Computational Methods for the Study of Immunoglobulin Aggregation



David Shorthouse

Department of Biochemistry

University of Oxford

This dissertation is submitted for the degree of

Doctor of Philosophy

Wolfson College

April 2015

Acknowledgements

Primary thanks go to Professor Mark Sansom, not only for giving me the opportunity to work in his lab, but for his exceptional support and encouragement throughout the project. A huge thanks also to my colleagues at Lonza, especially Thomas Gallagher who, although only being involved in the project for a short time, has made a huge impact on the direction of the project, and has acted as my mentor in all things relating to antibodies.

Closer to home, my friends and colleagues in Oxford deserve praise for their help and encouragement over the past 4 years. My cohort, Greg, Jemma, and Lukas, who all started together, learned together, struggled together, and (hopefully), will all finish succesfully. Acknowledgement must be made to those who helped me in the early stages of my study, when it took me an inordinately long time for me to get my feet - Ben, Antreas, both Phil(l)s, and Craig were all essential to this. Huge thanks go to my other close friends in the lab, Maria, Amanda, Nat, Erin, Ben, Anna, and Matteo who have made my time in the SBCB so enjoyable. A special mention must be made of Jo, who kept me sane during thesis writing with regular exercise by sharing my love of climbing, and by battering me around the squash court every week. Finally, and importantly - Heidi, whose support has shown me that I can in fact, do science, which means a great deal to me.

Finally, a huge thanks to my partner Annie, who has encouraged me throughout the project, and supported me through the stressful time of writing this thesis.

Thank you all.

Abstract

Protein aggregation is a major challenge in the development of antibody-based therapeutics. Therapeutic antibodies are produced and stored in high concentrations and under fluctuating conditions unfavourable for their stability. Aggregation of these proteins in solution leads to serious consequences for patients, with the initiation of immune reactions, which have the potential to be fatal, and in the loss of clinical potency. The types of aggregates formed by antibodies, and the processes that lead to their propagation are poorly understood. By studying these molecules via computational approaches, we are able to simulate and probe their tendency to aggregate on experimentally comparable timescales.

By performing small numbers of coarse grained simulations of immunoglobulin fragments it is shown that specific regions of proteins are involved in self-self interactions, and these regions are targets for reducing the self-association of experimental molecules. Techniques developed here are integrated within a high throughput approach that is able to generate information on aggregation for a large number of candidate antibody structures.

The methodology was refined via development of a novel technique for coarse grained simulations of oligosaccharides. This method was initially tested on glycolipids, and then extended to glycoproteins. The primary outcome is a coarse grained model for a glycosylated antibody Fc fragment. The glycosylated Fc was then simulated, and compared to experimental data. Coarse grained simulations support the hypothesis that the protein becomes more flexible in the absence of glycosylation.

Table of contents

Table of contents vii

List of figures xiii

List of tables xvii

1 Introduction 1

1.1 Antibodies 2

1.1.1 The Evolutionary History of Antibodies 2

1.1.2 Antibodies as Biological Molecules 6

1.1.3 Pharmacological Relevance of Monoclonal Antibodies 9

1.1.4 Aggregation of Antibodies 9

1.1.5 Physical Approaches to Antibody Aggregation 11

1.1.6 Computational Approaches to Antibody Aggregation 12

1.2 Carbohydrates 13

1.2.1 Carbohydrates as Biological molecules 13

1.2.2 Glycolipids 16

1.2.3 Protein Glycosylation 17

1.2.4 Current Approaches to Simulations of Carbohydrates 18

1.3 Thesis Overview 20

2 Materials and Methods **23**

2.1 Classical Molecular Dynamics 24

2.1.1 Forcefields 25

2.1.2 Bonded Interactions 26

2.1.3 Nonbonded Interactions 29

2.1.4 Classical Forcefield Development 31

2.2 Simulation Techniques 32

2.2.1 Timesteps 32

2.2.2 Periodicity and Periodic Boundary Conditions 33

2.2.3 Temperature and Pressure Coupling 35

2.2.4 The Berendsen Thermostat 35

2.2.5 The Berendsen Barostat 36

2.2.6 Particle Mesh Ewald 36

2.2.7 Particle Exclusions 37

2.2.8 Constraints 37

2.2.9 Position Restraints 38

2.3 Energy Minimization 38

2.4 Coarse Grained Forcefields 39

2.4.1 MARTINI lipids 40

2.4.2 MARTINI Extension to Proteins 43

2.4.3 Sweet MARTINI 45

2.4.4 MARTINI v 2.2 48

2.4.5 MARTINI Glycolipids 48

2.4.6 MARTINI Secondary Structure Restraints 49

2.5 Markov State Modelling 50

2.6 Software Used for Simulations 51

3 Simulations of Antibody Aggregation 53

3.1 Introduction . . . 53

3.1.1 Antibody Aggregation . . . 53

3.1.2 Previous Work . . . 54

3.1.3 Coarse Grained Molecular Dynamics . . . 55

3.2 Methods . . . 57

3.3 Results and Discussion . . . 61

3.3.1 Martini v 2.1 vs Martini v 2.2 - comparing forcefields . . . 61

3.3.2 Simulations of Pairs of Proteins . . . 63

3.3.3 Simulations of Larger Scale Systems . . . 78

3.3.4 Discussion . . . 79

4 Novel Methodologies for the Study of Antibody Aggregation 87

4.1 Introduction . . . 88

4.1.1 High Throughput Simulations . . . 88

4.1.2 Markov State Models . . . 88

4.1.3 Chapter Aims . . . 89

4.1.4 The Lonza Protein Dataset . . . 89

4.2 Methods . . . 93

4.2.1 High Throughput Simulations . . . 93

4.2.2 Simulation Setup . . . 93

4.2.3 Simulation Analysis . . . 95

4.2.4 Markov State Modelling . . . 98

4.3 Results and Discussion . . . 102

4.3.1 Pipeline Performance . . . 104

4.3.2 Results from Regional Analysis . . . 104

4.3.3 Study of Association Rates . . . 113

4.3.4 Markov State Models 113
4.3.5 Discussion 115

5 Coarse Grained Modelling of Oligosaccharides 121
5.1 Introduction 122
5.1.1 Carbohydrates 122
5.1.2 Cerebrosides 122
5.1.3 Gangliosides 123
5.1.4 Previous Work 126
5.2 Methods 129
5.2.1 Atomistic Simulations of Sugar Molecules 129
5.2.2 Coarse Grained Simulations of Sugar Molecules 130
5.2.3 Iterative Force-Matching 131
5.2.4 Simulations of Membrane Systems 132
5.3 Results and Discussion 132
5.3.1 Simulating single Carbohydrates 132
5.3.2 Simulating Disaccharides 139
5.3.3 More complex Carbohydrates 141
5.3.4 Applications of Carbohydrates - Glycolipids 143
5.3.5 Glucocerebroside and Galactocerebroside 146
5.3.6 GM3 147
5.3.7 Forssman Glycolipid 151
5.3.8 Discussion 151

6 Coarse Grained Simulations of Protein Glycosylation 157
6.1 Introduction 158
6.1.1 IgG1 Fc Fragment Glycosylation 158

6.1.2 Previous Computation Studies of Glycosylation 159
6.1.3 Coarse Grained Secondary Structure Restraints 160
6.2 Methods 161
6.2.1 Atomistic Simulations 161
6.2.2 Coarse Grained Simulations 161
6.2.3 Iterative Force Matching 162
6.2.4 Elastic Network Modelling 162
6.3 Results and Discussion 163
6.3.1 Secondary Structure Restraints and Antibody Fragments 163
6.3.2 Development of Glycosylation parameters 167
6.3.3 Structural Studies of the Immunoglobulin G Fc 177
6.3.4 Discussion 187

7 Conclusions and Future Directions **191**
7.1 Major conclusions 191
7.2 Future directions 193

References **197**

Appendix A Appendix **213**
A.1 Script for the Automation of Simulation Setup 213

List of figures

1.1 A brief history of immunoglobulins 5

1.2 Schematic of an immunoglobulin IgG 7

1.3 Generation of monoclonal antibodies 10

1.4 Computational methods for studying protein aggregation 14

1.5 Schematic of a biological membrane 17

1.6 Current computational work on glycolipids 20

2.1 Schematic of the main interactions within a classical molecular dynamics forcefield 27

2.2 Potential box shapes for use with periodic boundary conditions 34

2.3 Schematic of periodic boundary conditions 34

2.4 Schematic of MARTINI lipid parameters 40

2.5 Interaction Table for Martini Lipid Forcefield 43

2.6 Schematic of MARTINI amino acid parameters 46

2.7 Schematic of MARTINI carbohydrate parameters 47

2.8 Schematic of MARTINI glycolipid parameters 49

2.9 Schematic of Markov State Model state assignments 52

3.1 Previous computational studies of Antibodies 56

3.2 Figure showing setup of CG structures and systems 58

3.3 A comparison of Martini 2.1 and Martini 2.2 forcefields when used for soluble protein-protein interaction 62
3.4 Interaction Frequencies for 50 simulations of Fab or Fc self association . . 67
3.5 Cluster analysis for 50 Fab or Fc self association simulations 68
3.6 Interaction Landscape plot for 50 simulations of Fab-Fab self associations . 69
3.7 Specific interactions for 50 simulations of Fab-Fab self associations 70
3.8 Comparison of results for 50 simulations of Fc, with studies in the literature 73
3.9 Interaction landscape plot for 50 simulations of Fc - Fc self associations . . 75
3.10 Specific interactions for 50 simulations of Fc - Fc self associations 76
3.11 Residue type composition for interaction regions on Fab and Fc proteins . . 77
3.12 Comparison of interaction frequencies for different scales of system 80
3.13 Aggregates from simulations containing 8 proteins 81
3.14 Interaction landscape plots for all pairs of protein within simulations of Fab-Fab interactions 82
3.15 Interaction landscape plots for all pairs of proteins in simulations of Fc-Fc interactions 83

4.1 Examples of Simple Markov State Models 90
4.2 Schematic of a High Throughput Methodology for the Study of Antibody Aggregation 91
4.3 Flowchart for the Setup and Simulation of High Throughput Systems 96
4.4 Schematic of Different Analysis Levels Performed by High Throughput System 99
4.5 Schematic showing Euler Angles for Fab Proteins 101
4.6 Schematic for a Markov State Model Constructed with High Throughput Data 103
4.7 Regional Analysis for Lonza dataset Protein H9K7 106
4.8 Regional Analysis for Lonza dataset Protein H8K8 108

4.9 Regional Analysis for Lonza dataset Protein H31K5 109
4.10 Regional Analysis for Lonza dataset Protein H17L1 111
4.11 Regional Analysis for Lonza dataset Protein H33L2 112
4.12 Regional Analysis for Lonza dataset Protein H28L2 114
4.13 Comparison of Experimental Data and Markov State Model Results 116

5.1 Ceramide Structure 123
5.2 Glucocerebroside Structure 124
5.3 GM3 Structure 125
5.4 Forssman Structure 126
5.5 Previous work performed on simulations of carbohydrates 128
5.6 RMSD for a single Glucose molecule in silic 130
5.7 Schematic for Iterative Force Matching of parameters from Atomistic to Coarse Grained structures 133
5.8 Mapping Schemes for various levels of Carobhydrate models 134
5.9 Bonding Schemes for various levels of oligosaccharide models 135
5.10 Parameterization of basic structures 137
5.11 Advanced Paramerization of Basic Sugar Structures 138
5.12 Emergent properties of Maltose 142
5.13 Emergent properties of Tetraglucose 144
5.14 Structures of glycolipids studied 145
5.15 Parameters restrained for the glycosidic linkage of glycolipids 147
5.16 Analysis of emergent properties for GM3 149
5.17 Analysis of the properties of GM3 in a bilayer 150
5.18 Analysis of the emergent properties of the Forssman glycolipid 152
5.19 Analysis of the properties of Forssman in a bilayer 153
5.20 Other work performed using CG glycolipid parameters 155

6.1 Protein glycosylation of Immunoglobulin G 158
6.2 A comparison of secondary structure restraints for Fab proteins within coarse-grained molecule dynamics 165
6.3 A comparison of secondary structure restraints for Fc proteins within coarse-grained molecule dynamics 166
6.4 Sequence and Structure of the Immunoglobulin G Fc glycoform G2F 169
6.5 A Comparison of CG and AT Glycosylated Fc Structures 170
6.6 Atomistic, Schematic Coarse Grained, and Coarse Grained structures of the Immunoglobulin Fc Carbohydrate Molecule 171
6.7 A Comparison of Atomsitic and Coarse Grained Glycosydic Bonds 172
6.8 Force Matching for Immunoglobulin Fc Carbohydrate Molecule 174
6.9 Force Matching for Immunoglobulin Fc Carbohydrate Attachment to Protein 175
6.10 Solvent Accessible Surface Area for Atomistic and Coarse Grained IgG Fc Proteins 178
6.11 Comparison of Glycosylated and Unglycosylated RMSD 180
6.12 Comparison of Glycosylated and Unglycosylated Radius of Gyration 181
6.13 Distance between ASN 297 for Simulations of IgG Fc 182
6.14 Angle of IgG Fc Extended Region in Simulations 184
6.15 Distance between SER 258 residues for Simulations of IgG Fc 185
6.16 Distance between PRO 329 residues for Glycosylated and Unglycosylated IgG Fc Structures 186

List of tables

3.1 Table showing simulations performed for comparison of Martini forcefield types 60
3.2 Table showing simulations performed for investigation of aggregation . . . 60

4.1 Table comparing experimental characteristics of Lonza proteins 92
4.2 Comparison of Association Rate through Frame Analysis to Experimental Aggregation Propensity 115
4.3 Markov Model Results for Lonza dataset Protein Simulations 116

6.1 Simulations performed for comparison of Secondary Structure Restraints . 164
6.2 Simulations performed for comparison of Secondary Structure Restraints, and Immunoglobulin Fc Glycoforms 179

Chapter 1

Introduction

"I not only think we will tamper with Mother Nature, I think Mother wants us too."

Willard Gaylen

Antibodies are the culmination of nearly 500 million years of evolution (Cooper and Alder [33], Litman et al. [116]). They represent a paragon of cell defence against attack. A molecule able to be produced in such massive diversity that it can be relied on to detect almost any invading threat innately, yet specific structures can be manufactured in such huge numbers by the immune system that they can effectively annihilate invasions of the entire human body - by acting as the interface and targeting mechanism of cells involved in immunity, causing little collateral damage (Baumgarth et al. [10]). When antibodies and their unique properties were originally discovered, it was not long before the concept of immunoglobulins as therapeutic agents was born, the concept of antibodies as a "molecular magic bullet" was proposed by a pioneer of immunology, Paul Ehrlich, in the late 19TH century (Strebhardt and Ullrich [183]). The dream of Ehrlich was realised in 1980, when the first patient was successfully treated with a monoclonal antibody solution specific for antigens highly expressed in non-Hodgkins lymphoma (Nadler et al. [132], Oldham and

Dillman [136]). Nowadays monoclonal antibodies are routinely used for the treatment of numerous conditions, with the first FDA approved monoclonal antibody being rituximab in 1997 (Dillman [46]), but the generation rate of novel, approved, structures has plummeted, mostly due to issues with stability and manufacture (Kozlowski and Swann [105]). Proteins raised against particular targets are generally easy to develop, but finding structures that are stable enough to be stored in high concentrations, and given to patients, has proven extremely difficult, despite numerous techniques to make the proteins as immuno-friendly as possible (Roguska et al. [152]).

This thesis focusses on the development of computational techniques for the study and prevention of aggregation of monoclonal antibody populations, and the development of a novel methodology for simulating carbohydrates, with a focus on antibody glycosylation.

1.1 Antibodies

1.1.1 The Evolutionary History of Antibodies

The immune system is a vital part of multicellular organism function. It is estimated that 98% of multicellular life has some form of innate defence against attack by pathogens (Rodríguez et al. [151]). It is likely that the first cellular organisms to evolve billions of years ago would have rapidly gained some form of immune defence as a reaction to the intense competition and quick advancement of early life. This is generally thought to be the generation and production of antimicrobial peptide type enzymes (Li et al. [112]) or ribozymes (Alberts et al. [2]) that conferred evolutionary advantages to cells that were able to destroy and protect against interference from other bodies. Antimicrobial peptides have now been found in every classification of life (Jenssen et al. [92], Tennessen [187]), and although it is likely they have evolved multiple times, this indicates strongly how essential they would have been for the development of early organisms, possibly even before true cellular diver-

sity developed. They are likely the evolutionary ancestors of the true immune system.

The first semblance of a true innate immune system, that is, a system comprising of static and unchanging mechanisms for defence against pathogens, is thought to be the emergence of Pattern Recognition Receptors (PRRs), which are found in almost all classes of life, including plants and all Metazoa (Janeway and Medzhitov [89], Hoffman1999). All organisms exhibiting PRRs contain within them a set of genes encoding for receptors that recognize a variety of extremely specific targets found on other organisms, such as the lipopolysaccharide coating found on pathogenic bacteria (Rosenstiel et al. [156], Nurnberger2004). Upon activation, PRRs generate downstream signals that propagate host defenses, including the production of antimicrobial peptides, and the activation of an inflammatory response. PRRs are extremely evolutionarily conserved, and recognize highly specific pathogenic targets that cannot be confused with self-generated molecules. It is thought that these receptors have evolved to target areas in the structure of a pathogen that are critical for it to perform its function, and so are rarely, or never, able to be evolutionarily mutated in order to evade detection. These extensive defences make up the bulk of the innate immune systems that are shared by almost all forms of life.

A dominant form of PRRs, Toll Like Receptors (TLRs), are shared by all Metazoan life, as they have recently been found in sea sponges (Wiens et al. [200]), an organism at the root of the Metazoan branch of the tree of life, these have been conserved through at least 600 million years of evolution, though they also share amino acid motifs with even more evolutionarily ancient receptors (Hemmrich et al. [72]). Some organisms, stemming from early invertebrates, have since developed the ability to differentially splice PRR genes, and thus gain the ability to recognize a wider range of pathogenic targets, but at the cost of potential self-recognition. These mechanisms mark the first steps in the evolution of an adaptive immune system (Cooper and Alder [33]), and are a precursor to the generation of antibodies, and it is likely that these advancements were necessary for vertebrate animals,

which have long development and maturation times where they are vulnerable to infection, to successfully evolve to the complexity they currently have.

An adaptive immune system is characterized by the recombination and active selection of biological molecules for the explicit purpose of targeting pathogenic organisms or structures. The adaptive immune system as it is known today originally evolved in jawed fish around 500 million years ago (Flajnik and Kasahara [58]), although unjawed fish have recently been shown to undergo a similar process of receptor mutation through convergent evolution (Pancer et al. [141], Alder2005). It is thought that the antibody molecule – the workhorse of the adaptive immune system, evolved from the structure of a cell-cell adhesion molecule, that shares the immunoglobulin fold (Steiner [181]). Antibodies form an integral part of the adaptive immune system, reacting initially to pathogenic molecules and activating classical complement cascades, as well as being spliced, recombinated, and selected for their ability to target specific pathogens, before they are produced in huge quantities by B-cells. The most historically ancient immunoglobulin-like molecule is thought to be that of Immunoglobulin New Antigen Receptor (IgNAR) (Feige et al. [56]), or Immunoglobulin W (IgW) (Berstein et al. [13]), both of which are found in sharks and all cartilaginous fish. IgNAR and IgW are thought to have evolved around 500 million years ago with the emergence of the adaptive immune system, and although they are sequentially very different from modern antibodies, their tertiary structure has been shown to be remarkably similar to human antibodies. Cartilaginous fish have been shown to produce different types of immunoglobulin molecules, characterized by different quantities of structural chains, differing structures, and differing splice opportunities, and these have diversified to produce the myriad of antibody structures known today, (Flajnik and Kasahara [58]). An overview can be seen in Figure 1.1. Humans have evolved to have multiple different antibodies, each with different functional roles and structures. For the purpose of this thesis, immunoglobulin G (IgG), will be focussed on, as this is most commonly used for therapeutics.

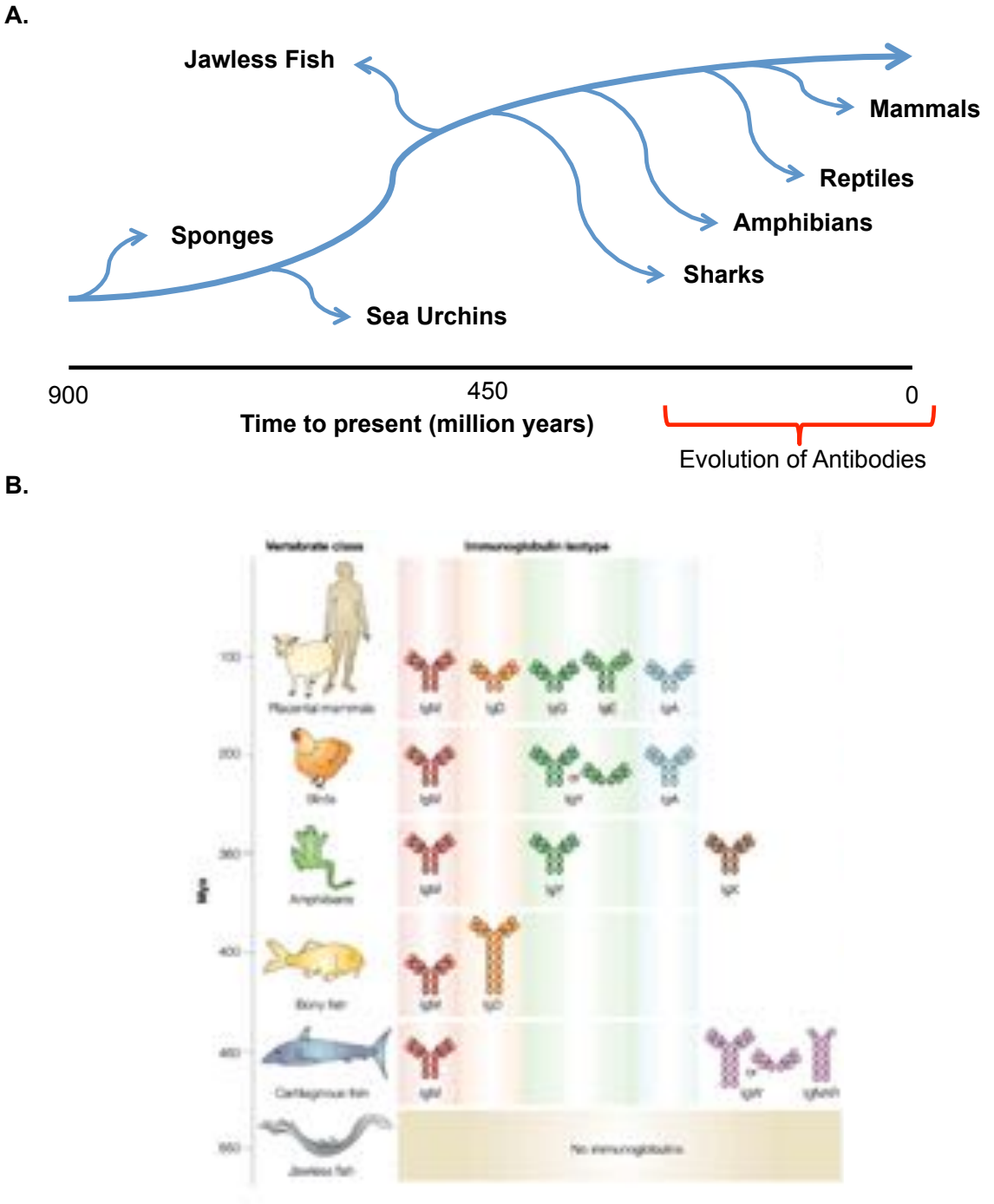


Fig. 1.1 Schematic for the evolution of antibodies, showing A) A timeline of eukaryotic evolution, with the emergence of immunoglobulins occuring with sharks B) Different spcecies, with the different classes of immunoglobulins they produce, image adapated from ref Flajnik and Kasahara [58].

1.1.2 Antibodies as Biological Molecules

The structure of Immunoglobulin G (IgG) molecules, the most common immunoglobulin, is extremely well defined (Schroeder and Cavacini [163]). In short they consist of 2 heavy (H) chains and 2 light (L) chains each of which contains a constant (C) region and variable (V) region (Figure 1.2). Each antibody is made of two halves, each of which consists of a heavy chain linked via disulphide bridges to its corresponding light chain. Each half is then further linked to the other via disulphide bridges. The variable domains contain numerous Complimentarity Determining Regions CDRs, domains that contain large amounts of hydrogen bonds, which make them extremely rigid, and consequently, these are involved in highly specific antigen recognition. Due to their functional role in the molecule, this region is hypervariable with respect to structure and sequence. The Fragment Crystalizable (Fc) domain of the protein is involved in effector function (the coupling of immunoglobulins to signaling proteins on immune cells) and consists of CH2 and CH3 domains, both of which are glycosylated on ASN 297. Glycosylation is known to play a role in maintaining the structure of the molecule, and glycans are also involved with interactions with the Fc (Nimmerjahn and Ravetch [134], Woof and Burton [201]), whilst the Fragment Antigen Binding (Fab) region, containing the CDRs, is used for antigen binding. A schematic can be seen in Figure 1.2.

Antibodies undertake a variety of functions within the body, and are an extremely diverse classification of molecule. The main immunoglobulin constituent of mammalian blood is IgG (Wachholz et al. [196]). This molecule generally performs initial binding events with recognized pathogens, and is then responsible for activating downstream signalling events and the activation of the adaptive immune system response. In humans, each person contains a large array of pre-expressed immunoglobulins called the *preimmune antibody repertoire* (Alberts et al. [2]), this is an array of around 10^5 molecules that are thought to be able to react weakly with almost any non-self antigen. These are thought to be used as a primary immune

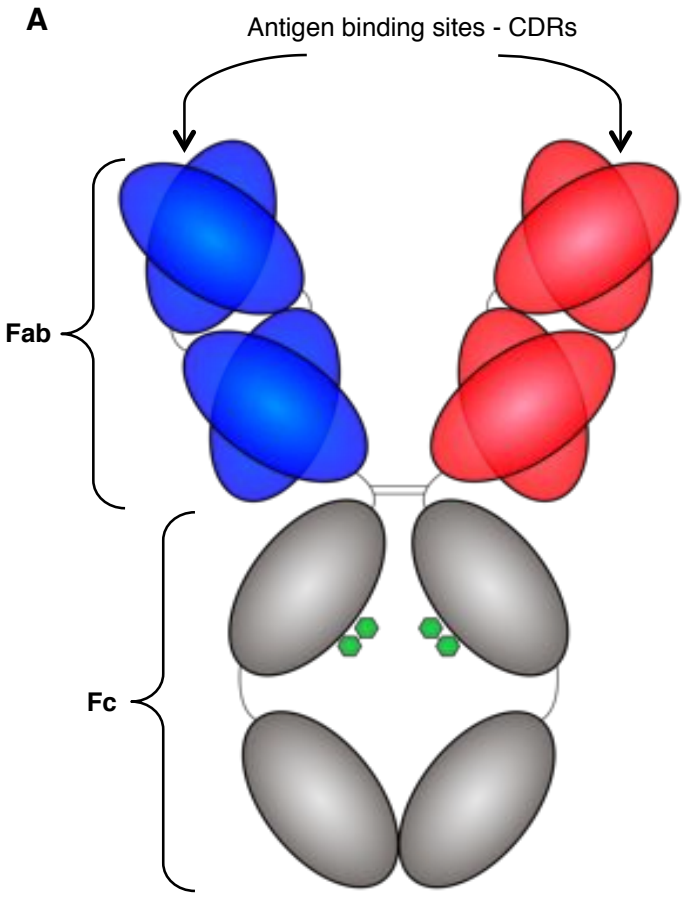


Fig. 1.2 Schematic structure of an immunoglobulin IgG.

response, much like Toll Like Receptors (TLRs) in the innate immune system, before the adaptive immune system is activated, and highly specific antibodies for a particular pathogen are generated in huge quantities (Kindt et al. [100]).

Whilst antibodies seemingly have roles in the innate immune system, they are primarily known for their role as the predominant agent of the adaptive immune system. They are produced in huge quantities by B-cells when stimulated by the identification of a pathogen by the immune system. Each B-cell produces only a single type of antibody, with trillions of B-cells existing, each producing an antibody with different chain structures, and different CDR regions (Mayforth and Quintans [125]). B-cells are selected for their ability to produce antibodies that bind highly specifically to antigens on target pathogens, and are stimulated

to replicate and start producing vast quantities of antibodies. Once antibodies are produced, they perform numerous functions that result, generally, in the destruction of an invading pathogen.

By binding to antigens on target molecules, antibodies are able to activate the complement system (Clark [32]), in which they bind a target, before interacting with numerous proteins on the cell surface of other phagocytic cells, activating protein cascades that result in the phagocytic cell attracting other immune cells, engulphing the target pathogen, or puncturing holes in the membrane to lyse the target. Apoptosis is also sometimes induced through the complement system (O'Mahony and Bishop [137]), in which antibodies direct the immune system to trigger self destruction of invading organisms.

A major ability of immunoglobulins is the agglutination of targets (Clark [32]), due to their structure containing two Fab regions. Each CDR is able to bind the same target on a different pathogenic cell or molecule, which, in large quantities can form a mesh of attached molecules, which are then easier for the immune system to clear. Another advantage to antibodies binding a specific target on the surface of invasive cells is that this may disrupt the functions of certain features, for example, proteins on the surface of viruses are known to be involved in their ability to bind to, and enter a cell (Ekiert et al. [54]).

Antibodies also act via Antibody dependent cell-mediated cytotoxicity (ADCC) (Wang et al. [198]) whereby other immune cells are recruited by the molecule to actively lyse the target pathogenic cell. The ability for antibodies, specifically IgG to perform these functions, as well as the ability for them be highly specific in their targets, makes them an extremely attractive molecule from a pharmacological perspective, as if these responses can be directed against targets that we wish to clear from the body, eg, cancers, the immune system can be stimulated to act against only a very specific target.

1.1.3 Pharmacological Relevance of Monoclonal Antibodies

Monoclonal Antibodies represent a significant proportion of newly developed drugs (Hudson and Souriau [82]). The ability to target a specific antigen within the body to the exclusion of all others, a 'molecular magic bullet' (Strebhardt and Ullrich [183]), is a highly attractive prospect. Antibodies are one of the most intensively investigated classes of proteins due to their potential use as pharmaceuticals (Padlan [140]). They are potential treatments for many otherwise hard to treat illnesses such as: cancer (Adams and Weiner [1]), arthritis (Arnold et al. [5]), and even psoriasis (Laws and Warren [109]) and food allergies (Leung et al. [111]). The natural functions of antibody molecules in immunity are properties that can be exploited by the production of antibodies with a target of our specification that retains abilities to recruit immune cells and activate ADCC, there is also the potential to use antibodies as specific carriers for molecules. There are currently over 25 commercial therapeutics that are monoclonal antibodies, and hundreds of protein drugs (Carter [24]). Currently all monoclonal antibody therapeutics are derived from the IgG subclass of antibodies, a majority being IgG1.

There are numerous mechanisms for the production of monoclonal antibodies, including the use of murine hybridomas (Zettlitz et al. [211]) and more recently, second generation monoclonal antibodies are produced using either phage display systems (Hoogenboom [77]) or with genetically engineered mice (Roque et al. [153]) – allowing the avoidance of a majority of side effects from grafted proteins (Hwang and Foote [84]). A timeline for generation of monoclonal antibodies can be seen in Figure 1.5

1.1.4 Aggregation of Antibodies

It is widely accepted that aggregation is a "general property of proteins" (Sanchez de Groot et al. [160]). Without aggregating, many proteins would not be able to correctly fold, associate with partners into complexes, or form amyloid structures - which have recently been

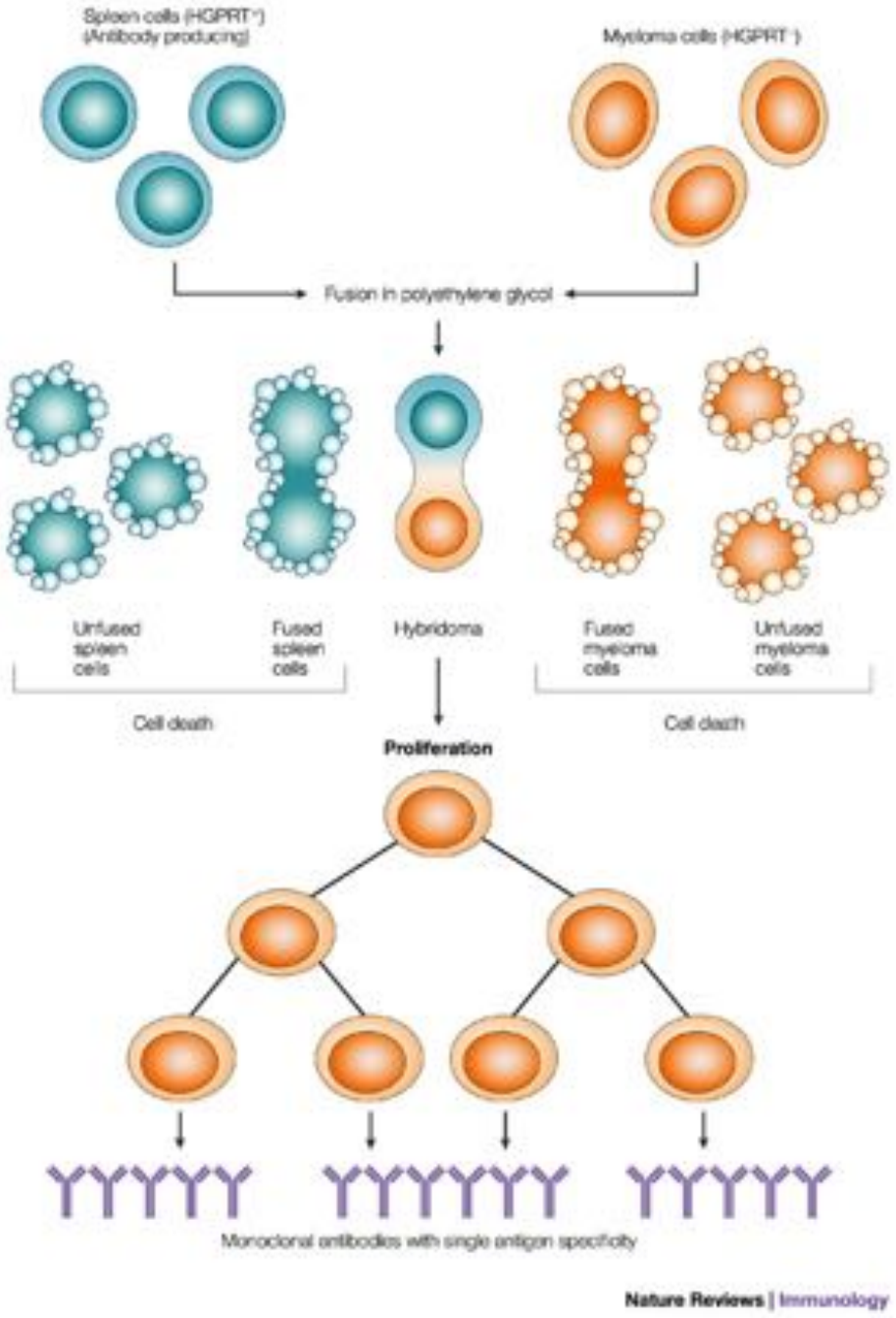


Fig. 1.3 The generation of monoclonal antibodies through the use of hybridoma cells. Figure adapted from Alkan [3].

shown to have biological function (Fowler et al. [59]). Aggregation prone regions in proteins are thus found to be evolutionarily conserved in many ubiquitous cases, and cells have

evolved to balance the risks associated with adverse protein aggregation, with the benefits granted by their presence. The proteome has evolved to overcome many problems associated with adverse aggregation of proteins by utilizing systems such as chaperones (Kinjo and Takada [101]), enzymes, and post-translational modification (Mirazimi and Svensson [129]) - among other things. The aggregation of proteins is still, however, implicated in over 40 major diseases, including Parkinson's disease, Alzheimer's disease, and Diabetes (Auer et al. [7], Soong2009) whereby large aggregates disrupt biological structures, shift the delicate equilibrium of stable proteins, or initiate costly and damaging immune reactions. Antibodies, by comparison to many other proteins, are relatively stable under physiological conditions, and their aggregation is rarely an issue biologically.

When antibodies are removed from a physiological context however, aggregation is a much more common phenomena. As the concentrations in which antibodies are manufactured are extremely high, and involve adverse temperatures, pressures, materials, and chemicals (Vázquez-Rey and Lang [190]), most natural, and a huge proportion of synthetically generated antibody structures aggregate to the point of being non-functional when generated in high enough quantities for clinical use. Due to this fact, the aggregation of monoclonal antibodies is the major bottleneck in the generation of novel antibody based drugs (Sharma [167]).

1.1.5 Physical Approaches to Antibody Aggregation

The aggregation of antibodies is a heavily studied topic. With the development of antibody based drugs a priority for many pharmaceutical companies, the stabilization of otherwise unusable protein structures is a very attractive prospect. Whilst most antibodies are similar in sequence, and only differ in the CDR region – it would be assumed that by the discovery of a few stable scaffolds, many different stable drug structures would be able to be easily developed. This has not proven to be the case however, as each structure can be described

as having a distinct "personality" (Brych et al. [21]), and thus needs careful investigation in order to develop a protein that is suitable for use as a therapeutic. Currently the development of antibody structures for therapeutic use has to be addressed in a case-by-case manner, and development generally involves experimental heuristic techniques (Wang et al. [199]) that generally revolve around trial and error.

1.1.6 Computational Approaches to Antibody Aggregation

There has been some recent work on the development of tools to address protein aggregation in silico. Most notably, work by Trout (Chennamsetty et al. [28, 29, 30, 31]), has developed a series of tools aimed at the identification of aggregation prone regions of immunoglobulin structures through analysis of a series of parameters, including Spatial Aggregation Propensity (SAP), a measure of aggregation propensity based on the hydrophobicity and solvent accessible surface area of regions of a static immunoglobulin structure. This has been somewhat successful, mutant proteins developed with the aid of the results from these tools have been shown to have a lower aggregation propensity. There are limitations to these techniques however - by not catching any of the dynamics of actual interactions, and only investigating the surface properties, information is able to gained quickly, but without context, and without any specific information about mechanisms of interactions. Recent additions to the SAP toolkit include the ability to gain information from a dynamic trajectory of a single protein taken from a molecular dynamics trajectory, and this alleviates some of the issue, though is still not a true simulation of aggregation.

Studies have also been performed on the self interactions of other proteins, notable examples include the aggregation of pre-amyloid protein constructs using a coarse grained methodology, and the self association of the cohesin-dockerin protein complexes (Hall and Sansom [70], Sø rensen et al. [173]). The Barnase-Barstar protein association complex has also been studied using the coarse grained methodology, and accurately reproduces experi-

mental interactions (Basdevant et al. [9]).

Further work has been done with the use of ultra-coarse simulations of full antibody molecules by Voth. In these systems, a single antibody is reduced to 12 or 26 particles, and general characteristics of the atomistic molecule are transferred. Different aggregation rates were observed in silico for two different classes of antibodies that correlated with experimental characteristics (Chaudhri et al. [27]).

These methods can be seen in Figure 1.4

1.2 Carbohydrates

1.2.1 Carbohydrates as Biological molecules

Sugars are a unique biological molecule. Whilst nowadays a substantial amount of research is performed on carbohydrates, they are still understudied compared to other main biological molecules – proteins, nucleic acids, and lipids (Lowe and Marth [120]). All organisms synthesize and utilize carbohydrates to some extent (Varki [189]), and they are arguably the most abundant biomolecule found in nature (Springer and Gagneux [178], Varki [189]), but the differences between each organism, each having a unique and dynamic glycome (Zoldoš et al. [212]), mean that they are often difficult to study, especially structurally, and are subsequently reduced, eliminated, or often even ignored within biological systems, simply because they are too difficult to study (Lauc et al. [107]). The makeup of sugars within an organism, the glycome, of any individual organism is generally in constant flux, the sugar content being maintained by a series of enzymes, including glycosyltransferases and glycosidases, whose main function is to catalyse the generation of glycosydic linkages, or break them (Ghazarian et al. [63]). In this way sugars are constantly being added, removed, reduced, or cleaved, therefore, getting a biological "snapshot" of the carbohydrate makeup of a cell, or cell component, is particularly difficult. The makeup of the glycome, and the

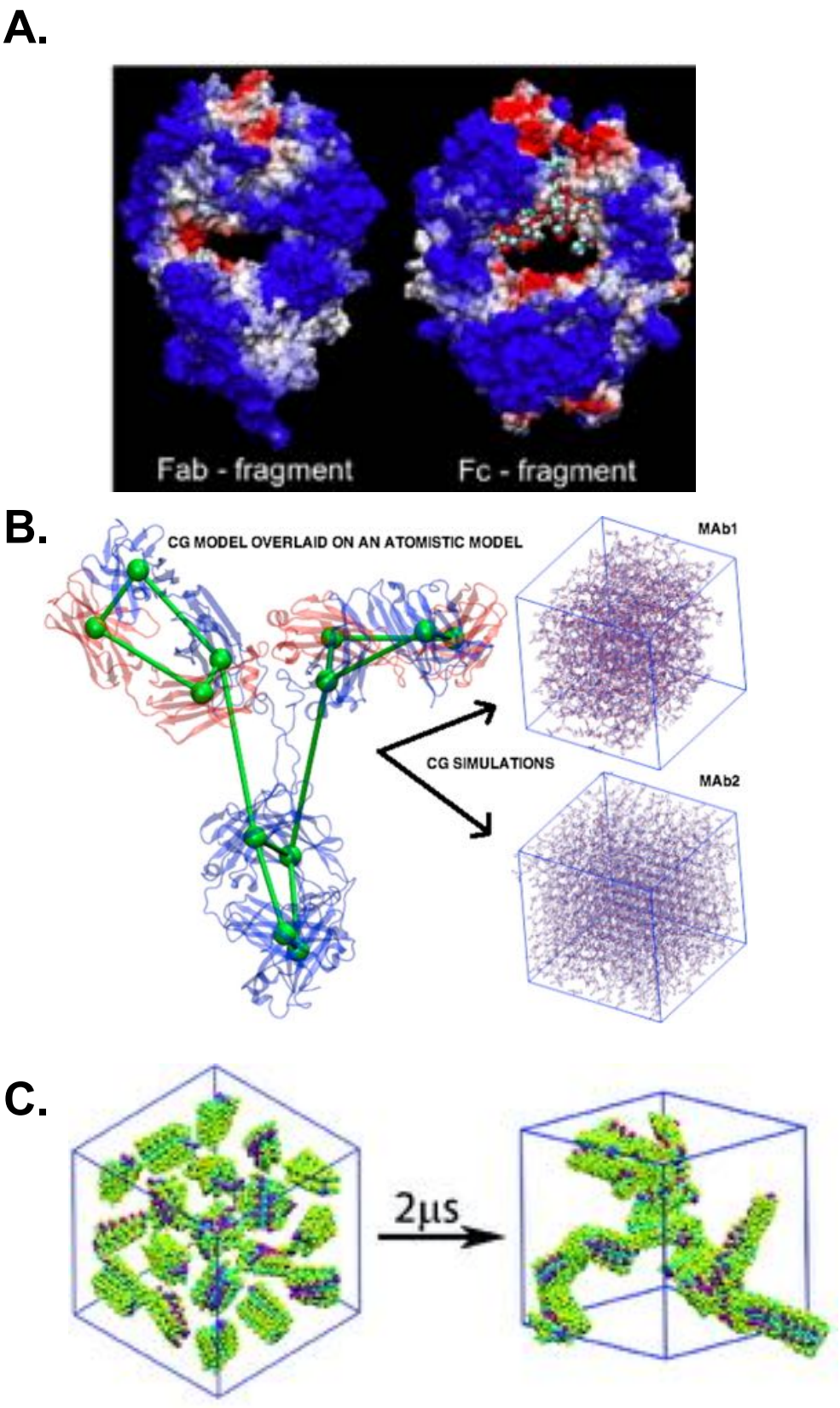


Fig. 1.4 Computational methods for studying protein aggregation, showing A) Spatial Aggregation Propensity (Chennamsetty et al. [31]). B) Ultra coarse grained models (Chaudhri et al. [27]). C) Amyloid protein aggregation (Sø rensen et al. [173]).

degree to which it is dynamic, is known to be different in almost all incidences of cancer, and thus, has potential to be a marker for disease (Dube and Bertozzi [49]).

Carbohydrates can be found on the membranes of most organisms, either in the form of gycolipids (Cantù et al. [23]), glycoproteins (Rudd and Dwek [157]), or standalone structures such as the cell wall in bacterium (Vollmer et al. [193]). The coating of cells due to carbohydrates is named the glycocalyx (Roseman [154]), and this fuzzy, sugary, shell is involved in cell to cell recognition (Hakomori [66]), cellular protection and immune responses (Sigvard Olofsson [172]), and cellular adhesion (Hakomori et al. [65], Taylor and Drickamer [186]), among other functions.

Complexity of carbohydrate molecules is almost certainly a key reason for the difficulty of their study. Unlike protein and DNA, which has a distinct and fairly ubiquitous code, carbohydrates have no such biological blueprints, and their uses are generally in the post-translational, and post-productional modification of proteins or lipids, and are hard to monitor (Dove [48]). Glycosylation is also often transient, with a certain sugar structure being used only, for example, to transfer a protein from certain regions of the golgi to others, before being cleaved or removed entirely (Reynders et al. [149]). Sugar molecules are also extremely diverse with respect to sequence (Meier and Duus [126]), and can form many more structures than a similar number of amino acids, or DNA bases, which have only a single linkage type (e.g. a peptide bond). Whilst the possible types of combinations of two amino acid residues is only 1, two sugars, however, can potentially be combined in 11 different ways, with glycosydic bonds linking different carbons in each molecule in each case. Expanding the number of components, four different amino acids can potentially form 24 different peptides if allowed every possible combination, whereas 4 different hexose sugar molecules can be combined into a staggering 35,560 different potential linkages, each with multiple different conformations, some of which would include branching structures, with multiple carbohydrates attached to a single preliminary one, though in reality this number

is less due to major glycoforms being favourably adopted by the cell due to enzyme specificity (Sharon and Lis [168]). The fact that many sugars have stereoisomers with distinct, recognizable differences in function, despite extremely similar structures also adds to the difficulties in studying carbohydrates experimentally. Glucose and galactose, for example, are extremely similar structurally, the difference between the two simply being the chirality of a single carbon on the ring, however, biologically the two are extremely different, glucose being used as a primary fuel source, and galactoses only known function in humans is as a component of much more complex structures such as glycolipids or glycoproteins (Neutra [133]).

1.2.2 Glycolipids

Glycolipids are a combination of two of the four main types of biological molecules – lipids and carbohydrates. Whilst glycolipids are present in a huge variety of cell membranes, their study only really started in earnest around 25 years ago, and comparatively little is known about them. Whilst the study of glycolipids is in its infancy, their misregulation or degradation is involved in numerous diseases (Buccoliero et al. [22]), such as Gaucher disease (Hruska et al. [79]), and Krabbes disease (Jacob et al. [88]), both of which involve the accumulation of glycolipids within tissues. They have been shown to have interactions with numerous membrane proteins, including EGF receptors (Coskun et al. [35]), insulin receptors (Vainio et al. [188]), and even numerous ion channels (Dart [39]). They are also a major component of the lipid raft hypothesis, the thought that certain lipid species may segregate within a fluid membrane to form regions of high concentrations of certain molecules, including proteins (Cantù et al. [23], Dart [39], Ghazarian et al. [63]). The stability of the membrane is also effected by the concentration of glycolipids, they are known to stabilize membranes in which they are present, this is a key mechanism used by epithelial cells to protect against hostile fluids (Edidin [52]). There are also implications that glycolipids may

be involved in changing the shape of certain membrane regions, leading to increasing or decreasing concentrations of proteins and lipid species and thus physiological changes in activity (Hakomori et al. [65]).

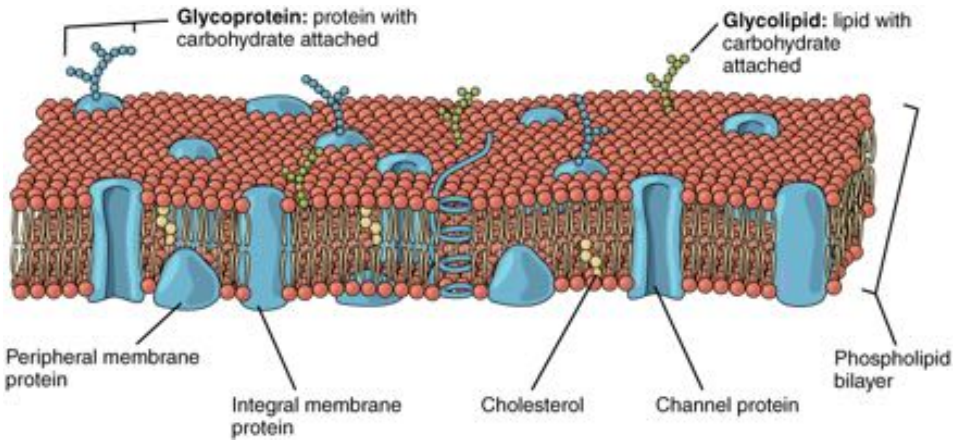


Fig. 1.5 Schematic of a biological membrane, showing glycolipids and glycoproteins. Image adapted from wiki commons.

1.2.3 Protein Glycosylation

Protein glycosylation is a comparatively understudied area of biochemical research – particularly through computational means. The difficulty in characterizing protein glycosylation, and the dynamic nature of carbohydrate sequences makes their study difficult. Many proteins are glycosylated at some point in their lifespan, either as tags for movement to different sections of the cell, or as markers for recognition of self/non-self, and glycosylation is by far the most common post-translational modification for proteins (Lis and Sharon [115]). Most membrane proteins are glycosylated on their extracellular domains, and most proteins that are secreted from the cell do so with post-translational glycosylation. Whilst glycans are known to stabilize and be intrinsically involved in the fully folded structures of numerous proteins (Wyss and Wagner [204]), many simulations simply do not take this into account.

1.2.4 Current Approaches to Simulations of Carbohydrates

Currently, there are numerous methods available for the simulation of carbohydrate molecules. Atomistic forcefields have existed for some time, though have only recently started to accurately represent the dynamic nature of sugar molecules, and account for the huge diversity in structure and linkage (DeMarco and Woods [43]). Simulating carbohydrates is generally very difficult due to their huge conformational heterogeneity, caused by their unusual steric, electrostatic, and solvation properties. Furthermore, crystal structures that are (rarely) obtained containing carbohydrate structures are often truncated due to structural flexibility, and it is often uncertain if these structures are adopted in physiological conditions. Furthermore, Nuclear Magnetic Resonance (NMR) data only gives information on average structures, over huge time periods, and over large arrays of structures.

Despite these issues, libraries exist for many carbohydrate molecules in different forcefield contexts, GROMOS forcefields for example, have a limited pool of carbohydrate structures that can be added to existing molecules (Lins and Hünenberger [114]), though linkages for eg. protein glycosylation, do not currently exist in this scheme. Recently, development of a new forcefield named GLYCAM (Kirschner et al. [102]), developed by Robert Woods' group, has set the standard for carbohydrate parameters. GLYCAM is parametrized based primarily on NMR experiments, and quantum mechanical simulation data, and includes details required for the simulation of glycoproteins, and glycolipids, due to the modular nature of the AMBER forcefield on which it is based.

Simulations of both glycolipids, and glycoproteins have since been reported in the literature, though the addition of carbohydrate groups or glycolipids to a simulation is rare, and generally is not standard procedure unless carbohydrates are explicitly necessary in a system. Despite the difficulty in simulating cabohydrate structures, some simulations have been performed. Notably, a simulation has been performed of a GM3 lipid bilayer, in which GM3 is observed clustering together (DeMarco and Woods [44]). Atomistic simulations of

galactocerebroside containing lipid bilayers also show the formation of lipid nanodomains (Hall et al. [67]).

Atomistic simulations of protein glycosylation are also rare, though some have been reported. Notable simulations include those of prion proteins (Zuegg and Gready [213]), membrane protein receptors (Mandal and Mukhopadhyay [122]), and protein folding (Shental-Bechor and Levy [169])

Simulations run using coarse-grained molecular dynamics rarely contain carbohydrate groups. Carbohydrate parameters were released as a part of the MARTINI coarse grained forcefield in 2009 (Lopez et al. [117]), though only a very small number of monosaccharides were included, and no parameters for the attachment of molecules to either proteins or lipids were provided. This first iteration of the carbohydrate forcefield, named Sweet MARTINI, was generally parametrized for the simulation of multimeric carbohydrate molecules such as amylose. Recently however, parameters for a number of glycolipid species were released for general use (López et al. [118]). There were however, some omissions, notably GM3, a common component of biological membranes. Also of note with this work, is that structures are parametrized with the use of an extensive elastic network, this results in parameters that are non-transferable to other similar structures, and the dynamics are not able to be explicitly controlled, which may be useful in some cases. Another current issue with the parameters released by the MARTINI group is that lipids structures are not stable enough to be simulated with large time steps. Coarse grained simulations generally rely on larger time steps in order to produce simulations of a much longer timescale, traditionally 20, or sometimes even 40 fs timesteps are taken for a coarse grained system, the time steps reported to be stable for the recently glycolipids are generally around 5 fs, and so, gaining long term behavioural information from these systems is extremely computationally expensive.

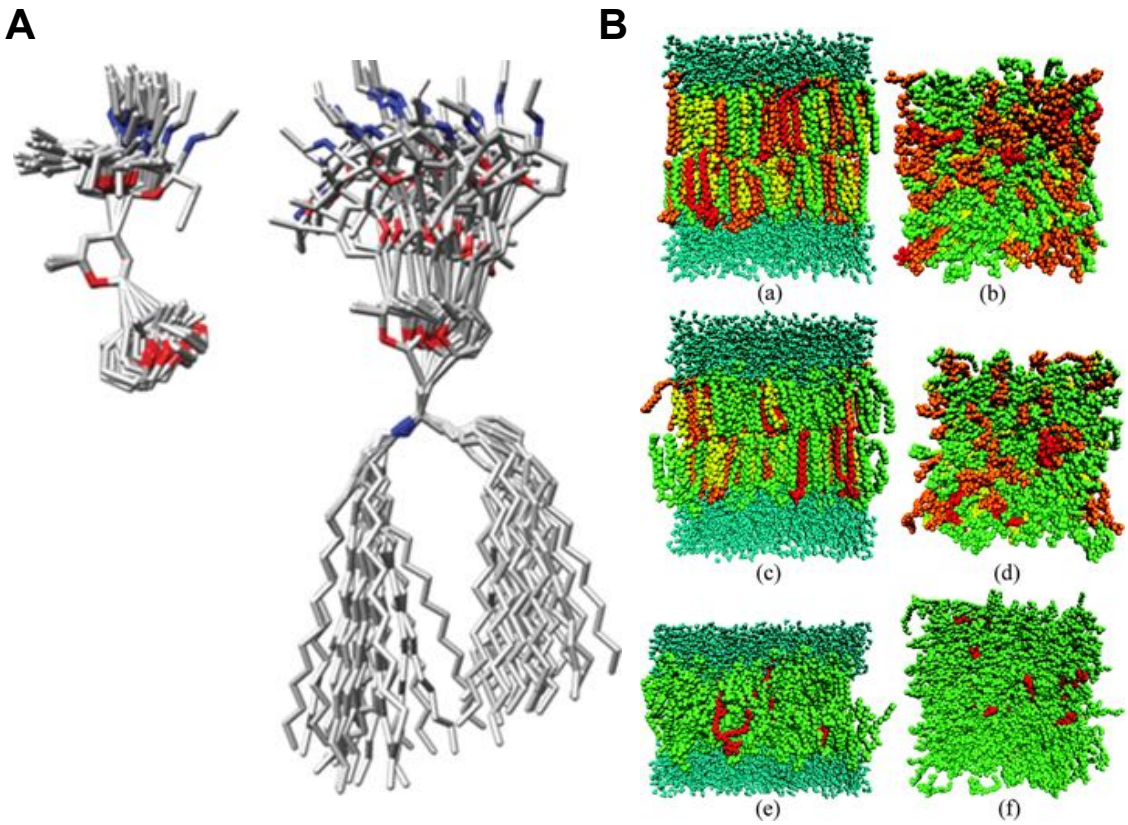


Fig. 1.6 Current computational work on glycolipids, showing A) Different conformations adopted by GM3 in an atomistic bilayer simulations, adapted from DeMarco and Woods [44]. B) Atomistic bilayers showing the partitioning of galactocerebroside. Galactocerebroside is shown in red. Adapted from Hall et al. [67].

1.3 Thesis Overview

Presented in this thesis are 4 bodies of work, focussing on computational studies of antibody aggregation and the simulation of carbohydrates.

Chapter 2 focusses on the methods used for simulation and analysis of systems described in other chapters, and aims to be a comprehensive reference for the way in which molecular dynamics is performed.

Chapter 3 describes initial forays into the study of antibody aggregation, using publicly available structures and large scale coarse grained molecular dynamics simulations. Differences can be seen in the self-self interaction behaviours of two different regions of the fragmented antibody, and regions of high interactivity are suggested for mutational analysis.

Chapter 4 extends the methodologies described in chapter 3 to a high-throughput approach, whereby a series of scripts is presented that automates the system setup, large-scale simulation, and analysis of protein-protein interactions for protein structures. This script is highly customizable and modular, and incorporates analysis techniques developed in the previous chapter, and utilizes more statistical techniques to extract information from large arrays of data. The script is tested on a set of 6 antibody fragment structures provided by Lonza Biologics, with corresponding experimental data. Results from the high-throughput analysis match well with known experimental properties, the system being able to rank the 6 proteins in order of experimental aggregation propensity. Regions of the structures are also suggested for mutational studies.

Chapter 5 introduces the development of a novel methodology for the simulation of carbohydrate molecules in coarse grained systems. Principles of the force field are introduced, and simple examples demonstrated, before moving to more complex carbohydrate molecules. Last, in order to show the success of the methodology, glycolipid molecules are parametrized, emergent properties are analysed, and structures made available for others to use.

Chapter 6 expands the previously developed carbohydrate forcefield to the much more complex glycan found within the immunoglobulin Fc, with an emphasis on the development and parametrization of the carbohydrate structure, and it's attachment to the protein. Once a

parametrized model is completed, emergent properties are compared to atomistic values for validation, and different secondary structure restraints are tested with the methodology. The final project involved using the glycosylated protein model to investigate structural fluctuations of the immunoglobulin Fc with and without carbohydrates, to attempt to resolve how the molecule reacts to deglycosylation and how this impacts interactions with Fcγ receptor.

Finally, some future directions are examined in Chapter 7, with ideas for combining the ideas from all of the chapters into a single study in the future.

Chapter 2

Materials and Methods

"The history of atomism is one of reductionism – the effort to reduce all the operations of nature to a small number of laws governing a small number of primordial objects."

Leon M. Lederman

This chapter aims to demonstrate and examine the materials and methodology employed for the generation of data in this thesis. The principle methodology used is molecular dynamics (MD) simulations, in which a system of particles is set up, and dynamics applied, whereby the positions and energies of atoms are recalculated for every timestep t, until a target time is reached.

For an excellent overview of the methods described here, the textbook written by (Leach [110]), is recommended.

2.1 Classical Molecular Dynamics

With the description of quantum physics, it becomes possible to describe the intricate motion of all components of a system from basic principles. In practise, however, for the purposes of simulations, describing the location and characteristics of every elemental particle for any system more complex than a hydrogen atom becomes extremely computationally taxing. Today, quantum mechanical simulations, that take into account all elementary particles, can rarely be expanded to system sizes larger than, for example, a ligand binding site comprised of 3 or 4 amino acid residues. In order to overcome this problem, the principles based on the Born-Oppenheimer approximation are used. The Born-Oppenheimer approximation, simply, allows the description of a system of atoms as a series of Cartesian coordinates, one for each atom, by using the huge disparity between the mass of the nucleus of an atom, and the electrons that surround it to assume that the electrons are able to react instantly to any changes in the nucleus. Due to this, the atom can be represented as a single point in space - this is known as the classical representation (Leach [110]).

By using the classical treatment for every atom in a system, it is possible to treat particles very simply, using Newtonian equations of motion to evolve the system over time. In its simplest form, Newtons second law for motion can written, for particle i, as:

$$F_i(t) = m_i \ddot{r}_i(t) \tag{2.1}$$

Where F_i represents the force on a particle, m_i represents the mass of each particle, and r_i represents the acceleration at a certain time. By solving this equation for every particle in the system, we are able to generate their locations and velocities, and thus by solving this equation for windows of time were are able to observe predicted changes to the locations of

particles over time. This law can be rearranged as:

$$\frac{d^2 r_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{2.2}$$

To find the motion of a particle of mass m_i along coordinate x_i, given a force F_{x_i} acting in the x direction. In systems the size of those explored in molecular dynamics, it is not possible to solve these equations for multiple particles analytically, as the system is coupled, and complex, therefore an algorithm is applied that iterates over the equations in very small steps, until a target time period is reached.

The method used to iterate over the system is a finite difference methodology, in which differential equations can be solved, with some error, in a relatively efficient manner. The specific algorithm utilized by GROMACS is based on a technique known as the Verlet method, and is called the "Leap Frog Algorithm" (Hockney et al. [75]). The Leap Frog Algorithm is used to calculate velocities at half timestep intervals from the initial defined timestep Δt. It takes the following forms:

$$r_i(t + \frac{\Delta t}{2}) = r_i(t) + v_i(t + \frac{\Delta t}{2})\Delta t \tag{2.3}$$

$$v_i(t + \frac{\Delta t}{2}) = v_i(t - \frac{\Delta t}{2} + \frac{F_i(t)}{m_i})\Delta t \tag{2.4}$$

2.1.1 Forcefields

In classical molecular dynamics, the forcefield is a description of the interactions between all atoms present in any system (Leach [110]). They can be described as a potential energy function based on the derivative of Newtons second law of mechanics, coupled with parameters for each atom type. In classical molecular dynamics, atoms are treated as points of mass, that are charged, and are bonded to other atoms via a spring constant, or harmonic

potential. This representation is called a "ball and spring" model, and can be thought of as being similar to a physical model of spheres attached to each other via springs. The potential energy of any atom within a system is made up of the sum of contributions from bonded and non-bonded interactions within this model. Bonded interactions being those between covalently linked particles, and non-bonded interaction being interactions that act between all atoms in a system. The potential energy of a system can be thus calculated as:

$$U(r^N) = U_{BONDED} + U_{NON-BONDED} \tag{2.5}$$

Where N is the number of particles, and r is their position.

2.1.2 Bonded Interactions

The bonded interactions component of a forcefield includes not only pairwise interactions, but also takes into account 3 body, and 4 body interactions, in the form of angle and dihedral values. In total, the contributions can be broken down into:

1. Bond Stretching (U_{BONDS}).
2. Bond Angle Bending (U_{ANGLES}).
3. Bond Torsion, or Proper Dihedral Angles ($U_{DIHEDRALS}$).
4. Improper Dihedral Angles ($U_{IMPROPERS}$).

