Towards molecular simulations of glycans, glycoproteins, and membrane-membrane interactions

Thesis by
Elizabeth Eva Jefferys

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy



UNIVERSITY OF OXFORD
Oxford, United Kingdom

Submitted January 26, 2021

©

Elizabeth Eva Jefferys
ORCID: https://orcid.org/0000-0002-0097-0520

ACKNOWLEDGEMENTS

With thanks to Professor Mark Sansom for supervising me during this research, and to all the members of the Structural Biology and Computational Biochemistry Unit in the University of Oxford, in particular Doctor Robin Corey, Doctor Anna Duncan, and Joshua Sauer for offering help and advice at various times during this DPhil. Thanks also to Professor Siewert-Jan Marrink, Doctor Paulo Telles de Souza, and Bart Bruininks at the University of Groningen for allowing me to spend some time working there on Martini 3.

ABSTRACT

Close apposition of membranes is a frequent occurrence in biological systems, and one of considerable physiological and pathological significance. The geometric and compositional complexity of such associations, and in some cases their transient and/or rare occurrence, makes them challenging to study experimentally, and necessitates a multidisciplinary approach to their investigation. Molecular dynamics (MD) simulations have considerable potential to form part of the toolkit for studying membrane interactions, being able to provide molecular-level insight into the time-dependent behaviour of biological systems. However, there at present is no standard approach to prepare and simulate membrane-membrane interactions for MD simulations. In this work, I present progress towards the development of methods that could prove valuable in preparing simulation systems for this purpose.

TABLE OF CONTENTS

Acknowledgements . . . iii
Abstract . . . iv
Table of Contents . . . v
List of Figures . . . vii
Chapter I: Introduction . . . 1
1.1 Double bilayers in biology . . . 1
1.2 Studying membrane-membrane interactions . . . 3
1.3 Project overview . . . 10
Chapter II: Background to molecular dynamics simulations . . . 12
2.1 Introduction . . . 12
2.2 Theory . . . 13
2.3 Choosing a force field . . . 24
2.4 The Martini force field . . . 33
2.5 Running a simulation . . . 42
2.6 Setting up a simulation . . . 50
2.7 Progress in membrane simulations . . . 60
2.8 Molecular dynamics simulations presented in this thesis . . . 67
Chapter III: Methodological development of double bilayer simulations . 69
3.1 Introduction . . . 69
3.2 Method overview . . . 71
3.3 Basic pore properties . . . 75
3.4 Bilayer asymmetry . . . 76
3.5 Protein-containing bilayers . . . 78
3.6 Justification . . . 81
3.7 Concluding remarks . . . 81
Chapter IV: Parameterising glycans for Martini 3 . . . 85
4.1 Introduction . . . 85
4.2 Parameterising the Martini monosaccharides . . . 91
4.3 Parameterising the glycan building blocks . . . 96
4.4 Interresidue bonded parameters . . . 113
4.5 Model evaluation . . . 116
4.6 Conclusions and further work . . . 132
Chapter V: Modelling glycoproteins . . . 134
5.1 Introduction . . . 134
5.2 Computational glycosylation method . . . 136
5.3 Glycan construction . . . 140
5.4 Example applications . . . 152
5.5 Future work . . . 164
Chapter VI: Putting it all together . . . 166

6.1 Introduction 166
6.2 The system 166
6.3 Approach taken 167
6.4 Concluding remarks and future directions 171
Bibliography 179
Appendices 230
Appendix A: Structures and abbreviations of monosaccharides discussed in the text 231
Appendix B: Structures of lipids most often used in this work 233
Appendix C: Methodological details 235
Appendix D: Martini 3 monosaccharide trials 238

LIST OF FIGURES

Number *Page*

1.1 Schematic illustration of different types of membrane-membrane interaction that occur in biological systems. The immunological synapse between a T-cell and an antigen-presenting cell, which allows the generation of effector and memory T cells to promote elimination of a particular pathogen, the process of leukocyte rolling motility along endothelia to allow their movement out of the circulatory system into sites of tissue infection or damage, and association of an enveloped virion with a host cell, are all relatively transient cell surface associations. The interactions involved in tissue formation and maintenance, which may occur between cells of the same or different types, tend to be more stable and prolonged, as do the junctions between epithelial cells that are important in the maintenance of epithelial barriers. Membrane-contact sites arise within eukaryotic cells, between the various membranous compartments of the cell, allowing communication between them and better coordination of cellular activities. 2

1.2 Summary of the components that may be found in a membrane-membrane interaction, using the proteins involved in Herpes Simplex Virus 1 attachment to a host cell as an illustration. Constructing models of the protein components for simulation may require, in addition to a high resolution structure from e.g. X-ray crystallography, modelling missing regions such as loops and transmembrane domains. As such procedures are well-established in the simulation community, I do not focus on them here. Instead, I focus on how to simulate a double bilayer and how to construct models of the glycan components. Although neither is a novel notion to MD simulation, to my knowledge neither is routinely considered nor has available established robust protocols to follow. The proteins shown are the receptor-binding glycoprotein gD (blue) [PDB[58] ID 2C36][59], and gH/gL (violet and cyan)[3M1C][60] from the virus, and a complex of gD with host protein Nectin-1 (blue and yellow, respectively) [3U82][61]. The proteins are shown in surface representations, and missing regions indicated with dark grey lines. The bilayers are shown as simple slabs and glycans schematically with hexagons. 10

2.1 Illustration of the bonded and nonbonded interactions commonly included in MD force fields. The bond length term (top left) relates to particles connected by a covalent bond, modelled using a harmonic potential in many cases. The bond angle term (top middle) exists between three bonded particles, and is often also modelled using a harmonic potential. Both the dihedral (top right) and improper (bottom left) terms involve four bonded particles, and are defined as the angle between the planes containing the first three and last three of these. The two nonbonded terms (bottom middle and bottom right) described interactions taking place between particles not directly linked, and include the van der Waals interactions that arise between any two particles, and the electrostatic attraction or repulsion between pairs of charged particles. 22

2.2 Simple illustration of the range of resolutions used in MD, from including all atoms to simplifying all components to either single interaction sites, simple geometric approximations of the true shape, or continuum environments. 25

2.3 A representation of some of the differences between atomistic and coarse-grained forcefields. The energy landscape for the simulated system, such as the hypothetical simple one shown, is smoothed upon coarse-graining so that it becomes easier to explore and the system is less likely to become trapped in local energy wells. Nonbonded interactions are also cutoff in many cases, rather than computationally expensive methods being required to model them. Combined with there simply being fewer particles for a given molecule, this reduces the pairwise interactions that need to be calculated at each timestep. Additionally, the loss of high frequency motions of atomic bonds allows use of a larger timestep, further helping to render more accessible longer timescales on coarse-graining. 29

2.4 Comparison of a protein (PDB code: 6K42[206]), a lipid (dipalmitoylphosphatidylcholine (DPPC)), and solvent with ions between CHARMM36 with TIP3P water (left-hand images) and the Martini force field (right-hand images). The molecules are shown as spheres coloured according to element or (for Martini) elements 'dominating' the properties of the underlying chemical groups (grey: C, red: O, blue: N, orange: P, yellow: S, Na purple). CG water is shown as pale blue spheres for consistency with later figures. 36

2.5 Two dimensional illustration of periodic boundary conditions, as applied in this example to a simple system consisting of a single lipid bilayer (grey) in a box of water (blue) that contains a number of solute molecules (orange/red). By effectively replicating the simulation cell infinitely in x, y, and z, it is possible to remove boundary effects that otherwise arise due to the finite size of the modelled system. According to the minimum image convention, the particle shown in red only interacts with the nearest neighbours, as indicated with dashed lines. 45

2.6 Illustration of a generalised MD workflow, starting from a protein structure to its solvation and ultimately the MD simnulation. The protein is shown in cartoon representation, with the inital structure in orange and modelled regions in red. Water is shown as greyish blue sticks and ions as pink and lilac spheres. The protein is a salutaridine reductase from *Papaver somniferum*, PDB code 3O26[260]. 51

2.7 Illustration of two approaches to constructing a protein-containing bilayer for MD simulations. The top illustrates the formation of a bilayer around a position-restrained protein, starting from a randomly distributed box of lipids intermixed with water. This specific example is the assemble of a DPPC bilayer about the $\alpha 2\beta$ adrenoceptor (PDB code 6K42, carried out using the MemProtMD[277] pipeline). Below is an example where a protein is inserted into a bilayer arrangement of lipids, which is allowed to equilibrate around it. This example is the KALP peptide being incorporated into a DPPC bilayer using *insane*[190]. Lipid headgroups are shown as light grey spheres, lipid tails as dark grey spheres, proteins as gold spheres, and the solvent in pale blue, as a translucent surface (top) or spheres (bottom), respectively. 57

3.1 Illustration of the problem created by having two bilayers in a single simulation cell. The issue derives from the fact that the bilayers are (unless in possession of one or more open pores) impermeable to water and ions, and due to the periodic boundary conditions are essentially infinite in the xy (lateral) dimensions. This creates two incompressible water compartments between the two bilayers, which has the potential to deform the bilayers and/or impede their dynamics. 71

3.2 Illustration of the method used to create pores in lipid bilayers. First, the desired x-y coordinate of the pore is chosen and one or more particles in the bilayer lipids are selected e.g. the C3A and C3B beads in a Martini lipid tail (top). A reference structure file is then created in which all of these beads in the system are mapped to the desired site of the pore (middle). During the simulation, these beads are then repelled from that specified point in the bilayer plane using a flat-bottomed repulsive potential, which acts only on those lipids in close proximity to the defined site of the pore (bottom). This has the effect of inducing the lipids to form a toroidal, headgroup-lined pore at the defined position. 73

3.3 Illustration of the functional form of the flat-bottomed potential used to form pores in the bilayers. The regions where the potential increases would correspond to the pore, where the lipids feel the repulsive potential, whilst the rest of the bilayer would be the flat part of the potential. 74

3.4 Final frames from 1 μs simulations of POPC bilayers in which pores were formed by applying repulsive potentials of force constant 1000 kJ mol^{-1} nm^{-2} and radius 1.5 nm to different tail beads of the POPC lipid. The tail beads are indicated in the right-hand image, and colour-coded with the corresponding snapshot on the right-hand-side. In each snapshot, the system is viewed along the plane of the bilayer and shows only the waters, highlighting the presence or absence of a complete water conduit. The image at the top shows a cross-section of the bilayer itself, with lipid tail beads in dark grey, the lipid headgroup choline, phosphate, and glycerol groups in blue, pink, and light grey, respectively. 77

3.5 Loss of asymmetry *via* a lipidic toroidal pore. When there is no pore present in this small, asymmetryic POPC/POPS bilayer, the two leaflets remained separated (top). Upon addition of a pore, the leaflets quickly mix and asymmetry is lost (middle). Applying a repulsive potential to the lipid headgroups to discourage such mixing allows asymmetry to be maintained (bottom). 79

3.6 Simulations of a KALP peptide in a POPC bilayer containing a single pore. The top image shows the peptide associating with the pore edge when no repulsive potential is applied to it. This can be seen more clearly in the middle plot, showing the x-y (in-plane) protein coordinates over the course of the simulation. Each point represents the protein centre of geometry at a single frame in the simulation, and the points are translucent such that a dense colour indicates repeated occupancy of the same position. Applying a repulsive potential allowed the peptide to explore the simulation cell more freely, as can be seen in the lower plot, showing the protein to lo longer circumnavigate the pore. 80

3.7 Final frames from 5 μs simulations of double bilayers with no pore (top), with a pore formed using a repulsion radius of 1.5 nm (middle), or a pore formed using a repulsion radius of 2.5 nm (bottom). In each case, the surface of each bilayer is shown in grey and the waters are ommitted. 82

4.1 Schematic illustration of the occurrence of glycans on the surfaces of Gram positive and Gram negative bacteria, animal, plant, and fungal cells, and viruses. The glycan components of the surfaces are shown in red, whilst the other constituents (lipids, proteins, cytoplasm etc.) are in grey. Note that these representations are illustrative and, for instance, there are variations in the specific compositions, some of the glycan types shown may not always be present whilst others not indicated may be found. 87

4.2 Skeletal structures illustrating some of the basic features of carbohydrate chemistry, as discussed in more detail in the text. Top and right: examples of the differences between anomers and isomers, which differ in the orientations, but not identities, of substituents about the ring. Middle left: a comparison of some of the conformations a pyranose ring might adopt in solution. Bottom left: two examples of disaccharides of Glc differing in the nature of the glycosidic bond, and highlighting the reducing and nonreducing ends of the molecule. 90

4.3 Comparison between the structures of three plant polysaccharides of glucose. Both cellulose and amylose are linear polymers, differing only in the anomeric state of their consituents. This difference may lead to the formation of helical structures in amylose, as opposed to long linear chains in cellulose. Amylopectin differs from amylose in the presence of additional branches linked to C6 of some residues. Both components of starch, the combination of these two types of organisation allow the formation of semicrystalline granules that are held together by networks of hydrogen bonds and are insoluble in water. In contrast, the linear chains of cellulose hydrogen bond with each other to form microfibrils with high tensile strength, an important feature for this major structural component of plant cell walls. 92

4.4 Illustration of the approach taken in the first work extending Martini to model carbohydrates. Monosaccharides such as glucose were modelled using three connected CG beads, whilst in disaccharides and oligosaccharides such as trehalose and amylose, each constituent monomer was modelled as a linear set of beads linked by a single harmonic bond with two dihedral angle terms (not shown) to define their relative orientation. In the case of sequences of three or more monosaccharide units, an additional angle term was included to better model higher order oligosaccharide conformations. 95

4.5 Chart of some monosaccharides commonly identified in biological systems, showing the symbol and nomenclature scheme that is standard in glycobiology[414]. The shapes and colours are consistent with stereochemisty for hexoses, hexosamines, N-acetylhexosamines, hexuronates, and pentoses (rows 1–4 and row 8) e.g. Glc is an isomer of Ido and Alt, and has the same arrangement of substituents to the ring as GlcNAc, GlcN, Qui, and GlcA. The shape alone is consistent for deoxyhexoses, deoxy-N-acetylhexosamines, dideoxyhexoses, and dideoxynonulosonates (rows 5–7 and row 10) e.g. Fuc is an isomer of Rha and Qui. . . . 97

4.6 Overview of the grouping (left), preliminary (middle), and final (right) mappings of the CG monosaccharides corresponding to the hexoses (e.g. Glc), *N*-acetylhexosamines (e.g. *N*-acetylglucosamine (GlcNAc)), deoxyhexoses (e.g. fucose (Fuc)), and Xyl. In each case, a general skeletal structure for the group is shown, with the heavy atoms to be grouped into CG beads boxed together. The numbers in the left-hand column indicate the CG bead number, whilst the bead types in the middle and right-hand columns are the preliminary and final, respectively, bead types chosen for each of those groups. The name of the CG residue is given: HXS, NHX, DHX, and XYL, respectively. 101

4.7 Overview of the grouping (left), preliminary (middle), and final (right) mappings of the CG monosaccharides corresponding to the hexosamines (e.g. glucosamine (GlcN)), hexuronates (e.g. iduronic acid (IdoA)), and sialic acids (specifically, *N*-acetylneuraminic acid (Neu5Ac)). In each case, a general skeletal structure for the group is shown, with the heavy atoms to be grouped into CG beads boxed together. The numbers in the left-hand column indicate the CG bead number, whilst the bead types in the middle and right-hand columns are the preliminary and final, respectively, bead types chosen for each of those groups. The name of the CG residues is given: HXN, HXA, and SIA, respectively. 102

4.8 Schematic illustration of the theory behind the octanol-water partition coefficient. The parameter reflects the relative occupancy of a solute in octanol verus water phases at equilibrium (top), whilst my computational approach involves simulating in octanol or water systems and calculating free energy differences between the solute in the different solvents. This is done using the thermodynamic cycle shown (bottom). 107

4.9 Simple illustration of the free energy calculation process, in which the solute is slowly 'phased' into existence within the given solute by increasing the value of λ over a series of simulations. The overall free energy difference between the end states at λ of zero and one, respectively, is calculated by combining the smaller energy differences between intermediate states. 109

4.10 Scatter plot of the computed partition energies for my preliminary CG models (turquoise), compared with those predicted from the AT structures (grey). The standard deviations on the computed values (less than 0.1 kJ mol^{-1}) are too small to be seen on the plot. 110

4.11 Scatter plot of the computed partition energies for my CG models (shades of blue/green), compared with those predicted from the AT structures (grey). The standard deviations on the computed values (less than 0.1 kJ mol^{-1}) are too small to be seen on the plot. The models chosen to be used for each monosaccharide are those with the closest match to the predicted partition energy. Full details on all the models tested are given in Appendix D. . . 111

4.12 Illustration of how the bonds and angles between Martini monosaccharides are to be described. Between each pair of residues there is a single harmonic bond length potential linking virtual sites (VS) at their centre of geometry, and six harmonic bond angle potentials describing the spatial relationships of the beads from each other. Both the bond length and angles vary between different pairs of residues and bond types (α, β, number of the carbon involved). 114

4.13 Plots of bond lengths obtained from the ATMD simulations of disaccharides. In each case, the bond length is illustrated with a horizontal bar proportional to the mean bond length, with a surrounding blue bar of length indicating the standard deviation. The transparency of this bar is proportional to the number of simulations of this particular pair, with a darker bar indicative of more simulations (i.e. higher n). The value of the mean length is given above the bar, and the name of the CG disaccharide pair shown beneath it. The data are given in full in the Supplementary Information. 116

4.14 Bar blots indicating the angles associated with each type of bond. Each bar corresponds to a single angle, defined as previously discussed, and the grey thinner bar to the standard deviation for that angle. 117

4.15 Plot indicating the values of individual bond types between pairs of CG residues. At the bottom of each column is the collected data for all pairs of that type (as shown in Figure 4.13). The data are presented as in Figure 4.13. 118
4.16 SNFG structure of the A2 glycan. 119
4.17 Plots of the radius of gyration, averaged over all A2 molecules present, and the total number of contacts between the A2 molecules, for simulations of 1, 5, 10, 50, or 100 and 5, 10, 50, or 100 A2 molecules, respectively. The A2 glycans in the final snapshot are shown in coloured spheres in the images on the right, with the neighbouring simulation cells also shown to illustrate any interactions that may be taking place across the periodic boundaries. . 121
4.18 Plot of area per lipid against temperature for four systems, containing no sugars and a monosaccharide HXS at concentrations 1M, 2M, and 4M. Above the graph are images of the final frame of each simulation at temperatures 200K, 275K, 325K and 400K, with waters shown in blue, bilayer lipids in grey, and sugar molecules in slate. 124
4.19 Final snapshot of a DPPC/DLPC/cholesterol bilayer after 10 μs simulation (left), and of the same bilayer simulated for a further 5 μs following addition of my trehalose model to the surrounding solution (right). The waters and ions are omitted for clarity, whilst the other components are shown as spheres (light grey and white for DPPC and cholesterol, respectively; dark grey for DLPC; cyan for trehalose. 125
4.20 Illustration of the synthetic pathways, and molecular relationships, between some of the major brain gangliosides, starting from addition of sialic acid to LacCer to form GM3. The gangliosides may be classified as asialo-, a-, b-, and c-series depending on whether 0, 1, 2, or 3 sialic acids are linked to the inner Gal residue. When sialic acid moieties are linked to the inner GalNAc residue, the α-series gangliosides are produced. 128

4.21 Comparison between simulations of different Martini 3 gangliosides, with the SNFG representation of each given on the right-hand side. Left: final (1 μs) snapshot, with gangliosides in red and DOPC in grey, and the mean ± standard deivation of the area per lipid given above. Middle: average number of contacts between ganglioside sugar moieties (red) and the interfacial regions of their ceramide groups (orange), with a contact being defined as any two particles belonging to the sugar/ceramide groups of a pair of gangliosides being within a 6.0 Åcutoff. Right: Distribution of maximum height differences between ganglioside headgroups and DOPC headgroups over the entire simulation. The height difference was taken between the mean height of the DOPC choline (NC3) bead and the ganglioside sugar beads, with the maximum of these differences being recorded at each timestep. 129

