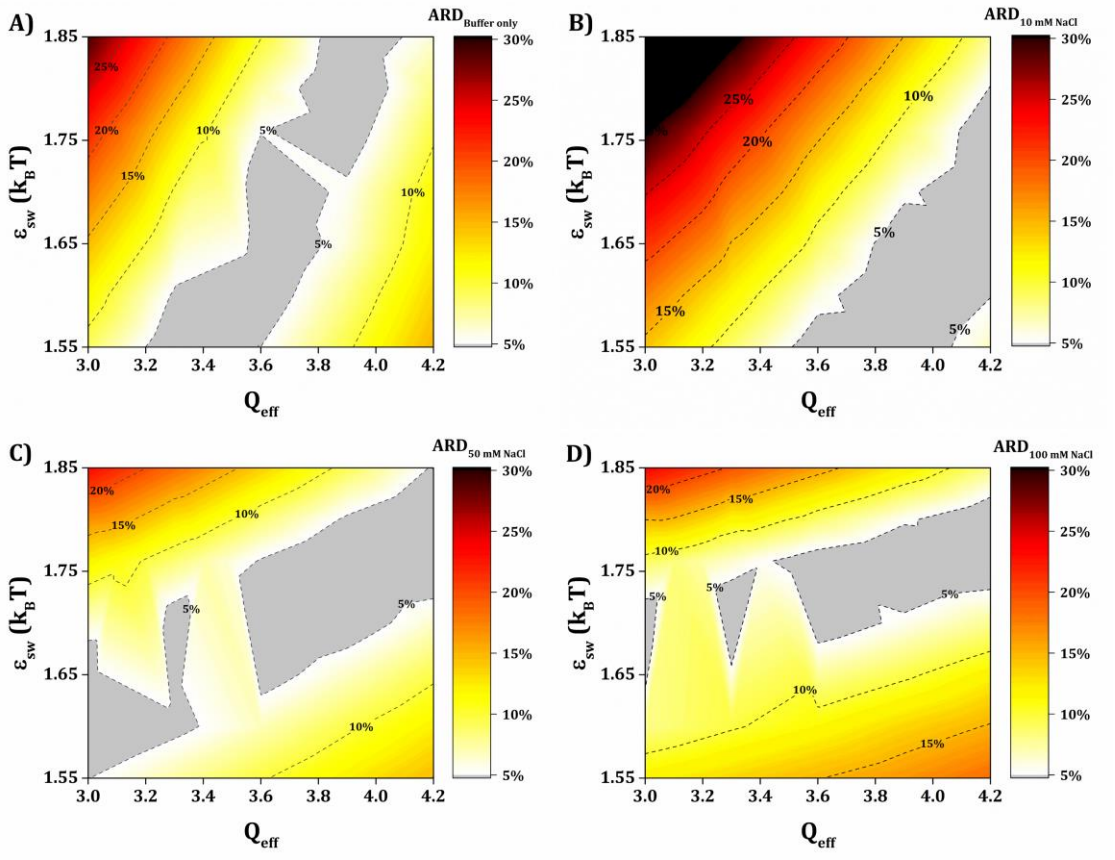

Figure A.2. Contour plots of ARD between experimental and predicted R^{ex}/K over all c_2 values for individual *TIS* values at pH 5 as follows: buffer only or *TIS* = 20 mM (panel A), 10 mM NaCl or *TIS* = 30 mM (panel B), 50 mM NaCl or *TIS* = 70 mM (panel C) and 100 mM NaCl or *TIS* = 120 mM (panel D). The gray area corresponds to ARD values below 5%.