The total bonded energy contribution to a system can be represented as:

$$U_{BONDED} = U_{BONDS} + U_{ANGLES} + U_{DIHEDRALS} + U_{IMPROPERS} \tag{2.6}$$

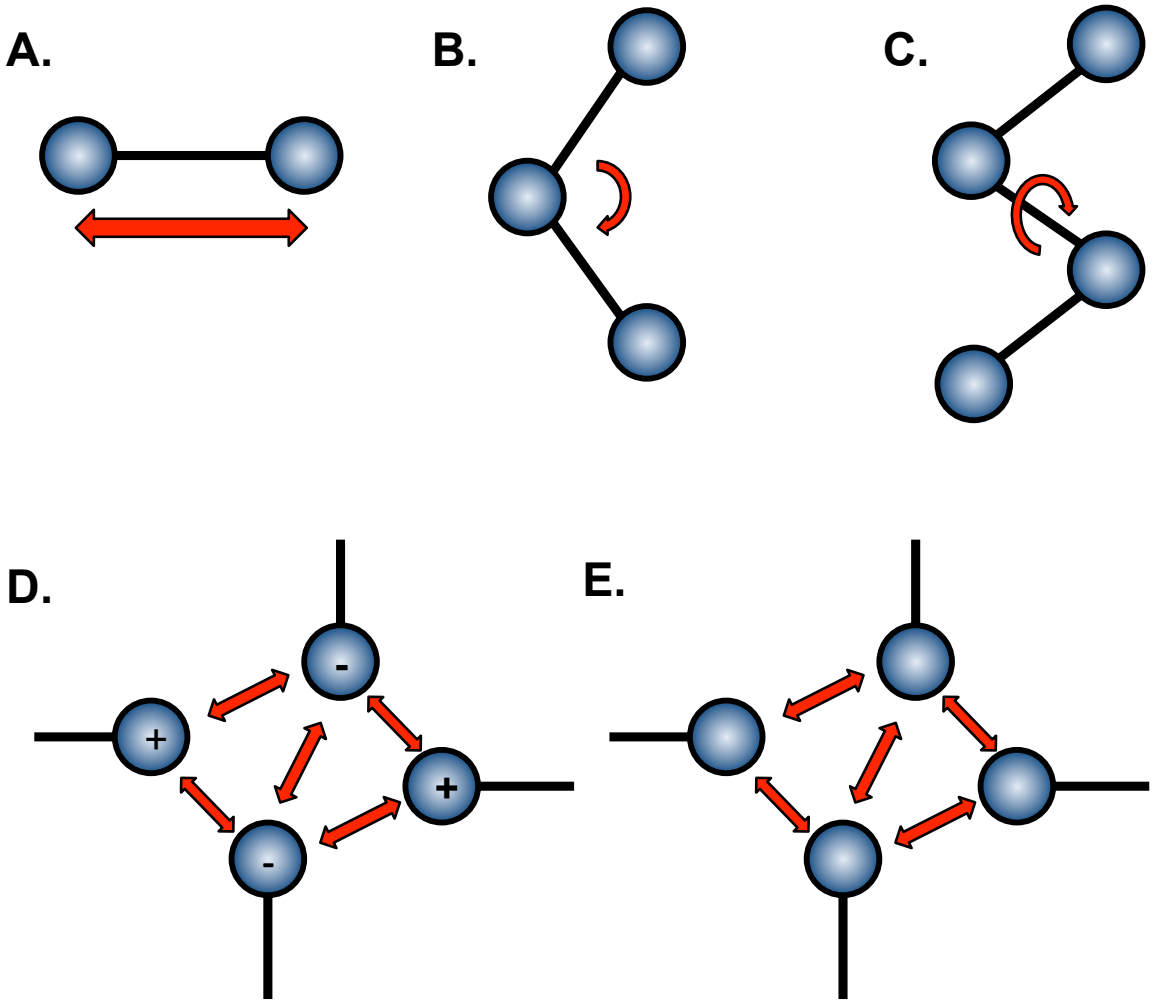


Fig. 2.1 Schematic of the main interactions within a classical molecular dynamics forcefield. A) Bond Stretching. B) Angle Bending. C) Torsion. D) Electrostatic interactions. E) Van der Waals interactions.

Bond Stretching

The bond stretching potential is a harmonic potential, in which the energy of a bond varies proportionally to the square of the displacement from the reference, or defined value. This means that bonds can oscillate from their equilibrium value, in accordance with the equation:

$$U_{BONDS}(r_{ij}) = \frac{1}{2}k_{ij}(r_{ij} - b_{ij}^{eq})^2 \tag{2.7}$$

In which r_{ij} represents the bond length for a particular time point within the simulation, b_{ij} represents the equilibrium bond length, and k_{ij} represents the spring constant for atoms i, and j. Increasing a force constant k_{ij} results in a bond that is much harder to stretch.

Bond Angle Bending

The bond angle potential is defined in a similar way to the bond stretching energy, it is a harmonic oscillator, in which the energy of an angle between three atoms, i, j, and k increases the further from the equilibrium value θ_{ijk}^{eq}, in accordance with:

$$U_{ANGLES}(\theta_{ijk}) = \frac{1}{2}k_{ijk}(\theta_{ijk} - \theta_{ijk}^{eq})^2 \tag{2.8}$$

Where k_{ijk} is the spring constant, and θ_{ijk}^{eq} is the equilibrium value for the angle between atoms i, j, and k.

Bond Torsion

The bond torsion energy is defined for a set of four atoms, and so is a four body potential. The potential for dihedrals is derived from a cosine function, and multiple potential equilibrium values can be assigned. The potential is show below, where ϕ_{ijkl}^{eq} is the equilibrium value, n the multiplicity of the function, and k_ϕ represents the force constant applied for four atoms, i, j, k, and l. The angle ϕ_{ijkl} represents the angle between the two planes ijk, and jkl.

$$U_{DIHEDRALS}(\phi_{ijkl}) = k_\phi(1 + cos(n\phi_{ijkl} - \phi_{ijkl}^{eq})) \tag{2.9}$$

Improper Dihedrals

Improper dihedrals are generally used in order to keep ring structures planar, and can be used to prevent the flipping of chiral molecules. There is no analogue in chemical bonding,

unlike the previous three potentials. They also use a harmonic potential to restrict potential angles along four atoms, based on the angle between the planes of atoms ijk, and jkl.

$$U_{IMPROPERS}(\xi_{ijkl} = \frac{1}{2}k_{\xi}(\xi_{ijkl} - \xi_{ijkl}^{eq})^2 \tag{2.10}$$

Where xi_{ijkl}^{eq} is the equilibrium value of the angle, ξ_{ijkl} is the current value of the angle, and k_{ξ} is the force constant.

2.1.3 Nonbonded Interactions

Nonbonded interactions are defined as the interactions between all particles in a system, whether they are covalently bonded to each other or not. Classical treatment of non-bonded interactions requires the inclusion of 3 terms:

1. Van der Waals Attraction.
2. Spatial Repulsion.
3. Electrostatic Interaction.

In most molecular mechanics methodologies, van der Waals interactions and spatial repulsions are combined into a single potential, the Lennard-Jones Potential, whereas electrostatics are treated separately, using a Coloumbic representation.

$$U_{NON-BONDED} = U_{LJ} + U_{ELECTROSTATICS} \tag{2.11}$$

The Lennard-Jones Potential

The Lennard-Jones Potential attempts to combine the attractive and repulsive terms into a single function. The repulsive term comes into effect when two particles move any distance within their cumulative contact radii, and attractive term is strongest when particles

are close together, and rapidly reduces as the molecules move a few Å apart. This behaviour is captured by a term that attempts to describe a weak attraction at distance, with attraction increasing as particles come together, into an "attractive well", before energy rapidly increasing as particles start to overlap their van der Waals (spatial) radii. This is captured with the term:

$$U_{LJ}(r_{ij}) = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.12}$$

In this equation ε_{ij} controls the well depth, and is a measure of the attractiveness of two particles i and j. r_{ij} represents the distance between the two particles, and σ_{ij} is the collision distance, or distance at which the energy between two particles is equal to zero (otherwise described as a VDW radii). The $(\frac{\sigma_{ij}}{r_{ij}})^{12}$ term controls the repulsive part of the equation, and scales with the power 12, whereas the $(\frac{\sigma_{ij}}{r_{ij}})^{6}$ component represents the attractive forces with respect to interatomic distance, and scales with the power 6. The attractive component is originally derived from London forces, and is generated from calculations of dipole attractions within gases.

In this thesis, all Lennard-Jones potentials are cut-off at a value of 12 Å, this is possible due to the exponential decay of the potential, and is necessary to reduce computational costs for atom pairs at large distances. As a result, to avoid artefacts caused by an abrupt cut-off, the potential is smoothed to 0, from an inter-particle distance of 9 Å, until 12 Å.

Coloumbic Interactions

Coloumbs law is assigned to calculate electrostatic interactions between any two particles in a simulation. This is necessary, as many biological molecules contain partial charges due to differences in electronegativity between atoms. The Coloumbic interaction between

particles i, and j, at distance r_{ij} is calculated with:

$$U_{ELECTROSTATICS}(r_i j) = \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \tag{2.13}$$

Where q_i, and q_j represent the partial charges of the particles, and ε is a term related to permittivity of free space, and the relative permittivity of the medium. Generally, electrostatic interactions are much more long range than Lennard-Jones potential interactions, and thus calculating them for each pair of atoms in a simulation system will rapidly become computationally too expensive for a system to make efficient progress. The treatment of long range electrostatics can be handled differently to reduce computational load. Most methods involve the generation of a cut-off, since all atoms interact with all others in the system, the number of calculations increases on the order of N^2, where N is the number of particles in a system. The key mechanism for dealing with this problem is discussed in Section 2.2.6.

2.1.4 Classical Forcefield Development

The parametrization of all the terms that make up a forcefield can occur in numerous ways. Generally, forcefields are generated using as much experimental data as possible, for example, thermodynamics data of organic molecules - as is the case with the OPLS forcefield (Jorgensen et al. [93]). Quantum mechanical calculation data is used as a supplement when either parameters are incomplete, or missing altogether. Forcefields must be utilized carefully - often parameters are generated for use under specific conditions, and using them outside of these conditions can lead to inconsistent results.

One primary way in which classical atomistic forcefields can differ, is in their approach to hydrogen atoms. Some forcefields treat hydrogens explicitly, and so they have mass, VDW radii, and electrostatic interactions, these are known as all-atom forcefields, and the AMBER FF series (Wang et al. [197]), along with GLYCAM (Kirschner et al. [102]), are

examples of these. Some classical forcefields attempt to increase simulation speed, and reduce calculations by not treating hydrogen atoms explicitly. In these systems, the hydrogen atoms are treated implicitly, and their parameters are combined with their parent carbon atom into a single site. This type of forcefield is known as a united-atom forcefield, the GROMOS series of forcefields (Oostenbrink et al. [139]) are an example of this treatment.

It is also worth noting, that forcefields are designed to be internally consistent, but unless otherwise mentioned, should not be mixed. Forcefields are often initially designed for specific molecules, for example, AMBER was initially designed for use with proteins and DNA fragments (Cornell et al. [34]), and OPLS (Kony et al. [104]) was initially designed with lipid molecules in mind. It would be a impractical, however, to mix these two forcefields, as they are not parametrized with the other in mind. Some forcefields are parametrized to be compatible with other, for example, the GLYCAM forcefield is parametrized as an extension to the AMBER protein, nucleic acid, and lipid parameters and thus, all parameters are compatible with each other.

2.2 Simulation Techniques

Many techniques employed by molecular dynamics simulations serve to reduce error, or increase efficiency (or very occasionally, both), the major techniques used in this thesis are listed here.

2.2.1 Timesteps

Timesteps employed during molecular simulations are crucial to their success. A timestep too large will make a system prone to crashing, and will increase the amount of error in calculations. Whereas a timestep that is too small will result in a stable and more accurate system, but will take a much longer time for the system to reach a time scale of interest. It is

generally suggested that a timestep that is 10 times less than the largest fastest motion within the system. In a fully atomistic system this is generally the vibration of a C-H bond, and so, a relevant timestep is on the order of 1 or 2 fs. Where a coarse grained, or reduced representation system is used, this rule still applies, and much larger timesteps can be employed. For commonly used schemes, timesteps of upwards of 40 fs are often utilized.

2.2.2 Periodicity and Periodic Boundary Conditions

Most simulations of biological systems must take into account periodicity. A box of particles may only be a finite size before calculations become too complex for a computer to perform them efficiently and effectively. This finite size means that there must be a cut off at the edge of a simulation box, and were we to simply treat this as a vacuum space, particles in contact with the vacuum would experience extremely unnatural forces; and we would find artefacts in the system. This problem is solved with the addition of Periodic Boundary Conditions (PBC) to the system, whereby the system is replicated in all dimensions around itself, thus eliminating any regions where molecules in a box are in contact with vacuum. Periodic Boundaries can be utilized with numerous box shapes, the most common of which is a cubic system, shown with periodicity in Figure 2.3. Other potential shapes include a rhombic dodecahedron, or a truncated octahedron (Figure 2.2), though the cubic-rectangular box is used for all systems in this thesis.

Particles within a periodic system are able to pass through the periodic boundaries, and so will appear on the opposite side of the original simulation box. Interactions will also work across periodic boundaries, though the systems use generally mean the short range interactions only take into account the primary periodic image. Schemes like Particle Mesh Ewald (PME) calculations are crucial to the accuracy and efficiency of periodic systems (Darden et al. [38], Essmann et al. [55]).

A final consideration is that care must be taken when selecting box sizes for periodic

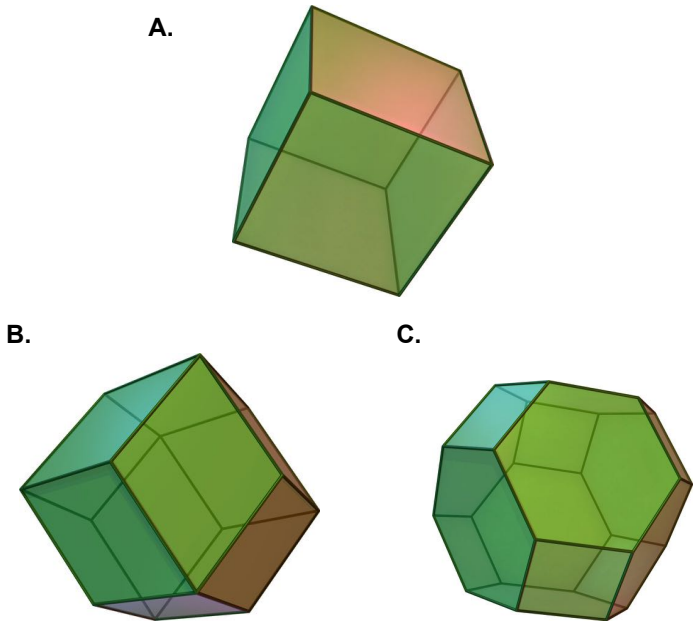


Fig. 2.2 Potential box shapes for use with periodic boundary conditions, showing A) Cubic-rectangular. B) Rhombic Dodecahedron. C) Truncated Octahedron.

systems, as particles in a box that is too small for them may be able to interact with their own periodic image, and thus introduce non-physiological conditions.

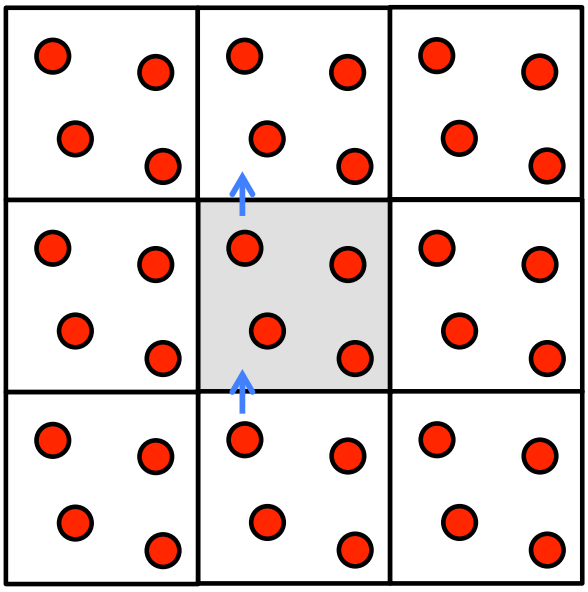


Fig. 2.3 Schematic of periodic boundary conditions, showing the "real" simulation box in the centre, grey, surrounded by periodic images. Particles are shown in red, and the movement of one (Blue arrows), shows how a periodic image will move to take it's place in the central box.

2.2.3 Temperature and Pressure Coupling

Traditional molecular dynamics simulations generally employ the constant NPT ensemble, in which the number of particles (N), the pressure (P), and the temperature (T), of a system are all kept constant. In order to keep the energy of the system from fluctuating by large amounts, a thermometer and/or barostat algorithm is generally applied to the system. These allow the system to be tailored to match experimental conditions, and numerous different thermostat and barostat algorithms are available. All simulations in this thesis were performed using the NPT ensemble, and were coupled using the Berendsen thermostat and barostat. Other schemes, including the Verlet scheme and Noose-hover coupling algorithms can also be used.

2.2.4 The Berendsen Thermostat

The temperature of a system is related to the kinetic energy of it's constituent particles. It follows from this therefore, that a useful way of adjusting and controlling the temperature of a given system is to adjust the velocities of particles to match those of a system of the correct temperature. The Berendsen thermostat employs this methodology by coupling to an external heat bath where the temperature is fixed at the required value (Berendsen et al. [12]). The velocities of each particle in the system are then rescaled each timestep by an amount proportional to the difference between reference temperature, T_{ref}, and the temperature of the system component T:

$$\frac{dT(t)}{dt} = \frac{1}{\tau_t}(T_{ref} - T) \tag{2.14}$$

Where τ is a parameter representing the strength of the coupling. If τ is equal to the timestep, the coupling will occur on every step, a larger value of τ will result in a weaker coupling, as the intervals between rescales are larger. Typically, the system is coupled

with different components separated, eg a protein will be coupled separately from the surrounding solvent. This is to prevent situations occurring where, for example, the protein temperature will increase dramatically, and the solvent temperature will decrease, resulting in a total system temperature that is correct, but components that are far from it.

2.2.5 The Berendsen Barostat

The Berendsen barostat works in a similar way to the Berendsen thermostat (Berendsen et al. [12]). By having an external "pressure bath", or set of reference pressures, the box size and coordinate locations for particles can be rescaled in order to accurately replicate the pressure required in the system. Whilst the mechanism of rescaling is more complex, the rate of change is essentially the same as for temperature coupling, given the reference pressure P_{ref}, and the pressure of the system component P, it can be shown as:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}(P_{ref} - P) \tag{2.15}$$

Again, τ represents the coupling parameter.

2.2.6 Particle Mesh Ewald

The Particle Mesh Ewald method, or PME, is an electrostatic summation method, which aims to speed up calculations of long range electrostatic interactions between particles (Darden et al. [38], Essmann et al. [55]). As discussed previously, long range electrostatics require vast computational power compared to other interactions (ie all bonded interactions), and the number of calculations scale exponentially with the number of particles in a system. A mechanism of preventing this, similar to the method employed with VDW interactions, would be to introduce a hard, or soft cut-off. However, due to the small rate of decay of the potential, introducing a cut-off would result in major inaccuracies in systems. The cur-

rent method of choice for reducing the computational burden of electrostatic calculations is PME, in which the interaction is split into two components, a short range potential, and a long range potential. The short range potential converges quickly in real space, whereas the long range potential converges quickly in Fourier space, due to the two components converging quickly in their respective spaces, it is possible to truncate them with only a small loss in accuracy, and a large increase in efficiency.

2.2.7 Particle Exclusions

In simulations it is possible to exclude particles joined to each other from self interactions other than those dictated by the bonded interactions within the forcefield. The number of bonds along particles can be excluded can be tailored to a particular molecule, in some cases for example, only atoms that are directly linked (number of exclusions = 1) need to be eliminated from interacting with each other. This allows the simplification of forcefield such that we can avoid problems caused by particles that are too close to each other, and would result in VDW or electrostatic potential values that are too large for a stable system.

In some cases, for example some cases of lipids, the exclusion number can be increased to a much larger number, to prevent the self interaction of particles held spatially close to each other, such as particles within lipid tails. Typically a value of 4 is used for classical molecular dynamics simulations, though in this thesis, values higher are used in the development of new models, and are described in each case they are adjusted.

2.2.8 Constraints

Constraining bonds and angles is a different methodology to the application of a harmonic potential. Constraints on a bond or angle define a set value that must be satisfied, and the value is simply reset each timestep, or timestep interval, so that it is correct. This is typically used for the constraint of rapidly vibrating bonds within an atomistic system,

such as bonds between a hydrogen and carbon, or for holding the distance between the two hydrogens within a water molecule the same, so as to reduce time calculating flexibility for large numbers of water molecules. In this thesis the LINCS algorithm (Hess [73]) is used in the parametrization of new structures, and in atomistic simulations where bonds or angles are constrained.

2.2.9 Position Restraints

Position restraints work in a similar way to constraints, in that they dramatically reduce the deviation of a bond or angle. Position restraints work by imposing a large energetic penalty to any deviation in the structure from the starting, or reference structure. Restraints are typically used in the setup of simulations, and are applied to the protein backbone in order to allow the rest of the system to relax around it. This typically occurs during the equilibration, though there is occasionally cause to apply restraints to a production simulation.

2.3 Energy Minimization

Energy minimization is an essential first step in the set up of any molecular dynamics simulation, though it's uses are not limited to simulations. Many tools that involve computational aspects, including crystallography, make use of energy minimization methods to improve structures. The concept of energy minimization is to allow structures to adjust their position in space, and internal spatial relationships, such that parameters from the forcefield are more accurately matched, without performing full molecular dynamics. In the case of a solvated protein, an energy minimization would reduce the energy of the system dramatically (hence the term energy minimization), by adjusting locations of protein sidechains to their minimal energy locations, and by allowing waters to move to reduce steric clashes, in an iterative manner. Two methodologies for the reduction of system energy are steepest descents

algorithm, and the conjugate gradient methodology (Leach [110]).

Steepest descents is able to reach a minimum most efficiently from a system that is far from an energy minimum, but will not always produce a structure that is completely minimized. This is the methodology employed in all system set up using the GROMACS software package.

The conjugate gradient algorithm is useful for reducing a structure down to a very low energy when starting from a position close to the energy minimum. Due to the different specializations of each algorithm, the AMBER software package typically uses both algorithms at once in a single minimization, performing steepest descents for the first half of the steps, and conjugate gradients for the second half.

2.4 Coarse Grained Forcefields

A major problem with the classical molecular dynamics methodology is that the timescales and system sizes are severely limited by the computational resources available for performing calculations on. Currently simulations are not often able to be run for much longer than a few hundred nanoseconds. Many biological behaviours that we wish to investigate occur on much longer timescales, often up to tens of microseconds, and so are totally inaccessible to an all-atom methodology. Many methods have been developed that reduce the system in question down into pseudo atoms, that are made up of constituent properties of many other atoms. This results in a huge reduction in the number of particles in the system, and hence the number of calculations, and the number of degrees of freedom are much lessened. Coarse grained forcefields vary from those representing hundreds of atoms as a single particle, through those that represent single lipids (around 20 atoms) in a single particle, to those representing around 4 atoms per particle. Most simulations in this thesis are run using the MARTINI coarse grained forcefield, which, on average, combines 4 non-hydrogen atoms into a single particle. A large part of the popularity of MARTINI stems from it's easy incor-

poration into the GROMACS software package. This thesis chronicles the development of, and current methodologies utilized in running simulations with the MARTINI model.

2.4.1 MARTINI lipids

The original MARTINI forcefield was parametrized exclusively for use with lipid molecules (Marrink et al. [123]). It employs a 4:1 ratio of atom mapping (approximately one coarse grained particle for every 4 non-hydrogen atoms). Most lipid molecule become on the order of 12 to 16 particles, depending on tail length. All particles in this scheme are given a mass of 72 Da, which is equivalent to four water molecules. Whilst it is certainly possible to use different masses for particles based on their constituent elements, for reasons of computational efficiency, all particles are given the same mass, and parameters are generated with this in mind.

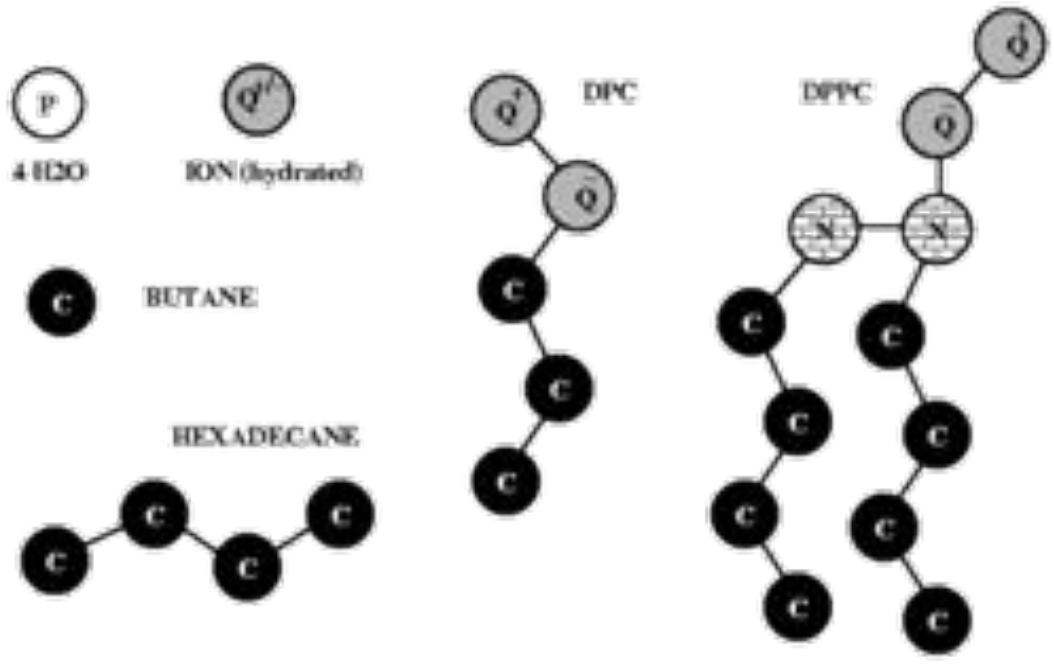


Fig. 2.4 Schematic of MARTINI lipid parameters, showing water molecules represented as a single polar (P) particle, and some examples of organic compounds, including DPPC. Figure adapted from (Marrink et al. [123]).

Particle Types

Four, distinct particle types are generated, and assigned to each component of a lipid based on the constituent atoms, the four particles types are:

- Polar (P) particles - Polar sites represent neutral groups that could dissolve readily into water, for example ethylene glycol.
- Apolar (C) particles - These represent completely hydrophobic groups, and are equivalent to chemically apolar groups.
- Nonpolar (N) particle - Nonpolar particles are used for groups where constituent atoms are partly polar, and partly apolar.
- Charged (Q) particles - Charged particles are reserved for ionized groups, such as ammonia.

Furthermore, charged (Q), and nonpolar (N) particles are further grouped into subtypes based on their hydrogen bonding properties, being denoted d and/or a properties if they are a donor or acceptor (or both). Groups that have no hydrogen bonding properties are given the type 0, and thus a nonpolar group that is able to both accept and donate to hydrogen bonds (for example 1-butanol), would be represented by an Nda particle.

Particle Interactions

Bonded, covalent interactions between particles are treated using the same basic principles as atomistic forcefields, but only pairwise, and three body restraints are taken into account, i.e. bond, and angle restraints. Bond distances are restrained with a harmonic restraint, with a force constant of 12.5 kJ mol^{-1} $Å^{-2}$, at a distance of 4.7 Å for all bonds. For each set of three consecutive particles that are bonded, an angle restraint is applied in the same manner, with an equilibrium value of 180°, and a force constant of 25 kJ mol^{-1} rad^{-2}. Exceptions to

the angle equilibrium value are made in some cases, for example, the headgroup-backbone linkage for one particle is adjusted to 120° to better reflect the physical structure.

Nonbonded interactions within this forcefield are described in the same ways as are generally used in atomistic simulations - a Lennard-Jones potential is used to replicate molecular attraction and repulsion, and electrostatics are treated with a Coloumb potential. In the case of the Lennard-Jones potential, five levels of attraction are described, and all particles are given the same cut-off, whereby particles are considered to be in contact when they are 4.7 Å apart. The interaction types go from level I, where particles are very attracted to each other, to level III, where particles are neither attracted, nor repulsive with respect to each other, up to level V, where particles are very repulsive. Intermediate values are also assigned at level II and IV. Each particle type is then assigned an interaction level for every other particle type, based upon the partitioning of these particles within a water/alkane mix, with comparisons to experimental and atomistic simulation data. The LJ potential is shifted to a value of zero between 9 Å and 12 Å. The interaction table for all particles can be seen in Figure 2.5.

Whilst the terms "attractive", and "repulsive" are used to describe the interactions, in isolation all of them will be attractive due to the nature of the LJ potential. Strength V interactions are named "repulsive" due to the ability for systems of oil and water to demix when these values are used.

Electrostatics in MARTINI simulations are generally treated in a similar way to those within fully atomistic forcefields, however, only particles with a total net charge of ≥ 1 are treated as charged. Other charges are implicitly included in the Lennard-Jones potential interactions. Coloumbic interactions are shifted to a value of zero between 0 Å and 12 Å, such that long range electrostatic interactions are not taken into account.

By matching these parameters to groups of atoms within lipids, parameters for a large set of lipids were generated. Upon simulation in water, these lipids spontaneously form bi-

		Q				P					N				C				
	sub	da	d	a	0	5	4	3	2	1	da	d	a	0	5	4	3	2	1
Q	da	O	O	O	II	O	O	O	I	I	I	I	I	IV	V	VI	VII	IX	IX
	d	O	I	O	II	O	O	O	I	I	I	III	I	IV	V	VI	VII	IX	IX
	a	O	O	I	II	O	O	O	I	I	I	I	III	IV	V	VI	VII	IX	IX
	0	II	II	II	IV	I	O	I	II	III	III	III	III	IV	V	VI	VII	IX	IX
P	5	O	O	O	I	O	O	O	O	O	I	I	I	IV	V	VI	VI	VII	VIII
	4	O	O	O	O	O	I	I	II	II	III	III	III	IV	V	VI	VI	VII	VIII
	3	O	O	O	I	O	I	I	II	II	II	II	II	IV	IV	V	V	VI	VII
	2	I	I	I	II	O	II	II	II	II	II	II	II	III	IV	IV	V	VI	VII
	1	I	I	I	III	O	II	II	II	II	II	II	II	III	IV	IV	IV	V	VI
N	da	I	I	I	III	I	III	II	II	II	II	II	II	IV	IV	V	VI	VI	VI
	d	I	III	I	III	I	III	II	II	II	II	III	II	IV	IV	V	VI	VI	VI
	a	I	I	III	III	I	III	II	II	II	II	II	III	IV	IV	V	VI	VI	VI
	0	IV	IV	IV	IV	IV	IV	IV	III	III	IV	IV	IV	IV	IV	IV	IV	V	VI
C	5	V	V	V	V	V	V	IV	IV	IV	IV	IV	IV	IV	IV	IV	IV	V	V
	4	VI	VI	VI	VI	VI	VI	V	IV	IV	V	V	V	IV	IV	IV	IV	V	V
	3	VII	VII	VII	VII	VI	VI	V	V	IV	VI	VI	VI	IV	IV	IV	IV	IV	IV
	2	IX	IX	IX	IX	VII	VII	VI	VI	V	VI	VI	VI	V	V	V	IV	IV	IV
	1	IX	IX	IX	IX	VIII	VIII	VII	VII	VI	VI	VI	VI	VI	V	V	IV	IV	IV

[a] Level of interaction indicates the well depth in the LJ potential: O, ϵ = 5.6 kJ/mol; I, ϵ = 5.0 kJ/mol; II, ϵ = 4.5 kJ/mol; III, ϵ = 4.0 kJ/mol; IV, ϵ = 3.5 kJ/mol; V, ϵ = 3.1 kJ/mol; VI, ϵ = 2.7 kJ/mol; VII, ϵ = 2.3 kJ/mol; VIII, ϵ = 2.0 kJ/mol; IX, ϵ = 2.0 kJ/mol. The LJ parameter σ = 0.47 nm for all interacion levels except level IX for which σ = 0.62 nm. Four different CG sites are considered: charged (Q), polar (P), nonpolar (N), and apolar (C). Subscripts are used to further distinguish groups with different chemical nature: 0, no hydrogen-bonding capabilities are present; d, groups acting as hydrogen bond donor; a, groups acting as hydrogen bond acceptor; da, groups with both donor and acceptor options; 1−5, indicating increasing polar affinity.

Fig. 2.5 Interaction table for Martini Lipid Forcefield, with interactions scaling from 0 (No interaction), to I (Very attractive), to V (Very repulsive). Table adapted from (Marrink et al. [123]).

layers. Simulations are able to run using a greatly increased timestep (MARTINI sometimes report 50 fs timesteps for these early works, but current MARTINI parameters recommend 20 fs), and result in simulations with greatly increased dynamics. As a result of this, MARTINI simulations are often quoted as being representative of 4 or 5 times more simulation time than was actually run, due to smoother dynamics. In this thesis, all times referenced for coarse grained simulations are "real" time, and are not multiplied. This work forms the foundation for the rest of the MARTINI parameters used in this thesis.

2.4.2 MARTINI Extension to Proteins

Shortly after the publication of the initial MARTINI model, parameters were released for use with protein structures. Generally, very little was changed in terms of the general mechanics of the forcefield, and so, it is completely compatible with lipid parameters. The MARTINI forcefield initial extension to proteins is named MARTINI v 2.1 (Monticelli et al. [131]).

Particle Types

The general rule of a 4:1 mapping scheme is retained for MARTINI proteins, though ring structures are an exception. Ring particles, such as the side chain of a histidine residue, use a less coarse mapping scheme, utilizing 3:1, or even 2:1 mapping. As a result of this, ring particle mass and size were adjusted to compensate for the reduced mapping scheme, the particle VDW radius reduced from the regular 4.7 Å, to 4.3 Å, and the depth of the Lennard-Jones potential was reduced by 75 %. Additional particle types were also added to the forcefield, whilst the principle particles are still used, namely: Polar (P), Nonpolar (N), Apolar (C), and Charged (Q), polar and apolar particles are further split into 5 subtypes, denoting the degree of polarity, from 1, which is low, to 5, which is high.

The mapping of amino acids is done by comparison of simulation results for coarse grained particles to experimental data on the water/oil partitioning of amino acid analogues. In this way, each amino acid side chain was mapped to it's correct coarse grained particle. PMF calculations were also used to distinguish between particles where different parameters gave similar results. The resulting mapping scheme is shown in Figure 2.6, where each amino acid is represented by a single backbone particle, mimicking the amino acid background, and a number of sidechain particles.

Particle Interactions

The bonded particle interactions within MARTINI v 2.1 were calculated from a dataset of 2000 protein crystal structures in the protein databank (PDB). A subset of 2000 protein structures was chosen, and the average positions of mapped coarse grained particles for each amino acid, in each type of secondary structure was calculated. Upon comparison with separate datasets of membrane proteins, the results were found to be the same. This database of distances and angles for each individual amino acid is used to generate the bonded parameters between backbone particles, with the distances being directly transferred. The distance

between backbone-backbone particles was set to 3.5 Å based upon the mean value from crystal structures. Force constants were applied to roughly replicate predicted vibrations for backbone-sidechain and sidechain-sidechain particles, and were iteratively matched based on small test simulations, and the crystal structure distributions of bonds and angles, and are totally independent of the secondary structure assigned to the protein. The values range from 12.5 kJ mol^{-1} $Å^{-2}$, up to 75 kJ mol^{-1} $Å^{-2}$ for harmonic bond restraints. Angle restraints were matched in the same way, and generally have a force of 25 kJ mol^{-1} rad^{-2}.

In order to attempt to restrain secondary structure of the protein, dihedral restraints were applied along backbone-backbone bonds when the atomistic structure is calculated through dssp as having remained the same for 4 residues in a row. At this point dihedral restraints are applied to attempt to hold the protein in roughly the starting conformation.

Nonbonded interactions were slightly updated from those generated for the lipid forcefield, novel polar (P), and apolar (C) particles, have values added for interaction using the same methodology as previously. Another change is that the potential energies of interactions, previously labelled in strength from I (attractive), to V (repulsive), are increased to ten total energies, I to X, such that better differentiation of energies for attraction can be distinguished.

Due to the extensive parametrization of protein molecules, the forcefield resulting from this work is unable to be run at timesteps of the same size as the lipid only iteration, and general timesteps employed range between 10 fs, and 20 fs.

2.4.3 Sweet MARTINI

After the success of the MARTINI forcefield extension to proteins, a small extension was made for the addition of carbohydrate parameters in 2009 (Lopez et al. [117]). In this case the 4-1 mapping scheme is retained, and each monosaccharide is generated as a triangle of coarse grained particles. The coarse grained particle types were selected for based

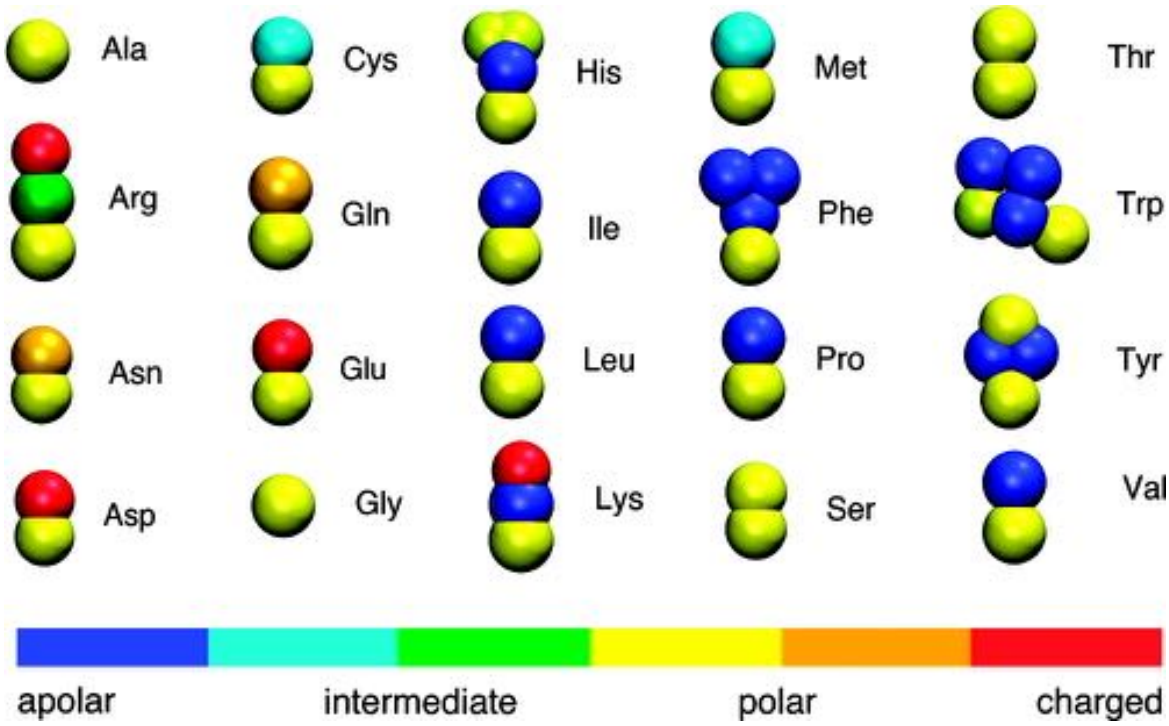


Fig. 2.6 Schematic of MARTINI amino acid parameters, showing amino acid side chains, in a coarse grained format, particles are coloured according to properties in the MARTINI v 2.1 forcefield. Figure adapted from (Monticelli et al. [131]).

upon comparison to experimental density and partitioning data, similar to the methods by which amino acid side chain particles were chosen. Bonded interactions for the carbohydrate molecules were generated based off atomistic simulation distributions. For carbohydrate molecules larger than a single monosaccharide, a different mapping scheme is used. Simply, a central particle is chosen, and linked to other carbohydrates that are covalently attached, and two "sidechain" particles are generated to match the rest of the properties of the molecule. In this case, a bonded interactions are determined by matching back to atomistic data, and a dihedral restraint is used along the carbohydrate chain, in a similar manner to protein dihedral restraints, in order to hold the system in place. This methodology has seen some success when used for the generation of large carbohydrate complexes, such as maltoheptaose.

Again, due to the extensive bond and angle restraints on these models, a large timestep

leads to instabilities in the model, and so carbohydrate simulations are generally run with a timestep of between 8 fs, and 10 fs.

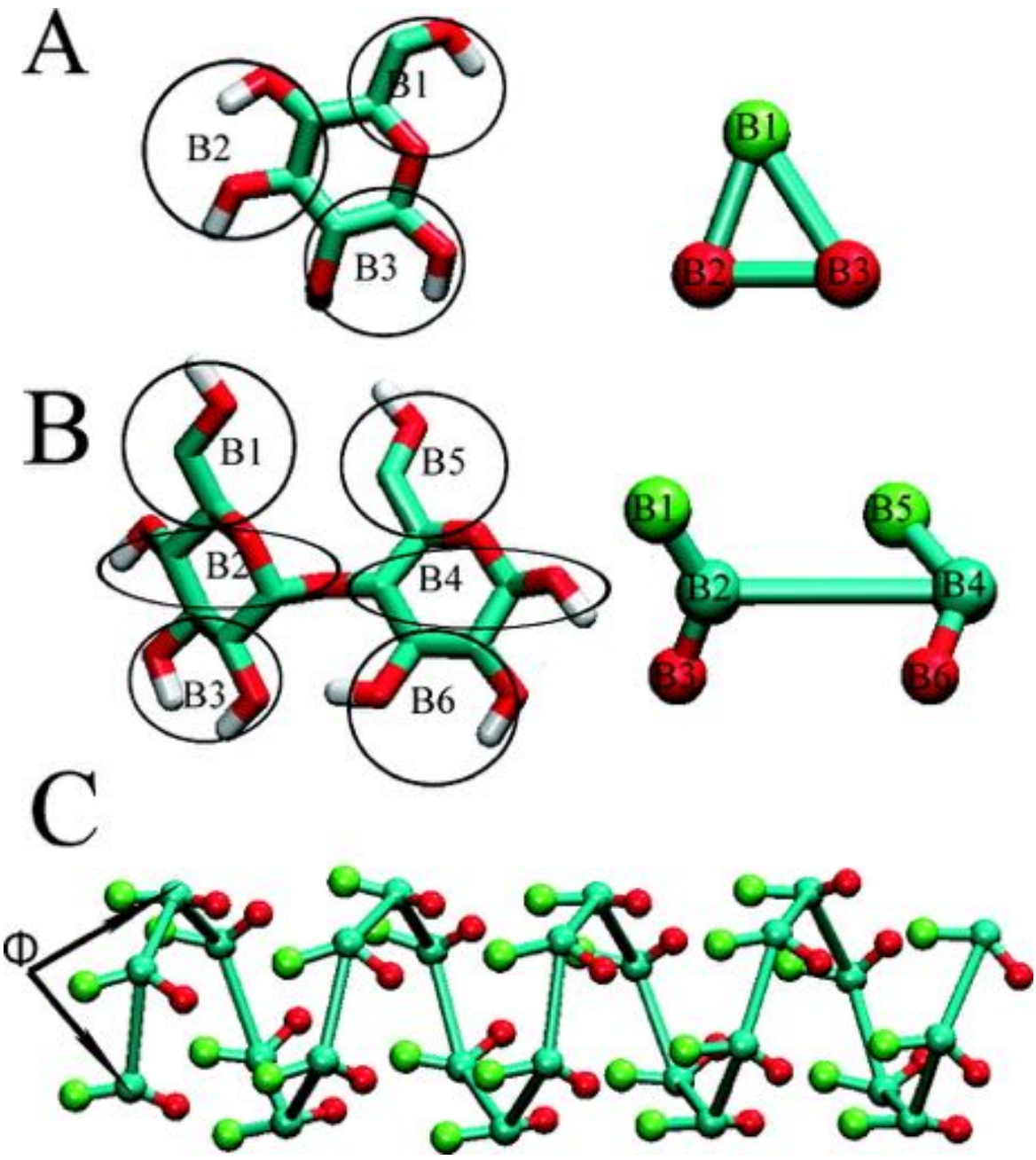


Fig. 2.7 Schematic of MARTINI carbohydrate parameters, showing A) Glucose. B) Trehalose. C) A large oligosaccharide. Figure adapted from (Lopez et al. [117]).

2.4.4 MARTINI v 2.2

Recently, the MARTINI group released improved parameters for the simulation of coarse grained proteins. This improvement to the forcefield was named MARTINI v 2.2 (De Jong et al. [41]), and seeks to correct some errors and inaccuracies in the earlier iteration. The key changes that effect protein structure and parametrization are the reparametrization of numerous side chain representations to more accurately match atomistic simulations - particularly side chains with rings, and adjustments to the way in which backbone-backbone bonds are defined for alpha helices, based upon experimental data.

These changes result in a model that more accurately represents proteins in a membrane environment, though extensive tests on soluble proteins have not been performed.

2.4.5 MARTINI Glycolipids

In 2013, a MARTINI forcefield extension containing glycolipids was released (López et al. [118]). In this iteration, glycolipids are generated in the same way as carbohydrate and lipid parameters, in which coarse grained particles are matched to atomistic groups of 4, 3, or occasionally 2, particles. Particle types are chosen through comparison to known mappings (i.e by comparison to chemical groups in amino acid sidechains or carbohydrates that include the same atoms). The structure of these molecules are generally held together with the use of an elastic network, whereby harmonic bonds are applied between all particles within a certain cut-off. In this way the general characteristics of the molecule are retained, without extensive parametrization.

The addition of these molecules to the forcefield again reduces the potential timestep that simulations can be run at, with certain glycolipid structures only being stable when run with a timestep of 5 fs.

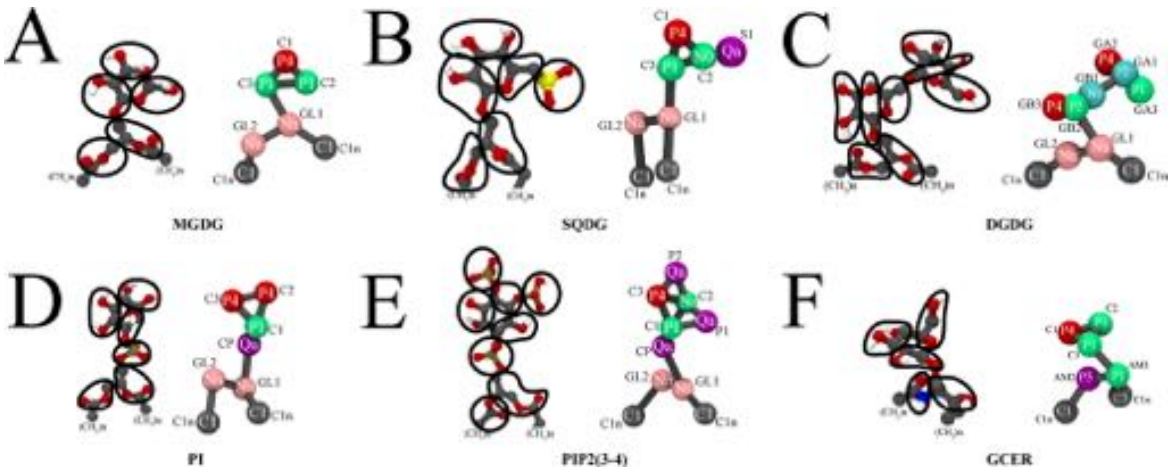


Fig. 2.8 Schematic of MARTINI glycolipid parameters, showing structures available for use. Structures are restrained with the use of an elastic network (Not shown). Figure adapted from (López et al. [118]).

2.4.6 MARTINI Secondary Structure Restraints

The use of dihedral restraints within the MARTINI protein forcefield, are generally unstable and function poorly when applied to larger proteins, or proteins not stabilized by a membrane. To overcome this, elastic networks can be applied to protein structures. Generally these system work by the application of multiple elastic bonds to all particles that are closer than a certain cut-off. Currently, for coarse grained system, two elastic network models can be used:

- Traditional Elastic Network Models - This model applies a harmonic bond to any two particles in the protein structure that are between 5 Å and 9 Å apart. The bonds are all the same, with no dependency on length, all of which have a force constant of 50 kJ mol^{-1} $Å^{-2}$. This bond system is applied to all particles excluding backbone particles that are linked within a chain.

- ElNeDyn Elastic Network - This model changes slightly the location of backbone particles in the mapping scheme, placing them on the Cα carbon, instead of the center of the mass of the backbone. Harmonic bonds are then applied between all particles

that are closer together than 9 Å, including backbone particles. In this way the elastic network is relied on to hold the backbone particles in position, rather than dihedral restraints. Force constants are set to 50 kJ mol^{-1} $Å^{-2}$, and there is no minimum cut-off for the network application.

Both methods have been shown to produce stable structures, but the aims of each is slightly different. The traditional elastic network model generally aims to look at global motions of the protein, whereas the ElNeDyn methodology is focussed on replicating the crystal structure. Comparisons between the two support this reasoning.

2.5 Markov State Modelling

Markov State Modelling (MSM) is a technique applied to data sets for measuring transitions between pre-defined states. Markov modelling can only be applied to stochastic systems that have "Markovian properties", that is, transitions between different system states depend on previous states in a non-determined way. Essentially, the immediate future state of a system can be influenced by the present state, but not the past state (Pande et al. [142]).

Biologically, this statistical method has been mainly applied to protein folding problems, by simulating small trajectories, and measuring the transitions between conformational states of the system, we are able to run arrays of simulations, pick out processes that appear to transition the fastest, and focus simulation efforts on them, as they likely show more important processes in the protein folding scheme. This is the principle of the folding@home (Shirts and Pande [170]) high throughput programme. The results from multiple simulations can also be combined using Markov models to map the exact pathways taken by molecules, and the rates at which changes occur, thus identifying key pathways.

Recently, however, tools have become available to allow the application of Markov State Models to any simulation data, in the form of software EMMA (Senne et al. [166]), and

more recently MSMbuilder (Bowman et al. [18]). Both pieces of software work in a similar way:

- Simulation data is fed into the model in the form of raw trajectories, or curated data to show behaviour of interest (e.g. distances system between components that represent conformational changes).
- Microstates are defined for the system whereby regions of relative stochastic isolation are identified.
- Transitions between microstates at lagtime T are counted into a transition matrix.
- Microstates are clustered into a series of N macrostates. These can be use defined, or defined with the use of a clustering algorithm.
- Transitions within a macrostate are ignored, and transitions between macrostates are aggregated to generate overall transitions between all N macrostates.

A schematic for this is seen in Figure 2.9. Currently, for biological simulations, Markov State Modelling has been applied to cases of molecule folding (Huang et al. [81], Pande et al. [142]), protein-drug interactions (González-Díaz et al. [64], Low-Nam et al. [119]), and protein conformational change (Noé and Fischer [135]).

2.6 Software Used for Simulations

Simulations were mostly performed using the GROMACS molecular dynamics program, which is open source, and publicly available. GROMACS versions 4.0 (Hess et al. [74]), and 4.5 (Pronk et al. [146]) were used in the generation of trajectory data.

For atomistic simulations requiring GLYCAM carbohydrate parameters (Kirschner et al. [102]), the AMBER11 (Case et al. [26]), and AMBER12 (Case et al. [25]) software package

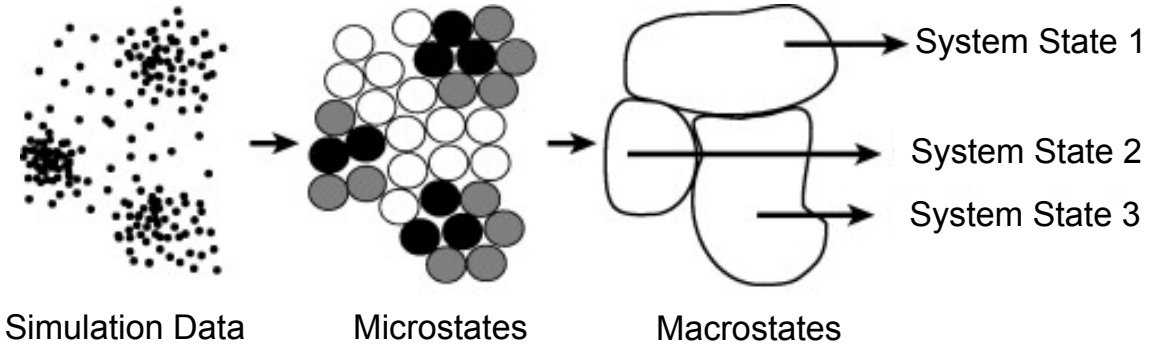


Fig. 2.9 Schematic of Markov State Model state assignments, showing microclusters, which are grouped into macroclusters, which then represent conformational states of a system. Figure adapted from (Bowman et al. [19]).

was used for performing molecular dynamics, along with AMBERtools12 software, which was used for the setup of simulations. AMBER was used due to the native compatibility between GLYCAM and the AMBER protein forcefields, and due to forcefield files being readily available for generation of structures within AMBER. This functionality is not yet implemented in GROMACS, and whilst it is theoretically possible to convert topologies for AMBER into a GROMACS format, it is unclear how similar the simulations would be.

For visualization and rendering of figures containing simulation particles, the software Visual Molecular Dynamics (VMD) was used (Humphrey et al. [83]).

Analysis was performed using tools available within the GROMACS software package, or with scripts written using Python, sometimes using the Python module MDAnalysis (Michaud-Agrawal et al. [127]), a module developed for easily dealing with trajectory data from a variety of simulation packages.

Chapter 3

Simulations of Antibody Aggregation

"All stable processes we will predict, all unstable processes we will control."

Freeman Dyson

In this chapter, I introduce mechanisms for the study of immunoglobulin aggregation. Focusing on the use of coarse grained molecular dynamics simulations, I firstly show that changes in the MARTINI protein forcefield, from version 2.1 to version 2.2, have made an improvement to the way in which protein-protein interactions are modelled. Studies are then performed on publicly available structures, to show that the methodology can predict regions of protein-protein interactions, and that specific details about how the two proteins interact can be investigated.

3.1 Introduction

3.1.1 Antibody Aggregation

Aggregation is a major challenge in the development of novel antibody-based therapeutics (Daugherty and Mrsny [40]). Aggregation is also a major bottleneck in the implementation

of many therapeutics that are otherwise clinically effective (Perchiacca and Tessier [144]). The large variation, and the extremities of the conditions that antibody therapeutics are produced in means they must be extremely stable in order to function without causing severe immune reaction (Rosenberg [155]). The reduction in the monoclonal protein population, and consequently, the increase of polymer population into aggregates, is a leading contributor to the ineffectiveness of many otherwise promising therapeutics (Sharma [167]). Generally, molecules can make it relatively far into trials before they are failed for excessive aggregation. Further to this, the types of aggregates formed by monoclonal antibody populations, and the processes that drive and propagate these aggregates are relatively poorly defined, though some progress has been made (Andersen et al. [4]). It is also well known that there are distinct differences in the aggregation propensities of commercially available monoclonal antibodies depending on their Immunoglobulin class, Complementarity Determining Regions (CDRS), and glycosylation patterns (Chennamsetty et al. [30], Voynov et al. [194, 195]) among other things. Furthermore, there have also been numerous physiochemical factors proven to play roles in the aggregation of monoclonal populations, including: pH environment (Hari et al. [71], Sahin et al. [158]), mechanical stresses (Kiese et al. [99]), interactions with nearby surfaces such as stainless steel (Bee et al. [11]), concentration (Chennamsetty et al. [28], Daugherty and Mrsny [40]), and temperature (Ishikawa et al. [86], Vermeer and Norde [191]).

3.1.2 Previous Work

Previous work has attempted to address the problem of antibody aggregation at different scales. Work from Chennamsetty (Chennamsetty et al. [28, 29]), has attempted to develop surface analysis tools that can assess the aggregation propensity of an antibody-based structure through sequence and surface alone (Spatial Aggregation Propensity), Figure 3.1, A, and through attempting to calculate hydrophobic interactions through a variety of surface

tools (Lauer et al. [108]). These tools are fast, and attempt to provide a rapid screening technique for the assessment of potential molecules. They are severely limited, however, in that they cannot take into account kinetics of interaction, or flexibility of the protein. They have thus far produced very mixed results. At the other end of the scale, large scale studies have been performed on highly reduced representation structures, whereby an entire antibody is converted into 12 or 26 "sites", and varying concentrations simulated to investigate mechanisms of aggregation. Though the granularity of the method means that specific mechanisms of interaction, and interaction pathways are unable to be discerned, they had some success in predicting aggregation rates for two different mab types (Chaudhri et al. [27]) Figure 3.1, C.

Fab and Fc proteins have been shown to be stable when separated from the whole antibody molecule (Vlasak and Ionescu [192]), there are also indications that Fab and Fc regions interact with each other with different characteristics in solution. In particular – studies suggest that for some structures, Fab-Fab interactions are reversible and specific (Kanai et al. [96]), Figure 3.1 B, and that they are caused by regions in the CDR, whilst Fc-Fc interactions tend to involve the CH2 domain of the protein, the same region that is implicated in the interactions with FcγR proteins (Wu et al. [203]) for IgG1.

3.1.3 Coarse Grained Molecular Dynamics

Coarsegrained molecular dynamics (CG-MD) has been used successfully to predict protein-protein interactions, exemplified by numerous model systems (Peng [143], Wu and Shea [202]) such as the Cohesin-Dockerin system (Hall and Sansom [70]), and Barnase-Barstar (Basdevant et al. [9]). Non-specific interactions have yet to be studied, and are too long to be investigated in depth using traditional atomistic simulations. By using CG-MD we are able to more successfully, and more comprehensively investigate non-specific aggregation of proteins. In particular, using CG-MD we are able to look at both specific surface residues

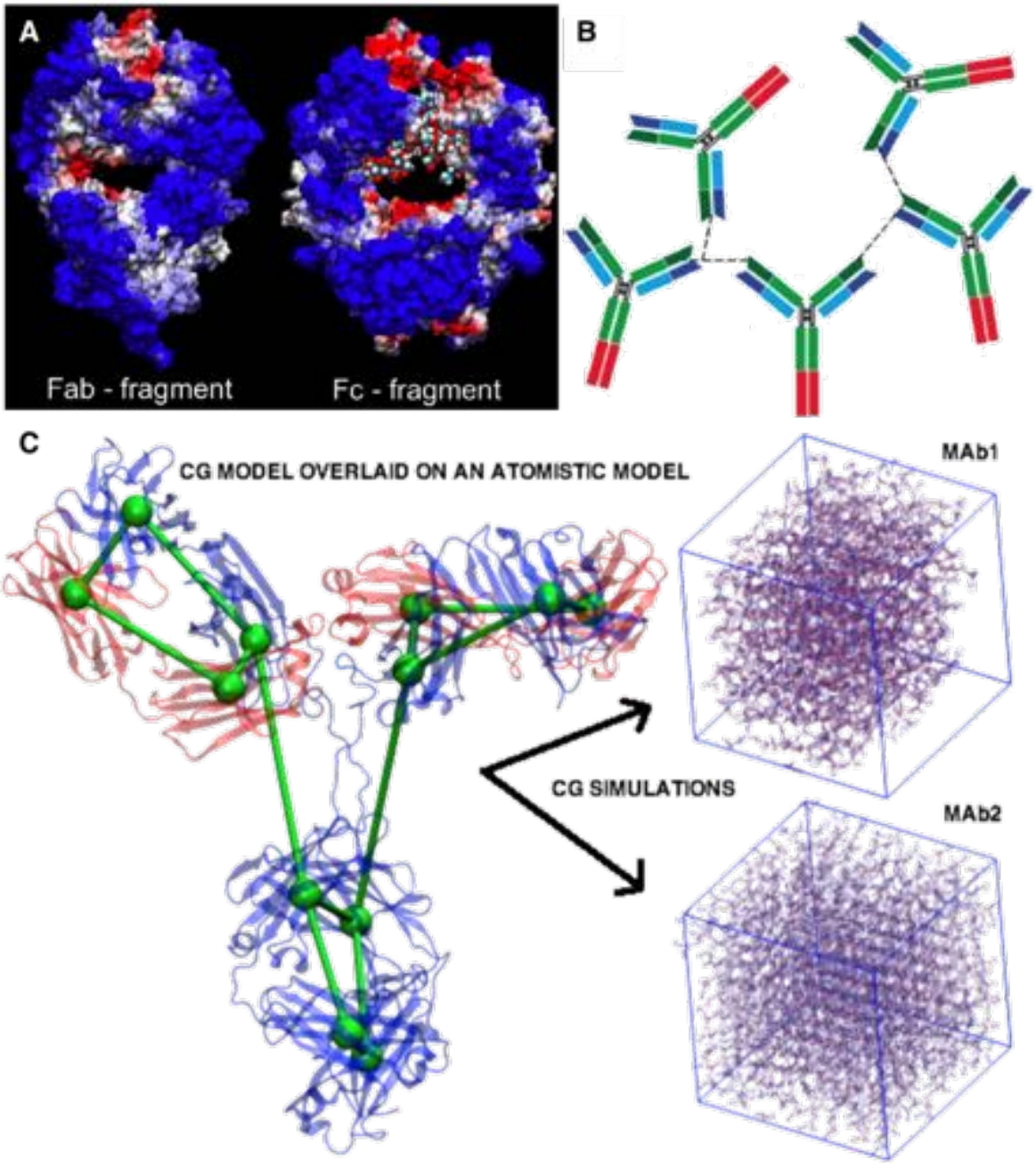


Fig. 3.1 Previous work on antibody aggregation, showing A) Spatial Aggregation Propensity (SAP) for a Fab protein, figure adapted from Chennamsetty et al. [29]. B) Transient interactions between Fab fragments, figure adapted from Vlasak and Ionescu [192]. C) Reduced representation simulations of huge systems figures adapted from Chaudhri et al. [27].

involved in interaction (Similar to those identified by Trout using SAP (Chennamsetty et al. [29])), and mechanisms of interaction between different proteins (Similar to those identified by Voth using reduced representations (Chaudhri et al. [27])). Currently CG-MD is an ideal method to gain both specific details, and larger-scale interaction and mechanistic information from the same technique.

We aim to gain new information about the self association of antibodies by utilizing CG-MD. By combining studies of small-scale simulations, with large-scale simulations containing many proteins, we can gain information about specific interactions between two proteins, and hopefully, information about more complex, multi-protein interactions. By combining these two approaches into a single study, we can observe a more comprehensive picture of the system dynamics and interactions. There is also potential to gain information about the pathways by which these molecules encounter each other - whether they simply come together and become locked together, or take time to "seek" the surface of each other until they find a region of interest.