4.22 Comparison between simulations of Martini 2 (top) and Martini 3 (bottom) simulations. Left: final (1 μs) snapshot, with GM3 in red and DOPC in grey, and the mean ± standard deivation of the area per lipid given above. Middle: average number of contacts between ganglioside sugar moieties (red) and the interfacial regions of their ceramide groups (orange), with a contact being defined as any two particles belonging to the sugar/ceramide groups of a pair of gangliosides being within a 6.0 Åcutoff. Right: Distribution of maximum height differences between ganglioside headgroups and DOPC headgroups over the entire simulation. The height difference was taken between the mean height of the DOPC choline (NC3) bead and the ganglioside sugar beads, with the maximum of these differences being recorded at each timestep. 131

5.1 Schematic illustration of some different types of glycoprotein found in vertebrates. Glycans can be attached to proteins through e.g. amide linkages with Asn side chains (*N*-glycosylation), glycosidic linkages to Ser/Thr side chains (*O*-glycosylation), and carbon-carbon linkages with the C2 position of Trp (*C*-mannosylation). Proteoglycans are defined by the covalent attachment of long glycosaminoglycan chains of repeating disaccharide units, often *via* a tetrasaccharide core to a Ser hydroxyl. *N*-linked glycans can be classified as complex, hybrid, or oligomannose which differ in their degree of processing. Mucin-type *O*-glycans are characterised by a GalNAc linked to a protein Ser or Thr side chain, which may be extended by addition of further monosaccharides to form oligosaccharides of considerable diversity in length. Other types of *O*-glycan, such as *O*-linked Fuc and *O*-linked Man, tend to be found only on specific proteins or protein domains, such as epidermal growth factor repeats. *O*-GlcNAc is a unique type of *O*-glycosylation that occurs on Ser and Thr residues of nucleocytoplasmic proteins. glycosylphosphatidylinositol (GPI)-anchored glycoproteins are linked at the carboxyl terminus through a phosphodiester linkage to phosphoethanolamine attached to an oligosaccharide core; the GlcN residue is linked to phosphatidylinositol, which is embedded in the cell membrane[452]. 135

5.2 Overview of my glycosylation method. The user provides a protein structure and topology, and information regarding the type of glycosylation desired (*N*-linked, mucin-type *O*-linked, glycosaminoglycan, or monosaccharide), as well as indicating the residue numbers of glycosylation sites (if known), and whether or not the protein is membrane-embedded. The code then uses this information to construct a glycan for each site to be modified, using knowledge-based rules and a degree of randomisation to achieve physiologically plausible and heterogeneous glycan structures. Because the glycan-construction component is managed by a separate code from the glycan attachment, it is also possible to use the same tools to construct free glycans. 137

5.3 Topology predictions were carried out for 35 different proteins, belonging to 12 distinct superfamilies. One example from each superfamily is shown (with the exception of the Kv voltage sensor, which is shown along with KcsA). The cartoon representations of the proteins are coloured according to their predicted topological location, with grey for transmembrane, purple for extracellular, and light blue for intracellular. The red and blue spheres indicate the OPM-predicted positions of the bilayer headgroup region for the same proteins, with red for the outer leaflet and blue for the cytoplasmic leaflet. The proteins are: 3J2P[460], 6H2F[461], 5ER7[462], 1R3J[463], 1M0L[464], 1ORS[465], 4P79[466], 5TCX[467], 5C1M[468], 2O9G[469], 3TDS[470], 1OQW[471], and 3HD7[472]. 141

5.4 Simplified outline of the stages in eukaryotic synthesis of *N*-linked glycans, highlighting the stages at which the glycan may deviate into formation of oligomannose or hybrid, rather than complex, structures. The precursor oligosaccharide $Glc_3Man_9GlcNAc_2$ that is transferred from a dolichol-phosphate lipid precurosor to the Asn side chain is trimmed in the ER before the glycoprotein can proceed to the Golgi. The removal and reattachment of the Glc residues plays an important role in protein quality control, which is not discussed here but may be studied in detail in any standard introduction to glycobiology[388, 485]. The glycan, on transit from the ER to the Golgi, will usually contain eight or nine Man residues, and is then further trimmed to yield the $Man_5GlcNAc_2$ intermediate. If this trimming does not occur, the glycan may be released as an oligomannose type. Processing of the intermediate in the *medial*-Golgi is initiated by the addition of β1–2 GlcNAc to the α1–3Man branch in the core unit, which allows removal of the terminal Man residues from the other branch to form $GlcNAcMan_3GlcNAc_2$, and is generally followed by similar attachment of GlcNAc to the α1–6Man arm of the glycan. If this does not occur, it may lead to the formation of a hybrid glycan, but if both branches are processed to yield terminal GlcNAc residues, they may be further extended to yield a variety of complex glycans. Additional branches may be initiated at C4 of the core α3 Man, the C6 of the core α6 Man or, at least in birds and fish, at the C4 of the core α6 Man. Complex and hybrid glycans can also carry a bisecting GlcNAc linked to the βMan of the core. The extension of hybrid and complex glycan precursors occurs through further additions to the core structure, elongation of the branches, and attachment of capping residues, such as α-linked sialic acid, Fuc, Gal, GlcNAc, or sulphate. The α linkage in a branch of β linkages means that the capping residue protrudes from the branch, therefore being more readily presented for interaction with e.g. lectins or antibodies. 145

5.5 SNFG illustrations of *N*-linked glycans produced using my approach, showing the output from five independent repeats for each glycan class, complex, hybrid, or oligomannose. 147

5.6 Overview of the general structure of mucin-type glycans, the eight core structures, and example glycans generated using my code. In the latter case, the output from five independent runs of the code are shown. All glycans are shown using the SNFG representation. 149
5.7 Examples of glycosaminoglycan chains produced using my code. For each GAG chain, the outputs from three independent runs of the code are shown in SNFG representation. 151
5.8 Structure and posttranslational modifications of MAG. The crystal structure (PDB ID 5LFU[38]) is shown in grey cartoon representations, with resolved sugar moieties as coloured spheres. Individual posttranslational modifications are indicated in the boxes. Note that the protein includes both Trp-mannosylation, and *N*-linked glycosylation, the former being in the vicinity of the ligand-binding site as shown in the top-left inset with the Man in sticks, and the ligand sugars resolved in the crystal structure as spheres. The ligands are gangliosides, with a preference towards GD1a and GT1b, as indicated in the top-right. 153
5.9 Results from the MAG simulations. Top: final frame from the CGMD of a glycosylated WT MAG dimer, showing the protein backbone as a grey tube and the glycans as spheres. Waters and ions have been omitted for clarity. Bottom: plots of mean radius of gyration and maximum interatomic distance obtained from simulation (orange) compared with the corresponding data from SAXS experiments (grey). 156
5.10 Final (1 μs) snapshots from four different simulations of glycoproteins in bilayers. In each case, the lipid headgroups (phosphate and choline beads) are shown in dark grey, the protein backbone as a light grey tube, and the glycans as spheres coloured according to CG monosaccharide type, as indicated in the key. For CXCR4, the GAG chain crossed the periodic boundary, and so appears to be coming back in from the other side of the simulation cell. . . . 158

5.11 Illustration of the product Smoothened glycoproteins when the -bilayer flag is not (left) and is (right) included, showing the exclusion in the latter case of the intracellular PNG. The protein backbone is shown as a grey tube and the glycans as spheres, coloured according to CG residue type, with sialic acid in red, the *N*-acetyl hexosamines in dark blue, the hexoses in slate, and deoxyhexose in cyan. 158

5.12 ACE2 simulation results. Top: final frame from CGMD of fully glycosylated ACE2 in a POPC bilayer, with the protein backbone as a grey tube and the glycans as spheres, coloured by residue type as indicated in the key. Waters and ions have been omitted for clarity. Bottom: plots of the proportion of the simulation each residue spends exposed to solvent, comparing fully glycosylated, fully deglycosylated, and partially deglycosylated (missing the glycan on Asn90) simulations. 161

5.13 Comparison of the SARS-CoV-2 Spike glycoprotein simulations with all (left) and with a subset (right) of the glycosites modified. Top: starting configurations. Bottom: configurations at the longest time the simulation got to/has been run thus far, as indicated. The protein is shown as grey tubes, the bilayer as grey spheres, and the glycans as spheres coloured according to the key. 164

6.1 In order to assemble a double bilayer system, I started by constructing and equilibrating each of the bilayers separately. A pore was incorporated into one during equilibration, and the unglycosylated proteins into the other. After equilibration of this latter, the proteins were glycosylated. The bilayers were then combined by translating them relative to each other, the system solvated and neutralised, thereby yielding the complete system. 168

6.2 The starting configuration of the MAG-bridged myelin-axon membrane-membrane association system simulated. The lipid, glycan, and protein components are shown as spheres, and waters and ions omitted for clarity. The components are coloured as indicated in the keys. 170

A.1 Skeletal structures, name, abbreviation, and SNFG symbol for the monosaccharides most frequently discussed in this work. 232

B.1 Skeletal structures and overviews of the lipid types most often used in this work. Top: the phospholipids, showing the tails dipalmitoyl (DP), palmitoyloleoyl (PO), dilauroyl (DL), and DO (dioleoyl), and headgroups ethanolamine (PE) and choline (PC); these are used in the simulations as e.g. DPPC, POPE. Middle: sphingolipids, which are composed of a sphingosine and fatty acid linked to a headgroup to form e.g. ceramide (H), sphingomyelin (choline), or glycosphingolipids (glycans). Bottom: cholesterol, composed of a steroid core, hydrocarbon tail, and polar hydroxyl headgroup. 234

Chapter 1

INTRODUCTION

1.1 Double bilayers in biology

All cells and many viruses, are surrounded by membranes, complex assemblies composed of a lipid bilayer with myriad associated proteins and carbohydrates. These membranes form the interface between cells and their immediate environment, and their direct juxtaposition occurs throughout biology[1] (Figure 1.1). This is perhaps most evident in the context of multicellular species, in which the direct association of cells forms a fundamental facet of their development and functional existence. For instance, tissue formation, maintenance and coordinated activity[2], and the operation of both the immune[3] and nervous systems[3–5] relies on numerous membrane-membrane interactions. In the multicellular assemblies of species such as filamentous cyanobacteria, direct cell-cell interactions play an important role in governing which cells differentiate to perform particular functions essential to the operation of the whole entity[6].

The membranes directly involved in cell-cell interactions may belong to the same or different species. For example, within bacterial biofilms, which may contain a diversity of different species, cell-cell communications *via* surface exposed proteins play an important role in establishing the 3D architecture of the biofilm following initial surface adhesion[7, 8], whilst enveloped viruses and other membrane-bound pathogens, parasites, and symbionts rely on membrane-membrane interactions to adhere to cells of their hosts[9, 10].

The various membranes within eukaryotic cells interact continuously and dynamically with each other, exchanging information in the form of signals, ions, and lipids. These membrane contact sites are critical to the proper functioning of cells and the organisms they make up, with identified roles including regulation of the immune response, apoptosis, and organelle dynamics[11]. Membrane-membrane interactions occur also within bacteria, most conspicuously in the case of Gram negative species, whose inner and outer membranes are connected *via* various proteins spanning both bilayers or interacting with each other directly or indirectly *via* components of the periplasmic space between.

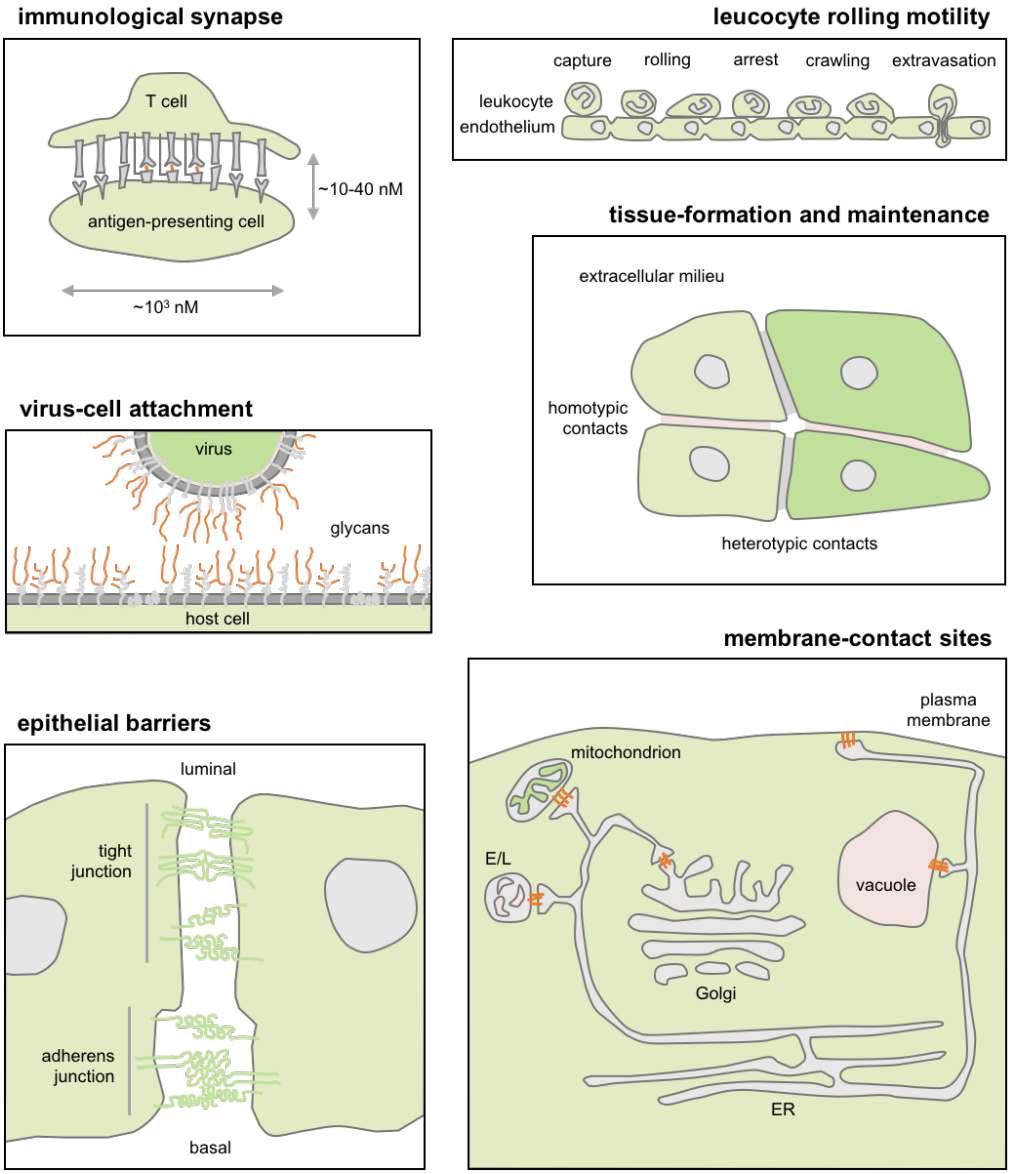


Figure 1.1: Schematic illustration of different types of membrane-membrane interaction that occur in biological systems. The immunological synapse between a T-cell and an antigen-presenting cell, which allows the generation of effector and memory T cells to promote elimination of a particular pathogen, the process of leukocyte rolling motility along endothelia to allow their movement out of the circulatory system into sites of tissue infection or damage, and association of an enveloped virion with a host cell, are all relatively transient cell surface associations. The interactions involved in tissue formation and maintenance, which may occur between cells of the same or different types, tend to be more stable and prolonged, as do the junctions between epithelial cells that are important in the maintenance of epithelial barriers. Membrane-contact sites arise within eukaryotic cells, between the various membranous compartments of the cell, allowing communication between them and better coordination of cellular activities.

1.2 Studying membrane-membrane interactions

Challenges

Given the aforementioned physiological and pathological importance of membrane-membrane associations, there is understandably considerable interest in being able to elucidate events taking place at such junctions. Doing so does, however, present a number of challenges for available experimental methods.

One of the main difficulties arises from the inherent complexity of any membrane-membrane interaction. Although they are, by definition, simply closely apposed lipid bilayers, in reality they are geometrically complex assemblies that may include a diversity of different lipids, proteins, sugars, and other molecules, all existing within the wider context of the cell or virion. Any attempt to study a membrane association *in vitro* or *ex vivo* will need to try and reproduce as closely as possible the native environment in which the interaction takes place, whilst studies carried out *in vivo* can come across challenges such as a multitude of difficult-to-control variables, and a higher degree of noise than encountered in more artificial scenarios.

Another challenge arises from the transient and/or rare nature of many membrane interactions, such as the previously noted pathogen-host adhesion events. In these cases, techniques that are low throughput, inherently static, or unable to achieve a high degree of sampling can struggle to capture the interaction events.

In any apposition of two bilayers, there may be several different factors at play governing the coming together of the membranes, their dynamics, and events that follow the association. Techniques for studying these associations need to be suited for examination of the particular features of interest. For instance, the mechanical properties of cell-cell interactions may help to determine morphogenic events such as epithelial cell movement and cell sorting[12–16], and a better understanding of these properties could help inform our understanding of morphogenesis in general. An experimental method that looks only at static structural details, even at high resolution, may not be of much value in understanding physical parameters such as these. Conversely, a purely physical model of an interaction could be of little utility in, for instance, trying to design drug molecules that impede specific interactions, as this often requires a precise understanding of the structures involved. Therefore, not only does one have to contend with challenges inherent to membrane-membrane associations, but it is also important to be able to apply methods appropriate to the question of interest.

Experimental approaches

Cell-cell and cell-virus contacts often take place on lengthscales in the nm–μm range[1, 3, 9]. These dimensions are amenable to various imaging methods, including both optical and electron microscopy. However, imaging of such interactions poses numerous challenges. In addition to difficulties noted previously regarding rare events or transient contacts, which can be particularly problematic given the acquisition bias that can be associated with imaging methods, one needs to have a means of distinguishing between individual components in an imaged sample. In a membrane-membrane interaction, this can include disambiguiation of those molecules belonging to each of the individual membranes, which will require the ability to spatially separate the signals belonging to each interaction partner. In fluorescence imaging, this may be achievable by using spectrally distinct fluorescent labels, although this is not invariably possible or straightforward[17].

Ultrastructural methods have the potential to provide information regarding molecular interactions and spatial organisations at contact sites between biological membranes. Ideally, it is hoped that we can identify the location of individual protein components within the 3D environment of a contact site, and correlate such localisations with real-time observations of the interaction dynamics. Some elements of associated membranes can be identified from morphology alone, whilst others require labelling.

The use of optical microscopy to examine subcellular organisation was revolutionised by the emergence of techniques that overcome the limits imposed by diffraction on spatial resolution[18] such as structured illumination microscopy and F-PALM/STORM. Such methods can now be frequently applied to characterise cell junctions[19–21]. For instance, using such approaches it was possible for researchers to demonstrate the nonhomogenous nature of cadherin (a class of transmembrane cell-cell adhesion proteins) molecules in adherens junctions (a type of multiprotein complex involved in cell adhesion and linking the actin cytoskeleton and membrane). These results were in support of previous suggestions of lateral clustering of cadherins suggested as a means to strengthen cadherin-based adhesion[22, 23]. However, in optical microscopy image acquisition is not fast enough to allow one to examine high enough numbers of cells to obtain an reasonably unbiased view of the biological situation being studied, or at least to be confident that this has been achieved.