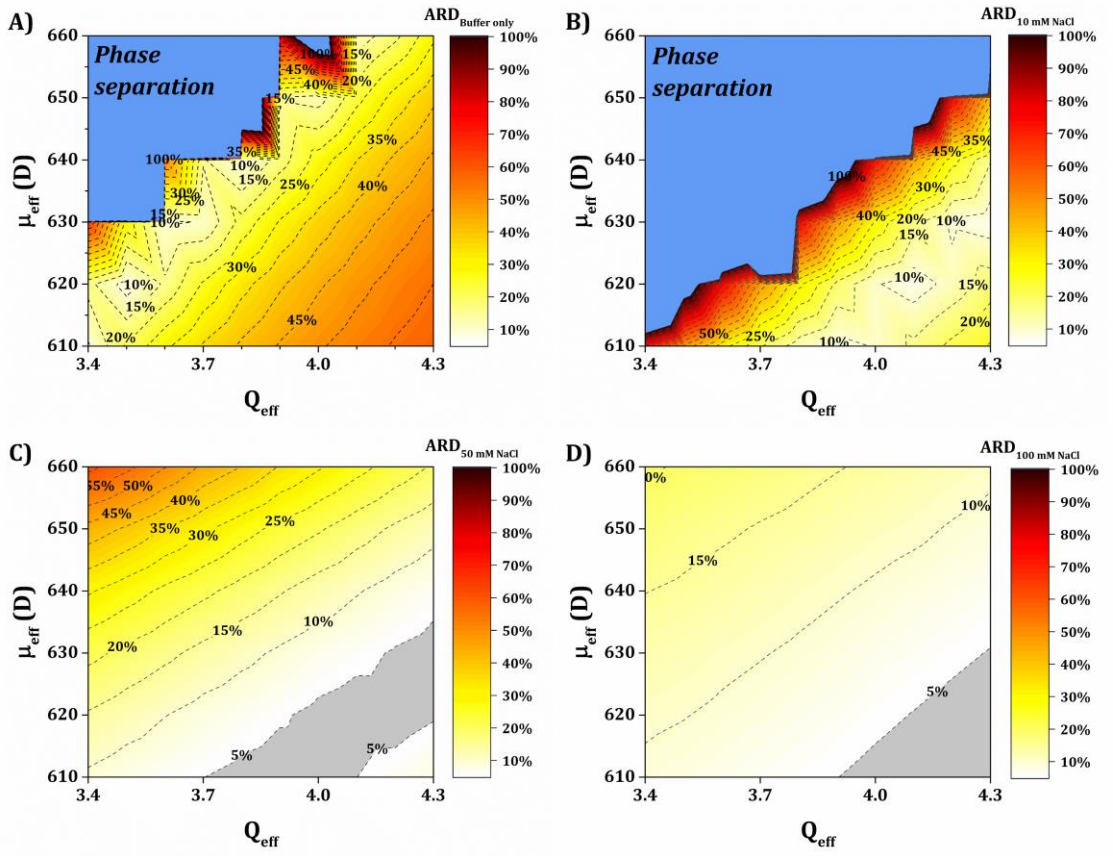

Figure A.3. Contour plots of ARD between experimental and predicted R^{ex}/K over all c_2 values and individual *TIS* values at pH 7 as follows: buffer only or *TIS* = 10 mM (panel A), 10 mM NaCl or *TIS* = 20 mM (panel B), 50 mM NaCl or *TIS* = 60 mM (panel C) and 100 mM NaCl or *TIS* = 110 mM (panel D). The gray area corresponds to ARD values below 5%.

Appendix B

DERIVATION OF PARTIAL SPECIFIC VOLUMES FOR MULTI-COMPONENT SOLUTIONS FROM THE PERSPECTIVE OF INVERSE KIRKWOOD-BUFF SOLUTION THEORY

Throughout the following derivations, $B_{(n)}$ is used to define an n-dimensional symmetric matrix with components $b_{(n)}^{ij}$ defined in equation 5.3, and whose cofactor matrix $C_{(n)}$ has components $c_{(n)}^{ij}$. The term $B_{(n\text{-}1|k)}$ will represent a (n-1)-dimensional matrix that is derived from the previous $B_{(n)}$ matrix by deleting component k (deleting the kth row and column) and further rearranging the matrix as is done in the calculation of cofactors. Similarly, $b_{(n-1|k)}^{ij}$represents the i-jth component of this new $B_{(n\text{-}1|k)}$ matrix. This new matrix will also have a new cofactor matrix $C_{(n\text{-}1|k)}$ with components $c_{(n-1|k)}^{ij}$. As an example, a system with n = 5 will be used, but the following derivation can be applied to any number of components. In this case, the $B_{(5)}$ matrix would be represented as in equation B.1, while the cofactor matrix $C_{(5)}$ is shown in equation B.2:

$$B_{(5)} = \begin{pmatrix} b_{(5)}^{11} & b_{(5)}^{12} & b_{(5)}^{13} & b_{(5)}^{14} & b_{(5)}^{15} \\ b_{(5)}^{21} & b_{(5)}^{22} & b_{(5)}^{23} & b_{(5)}^{24} & b_{(5)}^{25} \\ b_{(5)}^{31} & b_{(5)}^{32} & b_{(5)}^{33} & b_{(5)}^{34} & b_{(5)}^{35} \\ b_{(5)}^{41} & b_{(5)}^{42} & b_{(5)}^{43} & b_{(5)}^{44} & b_{(5)}^{45} \\ b_{(5)}^{51} & b_{(5)}^{52} & b_{(5)}^{53} & b_{(5)}^{54} & b_{(5)}^{55} \end{pmatrix}, \tag{B.1}$$