3.2 Methods

System Setup

Simulations were run using GROMACS version 4 (Hess et al. [74], Lindahl et al. [113]). Proteins were converted from atomistic detail into coarse-grained structures using an in-house protocol previously described in the literature (Bond et al. [15], Bond and Sansom [16]). Structures were built with the Martini v 2.1, and the more recent Martini v 2.2 coarse-grained forcefields (de Jong et al. [42], Monticelli et al. [131]). Non-bonded interactions between all particles are treated as a Lennard-Jones potential, which is shifted to zero between 9 and 12 Å. Electrostatic interactions are treated with a Columbic potential, and shifted to zero between 0 and 12 Å. Secondary structure of the protein is maintained

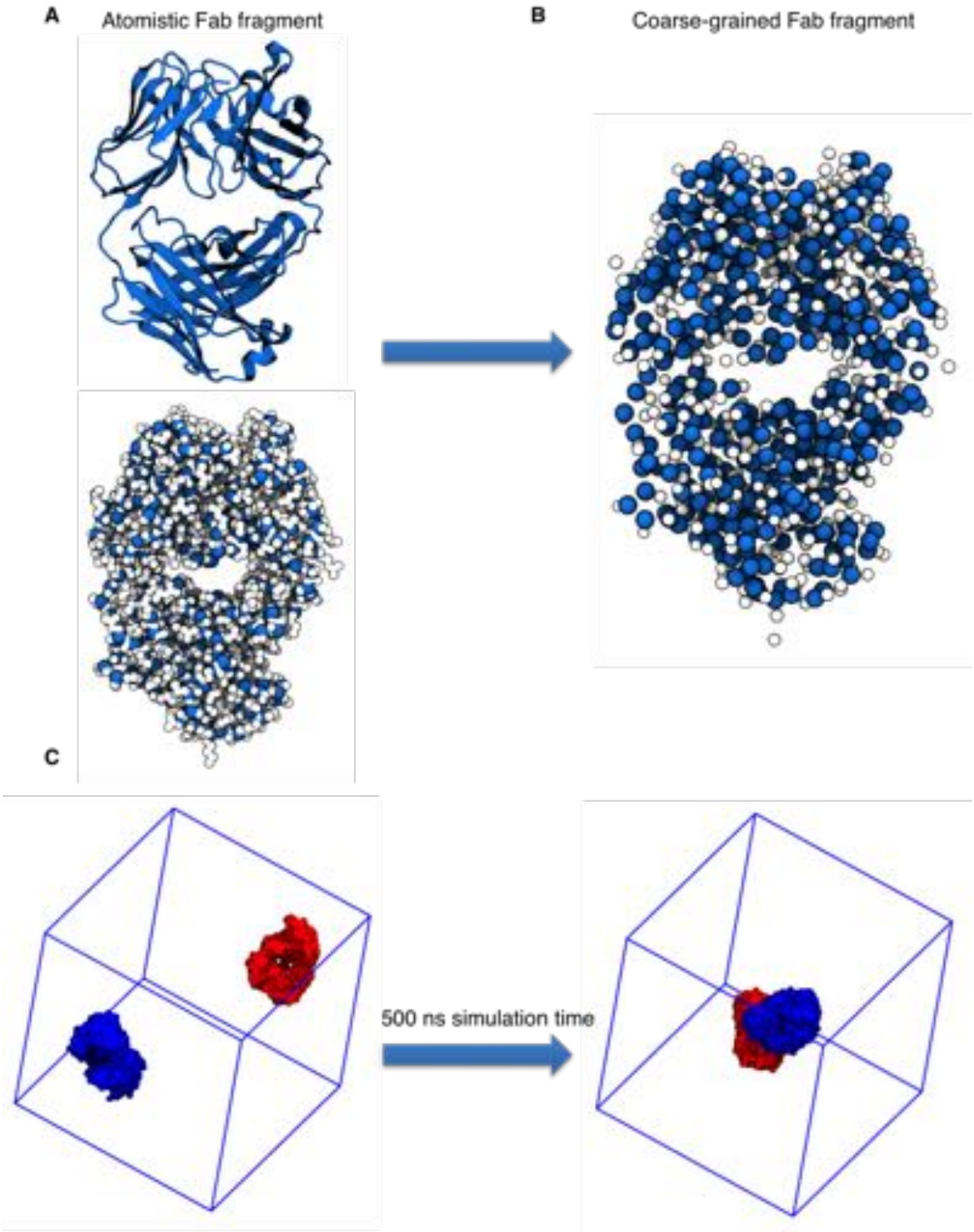


Fig. 3.2 Setup of proteins and systems for investigating aggregation, showing A) Atomistic structures of Fab fragment from structure 1CE1, in ribbon and spacefill representations, B) Coarse-grained representation of Fab structure from 1CE1 in spacefill representation, C) An example aggregation simulation for investigating interactions of two Fab proteins over 500 ns.

via the use of an elastic network, built with a cut-off of 7 Å (Atilgan et al. [6]), using the ElNeDyn methodology. The elastic network was then applied with a force constant of 10 kJ/mol. The elastic network is built by treating the entire protein as separate chains (Heavy and Light), meaning that the elastic network is applied within each chain individually - but not between chains. This allows the protein to undergo conformational changes, and means that it is not completely restricted to it's starting structure.

A Berendsen thermostat and Barostat (Berendsen et al. [12]) were used to couple the temperature and pressure at 310 K and 1 bar respectively.

In order to compare the way Fab and Fc molecules aggregate, we simulated each fragment of the protein in isolation from the entire IgG. The crystal structures that served as a basic for the Fab and Fc were 1CE1 and 3AVE. Atomistic and coarse grained representations of the structure 1CE1 can be seen in 3.2, A and B respectively.

To produce an aggregation simulation containing two proteins, both proteins were placed around 110 Å apart in a 160 x 160 x 160 Å cubic box, well beyond the cut-off distance for interactions, including taking into account periodic boundary conditions. Proteins were randomly orientated with respect to each other using the genconf command in GROMACS, and the system was energy minimized for 500 steps. After the addition of water and ions to the system, it was energy minimized for a further 1000 steps before CG-MD simulations were run with the protein backbone constrained for 50 ns. A 20 fs timestep was employed to then run a further 500 ns of CG-MD simulation on the completely unrestrained system, consisting of approximately 30,000 particles (Figure 3.2, C). For comparing forcefields, the setup protocol is exactly the same, aside from application of a different forcefield (.itp and .top files) to the system (Table 3.1).

The same procedure was performed for the setup of systems containing more than two proteins – with box sizes expanded appropriately, with the minimum protein-protein distance larger than the non-bonded cut-off. Larger simulations, however, require much larger

Table 3.1 Table showing simulations performed for comparison of Martini forcefield types

Forcefield	Proteins Simulated	Simulation Length (ns)	No. of Repeats
Martini v 2.1	2 x Fab	500	8
Martini v 2.1	2 x Fc	500	8
Martini v 2.2	2 x Fab	500	8
Martini v 2.2	2 x Fc	500	8

Table 3.2 Table showing simulations performed for investigation of aggregation

Proteins Simulated	Simulation Length (us)	Approx no. of Particles	No. of Repeats
2 x Fab	0.5	30,000	50
2 x Fc	0.5	30,000	50
4 x Fab	1.5	120,000	4
4 x Fc	1.5	120,000	4
8 x Fab	3	300,000	4
8 x Fc	3	300,000	4

amounts of CPU time, and extended simulation time, in order for all proteins to have a chance to interact (Table 3.2). Larger system sizes allow the investigation of the effects of multiple protein interactions on aggregation. Different system sizes were simulated, containing 2, 4, or 8 proteins. For systems containing 2 proteins, 50 simulations were performed for 500 ns each, for 4 proteins, 4 simulations were performed for 1,500 ns each, and for systems containing 8 proteins, 4 simulations were performed for 3,000 ns each. A larger number of 2 fragment simulations were performed in order to get better sampling of the interfaces with which the two proteins make contact. These simulations take comparatively less computational time, and thus are easier to run larger numbers of.

Analysis

Standard analysis metrics such as RMSD, RMSF, and distance, were generated using the appropriate GROMACS tool (eg. g_rms, g_rmsf or g_dist). For the analysis of interactions, first it is necessary to define when two molecules are interacting. Throughout the

simulations, two proteins are defined as interacting when any particle within one molecule (including amino acid side-chains) comes within 7 Å of any particle within the other. This distance criterion has to be held for a minimum of 5 ns before it is counted as an interaction.

Clustering of interaction surfaces was done initially by converting the data from each simulation into a binary string. If a residue on a single protein, within a single simulation was calculated as interacting with any residue of an opposing protein, the residue is counted as a 1, any other residue is counted as 0. This leads to the generation of a binary string for each protein in each simulation, which is a direct calculation of the interaction "fingerprint" for each protein in each simulation. The distance between each binary string is then calculated using the Manhattan distance – in which the number of "flips" for each value in the string (1's that must change to 0, and vice versa) that are needed to make the strings equal, are counted, and this count used as a distance between two compared strings. Strings can then be clustered based on these distances, and using the Ward clustering methodology to calculate inter-cluster distances.

To analyse mechanistic data about the interactions, points were chosen on the protein to represent regions of interest. The distances between these points were then used to populate a 2 dimensional surface, where regions are coloured based on how populated a region of the plot is. Distances chosen were, firstly, the center of mass of a region on the protein identified to be the major aggregation "hot spot" in simulations of 2 proteins, secondly, the center of mass of the protein.

3.3 Results and Discussion

3.3.1 Martini v 2.1 vs Martini v 2.2 - comparing forcefields

A comparison of the two Martini protein forcefields, Martini v 2.1 (Monticelli et al. [131]), and the newer Martini v 2.2 (De Jong et al. [41]) was performed, to look at performance

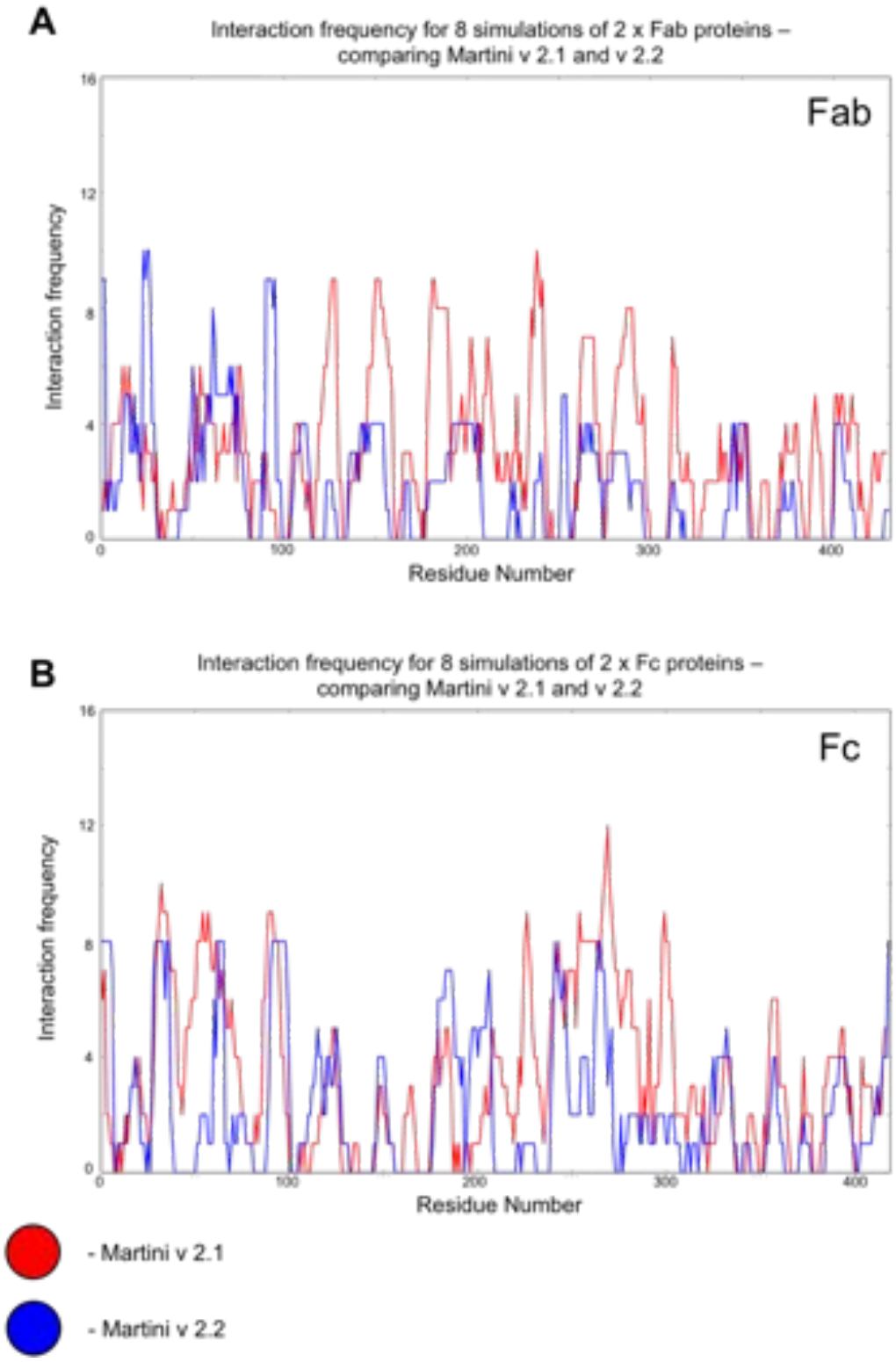


Fig. 3.3 A comparison of Martini v 2.1 (Monticelli et al. [131]) and Martini v 2.2 (De Jong et al. [41]) for the study of protein-protein interaction. An interaction is counted when two proteins are within 7 Å of each other for a minimum of 5 ns. Showing A) Fab protein interactions, there are only 3 peaks present for simulations with Martini v 2.2, compared to 7 for Martini v 2.1, the "base level" of interactions also appears lower, indicating that the interactions are more specific with Martini v 2.2. B) Fc protein interactions, showing similar peaks, but with slightly more refined peaks when using Martini v 2.2.

in studying protein-protein interactions. In order to compare the two forcefields, 8 simulations of each pair of proteins (Fab and Fc) were performed, using each forcefield (Table 3.1). The key improvement in Martini v 2.2 is the reparametrization of amino acids with ring side chains (phenylalanine and proline), and a reduction in the strength of some polar interactions.

Behaviour of the proteins was seen to be different in each forcefield, (Figure 3.3) by averaging the residues in contact with the opposing protein over all simulations, there is a distinct difference between the two behaviours for Fab protein. For both systems that were run using the Martini v 2.1 forcefield – the residue count does not highlight any distinctive regions. The opposite is true for the simulations run using Martini v 2.2, particularly for Fab proteins, in that a few distinct peaks are visible after a relatively small (8) number of simulations. For Fc proteins, the difference is not as large, but there are certainly a higher number of "non specific" interactions in 2.1 vs 2.2. The implication from these results is that the Martini v 2.2 forcefield results in more specific interaction surfaces compared Martini v 2.1, when used for the investigation of soluble protein-protein interactions, due to the fact that, within a relatively small number of simulations, there is a convergence in the most common interaction interfaces. For the investigation of protein-protein interactions, Martini v 2.2 was used in the remainder of the study.

3.3.2 Simulations of Pairs of Proteins

For the investigation of small systems containing 2 proteins, either 2 x Fab or 2 x Fc, 50 simulations of each system were run, as shown in Table 1. A large number of repeats were run, mainly because the systems being small (approx 30,000 particles), they are able to run in a comparatively short (24 hours on 12 CPU cores) time period.

Reversibility of Interactions

Throughout all simulations run using pairs of proteins, interactions were seen to generally be irreversible. All interactions resulted in dimers that did not separate at any point within the simulations. This is unlikely to be the case for similar systems in vivo, where is it generally assumed that thermal fluctuations will result in the molecules transiently binding in different orientations, until highly stable complexes are formed. Work from has shown that the energy required to break protein-protein complexes within Coarse Grained simulations is orders of magnitude higher than expected through comparison to atomistic level simulations (Stark et al. [180]). This increased "stickiness" of the forcefield is a known limitation of the methodology, though previous work on Cohesin-Dockerin (Hall and Sansom [70], and Barnase-Barstar (Basdevant et al. [9]) has shown that, although the forcefield parameters result in interactions that are too strong, the complexes formed, and residues involved in interactions can be biologically relevant.

Fab Proteins

Plotting the frequency of interaction against residue in each protein shows distinctive regions of high-frequency contact between the two proteins (Figure 3.4, A). In the case of Fab proteins, two regions of extremely high interactivity are seem at around residues 20-30 and 80-90 within the protein structure, with around 60 proteins (out of a potential 100 protein surface interactions – 2 from each of the 50 simulations) using these residues for a sustained interaction. There is also a smaller peak at around residue 250 in the protein that is also involved in numerous sustained interactions. Plotting these interaction frequencies on the surface of the protein allows us to more easily see where interactions are in relation to each other (Figure 3.4, B).

Clearly, for Fab proteins – a majority of the interactions with other Fab proteins occur between CDR regions. This matches extremely well with previous experimental work,

suggesting antibodies form transient networks through interactions between Fab domains (Vlasak and Ionescu [192]). This is also what we would expect, given that the CDRs are the main region of the molecule that changes, and that different CDR regions are associated with different aggregation propensities. This also matches with results from previous work (Chennamsetty et al. [30]), where the spatial aggregation propensity tool identified regions around residue 95 – 100 within an undisclosed IgG1 mab for mutational studies. These mutations were then shown to produce more stable proteins for expression and purification.

We can investigate further the regions of interaction by clustering similar interactions together with the aim to gain insights into the relationships between different interaction modes (Figure 3.5). For the 50 simulations of Fab proteins (part A), we see the two peaks around residues 20-30, and 80-90 are present, however there appears to be no single, or distinctive interaction mode – the other residues involved in interactions are very scattered and variable. This infers that interactions between two Fab proteins are centred on a small "hot spot" on the surface of the protein, but there are no interaction modes, or specific interaction poses, required for the proteins to adopt, in order for them to stick together. Interaction center around a specific site, but are varied in character.

In order to attempt to gain some mechanistic insight into how two proteins come together within the simulation, and if there is a conserved and distinctive pathway to interaction, an "interaction landscape" was generated. By measuring two distances in the Fab-Fab system we are able to plot the frequency at which each distance in populated within all 50 simulations. The two distances chosen are:

1. Distance between the two centres of mass of the separate proteins, d1.

2. Distance between residues 20-30, and 80-90 for each protein (i.e distance between two regions shown to interact in most simulations, d2.

The interaction landscape for Fab proteins is shown in figure 3.6. Firstly, the interaction landscapes reveal quite clearly that there is a "trap" type effect involved in the systems –

whereby the “background” frequency for distances that are not extremely high frequency (i.e. not interactions) is higher whilst the two proteins CoM are under 7 nm. This indicates that whilst the two proteins are within 7 nm of each other, more energy is required for them to drift further apart, and thus this is a fairly rare event. In fact, by visualizing most simulations, the proteins are seen “exploring” the surface of each other before coming into contact, rather than immediately interacting as soon as they come into contact with each other. This can be seen in the mid-high frequency “smear” seen in the interaction landscape, caused by the proteins coming close to one another, before somewhat exploring the landscape of the opposing protein, and then coming into contact, demonstrated by high density red spots.

Red regions of the landscape indicate regions at which the proteins spend a large frequency of simulation time, and therefore generally indicate interactions. By extracting the conformation that the two proteins adopt within these regions we are able to see what some of the most common interactions are. By extracting 6 of the most common interactions from the landscape (Figure 3.7), we can see that all of the interactions are centred around the previously identified interaction “hot spot”, albeit with very different, seemingly random orientations. Through the combination of analysis methods, and visualization of the trajectories, the way in which these two proteins interact seems to be:

- The two proteins drift randomly from the start.
- Upon coming within a close distance (7 nm) of each other, the proteins rarely move further apart.
- The two proteins “scan” the surface of each other looking for “hot spots” to interact with, indicated by the higher density "background" within this region of the plot.
- The two proteins interact with each other, often using the “hot spot” region identified, as shown by the snapshots of systems in high density regions of the plot.

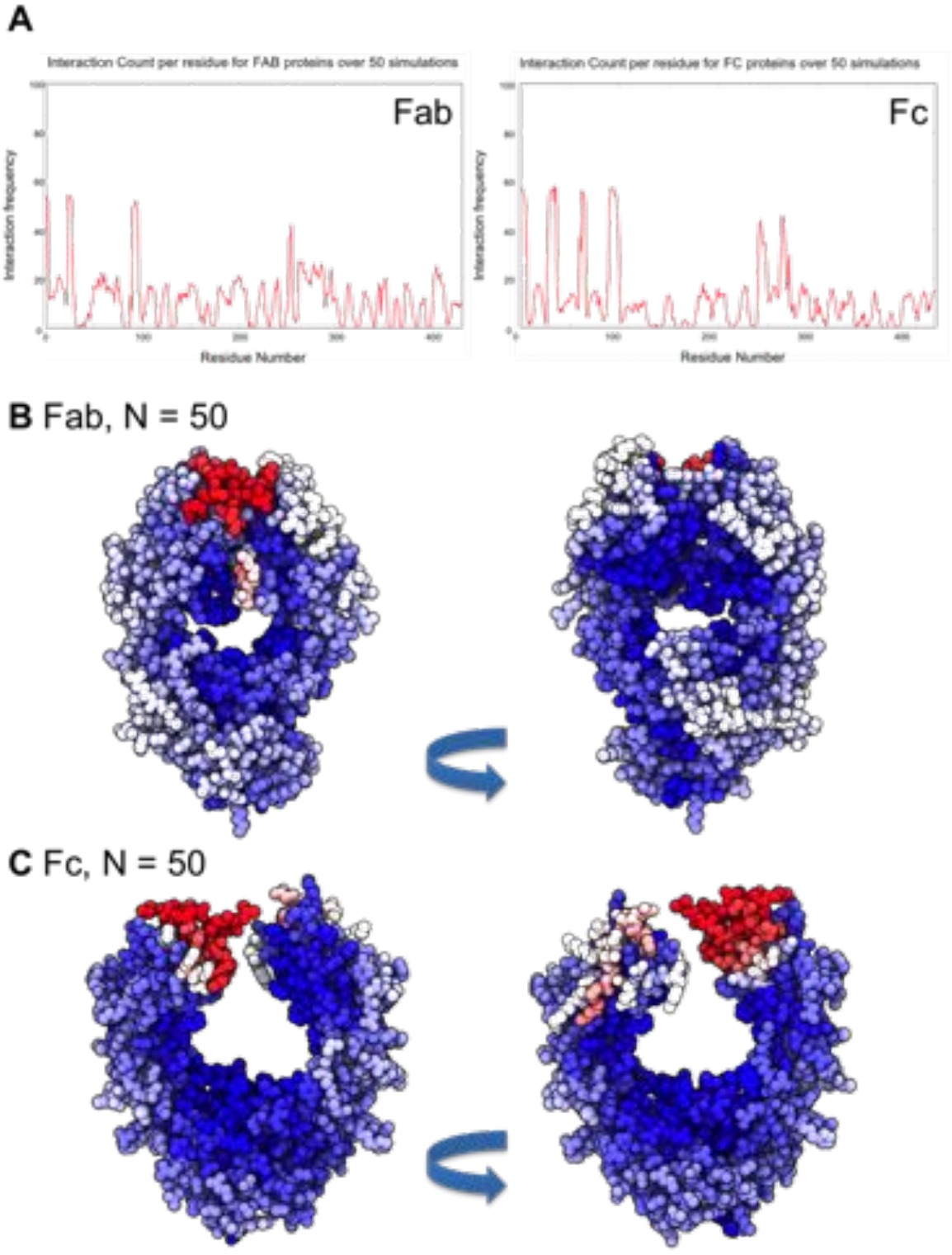


Fig. 3.4 Interaction frequencies for 50 simulations of Fab or Fc protein-protein interactions. A) Interaction frequency plots for Fab and Fc proteins. Peaks represent regions of the protein where large numbers of interactions with a partner protein occur B) Surface plot of Fab protein 1CE1 showing regions of high interaction by colour. Regions highlighted in red are inolved in high numbers of interactions (and correlate to the peaks on the graph). C)Surface plot of Fab protein 3AVE showing regions of high interaction by colour. Regions highlighted in red are inolved in high numbers of interactions (and correlate to the peaks on the graph).

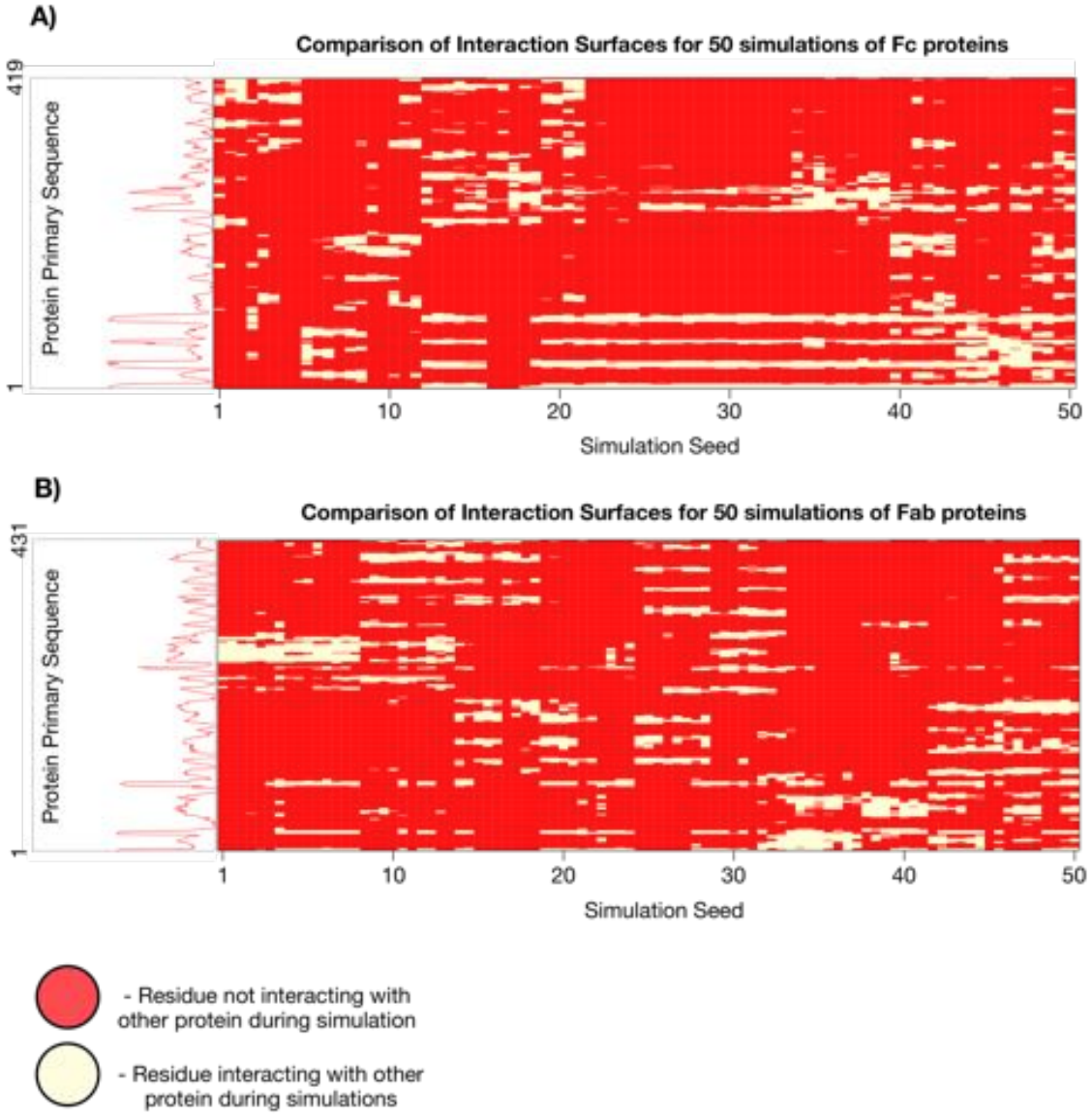


Fig. 3.5 Cluster analysis showing 50 simulations aligned next to each other for A) Fc, and B) Fab proteins, with each residue represented in red if it does not interact with the opposing protein within a simulation, and yellow if it does. An interaction is counted if the residue comes within 7 Åof any particle in the other protein for longer than 5 ns in the simulation. The order of interaction surfaces is generated based on ward clustering of the binary interaction fingerprint for each structure.

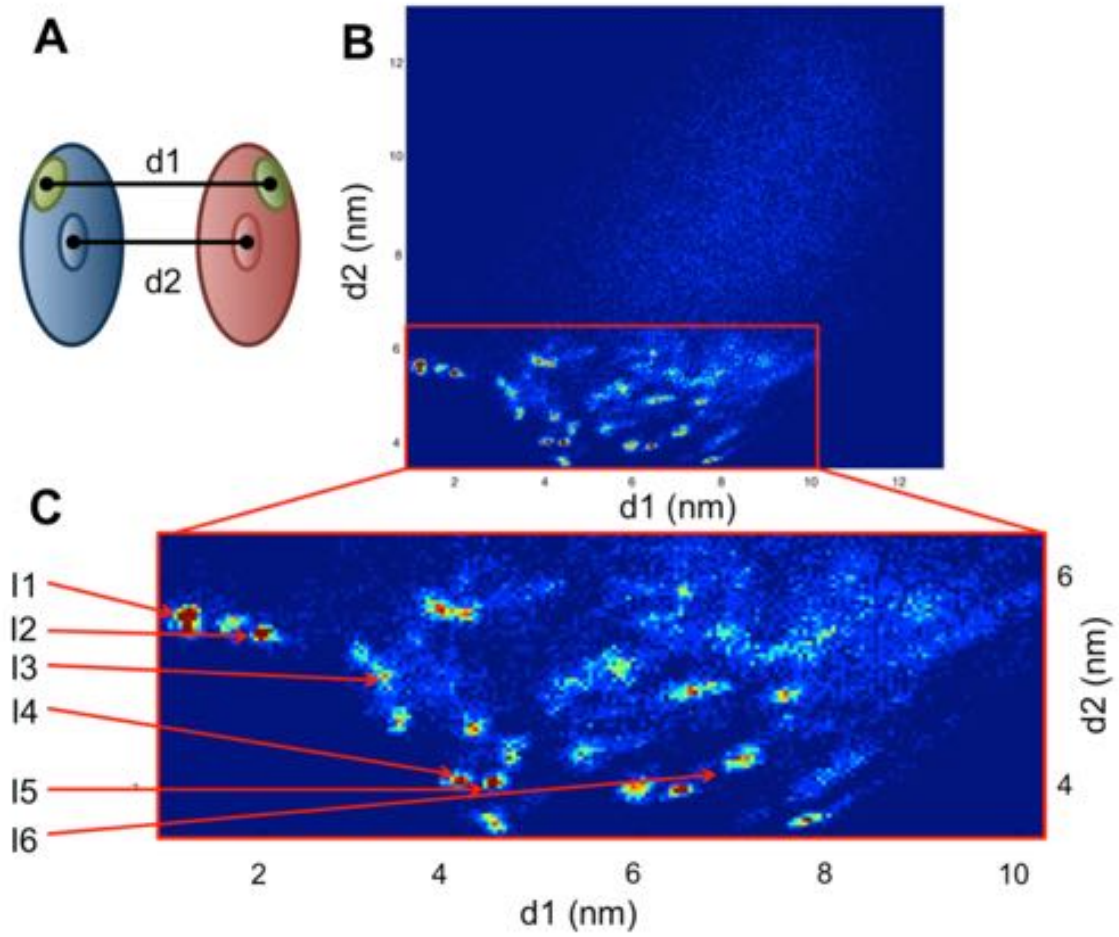


Fig. 3.6 Interaction landscape plot for 50 simulations of Fab - Fab interactions. A) Key showing two distances used to generate the plot: d1) Distance between the center of mass of each protein d2) Distance between the CDR region of the protein found to interact through frequency analysis (Figure 3.4). B) Interaction landscape for the two distances. Whereby blue represents a low frequency, and red "hot spots" indicate regions of space where the proteins spend a large amount of time. C) Zoomed in region of the plot where the protein come within a close distance to each other, showing highlighted regions of interest (I1 to I6).

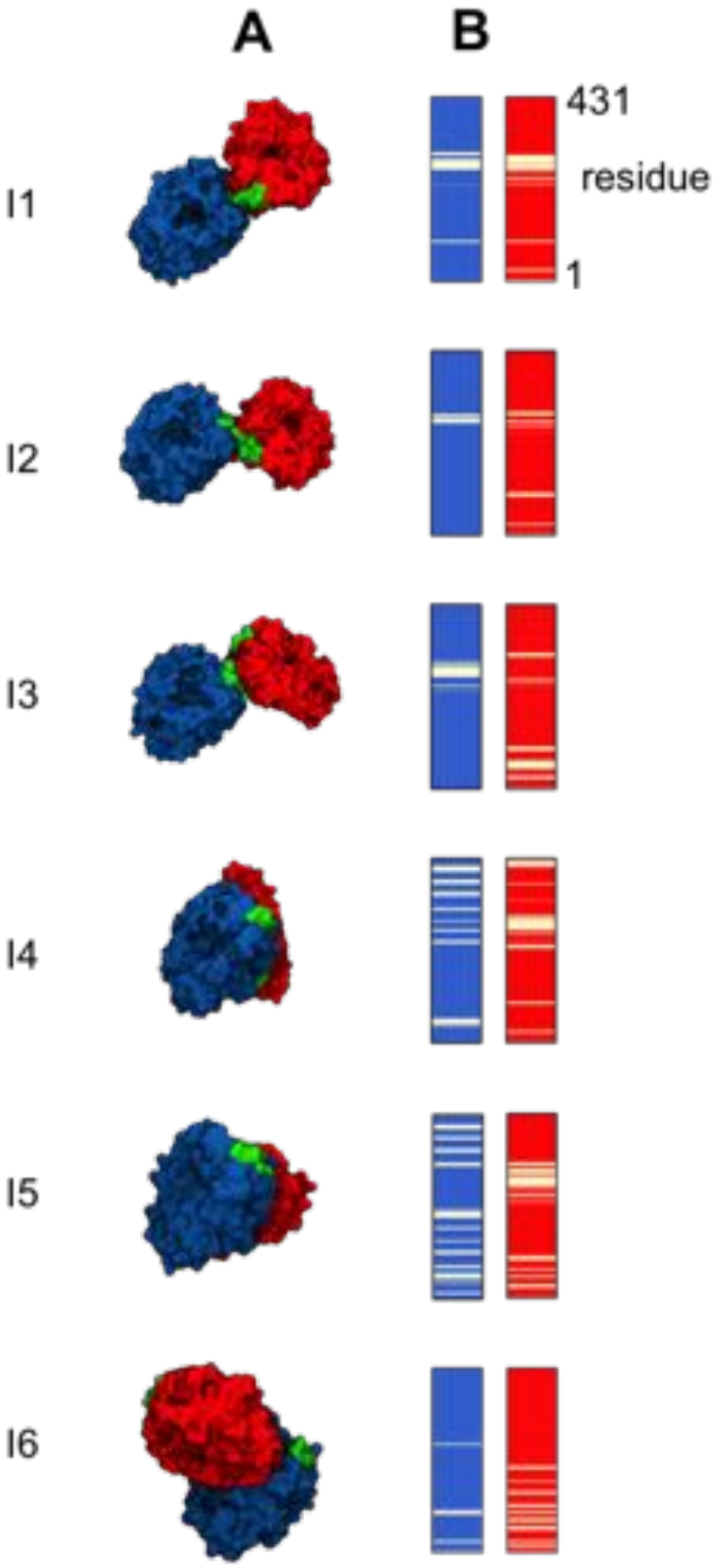


Fig. 3.7 Specific interactions for 50 simulations of Fab - Fab interactions. Each interaction is derived from a high density region within Figure 3.6. A) Snapshot of the two proteins at interaction IX, where green regions represent the area used to calculate D1 distance. B) Binary representation of each protein (Blue and Red), whereby yellow represents an interaction, and background colour represents no interaction.

Fc Proteins

When interaction frequency data is plotted for Fc proteins, a similar pattern to the one observed with Fab proteins emerges – there are distinct peaks, involved in around 60 of the 100 potential interactions. 4 peaks are found at roughly residues 1 – 10, 30 – 40, 60 – 7, and 90 – 100. Again, plotting these regions on the surface of the protein allows identification of specific surfaces on the protein that are involved in interaction, and in this case a single region is seen on one of the extended regions of the structure (Figure 3.4, C). With some, but reduced interactions on the opposing extended region. This is interesting, as the structure is biologically a homodimer, and thus should be perfectly symmetrical. It wouldn't be expected that there is any bias to one edge over the other. Looking closer at the structures however, can help to explain this result – the initial structure is not symmetrical, and by looking at the initial setup, and a few ns of simulation time, we can see there is an asymmetry within the structure, leading to one extended region having a larger range of motion during the simulation, and in some cases – fully collapsing into the structure, which is likely exacerbated by the absence of any carbohydrates from the center of the protein, which would be present in vivo. Another potential reason for this discrepancy is that the starting orientations of the proteins may influence the outcome of the simulations, and the random rotations of the starting structures with respect to each other may bias which half of the protein first encounters the other. Though this may be the case, we believe the asymmetry of the structure is the more likely reason for this, as the proteins are placed far enough apart that they are able to fully rotate before coming within a few Angstroms of each other. By comparing this "sticky" region with the results of the previously used SAP tool, we find a similar area again is highlighted using the tool and that the results from CG simulations match well (Figure 3.8, A and B). Its also worth noting that the region highlighted in both cases, is the region of the protein known to interact with the FcγR protein, and may interact due to the similar nature of the immunoglobulin fold that makes up the structures of the both the receptor, and

the Fc (Figure 3.8, C and D). It is worth noting, however, that many other proteins bind to immunoglobulins at the Fc, some such as Protein A, with a much higher affinity than the FcγR, and so this similarity may be entirely coincidental.

Clustering the results of the 50 simulations containing only Fc proteins shows that the same interaction "fingerprint" is present in almost all of the contacts. This indicates that the way in which the two proteins come together is similar or the same in most cases, and that there is a preferred "interaction mode", involving one of the extended regions of the protein, generally the more stable one (residues 1 – 100) in this structure (Figure 3.5, B). The clustering plot is different from the same plot for Fab simulations, in that many interaction surfaces are exactly the same, or extremely similar, whereas plots for Fab simulations show the same small number of residues interacting many times within very different larger interactions.

Looking at the interaction landscape for simulations containing 2 Fc proteins may inform more about the mechanism of interaction between the two proteins (Figure 3.9). Firstly, the electrostatic "trap" effect is much lessened when the population density is visualized, indicated by the absence of a large amount of "background" density in the plot below a certain distance between molecules, this absence indicates that the molecules do not spend a large amount of time close to each other exploring space like Fab proteins. This is potentially due to the shape of the molecule – however, a more likely explanation is that the "hot spot" of interaction is more attractive than the equivalent on the Fab protein, or that the rest of the protein surface is more unattractive for interaction, and the proteins don't come into close contact at all unless the "hot spots" are in proximity. This is shown in the landscape plot, where there is a distinct "line" of interactions where distance 1 (distance between previously shown interaction regions) is roughly similar, and distance 2 (distance between centres of mass of each protein) vary. This line suggests that in multiple simulations, the proteins are coming into contact with roughly the same interface (as shown in the clustering analy-

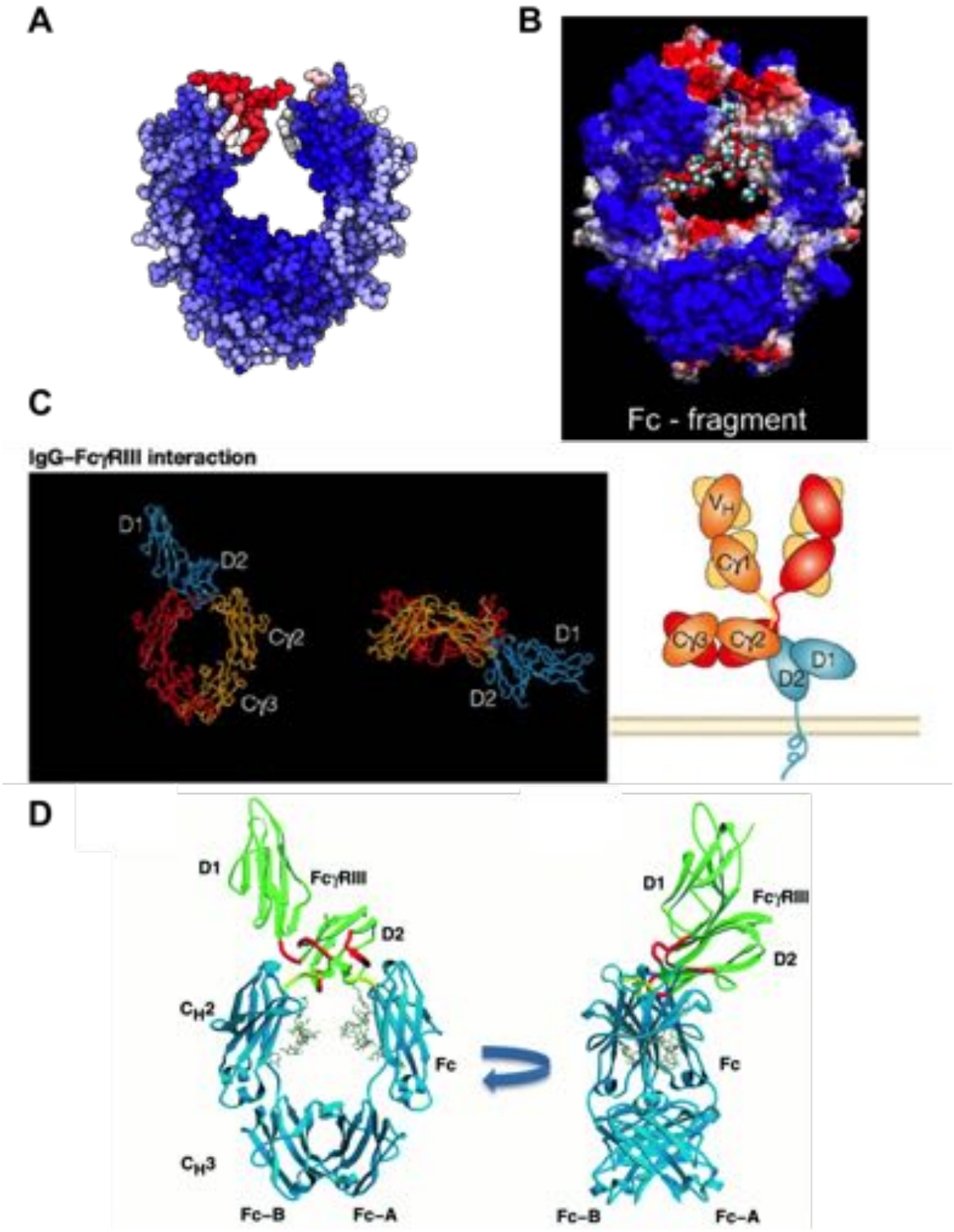


Fig. 3.8 A comparison of results from 50 simulations of Fc proteins with previous work on antibody-protein associations. Showing A) A single view of the interaction regions found to self associate in our studies. B) Results from an SAP analysis from Chennamsetty et al. [30] showing the same regions highlighted in red as being a major cause of self-association. C) Crystal structure, and schematic view of the IgG Fc in complex with the Fcγ R protein. Note that the same regions as identified in our studies, and with the SAP tool, are involved in interactions. From Woof and Burton [201]. D) Crystal structure of a glycosylated IgG Fc fragment in complex with an Fcγ R receptor, also showing the same region in interactions as our studies, and the SAP tool. Figure adapted from Radaev et al. [147].

sis), and conformed by the distance between "interaction surfaces" being the same, but with different protein conformations or orientations - possibly a “twisting” of the proteins with respect to each other (Figure 3.10), this is inferred from the difference in centre of mass value for each protein, when they are in contact using the same interface.

Comparison of Fab and Fc proteins

The two different protein analyses show two different modes of interaction. Whilst both interaction with a discrete “hot spot” on the surface over a large number of simulations, and both of these regions agree with previous work as being nucleating points for aggregation, the mechanism by which they interact is different in each case.

For Fab proteins, evidence suggests that interactions involve the two proteins exploring the surface of each other, often at length, before interacting around a small patch on the surface, with very different conformations of the cluster in each case. Fc proteins, on the other hand, appear to have a much more conserved interaction mode, as shown by the interaction clustering, and the interaction landscapes. This is possibly due to shape effects of the protein – the Fab protein is a very stable ovoid shape, whereas the Fc has the potential to be very elongated, and is much more dynamic throughout a simulation, the structure fluctuating by a large amount, as the extended regions move together and apart, occasionally collapsing. Different shapes of the molecules will lead to different behaviours when they come into contact with each other. The chemical nature of the two regions of interest are also shown to be quite different, perhaps accounting for this difference in behaviour. Looking at the interaction surfaces on the coarse grained protein structures (Figure 3.11), we can see that the region on the Fc protein that is involved in self-self interactions contains a patch of hydrophobic residues surrounded by a ring of charged amino acids (A). This region is known as an "O-ring", and is known to be involved in highly specific interactions (Bogan and Thorn [14]). The similar analysis for the Fab protein shows that there are also a large number of

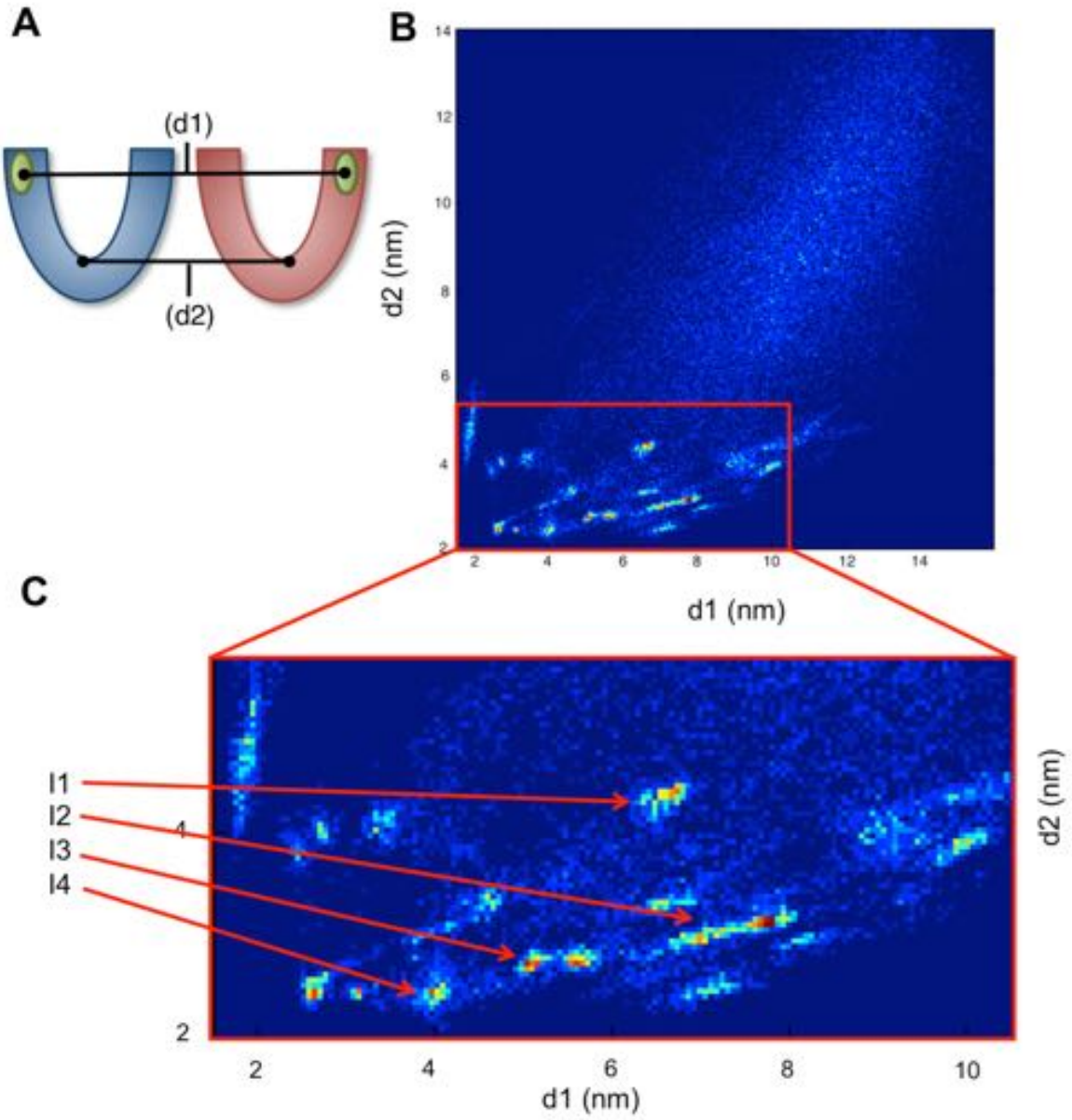


Fig. 3.9 Interaction landscape plot for 50 simulations of Fc - Fc interactions. A) Key showing two distances used to generate the plot: d1) Distance between the center of mass of each protein d2) Distance between the extended region of the protein found to interact through frequency analysis (Figure 3.4). B) Interaction landscape for the two distances. Whereby blue represents a low frequency, and red "hot spots" indicate regions of space where the proteins spend a large amount of time. C) Zoomed in region of the plot where the protein come within a close distance to each other, showing highlighted regions of interest (I1 to I4).

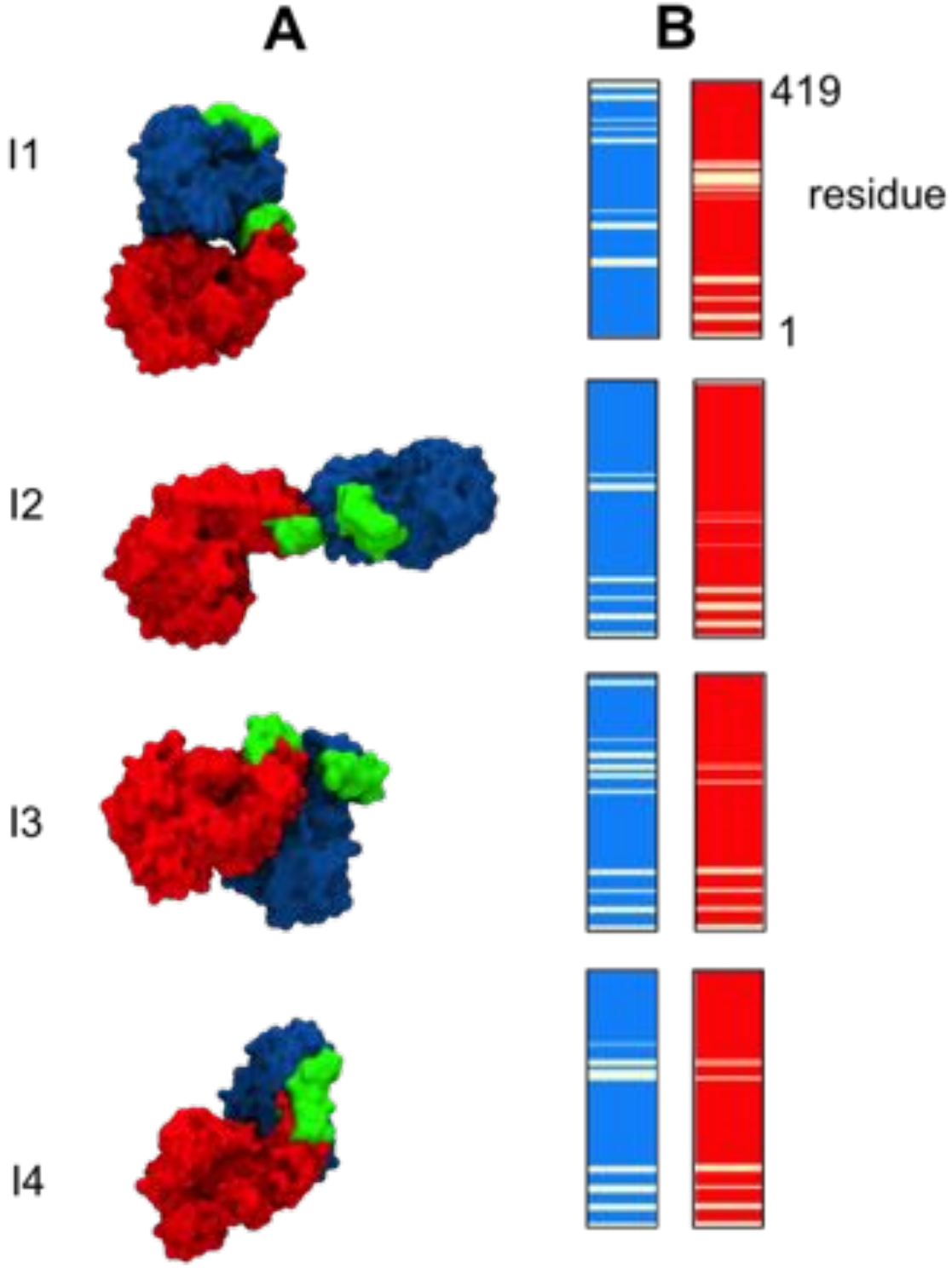


Fig. 3.10 Specific interactions for 50 simulations of Fc - Fc interactions. Each interaction is derived from a "hot-spot" within Figure 3.9. A) Snapshot of the two proteins at interaction IX, where green regions represent the area used to calculate d1 in figure 3.9. B) Binary representation of each protein (Blue and Red), whereby yellow represents an interaction, and background colour represents no interaction.

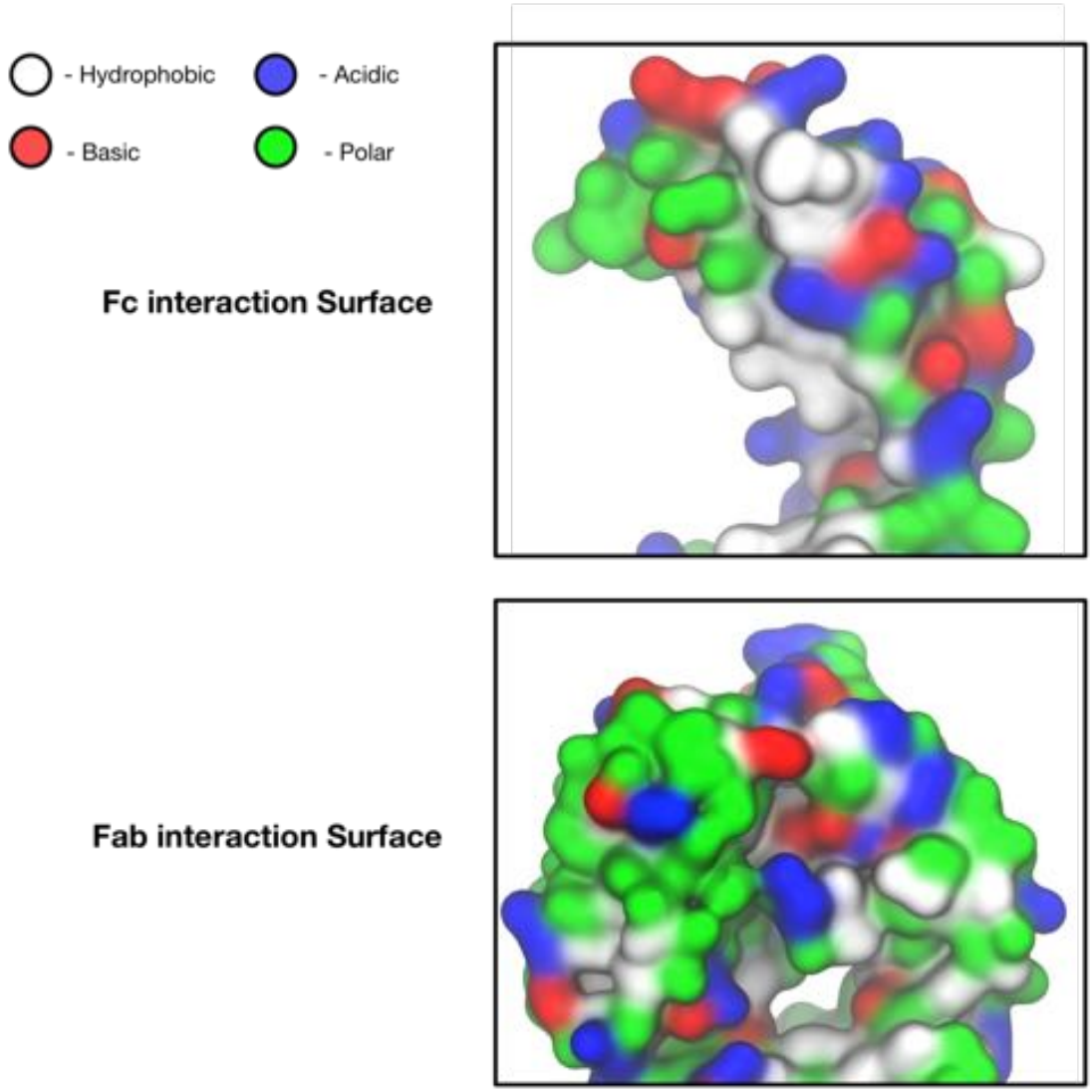


Fig. 3.11 Protein surfaces coloured according to amino acid composition, showing hydrophobic residues (white), polar (green), acidic (blue), and basic (red). The Fc protein, A) shows a hydrophobic patch surrounded by a ring of charged residues, a surface shown previously to be involved in protein-protein interactions (Bogan and Thorn [14]). The Fab protein, B) shows a large number of charged residues in the vicinity of the interaction region.

charged residues around the interactive region (B), but there is no large hydrophobic patch, like with the Fc protein.

This chemical difference in interaction regions however, could account for the differences in interaction behaviours, with the specific O-ring like structure resulting in the highly conserved interactions observed for Fc self-association, and Fab proteins showing only a few conserved residues.

3.3.3 Simulations of Larger Scale Systems

When simulations of larger numbers of proteins are performed, the ways in which the proteins interact is different both from each other, and from systems of the same protein with only 2 molecules. Simulations were performed as indicated in Table 3.2, taking care to place proteins far enough apart that nonbonded interactions are not present between molecules before the simulation is started. Larger scale aggregates formed by Fab proteins differ from Fc proteins in the locations at which interfaces form between the molecules. Looking at the final results of the simulations, the types of aggregates appear different. With the interaction hot-spots from smaller scale systems shown (Figure 3.4), the Fc proteins appear to cluster around a series of interacting spots, whereas the Fab proteins appear to show no real bias towards one region in interactions, and appear more linear in nature (Figure 3.13)

By looking at interaction frequencies for each residue, comparing between different simulation scales, we can see that the Fc proteins retain the same mode of interaction in each case – with the same regions showing a similar proportion of interaction frequency, when the data is normalized for how many potential interactions in a system (Figure 3.12, B). The Fab proteins show no such pattern, however, with each system larger than two proteins showing no particular hot-spot for interaction (Figure 3.12). This difference is also evident from plotting the interaction frequency on the surfaces of the proteins, and comparing different scales of interaction . Clearly Fc proteins interact using the same region – whereas the Fab

proteins appear to show no particular region or hot-spot for interaction. Larger numbers of less specific interactions appear to be favoured over smaller numbers of more specific interactions. Looking at the aggregates formed by the larger systems, they appear more linear in nature, with the proteins arranged end to end, rather than clustering around certain interaction sites, as the Fc proteins do.

By plotting spatial frequency landscapes for all levels of simulations together, we are able to compare interactions between them. This is shown for Fab proteins in Figure 3.14, and Fc proteins in Figure 3.15. Regions are highlighted that show interactions that occur primarily in different size simulations, with different colours indicating different sizes of simulations (yellow for two proteins, orange for four proteins, and red for eight proteins). There are regions of high frequency in both plots that show interactions that are common between the different granularities. There are also regions of high interactivity that represent proteins held in a similar relationship due to being part of a larger aggregate, but not due to being directly in contact with each other.

3.3.4 Discussion

In this chapter, we have demonstrated that coarse grained simulations are able to provide useful data on the interactions and aggregation of monoclonal antibody proteins. Data resulting from the study can be used to inform mutational studies on antibody aggregation, and we are able to learn about the mechanisms by which two proteins approach one another for non-specific interactions. Specific regions on the protein are identified as being involved in interaction, though their validity is determined through simulations of increasing numbers of proteins. In the case of Fab-Fab interactions, results from simulations of just 2 proteins are not necessarily useful for targeting protein mutational studies, as the specific residues no longer show as strong an impact on aggregate formation in larger system. Conversely, the region on the Fc protein is conserved throughout all simulations, and appears to contain

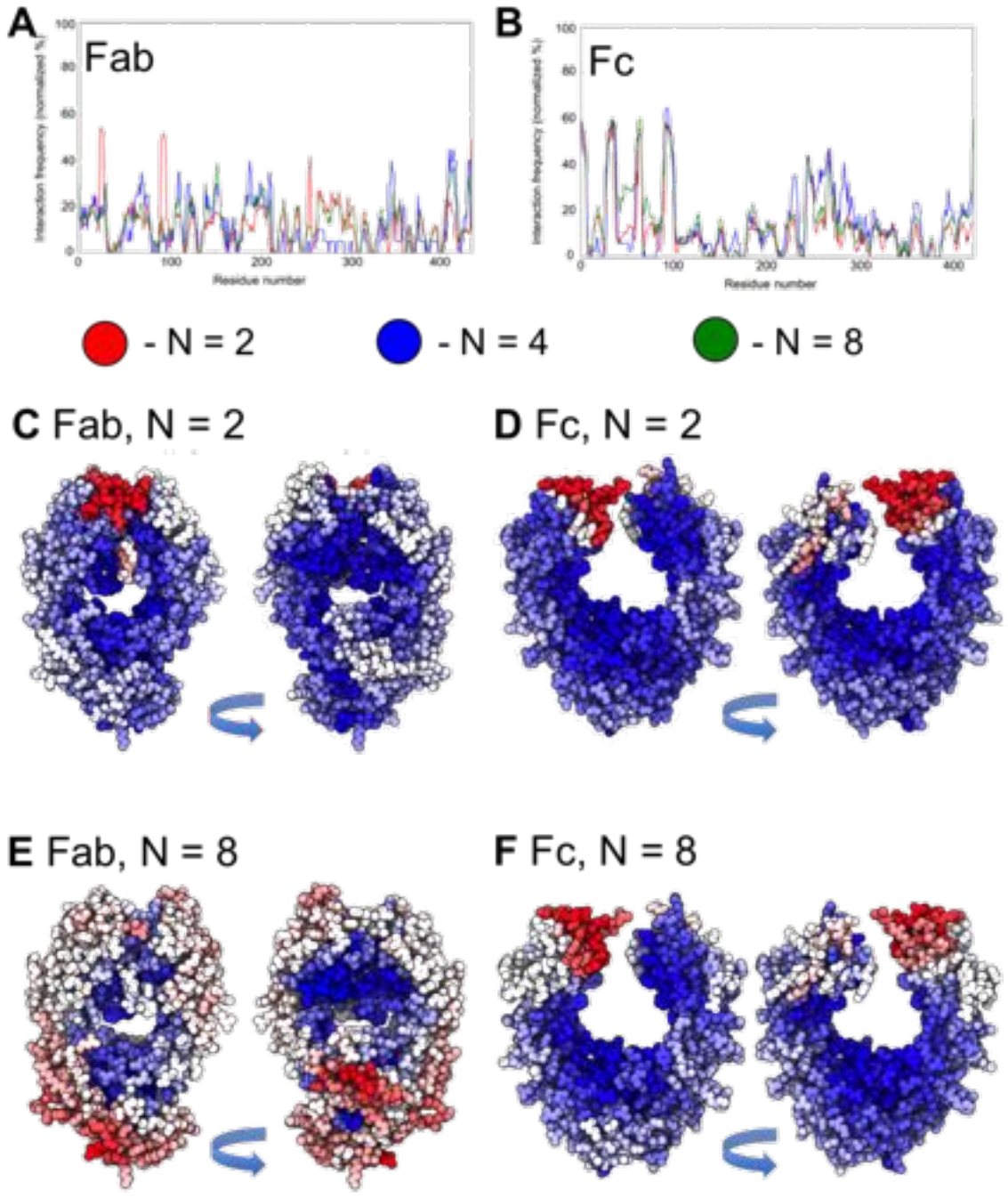


Fig. 3.12 Frequency of interaction per residue for Fab and Fc proteins over a range of simulations. A) Fab protein interactions for systems containing 2 (red), 4 (blue), and 8 (green) proteins. B) Fc protein interactions for systems containing 2 (red), 4 (blue), and 8 (green) proteins. C) Surface plot of Fab proteins when 2 proteins are present, showing regions of high interaction by colour. Regions highlighted in red are inolved in high numbers of interactions (and correlate to the peaks on the graph). D) Surface plot of Fc proteins when 2 proteins are present showing regions of high interaction by colour. Regions highlighted in red are involved in high numbers of interactions (and correlate to the peaks on the graph) E) Surface plot of systems where 8 Fab proteins are present. F) Surface plot of systems where 8 Fc proteins are present.