Flow cytometry, in which fluorescently labelled cells or other particles are passed single-file through a laser to allow their rapid detection and characterisation, overcomes the issues associated with low throughput in optical microscopy as it allows one to obtain quantitative data about a population of cells (ca. tens of thousands of cells per second[24] relatively quickly, albeit in the process reducing the information to a set of intensity measurements[25]. By combining high-speed automated image acquisition with quantitative image analysis, multispectral imaging flow cytometry (MIFC) effectively leverages the advantages of flow cytometry and fluorescence imaging. By permitting the analysis of hundreds of thousands of cells within minutes, MIFC facilitates the observation and quantification of potentially rare association events[17]. It has been successfully applied to examine the interactions between a variety of different cell types, including pathogens and host cells[26], and antigen-presenting cells and T-cells forming an immunological synapse[27].

The lifecycles of many viruses involve membrane-membrane interactions, for instance during host adhesion, entry *via* fusion, and new virion release[9]. Fluorescence microscopy can be used to examine the intracellular distribution of viruses or viral components[28], and if super-resolution methods are used it is possible to visualise some subviral features[28–30]. However, it is not generally possible to visualise individual membranes using this method at high enough resolution to spatially distinguish two membranes in close apposition.

Electron microscopy (EM) has a long history of application to high resolution characterisation of cellular and subcellular structures[31]. Traditional transmission EM, including chemical fixing and resin embedding, has provided considerable information on cellular ultrastructure, including membranes and cytoskeletal elements.

However, EM used in this manner comes with several limitations, including the need to use thin enough samples to permit penetration of the electron beam (usually <500 nm), a potential requirement of adding heavy metals to enhance contrast, and the necessity of the sample being stable in the high vacuum of the microscope column, which requires the sample to undergo special preparations incompatible with live studies. Furthermore, the field of view in this method is generally relatively small, on the order of a few microns at most, which can make the observation of rare events very difficult. It can also be challenging to distinguish components of a sample, unless immunogold labelling is carried out

or the constituents are morphologically distinct enough to be identified by shape alone.

Electron tomography, which involves tilting relatively thick sections of a specimen to obtain images at different angles, can be used to construct a 3D visualisation through the sample. This has been used, for instance, to study the organisation of actin filaments running parallel to adherens junctions[32], and has found particular utility in examining the organisation of membrane-contact sites within cells[33, 34]. High-resolution electron tomography has also allowed detailed examination of the changes taking place on the virion and host cell membrane surfaces upon association of vaccinia viruses and herpes simplex virus with their respective potential host cells[35].

Correlative light electron microscopy (CLEM) is a method that aims to overcome several of the limitations associated with electron and fluorescent microscopies, leveraging the advantages of both. Specifically, the fluorescently labelled sample is first examined using light microscopy, and then high-resolution EM is used on the same site; as the images obtained using both are correlated, it is possible to use the fluorescence microscopy data to localise components in the higher resolution EM. This can be very useful in allowing the location of rare events, as well as providing dynamic information from the live cell imaging to complement the high resolution EM data[36]. CLEM can help in the imaging of virus-host interactions by allowing the precise localisation of virions (them being much smaller than the host cells and therefore challenging to locate in electron micrographs) on the cell surface, and allows linking of the images to dynamic data obtained through fluorescence microscopy[37].

Many high resolution structures of the protein participants in membrane interactions have been solved using X-ray crystallography[38–40]. These structures can provide useful insight into the protein-protein interactions that often play a central role in bringing membranes together and/or holding them in close apposition at a particular separation. Measuring the dimensions of these structures can, for instance, provide a suggestion of the functional distance between the associated membranes[38, 39].

However, the structures in general do not include the membranes themselves, and in many cases can lack any transmembrane domains at all, as well as often not considering the native glycosylation state that can play an important role in membrane-membrane interactions. Cryo-electron microscopy and tomography

can provide some information about membrane interfaces including the bilayers themselves, but are often too low resolution to provide molecular-level insight into the structures involved. Additionally, structural methods such as these are essentially static techniques. Although there are ways to get some idea of time-dependent behaviours from them, they do not generally reveal the motions of their components directly, and so can struggle to capture the true dynamics of biological systems, such as the constant breaking-and-reforming contacts that take place during leucocyte rolling motility or the progression from attachment to adhesion during virus-cell association.

Computational approaches

It is likely that we can only hope to understand the behaviour of membrane juxtapositions through a multidisciplinary approach. Computational methods are becoming an increasingly important part of multidisciplinary bioscience as they can provide unique insight into the molecular and time-varying features of biological problems, and help to marry the results obtained using disparate experimental methods[41–44].

There have already been a number of studies applying computational methods to various types of membrane association, and these have varied considerably in their focus, level of resolution, and the precise techniques applied. Here I discuss only a few examples, illustrating a small part of the potential diversity of computational methods that might be used in such studies.

At the lower resolution end of the scale, in one recent study into cell-cell contacts during morphogenesis, researchers combined an approach derived from methods used traditionally to study nonliving systems with data from live cell imaging to examine the physical principles underpinning morphogenic events[45]. By considering cell-cell interactions in terms of simple mesoscale parameters wherein cells were considered as single particles, and using real-time live cell imaging data to statistically infer a set of effective forces and potentials describing cell adhesion, researchers carried out simulations based on these potentials that were able to accurately capture the 3D morphologies of basic structures formed during morphogenesis, including tubes, cavities, cups, and 2D sheets[45]. The work aimed to establish whether there might be an underlying physical principle that could explain the basic structural units that form during morphogenesis.

Molecular dynamics (MD) simulations have proven over many years to be an

effective means of providing molecular-level insight into the behaviours of biological systems [42, 44]. In the early years of the technique, beginning in the late 1970s [46], MD suffered considerable limitations in terms of accessible time- and lengthscales, which were typically much smaller than those over which the biological events of interest took place. Furthermore, the accuracy of the physical models used was relatively poor. Although issues of this nature will likely always be associated with MD simulations, their severity has lessened greatly in recent years, with advances in simulation software and computer hardware now allowing simulations to reach the microsecond, or even millisecond, regime for systems of tens of thousands of particles. This has rendered MD a powerful 'computational microscope'[41, 43].

Double bilayers are not a novel concept in the field of MD. Indeed, the simulation of two lipid bilayers has long been used in the field of computational electrophysiology[47]. Here the bilayers do not generally serve to represent any real membranes, but rather to facilitate the study of ion flow through channels resulting from charge or concentration gradients. The method is used to study ion channels, with a transmembrane voltage being established by the creation of an ion imbalance between the two water compartments produced by the presence of more than one lipid bilayer. When only one bilayer is present, due to the reasons discussed in detail in Chapter 3, there is in effect only a single, continuous water layer surrounding the bilayer, although visually it may appear as two. There can therefore be no transmembrane ionic gradient. When two bilayers are present, there now exist two distinct water compartments that cannot exchange ions or water molecules without a conduit to do so. This is exploited in computational electrophysiology to allow the creation of ion gradients that offer an alternative to applying an external voltage in order to study the behaviours of ion channels. However, in this case the presence of two bilayers is effectively a computational tool rather than intended to represent the organisation of any real biological system. There have, however, been some applications reported of MD simulations intended to model true membrane associations.

For example, a study was carried out to examine the behaviour of the interface between a T-cell and an antigen-presenting cell, specifically a TCR-pMHC-CD4 complex, representing the minimal model of an immunological synapse[48]. With over 300,000 atoms, at the time (over a decade ago now) the system could be simulated only for 10 ns. Although exhibiting some evidence of the lipids

in the bilayers having frozen, and being too short to reveal much information about the dynamics of the system, it represents one of the few reported attempts to address systems such as this using MD.

More recently, Hsu *et al*. modelled both the inner and outer membranes of *Escherichia coli* using a more simplified representation atoms (the Martini force field[49], discussed in Chapter 2) to facilitate simulating such a large and complex system[50]. This simulation included not only the bilayers themselves, but also various integral membrane proteins, including OmpA in the outer membrane, AqpZ in the inner membrane, and the AcrABZ-TolC complex that spans both bilayers and the periplasmic space, approximated in the study as a water filled region between them. In this simulation, it was demonstrated that a considerable correlation between the motions of lipid and protein components exists, particularly in the outer membrane, with the bilayer perturbations arising from the presence of the proteins being found to extend over a greater distance than had previously been thought[50].

There is a considerable literature of computational approaches to examine the energetics and pathways associated with membrane fusion events, such as those taking places during enveloped viral entry or during vesicular transport in eukaryotic cells. Such studies have provided insight into, for instance, the intermediate stages in membrane fusion and the factors that influence the energetic barriers to the process[51–57].

Through various different approaches, these studies have shown the existence of multiple competing pathways for membrane fusion, proceeding through structurally and topologically distinct intermediates. Indeed, by carrying out extensive MD simulations of closely apposed lipid vesicles, spontaneous fusion between lipid vesicles of different compositions was found to occur through shared pathways that were used regardless of the lipid composition, albeit to different extents[51].

Thus, there is a background of computational approaches to study events taking place at juxtaposed bilayers. However, in most cases, the systems investigated are greatly simplified and/or lacking several of the native components found in real situations. To my knowledge, there have been few instances of such investigations applied to systems closer to the true biological complexity, and there certainly does not at present exist any general procedure for studying them.

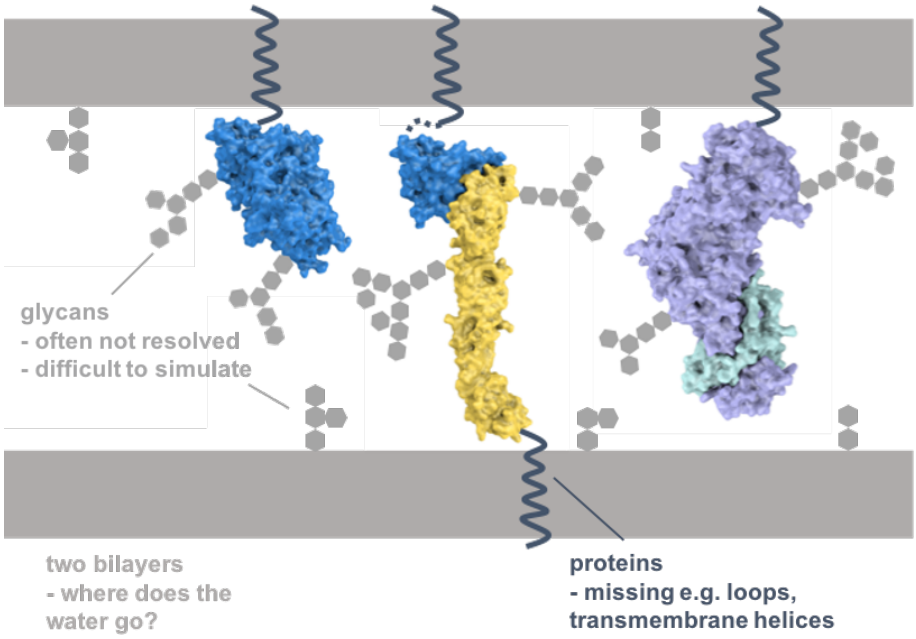


Figure 1.2: Summary of the components that may be found in a membrane-membrane interaction, using the proteins involved in Herpes Simplex Virus 1 attachment to a host cell as an illustration. Constructing models of the protein components for simulation may require, in addition to a high resolution structure from e.g. X-ray crystallography, modelling missing regions such as loops and transmembrane domains. As such procedures are well-established in the simulation community, I do not focus on them here. Instead, I focus on how to simulate a double bilayer and how to construct models of the glycan components. Although neither is a novel notion to MD simulation, to my knowledge neither is routinely considered nor has available established robust protocols to follow. The proteins shown are the receptor-binding glycoprotein gD (blue) [PDB[58] ID 2C36][59], and gH/gL (violet and cyan)[3M1C][60] from the virus, and a complex of gD with host protein Nectin-1 (blue and yellow, respectively) [3U82][61]. The proteins are shown in surface representations, and missing regions indicated with dark grey lines. The bilayers are shown as simple slabs and glycans schematically with hexagons.

1.3 Project overview

In this work, I aim to make progress towards developing a robust methodology for carrying out MD simulations of membrane-membrane juxtapositions. In order to do this, I consider each of the components that may frequently be observed in such a system and, if there are not existing ways to model that element in MD, work on developing a means to do so.

Conceptually, one may break the components of an interaction involving two cell surface membranes into the lipidic, proteinaceous, and saccharide parts. My approach has been to consider each of these separately, focusing on one at a time and then aiming to bring the individual methodologies together in assembly of a complete bilayer-bilayer complex.

As indicated in Figure 1.2 the proteins are in many cases the only such constituents for which we may already have structural data. The extent and quality of this can vary considerably. For instance, in the protein examples shown, there is a crystal structure of a complex between the two bridging proteins (yellow and blue), but this only includes their ectodomains. It would therefore be necessary to model in the transmembrane regions in this case, as well as any missing loops or membrane-proximal regions. In other cases, there may be full-length structures of components, structures only of the unliganded components, or very little structural data available at all. Although obtaining reasonable starting models of the protein participants in a membrane association is of considerable importance, I do not consider it in this work as there already exist a number of tools that can be used for this, including homology modelling and loop building methods[62, 63].

The following chapters therefore focus on the other elements indicated in Figure 1.2. Specifically, Chapter 3 considers how to simulate two bilayers vertically stacked in a simulation cell without potentially artificially restricting their dynamics, whilst Chapters 4 and 5 focus on the modelling of lipid- and protein-linked glycans. Each of these individual components, and the particular challenges they pose, are introduced in the relevant chapters, and therefore not discussed further here.

Chapter 2

BACKGROUND TO MOLECULAR DYNAMICS SIMULATIONS

2.1 Introduction

As noted in the previous chapter, the work presented herein centres around molecular dynamics (MD) simulations. Whilst far from the only computational method available for studying biological systems, MD in particular has garnered popularity in the field of structural biology. Before progressing to the methodological developments presented in this work, I first describe the theoretical basis of MD as well as some of the fundamental limitations and assumptions associated with the technique. It is neither possible nor, for understanding the present work necessary, to go into the finest of details here. For a thorough discussion of MD and other simulation methods, there exist several publications[64–66].

MD is a computational method in which the time-dependent behaviour of a system of particles is calculated using methods from molecular mechanics. If the input models can be assumed reasonably accurate and the time over which the simulation conducted reasonably long (the limit of 'reasonably' being dependent on the questions asked), then one can deduce from such simulations features about the biological system under investigation. In this chapter I briefly outline the theory behind MD simulations, the general steps in carrying out a simulation and in particular those features necessary for simulating lipid bilayers. Finally, I give a brief overview of the progress in simulating more realistic membrane-based systems using MD in order to put my work in context of the wider field.

Ultimately, the success of any modelling endeavour is critically dependent on defining the objective of the study, and then making use of appropriate simulation and analysis techniques to achieve the aims set out at its commencement. If such well-suited techniques are not already available, they must be developed, or the objectives reconsidered in view of what tools might instead be used. For example, I found that there are no clear and robust methods already established for modelling membrane-membrane associations using MD, although the underlying requirements are present, and therefore undertook the work presented in this thesis to help towards the goal of simulating such systems.

2.2 Theory

Introduction

By iteratively solving Newton's equations of motion for an ensemble of interacting particles representing the system of interest, one can build up a picture of how that system evolves over time. The resulting trajectory contains information on each individual particle and its interactions with the rest of the system. Through various different types of analyses, properties such as structural fluctuations, approximate energetic quantities, and individual intermolecular interactions can be assessed. Using statistical mechanics, the microscopic details evaluated over many timesteps can also be used to calculate values for macroscopic variables such as diffusion coefficients.

At the core of every MD simulation are two elements: an algorithm for evaluating the equations of motion for all particles present, and the force field describing the interaction potential between them.

Evaluating the equations of motion

In an MD simulation, Newton's equations of motion are solved for a system of interacting particles in order to deduce how their positions in the simulation cell vary over time. At timestep $n + 1$, we need to calculate the coordinates $\mathbf{r}_{n+1}$, velocities $\mathbf{v}_{n+1}$, and forces $\mathbf{F}_{n+1}$ of all particles, given the corresponding values of these quantities at timestep n.

For each particle i in a system of N particles:

$$\mathbf{F}_i = m_i \mathbf{a}_i \tag{2.1}$$

where $\mathbf{a}_i$ and m_i are the acceleration and mass of particle i, respectively, and $\mathbf{F}_i$ is the force acting upon that particle due to interactions with the other particles in the system. The force can also be obtained from the gradient of the potential U, which is a function of the positions of the N particles $\mathbf{r}^N$:

$$\mathbf{F}_i = -\frac{\delta U(\mathbf{r}^N)}{\delta \mathbf{r}_i} \tag{2.2}$$

Or, more generally, in terms of Hamilton's equations of motion, which relate the positions and momenta of the particles to the total energy of the system,

represented by the Hamiltonian $\mathcal{H}$. The instantaneous positions $\mathbf{r}$ and momenta $\mathbf{p}$ of all particles in a system describes a microscopic state of that system.

$$\mathbf{p}_i = -\frac{\delta \mathcal{H}}{\delta \mathbf{r}_i} \tag{2.3}$$

$$\mathbf{r}_i = \frac{\delta \mathcal{H}}{\delta \mathbf{p}_i} \tag{2.4}$$

$$\mathcal{H}(\mathbf{p}_i, \mathbf{r}_i) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}_i) \tag{2.5}$$

For a closed system of N particles, each of which has Cartesian momentum $\mathbf{p}_i$ and coordinate $\mathbf{r}_i$, the Hamiltonian $\mathcal{H}$ can be written as a sum of the kinetic energy E_K and potential energy U, each of which is a function of the coordinates and momenta of the constituent particles.

$$\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{2.6}$$

$$\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N) \tag{2.7}$$

$$H(\mathbf{r}, \mathbf{p}) = E_K(\mathbf{p}) + U(\mathbf{r}) \tag{2.8}$$

The kinetic energy usually takes the quadratic form:

$$E_K = \sum_{i=1}^{N} \frac{1}{2m_i}(p_{ix}^2 + p_{iy}^2 + p_{iz}^2) \tag{2.9}$$

where m_i is the mass of particle i, and p_{ix}, p_{iy}, and p_{iz} are its momenta in the x, y, and z directions, respectively. The functional form of the potential energy function is considerably more complicated, and is described using a force field, discussed in section 2.2. With both the kinetic and potential energy functions known, the time evolution of a given system governed by the Hamiltonian can be determined by solving the equations of motion.

That is, provided we have the potential function, it is possible to integrate the trajectory $x(t)$ of the system over time. However, this is an N-body problem,

which in the vast majority of cases cannot be solved, but rather an approximation to the solution must be obtained by applying one of various possible finite difference integration methods. These methods are based on a Taylor series expansion, illustrated below for the position **r**, velocity **v**, and acceleration **a**, where **b** is the third derivative. As Δt tends to zero, we get closer to the true values of the derivatives.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t) = \frac{1}{2}\Delta t^2\mathbf{a}(t) + \frac{1}{6}\Delta t^3\mathbf{b}(t) + \ldots \tag{2.10}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \Delta t\mathbf{a}(t) + \frac{1}{2}\Delta t^2\mathbf{b}(t) + \ldots \tag{2.11}$$

$$\mathbf{a}(t + \Delta t) = \mathbf{a}(t) + \Delta t\mathbf{b}(t) \tag{2.12}$$

There are several algorithms that might be used to perform these calculations. In order for any such algorithm to be of practical utility, it must not only be computationally efficient, but it must also both conserve energy and momentum (an important feature for linking MD to statistical mechanics) and be time-reversible. The practical consequence of time reversibility is that MD is deterministic. For a given set of initial positions and velocities, the time evolution of the system is entirely determined by the current state.