$$C_{(5)} = \begin{pmatrix} c_{(5)}^{11} & c_{(5)}^{12} & c_{(5)}^{13} & c_{(5)}^{14} & c_{(5)}^{15} \\ c_{(5)}^{21} & c_{(5)}^{22} & c_{(5)}^{23} & c_{(5)}^{24} & c_{(5)}^{25} \\ c_{(5)}^{31} & c_{(5)}^{32} & c_{(5)}^{33} & c_{(5)}^{34} & c_{(5)}^{35} \\ c_{(5)}^{41} & c_{(5)}^{42} & c_{(5)}^{43} & c_{(5)}^{44} & c_{(5)}^{45} \\ c_{(5)}^{51} & c_{(5)}^{52} & c_{(5)}^{53} & c_{(5)}^{54} & c_{(5)}^{55} \end{pmatrix}. \tag{B.2}$$

For simplicity, one can focus on component $\alpha = 1$, where the $B_{(4|1)}$ and $C_{(4|1)}$ matrices are written in equations B.3 and B.4. One needs to be cautious that the equality in equation B.3 does not apply to equation B.4 as the latter is a cofactor matrix and its components are derived from the rearranged $B_{(4|1)}$, not from $B_{(5)}$ or for that matter, $C_{(5)}$.

$$B_{(4|1)} = \begin{pmatrix} b_{(4|1)}^{11} & b_{(4|1)}^{12} & b_{(4|1)}^{13} & b_{(4|1)}^{14} \\ b_{(4|1)}^{21} & b_{(4|1)}^{22} & b_{(4|1)}^{23} & b_{(4|1)}^{24} \\ b_{(4|1)}^{31} & b_{(4|1)}^{32} & b_{(4|1)}^{33} & b_{(4|1)}^{34} \\ b_{(4|1)}^{41} & b_{(4|1)}^{42} & b_{(4|1)}^{43} & b_{(4|1)}^{44} \end{pmatrix} = \begin{pmatrix} b_{(5)}^{22} & b_{(5)}^{23} & b_{(5)}^{24} & b_{(5)}^{25} \\ b_{(5)}^{32} & b_{(5)}^{33} & b_{(5)}^{34} & b_{(5)}^{35} \\ b_{(5)}^{42} & b_{(5)}^{43} & b_{(5)}^{44} & b_{(5)}^{45} \\ b_{(5)}^{52} & b_{(5)}^{53} & b_{(5)}^{54} & b_{(5)}^{55} \end{pmatrix}, \tag{B.3}$$

$$C_{(4|1)} = \begin{pmatrix} c_{(4|1)}^{11} & c_{(4|1)}^{12} & c_{(4|1)}^{13} & c_{(4|1)}^{14} \\ c_{(4|1)}^{21} & c_{(4|1)}^{22} & c_{(4|1)}^{23} & c_{(4|1)}^{24} \\ c_{(4|1)}^{31} & c_{(4|1)}^{32} & c_{(4|1)}^{33} & c_{(4|1)}^{34} \\ c_{(4|1)}^{41} & c_{(4|1)}^{42} & c_{(4|1)}^{43} & c_{(4|1)}^{44} \end{pmatrix}. \tag{B.4}$$

For $C_{(5)}$, its component $c_{(5)}^{11}$ can be written as:

$$c_{(5)}^{11} = \left|B_{(4|1)}\right|, \tag{B.5}$$

where $|B_{(4|1)}|$ represents the determinant of the 4-dimentional $B_{(4|1)}$ matrix shown in equation B.3. Next, the cofactors $c_{(5)}^{21}$ and $c_{(5)}^{31}$ can be written as:

$$c_{(5)}^{21} = -\left[b_{(5)}^{12}\, c_{(4|1)}^{22} + \left(-b_{(5)}^{13}\right)\left(-c_{(4|1)}^{23}\right) + b_{(5)}^{14} c_{(4|1)}^{24} + \left(-b_{(5)}^{15}\right)\left(-c_{(4|1)}^{25}\right)\right], \tag{B.6}$$

$$c_{(5)}^{31} = b_{(5)}^{12}\left(-c_{(4|1)}^{32}\right) + \left(-b_{(5)}^{13}\right) c_{(4|1)}^{33} + \left(-b_{(5)}^{14}\right) c_{(4|1)}^{34} + b_{(5)}^{15}\left(-c_{(4|1)}^{35}\right). \tag{B.7}$$