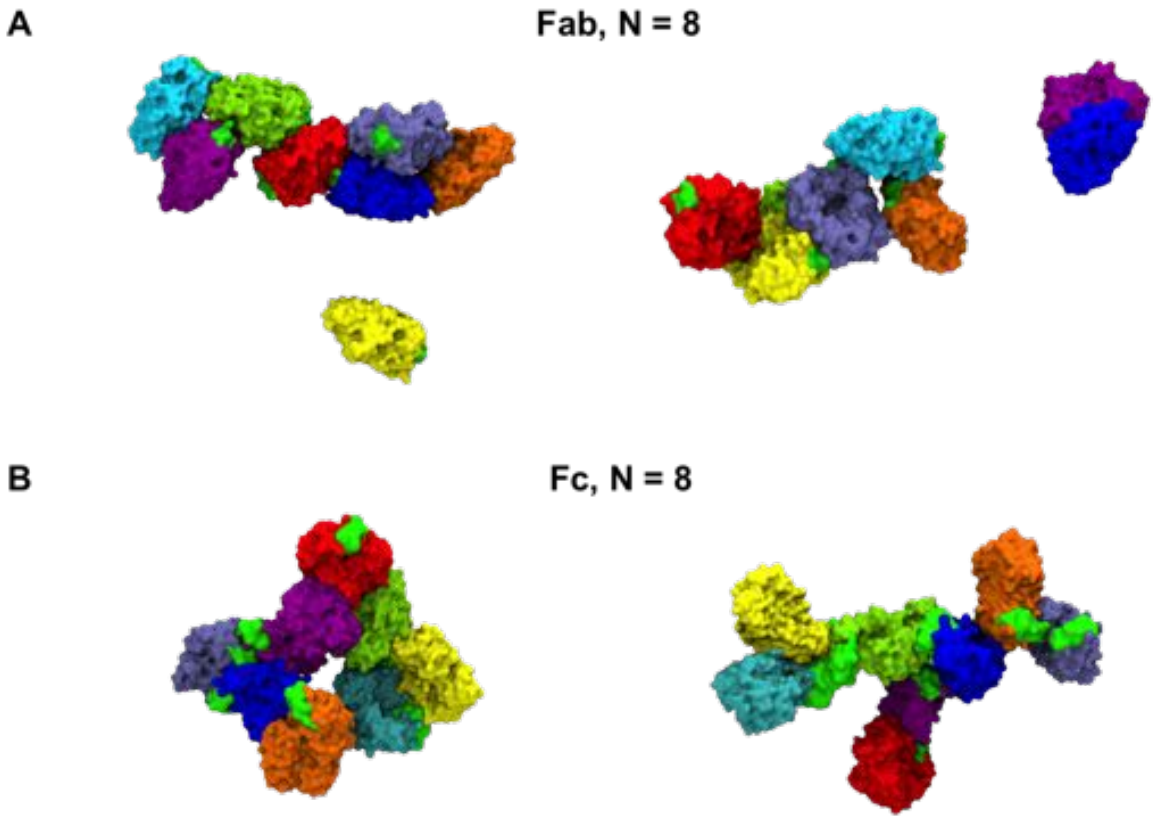


Fig. 3.13 Aggregates from simulations containing 8 proteins, showing A) Fab protein aggregates, which appear somewhat linear in nature. B) Fc protein aggregates, which appear to cluster around similar interaction sites.

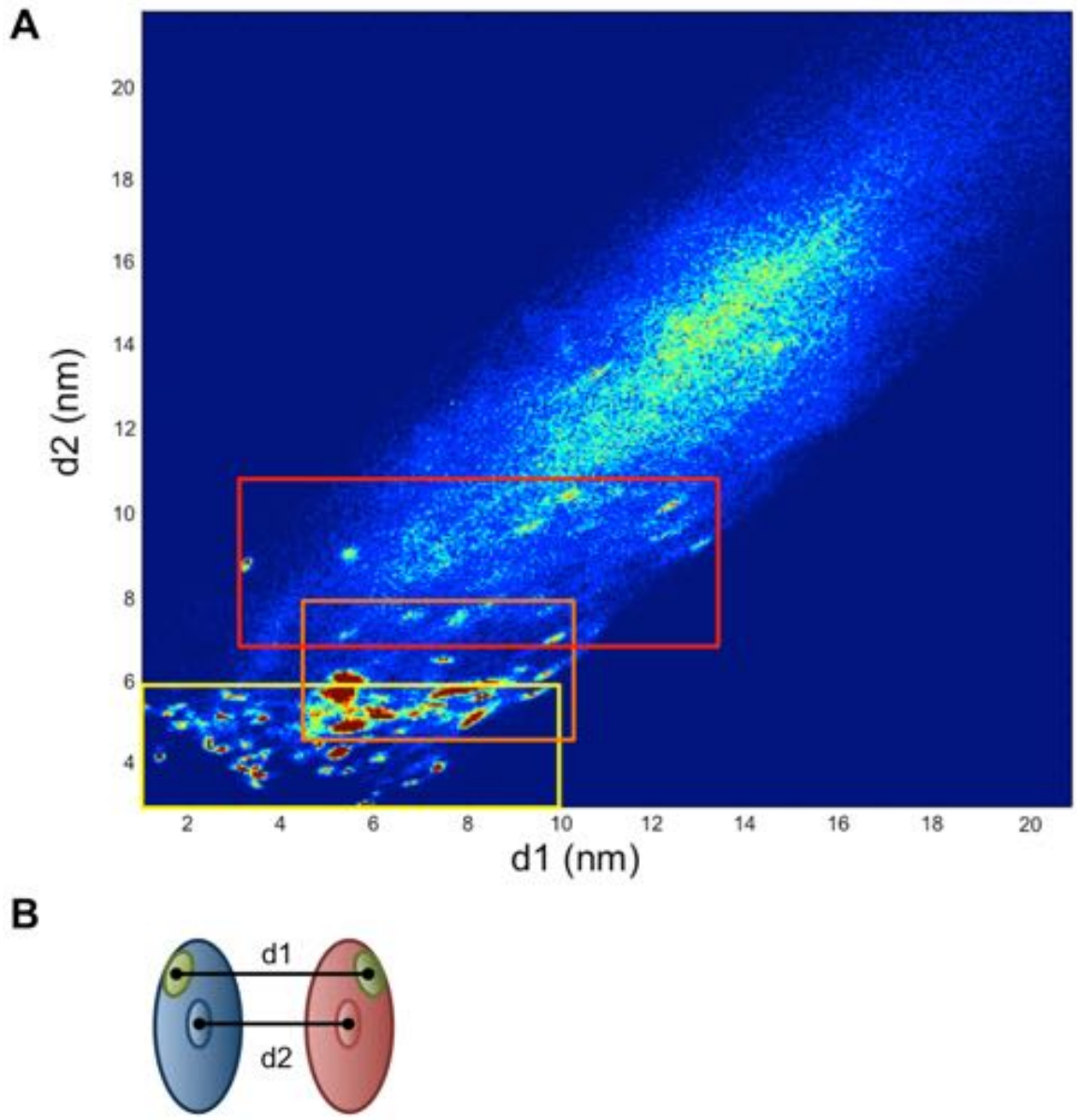


Fig. 3.14 Interaction landscape plots for all pairs of protein in simulations containing Fab proteins. Highlighted are regions mostly occupied by proteins in complex with a single other protein (Yellow), 3 other proteins (Orange), and 7 other proteins (Red).

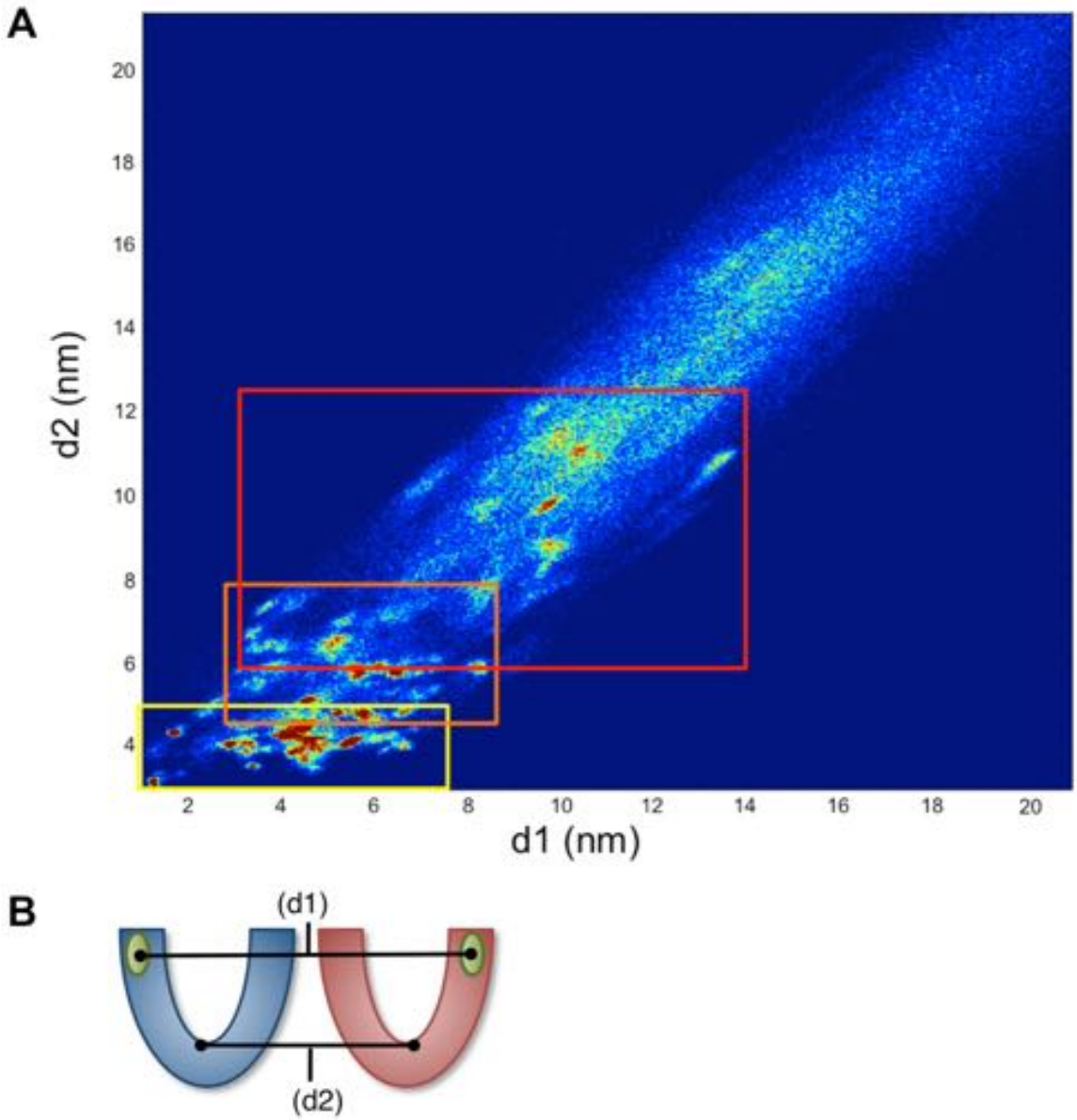


Fig. 3.15 Interaction landscape plots for all pairs of proteins in simulations containing Fc proteins. Highlighted are regions mostly occupied by proteins in complex with a single other protein (Yellow), 3 other proteins (Orange), and 7 other proteins (Red).

a hydrophobic region that is surrounded by charged residues, a makeup known to promote protein-protein association (Bogan and Thorn [14]). Although the region looks like it may be occluded by the presence of the entire immunoglobulin molecule, the flexible nature of the protein loops involved in linking Fab and Fc would allow access to this region of the protein for interaction in a full molecule. The results from simulations of small systems indicate regions of the molecule that are very similar to those suggested by the SAP tool, however, these regions are not seen once more proteins are introduced into the system, indicating that a larger number of weaker and less specific interactions are favoured over a small number of specific interactions in this case. In the case of Fc proteins, the region of interest in the extended region of the structure has enough surface area to allow interaction with multiple other proteins at once, which may explain why the interaction is retained over multiple simulations.

By looking at the interaction landscape of all systems together, we can potentially identify common pathways molecules take to interaction. Although no such pathway is obvious from our results, it is possible that more data is necessary in order to extract these pathways clearly. We can, however, infer information about the way in the which the two molecules approach and interact with each other. From the data, there is an increase in frequency within (7 nm distance) between two Fab molecules, indicating that the proteins get trapped together, before the two proteins sample the opposing surface from a low distance (indicated by high frequency regions on the plot), before coming together into one of a cluster of preferred interactions.

The FC fragment was also observed to be highly flexible in silico, with the two extended regions of the protein sampling a great deal of conformational space, potentially allowing a greater ease of interaction, especially with regions on the end of the extended regions. The reason for this flexibility is doubtlessly due to the lack of the well-characterized carbohydrate groups on the inside of the structure, which are known to increase stability (Arnold

et al. [5]).

The advantages of using CG-MD over traditional techniques are clearly highlighted in this study – as systems have to be simulated for much longer time periods than are accessible to atomistic simulations in order to gain insights into protein-protein interaction. Systems are also of a much larger size than can be usefully investigated using explicit models. Another advantage is that large numbers of smaller systems can be run relatively efficiently, allowing the potential for high-throughput screening of compounds for the investigation of specific structures, as shown by the generation of “hot-spots” on the proteins after a large number of small simulations.

It is clear that simulating small and large numbers of proteins are complimentary techniques, in that the behaviour of proteins in one methodology do not necessarily represent the behaviour of the same protein in the other. Therefore, in order to correctly identify interaction "hot spots" for some molecules, simulations of larger arrays of proteins are necessary. Though this is not the case for Fc proteins in this case. It may be possible, however, that many simulations of only two proteins is enough to gain accurate information about the relative aggregation propensity of proteins.

There are, however, disadvantages to using CG-MD in this context, the coarse-grained methodology is unable to account for protein unfolding effects, a major contributor to aggregation of most proteins, particularly once an initial complex has been formed. This is something that can potentially be studied in the future by simulating unfolding events in atomistic detail, before looking at larger scale interactions in coarse grained simulations. Another limitation is the well documented “stickiness” associated with coarse grained simulations in solution, although great progress has been made with the release of Martini v 2.2 (Baaden and Marrink [8]), the forcefield is still not entirely accurate when reproducing experimental association values for non-membrane proteins.

Our data provides more detail on the specifics of antibody aggregation than previous

studies (Chaudhri et al. [27]), in that our models are detailed enough to find specific regions of interaction, as well as emergent, large-scale properties and includes structural data on interaction mechanisms, and the way in which the proteins approach one another, an advantage over docking and surface analysis tools (Chennamsetty et al. [30]). Though the models we use are not complete (the antibody is reduced into domains), they provide enough detail for the purposes of studying non-specific interaction, especially as the protein has been shown to have different behaviour for each protein fragment experimentally.

It would be of huge interest to investigate the specific molecules used in this study experimentally, to attempt to corroborate on the findings from simulations, and mutate the regions of interest from Figure 3.12. Another source of potential validation for interaction surfaces is through analysis of crystal contacts. There are many crystal structures of both Fab and Fc proteins, and analysis of where these proteins are in contact within crystals may provide further validation that the "hot spots" identified in silico are genuine. Additionally, using a high-throughput approach, as a potential extension and more accurate alternative to docking studies, as was successful with sidekick (Hall et al. [68]) would be of interest. Sidekick was successfully used to automate and streamline the setup, simulation, and statistical analysis of large numbers of helix/helix interactions within Coarse Grained membranes. It would also be interesting to apply better statistical methods to study the mechanisms of interaction, looking for pathways of interaction using techniques such as Markov State Modelling.

Overall the CG methodology has yielded a greater understanding of monoclonal antibody aggregation, both with specific interaction sites being identified, and with an insightful window into initial aggregate formation. It would also be of further interest to also look at applying this methodology to other proteins, for example, more extensively studied amyloid proteins (Sø rensen et al. [173]), or a larger range of antibody fragments, with known differing interactions.

Chapter 4

Novel Methodologies for the Study of Antibody Aggregation

"I do not fear computers, I fear the lack of them."

Issac Asimov

In this chapter I extend the work initially done on protein-protein interactions, to much larger data sets. From work done in Chapter 3, it is evident that large ensembles of simulations are required for the generation of enough data to gain a statistical insight into antibody aggregation, and that more advanced statistical techniques are required in order to correctly compare aggregation propensities of protein structures. Here I present a method for the automation of setup, analysis, and ranking of aggregation propensities for proteins that is modular, easy to use, and highly customizable.

4.1 Introduction

4.1.1 High Throughput Simulations

High throughput simulations are characterized by the running of large numbers of replicates, of the same or similar systems, in order to gain a statistical picture on the behaviour of the objects within. Simulations that are run are typically short, or system sizes are small enough, that each replicate requires a small number of CPU hours to run to completion. This approach has been popularized with the release of various platforms such as the publicly assisted Folding@home (Shirts and Pande [170]), and the molecular docking tool Screensaver (Richards [150]). In recent years, however, high throughput methodologies have turned to more specific investigations, such as the helix helix dimerization (Hall et al. [69]), and protein docking (Salwinski and Eisenberg [159]), whereby sets of simulations can be run to address an extremely specific biological or chemical question. These wrapper scripts massively reduce the amount of user input, and setup time required for the simulation of large numbers of systems, and allow the easy acquisition of statistical data on the bulk behaviour of systems.

4.1.2 Markov State Models

Markov state models are a relatively recent addition to the array of tools available for the analysis of molecular simulations (Pande et al. [142]). Markov state models employ the theory first suggested by Andrei Markov in the 19th century, that any system can be described as a series of states, with transition probabilities between them, depending on the Markov property (Sarich et al. [161]). In the case of biology and biochemistry, this has generally been utilized for the explanation of kinetic phenomena, such as molecule folding (Huang et al. [81]), drug-receptor interactions (González-Díaz et al. [64]), and protein conformational change (Noé and Fischer [135]). A simplified schematic example of a two

state Markov model can be seen in Figure 4.1, A. Two states are defined, and the probably of moving along a certain pathway (arrow) is shown, where a probability of 1 represents a 100% chance of change within the specified time window. A biological example of a markov state model application can be seen in Figure 4.1, B, where a Markov state model is applied to the folding of NPL9 protein (Pande et al. [142]), where arrow size indicates probability, with larger arrows indicating more popular pathways between the unfolded and folded states. More details of Markov models are discussed in Chapter 2.

4.1.3 Chapter Aims

In this Chapter I present an efficient, autonomous, and modular pipeline devised for the setup, simulation, and analysis of large numbers of molecular dynamics trajectories (Figure 4.2). The primary focus of the high-throughput pipeline is to generate data for the exploration of the aggregation propensity of monoclonal antibody fragments. In order to accurately assess the simulation data, markov state models were applied to the large amounts of data in order to attempt to gain information on the kinetics and mechanics of non specific protein-protein interaction.

The pipeline was assessed by running simulations on the structures of 6 different antibody Fab structures provided by Lonza Biologics (Table 4.1). The structures have a variety of heavy and light chain types, meaning they do not all have a similar sequence aside from CDR regions. It will be expected that aggregation could occur due to any region on the structure, and so, they are a good model for testing the high throughput methodology.

4.1.4 The Lonza Protein Dataset

The Lonza set of proteins is part of a dataset of structures owned by Lonza biologics, of which there are known experimental values for aggregation propensity, stability, and expression levels. Summaries of these values for 6 Lonza dataset structures can be seen in table

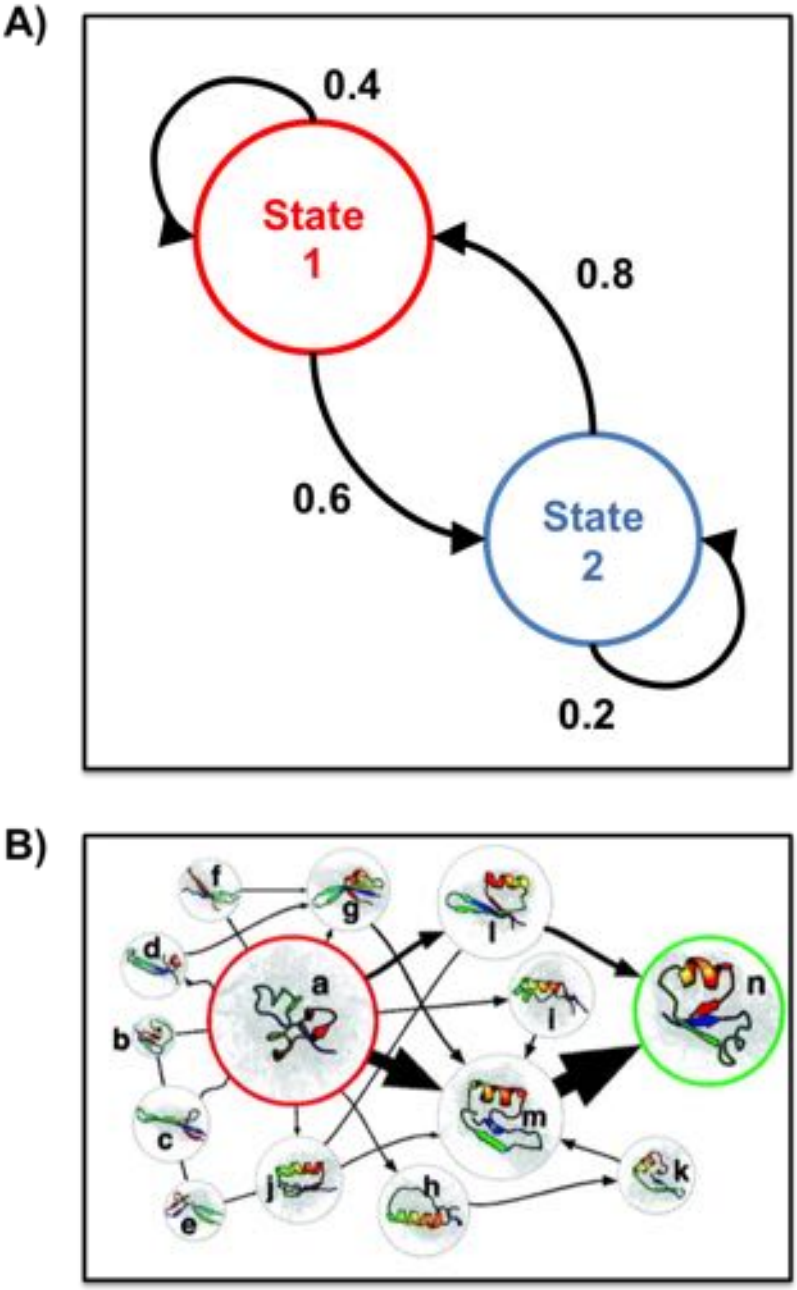


Fig. 4.1 Markov State Model (MSM) examples, showing A) A simple two state MSM, where state 1, and state 2 have different probabilities of switching between themselves, or staying the same. Probabilites are represented as values between 1 and 0, where 1 is a 100% chance for the state to change in a given window, and 0 is a 0% chance. B) An example of a MSM applied to a biological system. Folding of the NPL9 protein is represented with the unfolded protein on the left, and folded on the right. The size of each arrow represents probability, and thus more common pathways between states are bolder. Figure adapted from Pande et al. [142].

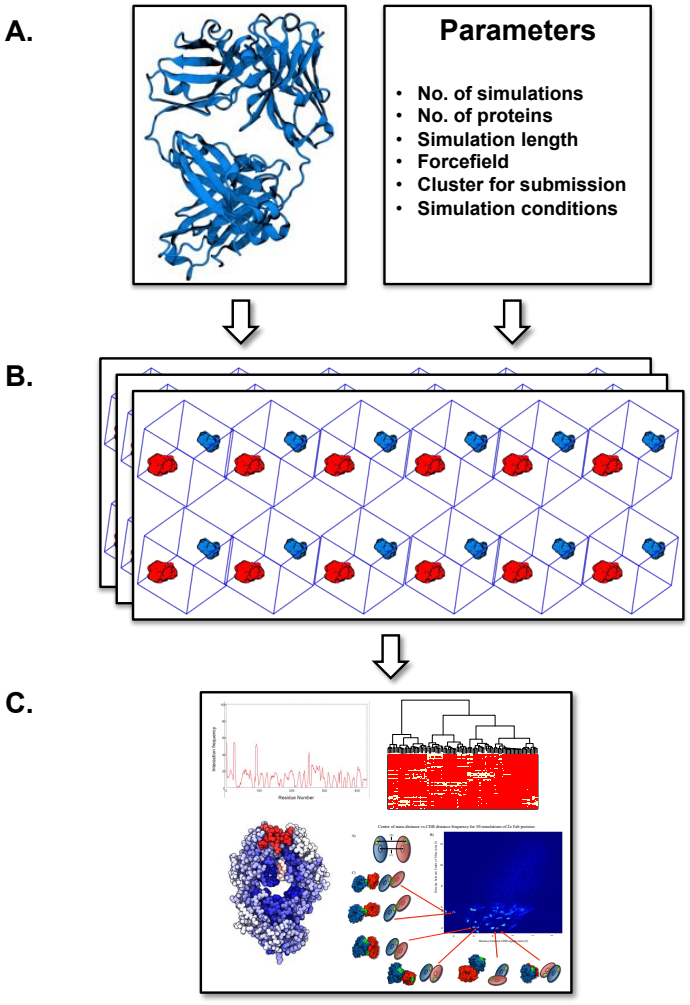


Fig. 4.2 Schematic of the pipeline for a high throughput system for investigating protein aggregation, showing A) Inputing a single structure, and specification of system parameters, B) Setup and simulation of a large number of systems, C) Anaylsis of group properties resulting from multiple simulations.

Table 4.1 Table comparing experimental characteristics of Lonza proteins

Lonza dataset Name	Experimental Titre (μg/ml)	Percentage Monomer Concentration (%)
Kappa VL chain		
H31K5	77.54 (High)	99.88 (Low)
H8K8	29.72 (Medium)	94.34 (Medium)
H9K7	6.35 (low)	60.1 (High)
Lambda VL chain		
H28L2	80.31 (High)	100 (Low)
H33L2	18.08 (Medium)	96.64 (Medium)
H17L1	8.44 (Low)	83.55 (High)

4.1. For validation of the high throughput methodology, I was given access to 6 of these structures, and their experimental data, to look for correlations between computational and experimental approaches. Experimental data provided was:

- Structure - A homology model of each structure is provided

- Experimental Titre - Measure in μg/ml, this is a measure of how much antibody that recognizes a chosen epitope is present in solution. This is measured using Enzyme Linked ImmunoSorbent Assays (ELISA)

- Percentage Monomer Concentration - A measure of how much of the purified solution of antibodies is monomer, and typically, non-aggregated. This is measured by passing solution through a 0.22 μm pipette tip. Typically, everything that passes through this sized pipette tip, is non-aggregated, and everything that does not pass through is aggregated protein.

4.2 Methods

4.2.1 High Throughput Simulations

In order to correctly automate the number of jobs required for the package, a modular strategy was employed. A series of different python scripts and self-written python modules and libraries are employed. A master script deals with general workflow (e.g energy minimization, solvatation, submission to clusters etc.), and calls on a series of libraries, containing commands with their resultant options. Files such as simulation parameter files (.mdp files) are stored in a toolbox, along with analysis scripts, and cluster submission scripts, allowing different submission, setup, and simulation methodologies to be swapped and changed easily. Parts of the script can be substantially adjusted for a specific function, without compromising the general usage. Simulations are designed to be robust with the addition of a restart mechanism included with job submission, and job submission can be controlled to take into account usage quotas on different machines.

The workflow is not truly fully automated, as currently the user is required to copy data from completed simulations to scratch space for analysis, before initiating the next step in the pipeline. This is to ensure complete data is analysed, as there are cases where crashed simulations will mean that the full number of simulations requested will not be analysed, and analysis time is wasted.

So example code from the system can be found in the appendix.

4.2.2 Simulation Setup

Simulation setup is performed with the use of the GROMACS package (Pronk et al. [146]). Tools such as trjconv, editconf, grompp, and mdrun are extensively utilized through python scripts in order to set up simulation systems. The scripts take a series of user inputs to generate systems for simulations. Parameters able to be modified by the user are:

- Protein structure of interest - a .pdb file is required as input containing the structure of interest.
- Forcefield of interest - Various option for forcefield are available, including Bond and Martini iterations, elastic and ElNeDyn network restraint models, additionally, if a customized parameter set is required (such as one containing unique structures such as carbohydrates), a topology file can be given as input instead.
- Number of proteins to be simulated - The setup process will take into account the number of proteins and will adjust box sizes and spacing to account for periodicity of the system.
- System parameters - Many system parameters can be adjusted, including ion concentration, temperature, pressure, and various coupling schemes.
- Simulation settings - The length and number of simulations can be customized.
- Cluster submission - The compute cluster to submit jobs to can be controlled, and the scripts contain a library of submission methods for each cluster with differing versions of GROMACS, the protocol can also be flagged to control how many cores to run each job on, and the number of simulations to submit at once, with a holding mechanism used to prevent more than a chosen number of jobs being submitted. This prevents the breaching of quotas on shared clusters, but without the downside of having to manually control submission.

The default behaviour is to take a structure, set up a system containing a pair of proteins, randomly oriented with respect to each other, then perform a short (500 ps) energy minimization, solvate the system, and add ions to the system until a system concentration of 0.15 M is reached, before equilibrating the system with position restraints on the proteins. In order to ensure that results are not biased by the starting structure of the proteins (even

though they are unable to interact with each other at the start), 4 different starting conformations can be generated for each setup, with proteins rotated randomly in the X, Y, and Z directions. These systems can then be independently equilibrated, then setup conformations are cycled through for each submission, ensuring that not all systems start from the exact same conformation.

Simulations typically contain two proteins, and are run for 500 ns, as previous work (Chapter 3) has determined that this is a reasonable amount of time for a pair of proteins to interact. The protocol for two proteins is as follows: proteins are placed in an 18 x 18 x 18 Å box, with sufficient distance between them that they cannot interact, including through periodic boundary conditions. The system is energy minimized for 1000 steps using the steepest descents algorithm, before the box is solvated, and counterions added to generate a salt concentration of 0.15 M (physiological conditions). The system is then energy minimized again for a further 1000 steps to reduce steric clashes in the water molecules, before an equilibration simulation is performed for 50 ns with the protein backbone restrained, in order to relax the water in the system. The system is then run for 500 ns unrestrained, before analysis is performed.

In order to keep the system robust, the system is designed to restart failed jobs the first time they crash, by logging the wallclock applied to the cluster in question at the start of each simulation. When the wallclock time has been reached, the system checks for the presence of files indicating completion of the simulation, and if the job is not complete, it is restarted. A flowchart representation of this can be seen in Figure 4.3.

4.2.3 Simulation Analysis

Simulation analysis is performed by the running of a separate script on completed setup folders, the script is able to analyse any number of trajectories, with any number of proteins present. Tools used for analysis are those available through GROMACS - such as g_dist

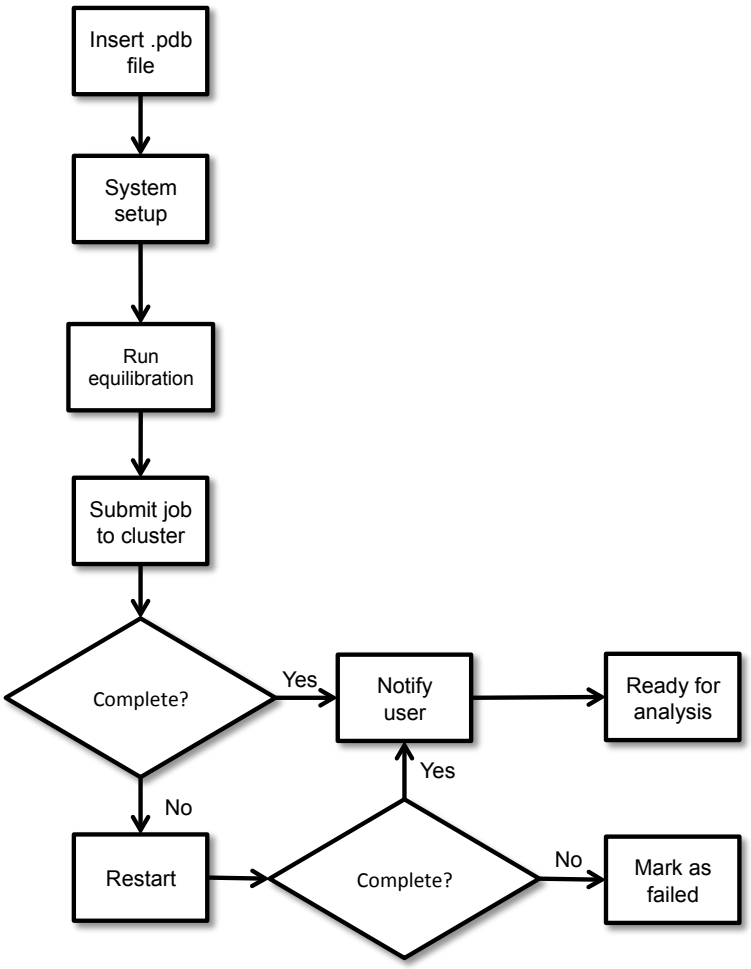


Fig. 4.3 Flowchart for the setup and simulation of multiple repeats of a single system, showing the decision making process. Note that the first time a job fails, it is restarted, the second time, it is logged as a failure.

and g_rmsf. Python, numpy, and MDAnalysis (Michaud-Agrawal et al. [128]) are also used. Currently the default behaviour for the analysis of a single simulation is as follows:

- Generate reduced size trajectories to ease analysis by the removal of water, and skipping of frames.
- Analyse the Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) properties of each individual protein within the system to check stability.
- Perform distance and minimum distance analysis between all pairs of proteins within

a system.

- Calculate contact mappings of interaction surfaces between each pair of proteins in a system.
- Extract information on the orientation of each protein with respect to every other protein in the system over the course of the simulation.
- Perform a clustering analysis on all results to extract the most common interactions.
- Combine residue interaction counts for all simulations, resulting in interaction frequency graphs.
- Generate spatial interaction landscapes for all systems.
- Generate default Markov State Models (though some user input is generally required to get useful models).

This analysis is performed on each individual simulation trajectory found. Due to the directory structure set up for analysis, we can group the simulations into different categories of analysis depending on their relationship to each other:

- Local Analysis - Data for a single simulation, e.g. RMSD values, interaction residues for a single simulation.
- Regional Analysis - Data for a set of simulations of the same proteins, e.g. an interaction frequency plot for all simulations of Lonza dataset protein H9K7.
- Global Analysis - Analysis performed on all protein-protein interaction data.

Analysis is designed to be as simple to run as possible, and generates a "Regional Analysis" folder for each set of simulations, containing information on the frequency of interaction

for every residue over all simulations, the top cluster from an analysis combining all trajectories, and interaction landscapes. It also attempts to set up markov state models for each protein found.

A global analysis can be specific within the analysis portion of the script, whereby all simulation data, regardless of protein can be combined and analysed together. This can be useful, for example, for looking at general behaviour of two molecules interacting.

The analysis structure in schematic form can be seen in Figure 4.4.

4.2.4 Markov State Modelling

Markov state models are constructed with the program EMMA (Senne et al. [166]). Markov states are determined by the extraction of data on the spatial relationship between each pair of proteins within each simulation. For each protein in each simulation, the rotation matrix of the protein with respect to each other protein is extracted. The rotation matrix calculated is based upon the the Cartesian coordinates of the box in which the molecule is placed, and so, systems with different starting conformations, or different target conformations (i.e the structure on which the rotation is calculated against) will generate non-compatible results. In order to homogenise rotation matrix results, the Cartesian rotation matrix values were converted to Euler angles. Euler angles represent the rotation of a molecule with respect to a starting conformation, and so have no relation to box coordinates. Euler angles can be thought of as being similar to the values of Yaw, Roll, and Pitch that are used to describe the motion of things such as aircraft, Figure 4.5, A. In this case, Euler angles are vectors describing the motion of the protein if it were to be rotated to match the exact location of it's target (the other protein in the simulation), each value is directly comparable to it's equivalent value from other simulations of the same protein.

Combining Euler angle with a distance metric (the distance between the centre of mass of each protein), allows the exact position and rotation of a protein to be represented using

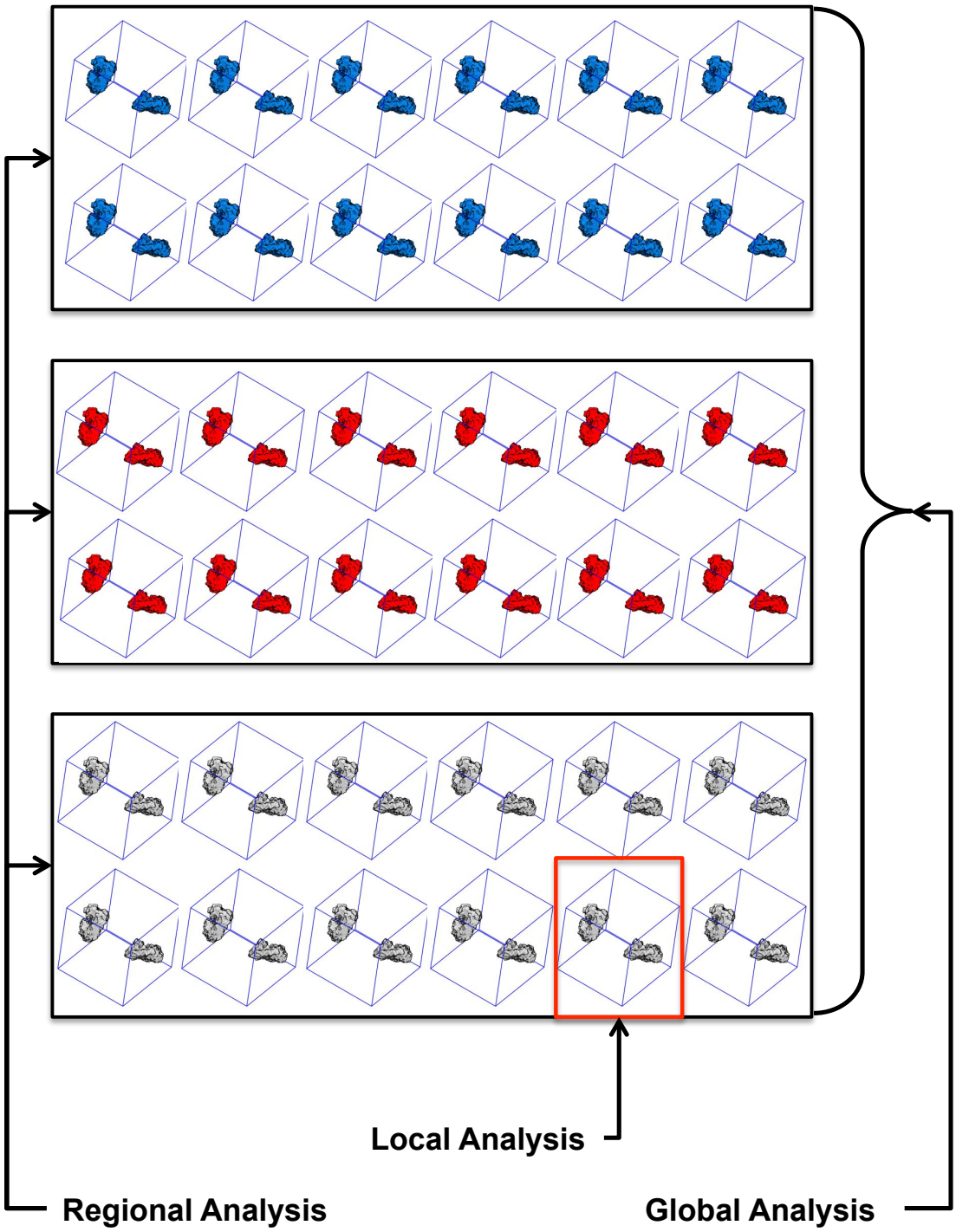


Fig. 4.4 A schematic showing the different levels of analysis performed by the high throughput system, showing local analysis, where individual simulations are analysed independantly of others, regional analysis, where sets of the same protein are analysed together, looking for group properties, and global analysis, whereby all simulations able to be found are analysed together, and global information on interactions are extracted.

only 4 values - Distance, Euler angle 1, Euler angle 2, and Euler angle 3. As these metrics contain only information about the relative positions and orientations of each protein with respect to the other, it allows direct comparison of data from similar simulations.

Another option for when smaller datasets are analysed is fitting trajectories. By fitting a single molecule in a trajectory to a set point, and allowing only the other protein to move within a simulation, two systems where the protein interacts with a similar region on the fixed protein should show similar results, though this method required manual checking that the fitting process has completed successfully.

EMMA is used to cluster the large amounts of angle and distance data into microstates, using the k-centres algorithm (Sun et al. [184]). It was found the 1,000 microstates produced adequate results, without using excessive CPU time (Around 10-15 minutes on 8 CPU cores).

Microstate results are validated by calculating the implied timescales (Bowman et al. [18]) for the Markov model. By counting the number of transitions between microstates at different lagtimes (Windows of frames within the data), it is possible to choose a timescale within the simulations in which the number of transitions between states is as high as possible, and the window of time is a small as possible. This allows the maximal sampling of the number of transitions in the data. When the number of transitions between microstates remains constant even with increased timeframes, the system has reached a "Markovian State", whereby transitions are generally memoryless ie. Not dependant on the state it was in before the previous state. The timeframe selected for counting the maximal number of transitions is generally evaluated by eye from plots of transition count vs lagtime - in the automation of models, an estimate of the lagtime that has worked for previous models is used, but this value is not expected to be useful for all instances without evaluation.

The Markov state model is then constructed by counting transitions at the defined lagtime between all microstates, and assigning a probability for each state change based on the

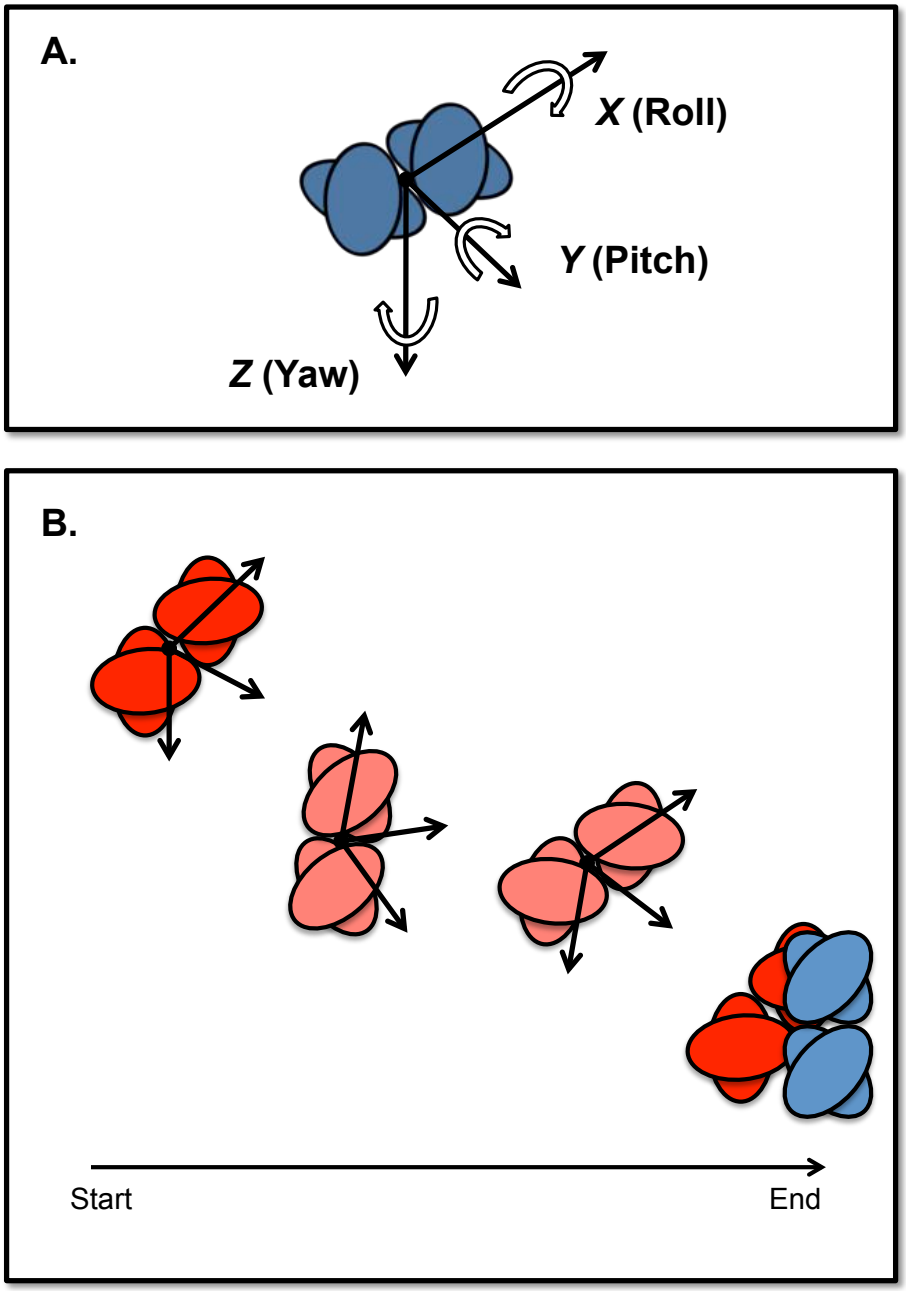


Fig. 4.5 A schematic how Euler angles are assigned for rigid bodies, showing: A) Euler angles assigned to the principle axes of the rigid protein, representing Roll, Pitch, and Yaw (X, Y, and Z). B) A schematic simulation, showing a protein held static (Blue), with the other moving towards it, and eventually contacting (Red). At each snapshot of the simulations, the Euler angles required to map the red protein to the blue one are taken, and used for generation of microstates for Markov modelling.

total number of transitions.

In order to gain meaningful, and understandable data from the model, it must be coarse grained. Microstates are assigned to a user defined macrostate, before their properties and transitions are consolidated and the bulk properties for a certain macrostate are calculated. This way data from thousands of microstates (in this case 1,000), can be consolidated into transitions between 3 macrostates, that capture system behaviour of interest. In this case, distances between the centre of mass of each protein were used for the definition of 3 macrostates:

- A) Proteins Apart - Microstates where the centre of mass distance between proteins is greater than 70 Å.
- B) Proteins Close Together - Microstates where the centre of mass distance between proteins is less than or equal to 70 Å, and greater than 40 Å.
- C) Proteins Completely Together - Microstates where the centre of mass distance between proteins is less than 40 Å.

These macrostates allow the investigation of transitions from proteins being apart at the start of the simulation, to them being together. With rates extracted for transitions between each state (eg. A-B), they can be compared between different molecules, to get estimates of relative rates of aggregation.

A schematic for a model constructed using these three states is shown in Figure 4.6.

4.3 Results and Discussion

The high-throughput pipeline was tested with 6 monoclonal antibody Fab fragments, known to aggregate with different propensities. 40 simulations of each protein were performed and

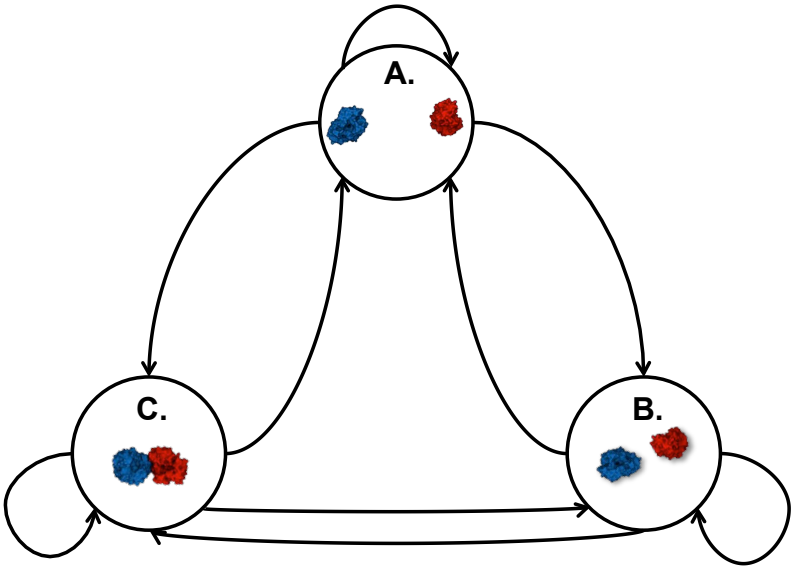


Fig. 4.6 A schematic showing a markov state model constructed using the three states described. Based upon cutoffs of the distance between centre of masses of two proteins, and using Euler angle data for the clustering of microstates, this representation allows the measuring of state transition probabilities.

analysed, and results compared to experimental values for aggregation. The experimental information about each protein can be seen in Table 4.1.

In this chapter, I will focus on regional data, as this provides information about the behaviour of individual proteins, and can potentially be used for the comparison of protein structures.

4.3.1 Pipeline Performance

The pipeline performed well for setup of all 6 protein structures provided by Lonza Biologics. Each protein was successfully setup, and simulated for 40 simulations of 500 ns, for a total of 240 simulations. Each simulation takes around 24 hours for completion when run on 12 CPUs.

Out of 240 simulations, 22 in total failed initially, and after restarts, 7 failed a second time. These systems were re-run from scratch (a new starting configuration and equilibration), another option in this case is to re-run the crashed simulation with a much lower timestep, as performed by sidekick, a similar software package for the automated high throughput submission of simulations (Hall et al. [69]), this would require an increase in simulation time required however.

4.3.2 Results from Regional Analysis

Regional analysis was performed on the results of each set of 40 simulations. Out of the 6 structures run, the three structures with Kappa light chains are presented, (H9K7, H8K8, and H31K5) along with the results for three structures containing Lambda light chains (H17L1, H33L2, and H28L2).

Kappa Light Chain Structures

Structure H9K7 - A High Aggregator

The structure H9K7 is known to be a high aggregating protein - in solution, only 60% of the solution is shown to be soluble monomer. This is comparatively an extremely low number, and so the solution would be extremely immunogenic. Results for the regional analysis of 40 simulations are shown in Figure 4.7. Interaction frequency for the 40 simulations shows two main regions of interest on the structure, one within the CDR region, and one within the opposite end of the protein. Interaction regions are not well defined, and the "hot spots" tend to be spread across a large region. The interaction landscape for this protein also appears to show a small number of high frequency interactions, which indicate that these two interaction areas may be involved in interactions with their counterparts on the opposing protein primarily. Looking at the top two cluster positions from the 40 simulations, they indicate that the two most common interaction "poses" for the proteins involve interactions of the two "hot spots" seen in part A. Whilst the CDR region is expected to interact with itself, from evidence seen in previous work, the region around residue 170 - 200 is a potential site for mutational studies.

Structure H8K8 - A Medium Aggregator

The structure H8K8 is known to be a medium level aggregating protein, exhibiting 94% monomeric content in solution. This is still enough for the protein to be immunogenic, but is potentially a useful structure, as it also has good expression levels. Results for the regional analysis of the H8K8 drug structures are showing in Figure 4.8. Similar regions as seen in the high aggregator H9K7 are highlighted as being high relative frequency interaction "hot spots", though they are not as specific as those seen previously. These regions are also then seen interacting, or at least spatially close to each other in the top two cluster results. There also appears to be a large frequency region in the interaction landscape that

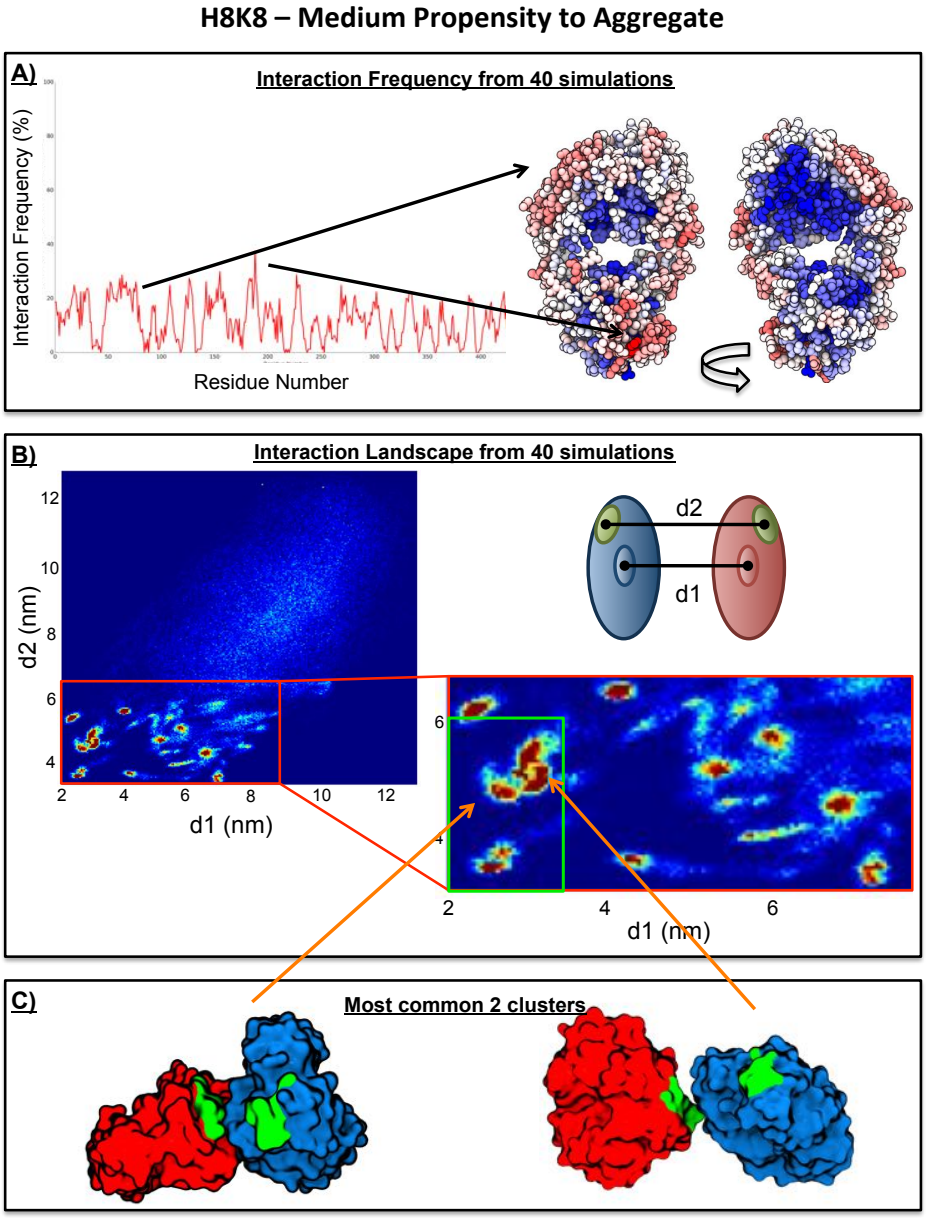


Fig. 4.7 Analysis for Lonza dataset protein H9K7, showing A) Interaction frequency, with relative interaction frequency coloured on the surface of the protein. Two particular regions can be seen as highly interactive on the surface, one in the CDR region, and one on the lower half of the protein. B) Interaction Landscape, showing that there are a few large regions of interactivity, where proteins interact in multiple simulations. In particular one region (Within the green box) indicates a large amount of CDR-CDR contact. C) Two top clusters resulting from the combination of all simulation. These match well with parts A) and B), as they show the same regions, and can be matched to the interaction landscape.

represents CDR-CDR type interactions, which result in these being the top two clusters. This behaviour is slightly different to the high-aggregate prone protein, which shows more interactions involving the non-CDR hotspots. This may explain the different behaviours of the proteins in experimental work, with very different monomer concentrations (60% compared to 94%).

Structure H31K5 - A Low Aggregator

Protein H31K5 is classified as a low aggregator, as it stays at roughly 100% monomer concentration in solution, making it an excellent candidate for potential further use, it also has a high expression rate. Results from the regional analysis of 40 simulations are shown in Figure 4.9. The relative interaction frequency analysis shows only a single "hot spot" on the protein surface that is involved in a high number of interactions. This region is within the CDR, and thus is expected to be involved in self/self interactions, from previous work. The interaction landscape for the molecule appears to have one large region where multiple high frequency regions are close, and clustering analysis indicates that this region is mostly CDR-CDR, and CDR-other interactions.

Lambda Light Chain Structures

Structure H17L2 - A High Aggregator

The structure H9K7 is known to be a high aggregating protein - in solution, roughly 84% of the solution is shown to be soluble monomer. This is comparatively an extremely low number, and so the solution would be extremely immunogenic. Results for the regional analysis of 40 simulations are shown in Figure 4.10. Interaction frequency for the 40 simulations shows two main regions of interest on the structure, one within the CDR region, and one within the opposite end of the protein. Interaction regions are not well defined, and the "hot spots" tend to be spread across a large region. The interaction landscape for this

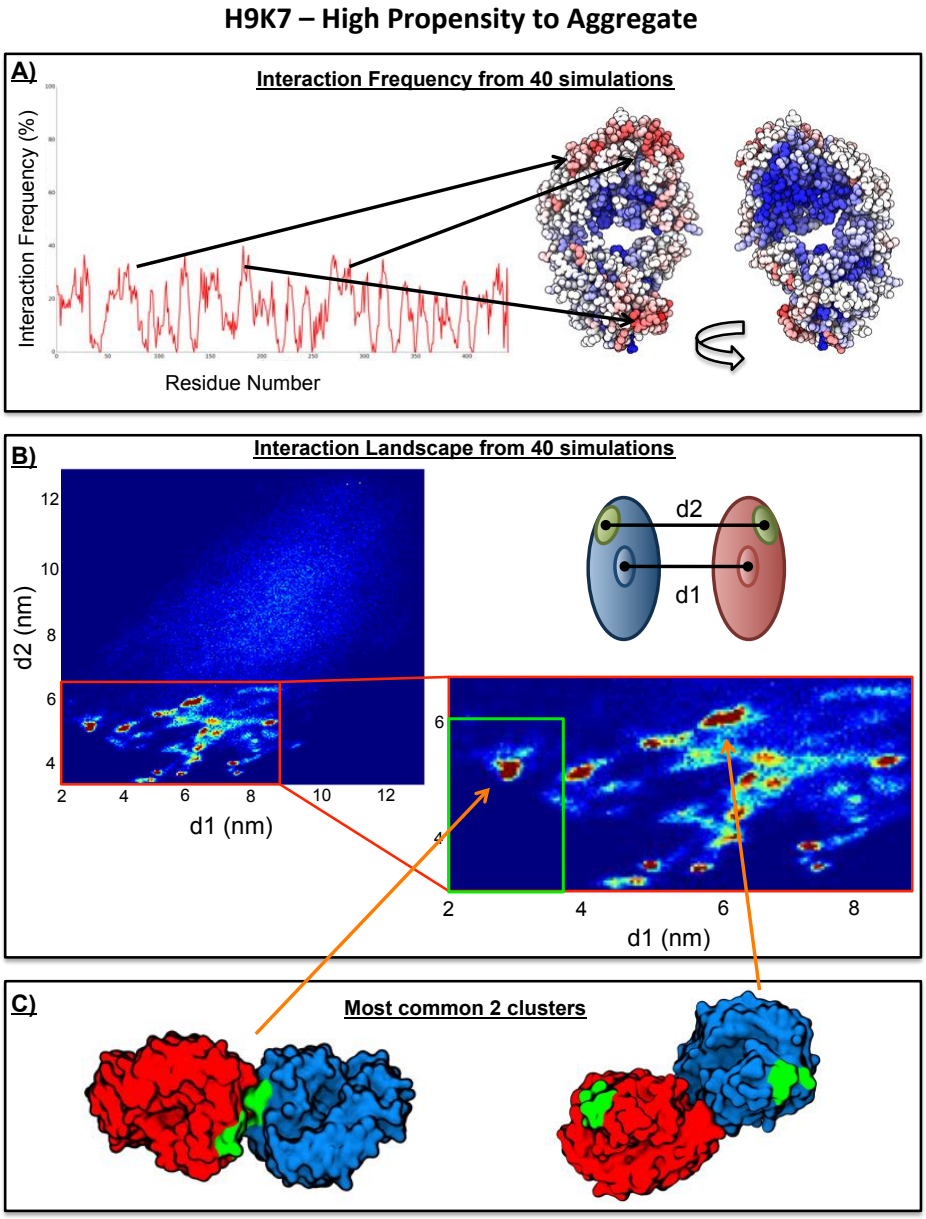


Fig. 4.8 Analysis for Lonza dataset protein H8K8, showing A) Interaction frequency, with relative interaction frequency coloured on the surface of the protein. The whole surface of the protein has regions of interactivity, but two particular regions can be seen as highly interactive on the surface, one in the CDR region, and one on the lower half of the protein, similar to the H9K7 protein. B) Interaction Landscape, showing that there are a few large regions of interactivity, where proteins interact in multiple simulations. In particular there is a large region (Within the green box) indicating CDR-CDR contact. C) Two top clusters resulting from the combination of all simulation. These match well with parts A) and B), as they show the same regions, and can be matched to the interaction landscape.

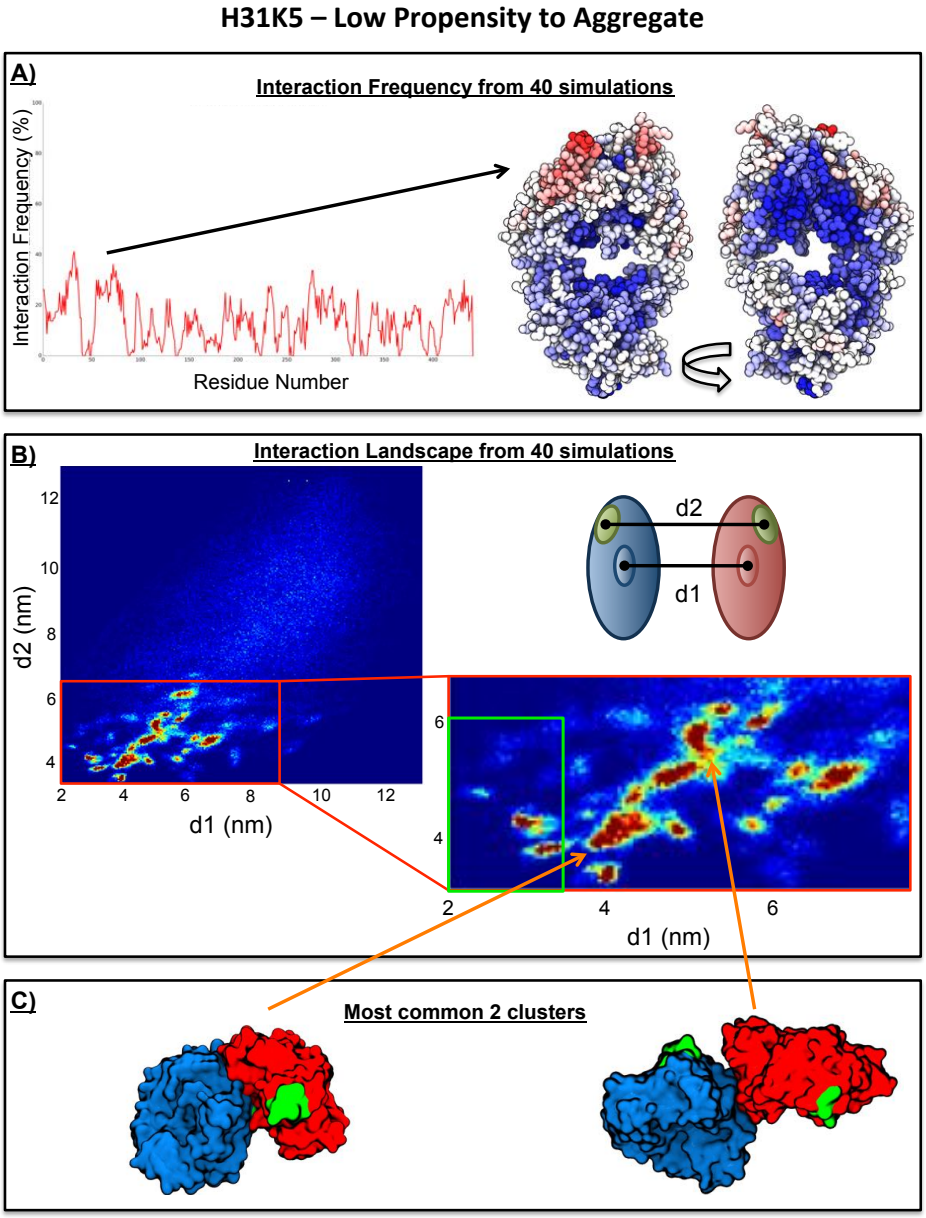


Fig. 4.9 Analysis for Lonza dataset protein H31K5, showing A) Interaction frequency, with relative interaction frequency coloured on the surface of the protein. There is a single patch of high interactivity in the CDR region of the protein B) Interaction Landscape, showing a large region of interactivity made up of lots of smaller regions. There is also very little density within the self-self region (green box), a majority of interactions appear to be CDR-other interactions C) Two top clusters resulting from the combination of all simulation. These match well with parts A) and B), as they show the same regions identified in part A, and can be matched to the interaction landscape plot.

protein also appears to show that the proteins favour interactions with the non-CDR regions of their structure, as confirmed by the poses adopted in the top two clusters. Whilst regions in the CDR are known to self interact from previous studies, and from experimental data, the region highlighted in red on the protein structure, on the bottom face of the protein, is a potential region of interest, especially as it seems to dominate interactions based on the interaction landscape and top two clusters.