To illustrate the process of carrying out these integrations, I discuss the Verlet algorithm, in which the Taylor expansions for the positions at times t and $t - \Delta t$ (the current and previous timestep, respectively) are combined.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t) + \frac{1}{2}\Delta t^2\mathbf{a}(t) + \ldots \tag{2.13}$$

$$\mathbf{r}(t - \Delta t) = \mathbf{r}(t) - \Delta t\mathbf{v}(t) + \frac{1}{2}\Delta t^2\mathbf{a}(t) - \ldots \tag{2.14}$$

Summing equations 2.13 and 2.14 together yields:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \Delta t^2\mathbf{a}(t) \tag{2.15}$$

Thus, using the positions of particles in the system at the current and previous timesteps (t and $t - \Delta t$), and the accelerations at the current timestep (t), the

Verlet algorithm is able to approximate the new positions of the system $\mathbf{r}(t+\Delta t)$. The accelerations themselves can be obtained by dividing the forces acting on the particles by the relevant masses, and the velocity may be computed using:

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} \tag{2.16}$$

Because the position vector $\mathbf{r}$ at time $t + \Delta t$ is calculated from the position vector at time t and $t - \Delta t$, the Verlet algorithm is a two step method, and the computer must store positions from three consecutive timesteps in order to do the velocity calculations, which consumes additional computer memory. Furthermore, the Verlet scheme treats velocities as less important than positions, which is inconsistent with an ergodic system, which requires that the phase space trajectory depend equally on positions and velocities.

An improvement on this is the velocity Verlet algorithm. In this formulation, the velocity is first calculated at step $n + \frac{1}{2}$, and then the coordinates at step $n + 1$. Using this new set of coordinates, the potential energy function is evaluated, allowing new forces to be obtained, and the velocity then advanced by another half timestep in order to produce the new value at step $n + 1$.

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \Delta t\mathbf{v}(t) + \frac{1}{2}\Delta t^2\mathbf{a}(t) \tag{2.17}$$

$$\mathbf{v}(t+\Delta t) = \mathbf{v}(t) + \frac{1}{2}\Delta t[\mathbf{a}(t) + \mathbf{a}(t+\Delta t)] \tag{2.18}$$

The velocity Verlet algorithm does away with the problem of initial velocities that arises in the Verlet algorithm as a result of its two-step formulation.

First, the positions at the next timestep are calculated as in Equation 2.17. This is then used to obtain the accelerations at this new time from the potential (which is dependent on r), and from this, the velocities are obtained using Equation 2.18. This assumes the accelerations are dependent only on positions and not too on velocities.

Another modified variant of the Verlet algorithm is the leapfrog algorithm. Here, the velocities $\mathbf{v}$ are calculated at time $t + 1/2\Delta t$. These are used to calculate the positions $\mathbf{r}$ at time $t + \Delta t$. The velocities therefore leap over the positions,

and the positions leap over the velocities. This holds the advantage of explicitly calculating the velocities, but not at the same time as the positions.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{1}{2}\Delta t)\Delta t \tag{2.19}$$

$$\mathbf{v}(t + \frac{1}{2}\Delta t) = \mathbf{v}(t - \frac{1}{2}\Delta t) + \mathbf{a}(t)\Delta t \tag{2.20}$$

Ultimately, what this amounts to is a procedure in which, at each step, the forces on the constituent particles are computed and combined with the positions and velocities to determine the positions and velocities at the next timestep, assuming the force acting on each particle to be unvarying in this interval. The particles are then moved to the new positions, the forces updated, and a new cycle commenced. One of the advantages of the leapfrog algorithm is that only one set of positions and one set of velocities need to be stored, as compared with the Verlet algorithm in which if velocities are desired, positions for three consecutive timesteps must be stored. It is the leap frog algorithm that I use for all simulations presented in this work, as although it is less accurate for some applications than other integrators, it is more computationally efficient and for the work presented here expected to be sufficiently accurate. Indeed, it is the default integrator for the GROMACS[67, 68] suite, used throughout this work, due to the high computational efficiency and high accuracy for the majority of applications.

A component of the above example algorithms that is of considerable importance when preparing and running MD simulations is the timestep Δt. The value of this parameter governs the frequency with which the integration is performed. By increasing the timestep, it is possible to simulate for longer times, suggesting a larger timestep to be desirable. Indeed, in general one aims to use as large a timestep as possible in order to push the limit on simulation time as much as possible. However, it is in practice necessary to limit the timestep due to instabilities that arise in the MD integration algorithms at large timesteps. If the timestep chosen is too high, it results in a large truncation error in the integration, which results from the use of finite difference methods that only approximate the true solution. Such errors can cause the total energy of the system to increase rapidly with time. One consequence of this is the system exploding: two particles collide as a consequence of the too-high timestep propagating the particles into

a near-overlapping configuration, resulting in strong repulsive forces that propel the particles apart.

Often for MD simulations, the timestep must be in the range of 1 fs or lower in order to ensure accurate evaluation of the fastest motions in the system: the vibration of bond length between two atoms. If constraints such as LINCS[69] or SHAKE[70] are used to fix the bond lengths between all atoms, it is possible to increase the timestep (e.g. from 1 fs to 2 fs), but even with such measures the low timestep is one reason why MD simulations cannot routinely access timescales beyond the range of a few hundred nanoseconds without additional steps taken to simplify the system, as discussed later in this chapter.

Statistical mechanics

During an MD simulation, one obtains very detailed information about the positions and velocities of each individual particle present in the system. Whilst such data can be of considerable interest, it is often not possible to directly relate it back to phenomena that can be observed or measured in experiments, which frequently monitor macroscopic features of the system as a whole. For instance, the exact locations of waters inside a pore is likely to be of less interest than overall rate of their passage through it when considering the pores degree of openness.

The relationship between macroscopic and microscopic properties of a system is the purview of statistical mechanics. Here, I discuss only some basic concepts that are of immediate relevance to the field of MD simulation. For more information, there are a great many publications that deal with the field as a whole, and its relation to computer simulations and biomolecular investigations in particular[65, 66, 71, 72].

In general, MD simulations scale by either $O(N \log N)$ or $O(N)$, meaning that simulations with macroscopic numbers of particles are computationally intractable – they would simply take far too long and far too much computational resource to carry out in a realistic time. It is therefore necessary to employ statistical mechanical methods to extract information on macroscopic properties from MD simulations.

Statistical mechanics is able to provide a link between the microscopic details of a system and its macroscopic observables, which reflect the time average behaviour of atoms at equilibrium. The underlying concept is that a given macroscopic

observable could arise from any of a multitude of distinct microscopic states of the system. An ensemble is all of these possible microstates considered at once. Macroscopic observables are thus formulated in terms of ensemble averages, obtained by averaging over many copies of the same system, each in a different configuration.

An ensemble that does not evolve over time can be said to be in statistical equilibrium. This characteristic, along with certain fixed macroscopic observables, defines the thermodynamic ensembles that are considered when carrying out MD simulations. In particular, the fixed variables might be energy E, temperature T, pressure P, volume V, number of particles N, or chemical potential μ. An ensemble where, for instance, number of particles, volume, and energy are constant (the microcanonical ensemble) is denoted by NVE. Other examples include the NVT (canonical), NPT (isothermal-isobaric), and μVT (grand canonical) ensembles. In experimental terms, the variables in brackets might be viewed as control parameters, specifying the experimental conditions. These different statistical ensembles are associated with distinct statistical characteristics, and can be viewed as akin to certain experimental conditions. For example, the canonical ensemble might be seen as a closed system that is in weak thermal contact with a heat bath so that it cannot exchange particles, but the total recorded temperature does not vary.

If the positions and momenta of all particles in the system are considered as coordinates in a $6N$-dimensional space, the phase space, then at any given time the system corresponds to a point in this space. The time evolution of the system is thus a trajectory in the phase space. A given ensemble can be viewed of as a collection of systems with the same Hamiltonian, and which therefore may be considered replicas of each other. Each corresponds to a point in phase space, and the collection of these points is a statistical ensemble. An observable A may then be computed by averaging its values across all members of the ensemble.

$$A_{obs} = \langle A \rangle_{ens} \tag{2.21}$$

where $\langle A \rangle_{ens}$ is the ensemble average of A

A central assumption behind the equations used to calculate ensemble averages is that the system has an equal probability of occupying any of the microstates corresponding to a given macrostate. In an MD simulation, it is sought that the

proper phase space distribution should be generated, from which the ensemble averages of various observables can be computed using equations appropriate for the particular ensemble. This relies upon the assumption that we have, in the course of the MD simulation, generated the correct phase space distribution so that an observable can be calculated as a time average from the simulation trajectory. This equivalence is called the ergodic assumption[65]. Therefore, a principle aim in MD simulations is the generation of the desired phase space distribution.

One of the most common limitations faced in MD simulations is that of computational power. The duration of a simulation depends on the number and size of the timesteps, between each of which the forces are recalculated. The timestep must be small enough to avoid discretisation errors. However, the number and length of timesteps must also be large enough that the overall simulation time is sufficient to model the phenomena of interest, without taking intractably long real-time. Ideally, it is hoped that the simulation is run long enough to explore all possible phase space points in the ensemble.

Force fields

The actual form of the potential energy component of the Hamiltonian of a system is far too complicated for its computation to be achieved in any practicable period of time. It is therefore necessary to approximate the form of the potential energy before we can hope to efficiently calculate it.

A force field, in the context of MD, can be defined as an equation describing the formulation for the potential energy for the system – it gives the potential energy U of a set of particles according to their positions r. Conceptually, force fields can be split into at least two parts: the intramolecular (bonded) and intermolecular (nonbonded) interactions. Among the former are the stretching, angular opening, and rotation about bonds, whilst the latter includes the van der Waals and electrostatic interactions (Equations 2.22–2.28, with parameters defined in Figure 2.1).

$$U = U_{bond} + U_{angle} + U_{dihedral} + U_{improper} + U_{vdw} + U_{elec} \tag{2.22}$$

$$U_{bond} = \frac{1}{2} K_b (b - b_0)^2 \tag{2.23}$$

$$U_{angle} = \frac{1}{2}K_{\theta}(\theta - \theta_0)^2 \tag{2.24}$$

$$U_{dihedral} = \frac{1}{2}K_{\chi}[1 + cos(n\chi - \delta)] \tag{2.25}$$

$$U_{improper} = \frac{1}{2}K_{\psi}(\psi - \psi_0)^2 \tag{2.26}$$

$$U_{vdw} = \epsilon[(\frac{R_{ij}}{r_{ij}})^{12} - 2(\frac{r_{ij}}{r_{ij}})^6] \tag{2.27}$$

$$U_{elec} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.28}$$

where r_{ij} is the distance between the interacting particles, the vacuum permittivity ϵ_0 is a constant of value $8.854x10^{-12}Fm^{-1}$, σ is the finite distance at which the potential energy becomes zero, ϵ the depth of the potential well, $K_{parameter}$ a force constant, $parameter_0$ the equilibrium value of $parameter$, and all other parameters illustrated in Figure 2.1.

Here, the first four equations (2.23–2.26, illustrated in orange in Figure 2.1) describe the intramolecular or local contributions to the overall energy. These include bond stretching and angle bending terms, as well as proper and improper torsional terms for systems with more than four atoms in a row and situations where planarity needs to be maintained. The second part of the force field (equations 2.27 and 2.28, blue in Figure 2.1) is the intermolecular term, which includes the van der Waals interactions between atom pairs (often described using the 6–12 Lennard–Jones potential), and the electrostatic interactions between charged/partially charged particles (generally described according to Coulomb's law). In addition to these two terms, there may be additional elements of the force field to, for instance, help correct deviations from experimental spectral data or improve the treatment of hydrogen bonding interactions.

A force field also has to have a means to handle nonbonded interactions between atoms of different types, which may have, for instance, dissimilar Lennard–Jones parameters. This is achieved by means of combination rules, sets of equations that dictate how the nonbonded interaction potential is calculated from the parameters of the respective atoms[73, 74]. There are several different combination

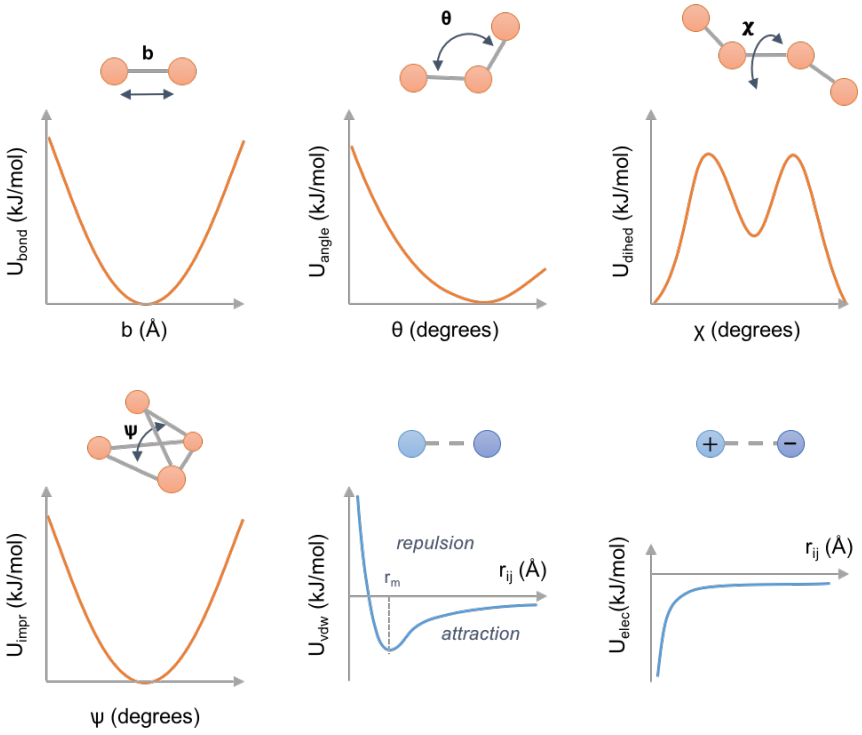


Figure 2.1: Illustration of the bonded and nonbonded interactions commonly included in MD force fields. The bond length term (top left) relates to particles connected by a covalent bond, modelled using a harmonic potential in many cases. The bond angle term (top middle) exists between three bonded particles, and is often also modelled using a harmonic potential. Both the dihedral (top right) and improper (bottom left) terms involve four bonded particles, and are defined as the angle between the planes containing the first three and last three of these. The two nonbonded terms (bottom middle and bottom right) described interactions taking place between particles not directly linked, and include the van der Waals interactions that arise between any two particles, and the electrostatic attraction or repulsion between pairs of charged particles.

rules available for the Lennard–Jones, and other nonbonded, potentials, whose choice can have an influence on the outcome of a simulation[74]. Among the most commonly employed, and perhaps simplest, are the Lorentz-Berthelot rules[75, 76], in which the modification involves simply taking the geometric and arithmetic means of the individual σ and ϵ values, respectively. Alternatives intended to improve upon the limitations of these[77, 78] and/or for application to different interaction potential forms have also been developed[79, 80].

The values that populate a given forcefield are generally obtained empirically. For instance, the equilibrium values for bond lengths and angles may come from high resolution structural data from X-ray or neutron diffraction, while force constants may be obtained from e.g. vibrational spectra and heats of vaporisation. As there are often limited experimental data, *ab initio* methods may also be used

as a major source of input data.

An important consideration for any MD force field is computational efficiency, and for this reason it is relatively common for many of the elements in the potential function to be relatively simple. For instance, both bond stretching and bond angle opening/closing are often described using a harmonic potential (equations 2.23 and 2.24). Often, the spring constants in these harmonic equations are relatively high, meaning that large deviations from the equilibrium values are strongly disfavoured, and therefore the bond lengths and angles tend to be kept relatively invariable. This is true also for many improper dihedrals equation 2.26.

Some improper potentials and many dihedral potentials tend to be somewhat 'softer' than the bond lengths and angles in biomolecular simulations, thereby endowing simulated molecules with greater flexibility. This, for instance, allows molecules such as proteins to undergo potentially substantial conformational changes. A dihedral potential controls rotation around the central bond of a group of four linked particles, and is used, for instance, for the Φ-Ψ backbone dihedrals of peptides. An improper, however, controls the planarity of the entire group of four particles, so might therefore be used to maintain the planarity of aromatic rings or of the peptide bonds themselves.

Particles that are not directly connected by a covalent bond also interact with each other, and indeed may do so in various different ways. All atoms interact through a short-range repulsion related to electron orbital overlap. Over longer ranges, there can be attractive associations due to dispersive forces or, in the case of charged or partially charged species, electrostatic interactions that may be attractive or repulsive depending on the relative properties of the particles.

The two nonbonded terms in the potential function reflect these interactions. The electrostatic term relates to interactions between two charged atoms (equation 2.28). The values for the point charges belonging to polar molecules can be derived from quantum mechanical simulations, often with subsequent adjustments using empirical data. The van der Waals term reflects those interactions that arise from induced dipoles and excluded volumes of pairs of atoms (equation 2.27).

2.3 Choosing a force field

General considerations

A complete force field includes the functional forms of the terms contributing to the potential energy of the system, and values for all of the parameters in these terms. Although often taking a similar functional form, the constituent equations, parameters, and their means of derivation can vary significantly between different force fields. It is of critical importance to the success of any MD endeavour that the force field chosen is well-suited for the system and phenomena of interest.

A complete discussion of the many biological force fields available at present is beyond the scope of this thesis – for more information the reader is referred to recent reviews on the subject [81–85]. However, I briefly discuss here some of the considerations that need to be made when electing a force field to be used. Later in this chapter I discuss in more detail the Martini force field[49], which is employed throughout much of the present work.

One of the first decisions when considering which force field to use is the resolution, that is, the level of molecular detail with which the system components are to be described. Within classical mechanics simulations, this can range from the inclusion of all atoms in every molecule, to systems in which the solvent is only implicitly present and proteins are modelled as single interaction sites (Figure 2.2). Higher resolution offers greater insight into molecular details, and in theory is associated with more faithful reproductions of the properties of the system. Lower resolution, on the other hand, allows one to access longer simulation times and larger systems, and so might facilitate the examination of phenomena more relevant to the biological events or experimental observations.

The force field must also be compatible with the simulation software to be used, and must either already include parameters for all components of the system of interest, or provide a basis upon which these can be parameterised. For example, many lipid-containing force fields might not cover all headgroup or tail types required for a biologically realistic simulation, particularly when one is looking at very specialised membranes or those belonging to less-frequently studied species. If parameters are not available and cannot be developed, then it is either necessary to opt for a different force field or to adjust the planned system to account for the limitations imposed by the available parameter set.

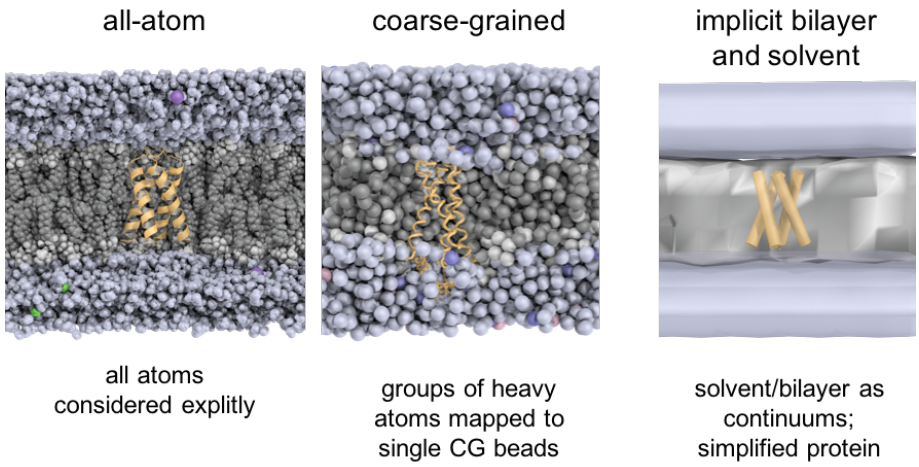


Figure 2.2: Simple illustration of the range of resolutions used in MD, from including all atoms to simplifying all components to either single interaction sites, simple geometric approximations of the true shape, or continuum environments.