These two cofactors can be simplified as follows:

$$c_{(5)}^{21} = -\left(b_{(5)}^{12} c_{(4|1)}^{22} + b_{(5)}^{13} c_{(4|1)}^{23} + b_{(5)}^{14} c_{(4|1)}^{24} + b_{(5)}^{15} c_{(4|1)}^{25}\right) = -\sum_{j \neq 1}^{5} b_{(5)}^{1j} c_{(4|1)}^{2j}, \tag{B.8}$$

$$c_{(5)}^{31} = -\left(b_{(5)}^{12}c_{(4|1)}^{32} + b_{(5)}^{13}c_{(4|1)}^{33} + b_{(5)}^{14}c_{(4|1)}^{34} + b_{(5)}^{15}c_{(4|1)}^{35}\right) = -\sum_{j\neq 1}^{5} b_{(5)}^{1j}c_{(4|1)}^{3j}. \qquad \text{(B.9)}$$

This process can be repeated for the remaining two cofactors, and one obtains that for any $i \neq 1$,

$$c_{(5)}^{i1} = -\sum_{j\neq 1}^{5} b_{(5)}^{1j}c_{(4|1)}^{ij}. \qquad \text{(B.10)}$$

This can be easily generalized for any other component α. Similarly, this can be extended to any value of n since one only needs to keep adding B_{1i} components to the summation in equation B.10. Consequently, any cofactor $c_{(n)}^{i\alpha}$ can be generalized as:

$$c_{(n)}^{i\alpha} = \left|B_{(n-1|\alpha)}\right|, \qquad if\ i = \alpha \qquad \text{(B.11)}$$

$$c_{(n)}^{i\alpha} = -\sum_{j\neq\alpha}^{n-1} b_{(n)}^{\alpha j}c_{(n-1|\alpha)}^{ij}, \qquad if\ i \neq \alpha \qquad \text{(B.12)}$$

It is useful to remember that computing $c_{(n-1|\alpha)}^{ij}$ involves a decrease in dimensionality from n to n-1, and further rearrangement of the $B_{(n\text{-}1)}$, which is responsible for the minus (-) sign in front of the summation in equation B.12. These two final expressions are useful for the definition of β_α in equation 4.5:

$$\beta_\alpha = \sum_{i}^{n} \rho_i c_{(n)}^{i\alpha} = \rho_\alpha \left|B_{(n-1|\alpha)}\right| - \sum_{i\neq\alpha}^{n-1} \rho_i \sum_{j\neq\alpha}^{n-1} b_{(n)}^{\alpha j}c_{(n-1|\alpha)}^{ij}. \qquad \text{(B.13)}$$

Equation 5.3 can be used for the formal definition of $b_{(n)}^{\alpha j}$, while an additional reorganization of the summations on the right-hand side of equation B.13 will result in equation B.14:

$$\beta_\alpha = \rho_\alpha \left|B_{(n-1|\alpha)}\right| - \sum_{i\neq\alpha}^{n-1}\left(\rho_i \sum_{j\neq\alpha}^{n-1} \rho_j \rho_\alpha G_{j\alpha} c^{ij}_{(n-1|\alpha)}\right)$$
$$= \rho_\alpha \left[\left|B_{(n-1|\alpha)}\right| - \sum_{j\neq\alpha}^{n-1}\left(\rho_j G_{j\alpha} \sum_{i\neq\alpha}^{n-1} \rho_i c^{ij}_{(n-1|\alpha)}\right)\right]. \quad \text{(B.14)}$$

Until this stage, no assumptions regarding the condition of any component have been made, so equation B.14 is an exact expansion that applies to any system regardless of its size and composition. However, this equation alone is not enough to obtain an expression for $\overline{V}_\alpha$, as equation 5.6 needs to be simplified as well. For this, it is easier to note a trend for $B_{(5)}$ if one explicitly displays its components in terms of densities and KB integrals and uses the former derivation:

$$B_{(5)} = \begin{pmatrix} \rho_1^2 G_{11} + \rho_1 & \rho_1\rho_2 G_{12} & \rho_1\rho_3 G_{13} & \rho_1\rho_4 G_{14} & \rho_1\rho_5 G_{15} \\ \rho_1\rho_2 G_{21} & \rho_2^2 G_{22} + \rho_2 & \rho_2\rho_3 G_{23} & \rho_2\rho_4 G_{24} & \rho_2\rho_5 G_{25} \\ \rho_1\rho_3 G_{31} & \rho_2\rho_3 G_{32} & \rho_3^2 G_{33} + \rho_3 & \rho_3\rho_4 G_{34} & \rho_3\rho_5 G_{35} \\ \rho_1\rho_4 G_{41} & \rho_2\rho_4 G_{42} & \rho_3\rho_4 G_{43} & \rho_4^2 G_{44} + \rho_4 & \rho_4\rho_5 G_{45} \\ \rho_1\rho_5 G_{51} & \rho_2\rho_5 G_{52} & \rho_3\rho_5 G_{53} & \rho_4\rho_5 G_{54} & \rho_5^2 G_{55} + \rho_5 \end{pmatrix}. \quad \text{(B.15)}$$

Using the definition of cofactors, one could generalize the definition of η (equation 5.6) for any given component *k* as follows:

$$\eta = (\rho_k^2 G_{kk} + \rho_k) * \sum_{i\neq k}^{n-1}\sum_{j\neq k}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|\alpha)} + \rho_k^2 * f_{cross} \quad \text{(B.16)}$$

The subscript *k* in equation B.16 has been used instead of α to indicate that this expansion for η can be achieved independently from equation B.14. The first term on the right-hand side of equation B.16 represents the sum over all the cofactors that will include the $(\rho_k^2 G_{kk} + \rho_k)$ term from the $b^{kk}_{(n)}$ component, like the way equation B.11

was derived. Any $c^{ij}_{(n-1|k)}$ cofactor with the prefactor $b^{kk}_{(n)}$will eliminate any *k* component from its own calculation. Therefore, the double summation is done over all the components except *k*. The second term on the right-hand side will involve all the other cofactors that include cross interactions with the *k*th component. Interestingly, this will result in a ρ_k^2 prefactor, which arises from one ρ_k from any $b^{ik}_{(n)}$ (equation 5.3) times a ρ_k from the summation on equation 5.6. This term has been denoted f_{cross} as it contains all the terms involving products of different G_{ik} terms multiplied by different ρ_i and ρ_k values. This is equivalent to multibody cross interactions when component *k* is concentrated enough so *k*-*k* interactions contribute to the solutions and affect the otherwise "independent" *i*-*j*, *k*-*i* and *k*-*j* interactions simultaneously. A simplified form for f_{cross} is yet known, but it is not needed for what follows below. Finally, equation B.16 can be rewritten as shown in equation B.17.

$$\eta = \rho_k \left[\sum_{i\neq k}^{n-1} \sum_{j\neq k}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|k)} + \rho_k \left(G_{kk} \sum_{i\neq k}^{n-1} \sum_{j\neq k}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|k)} + f_{cross} \right) \right] \qquad \text{(B.17)}$$

Once again, no assumption about the concentration of any component has been made, so this expression equally applies to any system. By combining both equations B14 and B.17, one can obtain an updated version for equation 5.4.

$$\overline{V}_\alpha = \frac{\rho_\alpha \left[\left| B_{(n-1|\alpha)} \right| - \sum_{j\neq\alpha}^{n-1} \left(\rho_j G_{j\alpha} \sum_{i\neq\alpha}^{n-1} \rho_i c^{ij}_{(n-1|\alpha)} \right) \right]}{\rho_k \left[\sum_{i\neq k}^{n-1} \sum_{j\neq k}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|k)} + \rho_k \left(G_{kk} \sum_{i\neq k}^{n-1} \sum_{j\neq k}^{n-1} \rho_i \rho_j c^{ij}_{(n-1|k)} + f_{cross} \right) \right]} \qquad \text{(B.18)}$$

It is worth to point out that no limitations exist for α and *k*. Consequently, they can take any value, including the one used in the main body of $k = \alpha$ as both derivations were obtained independent of each other.

Appendix C

REPRINT PERMISSION LETTERS

Title: Predicting unfolding thermodynamics and stable intermediates for alanine-rich helical peptides with the aid of coarse-grained molecular simulation

Author: Cesar Calero-Rubio,Bradford Paik,Xinqiao Jia,Kristi L. Kiick,Christopher J. Roberts

Publication: Biophysical Chemistry

Publisher: Elsevier

Date: October 2016

Title: Relating Protein–Protein Interactions and Aggregation Rates From Low to High Concentrations

Author: Ranendu Ghosh,Cesar Calero-Rubio,Atul Saluja,Christopher J. Roberts

Publication: Journal of Pharmaceutical Sciences

Publisher: Elsevier

Date: March 2016