Structure H33L2 - A Medium Aggregator

The structure H33L2 is known to be a medium level aggregating protein, exhibiting 97% monomeric content in solution. This is still enough for the drug to be immunogenic, but is potentially a useful structure, as it also has good expression levels. Results for the regional analysis of the H33L2 drug structures are showing in Figure 4.11. Regions in similar areas to H17L2, the Lamda high aggregator are observed - in particular, an area around the base of the structrue is highlighted again. This regions are also then seen interacting, or at least spatially close to each other in the top two cluster results. The interaction landscape appears to show a similar pattern to the high aggregator that is a Lamda structure, in that it is predominantly non CDR-CDR interactions, particularly involving the region shown at the bottom of the molecule. This high frequency interaction region at the bottom of both aggregating Lamda structures is a potentially useful area for interaction, particularly since this region is not observed aggregating in the low aggregator analysis.

Structure H28L2 - A Low Aggregator

Protein H28L2 is classified as a low aggregator, as it stays at 100% monomer concentration in solution, making it an excellent candidate for development, it also has a high expression rate. Regional analysis of 40 simulations are shown in Figure 4.12. The interaction frequency analysis shows interactions within two regions - one around the CDR region (though

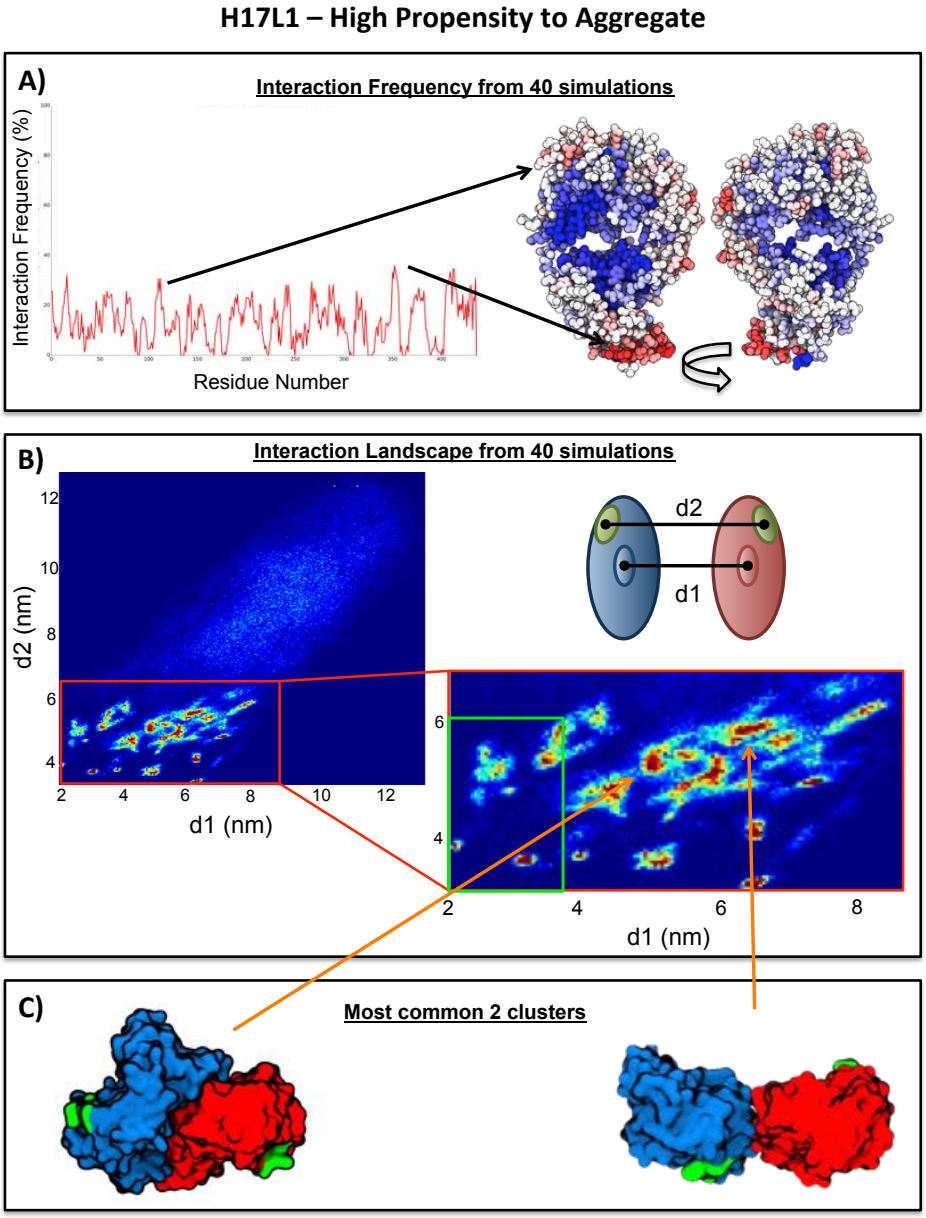


Fig. 4.10 Analysis for Lonza dataset protein H17L1, showing A) Interaction frequency, with interaction frequency coloured on the surface of the protein. Two particular regions can be seen as highly interactive on the surface, one in the CDR region, and one on the lower face of the protein. B) Interaction Landscape, showing that there is a large region of interactivity where the proteins interaction through their bottom edges, and not CDR regions. In particular there is a comparitively low amount of interaction with the green box, which represents CDR-CDR interactions, and a large desnity indicating poses where the proteins are "side on" to each other.

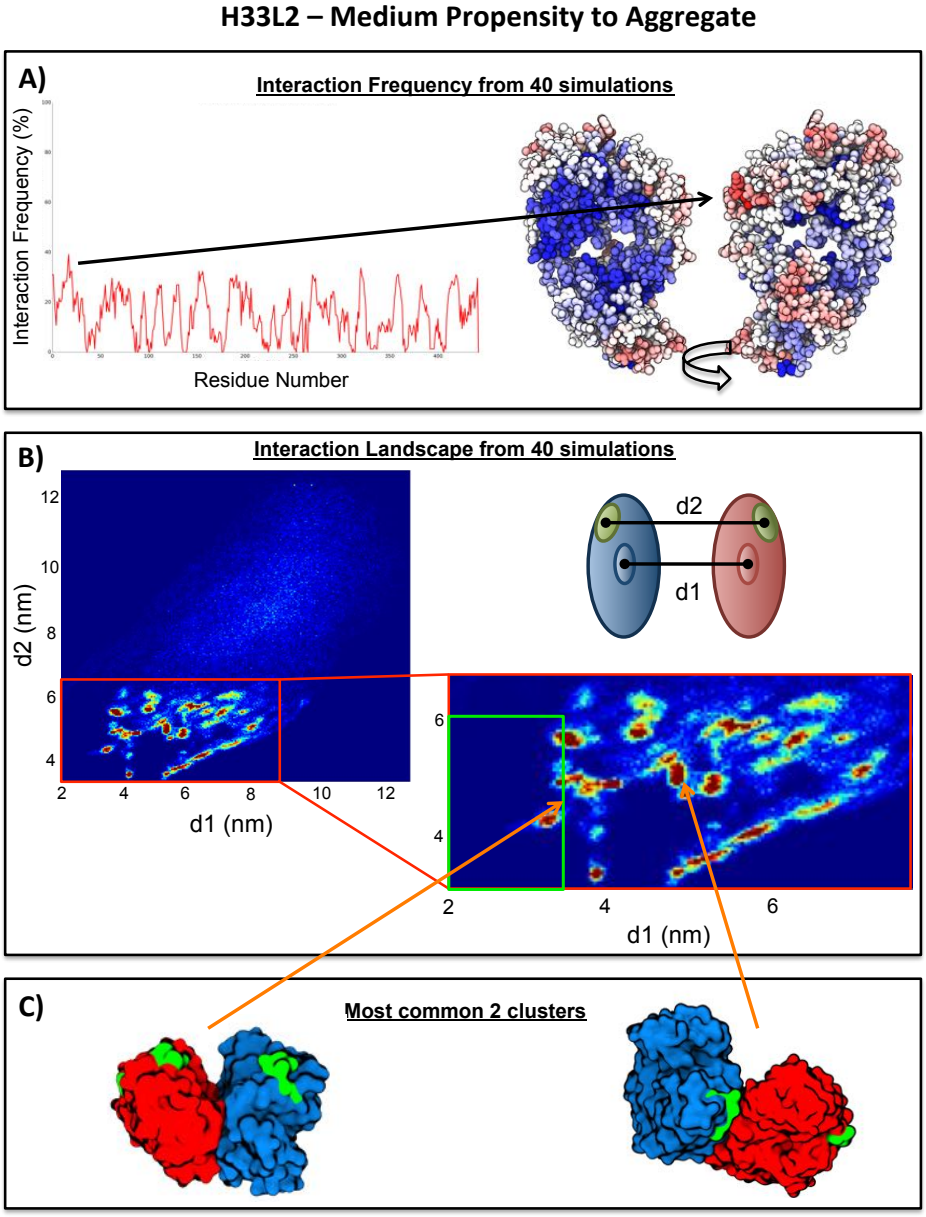


Fig. 4.11 Analysis for Lonza dataset protein H33L2, showing A) Interaction frequency, with interaction frequency coloured on the surface of the protein. The whole surface of the protein has regions of interactivity, but two particular regions can be seen as highly interactive on the surface, one in the CDR region, and one on the lower region of the protein, similar to the high aggregator H17L1. B) Interaction Landscape, showing that there are a large number of interactions in the space representing non CDR interactions, which is corroborated with the top two cluster results (Part C).

in a different location to the other low propensity aggregator), and one around the midpoint of the molecule. The landscape for the molecule is very sparse compared to the higher aggregator proteins, with a cluster occurring at the region where a single CDR interacts with the opposing protein, as shown by conformations of the top clusters (part C) .

4.3.3 Study of Association Rates

Studying the rate of association for proteins for a large number of simulations can be done in numerous ways. One extremely simple way of measuring association rates is to take the percentage of each simulation in which the two proteins are either together, or apart. Performing this with the simulations of the Lonza Dataset can be seen in Table 4.2. There is no correlation between these values for these simulations and the experimental polymer percentage. Another value that may correlate with the experimental values is the standard deviation for these frame analyses, as the proteins that are less likely to aggregate experimentally may be expected do have a larger variation in the time it takes them to interact, as they may occasionally interact very quickly in a simulation when the correct regions of each protein come into contact with each other, and often take a long time to interact (or not interact at all) due to the rarity of these surfaces. Calculating this value for the simulations of the Lonza Dataset shows that there is in fact, very little variation in these values, Table 4.2.

4.3.4 Markov State Models

Markov state models were constructed from the results of each set of simulations. A lagtime of 80 ns was to be most appropriate, whereby 80 ns windows within each simulation allowed transitions between states on a regular basis. Microstates were clustered into 3 macrostates, as discussed earlier - A, B, and C, shown in schematic form in Figure 4.6. The rates of transition between states are then measured, for each protein in every simulations. Results

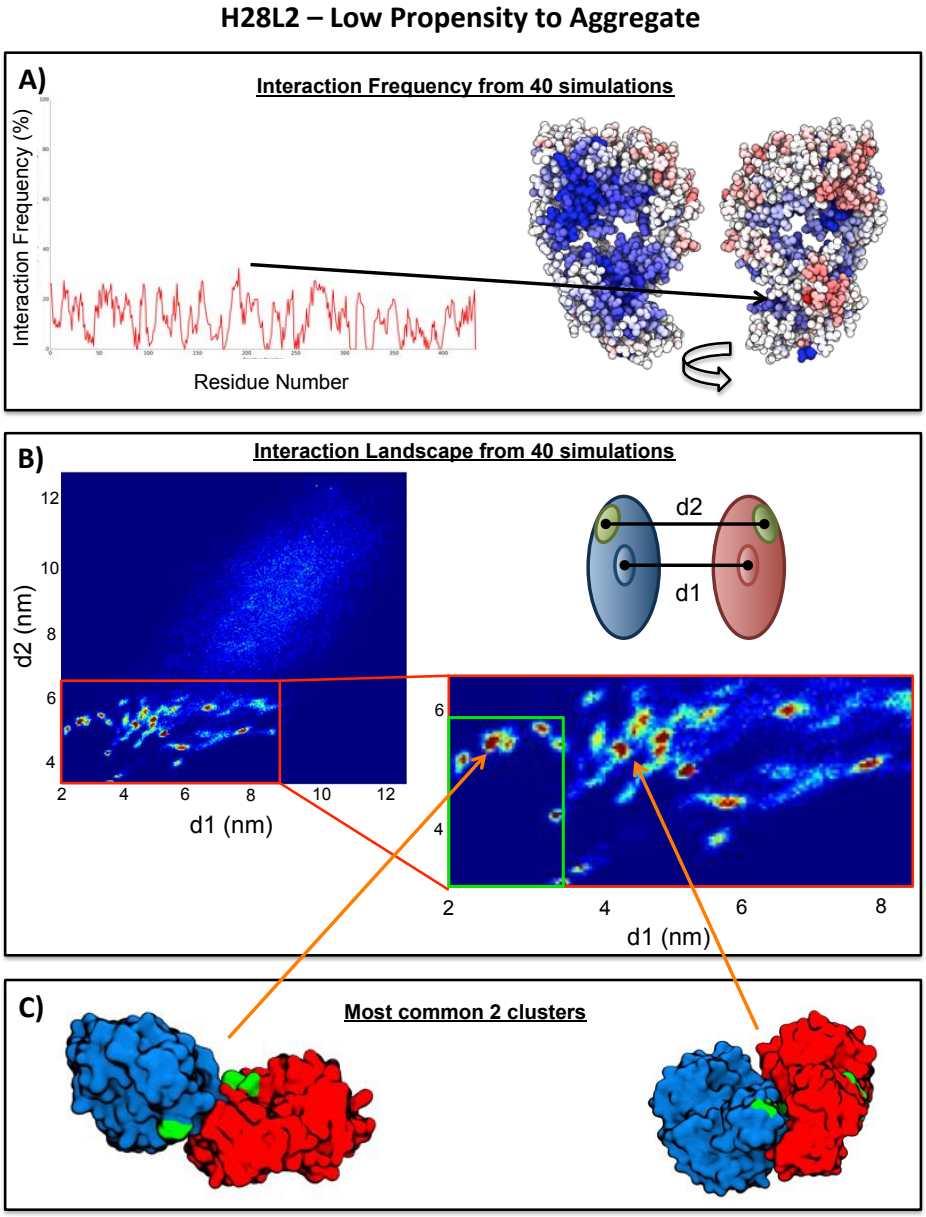


Fig. 4.12 Analysis for Lonza dataset protein H28L2, showing A) Interaction frequency, with interaction frequency coloured on the surface of the protein. There are two regions of high interactivity, highlighted on the protein. B) Interaction Landscape, showing sparse interactions. C) Two top clusters resulting from the combination of all simulations. These match well with parts A) and B), as they show the same regions identified in part A, and can be matched to the interaction landscape plot.

Table 4.2 Comparison of Association Rate through Frame Analysis to Experimental Aggregation Propensity

Lonza Dataset Name	Poly. Conc. (%)	% of Frames together (%)	Standard Deviation
Kappa VL chain	Experimental Data	Simulation Data	
H31K5	0.01% (Low)	48.2%	28
H8K8	6% (Medium)	41.3%	29
H9K7	40% (High)	43.6%	30
Lambda VL chain			
H28L2	0% (Low)	43.8%	28
H33L2	3% (Medium)	56.2%	26
H17L1	16% (High)	43.5%	30

for the 6 members of the Lonza dataset set are shown in table 4.3. There is a correlation between the A -> C transition probabilities and percentage polymer concentration. The A -> C transition represents regions of the simulations where the proteins go from being extremely distant, to close contact with each other within a single window of 80 ns. The higher the probability of this occurring, the more likely two proteins are to interact. The Markov State Model approach has, for a small dataset of 6 proteins, accurately predicted the rank order of aggregation propensity of protein structures from a large number of self-self interaction simulations, with a correlation value of 0.85. Though the fact that only the A -> C transition probabilities show this trend is a cause for concern, extension of the dataset is required to truly know of the MSM can accurately predict the correlation.

The correlation between the A -> C transition probability, and the Percentage Polymer Concentration from the experimental data can be seen in Figure 4.13. There is a strong correlation (with an R value of 0.85) between the two values.

4.3.5 Discussion

In this chapter I present a high-throughput methodology for predicting the relative aggregation propensity of immunoglobulin Fab structures. A high-throughput analysis procedure is

Table 4.3 Markov Model Results for Lonza dataset Protein Simulations

Lonza Dataset Name	Poly. Conc. (%)	A -> B (%)	B -> C (%)	A -> C (%)
Kappa VL chain	Experimental Data	Markov Model Data		
H31K5	0.01% (Low)	77%	68%	77%
H8K8	6% (Medium)	71%	64%	80%
H9K7	40% (High)	80%	64%	89%
Lambda VL chain				
H28L2	0% (Low)	77%	56%	73%
H33L2	3% (Medium)	76%	48%	73%
H17L1	16% (High)	77%	66%	82%

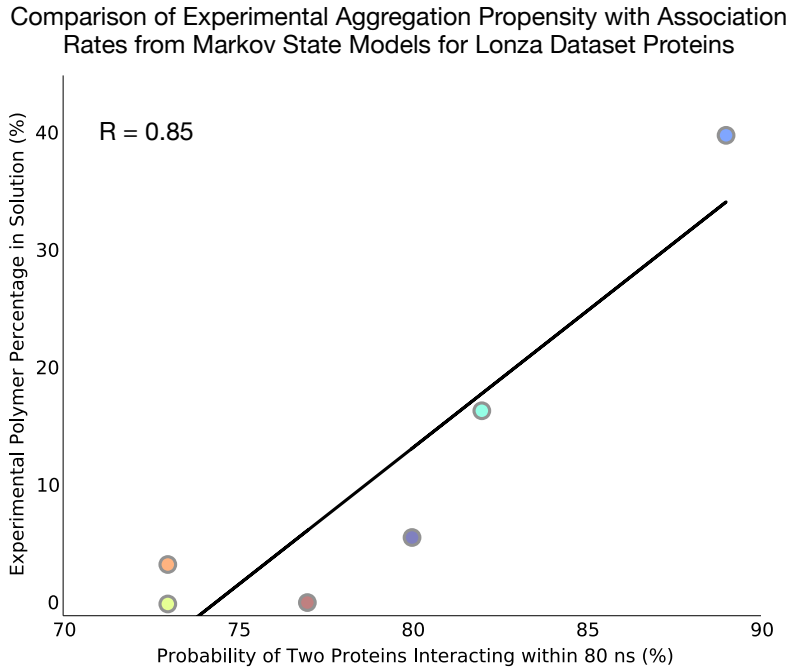


Fig. 4.13 Correlation between Experimental Percentage Polymer in Solution and Computational Markov State Model A -> C transitions. The correlation appears strong, with an R value of 0.85.

described, along with a robust setup and simulation method for acquiring large amounts of interaction data. The system is modular, robust, and customizable, and has been tested with the generation of 240 trajectories for 6 different proteins of different aggregation propensities. The script performed well, with few failures. Different analysis levels are identified, and separated, and analysis can be focussed on local, regional, or global characteristics. "Hot spots" of mutational interest can be identified for individual proteins, and the behaviour of structures can be predicted based on the frequency and number of the regions, and the ways in which they interact as seen in interaction landscape plots. General behavioural characteristics can be assessed, and there is a difference in the activity of proteins that have different experimental aggregation propensities, with defined "hot-spots" being larger, more spread across the protein surface, and located in non-CDR regions of the protein for structures that have a high aggregation propensity. This approach - generating ensembles of simulations at the press of a button, greatly simplifies the previously used work flows, which involve setting up, simulating, and analysing large numbers of simulations manually, which is very prone to error.

As a supplement to the general analysis of high-throughput data, I also present a novel application of Markov state modelling, to protein-protein interactions. This is the first time this technique has been applied to biological protein-protein interactions, and rate constants that have been extracted, whilst unable to be directly compared to experimental rate constants, can be comparatively matched to their equivalent drug structures. Results show that the Markov model technique replicates relative aggregation propensity when compared to percentage monomer concentration from experimental results. H9K7, and H17L1 - both protein structures known to aggregate to a large degree, show the largest probably of moving from a state where they are over 70 Å apart, to a state where they are in contact with each other, with the structure with the smallest monomer percentage (60%), also having the largest probability interactions occurring (89%) within 80 ns, when compared to the

other high aggregator structure (monomer percentage of 84%, interaction probability 82%). The sample size for this result, however, is extremely small. Having only 6 results reduces the confidence level for results applying to other protein significantly, and simulations on a much larger range of proteins will be required in order to fully validate the mechanisms.

Here, the advantages of the use of High-Throughput CG-MD are highlighted, allowing the simulation of much larger arrays of simulations, for much longer timescales, and allowing the analysis of much larger amounts of data, for better sampling. Whilst resource time required is still relatively high, requiring approximately 12,000 CPU hours per protein structure, this amount of resource time is available with the use of shared resources, such as the IRIDIS supercomputer. GPU clusters, such as EMERALD, are also highly suited to use in high-throughput molecular dynamics, as they are excellent at performing short simulations of small systems.

Extensions to the work will involve an expansion of the experimental Lonza dataset, to confirm that behaviour can be accurately predicted when a much larger number of structures are used as input. It would also be of interests to, as was suggesting in the previous chapter, perform mutational studies on the hot-spot regions identified from the surface frequency analysis. It would also be very useful to look at the convergence of the results for these studies, to assess whether 40 simulations are enough, or too many, to see convergence of rates and interaction residues. A starting point would be apply a sampling method such as bootstrapping, or jackknife analysis, whereby subsets of data are tested to determine how many simulations would be required for the same results to occur regularly (Efron [53]).

Limitations of this methodology include the large amount of computational resources required to run such a large number of simulations. Though this is much less than would be required for other methodologies, e.g. fully atomistic simulations, the amount of resources required is still high, and thus the methodology is not suitable for running on e.g. local compute resources. The results also need to be extended from a simple test set of 6 proteins

in order for it to be fully validated. There is also still a certain amount of manual input required for the accurate generation of Markov state models. Whilst a lagtime of 80 ns was sufficient for 6 structures featured in this chapter, it is perhaps not likely that this will be effective for all proteins in the Lonza dataset, and at this point, it will be difficult to directly compare the results from proteins that require a different lagtime in order for the Markov state model to be accurate.

Chapter 5

Coarse Grained Modelling of Oligosaccharides

"The beauty of a living thing is not the atoms that go into it, but the way in which those atoms are put together."

Carl Sagan

This chapter addresses the problems with performing coarse-grained simulations of carbohydrates with the development of a novel parameter set, and simulation methodology for their study. In this chapter I focus on the general development of the forcefield, starting with extremely simple biological sugars, and building up to larger structures of glycolipids, which are of intermediate complexity between simple carbohydrates, and much more complex protein glycosylation, and are of interest to other members of the research group.

5.1 Introduction

5.1.1 Carbohydrates

The main biological function of carbohydrates is in the post-translational modification of proteins or lipids, forming glycoproteins or glycolipids (Dwek [50]). These can directly influence the function of the molecule to which they are attached, as is the case with some proteins, such as the immunoglobulin G (IgG) (Arnold et al. [5]), and all glycolipids (Brandenburg and Holst [20]), or they are used as regulatory, signalling, or transport tags, as is the case with most glycoproteins (Furukawa et al. [61]). Given the importance of carbohydrates in glycoproteins and glycolipids in the regulation and function of biological molecules, there is a need for increased investigation utilizing computational studies. Better computational techniques are required for understanding the full picture of many biological systems, including most membranes.

5.1.2 Cerebrosides

The cerebrosides are one of the earliest discovered types of glycolipid, existing mainly in the brain and neural tissues. They are structurally characterized by two components: a carbohydrate headgroup, which is attached via a glycosidic bond to a ceramide molecule (Figure 5.1). In this study, we focus on two main types of monocerebrosides: glucocerebroside and galatocerebroside.

Glucocerebroside

Glucocerebroside is characterized by the linakge of a single glucose molecule to a ceramide molecule. It is mainly found within neural membranes, and is the known cause of Gaucher disease, an illness caused by the accumulation of glucocerebroside due to mutations in the enzyme that catalyses its degradation - glucocerebrosidase (Hruska et al. [80]). Gluco-

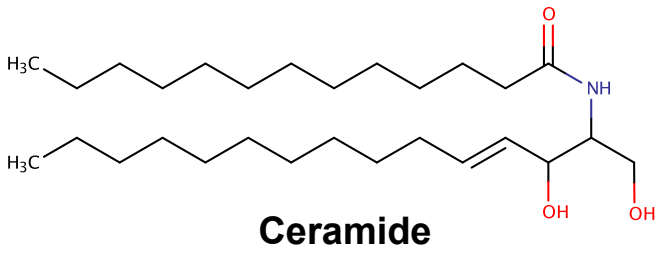


Fig. 5.1 Structure for the molecule ceramide, the common base between all cerebroside lipids.

cerebroside has also recently been implicated as a potential cause of Parkinsons disease in certain populations (Lwin [121]). Glucocerebroside is a major component in the myelin sheath, being the primary way in which a permeability barrier is maintained (Holleran et al. [76]).

Glucocerebroside is also found within plants, and is generally thought to be the most common glycolipid employed by plant life (Sperling et al. [177]).

Galactocerebroside

Galactocerebroside, similar to glucocerebroside, is characterized by a single galactose molecule attached via a glycosidic bond to a ceramide molecule. It is also found within neural tissue, and a defect in the enzyme catalysing their degradation is responsible for Krabbes disease (Suzuki and Suzuki [185]). They are also a major component of the myelin membrane (Stoffel and Bosio [182]). The structure can be seen in Figure 5.2.

5.1.3 Gangliosides

The most common form of glycolipids in mammalian membranes are gangliosides, characterized by the addition of a negatively charged N-acetylneuraminic acid (Neu5ac, or NANA) molecule at the end of the carbohydrate lipid headgroup. They carry a net negative charge, and are located on the outer membranes (Sonnino et al. [176]). They occur in almost all tis-

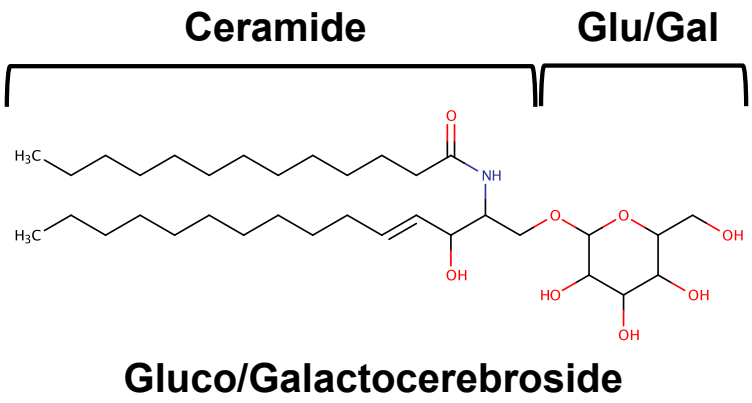


Fig. 5.2 Structure for Gluco and Galactocerebroside, showing ceramide tails, and a glucose or galactose carbohydrate head group.

sues within the mammalian body, though there are elevated levels within neural membranes in the brain and central nervous system (Sastry [162]), and their misregulation is generally implicated in neurological, and central nervous system disorders (Xu et al. [205]). Gangliosides were first isolated from ganglion cells within the brain (Yamakawa [206]), and are now the most studied subtype of glycolipid. Gangliosides are known to play some role in the differentiation and specialization of neural tissues, knock out mutants generally effect neural development, or are fatal (Yu et al. [210]).

GM3

Monosialodihexosylganglioside (GM3) is a glycolipid ganglioside with a headgroup consisting of Neu5Ac, Gal, and Glu carbohydrates. The headgroup is consists of a single molecule of each of the carbohydrates attached linearly to a ceramide molecule. GM3 is always found on the outer leaflet of lipid bilayers surrounding cells, and is known to form self-associated "lipid rafts" under certain conditions (Rajendran and Simons [148]). It is known to have interactions with the insulin receptor, and it is suggested that "metabolic disorders, such as type 2 diabetes, are membrane microdomain disorders caused by aberrant expression of gangliosides" (Inokuchi [85]). GM3 is also known to interact with EGFR

receptor proteins, being involved in their regulation and phosphorylation (Yamashita et al. [208]), and is known to be involved in cell adhesion and signal transduction in neural tissue (Iwabuchi et al. [87]). There is also a direct link known to exist between lack of GM3 and the degradation of the Organ of Corti, and therefore hearing loss in mice (Yoshikawa et al. [209]). The structure can be seen in Figure 5.3.

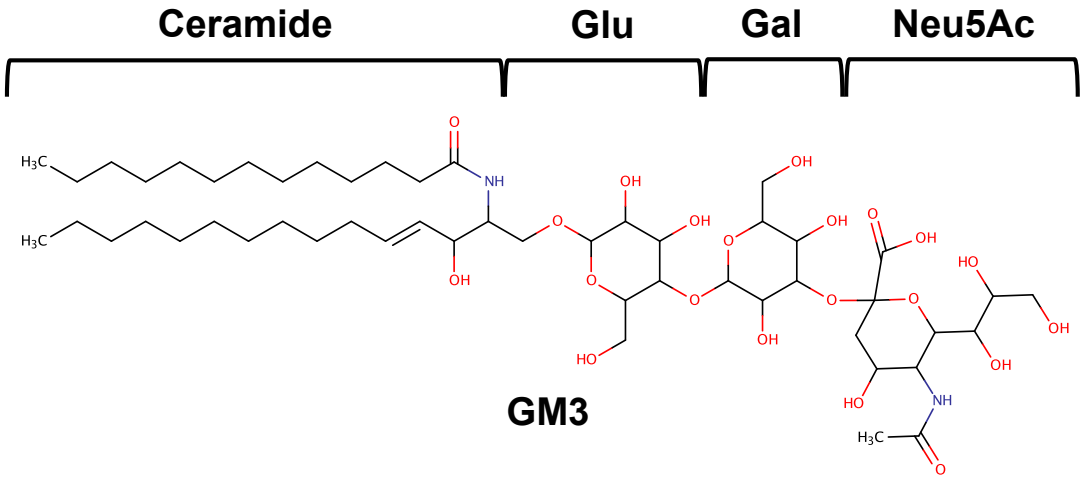


Fig. 5.3 Structure for glycolipid GM3, showing ceramide tails, and a glucose, galactose, N-acetylglucosamine carbohydrate head group.

Forssman Glycolipid

The Forssman glycolipid is a glycolipid containing a headgroup composed of the Forssman antigen, attached to a ceramide molecule. The Forssman antigen contains five residues, consisting of GalNac, Gal, and Glu carbohydrates, in a 2:2:1 ratio attached to a ceramide molecule. It was first isolated by J. Forssman in 1911 (Yamakawa [206]), and its structure was first accurately described in 1971 (Siddiqui and Hakomori [171]), since then, little further development has occurred. The Forssman glycolipid is structurally similar to the blood group A antigen, and this is generally thought to be a reason for it being highly immunogenic (Yamamoto et al. [207]). The Forssman antigen or glycolipid is thought to not be produced in human tissue - but has been found expressed in certain incidences of cancer

(Ono et al. [138]). The glycolipid has been isolated from sheep, mouse, cow, rabbit, and horse (Yamakawa [206]), and it is Flu virions budding from membranes in these organisms that contain the Forssman glycolipid. The glycolipid is known to be a major component of flu virus membranes, due to the flu virus budding from mammalian membranes itself (Gerl et al. [62]). Little else is known about the Forssman glycolipid, aside from the fact that it's expression levels change during embryonic development. The structure can be seen in Figure 5.4.

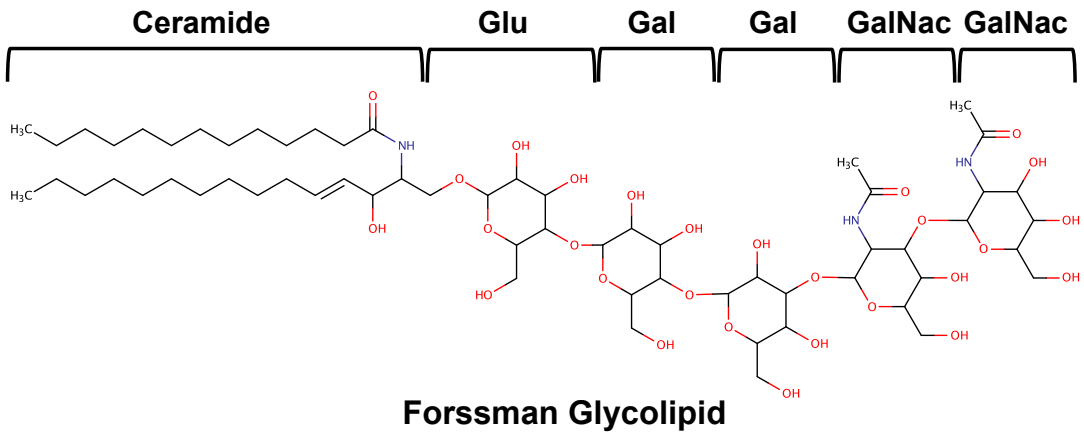


Fig. 5.4 Structure for the Forssman glycolipid, showing ceramide tails, and a glucose, galactose, N-acetylgalactosamine carbohydrate head group.

5.1.4 Previous Work

Whilst forcefields for simulating lipid and carbohydrate molecules have existed for some time (Kirschner et al. [102]), there has been relatively little work on combining the two, though some progress has been made. Noticeably, fully atomistic simulations of a GM3 bilayer have been performed using a variation of the GROMOS forcefield (Sega et al. [165]). Other atomistic simulations have been performed on simple glycolipid species, such as GM1 (Jedlovszky et al. [90]) and GM3 (DeMarco and Woods [45]) - but studies generally focus on a single lipid molecule in a bilayer, as simulating large membrane systems in full atomistic detail is computationally expensive.

Recently, parameters for carbohydrates, and some glycolipids, have been made available within the MARTINI CG forcefield (López et al. [118]), based on the "sweet MARTINI" sugar parameters (Lopez et al. [117]), though there are some notable exceptions, including GM3 and GCCE. The parameters are also generally restrained through the use of an elastic network, rather than a system of bond and torsion angles. Whilst this means that general shape and flexibility of the lipids are restrained, this means that structures are not modular, and that reparametrization of a similar structure will require extensive reparametrization of the elastic network each time. It also means that specific structural details, such as ring-ring positions are not controlled, emergent behaviours of the lipids, such as interactions with other membrane objects, may not be present, Figure 5.5.

In this chapter we present a novel methodology for the representation and simulation of carbohydrate molecules within the MARTINI coarse-grained forcefield. By starting off parametrizing small, simpler carbohydrates, and building up to larger, more complex glycolipids, we aim to show that the forcefield can replicate important emergent properties of complex molecules, and is able to overcome many of the complications involved in the simulation of oligosaccharides.

Future work in this area would involve extending the parameter sets to the glycosylation of proteins, relevant to this project is the glycosylation of antibody Fc fragments, whereby glycosylation is essential to the stability, and known to impact function. There are also numerous other glycolipids that could easily be parametrized using the methodology. GPI anchors are another target for the method, having essential physiological functions, but not yet being tackled with the CG methodology.

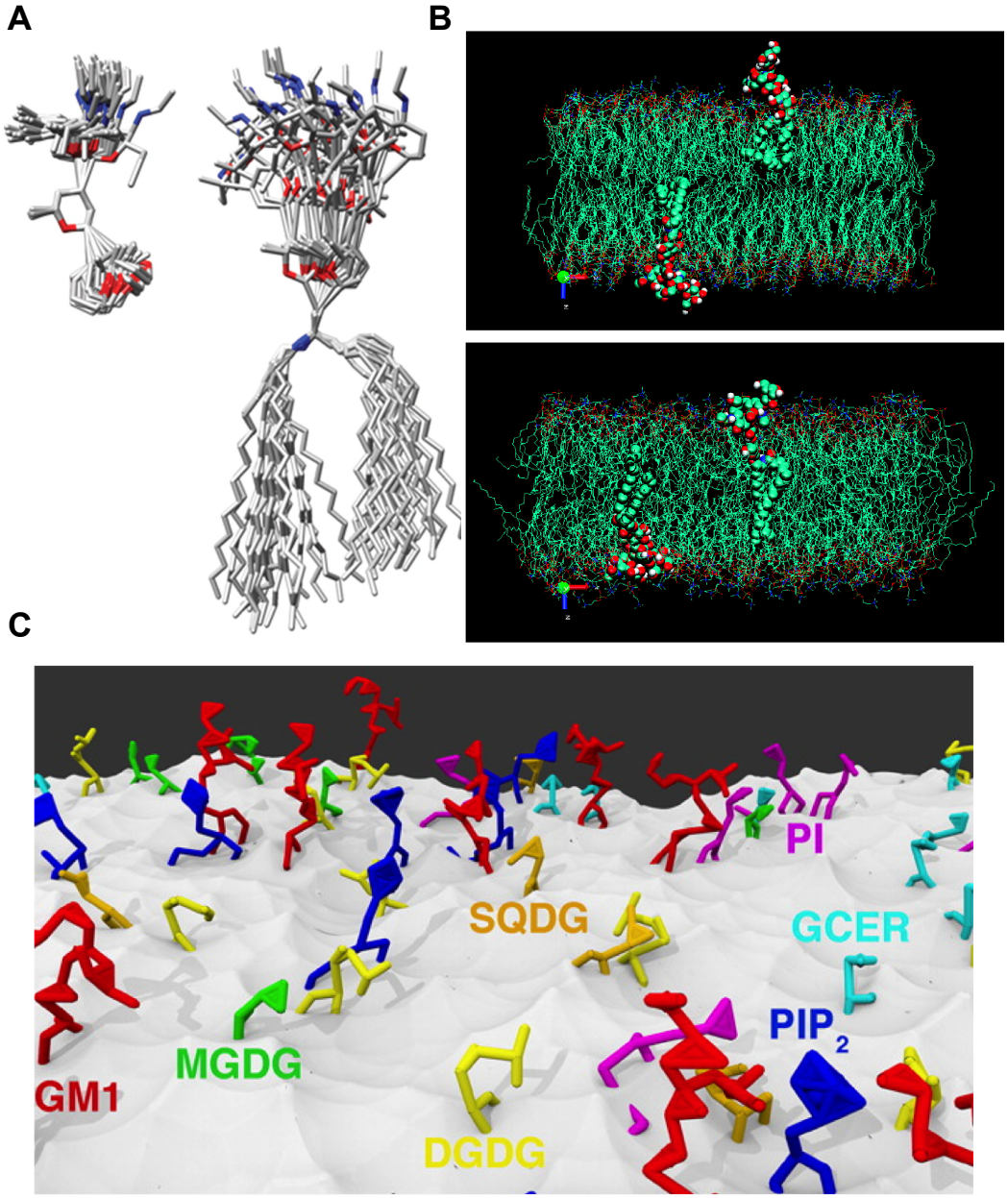


Fig. 5.5 Previous work done on simulations of carbohydrates and glycolipids, showing A) Simulations of AT detailed GM3 within a bilayer environment (DeMarco and Woods [45]), including conformational fluxuations of GM3 within it's bilayer environment. B) An atomistic bilayer containing GM1 lipids, figure adapted from (Jedlovszky et al. [90]). C) MARTINI glycolipid structures publicly available. Figure adapted from (López et al. [118]).

5.2 Methods

5.2.1 Atomistic Simulations of Sugar Molecules

Atomistic simulations of sugar molecules were run using the AMBER11 simulation package (Case et al. [26]), and the GLYCAM06 forcefield (Kirschner et al. [102]). Initially, carbohydrate structures were produced using the GLYCAM06 web server, and the most common rotamer conformation selected. Molecules were energy minimized for 200 ps before being solvated. The solvated system was then energy minimized for a further 200 ps, and over a further 5 ns ,the system was heated from 0 to 310 K, coupled to a Berendsen thermometer, and the carbohydrate molecule restrained. Afterwards the system was released and full molecular dynamics employed, using the NPT scheme for 200 ns, before bond distances, angle parameters, and dihedral parameters extracted using the MDAnalysis python package (Michaud-Agrawal et al. [127]).

Reliability of Atomistic Simulations

In order to reliably generate coarse grained structures, the atomistic simulations they are based upon should be either sufficiently long to allow the system to relax and sample all ranges of conformational motion, or biased in such a way as to force the sampling of all available conformations within a short time. Figure 5.6 shows the RMSD for all atoms (including hydrogens) in a 200 ns simulation. After an initial small rise in RMSD, the structure appears to become stable, with the RMSD not rising much higher than 1 Å. This indicates that the simulations run for 200 ns are able to fully sample the expected conformational space of the molecule. Any large scale shifts in the conformation of the ring, such as a switch from boat to chair conformations, are not expected to occur in MD simulations and would be accompanied by a much greater change in RMSD caused by the concerted motion of all atoms within the molecule. The relatively low RMSD value presented here gives

confidence that the conformational sampling in the atomistic simulations is sufficient for parametrisation of the coarse grained structures.

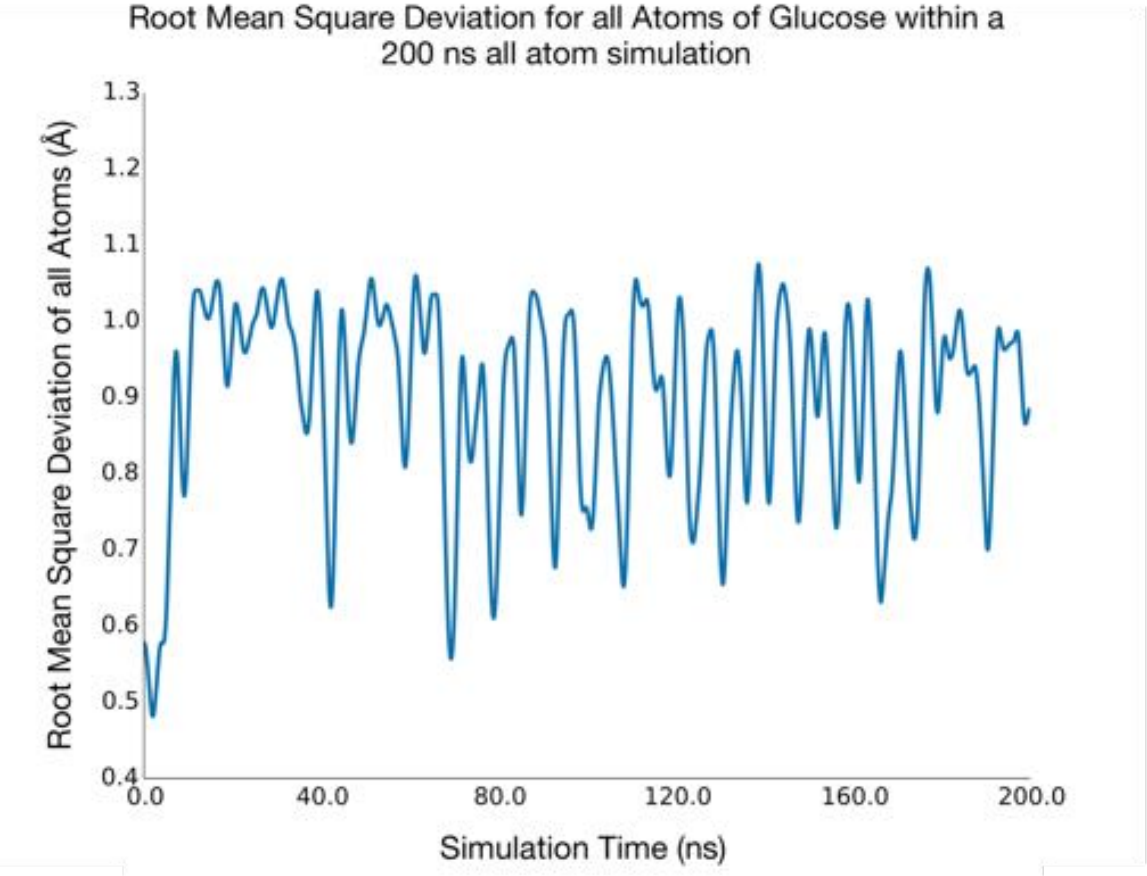


Fig. 5.6 Root Mean Square Deviation for all atoms in a single moleucle of Glucose in solvent, showing that the molecule does not undergo large scale conforamtional dynamics shifts, and that the simulations are sufficiently long enough for the parameterisation of coarse grained structures.

5.2.2 Coarse Grained Simulations of Sugar Molecules

Coarse-grained simulations of sugar molecules were run using the Martini v 2.2 forcefield (De Jong et al. [41]) within GROMACS version 4.5 (Pronk et al. [146]). Non-bonded interactions between all particles are treated as a Lenard-Jones potential, shifted to zero between 9 and 12 Å. Electrostatic interactions are calculated with a Columbic potential, and shifted to zero between 0 and 12 Å. Sugar molecules are energy minimized over 200 ps, before the

molecule is solvated, with enough room given for the molecule to not see its periodic image. The solvated system is then energy minimized for a further 200 ps, before carbohydrate regions are restrained via the central rotational node particle, and lipid tails are restrained, and the system is equilibrated over 50 ns. Afterwards the system is released and full molecular dynamics performed for 500 or 1000 ns. Parameters of interest are then extracted using the MDAnalysis python package.

5.2.3 Iterative Force-Matching

For transferring parameters from the AT systems to CG models, angle and distance distributions are extracted from the fully atomistic simulation trajectories. The aim of the force-matching protocol is to transfer the bulk properties of atoms within fully atomistic simulations, to the equivalent coarse grained structures, such that the coarse grained model will behave in a similar way to the fully atomistic structure, despite it containing fewer "physical" particles.

After the generation of a fully atomistic molecular dynamics trajectory, average bond lengths, angle parameters, and dihedral values are calculated from the centre of mass of constituent particles for each proposed CG bead. For example, if 2 carbon and 2 oxygen atoms are known to make up a single bead in the CG forcefield, their combined centre of mass is calculated, and used to define the location of the particle in the CG model. In order to calculate values for distances and angles, the distance/angles between combined centre of mass values are used. These values are transferred into the unparametrized CG model, and an arbitrary force constant applied. The CG system is set up, and run for a short (50 ns) unrestrained molecular dynamics simulation, and the same parameters are extracted and averaged again. We then have two different distributions for each value of interest - one from the AT simulation, and one from the CG simulation. Distributions of values from the AT simulation, and the novel CG system are compared, and force constants adjusted

in the CG model, with an aim to make the width of the two distribution histograms as similar as possible. The CG system is set up and run for a further 50 ns, and parameters reviewed and compared back to the AT results again. This process is iterated over until the two distributions are satisfactorily similar for all parameters of interest. Once this process is completed for all parameters, the full molecule is set up and simulated for a full 500 ns, before parameters are compared for a final time, to confirm that the parametrization is successful, and to examine emergent properties, namely distances or angle distributions that have not been explicitly included in the CG forcefield, but are a result of the correct inclusion of others, that may confirm that the model is successful at replicating the behaviour of the fully atomistic one.

A schematic for the parametrization methodology can be seen in Figure 5.7.

5.2.4 Simulations of Membrane Systems

For the setup of membrane systems, lipid molecules are placed randomly in a box, the box is solvated, and counterions added before the system is energy minimized, and molecular dynamics performed until a bilayer "slab" has formed. Once a bilayer has formed, the system is simulated for a further 500 or 1000 ns.

5.3 Results and Discussion

5.3.1 Simulating single Carbohydrates

Mapping Schemes

Single sugar molecules were initially built based on parametrization done by developers of the MARTINI forcefield. Most simple monosaccharides such as glucose, galactose, and fructose, have 12 non-hydrogen particles, following the principles of the CG methodology,

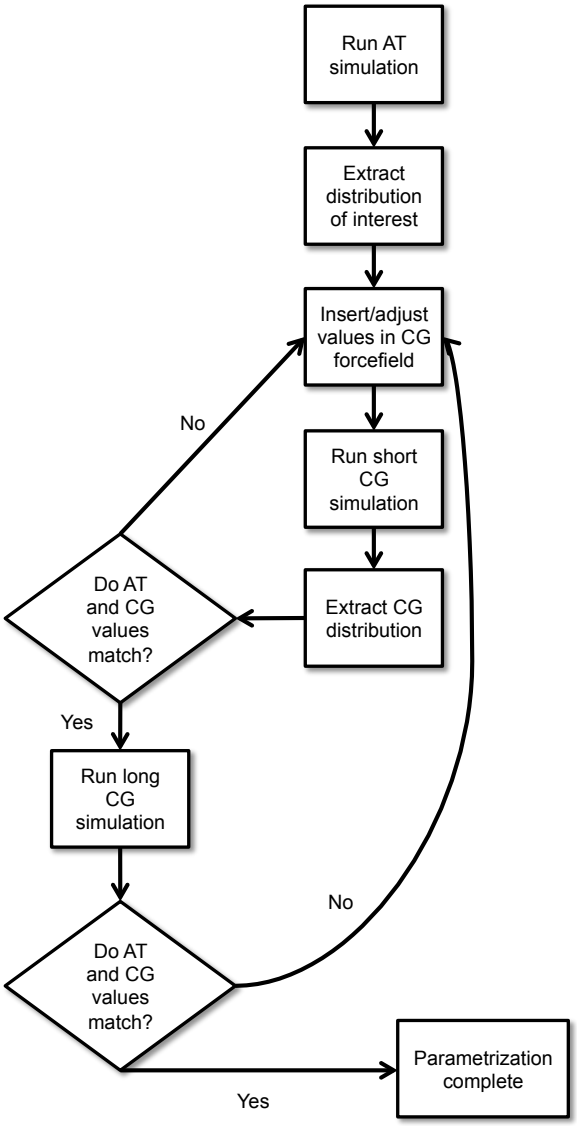


Fig. 5.7 Schematic showing the decision flow involved in parametrization of coarse grained carobhydrate stuctures.

this results in 3 CG particles per monosaccharide. Particle types for simple monosaccharides were based on those from the Sweet MARTINI forcefield. For glucose, polar (P1) particles are chosen, representing the polar nature of the hydroxyl groups attached, with a less polar particle (P4) representing the 5 and 6 positions of the ring, where carbon 6 is free from the ring, figure 5.8, A.

For more complex sugar molecules, functional group coarse grained beads were chosen

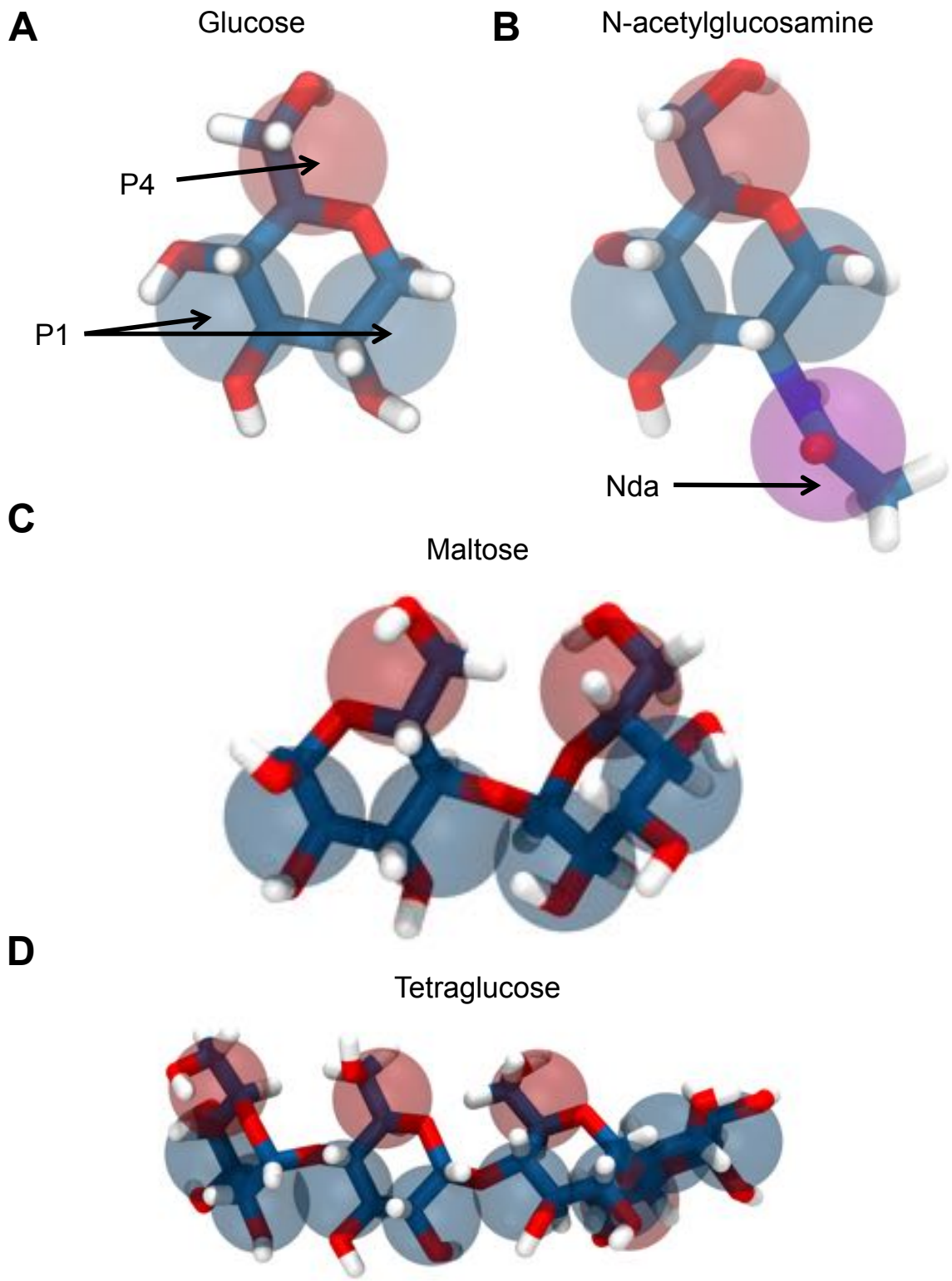


Fig. 5.8 Mapping Schemes for various carohydrate molecules, showing CG particles overlayed over fully atomistic structures. P4 particles are represented as blue beads, P1 as Red, and side chains of N-acetylglucosamine (Nda) as pink. Shown are: A) Glucose. B) N-acetylglucosamine. C) Maltose. D) Tetraglucose.

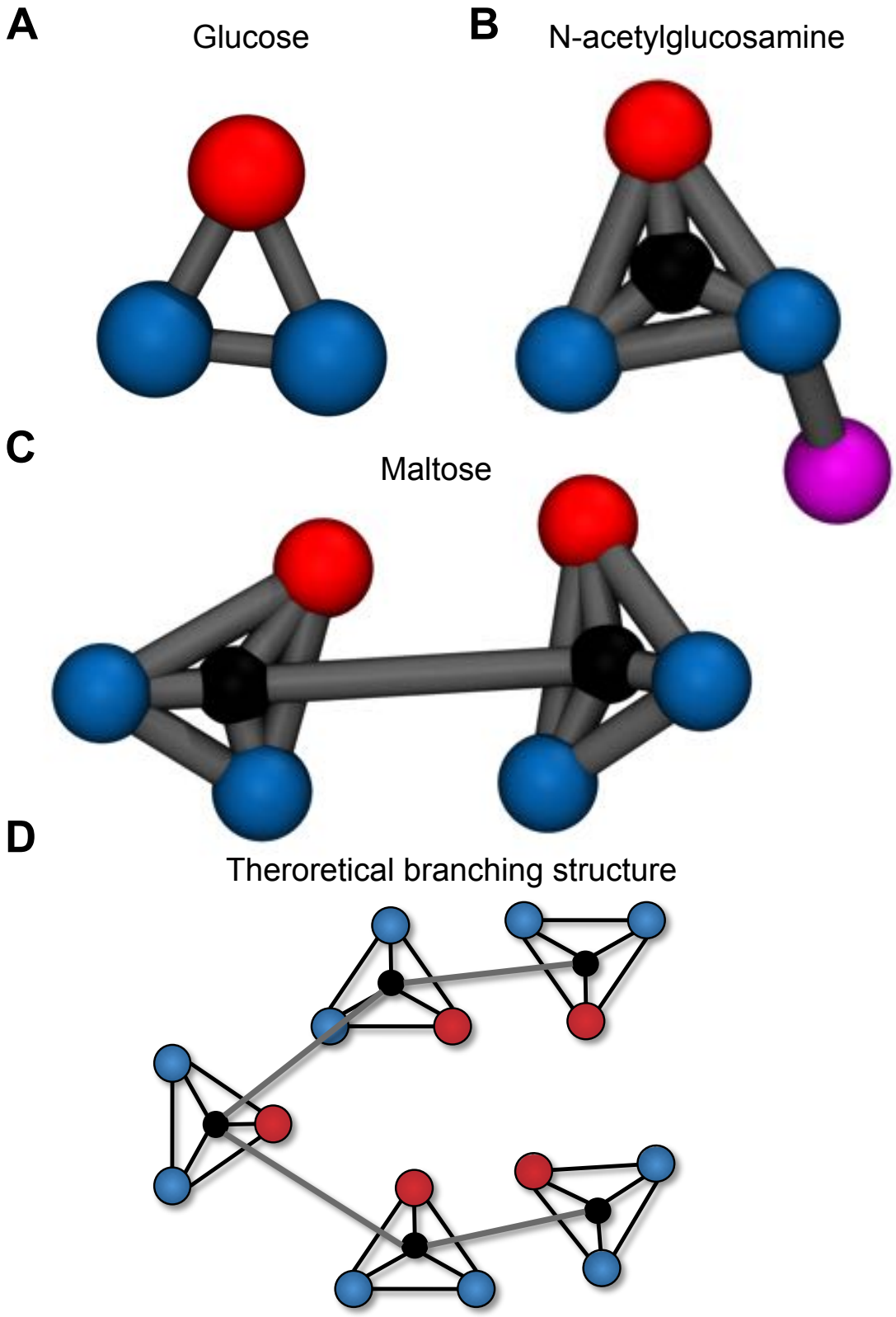


Fig. 5.9 Bonding Schemes for various oligosaccharide molecules, each ring is represented as a triangle, with sidechains added via bonds to the closest ring particle. For multimers, bonding is applied through "rotational node particles" (black), which have no representation on the interaction table, and no VDW radius, and are therefore functional only as a tethering point for bonds. The mapping scheme for maltose is shown in part C), and a theoretical branching sugar is shown in part D).

by comparison to already parametrized components of the MARTINI forcefield – notably amino acid parameters. Sidechain particles are then attached to the central triangle through the CG particle that contains the carbon they are attached too, figure 5.9, B.

Parametrization

Parameters for nonbonded interactions were calculated by comparison to atomistic simulations. Atomistic simulations were run using the GLYCAM forcefield in the AMBER simulation package. Individual sugar molecules were solvated and simulated for 200 ns in water, parameters were then extracted by analysing the center of mass of non-hydrogen atoms that would make up coarse-grained particles. Distance, angle, and dihedral parameters were extracted, and applied to the coarse-grained model. The coarse-grained model parameters were then iteratively matched to the atomistic model, figure 5.10, A. Distributions of the parameter in question were extracted from the atomistic simulation, then the same extracted from a short, 50 ns coarse-grained simulation. Parameters were compared, and the coarse-grained parameters adjusted to better represent the atomistic distribution, this was repeated until the parameters are satisfactorily matched. The number of iterations required varies greatly depending on the type of distribution investigated. Most distance values were matched to within 0.1 Å over 5 or less iterations, the angle distributions were slightly harder to fit within a few degrees, but were generally matched to within 5 deg of each other within 20 iterations. For each side chain atom, two angle parameters were chosen, one from each of the corner particles furthest from the corner that the sidechain is attached too, Figure 5.10, B. Note that for distance distributions around the central ring, distance d3 does not perfectly match between atomistic and coarse-grained simulations, this is due to the greater structural variability available to the fully atomistic molecule. In order to account for this, the average value for the distribution was taken, and the distribution matched to the highest peak in the atomistic distribution.

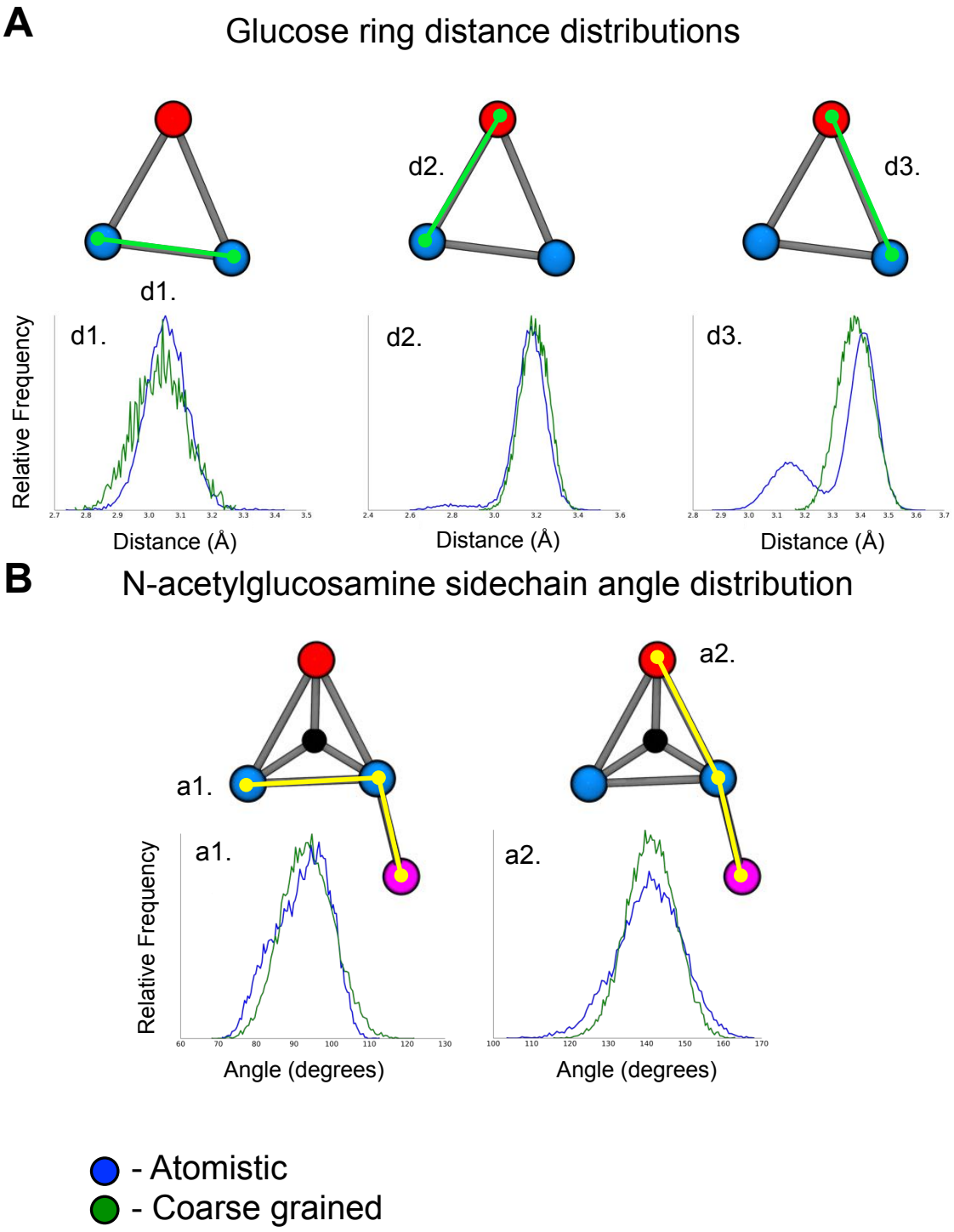


Fig. 5.10 Parameterization examples for some simple carbohydrates. A) Glucose distance parameters, showing each distance parameterized, and the atomistic (Blue), and coarse grained (Green) distributions for each distance. Notice that for certain distances, a perfect match cannot be made, this is likely due to the greater conformational flexibility afforded the atomistic structure, that cannot physically be included in the triangle structure B) N-acetylglucosamine sidechain angle parameterization, showing two angles that are restrained , and their distributions in atomistic (Blue), and coarse grained (Green).