Atomistic forcefields

Atomistic (AT) force fields, in which all heavy atoms are explicitly represented, can provide sub-nanometre information about individual atoms with femtosecond temporal resolution. However, whilst in theory associated with a higher degree of accuracy, the sheer number of particles that must be considered in calculations makes ATMD simulations extremely computationally expensive, generally limiting them to tens of nanometres and a few microseconds.

Broadly speaking, AT force fields can be classified into all-atom and united atom. In the former case, as the name suggests, all atoms are included in the models. This is true for the AMBER[86], CHARMM[87], and OPLS-AA[88] force field families, which have been used extensively for biomolecular MD simulations. Both CHARMM and AMBER have a number of compatible lipid models that allow their application to studies of membrane phenomena. Indeed, the CHARMM family of force fields includes parameters for proteins, nucleic acids, some small molecules, glycans, and lipids[89–91]. CHARMM36, a protein force field within the family, has associated parameter sets for various different lipids found in biological membranes, with a wealth of choice regarding tails and headgroups available. Furthermore, it can be used within the CHARMM-GUI online resource for constructing simulation systems, including bilayers and vesicles, making the forcefield very readily accessible for such simulations[87, 89, 92–96].

AMBER, however, is most often used for simulations of protein, nucleic acid, and druglike molecules, and has generally very advanced parameters for such molecules. Although there have been efforts to extend it to include lipids, including the Lipid14[97, 98] parameters, this force field has not been as widely extended into the lipidic realm as CHARMM. Slipids[99] are also parameterised to be consistent with the AMBER force field, and currently include several phosphatidylcholine (PC) phospholipids, as well as phosphatidylglycerol (PG), sphingomyelin, phosphatidylserine (PS), phosphatidylethanolamine (PE), and cholesterol[99, 100].

At a slightly less detailed level of resolution, united atom force fields remove some degrees of freedom by combining certain groups of atoms into pseudoatoms. For example, methyl groups and methylene bridges can be merged into a single particle, thus effectively incorporating the properties of the hydrogen atoms into the carbons to which they are bonded. The GROMOS family of force fields are a frequent example of this kind of force field, and include parameters for proteins and many lipids, as well as other molecule types[101–104].

Deriving accurate parameters for force fields is no trivial matter, and within any force field there will be biases and limitations that render it more accurate or appropriate for certain applications than others. One of the sources of difficulty is the limited amount and variable quality of experimental or *ab initio* data upon which the parameterisation might be based. Force field development generally involves first identifying suitable reference data, such as quantum mechanical calculations and experimental properties, then defining a means to measure force field quality, and refining the parameter set to improve the force field according to this assessment. The interdependency between the parameters is generally nontrivial, and so it is necessary to employ sophisticated methods to achieve optimisation, such as evolutionary algorithms. Ultimately, the quality of any force field depends critically on the data used in its development and the weights assigned to different reference properties, whilst overfitting to the data can yield to limited transferability.

It is ultimately assumed that similar chemical groups belonging to different molecules will behave in the same way. This assumption is necessary as it allows a force field developed based on reproducing data for a small set of molecules can be applied to other, generally more complex, systems. Furthermore, it would be impossible to derive a forcefield based on studies from, for instance,

a particular set of bilayer lipids as there is unlikely to be sufficient data for that system to populate all the needed parameters of the force field. The assumption that the parameters will hold for sets of molecules different from those used in parameterisation is referred to as the transferability assumption, and although required for one to use most force fields in practical applications, it does not necessarily hold perfectly true. Approximations are made when transferring data from small molecules used in parameterisation to larger biomolecules, and the experimental data used in parameterisation may have been collected under different conditions and it is not always easy to account for this. An important consequence of the limitations inherent in parameterisation of force fields is that energy functions are not transferable from one force field to another.

Force fields of this type also rely on assuming that the Born-Oppenheimer approximation holds. The Born-Oppenheimer approximation allows one to separate the electronic and nuclear motions and therefore consider only nuclear coordinates when evaluating the energy of a system. It is additionally assumed, when using these force fields, that the potential energy of the system can be written as a sum of individual potential energies.

All the force fields discussed thus far are referred to as non-polarisable force fields, meaning that dipole moments in the system are fixed at the average value. They are therefore unable to capture the effect of induced polarisation, the redistribution of electronic structure of an atom that arises due to changes in its local electrostatic environment[105]. For many applications, this is of little consequence. However, for some studies (e.g. examining hydrophobic gating in pores[106]) it can be a significant limitation. Over recent years, there has been both recognition of the limitations of non-polarisable force fields and efforts to remedy the problem, with the development of a number of polarisable force fields for application to water and biomolecular systems[107]. Examples include the CHARMM Drude force field[108], and the AMOEBA force field[109], which includes parameters for proteins and lipids, as well as water and ions[110–112].

In principle, in environments with variations in the dielectric, polarisable force fields ought to be more accurate and transferable than non-polarisable ones. However, difficulties in deriving parameters for these force fields may be more challenging than for classical MD forcefields, and if poorly developed parameters are used, it can have implications for performance of the force field under certain conditions. Furthermore, they come at a cost of computational time, which is an

increasing problem as the system gets larger and/more compositionally complex.

Coarse-grained forcefields

In a coarse-grained (CG) force field, sets of heavy atoms are grouped into single interaction sites in order to increase the efficiency of MD simulations, often by three or more orders of magnitude [113]. The popularity of CG models has been generally increasing as the simulation community moves towards trying to model ever more experimentally and biologically realistic systems, resulting in greater computational demands that cannot be met within the spatiotemporal limits of ATMD[114].

In coarse-graining, we reduce the number of interactions that need to be calculated in each timestep, and smoothen the potential energy landscape (Figures 2.3). This allows simulations of micron-sized systems for tens of microseconds, or even milliseconds. A consequence of reducing the degrees of freedom is a loss of much of the friction between the CG particles, and therefore results in the observed faster dynamics. Whilst advantageous from the point of view of efficient sampling of phase space, this presents a challenge for evaluating time-dependent properties: the extent of the thus achieved accelerations will vary with the particulars of the CG force field, and it can be difficult to accurately interpret timescales viewed in CG simulations. It is possible to address this concern by rescaling the simulation duration using a simple factor to convert the CG time to the 'true' time, this factor having been obtained by comparing diffusion coefficients obtained from CGMD with those from reference ATMD simulations.

The level of coarse-graining can vary from a mapping of 3-6:1 (i.e. 3-6 heavy atoms per CG particle), which retains some chemical specificity, to models in which solvent is present only as an implicit continuum and macromolecules may be reduced to single interaction sites. It is, of course, important that the coarse-graining is not so severe that the force field can no longer reproduce the physicochemical properties of the molecules that underlie the phenomena of interest, therefore a compromise must be made between speed up and accuracy.

Parameterising a coarse-grained force field

In principle, the process of developing a CG force field for molecular simulations requires two steps. The first is to decide on how the system of atoms shall be converted into a simpler system of CG sites, that is, the representation must be

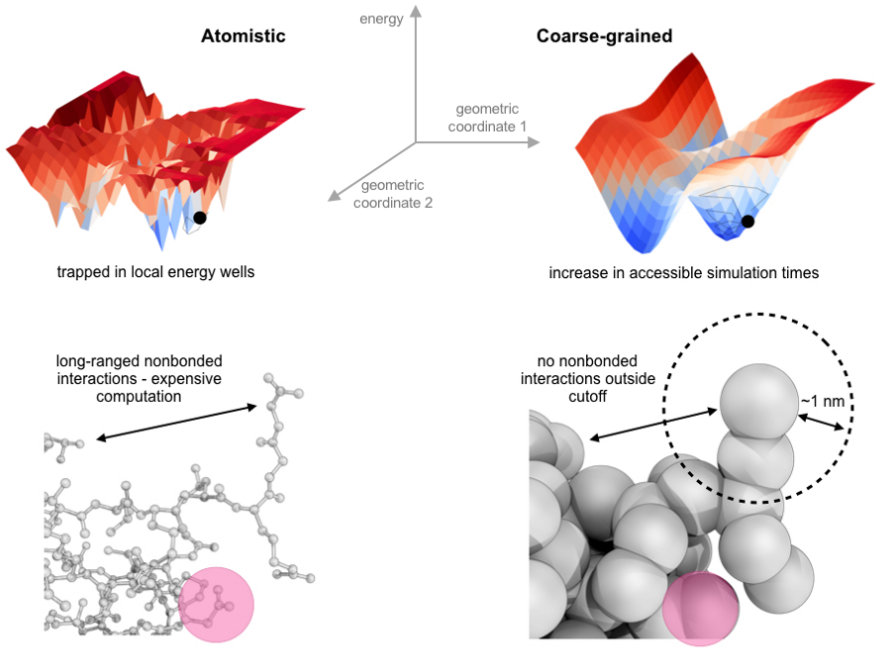


Figure 2.3: A representation of some of the differences between atomistic and coarse-grained forcefields. The energy landscape for the simulated system, such as the hypothetical simple one shown, is smoothed upon coarse-graining so that it becomes easier to explore and the system is less likely to become trapped in local energy wells. Nonbonded interactions are also cutoff in many cases, rather than computationally expensive methods being required to model them. Combined with there simply being fewer particles for a given molecule, this reduces the pairwise interactions that need to be calculated at each timestep. Additionally, the loss of high frequency motions of atomic bonds allows use of a larger timestep, further helping to render more accessible longer timescales on coarse-graining.

decided upon[115]. The manner in which this is done should be appropriate for the system and phenomena of interest. Within this process, there are two stages: system mapping, in which one specifies the number, types, and connectivity of the sites, and coordinate mapping, wherein the configuration of the CG model as a function of the AT model is deduced [116]. Here, I discuss only briefly the approaches that may be taken to parameterise CG force fields. For comprehensive reviews on CG force fields and their development, see [117–131].

The chosen mapping will be governed by the desired accuracy, efficiency, and transferability of the model. The ideal CG mapping should preserve all features necessary to describe the phenomena of interest and to capture the relevant slow, large amplitude motions of the molecules involved, while eliminating sufficient detail to significantly increase computational efficiency[130]. It is also important that high frequency, low amplitude fluctuations with only weak coupling to the

slower global motions be filtered out, and the force field should be such that the physical forces underlying the phenomena being investigated can be treated and interpreted clearly. Whilst in most cases, the CG interaction sites are considered point particles with no internal structure, more complex models are possible at the cost of reduced efficiency. For instance, by creating anisotropic sites it is possible to better recreate the geometry of the underlying atoms [124, 132–135].

There are several methods for determining the CG mapping with different underlying theories and levels of resolution. One can design high resolution mappings that allow accurate reconstruction of AT structures or potential energies [136–141]. For low resolution models of large biomolecules or complexes, one might employ systematic optimisation by considering correlated fluctuations and experimental density maps [131, 142–150]. In many cases, mapping is based on, or relies at least to some extent on, chemical intuition and understanding of the chemistry of the AT system. This is aided by the fact that in most cases, CG mapping is initially designed with a specific type of molecule in mind.

Conceptually, the parameterisation of a CG force field can take a top-down or bottom-up approach [115, 151, 152]. In the former case, the aim in developing the force field is to accurately reproduce experimental properties such as density, heat of vaporisation, and partitioning data. In the latter situation, parameters are extracted from references such as ATMD simulations and structural data banks, with the goal of reproducing structural features. This can lead to very accurate CG models, albeit with an accuracy limited to the state point at which those parameters were derived, and which can be difficult to interpret from a physicochemical point of view. Whilst potentially yielding less structural accuracy, the top-down method can better reproduce global partitioning of molecules over more state points. In practice, to maximise both accuracy and transferability, most CG force fields use a combination of these parameterisation approaches. However, it is important to note that it is impossible for a CG force field to achieve full transferability or accuracy – there will always need to be some compromise of detail owing to the lost degrees of freedom in coarse-graining, as well as some bias that results from the nature of the parameterisation data sources.

CG force fields for studing biomembranes

There have been efforts towards CGMD applications to membrane-like systems for many years. This started in the 1990s with simulations of micelles and sur-

factant bilayers by Smit *et al.*[153] and Goetz and Lipowsky[154], respectively, and since these early developments, CGMD has been widely applied to study features of biological membranes with considerable success[155–158]. A range of force fields have been developed with a view to studying such systems, or at least extended to include parameters required for their investigation.

Among the most common CG force fields used in studying membranes and associated phenomena is the Martini force field, which now includes parameters for lipids[49], glycolipids[159], proteins[160, 161], carbohydrates[162], and nucleic acids[163, 164]. Additionally, it is compatible with a polarisable water model[165], although use of this comes at some cost of efficiency, and it is not clear the extent to which its use is advantageous. Not only does Martini benefit from a modular construction and extensive force field, but it is well documented and associated with a number of freely available tools for use in analysis, resolution transformation, and system construction. The force field is discussed in considerably more detail later in this chapter as it is the force field that I use for much of the work presented hereafter.

There are a number of other CG forcefields at similar levels of resolution to Martini. For instance, the SDK model is based on a 3:1 mapping[166], and thus is slightly higher resolution. Whilst the model has been successfully applied to investigate the phase behaviour of lipid monolayers and vesicle fusion[167], it includes a relatively small range of lipid parameters and none for proteins. A more recent forcefield, SPICA, has been developed using the same basis, and this includes a more extensive array of biomolecules[168]. The ELBA model also uses a 3:1 mapping, but each water model is represented individually using soft, sticky dipole potentials[169]. The work done using this forcefield has mostly focused on studying the permeation of drugs and other compounds though bilayers, with only a few lipid types modelled thus far. This force field also incorporates dipoles into lipids and water beads, helping to improve the description of electrostatics.

SIRAH[170, 171] is a top-down CG force field with a similar mapping as Martini and explicit treatment of water and ions, currently including parameters for lipids, proteins, nucleotides, and aqueous solvent[172]. Within the parameter set, there is however a variation in the level of resolution depending on the specific molecule in question e.g. higher resolutions for polar or aromatic groups as compared with hydrophobic ones to reflect the distinct nature of interactions

in which such moieties are typically involved. Having been parameterised by fitting to structural data, it requires no constraints to fix protein secondary structures as required in e.g. Martini, and is therefore not forced to constrain protein conformations. Additional benefits derived from the use of long-range interactions and explicit water molecules with partial charged to create a dipole moment as required to place a dielectric permittivity.

An additional speedup can be achieved by combining coarse-graining approaches such as those used in the Martini and SIRAH forcefields with an implicit solvent. In this case, the water molecules that would usually be represented as individual particles are replaced with a continuum with properties that in some way mimic those of an aqueous milieu. Examples include the PLUM force field[173–175], which includes parameters for both lipids and proteins and can described protein folding, and Dry Martini[176].

Hybrid resolutions

It is not strictly necessary for the entire system being modelled to be simulated with a single resolution. Some parts can be more detailed than others, therefore helping to increase the accessible simulation time without losing information where it is in particular required. For instance, the united atom Berger lipids can be used with all atom protein force fields[177–179]. The PACE force field was designed to allow a CG membrane model to be used with a united atom protein model[180, 180], and the Martini force field has been combined with higher-resolution proteins[181].

At even greater disparity in the mixed resolutions, are those methods that combine an explicit protein representation with a mean-field model of the solvent and (if present) membrane, such as the generalised Born model in which the influence of membrane and aqueous environment are included implictly in the MD simulations[182, 183]. Here, the solvent and membrane are approximated as a mean-field environment with a specific dielectric constant, which will be different for the bilayer interior (typically 1-4) and for the surrounding water (usually 80). In some models, the dielectric constant varies along the bilayer normal to account for different layers, being higher at the interfacial regions[184].

Demonstrating the power of hybrid simulation approaches, a recent study used a hybrid multiscale approach with the SIRAH force field to study three different virus-like particles: the sesbania mosaic virus (SeMV) and tobacco rattle virus

(TrV) capsids, and the envelope of zika virus (ZIKV)[185]. In this approach, parts of the system of particular interest were modelled using AT and/or CG resolution, and then those of less direct interest used instead a supra-coarse-grained (SG) resolution. The results from the ZIKV simulations were in good agreement not only with experimental data, but also with CG simulations of other flaviviruses, providing support for the general applicability of this method. This hybrid multiscale approach allows much more efficient simulation of biomolecular systems that include large amounts of bulk solvent, as is the case for virions and virus-like particles, enabling simulations to be run on modest computational resources and thereby rendering computational virology accessible to a wide range of researchers without access to supercomputing facilities[185].

2.4 The Martini force field

Background

Among the myriad CG force fields available, Martini has perhaps enjoyed particular success in application to bilayer-containing simulations. It has been applied to detailed studies of individual membrane proteins and their interactions with lipids[186], right through to simulations of entire viral particles[44].

The force field itself was initially developed specifically for application to lipid simulations[187]. This early formulation treated all particles as having the same mass (72 Da, equivalent to four water molecules), which could interact *via* two types of harmonic potential describing bond-stretching and angle bending, and through Lennard-Jones and Coulombic nonbonded interaction potentials. The former bonded interactions were limited to a single bond length and only two bond angles, with only one option being used for the force constant in both cases. Regardless of the AT groups being captured by a CG bead, the value of the distance when the Lennard-Jones potential reaches zero (σ) was 4.7 Å, and the potential truncated between 9 Å and 12 Å . The Coulombic term was shifted to zero between 0 Å and 12 Å , with no long-range electrostatic approximation, and a relative dielectric constant of 20 was used, based on empirical data. This value, ϵ_r was included in the Coulombic potential:

$$U_{ij} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{2.29}$$

In order to capture the distinct behaviours of different atom types, this early iteration of the force field included four distinct CG bead types: polar (**P**),

nonpolar (**N**), apolar (**C**), and charged (**Q**).

bead type	represented atom types
P	partial charges, hydrophilic
A	neutral, hydrophobic
N	mix of polar and apolar
Q	charged

The nonpolar and charged particle types were subdivided further to account for hydrogen bonding capabilities: donor (*d*), acceptor (*a*), donor and acceptor (*da*), and no hydrogen bonding capacity (*0*).

Then, given these bead types, five distinct interaction levels were defined to account for the different ways in which pairs of particles would interact. For instance, the interaction level between two **P** type beads would be treated as more attractive than one between a **P** type and a **C** type. This was incorporated into the Lennard-Jones potentials listed for each pair of bead types.

The values of the various force field parameters were calculated using an alkane/ water system, and aimed to reproduce the experimental densities of the pure systems at room temperature, the mutual solubility of oil and water, and the relative diffusion rates[187].

By reducing the number of particles in the system, using short-range nonbonded interactions, and increasing the integration timestep, this CG force field could greatly increase the efficiency of MD simulations. A timestep of up to about 40 fs was facilitated owing to the various simplifications made to both atoms in general and to the nature of interactions arising between them.

It is often quoted that for this force field and its derivatives, the speed of sampling of processes are increased by a factor of four based on the increased sampling speed of water and lipid diffusion in bilayers[188], but this factor is likely not to be invariably true and in the present work all simulation times will be quoted without this correction factor of four.

This initial force field, at the time not yet having received its name, has been subsequently elaborated and refined by the same group to form the Martini force field[49, 160]. The range of particle types and their bonded and nonbonded interactions has increased substantially, while its modular philosophy, using few parameters and standard interaction potentials, has endowed it with considerable applicability and transferability, features borne out by both the diversity of

molecule types that it can be used to simulate, and the range of applications it has found in the literature[189]. Included in the molecular repertoire are a great range of lipids, including various phospholipids with different combinations of headgroup and tail types, glycolipids, and sterols[159, 190–201]. Martini is therefore ideally suited for studying various membrane-related processes and for modelling complex lipid bilayers that mimic those found in nature.

Indeed, Martini has enjoyed considerable success in both interpreting and informing experiments in this field[152, 197, 202–205]. Furthermore, there is a straightforward backmapping from Martini to an AT representation, which allows any features of the system that seem particularly interesting to be investigated in more detail using ATMD simulations.