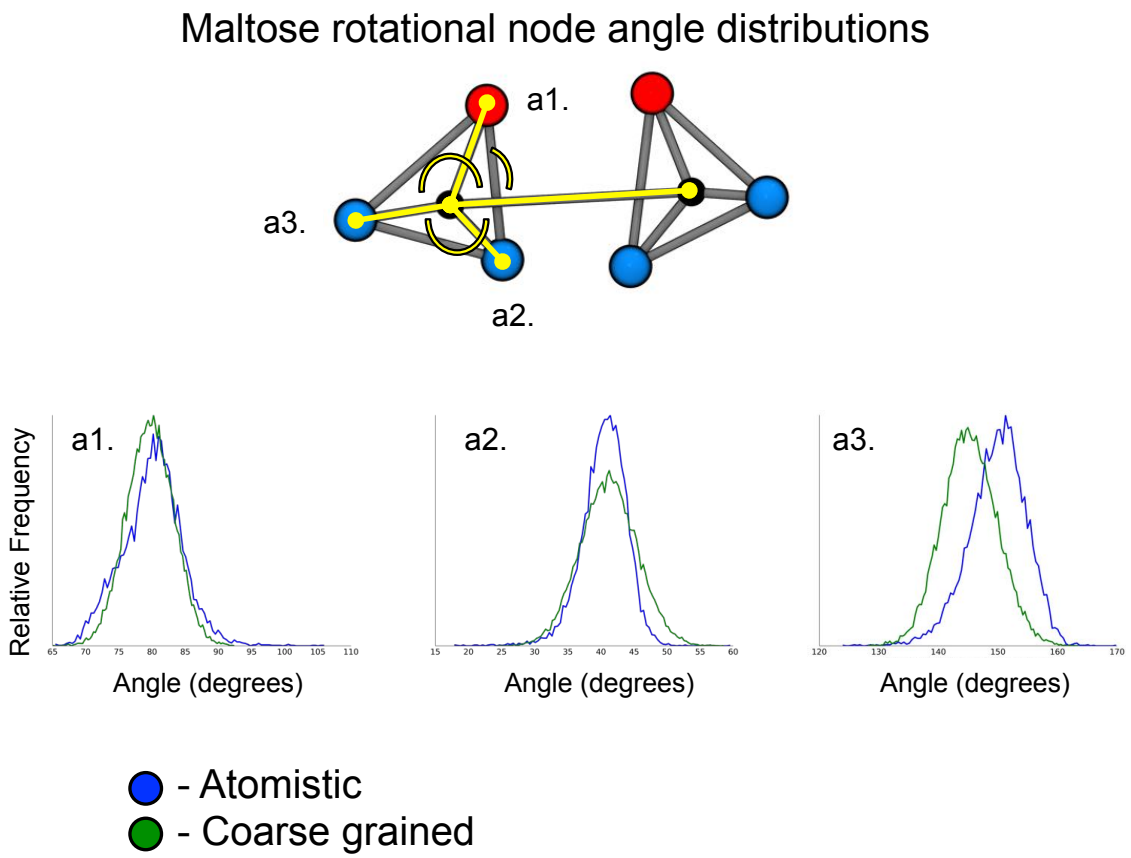


Fig. 5.11 More advanced parameters controlled for simple carbohydrates, showing the example Maltose. 3 angle terms are controlled across the glycosydic linkage to tailor the angle of each ring with respect to the next, a1, a2, and a3, shown from atomistic (blue), and coarse-grained (green) simulations. Note that not all parameters can be perfectly replicated, and the angle a3, for example, cannot be perfectly matched at the same time as having a1 and a2 matched.

5.3.2 Simulating Disaccharides

Mapping Schemes

For simulating disaccharides, two sets of parameters for monosaccharides are combined. The linkage between the two molecules is more complex. Initially we thought to simply link the two triangles, replicating the glycosidic bond , through the closest two B particles. However, once the structure is extended into more complex carbohydrates, this representation would not be efficient, as huge strength constraints would be required in order to keep the rings in their biological position, this leads to large instabilities and inaccuracies within the structure. To overcome this, the concept of rotational nodes was introduced. A new particle type was generated, with no VDW interactions with any other particle, and with no non-bonded interactions present on the interaction table, and a very low mass, a third of which was then deducted from each of the 3 central ring particles, in order to keep the mass of the ring as a whole the same. This particle can then be used as a rotational node, to link structures together.

Firstly, for each carbohydrate ring, a single rotational node particle was placed in the geometric center, the centroid, and held in location using constraints, so that there is no opportunity for it to move during the simulation. The rotational nodes can then be bonded together to form pseudo glycosidic linkages, and allow the twisting and rotation of each sugar with respect to each other.

Parametrization

In order to hold the central rotational node within the middle of each triangle (as even with constraints, the particle could fall out via the extension or contraction of the harmonic restraints between ring particles), an improper dihedral function was applied to the three ring particles, and the central rotational node, in order to keep all 4 particles in the same plane.

Parametrization of the linkage between two sugar molecules was performed in the same iterative manner as the distances for ring particles. The distance between the two molecules was parametrized, and 3 angle parameters were chosen for each ring, in order to hold it in the correct position. The angle parameters chosen are from each outer ring particle, through the central rotational node, and to the rotational node of the connected sugar ring, figure 5.11. Angle restraints across the central particles were much more difficult to iteratively match than those used for functional group restraints, as three angle parameters need to be synchronized. Upwards of 100 iterations were performed until a distribution match of within 5 deg was achieved, where possible. The use of 3 angles for each carbohydrate allows the adjustment of the exact orientation of each ring with respect to the other, and allows each ring to be tailored individually, keeping the methodology modular, for future extension. A further weak dihedral restraint was chosen to apply across the glycosidic linkage, in order to remove situations where the structure has two possible "relaxed" conformations. Note that for the example of Maltose - the angle d3 (Figure 5.11, A), does not completely match between atomistic, and coarse-grained simulations. This is likely due to structural differences between the rigid triangle structure, and the six membered ring that makes up the atomistic structure, whilst the coarse-grained structure is constrained to a single conformation due to geometry. In this case, matching all 3 angles perfectly is impossible due to physical constraints, and so, the first two angles (d1 and d2) were matched as closely as possible, and the third angle, d3, was matched such that the distribution had a similar shape, albeit slightly adjusted mean in order to stabilize the structure. The distance between the two rings was also matched between the two rings.

The nrexclusions value in the forcefield (defining how many covalent bonds along VDW interactions in the forcefield should be ignored for) was also changed for any structure containing more than a single triangle particle. In this case, the value was changed to a value of 4, such that particles 4 bonds away from each other are unable to interact with each other

through non-bonded interactions. This allows a ring to sit next to another group of particles that it is directly bonded too (such as another carbohydrate ring), without non-bonded interactions interfering with the structure.

In order to test the parametrization, the molecule maltose was chosen for conversion to CG. Both AT and CG simulations were performed of the molecule in solvent, the AT simulation was performed for 200 ns, and the CG simulation was performed for 500 ns. Distances, angles, and dihedrals were then compared. Whilst the angle of ring with respect to the other is explicitly described in the forcefield, the distance between ring particles in the structure were not - the distance between B1 particles of each ring was compared, along with the distance between the rotational nodes of each ring (which was explicitly described), (Figure 5.12)

5.3.3 More complex Carbohydrates

For the generation of more advanced carbohydrates, the same principles were applied as within disaccharides, whereby rings were chained together with the use of rotational nodes. As the forcefield was kept modular (i.e a GLU-GAL linkage is standalone, and will not be affected by the addition of an extra GLU on either end), the parameters for a more complex model are able to be built up from simple building blocks, and parameters can be reused for any structure that contains that particular set of sugars, with the same linkage. Branching can be accounted for simply with the addition of a second pseudo glycosidic bond between rotational nodes, as shown in Figure 5.9, D. An example of a larger, more complex carbohydrate is tetraglucose. The structure for tetraglucose was built up by re-employing the maltose parameters, multiple times, onto the pre-built tetraglucose scaffold. As the forcefield is modular, this should effectively replicate the general behaviour of the full molecule. In order to asses the parametrization of the CG model, simulations of the structure in AT, and using the new CG model were performed. The AT model was simulated for 200

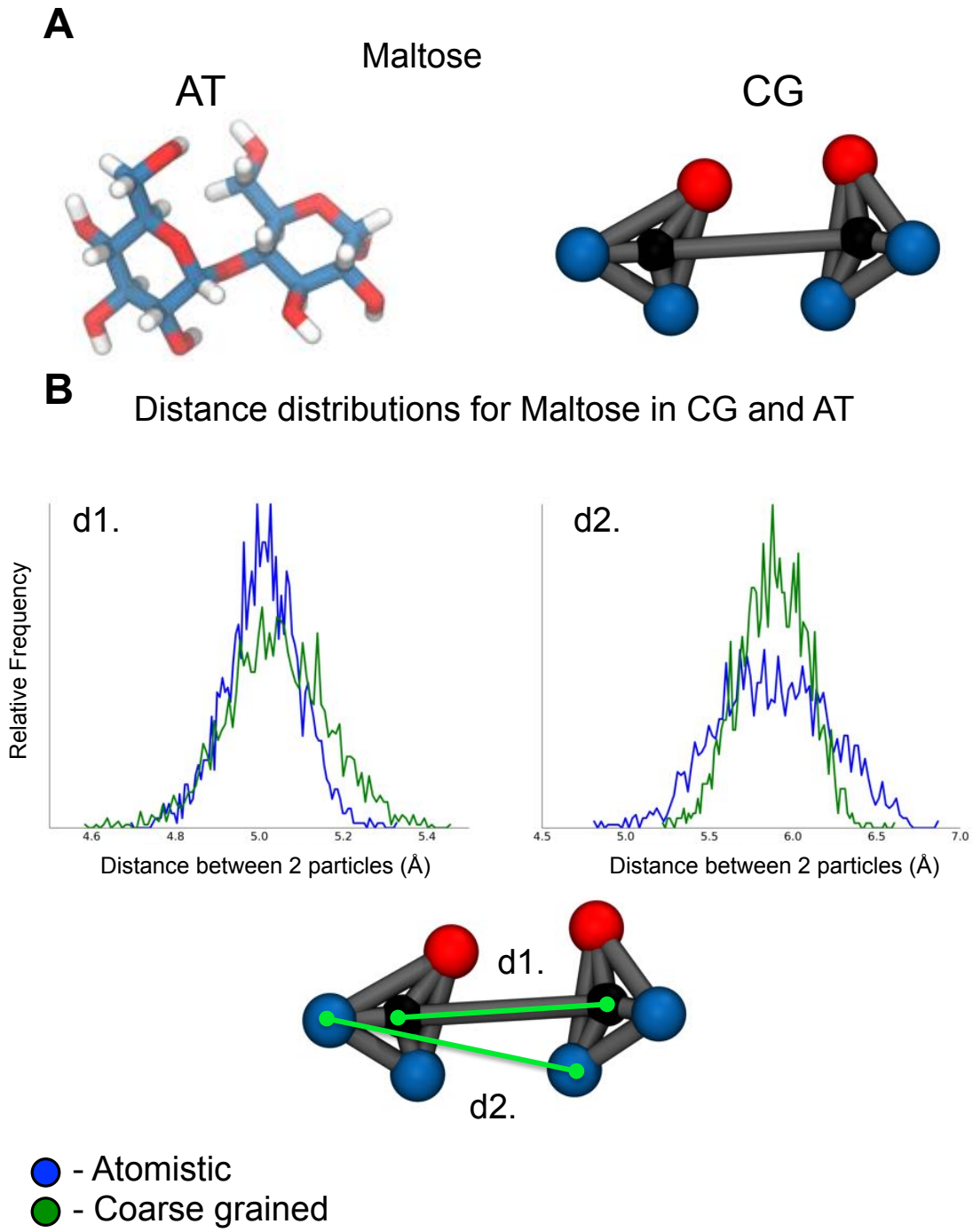


Fig. 5.12 The parameterization result for Maltose, showing atomistic and coarse grained structures, both adopting a similar rotation to each other (Explicitly controlled in the force-field). Also shown are the distributions for the distance between each rotational node particle (d1.), and the distance between B1 particles (d2. an emergent property), from atomistic (Blue), and coarse grained (Green) simulations.

ns, and the CG model for 500 ns, both solvated, and parameters extracted for comparison. In particular, the distance between the extreme ends of the molecules was extracted – as this was an emergent property, and was not explicitly included in the forcefield. The two values from the AT and CG representations match well, especially considering the distance is the emergent property of numerous angle and distance parameters (Figure 5.13). This suggests that the modular nature of the forcefield is effective in capturing the behaviour of larger molecules. It can also be seen from comparison of structures within the simulations, that CG tetraglucose structure adopts a correct helix-like structure, as expected (Figure 5.13, A).

5.3.4 Applications of Carbohydrates - Glycolipids

As a result of the effective development of carbohydrate parameters, selected glycolipid models, of differing complexity, were generated. Glycolipids are a good test case for this methodology, as the coarse-grained methodology has been shown previously to accurately represent the properties of other lipid species. They are also comparatively simple structurally, and so allow us to take a small step towards large, complex structures, with resultant structures that are of use for investigations of complex membrane systems. All of the glycolipids generated are sphingolipids, they have a similar backbone and tail region, with a differing carbohydrate headgroup. The lipids generated are as follows:

- Gluco and Galactocerebroside.
- GM3 (monosialodihexosylganglioside).
- Forssman glycolipid.

Resultant coarse grained structures, alongside their atomistic structures can be seen in Figure 5.14.

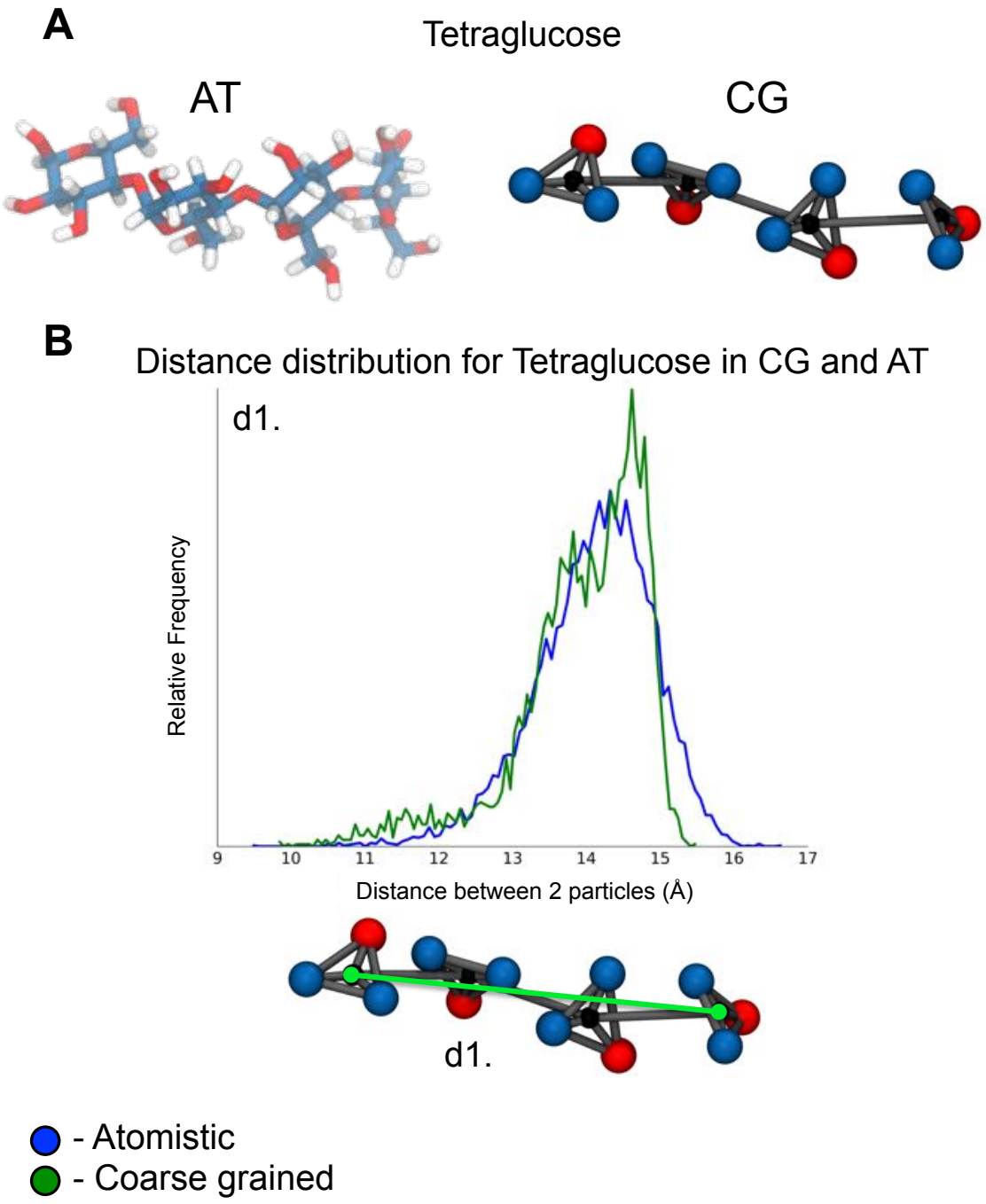


Fig. 5.13 The parameterization result for tetraglucose, showing atomistic and coarse grained structures, both adopting a stereotypical helix-like shape. Also shown are the distributions for the end-end distance (rings 1-4) between two rotational node locations shown from atomistic (Blue), and coarse grained (Green) simulations. The histogram distribution matches between the two well, the width of the distribution (representing the maximum and minimum distance in each simulation) are in good agreement. As this parameter is emergent, and a result of multiple different angle and distance parameters, it is a good indication that the CG model is accurate at reproducing the general structural behaviour of the atomistic counterpart.

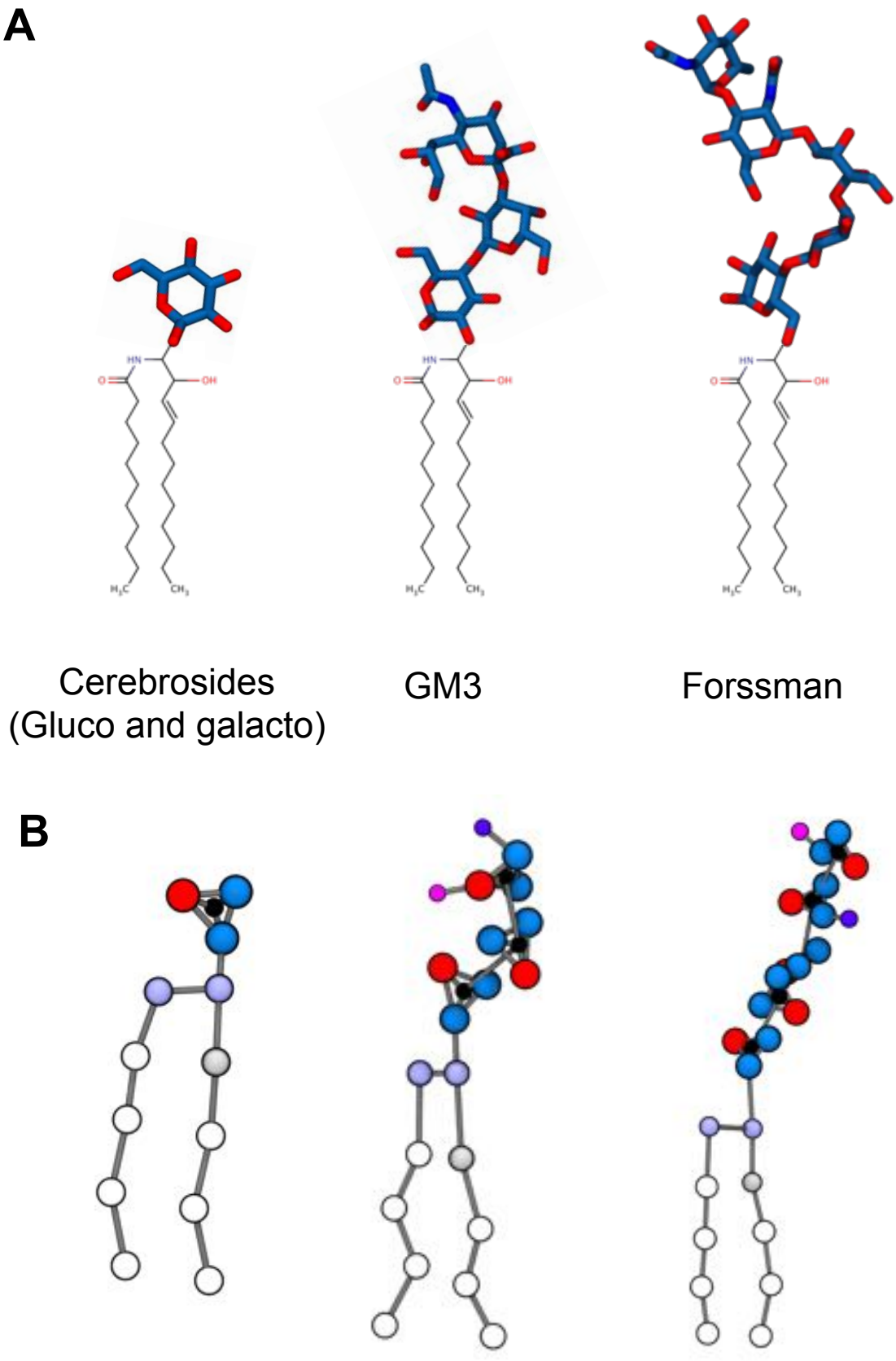


Fig. 5.14 A comparison of structures for all 3 glycolipids parameterized in A) atomistic detail, B) Coarse grained detail.

5.3.5 Glucocerebroside and Galactocerebroside

The lipids glucocerebroside and galactocerebroside are essentially the same in coarse grain – both involving a single carbohydrate, either glucose or galactose, attached to a ceramide molecule. As glucose and galactose are sterioisomers of each other, the CG structure will have very small changes between each lipid, and essentially no difference in the parameters making up the linkage to the ceramide tails. Ceramide parameters were taken from the already existing CG parameters for sphingomyelin from the MARTINI lipid forcefield. The current MARTINI parameters for sphingomyelin include a slight error in the numbering scheme for the topology, resulting in the headgroup for the molecule being attached to the AM2 bead, which has a saturated tail, rather than the AM1 bead, which contains the unsaturated lipid tail, which is connected to the lipid head in all other lipid structures. Whilst this error would not matter if parameters were adjusted accordingly to ensure the headgroup maintains the correct orientation, in this case, parameters are not adjusted, and the angle parameters for the lipid backbone and tails are thus incorrect, leading to a "curling up" effect on the molecule in simulations. This error has been reported to the MARTINI group and will be fixed in the next forcefield iteration. For our purposes, we chose to connect the nearest headgroup particle to the AM1 backbone particle, resulting in expected behaviour.

Both glucose and galactose monosaccharides were parametrized independently, and were represented as triangles, with a central rotational node particle. Although the triangle representing the headgroup was attached via the closest ring particle, rather than the central node particle, the node particle was left in place in the molecule, as the ring structure was parametrized with this present. This also preserves the modular nature of the forcefield. In order to accurately match the parameters for the linkage between the headgroup and tails, AT simulations of the gluco and galactocerebroside lipids were performed. A single lipid was placed in a box and solvated, and the CG equivalent parameters extracted for 3 repeats of 200 ns simulations. Details of the bond length were extracted, along with 3 angle param-

eters that would effectively hold the headgroup in a physiological position similar to the AT simulations (Figure 5.15).

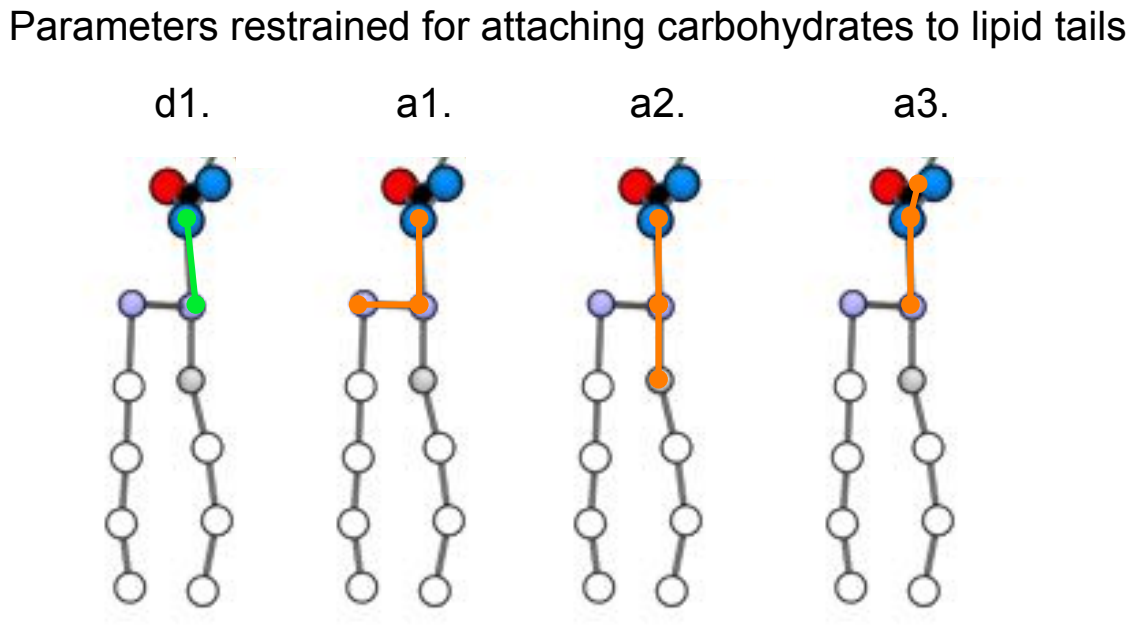


Fig. 5.15 Distance (Green) and angle (Orange) parameters restrained for representing the glycosidic linkage.

In order to test the parameters of the molecule, distances and angles of the carbohydrate rings with respect to lipid structure were compared between the AT and CG representations of the glycolipid. The CG model successfully retains the overall structure and characteristics of the AT model.

5.3.6 GM3

The molecule GM3 is slightly more complex than that of gluco or galactocerebroside, involving a headgroup consisting of three different sugar molecules – glucose (GLU), galactose (GAL), and n-acetylneuraminic acid (Neu5Ac). Using standard nomenclature, the molecule is represented as:

$$\text{Neu5Ac} - \alpha(2-3) - \text{Gal} - \beta(1-4) - \text{Glc} - \beta(1-1) - \text{Cer}$$

Each monosaccharide was parametrized individually, including sidechains for the NANA molecule. Once the individual sugar parameters were produced, a simulation was performed of a single solvated GM3 molecule, in atomistic detail, for 200 ns. AT parameters for the linkage between the headgroup and the tails, and between each pair of sugar molecules, were extracted. The CG model was iteratively matched to its AT equivalent as done with previous models, before the coarse grained model was simulated in a lipid environment.

Self assembly simulations were set up in order to assess the partitioning, stability, and diffusion properties of the lipids. By placing 50 CG GM3 molecules in a bilayer with 550 CG POPC molecules, we were able to look at the emergent properties over multiple molecules. Comparing the length and angles of the headgroup sugars (Figure 5.16), and the angle of the headgroup with respect to the lipid tails, between the atomistic simulations of a single AT lipid in water, and CG simulations of the bilayer system, we can see that the CG simulations still replicate the atomistic values. Comparing the partitioning density of GM3 to those of other lipids in bilayer simulations it is clear that GM3 behaves in a similar manner to other lipids in terms of its lateral location in the membrane. Calculating the radial distribution function for GM3, it also appears to diffuse at a much slower rate than other lipids in the system. This is potentially due to clustering of the lipids within the membrane, though further work, such as cross correlation studies, or a simulation of a single GM3 in a POPC bilayer would be required to prove this. Work performed by Dr Heidi Koldsoe using the GM3 model (Koldsøet al. [103]), has proven that the CG model for GM3 self associates in biological membrane system. Bilayer properties are shown in figure 5.17. The validated, and complete model was used in large membrane systems by Dr Heidi Koldsoe (Koldsøet al. [103]), and Dr Khairul Abd Halim.

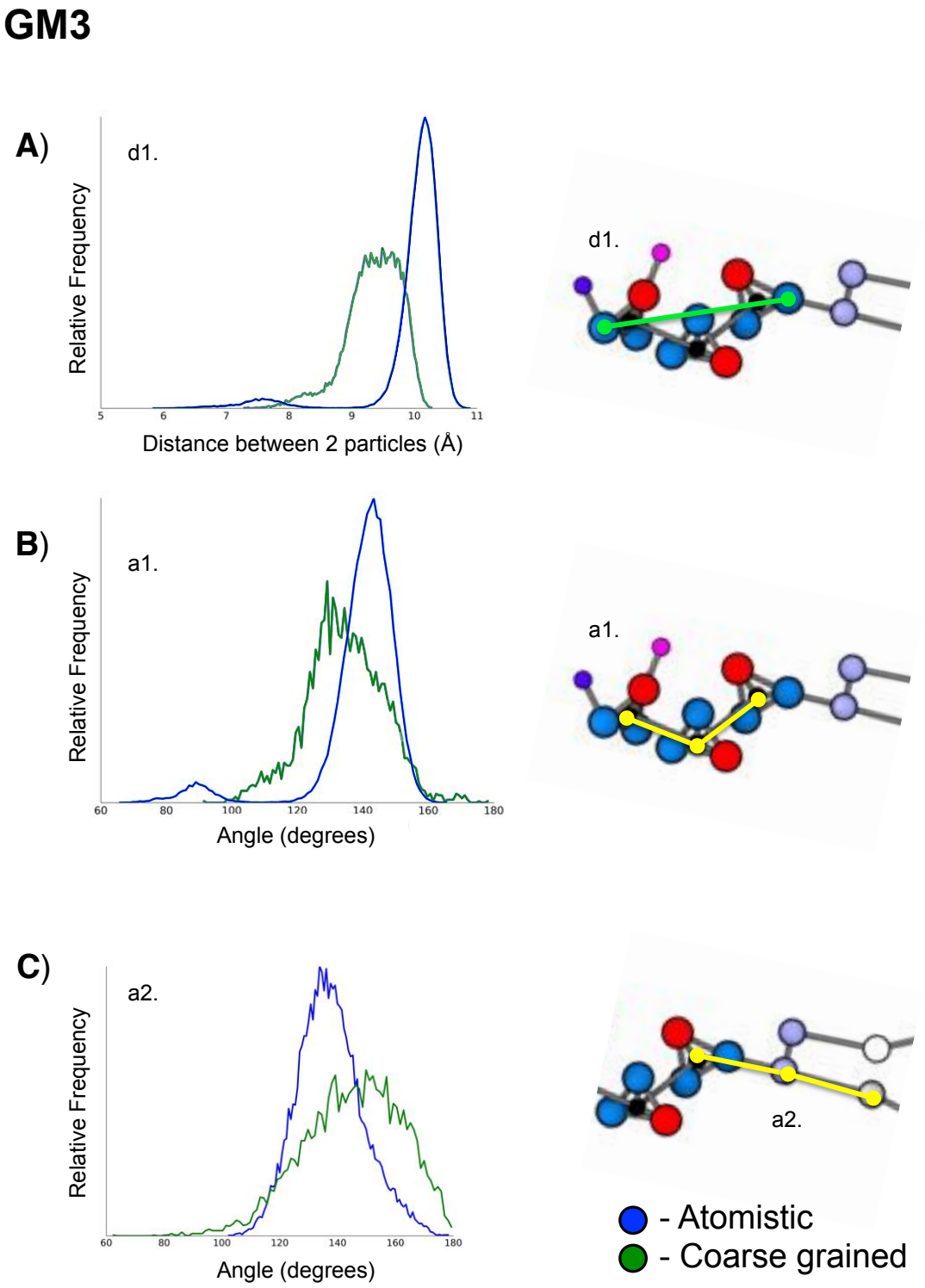


Fig. 5.16 Emergent properties for GM3 molecules in atomisitic simulations (Blue), and averaged over GM3 in a coarse grained bilayer (Green). A single molecule of GM3 was simulated in isolation, rather than in a bilayer environment, which could account for slight differences when the CG model is simulated with other lipids.

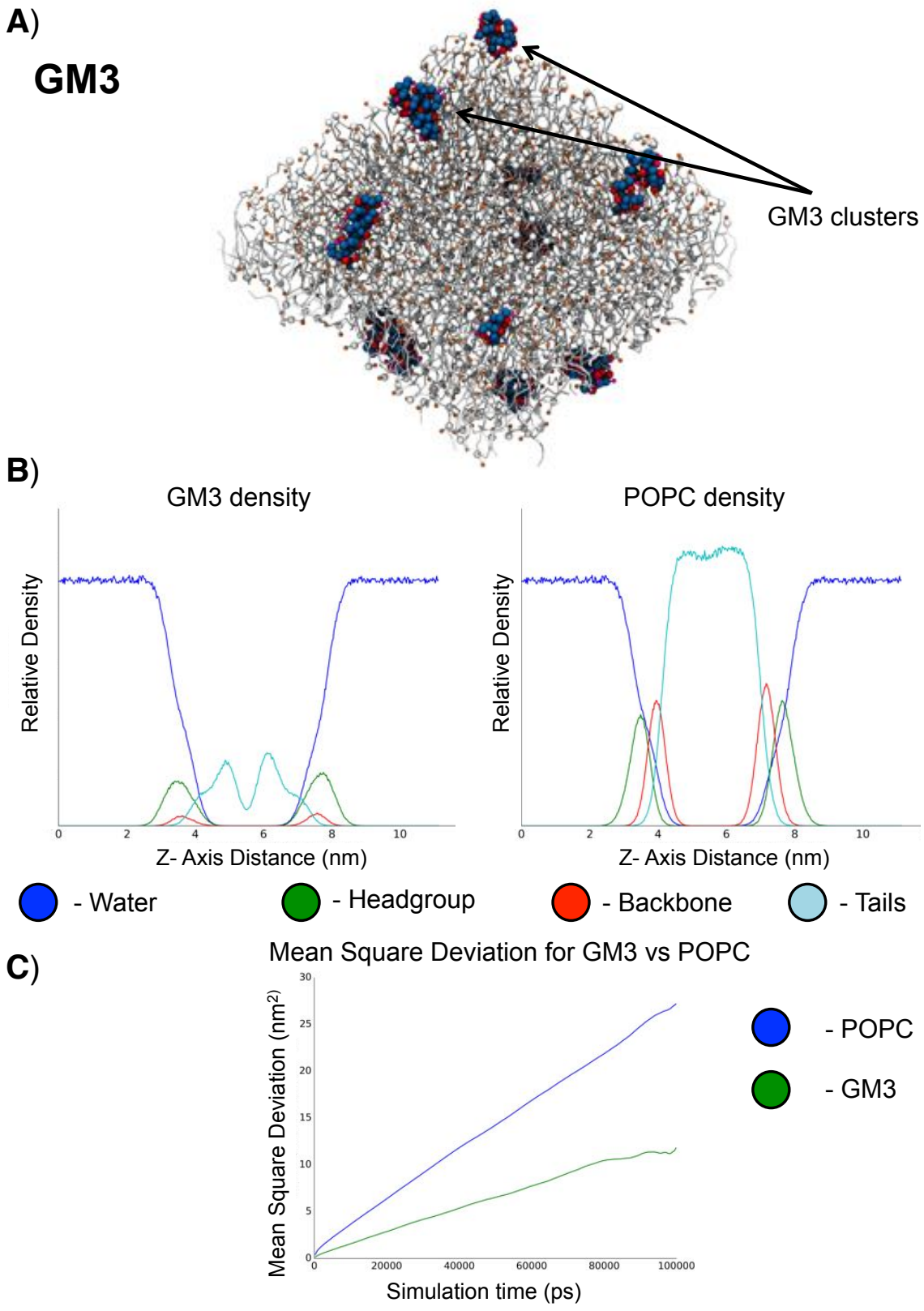


Fig. 5.17 Properties of the CG GM3 model in a POPC bilayer showing A) A screenshot of the bilayer after dynamics containing 50 GM3 molecules (Blue). B) Distribution function for GM3 and POPC lipids in a bilayer, showing that both lipids partition similarly. C) Mean Square Deviation for GM3 and POPC in a bilayer, showing average displacement of each lipid type throughout the simulation. GM3 clearly moves less during the simulation, and this is a good indication that it may be self associating and thus diffusing less.

5.3.7 Forssman Glycolipid

The Forssman glycolipid is another complex glycolipid – commonly found in the viral membrane of the flu virus. The headgroup consists of 5 sugar molecules, which can be represented using the standard nomenclature as:

$$\text{GalNac}-\alpha(1-3)-\text{GalNac}-\beta(1-3)-\text{Gal}-\alpha(1-4)-\text{Gal}-\beta(1-4)-\text{Glc}-\beta(1-1)-\text{Cer}$$

All the individual components required for this lipid were already parametrized for GCCE, and GM3, and so were already present, the linkage from ceramide to carbohydrate headgroup was taken from the GM3 structure, and ceramide parameters were duplicated as done previously. In order to parametrize the linkages between sugar molecules, a simulation was performed of the full atomistic structure of the forssman glycolipid, in water, for 200 ns. Parameters were extracted as before, and iteratively matched back to the AT simulations. The CG system was then tested by self-assembly simulation similarly to GM3, with an unrestrained simulation of 30 lipids simulated in a POPC environment for 100 ns. Bilayer properties can be seen in Figure 5.19. Emergent properties were tested against those in AT simulations for validation of the model, this can be seen in Figure 5.18.

5.3.8 Discussion

In this chapter I have demonstrated a novel methodology for the CG representation of carbohydrates. Using a stringent, iterative method for parametrization, and a novel mapping scheme that allows important structural details to be fine tuned, we are able to produce carbohydrate containing models of physiological significance. By starting with the parametrization of small molecules, and building up to more complex ones, we are confident that the model can reproduce the behaviours of atomistic simulations. A validation of the model for larger systems comes from the examination of emergent properties in both CG

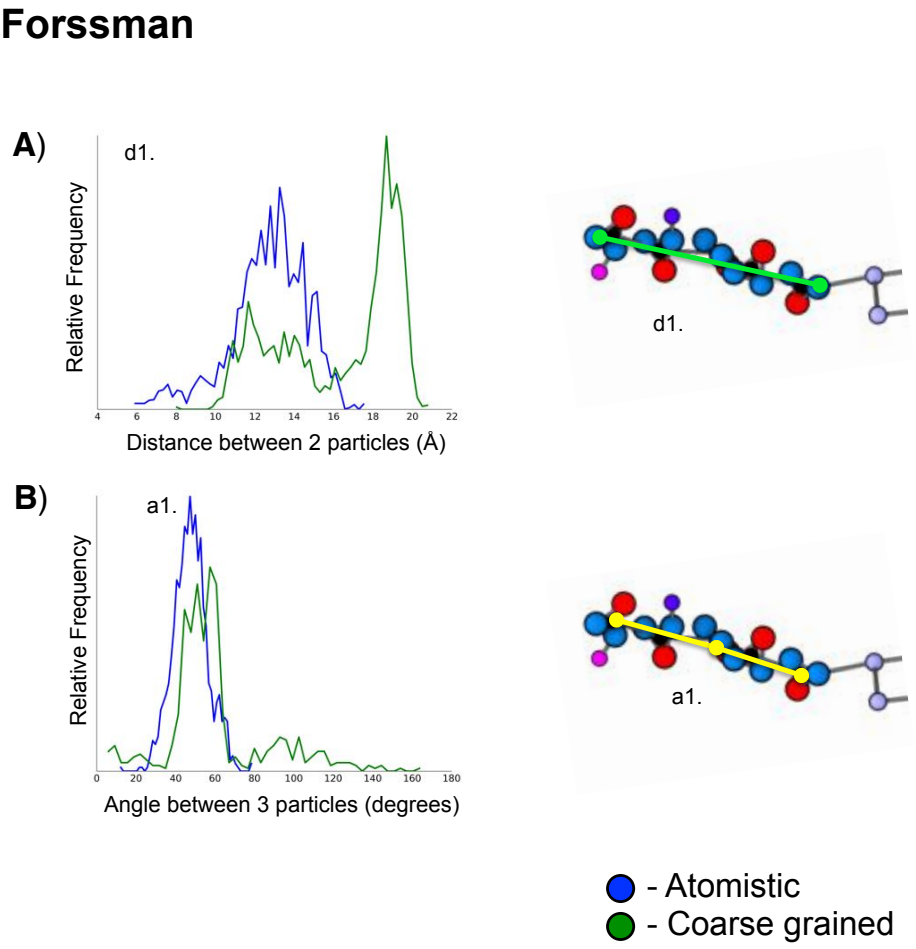


Fig. 5.18 Emergent properties for GM3 molecules in atomisitic simulations (Blue), and coarse grained simulations (Green). Showing comparisons between an angle and distance parameter that are not explicitly mentioned in the forcefield.

and AT, properties such as large scale distances, and angles that aren't explicitly included in the forcefield can be compared to confirm that the model is working as intended.

Disaccharides can be parametrized easily, and the system built up in a modular fashion, as has been proven with maltose and tetraglucose, whereby parameters from a disaccharide, multiplied into a larger model reproduce the correct behaviour of a tetrasaccharide, without ever having been parametrized explicitly from the AT tetraglucose structure.

Other results of the study include working models of the glycolipids, GCCE, GM3, and the Forssman glycolipid, which have all been tested in physiological bilayers with POPC.

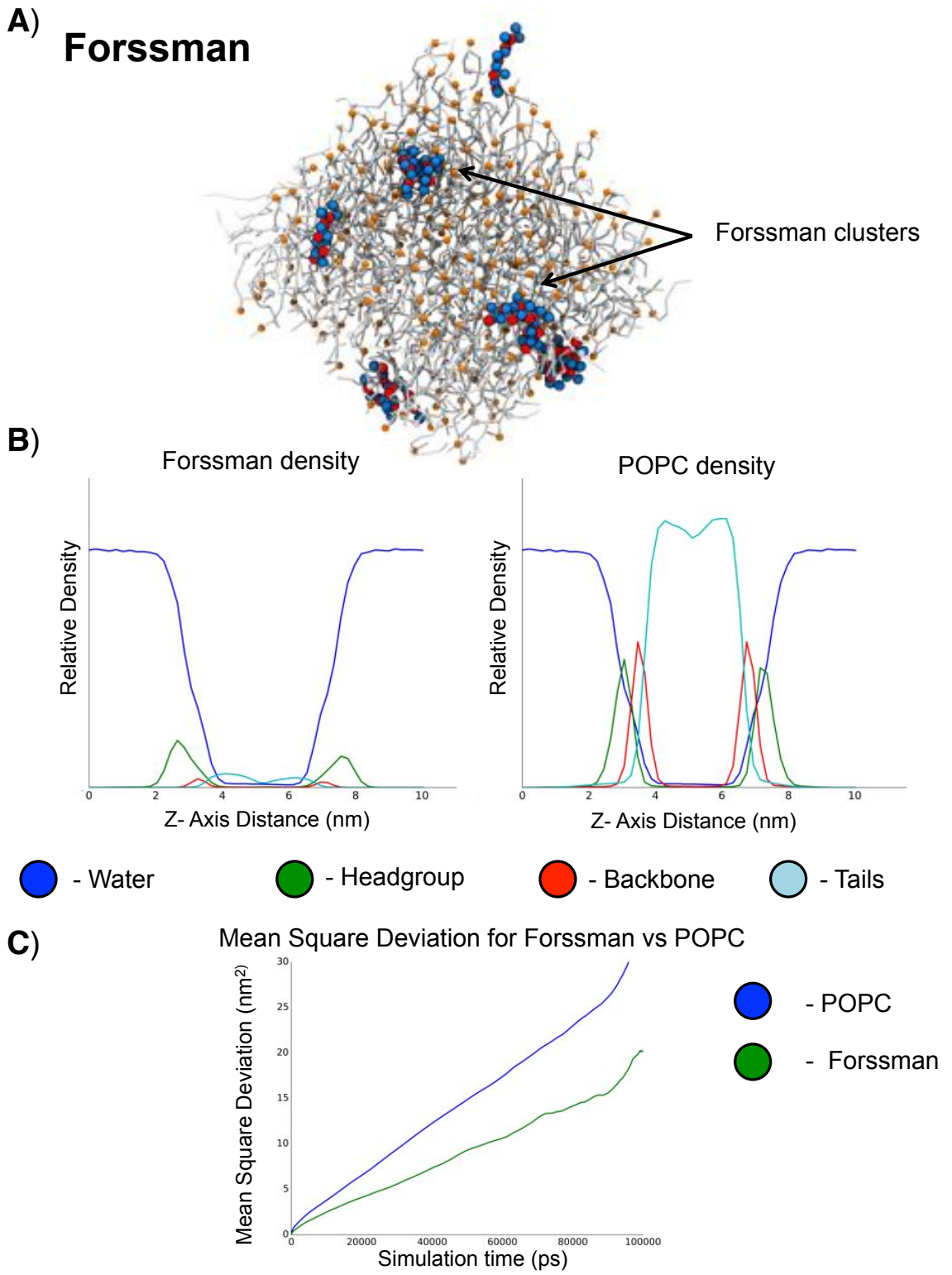


Fig. 5.19 Properties of the CG Forssman model in a POPC bilayer showing A) A screenshot of the bilayer containing 30 Forssman molecules after dynamics (Blue). B) Distribution function for Forssman and POPC lipids in a bilayer, showing that both lipids partition similarly. C) Mean Square Deviation for Forssman and POPC in a bilayer, showing average displacement of each lipid type throughout the simulation. The Forssman clearly moves less during the simulation, and this is a good indication that it may be self associating and thus diffusing less.

In the case of GM3 and the Forssman glycolipid, parameters have been used by others for published research, and in large, physiologically interesting systems. Figure 5.20.

Limitations of the method are that it is unable to capture minute structural details of carbohydrates. This is a general criticism of the CG methodology, as in instances such as glucose and galactose molecules being almost exactly the same in CG, we would not expect any simulation to distinguish between the two functionally. A way of tackling this is to allow conversion from CG structures back to atomistic details. This has been completed for the lipid GM3, whereby the molecule can be converted back to atomistic detail through the previously published CG2AT (Stansfeld and Sansom [179]).

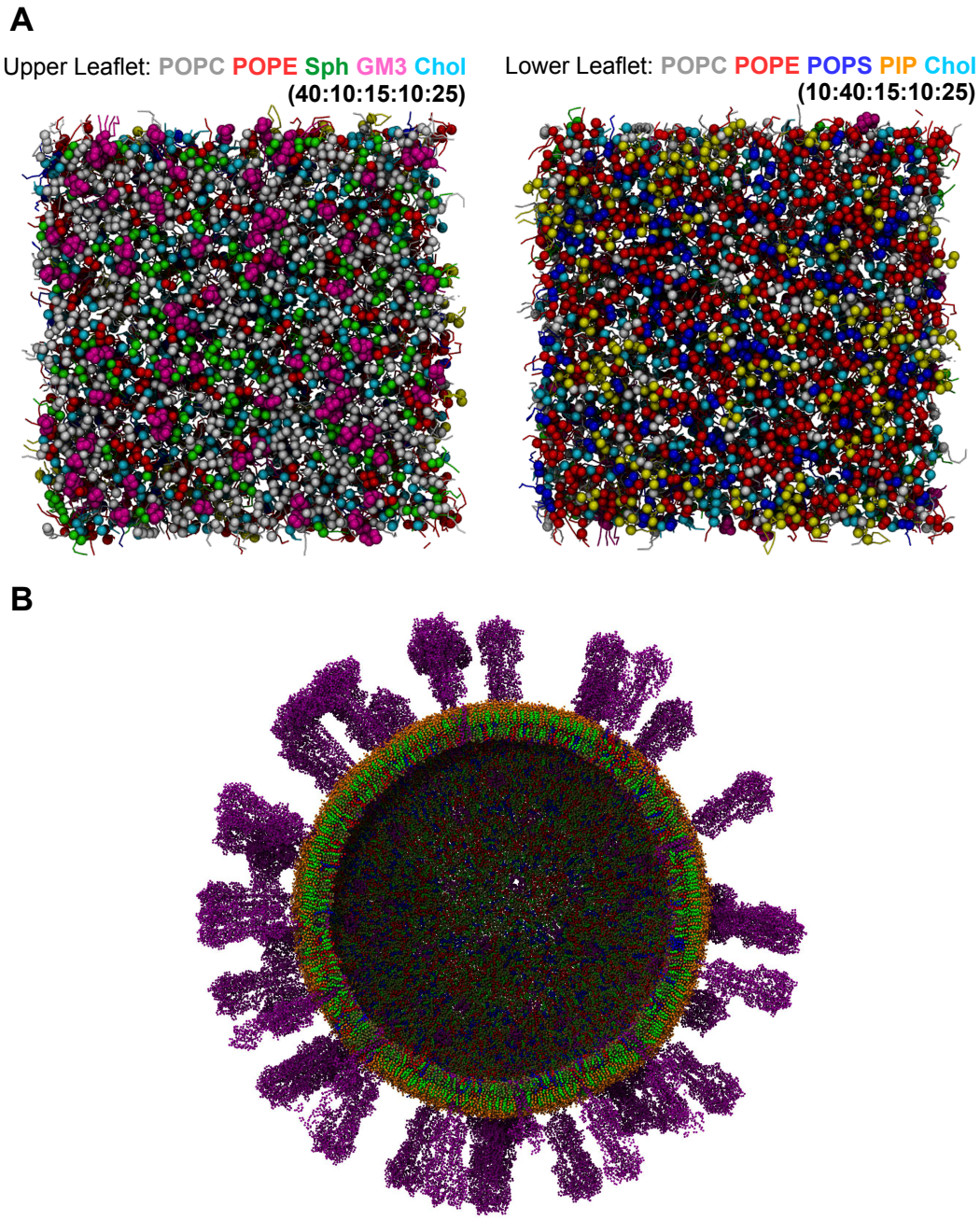


Fig. 5.20 Examples of other work performed using CG glycolipid parameters, showing A) GM3 in use as part of a mammalian red blood cell membrane, figure generated by Dr Heidi Koldsoe (Koldsøet al. [103]). B) The Forssman glycolipid in use as part of a Flu virion membrane, figure generated by Dr Tyler Reddys.

Chapter 6

Coarse Grained Simulations of Protein Glycosylation

"Your fair discourse hath been as sugar,
Making the hard way sweet and
delectable."

William Shakespeare

In this chapter, I present an extension of the previously developed methodology for modelling and simulating sugars to glycoproteins, with the goal of extending the glycosylation methodology to much more complex, branched, structures. The test case for this methodology is the immunoglobulin G (IgG) Fc fragment, as we have seen from previous work in Chapter 3, that that structure changes when an elastic network is applied without glycosylation. There is also some debate about how the glycosylation of the protein influences structure, we hope to be able to gain evidence for one mechanism over the other through comparison of glycosylated and unglycosylated coarse grained dynamics.

6.1 Introduction

6.1.1 IgG1 Fc Fragment Glycosylation

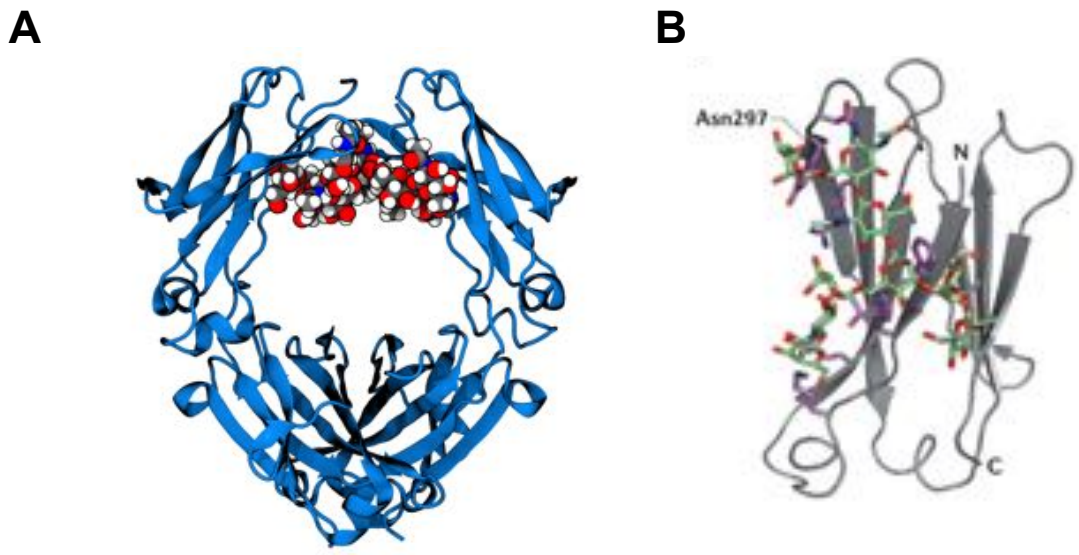


Fig. 6.1 Structures of glycosylation of IgG Fc, showing: A) Glycosylated Fc structure, including wild type glycosylation on ASN 297 residues, structure shown is PDB ID 3AVE. Carbohydrates are shown in spacefill representation. B) A close up of the immunoglobulin Fc glycosylation, adapted from Schwab and Nimmerjahn [164].

The glycosylation of the immunoglobulin G Fc fragment has seen a proportionately large amount of study compared to other glycoproteins. The carbohydrates attached to ASN 297 within the Fc "horseshoe" are known to play a large role in maintenance of structure, and by extension, are known to be intrinsically important to interactions between the immunoglobulin and the FcγR receptor (Arnold et al. [5], Krapp et al. [106]). The carbohydrates are thought to be somewhat dynamic, exploring the empty space inside the centre of the horseshoe shaped protein, but evidence also suggests that they sit tightly to the inner side of the protein, and can help to stabilize the structure (Meier and Duus [126]). The carbohydrates within the Fc of IgG1 are also known as targets for certain lectins within macrophages (Dong et al. [47]). It is also known that small differences in carbohydrate sequence can have

large effects on the physiological activity of the molecule (Mizushima et al. [130]). Due to the pharmacological relevance of immunoglobulins, particularly IgG1, and in particular, the attractiveness of being potentially able to modulate effector function levels has driven a greater understanding of how glycosylation influences antibody function (Dalziel et al. [37]). An example of IgG glycosylation can be seen in figure 6.1, A.

It is known that by removal of glycan molecules from within the Fc structure, that the immunoglobulin affinity for it's FcγR receptor is greatly reduced. There are two possible mechanisms by which this removal changes interactions with the receptor:

- Collapse - In this mechanism, the removal of oligosaccharides from the centre of the Fc protein results in a closing of the space between the two extended regions of the molecule. This collapse of the protein structure is heavily supported by crystal structures (Crispin et al. [36], Feige et al. [57], Krapp et al. [106]), though there are suggestions and debate about whether this is simply one possible conformation, or a physiologically stable state (Sondermann et al. [174, 175])

- Flexibility - This hypothesis proposes that the removal of the glycan structure from the Fc will destabilize the receptor extended regions, resulting in large conformational fluctuations. As a result, it will be more difficult for the FcγR to interact successfully with the structure. This hypothesis is supported by Förster Resonance Energy Transfer (FRET) (Kelliher et al. [98]), small-angle X-ray scattering (SAXS) (Borrok et al. [17]), and through mutational studies (Jung et al. [94]).

6.1.2 Previous Computation Studies of Glycosylation

Glycosylation of proteins in silico is not a heavily researched topic, sugars being complex molecules to model and characterize. Despite this fact, a simulation of a fully glycosylated antibody has been performed (Kayser et al. [97]), though the full structure was simulated for

only 30 ns, barely enough time to observe any conformational change, a large limitation of traditional atomistic molecular dynamics. Since this simulation, much improved parameters for carbohydrates, in the form of the GLYCAM forcefield, have been developed (Kirschner et al. [102]). Further simulations have also been done on the glycosylated Fc protein, indicating that carbohydrates have an influence on the aggregation propensity of the Fc molecule (Voynov et al. [194]), though these simulations again, were not performed for a long enough time scale to investigate conformational change of the molecule. There have also been some simulations of other glycoproteins, with carbohydrate molecules explicitly included (Frank and Schloissnig [60]), but these are not common.

Although within the MARTINI forcefield parameters exist that theoretically allow the glycosylation of proteins, no coarse-grained simulations containing a glycosylated protein have yet been reported.

6.1.3 Coarse Grained Secondary Structure Restraints

Coarse grained simulations may be an avenue with which to gain more information about the dynamics of the unglycosylated, and glycosylated Fc, as we are able to investigate time-scales upon which the conformational changes that the protein undertakes occur.

For the purposes of this study, different secondary structure restraint mechanisms can be applied to antibody fragments, in order to complement the addition of glycosylation, and attempt to solve the problem of Fc structural instability. In order to accurately represent crystal structures and to investigate conformational change of biological structures, elastic network modelling has been employed on reduced representation structures – whereby the structure is constrained by the addition of numerous spring constants to the structure. In the MARTINI forcefield (Monticelli et al. [131]), there are 3 mechanisms for the restraint of secondary structure employed - Dihedral restraints, Elastic Network modelling, and ElNeDyn elastic network modelling, these are explained in more details in Chapter 2, and in the

section 6.2.4.

6.2 Methods

6.2.1 Atomistic Simulations

As stated in chapter 4, atomistic simulations of carbohydrates, and glycoproteins were run using the AMBER11 simulation package (Case et al. [26]), with the GLYCAM06 (Kirschner et al. [102]), and ff12SB (an update of the well used ff99SB forcefield - Hornak et al. [78]) forcefields, applied to carbohydrates, and proteins respectively. Carbohydrate structures were initially build using the GLYCAM06 web server (www.glycam.org), and the most common rotamer selected, before the protein structure was built using the xleap programme within AMBER. The two molecules were linked using GLYCAM06 parameters, and the xleap tool within AMBER11, before the structure was spatially optimized in xleap. Molecules were then energy minimized for 200 ps before the system was solvated using Single Point Charge (SPC) water, and the system heated from 0 to 310 K over a further 5 ns. For systems containing protein molecules, the system was then equilibrated by restraining protein backbone atoms, and running molecular dynamics on the system for a further 5 ns to allow water to relax. Parameters of interest for CG modelling were extracted using the MDAnalysis python package (Michaud-Agrawal et al. [127]).

6.2.2 Coarse Grained Simulations

Coarse-grained simulations of sugar molecules were run as before in Chapter 4, using the Martini v 2.2 forcefield (De Jong et al. [41]) within GROMACS version 4.5 (Pronk et al. [146]). Molecules were energy minimized over 200 ps, before being placed in a box large enough that they cannot interact with any periodic image, before they are solvated. The solvated system is then energy minimized for a further 200 ps, before carbohydrate regions

are restrained via the central rotational node particle, and proteins are restrained by their backbone particles, and the system is equilibrated over 50 ns. Afterwards the system is released and full molecular dynamics performed for 500 or 1000 ns. Parameters of interest are then extracted using the MDAnalysis python package. The parameters explicitly extracted include all intra-ring distances, and all inter-ring angles, the distances between the central particle of each ring, and dihedral values across the psuedo-glycosydic linkage.

6.2.3 Iterative Force Matching

Force matching was performed as before, with successive AT and CG simulations performed to match parameters.

6.2.4 Elastic Network Modelling

Elastic network model simulations of antibody fragments were performed by producing the CG structure from an AT .pdb file using the martinize.py script available on the MARTINI website. The three restraint mechanisms were adjusted using appropriate flags within the script (martini22, elastic, and elnedyn for dihedral, elastic network, and elnedyn restraints respectively). The script was also modified to calculate disulphide distances more accurately (increasing the cut-off for disulphide distances from 2.4 to 2.5 Å, in order to account for situations where a bond may be slightly higher than the cut-off). Chains in the protein were kept separate, so that elastic networks were applied within a protein chain, but not between chains, allowing the model to attempt to replicate physiological conformational space. The three secondary structure restraints work in slightly different ways:

- Dihedral restraints - a methodology that restrains the protein backbone with dihedral angles. Dihedral angles are calculated from a database accessed by comparison with DSSP (Kabsch and Sander [95]). This methodology attempts to replicate the way in

which structures within a fully atomistic simulation are constrained, though anecdotal evidence suggests that it tends to be unstable with more complex, larger proteins.

- Elastic network model - a methodology that uses simple cut-off values to attach a series of spring constants to any backbone atom within a defined cut-off. This method aims to gain insights into potential structural changes the molecule may undergo. Cut-off distances, and spring strengths are able to be changed such that the dynamics better reflect behaviour expected, though default values are parametrized based on fitting to experimental data, and normal mode analysis (Monticelli et al. [131]).

- ElNeDyn elastic network - a more recent elastic network model, that uses a larger number of springs, and slightly different placement of backbone particles. The model attempts to retain the starting structure more accurately than other methods, and so is not designed for investigating large scale conformational change, but is useful for local changes in the structure of the protein (Periole et al. [145]). It is worth noting that the name used for the network in the original publication is ELNEDIN, this has recently, however, been changed within the literature, and the MARTINI website, to ElNeDyn.

6.3 Results and Discussion

6.3.1 Secondary Structure Restraints and Antibody Fragments

In order to assess the stability of the of antibody fragments, particularly the Fc fragment in the coarse-grained methodology, and due to observing collapse of the Fc fragment when using dihedral, and weak elastic network restraints, simulations were run of individual antibody domains in isolation, with varying restraint systems. Total simulations run are shown in Table 6.1

Table 6.1 Simulations performed for comparison of Secondary Structure Restraints

Restraint	Protein Simulated
Dihedral Restraints	Fab
Dihedral Restraints	Fc
Traditional Elastic Network	Fab
Traditional Elastic Network	Fc
Elnedyn Elastic Network	Fab
Elnedyn Elastic Network	Fc

Each protein was simulated for 500 ns in CG, and the RMSD for each particle in the system calculated for Fab proteins (figure 6.2), and Fc proteins (figure 6.3). 3 different methodologies for restraining secondary structure were assessed:

- Dihedral restraints - a methodology that simply restrains the protein backbone with dihedral angles, based on what type of secondary structure is expected from DSSP
- Elastic network model - a methodology that uses simple cut-off values to attach a series of spring constants to any backbone atom within the cut-off. This method aims to gain insights into potential structural changes the molecule may undergo.
- ElNeDyn elastic network - a more recent elastic network model, that uses a larger number of springs, and slightly different placement of backbone particles to attempt to retain the starting structure more accurately than other methods.

For FAB proteins, it is clear that there is little difference between the structure stability when restrained with the elastic network and ElNeDyn methodologies. The dihedral restraints however, show large increases in RMSD, and general loss of secondary structure. Looking at the structures there is a slight twisting of the structure when the standard elastic network model is applied. The ElNeDyn elastic network model appears to reliably retain the starting structure for Fab proteins, as shown by a much lower RMSD value for the entire protein in Figure 6.2, C.