Force field details

The basic organisation of Martini has not changed since its first preliminary variant. There remains a relatively small number of different CG bead types, each designed to represent particular sets of physicochemical properties, and a set of defined interaction potentials that can arise between these in either bonded or nonbonded associations. However, the bead types have been improved upon and extended, and the details of their interactions refined. Example uses of some of the current bead types are given in Table 2.1. Note the subcategories of **P** and **N** beads that account for hydrogen bonding capacity (d:can only donate H bonds; a:can only accept H bonds), so that even though directionality of hydrogen bonds cannot be accounted for it is possible to approximate the differences between groups capable of donating and/or accepting hydrogen bonds.

In addition to these variations in polarity and 'hydrogen bonding' capacity of the beads, each can be in any of three different sizes: normal (**N**), small (**S**), or tiny (**T**). These are typically used to account for variations from the 4:1 mapping scheme (e.g. three heavy atoms may be replaced with a small, rather than normal, bead). They can additionally be used to capture the reduced availability of atoms in branched or ring groups so that, as an example, three heavy atoms within a ring or branched structure would be represented with a **T**, rather than **S**, bead.

In practice, changing bead size alters the sigma value in the van der Waals equation, so that each pair of bead sizes has sigma appropriate to the number of underlying groups. In addition to this, the epsilon at the start of the equation is altered to account for effects of bead size on interactions specific to particular

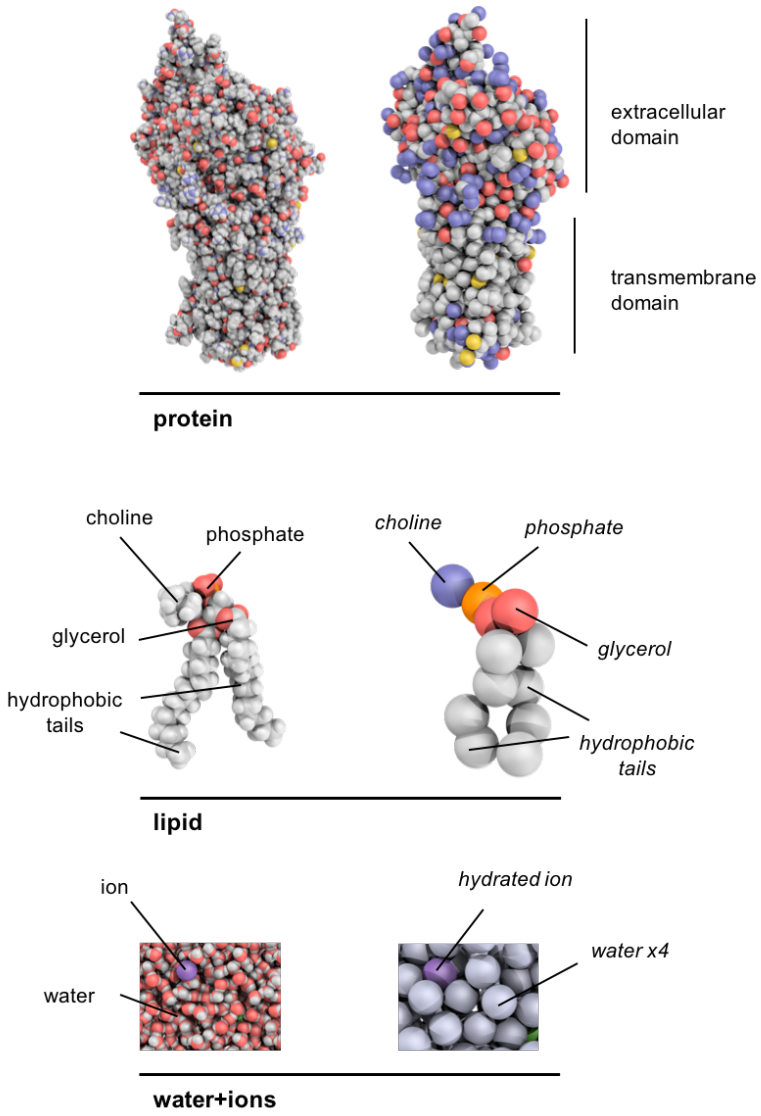


Figure 2.4: Comparison of a protein (PDB code: 6K42[206]), a lipid (dipalmitoylphosphatidylcholine (DPPC)), and solvent with ions between CHARMM36 with TIP3P water (left-hand images) and the Martini force field (right-hand images). The molecules are shown as spheres coloured according to element or (for Martini) elements 'dominating' the properties of the underlying chemical groups (grey: C, red: O, blue: N, orange: P, yellow: S, Na purple). CG water is shown as pale blue spheres for consistency with later figures.

	example groups		
bead type	**N**	**S**	**T**
C1	C–C–C–C	C–C(CH3)–C	–C(CH3)–CH3
C2		C–C–C	C–C
C3	C–C=C–C	C–C=C	C=C
C4	C=C–C=C	C–C–SH	–C-SH
C5	C–C≡C–C	C≡C–C	C≡C
N0	C–C–O–C–C	C–O–C	C–O
N1			
N2	C–NH–C–C	C–NH–C	
N3	C–C–C–OH	C–C(CH3)–C–OH	
N0d/a			
N1d/a	C–C(=O)–C–C	C–C(=O)–C	C–O
N2d/a	C–C–C=O	C–C=O	C=O
N3d/a		C–C–NH2	C–NH2
P1	C=C–C–OH	C–C–OH	C–OH
P2	C–C–COOH	C–COOH	
P3	C(OH)–C(OH)–C	C(OH)–C(OH)	
P4	C–C–CONH2	C–CONH2	
P5	NH3+–C–COO-		
P1d/a			C=O (nucleotide)
P2d/a	C(=O)–C–C(=O)		
P3d/a	C(=O)–N–(CH3)2		
P4d/a			
P5d/a			
Q0	C–C–N(C)3+	N(CH3)3+	M+
Qp	C–C–C–NH2+	C–C–NH2+	C–NH2+
Qn	C–C–COO-	C–COO-	
Q1	PO4(-1)	Cl- (H2O)2, Na+ (H2O)2	Cl-, Na+
Q2	PO4 (-2)	Ca2+ (H2O)2	Ca2+

Table 2.1: Table summarising some of the recommended bead types and sizes to use for different chemical groups. Each group would generally be part of a larger molecule, with the exception of ions. M+ refers to a heavy metal complex. Gaps in the table reflect a present lack of specific recommendation for certain bead types, not deficiencies in their parameterisation.

bead types. Thus, for example, reducing bead size will increase the strength of interactions between charged bead types, while decreasing self-interactions between beads representing hydrophobic moieties.

Each Martini bead can be considered a particle that participates in isotropic interaction potentials with surrounding particles, provided they lie within the cutoff radius. One limitation of Martini arises from this lack of directionality when it comes to intermolecular interactions – it prevents the decidedly directional na-

ture of hydrogen bonds from being captured. As hydrogen bonds play a critical role in maintaining the conformations of proteins, it is therefore necessary to incorporate an elastic network to hold proteins in their proper conformations, fixed based on that of the input reference structure[207]. Although it is possible to remove or weaken parts of the network to allow some conformational adjustments, in general the requirement for an elastic network does preclude large-scale conformational changes. Further more, the Martini beads may have an overall charge if representing a charged group (e.g. deprotonated carboxylate), but they are not endowed with partial charges of the component heavy atoms. The nonbonded interactions of uncharged molecules are therefore determined by Lennard-Jones potentials only. These interactions are all listed in a table, which describes the strengths of van der Waals interactions arising between beads of different types (Table 2.2).

There are twenty distinct levels of interaction possible, ranging from 0 (hyper attractive) to 19 (super repulsive). For every pair of bead types and sizes, there is a particular interaction strength that aims to reflect the underlying atomic properties. This is a considerable increase over the previous Martini versions, which included only nine different interaction levels, and for which only the sigma value of the van der Waals potential altered with bead size.

The bonded interactions are modelled using standard functions common to AT force fields, specifically harmonic bond and angle potentials, and multimodel dihedral potentials. Proper dihedrals are mostly used to assist in imposing secondary structure on protein backbones, which may otherwise be lost due to the aforementioned lack of hydrogen bond directionality, whilst improper potentials help to ensure planarity of planar groups is not lost.

The parameterisation of bonded interactions was based primarily on matching structural data and ATMD simulations, whilst the nonbonded interactions were parameterised based on comparison with experimental thermodynamic data, including the free energies of hydration and vaporisation, and the partitioning free energies between water and organic phases.

‘Sticky’ Martini

It has been demonstrated that several current force fields, both AT and CG, tend to overestimate the aggregation propensity of certain biological molecules, including in proteins and glycans[208–214]. It is generally thought that this

Table 2.2: Martini 3 (beta) interaction table. The row and column headings indicate the different bead types, and the numbers populating the table reflect the nature of the interaction between the indicated pair. The interactions cover the range from hyper attractive (0) to super repulsive (19). The colours correspond to different groups of interacting pairs. For each pair of bead types in a given group, the interactions scale in the same way as the bead sizes change. The groups are ion (pink), ion-other (lilac), organic (orange), and water (blue). For both the organic and water groups, the interaction strengths decrease with decreasing bead size, while the converse is true for the ion groups. Hydrophilicity increases with smaller bead sizes for the organic groups, but decreases for water and ion groups.

Q2	Q1	Qa	Qd	Q0
12	11	11	11	11
11	1	2	2	10
11	2	9	2	10
11	2	2	9	10
11	10	10	10	8

P5	P4	P3	P2	P1	Nda	Na	Nd	N0	C6	C5	C4	C3	C2	C1
0	1	2	3	4	4	4	6	7	8	9	10	11	12	13
1	2	3	4	5	5	5	7	8	9	10	11	12	13	14
2	3	4	5	6	6	6	8	9	10	11	12	13	14	15
3	4	5	6	7	7	7	9	10	11	12	13	14	15	16
4	5	6	7	8	8	8	10	11	12	13	14	15	16	17
5	6	7	8	9	9	9	11	12	13	14	15	16	17	18
6	7	8	9	10	10	10	12	13	14	15	16	17	18	19

W
0
1
2
3

Q2	Q1	Qa	Qd	Q0
11	10	10	10	8

P5	P4	P3	P2	P1	Nda	Na	Nd	N0	C6	C5	C4	C3	C2	C1
1	2	3	4	5	5	5	7	8	9	10	11	12	13	14
2	3	4	5	6	6	6	8	9	10	11	12	13	14	15
3	4	5	6	7	7	7	9	10	11	12	13	14	15	16
4	5	6	7	8	8	8	10	11	12	13	14	15	16	17
5	6	7	8	9	9	9	11	12	13	14	15	16	17	18
6	7	8	9	10	10	10	12	13	14	15	16	17	18	19

W
1
2
3

Q2	Q1	Qa	Qd	Q0
11	1	2	2	10
11	2	9	2	10
11	2	2	9	10

P5	P4	P3	P2	P1	Nda	Na	Nd	N0	C6	C5	C4	C3	C2	C1
2	3	4	4	5	6	9	9	11	12	14	15	16	16	17
3	4	5	6	7	7	8	8	11	12	14	15	16	17	18
4	5	6	7	7	7	7	7	10	11	12	14	15	16	17
5	7	7	7	7	7	7	9	10	11	12	13	14	15	16
6	7	7	7	7	7	7	9	10	11	11	12	13	14	15
9	8	7	7	7	7	9	10	10	10	11	11	13	14	15
9	8	7	7	7	7	7	10	10	10	10	11	11	13	14
11	11	10	10	9	9	10	10	10	10	10	11	11	12	12
11	11	10	10	10	10	10	10	10	9	9	10	10	12	12
12	12	11	11	11	11	10	10	9	9	10	10	10	12	12
14	14	12	12	12	11	11	11	10	10	10	10	10	11	12
15	15	14	14	13	12	11	11	11	10	10	10	10	11	11
16	16	15	14	14	13	13	13	12	12	12	12	11	10	10
17	17	16	15	15	14	14	14	12	12	12	12	11	10	10

W
3
4
5
6
7
8
8
8
10
12
14
15
16
17

Q2	Q1	Qa	Qd	Q0
12	11	11	11	11
11	1	2	2	10
11	2	9	2	10
11	2	2	9	10
11	10	10	10	8

P5	P4	P3	P2	P1	Nda	Na	Nd	N0	C6	C5	C4	C3	C2	C1
0	1	2	3	4	4	4	6	7	8	9	10	11	12	13
1	2	3	4	5	5	5	7	8	9	10	11	12	13	14
2	3	4	5	6	6	6	8	9	10	11	12	13	14	15
3	4	5	6	7	7	7	9	10	11	12	13	14	15	16
4	5	6	7	8	8	8	10	11	12	13	14	15	16	17
5	6	7	8	9	9	9	11	12	13	14	15	16	17	18
6	7	8	9	10	10	10	12	13	14	15	16	17	18	19

W
0
1
2
3

arises from an imbalance of solute-solute, solute-water, and water-water interactions, which are dependent on the water model used and the solute interaction potentials included in the force field.

This may arise from the general method by which these force fields are often parameterised. When interaction energies derived using small molecules are extrapolated to larger molecules, there is a risk of multiplying systematic errors in the parameterisation, which could account for the imbalance in nonbonded solute-solute, solute-water, and water-water interactions that may manifest as unrealistic aggregation properties. Indeed, the larger the solute, the more small errors in the interaction potentials may accumulate, and the worse the problem may become. These interactions are influenced by the water model being used, as well as the solute-solute interaction potentials as defined in the force field. Another possible source of the errors in AT protein force fields is the focus on maintaining protein native structures during parameterisation, which has been suggested could lead to overestimation of protein-protein interactions[214]. Additionally, the choice of water model has been observed to influence the interaction propensity, suggesting poorly parameterised water models might be partially the cause of such erroneous behaviour[209, 211].

The problem of over-aggregation has received particular attention when it comes to the Martini force field, earning it a reputation of being ‘too sticky’. This has been perhaps most conspicuously discussed in relation to protein-protein interactions in acqueous solution[215, 216], but has also been observed with both free and lipid-linked glycans[198, 214, 217]. One of the solutions that has been devised to manage this issue, and has been applied to both proteins[215, 216] and glycans[214], is to apply a scaling factor to the nonbonded interactions in order to achieve a better match to experimental and/or AT properties. This does not address the underlying problems with the force fields, and is rather a bespoke hack that may not be transferable between different systems being studied. Far better would be to improve the force field itself, and this was one of the focal points in development of the Martini 3 forcefield[Paulo de Souza, personal communication].

Parameterisation methods that include experimentally addressable solution properties such as osmotic pressure[218–220], Kirkwood-Bugg integrals[221, 222], or osmotic coefficient[211, 223] could in theory help to alleviate such problems. All of these methods can be used to evaluate the second virial coefficient in the

osmotic pressure describing the deviation from ideal behaviour in a solution with solute[224]. The value of B_{22} can be estimated experimentally through directly measuring the osmotic pressure[225], self-interaction chromatography[226], and diffraction experiments[227]. If $B_{22} > 0$, it indicates a net attraction, whilst if $B_{22} < 0$ it is a net repulsion between the solute molecules. The magnitude reflects propensity to aggregate. Additionally, there is a strong theoretical basis for getting corresponding parameters from MD simulations, therefore rendering the use of such data useful for refining MD forcefields[216, 224, 228–231].

It should be noted that the extent to which the aforementioned is a problem is not necessarily clear as there are not always suitable experimental data against which the observations can be compared. For instance, is a lack of disocciation between molecules observed during a simulation the consequence of overestimated interaction potentials or due to the simulation time being insufficient for dissociation to be observed? Simulation times are, in general, considerably shorter than experiments and so some of the discrepancy may arise from the substantial difference in observed timescales.

2.5 Running a simulation

In addition to the force field and integration algorithm, a simulation requires several other factors to be considered in order to ensure that it can be run without crashing, and that it models as accurately as possible the system under investigation. The choices of these additional parameters can have a considerable effect on the behaviour of the simulation. Many are chosen based on recommendations for the particular forcefield being used, whilst others might be chosen due to particulars of the work itself or availability within the simulation software.

Simulation time

One factor that is of obvious importance is the simulation time, or more specifically, the number of steps that the simulation will be run for, the simulation time being the product of the number of steps and the timestep, both of which the user must decide upon when commencing a simulation. Even on the largest supercomputers available today, the accessible simulation time and therefore sampling are limited. In a short simulation, most systems are very unlikely to explore extensively their configurational phase space. This limitation can be got around to some extent using methods that enhance sampling, allowing a more efficient exploration of phase space, such as replica exchange methods,

metadynamics, and umbrella sampling[232–239].

However, in general, the question of time becomes one of the major ones when running MD. How long does a simulation need to be run for to observe the phenomena of interest? How many repeats are required to evaluate it with any reasonable reliability? Is it better to run several short simulations or few long ones, given the same computational resources? Such questions ultimately depend upon the question being asked, and the available computational resources.

Long-range interactions

In general, it is rather efficient to calculate the bonded interactions as they involve particles connected by one to three covalent bonds. The nonbonded interactions in a system, however, present a greater challenge as they arise between every pair of particles present, and therefore scale with $O(N^2)$. Practically speaking, it is not feasible to calculate the pairwise interactions for the entire system. We therefore need some way to neglect some of the intermolecular interactions to improve the efficiency of the simulation, albeit without sacrificing the believability of the resulting trajectory.

Looking at equations 2.27 and 2.28 shows that, although both depend on r, their dependence differs considerably. The van der Waals interaction, for instance, is considered a short-range interaction as it decays faster than r^{-d} (d being the dimensionality of the system). The electrostatic potential, which decays as r^{-1}, is a long-range interaction.

For short-range interactions, such as the van der Waals potential, a cutoff scheme can be used, beyond which the potential is truncated. Essentially, any interactions between particles within a defined cutoff distance (typically of the order 8-12 Å) are calculated, whilst those occurring between more distant pairs are neglected from the calculation. Given that the interaction energy rapidly decays to zero as the distance between the particles increases, this approximation is acceptable. However, in practice, the method employed tends to be more complicated so as to avoid a sudden change in the van der Waals forces that would result from a sudden discontinuity in the derivative of the potential at the cutoff distance. It is also possible to add a correction term to account for the neglected van der Waals potential energy beyond the cutoff distance.

The situation is more complex with Coulombic interactions as the convergence to zero is much slower, reducing with the inverse of the distance rather than

the inverse sixth power. Therefore, an alternative method is needed as simple truncation would not be suitable. Most of the solutions developed aim to carry out an explicit evaluation of electrostatic forces within a cut-off, and then a more approximate evaluation over longer ranges. The most common of these solutions is the Particle Mesh Ewald (PME) method[240]. In PME, a neutralising charge distribution is introduced for every point charge in the system. The resulting electrostatic potential, which decays much faster than r^{-1}, can then be calculated using a cutoff scheme, and the electrostatic potential of the neutralising charge distribution removed from the final result. As the potential itself varies slowly, this part of the calculation can be performed in the reciprocal space *via* Fourier transform, and the cutoff scheme then used once again.

A popular alternative to PME the reaction field method[241], in which electrostatic interactions are computed explicitly for particle i only for those other particles j within a sphere of defined size, whilst beyond this the medium is treated as having a uniform dielectric constant. In general, PME is the preferred method, being more accurate, but for many applications the two approaches give similar results and the reaction field methods has advantages in terms of computational efficiency[65].