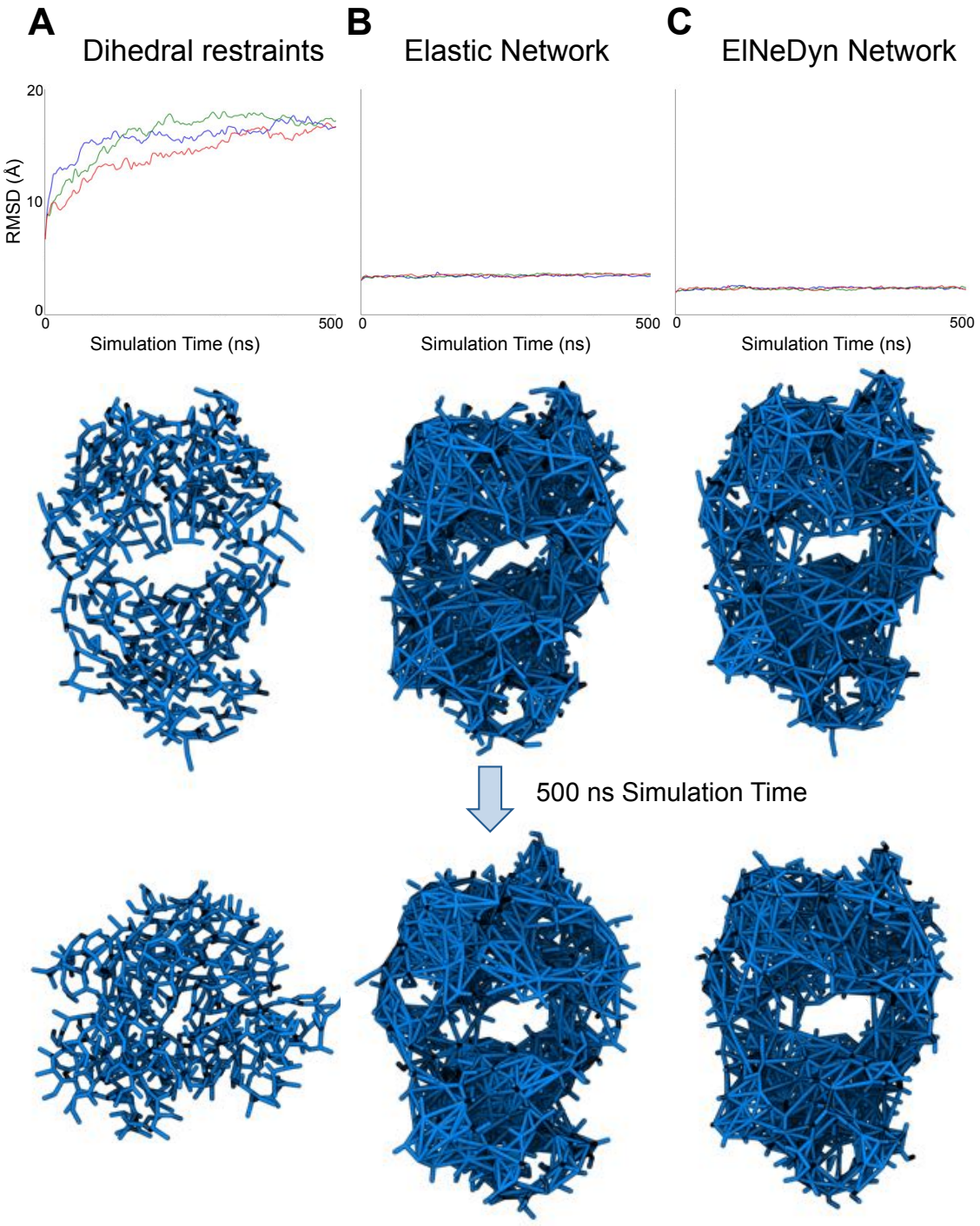


Fig. 6.2 Representative structures taken from simulations of Fab protein with different secondary structure restraints. RMSD plots are shown for 3 simulations of each of the 3 restraints: A) Dihedral restraint B) Elastic Network C) ElNeDyn network. Showing a clear collapse of the structure when using dihedral restraints. There is a difference between the two elastic network type restraints, with the ElNeDyn methodology more closely retaining the known crystal structure, though both show very little change structurally over the course of the simulations.

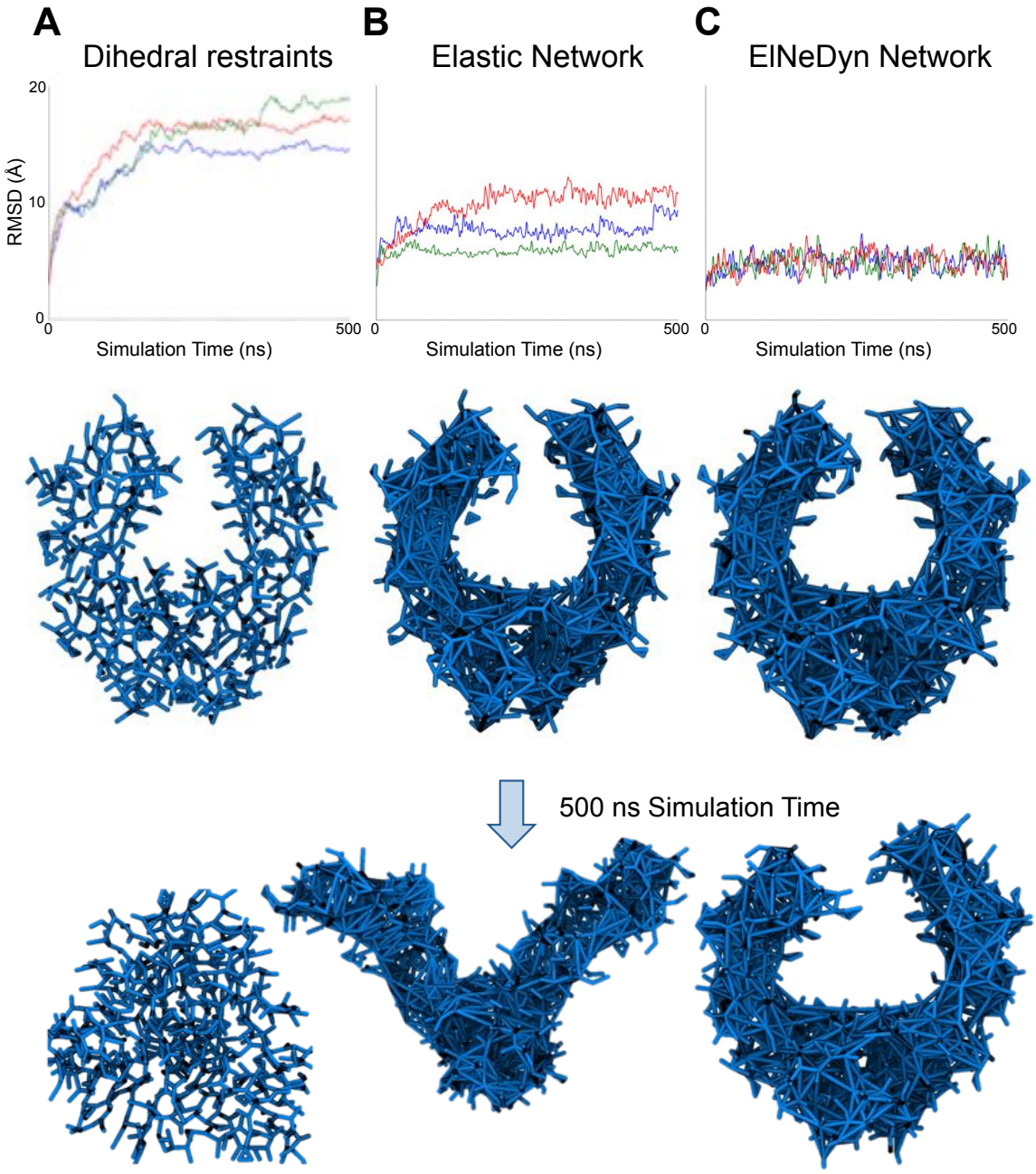


Fig. 6.3 Representative structures taken from simulations of Fc protein with different secondary structure restraints. RMSD plots are shown for 3 simulations of each of the 3 restraints: A) Dihedral restraint B) Elastic Network C) ElNeDyn network. It is clear that the dihedral restraints are extremely poor at reproducing the structure of the molecule. The other two restraint mechanisms give slightly different results, in that the elastic network model produces larger, extended structures as seen in aggregation simulations (Chapter 3), and it is currently unclear whether this is natural behavior of the unglycosylated Fc fragment. The ElNeDyn elastic network model produces a more stable structure, that is still flexible. Neither model shows a collapse inwards.

For Fc proteins as evident from the RMSD plots, each methodology treats the structure differently. The dihedral restraints clearly deviate a large amount from the starting structure, with it quickly falling in on itself, this methodology for restraining protein structures is no longer addressed.

The elastic network model retains the overall shape of the Fc molecule, but with large fluctuations of the extended regions of the protein. This may, however, represent the expected behaviour of the protein when it is unglycosylated. The elastic network model may be producing the most physiologically correct structures (even when one or both extended regions collapse), for a completely unglycosylated, or aglycosylated protein. The ElNeDyn elastic network is by far the most stable methodology, which is expected, due to the more extensive spring network, with the extended regions still being allowed some movement - particularly the region seen to collapse in the simulations of aggregation, however, the RMSD is much lower, and the structure remains much closer to the crystal structure throughout the 500 ns simulations..

In order to assess the effect of glycosylation on the structure of the antibody Fc, and also to attempt to characterize the behaviour and mechanisms of motion in the unglycosylated structure, glycosylated structures were generated with the protein restrained using both elastic network methodology, and the ElNeDyn methodology, and the behaviour between glycosylated and unglycosylated structures were compared. Experimental data for the movement of the domains is also available, and will play a key part in attempting to determine the physiological structural movement of the proteins.

6.3.2 Development of Glycosylation parameters

Glycosylation of the Immunoglobulin G Fc

As the parametrization of simple carbohydrates, such as glucose, and maltose, and the parametrization of more complex carbohydrates such as the headgroup of the Forssman

glycolipid, and GM3 (Chapter 5), has been performed, by combining the two we are able to attempt to study more complex glycosylation structures.

The most common type of glycosylation found on polyclonal, wildtype IgG are characterized by a branching structure, and consist of N-acetylglucosamine (GlcNac), mannose (Man), galactose (Gal), and fucose residues (Fuc). Structures that have been previously developed, or isomers of which have been parametrized previously for other structures such as the glycolipids. A schematic for the structure can be seen in figure 6.4, A, showing the branched structure, and variation of linkages between different carbohydrates. The core structure is made of the branched structure, with no galactose residues, it is denoted G0, whilst structures with one and two galactose residues are named G1 and G2 glycoforms respectively. When fucose is added to the primary GlcNAc, the glycoform becomes G0F, G1F, or G2F respectively. For this study, we chose to focus on the G2F glycoform (And by extension G1F, and G0F are easily accessible by the removal of Gal residues from the ends of each branch, and GX variants by the removal of the Fuc), as this has been shown to be the primary glycoform in Human populations (Dwek et al. [51], Jefferis [91]), and G0F has been shown to be most common in Chinese Hamster Ovary populations. The atomistic structure of the glycoform can be seen in figure 6.4, B, generated from the crystal structure 3AVE, with missing molecules docked using GLYCAM parameters.

Setup of Atomistic Structure

For the generation of the initial glycoprotein structure in atomistic detail, the atomistic structure of the carbohydrate was generated using the GLYCAM06 web server, and the protein using the ff12SB AMBER forcefield. Whilst the initial structures we combined with the use of the web server, the positioning of the carbohydrate groups was not satisfactory, and did not agree with the crystal structure data available on the location. The 3AVE crystal structure (Matsumiya et al. [124]) was used as the basis for the Fc protein structure is also

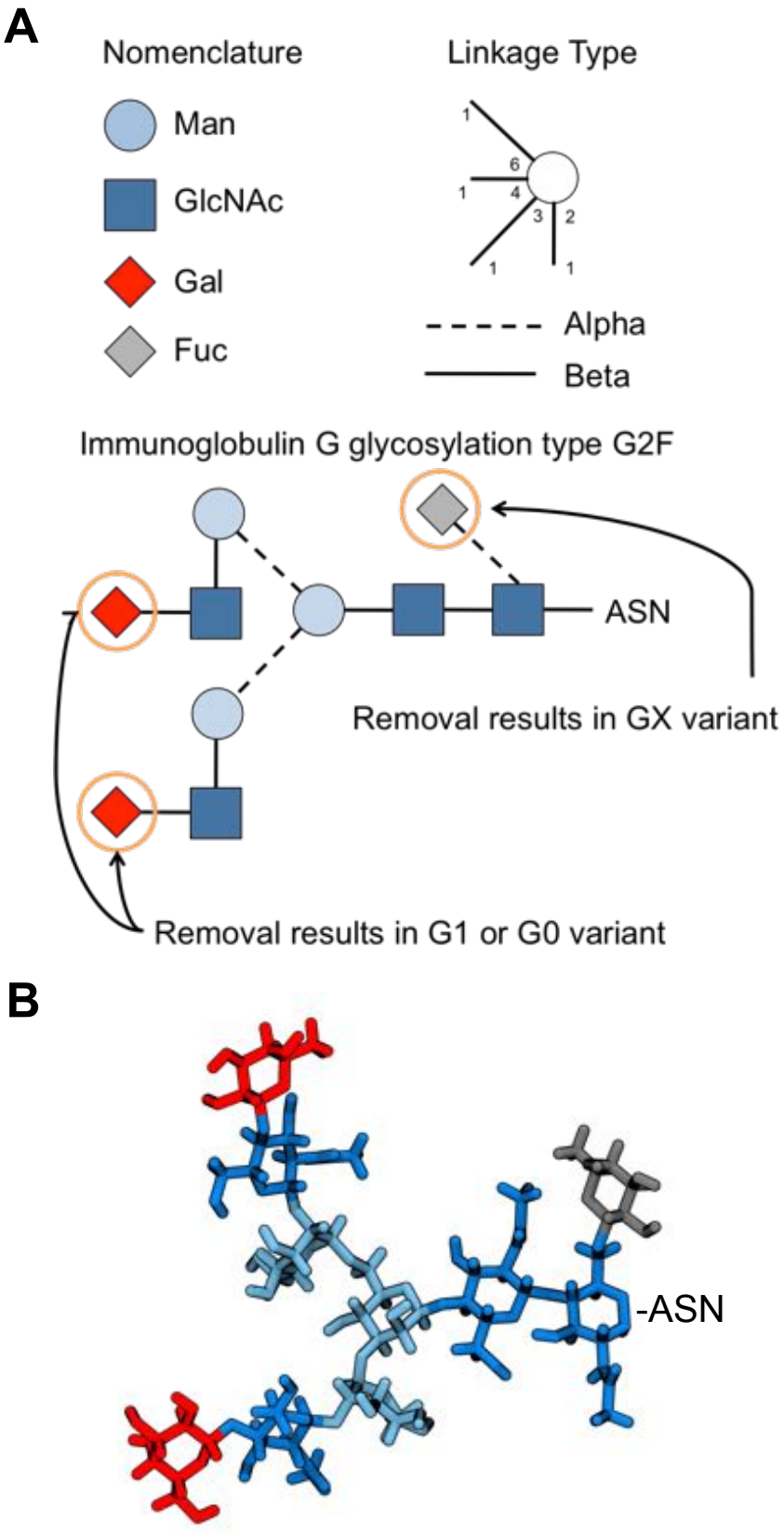


Fig. 6.4 Sequence and structure of the immunoglobulin G Fc carbohydrate. Showing A) Schematic representation of the sequence, including branching. The structure shown in the G2F variant, with the terminal galactose and fucose residues that make up the G1F and G0F variants, and the GX variants. B) Molecular structure of the carbohydrate molecule, coloured to match part A.

glycosylated, with the structure containing the G0F glycosylation. The structure generated by the GLYCAM web server was then docked onto the structure generated from 3AVE, using the 6 ring particles from each glycan residue as a target (6.5). The structure of the full glycoproteins was then run using the AMBER simulation package for a single repeat of 100 ns, which was then used to match coarse grained parameters.

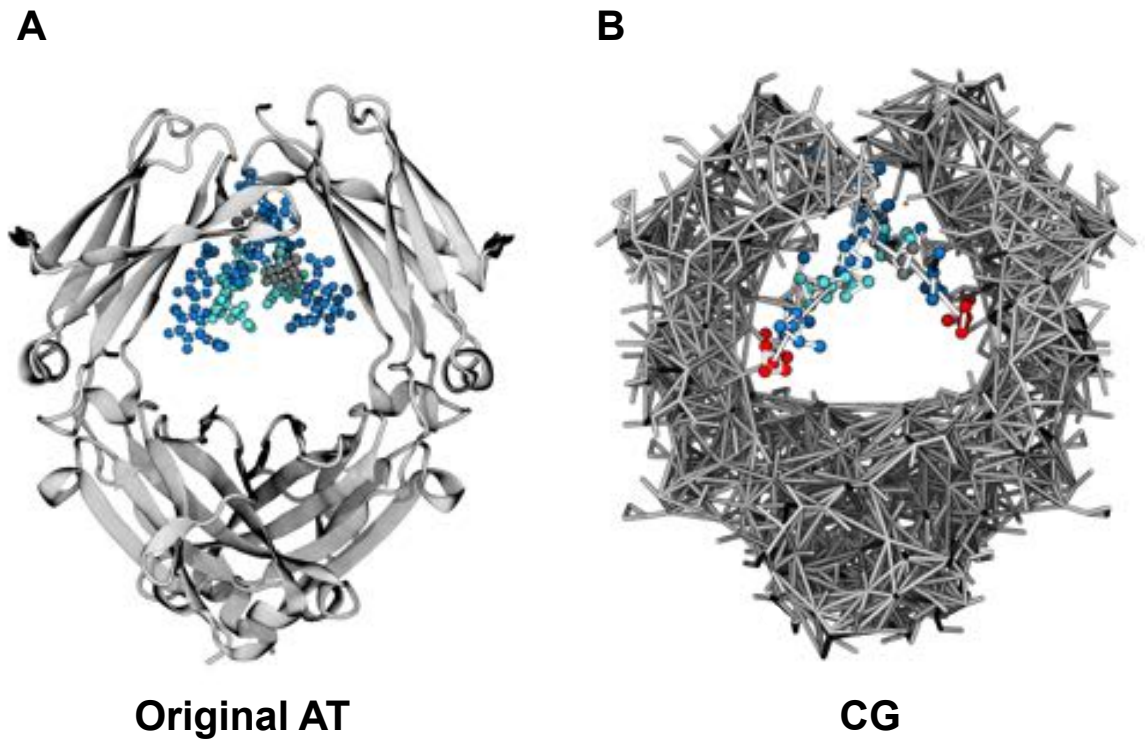


Fig. 6.5 Atomistic and coarse grained structures for the glycosylated immunoglobulin G Fc, showing A) Atomsitic crystal structure 3AVE. B) Resultant coarse grained structure after docking. Note that the coarse grained structure has an additional 4 galactose residues (Red). These are added by comparing the atomistic structure predicted from the GLYCAM and xleap programme.

Mapping of Coarse Grained Structure

The mapping of the carbohydrate structures was done in the same way as for previous molecules. The mapping scheme was retained, and each carbohydrate was represented as a triangle, with additional groups added as particles attached to the ring particles. The rings

were then linked to each other via a series of rotational nodes, placed at the centroid of the ring. The mapping and bonding scheme for the glycan structure can be seen in figure 6.6, B. The carbohydrate structure was then attached to the sidechain particle for the ASN residue via the closest ring particle (Figure 6.7), in a manner similar to the attachment of carbohydrate headgroups to glycolipids.

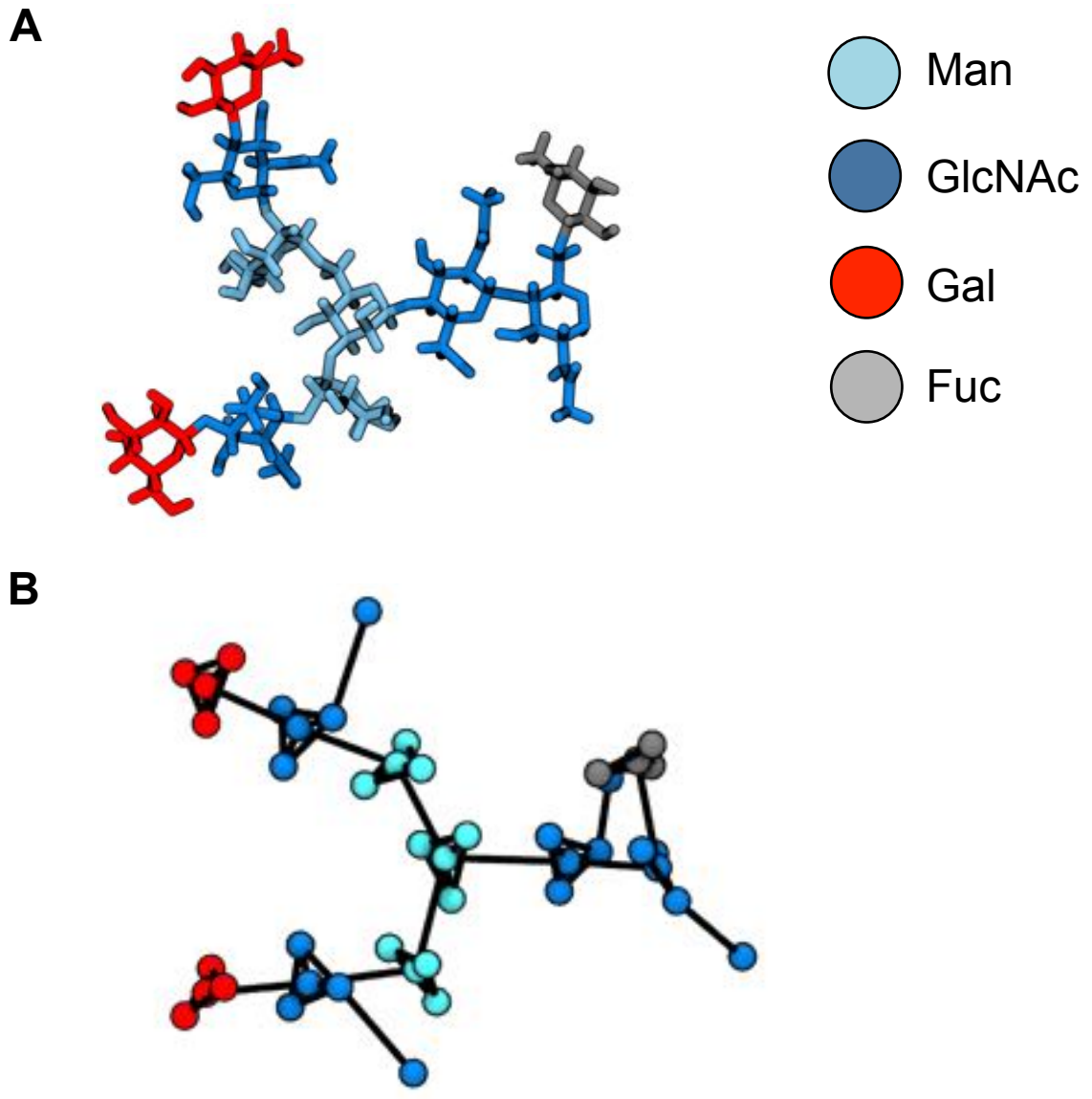


Fig. 6.6 Different structure representations for the IgG Fc G2F carbohydrate, showing A) Atomistic structure coloured according carbohydrate type. B) Resultant coarse grained structure coloured according to carbohydrate type.

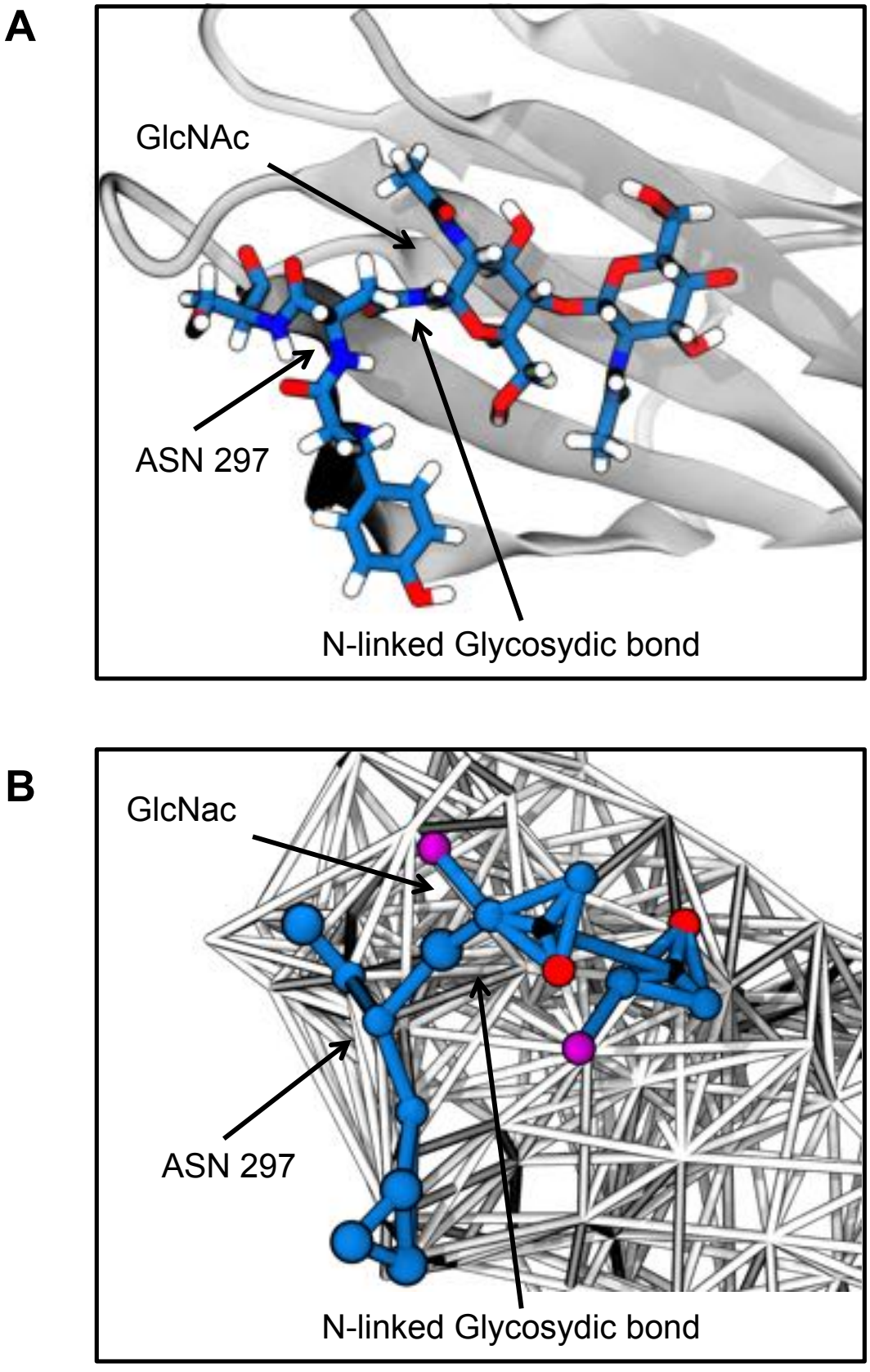


Fig. 6.7 Atomistic and coarse grained structures for the glycosydic bond attaching the primary N-acetylglucosamine to ASN 297. A) Atomistic structure. B) Coarse grained structure.

A comparison of the atomistic crystal structure, and the coarse grained structure resulting from it can be seen in figure 6.5.

Parametrization of the Coarse Grained Structure

Parametrization of the coarse grained structure was performed in the same way as all structures previously. Briefly, ring structures were connected to each other via bonds through the rotational node particles. For each pair of carbohydrates, angle parameters were then controlled across the bonds in order to match the orientation of each ring with their corresponding atomistic ring. Distances between each central node particle were also controlled, as well there being a single dihedral angle controlled between two ring particles, across the central ring, in order to shift the population of states when there are multiple possible conformations. Values for each parameter were initially selected from atomistic simulations, and force constants adjusted until the distribution of each restraint matches between atomistic and coarse grained simulations. Initially, the coarse grained structure was developed independently of any protein molecule, as this greatly increased simulation efficiency, and allowed complete control of the molecule without nonbonded interactions with the protein influencing the structure.

The molecule was then added to a system containing the full Fc protein, and the glycosydic linkage was parametrized in the same way as that from glycolipids, with a single bond placed between the sidechain particle for the ASN 297 residue, and the closest ring particle, and three angle restraints imposed in order to hold the carbohydrate group in the correct location, in a similar manner to the way glycolipid head groups were restrained previously in Chapter 5.

In order to confirm that the force matching was successful for the overall structure, and that the nonbonded interactions between the carbohydrate and protein did not drastically change the behavior of the molecule, emergent properties from the two different simulation

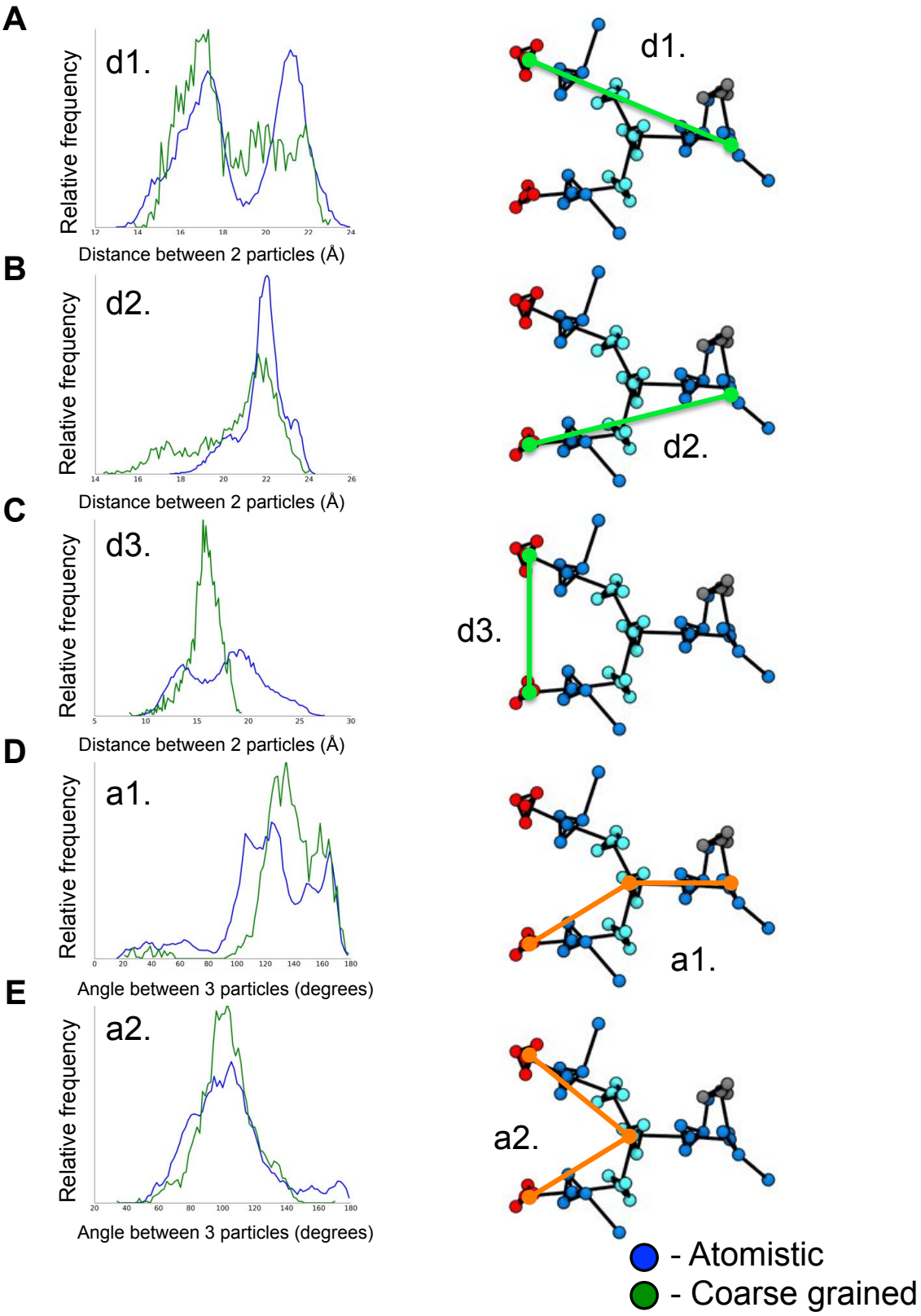


Fig. 6.8 Force matching plots for the emergent properties of the carbohydrate attached to an Fc fragment, with atomistic distribution shown in blue, and coarse grained distribution in green. A, B, and C represent distances made up of multiple parameters, D and E represent emergent angle parameters. In the carbohydrate structure, the lower branch in each case represents the 1-3 linkage arm, and the upper branch represents the 1-6 linkage arm

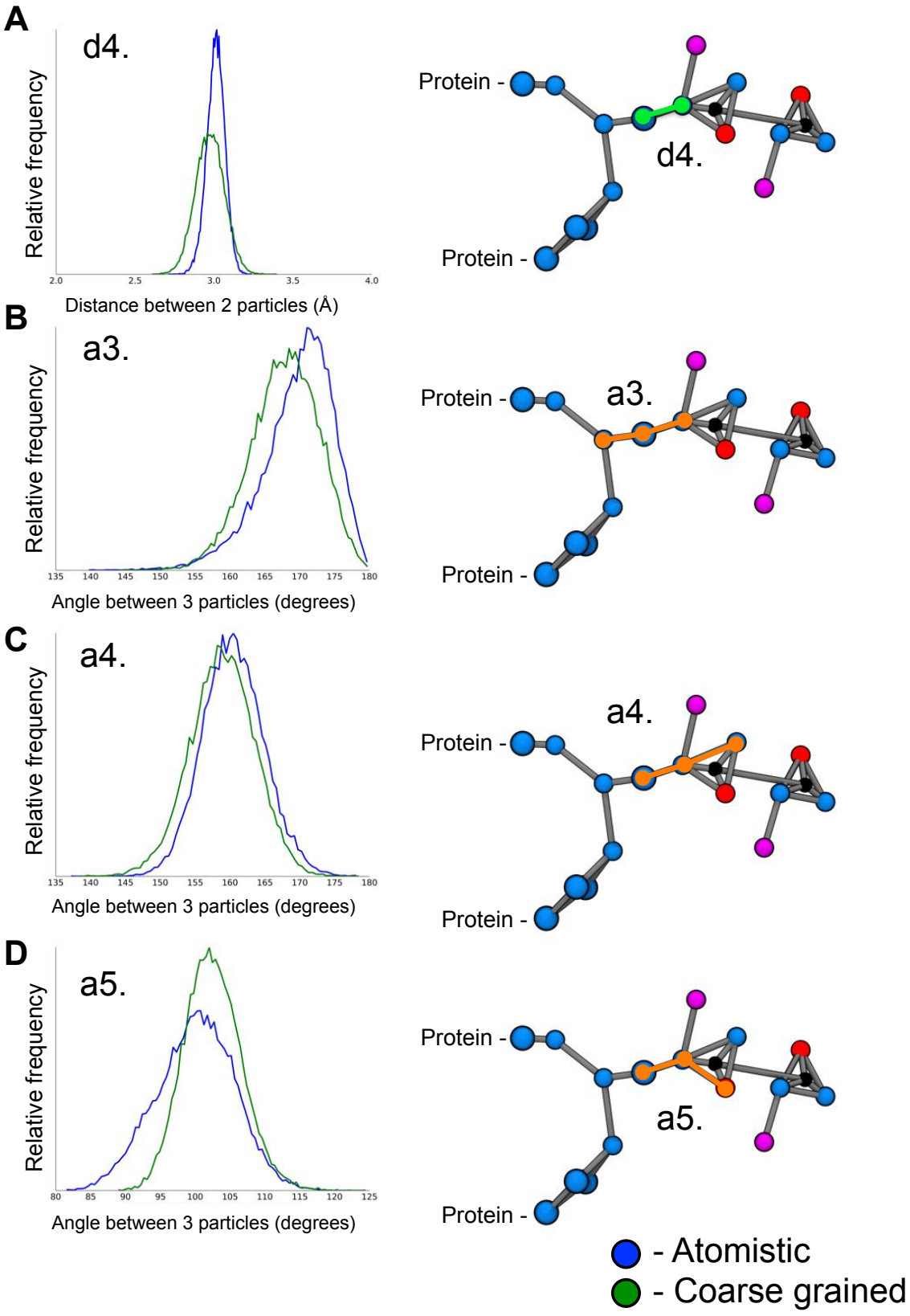


Fig. 6.9 Force matching plots for the N-linked glycosydic bond between ASN 297, and the carbohydrate structure, with atomistic distribution shown in blue, and coarse grained distribution in green. A) represents the distance between ASN sidechain and primary NAG molecule. B, C, and D represent angle parameters.

granularities were extracted (Figure 6.8). A 500 ns simulation was performed of the full CG glycoprotein, restrained with the ElNeDyn Elastic Network in solvent, with ions, and the distributions of emergent properties were extracted and compared to those within the 100 ns AT simulation. The distances between the first carbohydrate and end of each branch were measured - matching well between CG and AT (Figure 6.8, A and B), and the distance between each end carbohydrate measured (Part C). Whilst the exact shape of this distance distribution is not matched, the average values can be observed as being close. These parameters are made up of a combination of over 20 other distance and angle restraints - and the distance between each end particle is made up of the combination of each arm, meaning a relatively large number of parameters need to have been matched to within a few Åor degrees for these to compare correctly, showing that the force matching has generally been successful at reproducing Atomistic parameters.

Another set of emergent characteristics of the carbohydrate chains that were compared are angles formed by the branching chain, namely, the angle formed by the two branching ends (Figure 6.8, E), and the angle formed by the entire chain, start to end (Figure 6.8, F). These distributions also compare well, including replicating the dual peaks displayed in the atomistic simulation.

Parameters for the glycosydic linkage were also compared, the single distance distribution, and three angle parameters are compared in figure 6.9. The distance distribution compares well, and the angle values are all within 10 degrees of each other. These parameters appear stable over a 500 ns simulation, and thus indicate that the model can reproduce atomistic characteristics over useful timescales, and is not hugely influenced by the addition of the full protein structure.

In order to further validate that the model is successfully reproducing the behaviour of the atomistic structure, we also calculated the solvent accessible surface area (SASA) for each structure, using a probe radius of 1.4 Å for the atomistic structure, representing

the radius of a single water molecule, and a probe radius of 2.64 Å for the coarse grained structure, representing roughly the size of a CG water molecule. These two values have been shown previously to produce roughly comparable results between AT and CG structures by Sø rensen et al. [173]. The values for the comparison of the two structures are shown in figure 6.10, and the snapshot from each simulation shows that the carbohydrate groups fill roughly the same space in each molecule. The values for the two structures are similar - with the atomistic structure having a solvent accessible surface area averaging around 26,000 $Å^2$, and the coarse grained structure having an average of around 23,500 $Å^2$. Whilst these values don't match exactly, it is expected that the coarser structure would have a lower accessible surface area, when using a probe of appropriate width. Plotting the surface area over time also confirms that the structure is stable in both instances, as there are no large fluctuations in the values.

6.3.3 Structural Studies of the Immunoglobulin G Fc

Once the model of glycosylation on the IgG Fc fragment were completed, simulations were performed to assess the conformation of the glycoprotein, and to attempt to gain insights into the movement and conformational changes undertaken by the molecule when it is unglycosylated. Simulations were run using both previously described types of elastic network - the traditional MARTINI Elastic Network, with a slightly higher force constant compared to previous simulations (15 kJ/mol as opposed to 10 kJ/mol), and the newer ElNeDyn methodology, and unglycosylated, and fully glycosylated cases of each were run for 500 ns in water, three repeats of each were performed. A summary of the simulations performed can be seen in table 6.2.

Experimental data exists for possible properties of the Fc fragment with and without glycosylation. Crystal structures indicate that the structure undergoes a form of inward collapse once carbohydrates are removed, whereas SAXS and FRET data suggest that the

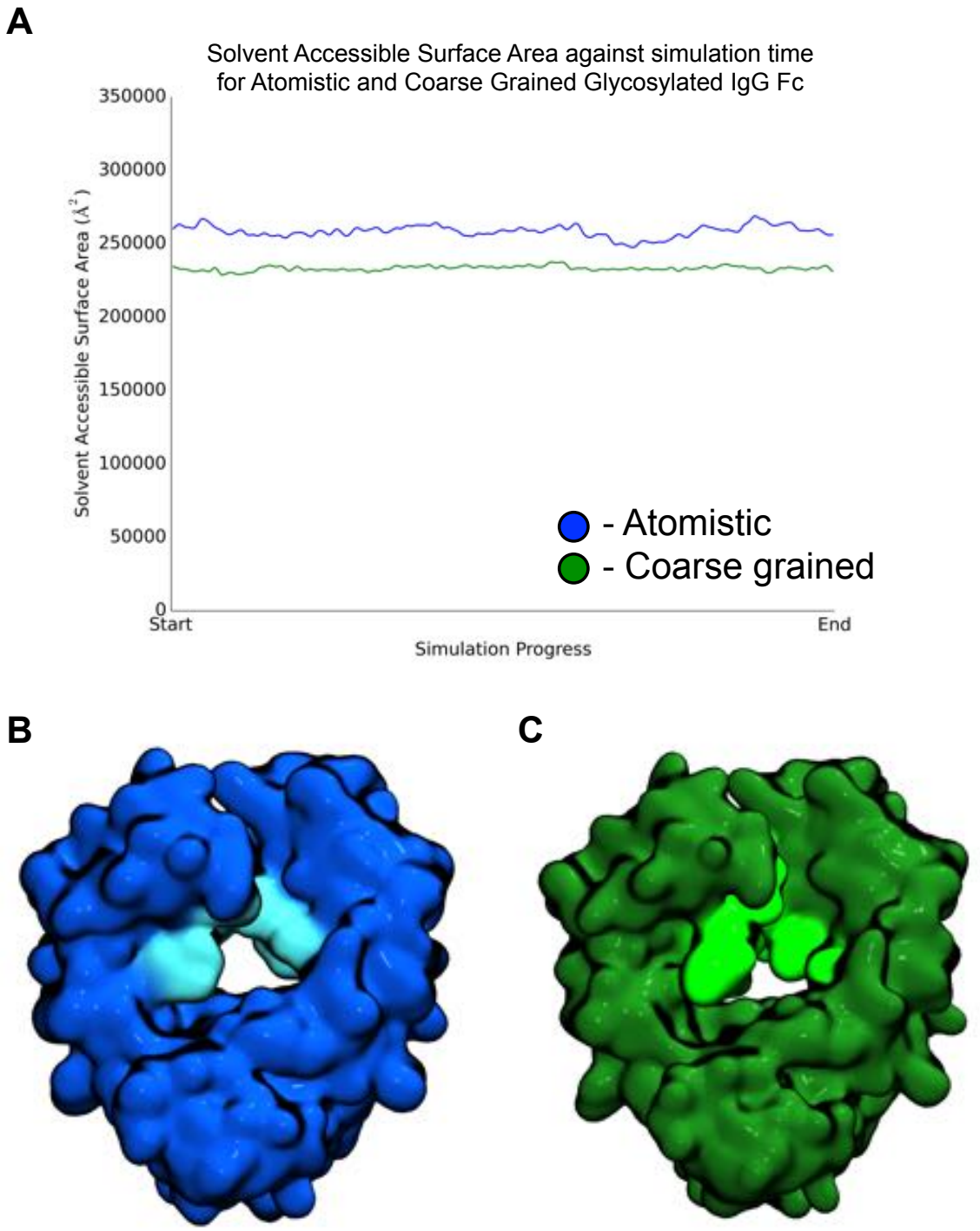


Fig. 6.10 Solvent Accessible Surface Area for atomistic and coarse grained IgG Fc structures, showing A) Solvent Accessible Surface Area over time, showing roughly comparible values, with no major fluxuations over 200 ns atomistic, and 500 ns coarse grained. B) Atomisitic surface area visualization, with carbohydrate molecules represented lighter, and in the center. C) Coarse grained surface area visualization, with carbohydrate molecules represented lighter, and in the center.

Table 6.2 Simulations performed for comparison of Secondary Structure Restraints, and Immunoglobulin Fc Glycoforms

Glycosylation Type	Secondary Structure Restraints
G2F	Elastic Network
G2F	ElNeDyn
Unglycosylated	Elastic Network
Unglycosylated	ElNeDyn

structure becomes more flexible, and samples many conformations, once glycosylation is removed. By comparing data from CG simulations with experimental data, we can gain insights and study how the protein appears to change it's conformation with and without glycosylation.

There are 6 main values used to assess the conformational dynamics of the Fc fragment in each case:

- Root Mean Square Deviation - A measure of the difference between the structure of the protein at the start of the simulation, compared to the structure at each frame.
- Radius of Gyration - a measure of the compactness of the structure. SAXS data calculates the radius of gyration for a population, and indicates roughly how far the protein extends from its center of mass, so can indicate the degree of flexibility.
- ASN 297 Distance - The distance between two ASN residues within the structure.
- Extended region angle - The angle formed from three selected ASN residues the represent the angle of the extended region of the protein.
- SER 265 Distance - a measure of the distance between two SER residues at the base of the extended regions of the structure. This is measured directly using FRET, and indicate the distance between the tips of the two extended regions of the horseshoe shaped Fc.

- PRO 329 Distance - Measured from crystal structure, this distance is seen to vary between glycosylated and unglycosylated/aglycosylated structures. Many glycosylated structures have a smaller distance between these two residues, and these generally support the collapse mechanism of antibody Fc conformational change.

These values were calculated for each of the three simulations of glycosylated, and unglycosylated protein structure, and, where applicable, compared to the experimental values, in order to evaluate the behaviour of the protein when glycosylation is present, and absent.

RMSD values indicate that, in both secondary structure restraint methodologies, the unglycosylated structure deviates more from the starting structure (Figure 6.11, A and B), indicating that they are more unstable, and more flexible. The elastic network model also appears more flexible, with both structures deviating by much more than either structure in the ElNeDyn simulations.

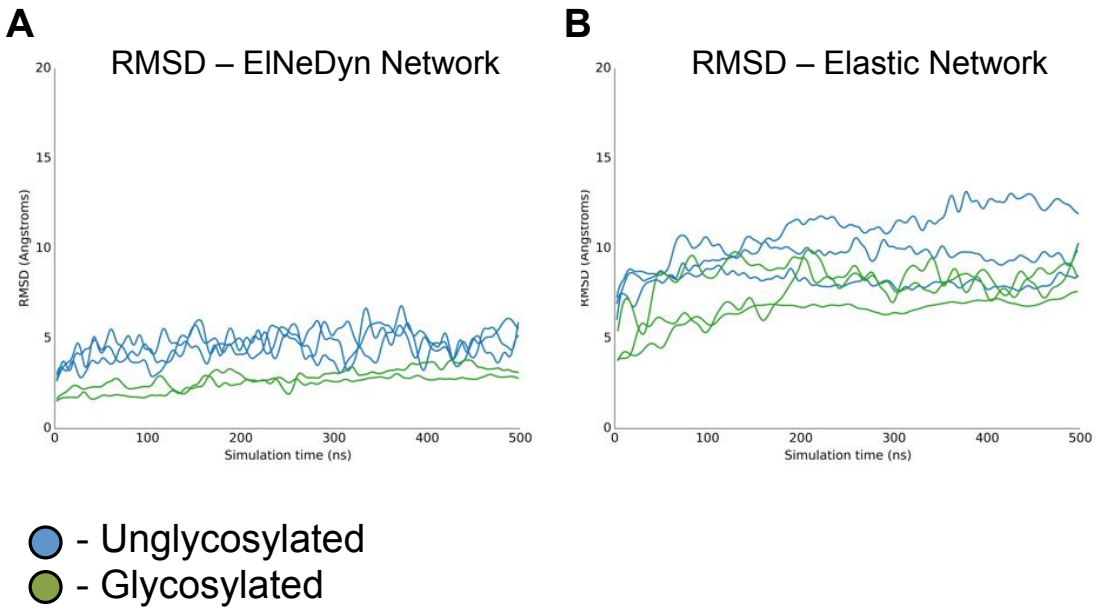


Fig. 6.11 RMSD values for glycosylated and unglycosylated IgG Fc. RMSD values show that unglycosylated Fc proteins (blue) generally deviate more from the initial structure, and that the ElNeDyn network model (A) generally holds the initial structure more stable than the elastic network model (B).

The radius of gyration values for all structures correlate very well with the experimen-

tal values derived from Borrok et al. [17], whereby glycosylated structures tend to be more compact, and less extended than structures that are unglycosylated (Figure 6.12). The ElNeDyn values for unglycosylated proteins match particularly well with the SAXS data, indicating that the range of motion exhibited by them is potentially the physiological behaviour of the molecule.

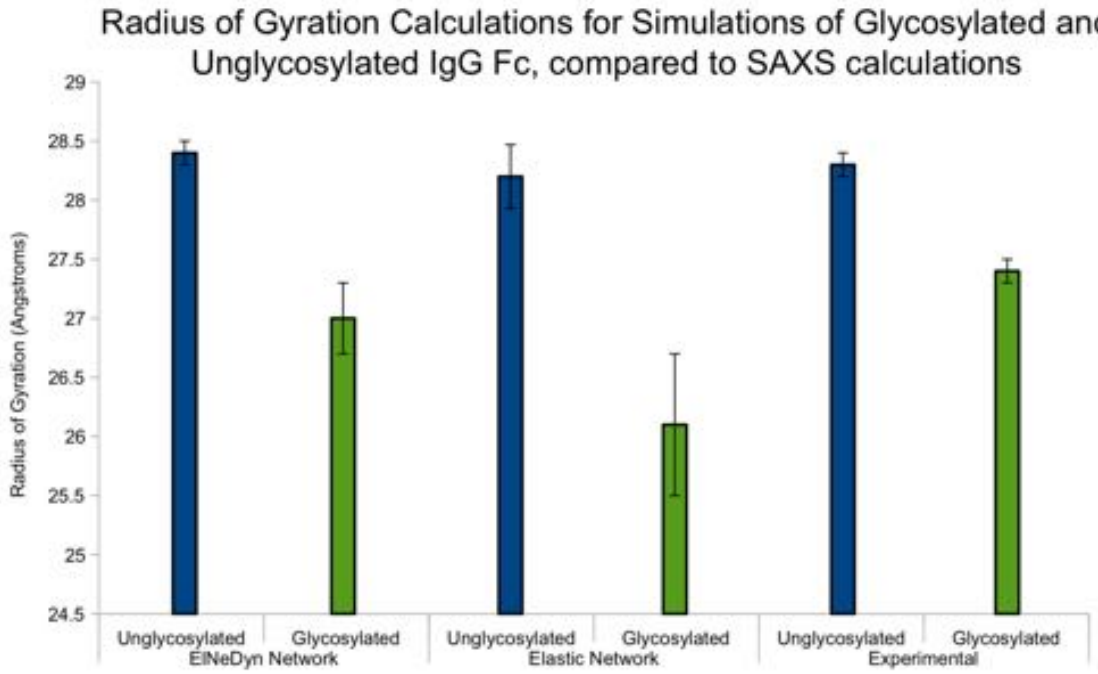


Fig. 6.12 Radius of gyration values for glycosylated (blue) and unglycosylated (green) IgG Fc, comparing elastic network models, and SAXS data. Glycosylated proteins generally have a smaller radius of gyration, indicating that they are more compact.

The distance between two ASN 297 residues indicates how close the starting point of each glycan attached to the molecule are (Figure 6.13, C). It is clear that the distance is more stable with the ElNeDyn methodology, and that the glycosylated structure distribution is much more stable, figure 6.13, A. This supports the hypothesis that once the carbohydrate groups are removed, the molecule becomes more flexible. The elastic network model however, seems to show little correlation between glycosylated and unglycosylated proteins (Figure 6.13, B), possibly due to the smaller number of secondary structure restraints, and

the forces exerted on the ASN residue by the bulky carbohydrate group.

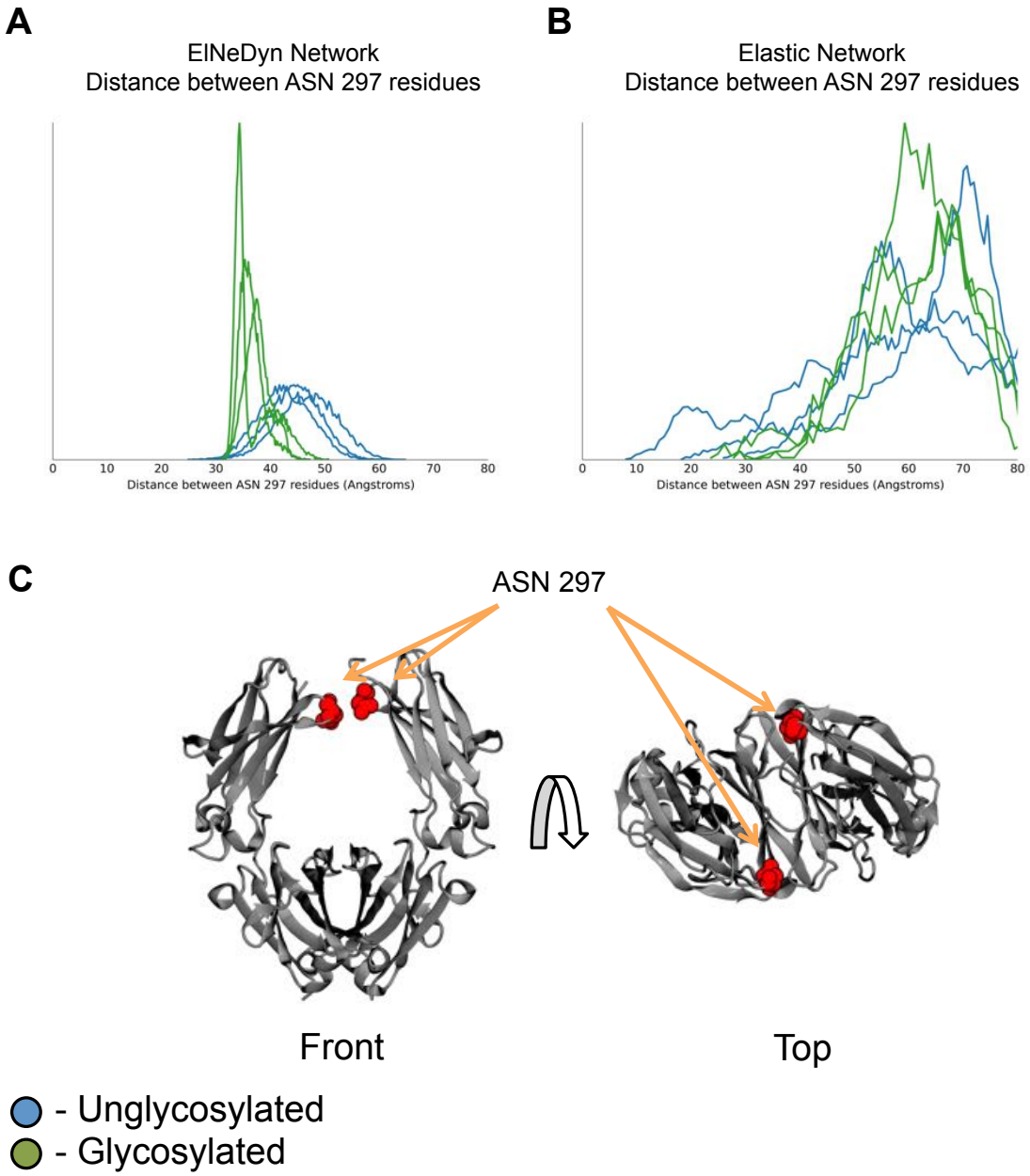


Fig. 6.13 Distance between ASN 297 residues for glycosylated and unglycosylated proteins, showing A) ElNeDyn network models, where the distance distribution is much wider for unglycosylated proteins, indicating that they are more flexible. B) Elastic network models, where the distance distributions for both structures are wide, indicating large degrees of flexibility, though the glycosylated structure is slightly more stable. C) Schematic of Fc structure with ASN 297 residues highlighted (Red).

An angle was measured down a single extended region of the protein through residues

359, 290, and 300 (Figure 6.14, C). Giving a rough idea of the conformation of the extended regions of the protein, though it's worth noting that twisting motions are not captured with this value. In both secondary structure restraint mechanisms, the glycosylated structure value is lower (though only slightly for ElNeDyn, Figure 6.14, A). The difference is much clearer with the elastic network model (Figure 6.14, B) - as the glycosylated structure stays in a similar conformation to the starting structure, whereas the unglycosylated structure has a large deal of deviation, including in each seed. This indicates that the structure is not reaching a single stable conformation, as each seed produces a different result.

Förster Resonance Energy Transfer (FRET) experiments have been performed on the glycosylated and unglycosylated Fc fragment (Kelliher et al. [98]). With fluorescent probes attached to the SER 258, shown highlighted on the protein structure in Figure 6.15, C. FRET data indicates that the distance has many more conformations when the structure is unglycosylated, as shown in part D. Our data indicates a similar relationship, particularly with the ElNeDyn network, Figure 6.15, A - the glycosylated protein has a much smaller distribution of potential distance values, when compared to the unglycosylated structure. The elastic network model shows a similar trend, though one seed of each simulation type show some form of collapse, and thus a lower distance, Figure 6.15, B.

Finally, in order to compare with the wealth of crystal structure data available, a further distance value was measured for each simulation. The distance between PRO 329, shown on the crystal structure in figure 6.16, B. This distance has been measured for 4 unglycosylated, and 8 glycosylated crystals of IgG Fc proteins. These structures suggest that unglycosylated proteins exhibit a "collapsing" motion, with the PRO 329 distance becoming smaller when glycans are removed, though it is argued that this is an artefact of the crystallization process (Sondermann et al. [175]). Our results contrast with this finding, and indicate that when the glycan moieties are removed from the inner surfaces of the Fc, that the distance between PRO 329 becomes larger on average (though there is little difference when the elastic net-

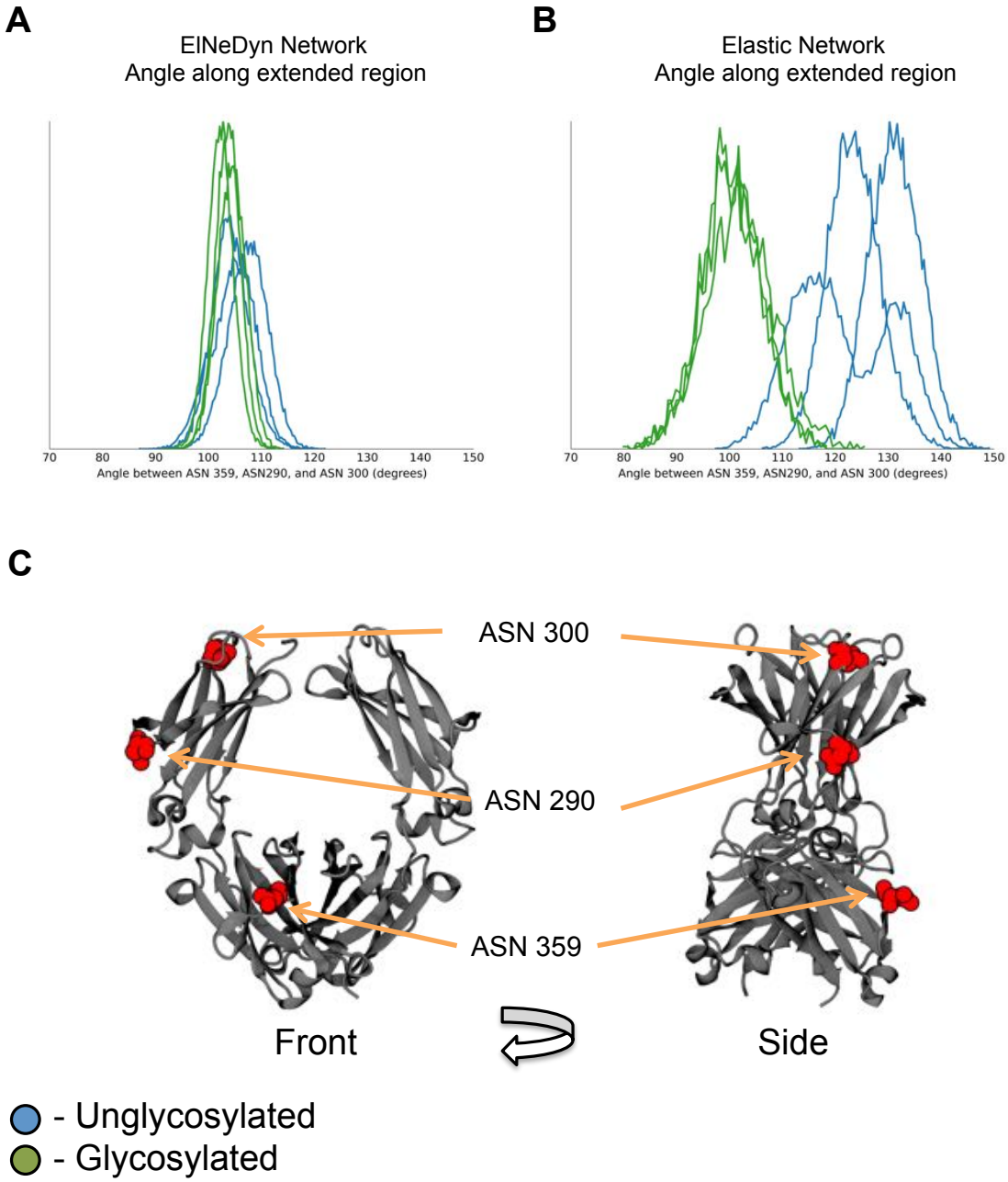


Fig. 6.14 Angle along three selected ASN residues in the extended region of one side of the IgG Fc, showing A) Relatively similar values for ElNeDyn network simulations, though the unglycosylated angle is slightly larger, and slightly more variable. B) Elastic Network simulations, where the glycosylated form of the protein appears much smaller, and more stable than the unglycosylated protein, which has a larger distribution of angles. C) Schematic of Fc structure with ASN 359, ASN 290, and ASN 300 highlighted (Red).