Periodic boundary conditions

It is usual for simulations to be carried out using periodic boundary conditions (PBC). Effectively, the simulation cell is repeated indefinitely in all dimensions, and any particle that leaves the cell from one side will reenter from the opposing side (Figure 2.5). This eliminates the surface effects that arise from a finite system size, as there is no surface on an infinite lattice, making what is in effect a small volume behave more like a small part of a much larger one. However, a significant limitation that arises from the use of PBCs is the inhibition of long-wavelength fluctuations[65]. This has been noted, for instance, in the case of lipid bilayer simulations in which long wavelength undulations are suppressed if the simulated patch is too small. This is not an apparent problem in many cases, as the effect on bilayer properties is relatively small [242] so as to be negligible, but could cause complications if one wishes to investigate large scale motions of a bilayer. It has also been noted that PBCs are associated with a possible reduction in magnitude of ionic solvation energy[243]. In general, however, it is believed that PBCs have little impact on the equilibrium thermodynamics of a system, and they are standard practice in MD simulations.

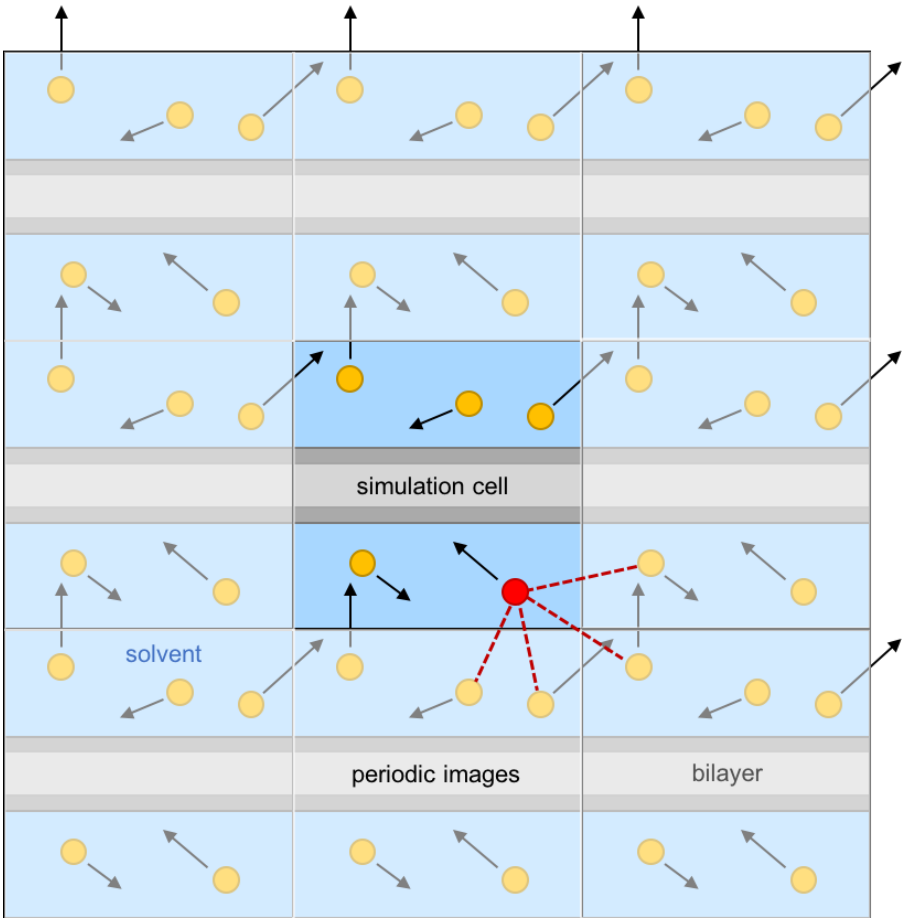


Figure 2.5: Two dimensional illustration of periodic boundary conditions, as applied in this example to a simple system consisting of a single lipid bilayer (grey) in a box of water (blue) that contains a number of solute molecules (orange/red). By effectively replicating the simulation cell infinitely in x, y, and z, it is possible to remove boundary effects that otherwise arise due to the finite size of the modelled system. According to the minimum image convention, the particle shown in red only interacts with the nearest neighbours, as indicated with dashed lines.

The interactions in the system are evaluated by applying the minimum image convention, which means that a particle i interacts only with the closest periodic image of any other particle j, thereby ensuring there are no duplicated calculations.

There are a number of different types of simulation cell that can be used. In many cases, a simple cubic cell suffices. However, for systems with an overall spherical topology, such as viruses, it can be desirable to use a rhombic dodecahedron or truncated octahedron, as in these cases the amount of bulk solvent needed to be simulated is significantly less than would be the case were a cubic cell used.

Energy drift

One of the problems that is often encountered in MD simulations is some degree of energy drift, arising from several potential sources. This is not always in practice a significant problem, but it does have the potential to greatly influence the resulting simulation, and should in general be sought to be minimised from the outset.

In some cases, this drift can arise from instabilities in the system itself, reflecting underlying problems with the model or configuration of particles. These errors can arise due to insufficient minimisation prior to running a simulation, or simply due to problems within the models being used. In any case, they can cause the simulation to blow up and thereby fail, and the underlying issues must then be addressed before the simulation can be successfully run.

The use of a finite timestep is another source of energy drift in MD. This means that the true Hamiltonian for the system is not reproduced, but rather a shadow Hamiltonian is, whose similarity to the true one is dependent on the timestep used[244]. This arises because the discrete timesteps will result in nonphysical sampling of those motions whose frequencies are similar to those of the velocity updates, thereby imposing upper limits on the timesteps that can be used. The simulation will not be stable if the timestep is greater than period of the fastest fundamental modes of the system's motion. In biological situtations, this is often the movement of hydrogen atoms. Use of a constraint algorithm such as LINCS[69] or SHAKE[70] to restrict these motions will allow the use of a higher timestep, but only by about twofold. AT simulations therefore often require timesteps no higher than 1-2 fs.

Another source of energy drift is the evaluation of the potential energy of the system. In many cases this is related to simplifications in the force field that improve simulation efficiency at the expense of accuracy. Use of simple cutoffs for van der Waals interactions without sufficient smoothing, for example, or using truncation schemes for the electrostatic interactions, can cause such errors.

Temperature and pressure coupling

MD simulations are effectively conducted in the microcanonical (NVE) ensemble. Experiments and biological processes, however, are generally carried out under more-or-less fixed temperature, thereby being consistent more with the canonical (NVT) ensemble. In order to relate simulation to experiment using

statistical mechanics, it is therefore desirable to carry out simulations at constant temperature i.e. to run the simulation so as to obtain the NVT ensemble. A fixed temperature also helps to minimise the energy drifts described in the preceding section.

It may also be desirable to maintain a constant pressure, and therefore to simulate in the NPT ensemble, in which the number of particles is fixed, but the volume and total energy can fluctuate.

The methods available for temperature and pressure coupling can be grouped into several categories based on their essential formulation. The stochastic methods, such as the Langevin and Andersen thermostats, work by constraining a system variable to a preset distribution function. The latter, for instance, assigns a random particle a new velocity from a Boltzmann distribution at the desired temperature T upon a collision event, proceeding at constant energy between these events[245]. This was the first thermostat developed for MD simulations, and it is generally not advised for applications other than those looking at time-independent properties.

Strong coupling methods constrain a variable (T,P) to an exact decided value, whereas weak coupling ones constrain it towards that desired value. The Berendsen thermostat[246] is a highly efficient example of the latter.

The temperature of the simulation system is related to the time average of the kinetic energy:

$$\langle E_K \rangle = \frac{1}{2} \sum_{i=1}^{N} m_i v_i^2 = \frac{3}{2} N k_B T \tag{2.30}$$

As the temperature T depends directly on the particle velocities v_i and masses m_i, by rescaling the velocities, the temperature can be controlled.

The Berendsen thermostat weakly couples the system to an external heat bath fixed at temperature T_{bath}. The scaling is softened with a time constant τ. At each timestep, the Berendsen thermostat scales all velocities in the system by a factor λ. If at time t the temperature is $T(t)$ and the velocities are multiplied by λ, then the temperature change can be calculated using:

$$\Delta T = \frac{1}{2} \sum_{i=1}^{N} \frac{2}{3} \frac{m_i (\lambda v_i)^2}{N k_B} - \frac{1}{2} \sum_{i=1}^{N} \frac{2}{3} \frac{m_i v_i^2}{N k_B} \tag{2.31}$$

$$\Delta T = (\lambda^2 - 1)T(t) \tag{2.32}$$

At each timestep, the velocities are scaled so that the rate of change of temperature is proportional to the difference in temperature between the heat bath and the system:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{bath} - T(t)) \tag{2.33}$$

τ is an adjustable parameter governing the strength of the coupling between the bath and the system. The change in temperature between timesteps is:

$$\Delta T = \frac{\delta t}{\tau}(T_{bath} - T(t)) \tag{2.34}$$

And the scaling factor is given by:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}(\frac{T_{bath}}{T(t)} - 1) \tag{2.35}$$

If τ is high, the coupling is weak, whilst if it is too low then the scheme becomes a direct rescaling one and the temperature fluctuations will be unrealistically low. In theory, if $\tau > 100\Delta t$, the system will have natural temperature fluctuations around the average.

The Berendsen thermostat is a highly efficient means to relax the system to the desired temperature, and for a large number of particles it allows for reasonable approximation of many quantities of the system[247]. However, it has been suggested that owing to its tendency towards energy drift and inability to generate a canonical ensemble, it not be used for any simulations now that alternatives are available[248]. Due to its efficiency in achieving the desired temperature, the Berendsen thermostat is often used to equilibrate a system to the desired temperature, and then a more accurate thermostat employed for production simulations.

An alternative, widely employed thermostat, is the Nosé -Hoover thermostat[249]. The approach here is to introduce a fictitious dynamic variable, ζ, which effectively acts as a friction acting to slow or accelerate particles until the temperature of the system is equalled to the desired value. This method therefore falls into

a fourth category of approach: those that extend the degrees of freedom in the system to include non-Newtonian temperature or pressure terms.

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = \mathbf{F}_i - \zeta m_i \mathbf{v}_i \tag{2.36}$$

$$\frac{d\zeta(t)}{dt} = \frac{1}{Q}\left[\sum_i^N m_i \frac{\mathbf{v}_i^2}{2} - \frac{3N+1}{2} k_B T\right] \tag{2.37}$$

where Q governs the strength of coupling of the friction $\zeta(t)$: larger Q gives weaker coupling. The temperature is therefore not fixed when using this thermostat, but rather tends towards the desired value.

Pressure can be maintained at or near to a desired value using method similar to those employed for temperature, except that rather than alter particle momenta, we can scale the dimensionality of the system, altering the positions r and the simulation box size. For example, the Berendsen barostat couples the pressure of the simulation system to a pressure bath[64, 246]:

$$\lambda = 1 - \kappa \frac{\delta t}{\tau_P}(P - P_{bath}) \tag{2.38}$$

$$\mathbf{r}_i = \lambda^{\frac{1}{3}} \mathbf{r}_i \tag{2.39}$$

Here, $\mathbf{r}_i$ represents the rescaled coordinates, and the parameters τ_P (time constant) and κ (isothermal compressibility) govern the strength of coupling to the pressure bath. The isothermal compressibility can be related to the magnitude of volume fluctuations[64]:

$$\kappa = \frac{1}{V}\left(\frac{\delta V}{\delta P}\right)_T = \frac{1}{k_B T} \frac{< V^2 > - < V >^2}{< V^2 >} \tag{2.40}$$

There are several options for how the system dimensions can be rescaled to maintain the pressure. In isotropic pressure coupling, all dimensions are scaled simultaneously. In semiisotropic pressure coupling, which is in particular of use in bilayer simulations, the x and y (in plane with the bilayer) dimensions are rescaled together, and independently from the z (normal) dimension. This allows for proper equilibration of the lateral pressure profile of the bilayer, and

enables self-assembly of lipid bilayers, as subsequently described. It is also possible to employ anisotropic pressure coupling, in which all three dimensions are independently rescaled.

Similar to the improvements of the Nosé -Hoover thermostat over the Berendsen thermostat in terms of maintaining the correct ensemble, the Parinello-Rahman barostat[250] extends the equations of motion to incorporate a term that allows for pressure coupling. As for the thermostats, this renders the approach to the desired pressure less efficient, but is considered more reliable for simulations conducted at equilibrium, and more appropriate for extracting from these simulations thermodynamic properties.

2.6 Setting up a simulation

Overview

As with any experiment, there are several steps that need to be taken in order to carry out an MD simulation once the desired force field and input parameters have been decided upon. Many of these can be carried out from within the software that is to be used for the simulation itself, whilst others require or can be done with external tools or servers. A number of software packages are available that implement MD, including CHARMM[251] and AMBER[252, 253] (not to be confused with the force field families that bear the same names), GROMACS[67, 68], NAMD[254], OpenMM[255], LAMMPS[256], ESPResSo[257], and the special purpose ANTON and ANTON2 machines with DESMOND[258, 259]. While differing in some respects, these all incorporate the basic functionality required for MD simulations, and in most cases additional features for system preparation and/or analysis.

The accuracy and reliability of the output of an MD simulation are reliant on how well the simulation system is set up and the parameters chosen for its performance. If the models used are poor, or the simulation protocol inherently flawed, then the results will be of little use, even if the simulation software takes account of no major errors. A critical part of all MD simulation studies is, therefore, to carefully prepare the initial system and choose the appropriate means of carrying the simulation out. In this section, I discuss some of the main steps in this process in a general sense – the particulars will of course vary with system and aim of the study, but many of the considerations discussed here remain relevant.

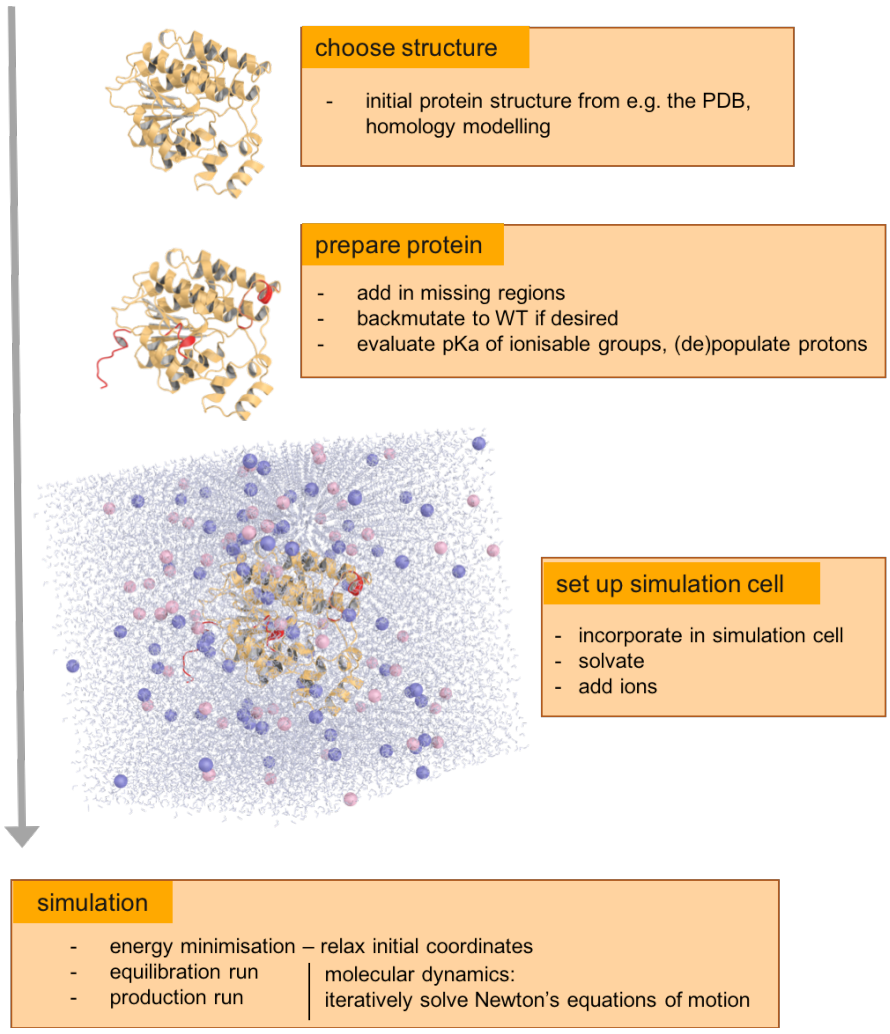


Figure 2.6: Illustration of a generalised MD workflow, starting from a protein structure to its solvation and ultimately the MD simnulation. The protein is shown in cartoon representation, with the inital structure in orange and modelled regions in red. Water is shown as greyish blue sticks and ions as pink and lilac spheres. The protein is a salutaridine reductase from *Papaver somniferum*, PDB code 3O26[260].

Protein structure preparation

As central players in many biological activities, much simulation work has focused on understanding the dynamic behaviours and interactions of specific proteins. Therefore, it is often the case that one desires to include one or more proteins in a simulation. Given the wealth of structural information available for proteins, the starting point these days for most MD simulation studies is an atomistic structure of the construct to be studied, often obtained from the Protein Data Bank (`www.rcsb.org`[58]), if not directly from experiment.

In most cases, the structure chosen in those cases where many are available, will be the most relevant and high quality one available. For instance, out of two structures of the same protein, one will often choose that which is more complete (fewer missing regions) and/or of higher resolution. If studying a particular mutant form or conformational state of a protein, then where possible it is the structure of that variant/state that ought to be selected. Homology modelling can be used to approximate the structure of a protein of interest if no structure for that protein is currently available, provided there is relatively high (>30%) sequence identity between the target and reference proteins[261], and one knows how to validate and refine the model[262].

The protein structure must then be prepared for simulation. In some cases this will involve the modelling in of regions not included in the structural model, for instance loops that may be too disordered to be visualised using X-ray crystallography or that might have been removed to facilitate crystallisation. There may also be some mutations introduced into the protein in order to enable its structural solution, and again these must be returned to the wild-type identities (unless the mutations are relevant to the investigation). Both of these changes can be carried out using tools such as MODELLER[62].

These steps taken to ensure a reasonably complete and high quality structure, the protonation states of all ionisable residues must be decided upon based on their pKa values. These, influenced by hydrogen bonding, solvation, and Coulombic interactions in the immediate environment, might vary over the course of a simulation, something that is difficult to account for. However, the values can be estimated and protonation states chosen based on these at the start of the simulation, and it is hoped that most of the initial pKa values will hold true for the duration of the simulation. In order to estimate the appropriate values, various tools are available, including MEAD[263], PROPKA[264], MCCE[265, 266],

and UHBD[267].

Once the protein has been prepared in this way, a topology must be generated that describes it in the context of the chosen forcefield. This can be done within the software being used for the simulation in some cases, as with the GROMACS pdb2gmx command that produces a topology for the protein, including atom definitions in line with the chosen forcefield[67, 68]. For other cases, including the CG Martini force field, a conversion must be made using additional software in order to produce a protein topology for use within the MD suite. In some cases, the protein, either before or after this step, might be modified with one or more type of posttranslational modification, such as attachment of a lipidic moiety or glycans. For AT structures, the CHARMM-GUI input generator can be used not only to add certain posttranslational modifications, but also to prepare the system as a whole, with compatibility with most simulation softwares, but limited in choice of AT forcefield to CHARMM36 and CHARMM36m[92, 268]. Modifications can, to some extent, be applied in the modular Martini forcefield manually, for instance, by taking the tail beads from a palmitoyl-containing lipid model, and adding them to the protein topology and structure to achieve palmitoylation. Then, the only 'new' bonding parameters that need to be added are those between the appropriate beads of the protein and lipid, which in the case of Martini will generally be some form of simply harmonic potential.

For simulations carried out in aqueous solution, the next step is solvation, as described subsequently, but for those requiring bilayers these must be constructed either around any proteins or prior to their incorporation.

Modelling lipid bilayers

There exist a number of tools for generating lipid bilayers for use in MD simulations. Some of these[190, 269–271] are limited to planar bilayers, whilst others such as CHARMM-GUI[95, 272], Packmol[273], LipidWrapper[274], and BumPy[275] can set up vesicles and other more topologically complex bilayers. These latter must tackle the challenge of interleaflet ratios that is presented by any curved bilayer topology. For instance, an imbalance in lipids per leaflet in a vesicular system could potentially lead to changes in shape and/or artefacts due to nonequilibrium packing.