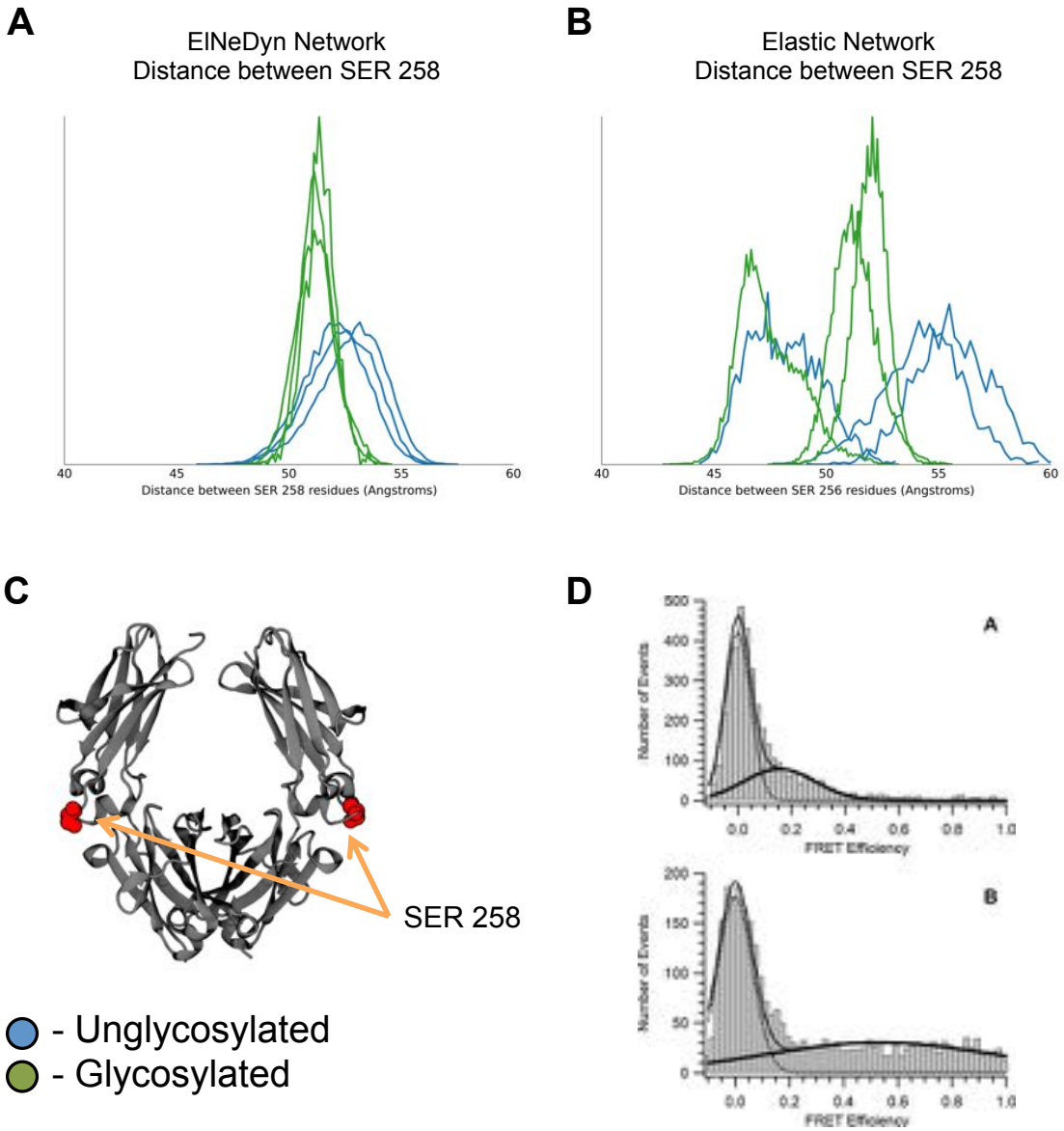


Fig. 6.15 Distance between ASN 297 residues for glycosylated and unglycosylated protein, showing A) ElNeDyn network models, where the distance distribution is larger for unglycosylated proteins, matching FRET data (C), B) Elastic network models, where the distance distribution for unglycosylated structures are generally larger. D) Schematic of Fc structure with SER 258 highlighted (Red). D) FRET data showing that the distribution of SER 258 distances are larger in unglycosylated structures. Figure adapted from Kelliher et al. [98].

work model is studied). The increase in distance average is due to the high flexibility of the protein regions. Distances are compared in Figure 6.16, A.

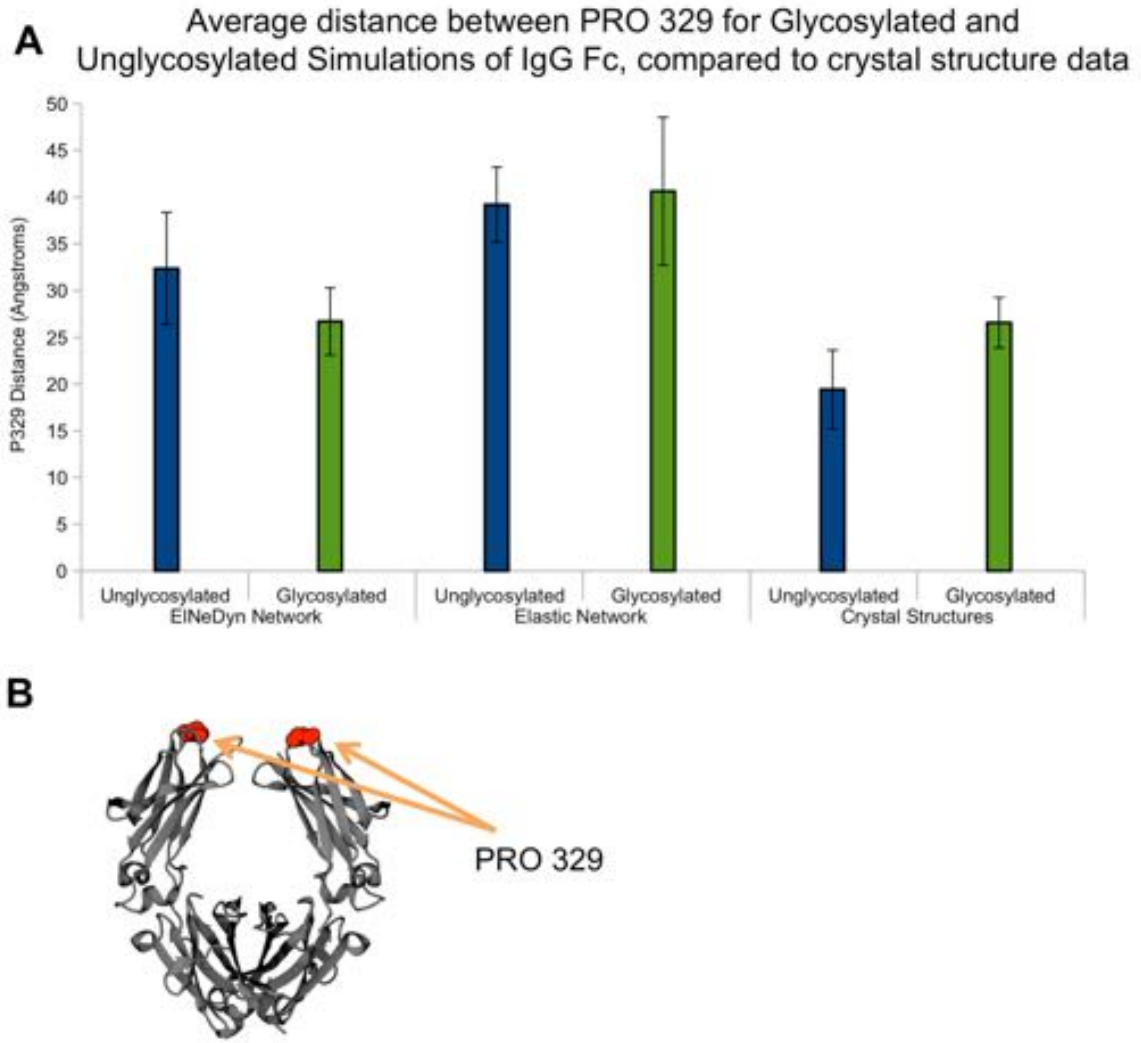


Fig. 6.16 Average distance between PRO 329 residues for structures of IgG Fc, showing glycosylated (blue), and unglycosylated (green) structures in A) Bar plots, and B) Distribution plots. Generally, for crystal structures the unglycosylated structures are more compact, whereas glycosylated structures are less compact.

Overall, our studies support the hypothesis suggested by mutational studies, FRET, and SAXS data - that once carbohydrates are removed from the IgG Fc, the structure becomes more flexible, and open to conformational change, rather than the collapse seen by crystal

structure studies.

6.3.4 Discussion

In this chapter we have demonstrated an extension of the previously used methodology for simulations of carbohydrates to a glycoprotein. By initially glycosylating a protein with very simple carbohydrate structures, we are able to deal with stability and parametrization issues, before glycosylating a much more complex structure, such as the immunoglobulin G Fc protein. We have demonstrated that the glycoprotein behaves in a similar manner when run in atomistic detail, and in coarse grained detail, before investigating structural effects of the Fc protein.

By using the same iterative force matching technique described previously, we were able to successfully generate a set of coarse grained parameters for the glycosylated immunoglobulin Fc, and show that emergent properties between the two models are very similar - this is a good measure of effectiveness of the technique, as many of the emergent properties measured are the results of over 20 separate parameters that must be accurately matched in order for them to resemble their atomistic equivalents. A further validation that the molecules are behaving as expected is examination of the Solvent Accessible Surface Area (SASA). The values for the atomistic and coarse grained systems are very similar, and stable - indicating that there are no major fluctuation in the structures.

Once the model was confirmed as behaving in a similar manner to the atomistic structure, studies were performed on the structure with two different secondary structure restraints, in glycosylated and unglycosylated forms. By comparing to numerous pieces of experimental data, we can see that, in both cases, the unglycosylated form of the protein is less structurally stable, and tends to undergo conformational flexibility, with the extended regions (the "arms" of the horseshoe shape) moving to a larger degree than in the glycosylated form, indicating that the glycosylation stabilizes the protein into a single conformation. Our

results do not support the hypothesis that removing glycans from the structure leads to a collapse, and the shortening of the PRO 329 distance. We also observe that the closed conformation seen in crystal structures, is not seen in simulations at all, with structures in silico generally becoming more open and variable once carbohydrates are removed.

This result is perhaps not surprising - a technique most utilized because it allows investigation into protein dynamics would possibly be expected to show a dynamic structure. This may not be the case however, as, due to the unique horseshoe structure of the antibody Fc - it may be expected that once carbohydrates are removed, the twin extended regions are pulled inwards as seen in crystal structures.

The sensitivity of elastic network models to changes in their force constants are also highlighted here, whereby an increase in the force constant used to define the springs appears to have a large effect on the behaviour of the structure by comparison to those simulations in Chapter 3, where collapse of the structure was observed. This sensitivity to changes in the elastic network means that further study is required to determine which values most accurately represent the physiological activity.

Another major limitation of this particular system, is that the immunoglobulin fragment is simulated without the hinge peptide regions, which are present within most of the experiments. This will result in a discrepancy between the simulations and experimental data, something that will be addressed in ongoing work before publication.

Now that the glycosylation methodology has been successfully applied to a complex protein as a case study, it is possible for the method to be extended to other glycosylated proteins. Examples include EGF receptor proteins, viral capsid proteins, or major histocompatibility complex (MHC) proteins.

An example where this has already been done is with the example of the Dengue virus envelope protein, which is known to be glycosylated, and this glycosylation is essential to the structure of the virus. The protein structure was glycosylated in the same manner as the

immunoglobulin Fc, before being used by Dr Tyler Reddy in the production of simulations of the Dengue virion.

Further work could be done on observing the effects of slight changes in the sugar sequence and structure, though this may not have a large effect in a coarse grained model. It would also be of interest to see if fully atomistic simulations confirm these dynamics, though this would be a particularly large undertaking, as the glycosylated protein is an extremely large system. This work could also be extended to look at simulations of complete, coarse-grained, immunoglobulins, and their aggregation, and also to investigate the aggregation of glycosylated Fc proteins, compared to previous (Chapter 3) work done with unglycosylated Fc proteins, as it would be expected that proteins that are glycosylated will aggregate in a different way to those that are unglycosylated. A region of interest is also to look at the interactions between the immunoglobulin Fc, and the respective Fc*gamma*R receptor protein, as we have shown previously that the experimentally implicated region of the Fc is shown to be highly self-interacting in Chapter 3.

There are also, however, limitations of the methodology. It is unclear, for example, exactly how well the coarse grained secondary structure restraint methodologies replicate atomistic simulation data, though the methods were developed by comparison to fully atomistic simulations. It is also possible that the coarse grained unglycosylated results do not accurately map to some experimental data, as it was performed on aglycosylated (ie still containing a single carbohydrate residue), rather than fully unglycosylated proteins, though we do not believe this is the case.

Chapter 7

Conclusions and Future Directions

This thesis represents a study of the aggregation of monoclonal antibodies from a computational perspective. Along the way, a novel methodology for simulating carbohydrates was developed to overcome hurdles encountered, and techniques were applied to particular questions for the first time, in the form of applying Markov State Models to the problem of protein self association.

A particular outcome of this thesis is the invention of a novel way of simulating coarse grained carbohydrates. A method that has resulted in the generation of structures and topologies that are currently being employed by others.

7.1 Major conclusions

Chapter 3 introduces the problem of protein aggregation from a coarse grained perspective, and initial studies focus on attempting to understand how self interactions occur with two publicly available immunoglobulin fragment structures. Initial work demonstrates that the newer MARTINI version 2.2 forcefield more accurately modelled soluble protein behaviour than the older MARTINI version 2.1. Investigation of self-self interactions then occurred, with simulations of a large number of small scale systems, and a small number of large

scale system for each fragment. Both proteins show distinct "hot spots" of self-self interaction on their surfaces. Results show that there is a difference in the behaviour of the two proteins, one Fab, and one Fc, when simulations are scaled upwards, with Fab proteins behaving in a different manner when a larger number of them are present, indicating that, for some proteins, small scale simulations may not accurately replicate the correct interactions. Finally, interaction landscapes can be generated, that give information about the mechanics of interactions, such as whether interactions are between "hot spots" or whether one spatial conformation of the two proteins is preferred when interacting.

The study described in Chapter 4 shows how the methods developed in Chapter 3 can be scaled up and automated. An automated series of scripts is presented that is able to take an atomistic level protein structure, and a set of input parameters, to generate, simulate, and analyse a large array of simulations on their self association behaviour. A novel application of Markov State Modelling is also applied here, whereby the results of many simulations can be further coarsened into spatial relationships, which are used as input for a Markov State Model. The system is tested on a set of six protein structures with known aggregation propensities, that were provided by Lonza Biologics. Results from the Markov model analysis suggest that the system is adequate at predicting relative aggregation propensities from 40 simulations, as the proteins are able to be ranked correctly in order of increasing aggregation propensity based on the results of the analysis alone. Using the methods described in Chapter 3, it is also possible to gain information about specific interaction regions on the proteins, and information about the mechanisms by which they interact.

Chapter 5 represents a shift in the focus of the project. Initial work on Fc proteins, with a variety of parameters, showed a destabilization of the structure, and an unfolding or denaturing of the protein in silico. To address this issue, it was decided that a set of parameters for the carbohydrates within the protein structure, that are known to stabilize it, are required. This Chapter represents the first steps to generating that model, with the development of

much simpler carbohydrate parameters for glycolipids. The model is generated by matching the behaviour of a fully atomistic model of a carbohydrate to a set of coarse grained particles, and a novel mechanism is presented, with "invisible" particles for holding multiple carbohydrate rings together. Initial parameter sets focussed on very simple mono, and disaccharides such as glucose and maltose, but eventually more complex structures such as multiple carbohydrate-containing glycolipids are developed, tested, and applied to biologically relevant systems by colleagues within the laboratory. The initial parametrization of molecules is successful, and more ambitious structures can be now be attempted.

Chapter 6 is an extension of the work done in Chapter 5 to a much more complex biological carbohydrate - the glycoform attached to the Immunoglobulin Fc. Glycosylation of a coarse grained immunoglobulin structure (another methodology that has not yet been observed in the literature) is shown. Parameters are developed in the same way as previous carbohydrates, and the resultant model is shown to accurately reproduce the behaviour of an atomistic structure. The fully glycosylated structure is then studied in order to investigate the effects of glycosylation on immunoglobulin Fc structure. There are two major potential results of the removal of carbohydrates from the Fc protein, either a collapsing conformational change, or a more conformationally open structure is predicted. Coarse grained simulations of the glycosylated and unglycosylated protein show that this model supports the conformationally destabilized hypothesis, as the protein is seen to become more flexible upon the removal of carbohydrates.

7.2 Future directions

Much of the work presented in this thesis could be extended further. Particularly with the increasing speeds and complexity of CPU and GPU technology, simulation sizes that took huge computational effort to run previously can now be run much more efficiently.

Work on immunoglobulin aggregation can initially be extended simply by increasing the

number of antibody structures simulated with the automated series of protocols. The PEP-SoP protein set contains many more structures (over 100) with known aggregation propensities, and a great validation for the methodology would be to run a much larger set of structures, to see if the Markov State Models can accurately predict relative aggregation propensity. Due to the modular nature of the high throughput methodology, novel analysis techniques can be quickly applied and tested against current data, before being included in the general protocol, leading to the evolution of the script as novel, more advanced methods become available.

Another region of interest from Chapter 3 and Chapter 4 is the analysis of global protein-protein interaction data. There has been very little study on the general ways in which non-specific interactions between proteins occur, and we have access to a large amount of data (240 simulations) that contains information on intermediate states, and rates between them, for protein-protein interactions. It would be interesting to apply more advanced Markov State Model construction techniques to the much larger dataset, to see if general properties of protein-protein interaction can be elucidated.

Another area of interest, is the application of Markov State Models to much more complex systems containing 8, or even 16 proteins.

Regarding the carbohydrate parametrization, a primary extension is simply increasing the library of available carbohydrates, and by extension glycolipids and protein glycosylation types, that are able to be simulated. By doing this, we can increase the complexity, and physiological relevance of coarse grained membrane and protein systems being generated within the lab. It would also be very easy to extend the methodology to more complex membrane structures, such as GPI anchors, and the glycosylation of membrane proteins.
Work combining both areas of the thesis is also of great interest - looking at how the glycosylated Fc protein behaves differently to the unglycosylated one, when they are simulated in larger numbers. Also, with the inclusion of carbohydrates, simulations of a full antibody

structure are potentially possible. Whilst the molecules are comparatively extremely large (ca. 1,200 amino acid residues), coarse graining means that these sizes are accessible. As of yet, no studies looking at the large scale dynamics of full immunoglobulins *in silico* have been reported.

References

[1] Adams, G. P. and Weiner, L. M. (2005). Monoclonal antibody therapy of cancer. *Nat. Biotechnol.*, 23(9):1147–1157.

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). The RNA World and the Origins of Life.

[3] Alkan, S. S. (2004). Monoclonal antibodies: the story of a discovery that revolutionized science and medicine. *Nat. Rev. Immunol.*, 4(2):153–6.

[4] Andersen, C. B., Manno, M., Rischel, C., Thórólfsson, M., and Martorana, V. (2010). Aggregation of a multidomain protein: a coagulation mechanism governs aggregation of a model IgG1 antibody under weak thermal stress. *Protein Sci.*, 19(2):279–290.

[5] Arnold, J. N., Wormald, M. R., Sim, R. B., Rudd, P. M., and Dwek, R. A. (2007). The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu. Rev. Immunol.*, 25:21–50.

[6] Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80(1):505–515.

[7] Auer, S., Dobson, C. M., Vendruscolo, M., and Maritan, A. (2008). Self-Templated Nucleation in Peptide and Protein aggregation. *Phys. Rev. Lett.*, 101(25):4.

[8] Baaden, M. and Marrink, S. J. (2013). Coarse-grain modelling of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 23(6):878–886.

[9] Basdevant, N., Borgis, D., and Ha-Duong, T. (2013). Modeling Protein–Protein Recognition in Solution Using the Coarse-Grained Force Field SCORPION. *J. Chem. Theory Comput.*, 9(1):803–813.

[10] Baumgarth, N., Tung, J. W., and Herzenberg, L. A. (2005). Inherent specificities in natural antibodies: a key to immune defense against pathogen invasion. *Springer Semin. Immunopathol.*, 26(4):347–62.

[11] Bee, J. S., Davis, M., Freund, E., Carpenter, J. F., and Randolph, T. W. (2010). Aggregation of a monoclonal antibody induced by adsorption to stainless steel. *Biotechnol. Bioeng.*, 105(1):121–129.

[12] Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684.

[13] Berstein, R. M., Schluter, S. F., Shen, S., and Marchalonis, J. J. (1996). A new high molecular weight immunoglobulin class from the carcharhine shark: implications for the properties of the primordial immunoglobulin. *Proc. Natl. Acad. Sci. U. S. A.*, 93(8):3289–3293.

[14] Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280:1–9.

[15] Bond, P. J., Holyoake, J., Ivetac, A., Khalid, S., and Sansom, M. S. P. (2007). Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.*, 157(3):593–605.

[16] Bond, P. J. and Sansom, M. S. P. (2006). Insertion and assembly of membrane proteins via simulation. *J. Am. Chem. Soc.*, 128(8):2697–2704.

[17] Borrok, M. J., Jung, S. T., Kang, T. H., Monzingo, A. F., and Georgiou, G. (2012). Revisiting the role of glycosylation in the structure of human IgG Fc. *ACS Chem. Biol.*, 7(9):1596–602.

[18] Bowman, G. R., Beauchamp, K. A., Boxer, G., and Pande, V. S. (2009a). Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.*, 131(12):124101.

[19] Bowman, G. R., Huang, X., and Pande, V. S. (2009b). Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods*, 49(2):197–201.

[20] Brandenburg, K. and Holst, O. (2001). Glycolipids: Distribution and Biological Function. *Life Sci.*, pages 1–9.

[21] Brych, S. R., Gokarn, Y. R., Hultgen, H., Stevenson, R. J., Rajan, R., and Matsumura, M. (2010). Characterization of antibody aggregation: role of buried, unpaired cysteines in particle formation. *J. Pharm. Sci.*, 99(2):764–81.

[22] Buccoliero, R., Bodennec, J., and Futerman, A. H. (2002). The Role of Sphingolipids in Neuronal Development: Lessons from Models of Sphingolipid Storage Diseases. *Neurochem. Res.*, 27(7-8):565–574.

[23] Cantù, L., Del Favero, E., Sonnino, S., and Prinetti, A. (2011). Gangliosides and the multiscale modulation of membrane structure. *Chem. Phys. Lipids*, 164(8):796–810.

[24] Carter, P. J. (2011). Introduction to current and future protein therapeutics: a protein engineering perspective. *Exp. Cell Res.*, 317(9):1261–1269.

[25] Case, D., Darden, T., Cheatham, T., Simmerling, C., Wang, J., Duke, R., Luo, R., Walker, R., Zhang, W., Merz, K., Roberts, B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Goetz, A., Kolossváry, I., Wong, K., Paesani, F., Vanicek, J., Wolf, R., Liu, J., Wu, X., Brozell, S., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Hsieh, M., Cui, G., Roe, D., Mathews, D., Seetin, M., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. (2012). AMBER 12.

[26] Case, D. A., Darden, T. A., Cheatham T.E., I. I. I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B. P., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvai, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. (2010). AMBER 11. *Univ. California, San Fr.*

[27] Chaudhri, A., Zarraga, I. E., Kamerzell, T. J., Brandt, J. P., Patapoff, T. W., Shire, S. J., and Voth, G. A. (2012). Coarse-grained modeling of the self-association of therapeutic monoclonal antibodies. *J. Phys. Chem. B*, 116(28):8045–8057.

[28] Chennamsetty, N., Helk, B., Voynov, V., Kayser, V., and Trout, B. L. (2009a). Aggregation-prone motifs in human immunoglobulin G. *J. Mol. Biol.*, 391(2):404–413.

[29] Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B. L. (2009b). Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci. U. S. A.*, 106(29):11937–11942.

[30] Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., and Trout, B. L. (2010a). Prediction of aggregation prone regions of therapeutic proteins. *J. Phys. Chem. B*, 114(19):6614–6624.

[31] Chennamsetty, N., Voynov, V., Kayser, V., Schneider, C., Shulka, D., Helk, B., Trout, B. L., Wood, G., Bondarenko, P., Dillon, T., and Ricci, M. (2010b). Modeling Protein Degradation Processes and the Development of Rational Approaches to Stabilization. *Engineering*.

[32] Clark, M. R. (1997). IgG effector mechanisms. *Chem. Immunol.*, pages 1–31.

[33] Cooper, M. D. and Alder, M. N. (2006). The evolution of adaptive immune systems. *Cell*, 124(4):815–22.

[34] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197.

[35] Coskun, U., Grzybek, M., Drechsel, D., and Simons, K. (2011). Regulation of human EGF receptor by lipids. *Proc. Natl. Acad. Sci. U. S. A.*, 108(22):9044–8.

[36] Crispin, M., Yu, X., and Bowden, T. A. (2013). Crystal structure of sialylated IgG Fc: implications for the mechanism of intravenous immunoglobulin therapy. *Proc. Natl. Acad. Sci. U. S. A.*, 110(38).

[37] Dalziel, M., Crispin, M., Scanlan, C. N., Zitzmann, N., and Dwek, R. A. (2014). Emerging principles for the therapeutic exploitation of glycosylation. *Science*, 343(6166):1235681.

[38] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089.

[39] Dart, C. (2010). Lipid microdomains and the regulation of ion channel function. *J. Physiol.*, 588(Pt 17):3169–78.

[40] Daugherty, A. L. and Mrsny, R. J. (2006). Formulation and delivery issues for monoclonal antibody therapeutics. *Adv. Drug Deliv. Rev.*, 58(5-6):686–706.

[41] De Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., and Marrink, S. J. (2013). Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.*, 9:687–697.

[42] de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., and Marrink, S. J. (2013). Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.*, 9(1):687–697.

[43] DeMarco, M. L. and Woods, R. J. (2008). Structural glycobiology: a game of snakes and ladders. *Glycobiology*, 18(6):426–440.

[44] DeMarco, M. L. and Woods, R. J. (2009a). Atomic-resolution conformational analysis of the GM3 ganglioside in a lipid bilayer and its implications for ganglioside-protein recognition at membrane surfaces. *Glycobiology*, 19(4):344–355.

[45] DeMarco, M. L. and Woods, R. J. (2009b). Atomic-resolution conformational analysis of the GM3 ganglioside in a lipid bilayer and its implications for ganglioside-protein recognition at membrane surfaces. *Glycobiology*, 19(4):344–55.

[46] Dillman, R. O. (1997). Magic bullets at last! Finally–approval of a monoclonal antibody for the treatment of cancer!!! *Cancer Biother. Radiopharm.*, 12(4):223–5.

[47] Dong, X., Storkus, W. J., and Salter, R. D. (1999). Binding and uptake of agalactosyl IgG by mannose receptor on macrophages and dendritic cells. *J. Immunol.*, 163(10):5427–34.

[48] Dove, A. (2001). The bittersweet promise of glycobiology. *Nat. Biotechnol.*, 19(10):913–7.

[49] Dube, D. H. and Bertozzi, C. R. (2005). Glycans in cancer and inflammation–potential for therapeutics and diagnostics. *Nat. Rev. Drug Discov.*, 4(6):477–88.

[50] Dwek, R. A. (1996). Glycobiology: Toward Understanding the Function of Sugars. *Chem. Rev.*, 96(2):683–720.

[51] Dwek, R. A., Lellouch, A. C., and Wormald, M. R. (1995). Glycobiology: 'the function of sugar in the IgG molecule'. *J. Anat.*, 187 (Pt 2:279–92.

[52] Edidin, M. (2003). The state of lipid rafts: from model membranes to cells. *Annu. Rev. Biophys. Biomol. Struct.*, 32:257–83.

[53] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.*

[54] Ekiert, D. C., Bhabha, G., Elsliger, M.-A., Friesen, R. H. E., Jongeneelen, M., Throsby, M., Goudsmit, J., and Wilson, I. A. (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science*, 324(5924):246–51.

[55] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577.

[56] Feige, M. J., Gräwert, M. A., Marcinowski, M., Hennig, J., Behnke, J., Ausländer, D., Herold, E. M., Peschek, J., Castro, C. D., Flajnik, M., Hendershot, L. M., Sattler, M., Groll, M., and Buchner, J. (2014). The structural analysis of shark IgNAR antibodies reveals evolutionary principles of immunoglobulins. *Proc. Natl. Acad. Sci. U. S. A.*, 111(22):8155–60.

[57] Feige, M. J., Nath, S., Catharino, S. R., Weinfurtner, D., Steinbacher, S., and Buchner, J. (2009). Structure of the murine unglycosylated IgG1 Fc fragment. *J. Mol. Biol.*, 391(3):599–608.

[58] Flajnik, M. F. and Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.*, 11(1):47–59.

[59] Fowler, D. M., Koulov, A. V., Balch, W. E., and Kelly, J. W. (2007). Functional amyloid–from bacteria to humans. *Trends Biochem. Sci.*, 32(5):217–24.

[60] Frank, M. and Schloissnig, S. (2010). Bioinformatics and molecular modeling in glycobiology. *Cell. Mol. Life Sci.*, 67(16):2749–72.

[61] Furukawa, K., Ohkawa, Y., Yamauchi, Y., Hamamura, K., Ohmi, Y., and Furukawa, K. (2012). Fine tuning of cell signals by glycosylation. *J. Biochem.*, 151(6):573–8.

[62] Gerl, M. J., Sampaio, J. L., Urban, S., Kalvodova, L., Verbavatz, J.-M., Binnington, B., Lindemann, D., Lingwood, C. A., Shevchenko, A., Schroeder, C., and Simons, K. (2012). Quantitative analysis of the lipidomes of the influenza virus envelope and MDCK cell apical membrane. *J. Cell Biol.*, 196(2):213–21.

[63] Ghazarian, H., Idoni, B., and Oppenheimer, S. B. (2011). A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. *Acta Histochem.*, 113(3):236–47.

[64] González-Díaz, H., Cruz-Monteagudo, M., Molina, R., Tenorio, E., and Uriarte, E. (2005). Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. *Bioorg. Med. Chem.*, 13(4):1119–29.

[65] Hakomori, Senitiroh MD, and PhD (2003). Structure, organization, and function of glycosphingolipids in membrane. *Curr. Opin. Hematol.*, 10(1):16–24.

[66] Hakomori, S. (1990). Bifunctional role of glycosphingolipids. Modulators for transmembrane signaling and mediators for cellular interactions. *J. Biol. Chem.*, 265(31):18713–18716.

[67] Hall, A., Róg, T., Karttunen, M., and Vattulainen, I. (2010). Role of glycolipids in lipid rafts: a view through atomistic molecular dynamics simulations with galactosylceramide. *J. Phys. Chem. B*, 114(23):7797–807.

[68] Hall, B. A., Armitage, J. P., and Sansom, M. S. P. (2011). Transmembrane Helix Dynamics of Bacterial Chemoreceptors Supports a Piston Model of Signalling. *PLoS Comput. Biol.*, 7(10).

[69] Hall, B. A., Halim, K. B. A., Buyan, A., Emmanouil, B., and Sansom, M. S. P. (2014). Sidekick for Membrane Simulations: Automated Ensemble Molecular Dynamics Simulations of Transmembrane Helices. *J. Chem. Theory Comput.*, 10(5):2165–2175.

[70] Hall, B. a. and Sansom, M. S. P. (2009). Coarse-Grained MD Simulations and Protein-Protein Interactions: The Cohesin-Dockerin System. *J. Chem. Theory Comput.*, 5(9):2465–2471.

[71] Hari, S. B., Lau, H., Razinkov, V. I., Chen, S., and Latypov, R. F. (2010). Acid-Induced Aggregation of Human Monoclonal IgG1 and IgG2: Molecular Mechanism and the Effect of Solution Composition. *Biochemistry*, pages 9328–9338.

[72] Hemmrich, G., Miller, D. J., and Bosch, T. C. G. (2007). The evolution of immunity: a low-life perspective. *Trends Immunol.*, 28(10):449–54.

[73] Hess, B. (2008). P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.*, 4(1):116–122.

[74] Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.*, 4(3):435–447.

[75] Hockney, R., Goel, S., and Eastwood, J. (1974). Quiet high resolution computer models of a plasma. *J. Comput. Phys.*, 14(2):148–158.

[76] Holleran, W., Takagi, Y., Menon, G., Jackson, S., Lee, J., Feingold, K., and Elias, P. (1994). Permeability barrier requirements regulate epidermal beta- glucocerebrosidase. *J. Lipid Res.*, 35(5):905–912.

[77] Hoogenboom, H. R. (2005). Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.*, 23(9):1105–1116.

[78] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–25.

[79] Hruska, K. S., LaMarca, M. E., Scott, C. R., and Sidransky, E. (2008a). Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum. Mutat.*, 29(5):567–83.

[80] Hruska, K. S., LaMarca, M. E., Scott, C. R., and Sidransky, E. (2008b). Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum. Mutat.*, 29(5):567–83.

[81] Huang, X., Yao, Y., Bowman, G. R., Sun, J., Guibas, L. J., Carlsson, G., and Pande, V. S. (2010). Constructing multi-resolution Markov State Models (MSMs) to elucidate RNA hairpin folding mechanisms. *Pac. Symp. Biocomput.*, pages 228–39.

[82] Hudson, P. J. and Souriau, C. (2003). Engineered antibodies. *Nat. Med.*, 9(1):129–34.

[83] Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.*, 14(1):33–8, 27–8.

[84] Hwang, W. Y. K. and Foote, J. (2005). Immunogenicity of engineered antibodies. *Methods San Diego Calif*, 36(1):3–10.

[85] Inokuchi, J.-i. (2010). Membrane microdomains and insulin resistance. *FEBS Lett.*, 584(9):1864–71.

[86] Ishikawa, T., Ito, T., Endo, R., Nakagawa, K., Sawa, E., and Wakamatsu, K. (2010). Influence of pH on heat-induced aggregation and degradation of therapeutic monoclonal antibodies. *Biol. Pharm. Bull.*, 33(8):1413–1417.

[87] Iwabuchi, K., Yamamura, S., Prinetti, A., Handa, K., and Hakomori, S.-i. (1998). GM3-enriched Microdomain Involved in Cell Adhesion and Signal Transduction through Carbohydrate-Carbohydrate Interaction in Mouse Melanoma B16 Cells. *J. Biol. Chem.*, 273(15):9130–9138.

[88] Jacob, J. C., Kutty, K. M., Islam, M., Dominic, R. G., and Dawson, G. (1973). Krabbe's disease: globoid cell leukodystrophy. *Can. Med. Assoc. J.*, 108(11):1398–400.

[89] Janeway, C. A. and Medzhitov, R. (2002). Innate immune recognition. *Annu. Rev. Immunol.*, 20:197–216.

[90] Jedlovszky, P., Sega, M., and Vallauri, R. (2009). GM1 ganglioside embedded in a hydrated DOPC membrane: a molecular dynamics simulation study. *J. Phys. Chem. B*, 113(14):4876–86.

[91] Jefferis, R. (2009). Glycosylation as a strategy to improve antibody-based therapeutics. *Nat. Rev. Drug Discov.*, 8(3):226–34.

[92] Jenssen, H. v., Hamill, P., and Hancock, R. E. W. (2006). Peptide antimicrobial agents. *Clin. Microbiol. Rev.*, 19(3):491–511.

[93] Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236.

[94] Jung, S. T., Reddy, S. T., Kang, T. H., Borrok, M. J., Sandlie, I., Tucker, P. W., and Georgiou, G. (2010). Aglycosylated IgG variants expressed in bacteria that selectively bind FcgammaRI potentiate tumor cell killing by monocyte-dendritic cells. *Proc. Natl. Acad. Sci. U. S. A.*, 107(2):604–9.

[95] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

[96] Kanai, S., Liu, J., Patapoff, T. W., and Shire, S. J. (2008). Reversible self-association of a concentrated monoclonal antibody solution mediated by Fab-Fab interaction that impacts solution viscosity. *J. Pharm. Sci.*, 97(10):4219–4227.

[97] Kayser, V., Chennamsetty, N., Voynov, V., Forrer, K., Helk, B., and Trout, B. L. (2010). Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies. *Biotechnol. J.*, pages 1–7.

[98] Kelliher, M. T., Jacks, R. D., Piraino, M. S., and Southern, C. A. (2014). The effect of sugar removal on the structure of the Fc region of an IgG antibody as observed with single molecule Förster Resonance Energy Transfer. *Mol. Immunol.*, 60(2):103–8.

[99] Kiese, S., Papppenberger, A., Friess, W., and Mahler, H.-C. (2008). Shaken, not stirred: mechanical stress testing of an IgG1 antibody. *J. Pharm. Sci.*, 97(10):4347–4366.

[100] Kindt, T. J., Goldsby, R. A., Osborne, B. A., and Kuby, J. (2007). *Kuby Immunology*. W. H. Freeman.

[101] Kinjo, A. R. and Takada, S. (2003). Competition between protein folding and aggregation with molecular chaperones in crowded solutions: insight from mesoscopic simulations. *Biophys. J.*, 85(6):3521–31.

[102] Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., and Woods, R. J. (2008). GLYCAM06: A generalizable biomolecular force field. carbohydrates. *J. Comput. Chem.*, 29:622–655.

[103] Koldsø, H., Shorthouse, D., Hélie, J., and Sansom, M. S. P. (2014). Lipid Clustering Correlates with Membrane Curvature as Revealed by Molecular Simulations of Complex Lipid Bilayers. *PLoS Comput. Biol.*, 10(10):e1003911.

[104] Kony, D., Damm, W., Stoll, S., and Van Gunsteren, W. F. (2002). An improved OPLS-AA force field for carbohydrates. *J. Comput. Chem.*, 23(15):1416–1429.

[105] Kozlowski, S. and Swann, P. (2006). Current and future issues in the manufacturing and development of monoclonal antibodies. *Adv. Drug Deliv. Rev.*, 58(5-6):707–22.

[106] Krapp, S., Mimura, Y., Jefferis, R., Huber, R., and Sondermann, P. (2003). Structural analysis of human IgG-Fc glycoforms reveals a correlation between glycosylation and structural integrity. *J. Mol. Biol.*, 325(5):979–89.

[107] Lauc, G., Krištić, J., and Zoldoš, V. (2014). Glycans - the third revolution in evolution. *Front. Genet.*, 5:145.

[108] Lauer, T. M., Agrawal, N. J., Chennamsetty, N., Egodage, K., Helk, B., and Trout, B. L. (2012). Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J. Pharm. Sci.*, 101(1):102–115.

[109] Laws, P. M. and Warren, R. B. (2011). Ustekinumab for the treatment of psoriasis. *Expert Rev. Clin. Immunol.*, 7(2):155–64.

[110] Leach, A. (2001). *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall.

[111] Leung, D. Y. M., Sampson, H. A., Yunginger, J. W., Burks, A. W., Schneider, L. C., Wortel, C. H., Davis, F. M., Hyun, J. D., and Shanahan, W. R. (2003). Effect of anti-IgE therapy in patients with peanut allergy. Technical Report 11, National Jewish Medical and Research Center, Denver, USA.

[112] Li, Y., Xiang, Q., Zhang, Q., Huang, Y., and Su, Z. (2012). Overview on the recent study of antimicrobial peptides: origins, functions, relative mechanisms and application. *Peptides*, 37(2):207–15.

[113] Lindahl, E., Hess, B., and Van Der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7(8):306–317.

[114] Lins, R. D. and Hünenberger, P. H. (2005). A new GROMOS force field for hexopyranose-based carbohydrates. *J. Comput. Chem.*, 26(13):1400–12.

[115] Lis, H. and Sharon, N. (1993). Protein glycosylation. Structural and functional aspects. *Eur. J. Biochem.*, 218(1):1–27.

[116] Litman, G. W., Anderson, M. K., and Rast, J. P. (1999). Evolution of antigen binding receptors. *Annu. Rev. Immunol.*, 17:109–47.

[117] Lopez, C. A., Rzepiela, A. J., de Vries, A. H., Dijkhuizen, L., Hunenberger, P. H., and Marrink, S. J. (2009). Martini Coarse-Grained Force Field: Extension to Carbohydrates. *J. Chem. Theory Comput.*, 5(12):3195–3210.

[118] López, C. A., Sovova, Z., van Eerden, F. J., de Vries, A. H., and Marrink, S. J. (2013). Martini Force Field Parameters for Glycolipids. *J. Chem. Theory Comput.*, 9(3):1694–1708.

[119] Low-Nam, S. T., Lidke, K. A., Cutler, P. J., Roovers, R. C., van Bergen En Henegouwen, P. M. P., Wilson, B. S., and Lidke, D. S. (2011). ErbB1 dimerization is promoted by domain co-confinement and stabilized by ligand binding. *Nat. Struct. Mol. Biol.*, 18(11):1244–1249.

[120] Lowe, J. B. and Marth, J. D. (2003). A genetic approach to Mammalian glycan function. *Annu. Rev. Biochem.*, 72:643–91.

[121] Lwin, A. (2004). Glucocerebrosidase mutations in subjects with parkinsonism. *Mol. Genet. Metab.*, 81(1):70–73.

[122] Mandal, T. K. and Mukhopadhyay, C. (2001). Effect of glycosylation on structure and dynamics of MHC class I glycoprotein: a molecular dynamics study. *Biopolymers*, 59(1):11–23.

[123] Marrink, S. J., de Vries, A. H., and Mark, A. E. (2004). Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B*, 108(2):750–760.

[124] Matsumiya, S., Yamaguchi, Y., Saito, J.-i., Nagano, M., Sasakawa, H., Otaki, S., Satoh, M., Shitara, K., and Kato, K. (2007). Structural comparison of fucosylated and nonfucosylated Fc fragments of human immunoglobulin G1. *J. Mol. Biol.*, 368(3):767–79.

[125] Mayforth, R. and Quintans, J. (1990). Designer and Catalytic Antibodies — NEJM. *N. Engl. J. Med.*, (323):173–178.

[126] Meier, S. and Duus, J. (2011). Carbohydrate dynamics: Antibody glycans wiggle and jiggle. *Nat. Chem. Biol.*, 7(3):131–132.

[127] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011a). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32:2319–2327.

[128] Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011b). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32(10):2319–2327.

[129] Mirazimi, A. and Svensson, L. (1998). Carbohydrates facilitate correct disulfide bond formation and folding of rotavirus VP7. *J. Virol.*, 72(5):3887–92.

[130] Mizushima, T., Yagi, H., Takemoto, E., Shibata-Koyama, M., Isoda, Y., Iida, S., Masuda, K., Satoh, M., and Kato, K. (2011). Structural basis for improved efficacy of therapeutic antibodies on defucosylation of their Fc glycans. *Genes Cells*, 16(11):1071–80.

[131] Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.*, 4(5):819–834.

[132] Nadler, L. M., Stashenko, P., Hardy, R., Kaplan, W. D., Button, L. N., Kufe, D. W., Antman, K. H., and Schlossman, S. F. (1980). Serotherapy of a Patient with a Monoclonal Antibody Directed against a Human Lymphoma-associated Antigen. *Cancer Res.*, 40(9):3147–3154.

[133] Neutra, M. (1966). Radioautographic comparison of the uptake of galactose-H3 in the golgi region of various cells secreting glycoproteins or mucopolysaccharides. *J. Cell Biol.*, 30(1):137–150.

[134] Nimmerjahn, F. and Ravetch, J. V. (2008). Fcgamma receptors as regulators of immune responses. *Nat. Rev. Immunol.*, 8(1):34–47.

[135] Noé, F. and Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.*, 18(2):154–62.

[136] Oldham, R. K. and Dillman, R. O. (2008). Monoclonal antibodies in cancer therapy: 25 years of progress. *J. Clin. Oncol.*, 26(11):1774–7.

[137] O'Mahony, D. and Bishop, M. R. (2006). Monoclonal antibody therapy. *Front. Biosci.*, 11:1620–1635.

[138] Ono, K., Hattori, H., Uemura, K., Nakayama, J., Ota, H., and Katsuyama, T. (1994). Expression of Forssman antigen in human large intestine. *J. Histochem. Cytochem.*, 42(5):659–65.

[139] Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25(13):1656–1676.

[140] Padlan, E. A. (1994). Anatomy of the antibody molecule. *Mol. Immunol.*, 31(3):169–217.

[141] Pancer, Z., Amemiya, C. T., Ehrhardt, G. R. A., Ceitlin, J., Gartland, G. L., and Cooper, M. D. (2004). Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature*, 430(6996):174–80.

[142] Pande, V. S., Beauchamp, K., and Bowman, G. R. (2010). Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105.

[143] Peng, R. D. (2011). Reproducible Research in Computational Science. *Science (80-.).*, 334(6060):1226–1227.

[144] Perchiacca, J. M. and Tessier, P. M. (2012). Engineering Aggregation-Resistant Antibodies. *Annu. Rev. Chem. Biomol. Eng.*, 3(March):263–286.

[145] Periole, X., Cavalli, M., Marrink, S.-J., and Ceruso, M. A. (2009). Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.*, 5(9):2531–2543.

[146] Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., Van Der Spoel, D., Hess, B., and Lindahl, E. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29:845–854.

[147] Radaev, S., Motyka, S., Fridman, W. H., Sautes-Fridman, C., and Sun, P. D. (2001). The structure of a human type III Fcgamma receptor in complex with Fc. *J. Biol. Chem.*, 276(19):16469–77.

[148] Rajendran, L. and Simons, K. (2005). Lipid rafts and membrane dynamics. *J. Cell Sci.*, 118(Pt 6):1099–102.

[149] Reynders, E., Foulquier, F., Annaert, W., and Matthijs, G. (2011). How Golgi glycosylation meets and needs trafficking: the case of the COG complex. *Glycobiology*, 21(7):853–63.

[150] Richards, W. G. (2002). Virtual screening using grid computing: the screensaver project. *Nat. Rev. Drug Discov.*, 1(7):551–5.

[151] Rodríguez, R. M., López-Vázquez, A., and López-Larrea, C. (2012). Immune systems evolution. *Adv. Exp. Med. Biol.*, 739:237–51.

[152] Roguska, M. A., Pedersen, J. T., Keddy, C. A., Henry, A. H., Searle, S. J., Lambert, J. M., Goldmacher, V. S., Blattler, W. A., Rees, A. R., and Guild, B. C. (1994). Humanization of murine monoclonal antibodies through variable domain resurfacing. *Proc. Natl. Acad. Sci. U. S. A.*, 91(3):969–973.

[153] Roque, A. C. A., Lowe, C. R., and Taipa, M. A. (2004). Antibodies and genetically engineered related molecules: production and purification. *Biotechnol. Prog.*, 20(3):639–654.

[154] Roseman, S. (2001). Reflections on glycobiology. *J. Biol. Chem.*, 276(45):41527–42.

[155] Rosenberg, A. S. (2006). Effects of protein aggregates: an immunologic perspective. *AAPS J.*, 8(3):501–507.

[156] Rosenstiel, P., Philipp, E. E. R., Schreiber, S., and Bosch, T. C. G. (2009). Evolution and function of innate immune receptors–insights from marine invertebrates. *J. Innate Immun.*, 1(4):291–300.

[157] Rudd, P. M. and Dwek, R. A. (1997). Glycosylation: Heterogeneity and the 3D structure of proteins. *Crit. Rev. Biochem. Mol. Biol.*, 32(1):1–100.

[158] Sahin, E., Grillo, A. O., Perkins, M. D., and Roberts, C. J. (2010). Comparative Effects of pH and Ionic Strength on Protein – Protein Interactions , Unfolding , and Aggregation for IgG1 Antibodies. *Sci. York*, 99(12):4830–4848.

[159] Salwinski, L. and Eisenberg, D. (2003). Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 13(3):377–82.

[160] Sanchez de Groot, N., Torrent, M., Villar-Piqué, A., Lang, B., Ventura, S., Gsponer, J., and Babu, M. M. (2012). Evolutionary selection for protein aggregation. *Biochem. Soc. Trans.*, 40(5):1032–7.

[161] Sarich, M., Noé, F., and Schütte, C. (2010). On the Approximation Quality of Markov State Models. *Multiscale Model. Simul.*, 8(4):1154–1177.

[162] Sastry, P. S. (1985). Lipids of nervous tissue: composition and metabolism. *Prog. Lipid Res.*, 24(2):69–176.

[163] Schroeder, H. W. and Cavacini, L. (2010). Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.*, 125(2 Suppl 2):S41–52.

[164] Schwab, I. and Nimmerjahn, F. (2013). Intravenous immunoglobulin therapy: how does IgG modulate the immune system? *Nat. Rev. Immunol.*, 13(3):176–89.

[165] Sega, M., Vallauri, R., Brocca, P., and Melchionna, S. (2004). Molecular Dynamics Simulation of a GM3 Ganglioside Bilayer. *J. Phys. Chem. B*, 108(52):20322–20330.

[166] Senne, M., Trendelkamp-Schroer, B., Mey, A. S., Schütte, C., and Noé, F. (2012). EMMA: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput.*, 8(7):2223–2238.

[167] Sharma, B. (2007). Immunogenicity of therapeutic proteins. Part 3: impact of manufacturing changes. *Biotechnol. Adv.*, 25(3):325–331.

[168] Sharon, N. and Lis, H. (1993). Carbohydrates in Cell Recognition. *Sci. Am.*, 268(1):82–89.

[169] Shental-Bechor, D. and Levy, Y. (2008). Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc. Natl. Acad. Sci. U. S. A.*, 105(24):8256–61.

[170] Shirts, M. and Pande, V. S. (2000). COMPUTING: Screen Savers of the World Unite! *Science*, 290(5498):1903–4.

[171] Siddiqui, B. and Hakomori, S.-I. (1971). A Revised Structure for the Forssman Glycolipid Hapten. *J. Biol. Chem.*, 246(18):5766–5769.

[172] Sigvard Olofsson, J.-E. S. H. (1998). Host Cell Glycosylation of Viral Glycoproteins - a Battlefield for Host Defence and Viral Resistance. *Scand. J. Infect. Dis.*, 30(5):435–440.

[173] Sø rensen, J., Periole, X., Skeby, K. K., Marrink, S.-J., and Schiø tt, B. (2011). Protofibrillar Assembly Toward the Formation of Amyloid Fibrils. *J. Phys. Chem. Lett.*, 2(19):2385–2390.

[174] Sondermann, P., Pincetic, A., Maamary, J., Lammens, K., and Ravetch, J. V. (2013a). General mechanism for modulating immunoglobulin effector function. *Proc. Natl. Acad. Sci. U. S. A.*, 110(24):9868–72.

[175] Sondermann, P., Pincetic, A., Maamary, J., Lammens, K., and Ravetch, J. V. (2013b). Reply to Crispin et al.: Molecular model that accounts for the biological and physical properties of sialylated Fc. *Proc. Natl. Acad. Sci. U. S. A.*, 110(38):E3547–E3547.

[176] Sonnino, S., Mauri, L., Chigorno, V., and Prinetti, A. (2007). Gangliosides as components of lipid membrane domains. *Glycobiology*, 17(1):1R———13R.

[177] Sperling, P., Franke, S., Lüthje, S., and Heinz, E. (2005). Are glucocerebrosides the predominant sphingolipids in plant plasma membranes? *Plant Physiol. Biochem.*, 43(12):1031–8.

[178] Springer, S. A. and Gagneux, P. (2013). Glycan evolution in response to collaboration, conflict, and constraint. *J. Biol. Chem.*, 288(10):6904–11.

[179] Stansfeld, P. J. and Sansom, M. S. P. (2011). From coarse grained to atomistic: A serial multiscale approach to membrane protein simulations. *J. Chem. Theory Comput.*, 7:1157–1166.

[180] Stark, A. C., Andrews, C. T., and Elcock, A. H. (2013). Toward optimized potential functions for protein-protein interactions in aqueous solutions: Osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J. Chem. Theory Comput.*, 9:4176–4185.

[181] Steiner, L. A. (1996). Immunoglobulin evolution, 30 years on. *Glycobiology*, 6(7):649–656.

[182] Stoffel, W. and Bosio, A. (1997). Myelin glycolipids and their functions. *Curr. Opin. Neurobiol.*, 7(5):654–61.

[183] Strebhardt, K. and Ullrich, A. (2008). Paul Ehrlich's magic bullet concept: 100 years of progress. *Nat. Rev. Cancer*, 8(6):473–80.

[184] Sun, J., Yao, Y., Huang, X., and Pande, V. (2012). A Fast Geometric Clustering Method on Conformation Space of Biomolecules. *learning*, 24(2).

[185] Suzuki, K. and Suzuki, Y. (1970). Globoid cell leucodystrophy (Krabbe's disease): deficiency of galactocerebroside beta-galactosidase. *Proc. Natl. Acad. Sci. U. S. A.*, 66(2):302–9.

[186] Taylor, M. E. and Drickamer, K. (2007). Paradigms for glycan-binding receptors in cell adhesion. *Curr. Opin. Cell Biol.*, 19(5):572–7.

[187] Tennessen, J. A. (2005). Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *J. Evol. Biol.*, 18(6):1387–1394.

[188] Vainio, S., Heino, S., Mansson, J.-E., Fredman, P., Kuismanen, E., Vaarala, O., and Ikonen, E. (2002). Dynamic association of human insulin receptor with lipid rafts in cells lacking caveolae. *EMBO Rep.*, 3(1):95–100.

[189] Varki, A. (2011). Evolutionary forces shaping the Golgi glycosylation machinery: why cell surface glycans are universal to living cells. *Cold Spring Harb. Perspect. Biol.*, 3(6).

[190] Vázquez-Rey, M. and Lang, D. A. (2011). Aggregates in monoclonal antibody manufacturing processes. *Biotechnol. Bioeng.*, 108(7):1494–508.

[191] Vermeer, a. W. and Norde, W. (2000). The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. *Biophys. J.*, 78(1):394–404.

[192] Vlasak, J. and Ionescu, R. (2011). Fragmentation of monoclonal antibodies. *MAbs*, 3(3):354–366.

[193] Vollmer, W., Blanot, D., and de Pedro, M. A. (2008). Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.*, 32(2):149–67.

[194] Voynov, V., Chennamsetty, N., Kayser, V., Helk, B., Forrer, K., Zhang, H., Fritsch, C., Heine, H., and Trout, B. L. (2009). Dynamic fluctuations of protein-carbohydrate interactions promote protein aggregation. *PLoS One*, 4(12):e8425.

[195] Voynov, V., Chennamsetty, N., Kayser, V., Wallny, H.-J., Helk, B., and Trout, B. L. (2010). Design and application of antibody cysteine variants. *Bioconjug. Chem.*, 21(2):385–392.

[196] Wachholz, Petra, A., Durham, and Stephen, R. (2004). Current Opinion in Allergy and Clinical Immunology. *Curr. Opin. Allergy Clin. Immunol.*, 4(4):313–318.

[197] Wang, J., Cieplak, P., and Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074.

[198] Wang, W., Singh, S., Zeng, D. L., and King, K. (2007). Antibody structure, instability, and formulation. *J. Pharm. Sci.*, 96(1):1–26.

[199] Wang, X., Das, T. K., Singh, S. K., and Kumar, S. (2009). Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. *MAbs*, 1(3):254–267.

[200] Wiens, M., Korzhev, M., Perovic-Ottstadt, S., Luthringer, B., Brandt, D., Klein, S., and Müller, W. E. G. (2007). Toll-like receptors are part of the innate immune defense system of sponges (demospongiae: Porifera). *Mol. Biol. Evol.*, 24(3):792–804.

[201] Woof, J. M. and Burton, D. R. (2004). Human antibody-Fc receptor interactions illuminated by crystal structures. *Nat. Rev. Immunol.*, 4(2):89–99.

[202] Wu, C. and Shea, J.-E. (2011). Coarse-grained models for protein aggregation. *Curr. Opin. Struct. Biol.*, 21(2):209–220.

[203] Wu, H., Kroe-Barrett, R., Singh, S., Robinson, A. S., and Roberts, C. J. (2014). Competing aggregation pathways for monoclonal antibodies. *FEBS Lett.*, 588(6):936–941.

[204] Wyss, D. F. and Wagner, G. (1996). The structural role of sugars in glycoproteins. *Curr. Opin. Biotechnol.*, 7(4):409–416.

[205] Xu, Y.-H., Barnes, S., Sun, Y., and Grabowski, G. A. (2010). Multi-system disorders of glycosphingolipid and ganglioside metabolism. *J. Lipid Res.*, 51(7):1643–75.

[206] Yamakawa, T. (1996). A reflection on the early history of glycosphingolipids. *Glycoconj. J.*, 13(2):123–6.

[207] Yamamoto, M., Cid, E., and Yamamoto, F. (2012). Molecular genetic basis of the human Forssman glycolipid antigen negativity. *Sci. Rep.*, 2:975.

[208] Yamashita, T., Hashiramoto, A., Haluzik, M., Mizukami, H., Beck, S., Norton, A., Kono, M., Tsuji, S., Daniotti, J. L., Werth, N., Sandhoff, R., Sandhoff, K., and Proia, R. L. (2003). Enhanced insulin sensitivity in mice lacking ganglioside GM3. *Proc. Natl. Acad. Sci. U. S. A.*, 100(6):3445–9.

[209] Yoshikawa, M., Go, S., Takasaki, K., Kakazu, Y., Ohashi, M., Nagafuku, M., Kabayama, K., Sekimoto, J., Suzuki, S.-i., Takaiwa, K., Kimitsuki, T., Matsumoto, N., Komune, S., Kamei, D., Saito, M., Fujiwara, M., Iwasaki, K., and Inokuchi, J.-i. (2009). Mice lacking ganglioside GM3 synthase exhibit complete hearing loss due to selective degeneration of the organ of Corti. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9483–8.

[210] Yu, R. K., Tsai, Y.-T., and Ariga, T. (2012). Functional roles of gangliosides in neurodevelopment: an overview of recent advances. *Neurochem. Res.*, 37(6):1230–44.

[211] Zettlitz, K. a., Seitter, J., Müller, D., and Kontermann, R. E. (2010). Humanization of a mouse monoclonal antibody directed against a cell surface-exposed epitope of membrane-associated heat shock protein 70 (Hsp70). *Mol. Biotechnol.*, 46(3):265–278.

[212] Zoldoš, V., Novokmet, M., Bečeheli, I., and Lauc, G. (2013). Genomics and epigenomics of the human glycome. *Glycoconj. J.*, 30(1):41–50.

[213] Zuegg, J. and Gready, J. E. (2000). Molecular dynamics simulation of human prion protein including both N-linked oligosaccharides and the GPI anchor. *Glycobiology*, 10(10):959–974.

Appendix A

Appendix

This appendix presents supplemental data for Chapter 4 of the thesis. Specifically an example of some of the code used for the automated setup of high-throughput simulations can be seen.

A.1 Script for the Automation of Simulation Setup

[fragile]

```
#! /usr/bin/python
import sys
import os
import numpy
import math
import re
from os.path import basename
import argparse
sys.path.append(r"/sansom/sc3/wolf3482/AUTOMATION/Dependancies")
from gromacscom import *
```

```
import subprocess
from random import *
import time

#Take inputs from sys.argv - if CG structure, take .itp and pass on to next stage (sy
#If AT structure - Coarsegrain using MARTINIZE.py, then take .itp and .top and pass o

#Variables:

dssppath = "/sbcb/packages/opt/Linux_x86_64/dssp/bin/dsspcmbi"
martini_forcefield = "/sansom/sc3/wolf3482/AUTOMATION/Toolbox/martini_v2.2.itp"
toolbox = "/sansom/sc3/wolf3482/AUTOMATION/Toolbox/"

#Take comand line input
parser = argparse.ArgumentParser(description='Input number of proteins to simulate')
parser.add_argument('-np', '--protnum', default=2)
parser.add_argument('-sim', '--simnum', default=2)
parser.add_argument('-time', '--simtime', default=500)
parser.add_argument('-cores', '--cores', default=4)
parser.add_argument('-sub', '--sub', default="workstation")
parser.add_argument('-eq', '--eq', default="yes")
parser.add_argument('-temp', '--temperature', default="310")
parser.add_argument('-wait', '--wait', default="0")
parser.add_argument('-gpu', '--gpu', default="no")

args, unknown = parser.parse_known_args()
```

```
argsdict = vars(args)
protnum = argsdict['protnum']
simnum = argsdict['simnum']
simnum = int(simnum)
cores = argsdict['cores']
simtime = argsdict['simtime']
submission = argsdict['sub']
eq = argsdict['eq']
temperature = argsdict['temperature']
sleeptime = float(argsdict['wait'])
gpustat = argsdict['gpu']
print "Will set up %s systems with %s proteins - taking first .pdb file" % (simnum,
eqrun = str()
if eq != "yes":
eqrun = "no"
if eq !="no":
eqrun = "no"
print "Equilbration state not recognised - will default to no, no equilibration wil
if eq == "yes":
eqrun = "yes"
#Find protein in input
protein = list()
endfile = str()
for item in sys.argv:
if item.endswith(".pdb") or item.endswith(".gro"):
protein.append(basename(item))
```

```
endfile = item


topology = list()
for item in sys.argv:
if item.endswith(".itp"):
topology.append(item)
#Determine CG or AT file
ATprotein = str()
for item in protein:
if "_cg_" in item or "_CG_" in item:
pdbtype = "CG"
print "%s read as Coarse-grained file - if this is not the case, rename to contain ei
if not topology:
print "No topology given, this is required with a CG structure - dying"
sys.exit()
elif "_at_" in item or "_AT_" in item:
pdbtype = "AT"
print "%s read as Atomistic file - will now attempt to convert using martinize.py - c
ATprotein = item
else:
print "Unable to recognise %s as CG or AT, please rename with _AT_ or _CG_ for clarif
sys.exit()
if os.path.isdir("Coarsegrain"):
pass
else:
os.system("mkdir Coarsegrain")
```

```
os.chdir("Coarsegrain/")

#Produce CG protein from AT structure, or copy CG protein
if pdbtype == "CG":
for item in topology:
topology = item
for item in protein:
protein = item
print "Passing on %s and %s to simulation setup, for %s proteins." % (protein, topo
os.system("mv ../" +protein + " .")
os.system("mv ../" +topology + " .")
elif pdbtype == "AT":
os.system("python " +toolbox +"martinize.py -f ../" +ATprotein + " -o topol.top -x

#Now have CG protein .pdb and .itp files in Coarsegrain directory

#Have :CG_protein.pdb, number of proteins to put in box.
# If started off AT, now have specified ITP file (output from martinize script)
# If started off CG, now have specified ITP file from input

if pdbtype == "CG":
itpfile = topology
if pdbtype == "AT":
itpfile = "CG_topology.itp"
```

```
topolbuild(itpfile, protnum, martini_forcefield)



proteinCG = "CG_protein.pdb"

systemout = "System_preem"

editconf_system(proteinCG, protnum, systemout)

#genbox(proteinCG, protnum, systemout)

os.chdir("../")

emfile = "%sprotem.mdp" % (toolbox)

protein_prewat = "Coarsegrain/%s_conf.pdb" % (systemout)

topology = "Coarsegrain/CG_top.top"

outfile = "em_protonly.tpr"



## Grompp files into pre em ##

grompp(protein_prewat, topology, emfile, outfile)

mdrun(outfile)



outfilestrip = os.path.splitext(outfile)[0]

minprot = "%s.gro" % (outfilestrip)

solvatedprotein = "solvated_preions.gro"

solvatebox(minprot, topology, solvatedprotein)
```

```
#Now have solavted box with no ions and pre-em

emfile = "%swatem.mdp" % (toolbox)
solvated_grompp = "solvated_postgrompp.tpr"
grompp(solvatedprotein, topology, emfile, solvated_grompp)
ions_out = "solvated_ions.gro"
option = "-neutral -conc 0.05"
genion(solvated_grompp, ions_out, topology, option)

emfile = "%sionsem.mdp" % (toolbox)
ions_em = "protein_ions_preem.tpr"
grompp(ions_out, topology, emfile, ions_em)
mdrun(ions_em)

#generate position restraints
outfilestrip = os.path.splitext(ions_em)[0]
ions_em = "%s.gro" % (outfilestrip)

forceconstant = "1000"
posre = "Coarsegrain/POSRES.itp"
proteinCG = "Coarsegrain/CG_protein.pdb"
genrestr(proteinCG, posre, forceconstant)

#Generate and run equilibration
if gpustat == "yes":
emfile = "%s4_6_eq.mdp" % (toolbox)
```

```
else:
emfile = "%s4_5_5_eq.mdp" % (toolbox)
equilibration_ready = "equilibration.tpr"
grompp(ions_em, topology, emfile, equilibration_ready)
time.sleep(sleeptime)
if eqrun == "yes":
print "##############################################\n"
print "Running equilibration - see below for output\n"
print "##############################################\n"
mdrun(equilibration_ready)
endfilename = os.path.splitext(endfile)[0]
outfilestrip = os.path.splitext(equilibration_ready)[0]
equilibration_output = "%s.gro" % (outfilestrip)


#Architecture for running simulations - below is for workstation queue
if submission == "workstation":
workstationsubmit(simnum, simtime, equilibration_output, topology, cores, temperature
elif submission == "neuron":
neuronsubmit(simnum, simtime, equilibration_output, topology, cores, temperature)
elif submission == "astrocyte":
astrocytesubmit(simnum, simtime, equilibration_output, topology, cores, temperature)
elif submission == "dendrite":
dendritesubmit(simnum, simtime, equilibration_output, topology, cores, temperature)
elif submission == "iridis":
iridissubmit(simnum, simtime, equilibration_output, topology, cores, temperature)
```

```
elif submission == "none":

pass

else:

print "submission cluster not recognised - try again"

elif eqrun == "no":

print "Equilibration ready to run: %s" % (equilibration_ready)
```