In the early days of bilayer simulations, it was necessary to manually construct bilayers from known structures of the component lipids, an arduous, inefficient,

and error-prone task[276]. Once preequilibrated bilayer structures for some force fields and lipid compositions became available, these were often desirable alternatives to manual bespoke construction. However, this approach was of course limited as only those bilayers already constructed, validated, and made freely available on the internet were options, and if the bilayer needed to have proteins added this had to be done with some perturbation of the assembled bilayer structure.

Perhaps the most generally opted route now is to build each bilayer bespoke for its particular application using one of the many well-documented and well-validated methods available, either through web servers or using downloadable tools. Some of these start with a bilayer with the chosen composition (e.g. CHARMM-GUI[272], where the bilayers already have lateral densities close to the equilibrium value, the Martini-specific tool *insane*[190]), whilst others (e.g. MemProtMD[277]) employ a self-assembly approach in which a short simulation is used to allow the bilayer to form. A small number of methods allow construction of non-standard bilayer topologies (e.g. PackMol[273], which constructs complex organisations by optimising the packing molecules into user-defined geometric volumes).

Both for ease of repeated use and to facilitate the reliable generation of ever more complex bilayer systems, a number of more-or-less automated tools and protocols have been developed for constructing bilayers for use in MD simulation. In recent years, it has become increasingly regular to find publications of large, multicomponent bilayers, as well as some in complex topological arrangements, in general with other classes of biomolecule or polymer present[152, 278, 279]. Whereas simpler bilayers containing but a few different lipid types might have been possible to construct using manual methods or simple scripts, such complex assemblies require more specialised tools.

One of the most important factors to consider when constructing a bilayer for simulation is the nature and proportions of lipids that need to be present in each leaflet. This is often based on experimental data for the membrane in question, where available, or best estimates based on e.g. biological understanding where not. Sometimes the desired lipid types may not all be available in the force field to be used, and so some degree of simplification in composition may be required. Simplification may also be desirable if the true extent of complexity is unlikely to impact significantly on the focus of the investigation, as more complex bilayers

can take far longer to equilibrate, thereby imposing an additional delay on the work.

Another consideration is the non-lipidic components of the bilayer, generally proteins or peptides, which need to be added to or incorporated within the bilayer in an appropriate manner. In general, one can conceive two means of the latter: either insert a protein into a preexisting bilayer, or assemble the bilayer around the protein (Figure 2.7).

When inserting a protein into a bilayer for simulation, it is necessary to consider the orientation and position of the protein with respect to the bilayer normal. The early methods generally involved assembling a bilayer around a protein[280, 281], using packing algorithms that tended to leave systems that would need long equilibration times to settle, with restraints on the protein structure. There was also a risk of lipid tails being positioned such that they thread through the rings of protein side chains, and these methods offered no way to avoid or deal with this, other than to delete the offending lipid(s) and thereby decrease overall packing density further. Deleting lipids from an existing bilayer to create a void into which the protein can fit, either directly or with aid of a cylindrical repulsion[282] leads to the formation of holes with poor match to the protein shape, leading to long equilibration times. Such methods are compatible with essentially any force field or simulation software, but are not very user-friendly and require lengthy equilibrations, which is not ideal for large systems.

Several of the methods for inserting proteins into membranes are included as a part of, or rely on features of, MD simulation software packages. A number of these are incorporated within or based upon GROMACS tools, including make_ hole[283], which creates a hole in the bilayer based on the protein shape, InflateGRO[284] and InflateGRO2[285], which scale up the bilayer coordinates before inserting the protein and removing the few overlapping lipids so as to minimise perturbations to the bilayer structure once it is scaled back to the normal lipid packing density. g_ membed[286], which is built into the GROMACS mdrun functionality, is included in the most recent GROMACS distribution. It works by embedding a shrunken protein into a hole created by the removal of a small number of lipids, and then gradually scaling the protein up over a short simulation. This procedure should lead to a configuration with small perturbations to the bilayer, and therefore will reduce equilibration times as compared to the previous methods.

GRIFFIN, which is compatible with both NAMD and GROMACS, is based on the same general notions as make_hole[287]. However, it makes an improvement on the methodology in that a combination of van der Waals and Coulombic forces are used alongside the shape-dependent once, to push lipid and solvent molecules away from the protein surface, thereby better optimising the interface between the protein and its environment and reducing equilibration times. It is also able to handle proteins with large internal cavities, such as porins, by allowing for incorporation of restraints to exclude lipids from particular parts of the protein volume.

Any method that involves incorporation into a preexisting lipid bilayer requires some prior knowledge or prediction of where along the bilayer normal the protein would naturally be positioned. This can, of course, be predicted from the positions of certain amino acid types along the predicted membrane-spanning region of the protein, but there is invariably some risk of an error that might take long simulation times to alleviate.

One of the advantages of building bilayer systems in CG is the much faster equilibration than with AT assemblies. This means that there is some leeway in the initial construction as errors such as large voids between lipids, poor packing around a protein, or slight errors in orientation can be quickly relaxed. The Python-based tool for the Martini force field, *insane*[190], constructs a bilayer (with or without a protein) by geometrically placing lipids with linear tails at an even spacing. Any proteins must first be coarse-grained, which can be done with the martinize script also available from the group developing Martini, and the secondary structure is maintained as part of this protein topology. *insane* can construct bilayers with complex compositions, with any of the lipids included in the Martini force field being potentially incorporated, and with a defined ratio of lipids in each bilayer. If AT resolution is required, various tools can be used to backmap an equilibrated CG system to AT once the bilayer has assembled and equilibrated, provided the corresponding AT parameters are available.

Some methods for protein insertion are based on tools available in simulation software, but are not directly built into the software themselves. This in theory means that such approaches are not restricted by choice of MD software or force fields. In one method, described by Javanainen[288], a protein is placed next to a bilayer and then inserted into it by the application of a high lateral pressure. Although no lipids need to be deleted in principle, in some cases this may in

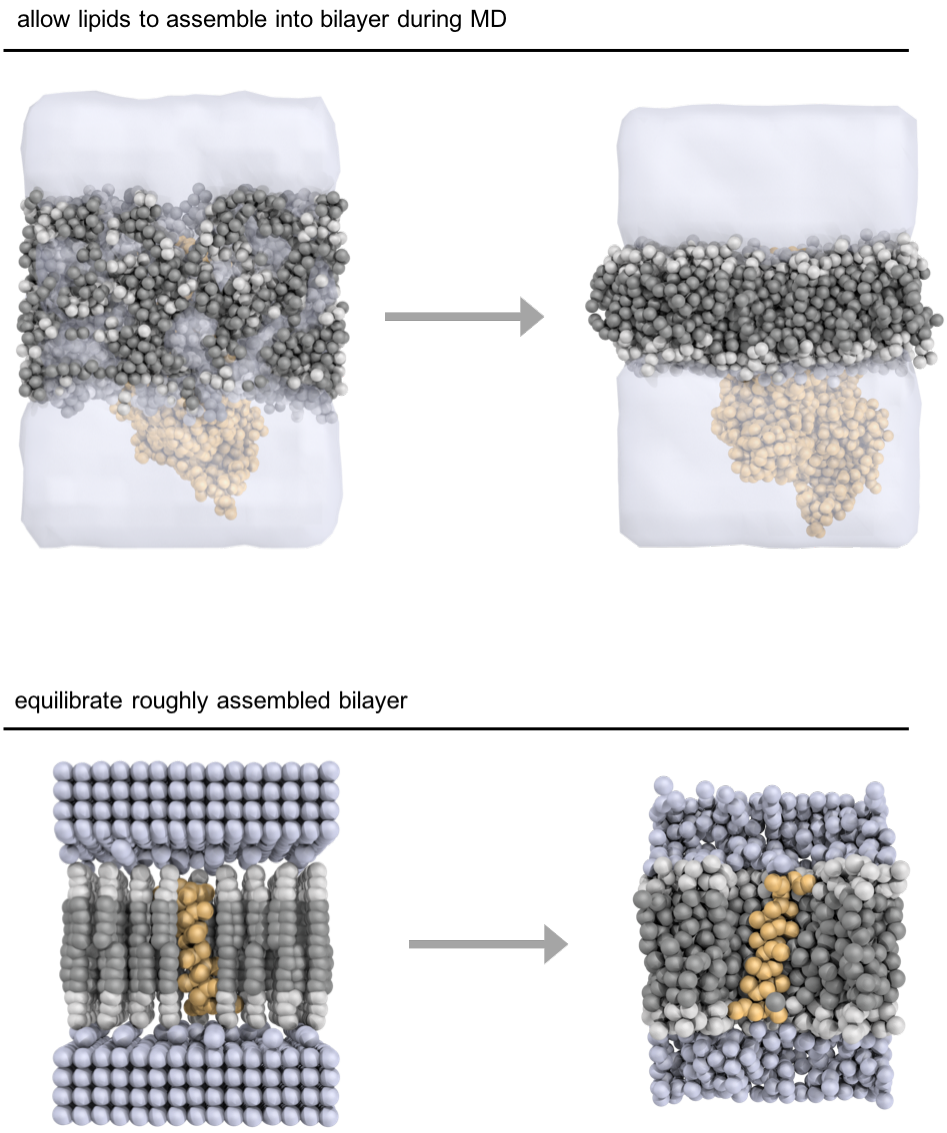


Figure 2.7: Illustration of two approaches to constructing a protein-containing bilayer for MD simulations. The top illustrates the formation of a bilayer around a position-restrained protein, starting from a randomly distributed box of lipids intermixed with water. This specific example is the assemble of a DPPC bilayer about the $\alpha 2\beta$ adrenoceptor (PDB code 6K42, carried out using the MemProtMD[277] pipeline). Below is an example where a protein is inserted into a bilayer arrangement of lipids, which is allowed to equilibrate around it. This example is the KALP peptide being incorporated into a DPPC bilayer using *insane*[190]. Lipid headgroups are shown as light grey spheres, lipid tails as dark grey spheres, proteins as gold spheres, and the solvent in pale blue, as a translucent surface (top) or spheres (bottom), respectively.

fact be necessary as the insertion of the protein can alter the relative volumes of the two leaflets, which can cause deformations in the bilayer. Additionally, the approach is limited to planar bilayers, and requires some manual work to set up.

Another option is to employ tools packaged with free energy methods in most MD distributions according to a protocol previously developed by my lab. In this method, alchembed[289], lipids and other molecules in the environment are pushed away from a protein volume by slowly 'phasing in' the protein from non-existent (no interactions with the rest of the system) to fully present (interacting normally with the system). The surrounding molecules move out of the way as the protein comes into being, and therefore in theory the short simulation should yield a well-packed assembly that will require only short equilibration times. As the method relies only on the presence of soft-core interactions and basic free energy perturbation functionalities, it can be done within most MD software packages and is independent of force fields used. However, it relies on manual positioning of the protein in a preexisting bilayer, and some lipids may need to be deleted from the protein volume. It is applicable to nonplanar bilayer geometries, and is easily extended to multiple proteins being inserted at once.

Solvation

If not already done (as part of a bilayer construction method, for example), it is necessary to place the molecules being simulated within a simulation cell, effectively the container within which the simulation will take place. As noted previously, it is standard practice that PBCs be used to eliminate surface effects. A requirement that arises from the use of PBCs is that each particle in the simulation cell must not be able to 'see' its periodic image. The size of the simulation cell must be chosen to ensure that this is the case i.e. two neighbouring periodic images of the molecule must be separated by at least the cutoff distance for nonbonded interactions. For AT simulations, this generally requires at least 10-15 Åspace to be present between each wall of the simulation cell and the molecules placed within it, although a larger space may be left if it is expected that the molecules may undergo large conformational changes over the course of the simulation, as may be expected for some disordered proteins or peptides.

The space left in the simulation cell around the biomolecular components is, in most cases, then filled with water molecules. There are a great many different models available for water. A discussion of these is out of the scope of the

present work, and of no immediate relevance to the chapters that follow. One important point is, however, that the water model chosen must be appropriate to the force field being used for the rest of the system as each water model was developed with a particular force field in mind. Thus, for instance, the SPC/E water model is to be used with the GROMOS force fields, and the TIP3P and TIP4P models with CHARMM. The various water models available have been reviewed in the literature[106, 290].

In addition to the water, it is necessary to include ions in the simulation system. This is of immediate physiological relevance as ions are present in biological media in various concentrations, but the parameters are not available to allow a realistic ionic solution to be modelled. In general, the best approximation one can do is to model the correct ionic strength using ions such as sodium, chloride, and potassium. The added ions should neutralise the net charge of the biomolecules present, ensuring a total charge of zero across the simulation cell. Technically, the Ewald summation method introduces a neturalising background charge with no apparent error, but it is still considered best practice to maintain the simulation cell neutrality using added ions.

Simulation conditions

At least two preparatory simulations tend to precede any production runs after the system has been set up. The first is an energy minimisation, which is a short process (not an MD simulation) during which any major steric clashes or too-close contacts in the system are eased by applying an algorithm that aims to change the coordinates in such a way that the total potential energy is reduced. This is followed by one or more equilibration simulations, which are carried out using an MD integrator. The purpose of the equilibration is to bring the system as a whole to the desired conditions, such as the correct temperature and pressure. Effectively, the surroundings are allowed to relax around the biomolecule(s) of interest. The precise time and number of equilibrations vary with the set up, with in general more complex or initially unstable configurations requiring longer times to equilibrate.

The equilibration brings the system to the desired conditions according to the statistical ensemble being used. In most cases, one uses the canonical (NVT) or isothermal-isobaric (NPT) ensembles, during which the number of particles, temperature, and either volume or pressure, respectively, are maintained

constant. These are chosen as they are a better representation of experimental conditions than, for example, the microcanonical (NVE) ensemble. It has been shown that, for a system at equilibrium, the various different ensembles yield similar results for calculations of thermodynamic properties[65].

At the start of a simulation, the positions of the particles are given by the input coordinates, and each must be given a velocity to start the system in motion. The initial velocities are set by assuming a Maxwell-Boltzmann distribution for velocities along the three dimensions. In practice, this is done by multiplying Gaussian distributed random numbers by a mean square velocity $\sqrt{2k_BT/m}$ in each dimension, and ensuring a total momentum for the system of zero. The velocities are often initiated during the equilibration(s), and then the simulations simply extended into the actual production stage of the MD from these initial runs. The simulation then proceeds for the chosen number of steps, or until manually halted sooner by the user (or an unfortunate event within the system).

2.7 Progress in membrane simulations

Overview

Biological membranes are, in general, extremely complex. Not only may each contain hundreds of different lipid species and a variety of proteins, but the bilayer itself can adopt various dynamic or stable morphologies, such as buds or curved arrangements, and might interact with elements of the surrounding media on either side, including the cytoskeleton and the extracellular matrix of glycosylated protein fibres. Membranes can also interact with each other, forming temporary or stable junctions, in some cases precipitating signalling or fusion events. Invariably, these membranes and their milieu exist in ever-changing, nonequilibrium conditions.

A major challenge of membrane bioscience, both experimental and computational, is to capture at least a small degree of this vast complexity. In the context of simulations, this means that we must ever be working towards the modelling of more realistic, and thus more complex, systems. It should be noted however that, although there is considerable effort to advance our ability to study more complex systems using simulation, there continue to be many simulation studies whose focus is the fine details of interactions and properties of small numbers of biological molecules in specific contexts. For example, studies of ion channels might focus on the protein itself in great detail, examining the interactions of

specific amino acids with particular lipids, the effects of specified mutations, and the minutiae of the channel pathway. Such studies represent another branch, in effect, of progress in simulation work that is not discussed further here, but is no less important for advancing our understanding of biological systems.

Experimental advances in areas such as single-particle tracking, fluorescence correlation spectroscopy, super-resolved imaging, and mass spectrometry are allowing the properties of membranes and membrane mimetics to be examined in ever greater detail[291–298]. However, it remains difficult to probe the molecular details underlying observable properties, which is precisely the type of information that MD simulations have the potential to reveal. From MD, it is in theory possible to examine the interactions between all components of a system at the atomic or near-atomic level, effectively allowing it to act as a 'computational microscope'[205, 299].

The history of MD simulations being applied to study membrane-associated phenomena dates back to the 1980s, not long after the first report of MD applied to any biomolecular system was published. In these early simulations, the complexity was limited to ordered bilayers of decanoate with harmonic restraints to fix the positions of all headgroups[300]. Rapid progress was made, however, and by the early 1990s there were several reported simulations of phospholipids including explicit water, which addressed questions that are still under scrutiny today, including transport of small molecules through bilayers[301], the effect of cholesterol[302], and interactions of membrane-active peptides[303].

As computational power has increased, allowing larger membrane-based systems to be simulated, it has been possible to extend studies into the realms of bilayer undulations[304], spontaneous bilayer assembly[305], fusion[53], and the formation of bilayer domains[306]. For more detailed discussion on many of these pioneering developments can be found in various publications[307–311].

The 'standard' setup for an MD simulation studying membrane-based phenomena is a single lipid bilayer in a cubic simulation cell, simulated at constant temperature and pressure. Although this general scheme has enjoyed considerable success on many levels, it suffers from an inability to model large bilayer deformations or high curvatures. Many biological membranes either exist in complex topologies, either over long periods or transiently as part of their function e.g. mitochondrial christae, vesicles, the endoplasmic reticulum, eukaryotic chemotaxis. One branch of progress in biomembrane simulations

has therefore been towards modelling bilayers with more complex topologies than what has been the norm. Such work thus far has included vesicles[312–316], cylinders[312, 316, 317], and membrane buckles[275, 318–320]. Another, not independent, branch of progress has been towards the modelling of complex membrane compositions, trying to capture the properties that arise when a bilayer is composed of myriad chemically distinct lipid and protein species[321–323].

Realistic biological membrane models

With the availability of parameters for a diverse array of biological molecules to be used in simulations, both AT and CG, and increasing experimental data on the precise compositions of such environments, it is becoming increasingly possible to simulate ever more realistic cell and viral membranes. One major branch of such advances has been towards simulating large patches of plasma membrane-like bilayer, using data from mass spectrometry to guide the range and proportions of different lipids types[152]. One such model, from Ingolfsson et al., included 63 different lipid types, with 14 distinct headgroups and 11 distinct tails, with an asymmetric distribution between the two bilayer leaflets. The complexity exploited the wide variety of distinct lipids available with the Martini force field, and the CG nature meant that a system with approximately 20 000 lipids could be simulated for 80 µs[152, 323]. Such work allows one to investigate the interactions between all the lipid species, as well as the overall organisation of the bilayer, including leaflet coupling, lipid flip-flop, and formation of both large, stable domains, and smaller, transient arrangements.

In addition to other PM models[321–323], there have been many simulation studies aiming to model the membranes of specific organelles[324], the bacterial outer membrane[200], and the envelopes of viruses[185, 202].

There are several significant challenges presented by efforts to simulate more realistic biological membranes. On the one hand, the greater complexity of composition means that much longer sampling times are required as compared to simple lipid bilayers. Bilayers crowded with proteins exibit slowed lateral diffusion rates[325], and perturbation of bilayer properties may take a long time to reach an equilibrium[326, 327]. In addition, the actual construction of accurate models is difficult. Even if fine quantitative details are known about the nature of component lipids, such information rarely extends to the asymmetry of this composition. Some lipid types are known to only occur on the inner or

outer leaflet, but the relative amounts of those that are found in both is generally not known. The PBCs additionally mean that determining the compositions of each leaflet can be difficult as the area per lipid must be matched between the two. With increasing complexity of a model, the sampling required increases exponentially, and the analysis becomes less straightforward.

Simulating viruses

Given the medical significance and potential biotechnological applications of viruses, it is no surprise that viruses have long been the foci of computational modelling endeavours, and often at the forefront of developments in the field. For example, capsid assembly and disassembly have been extensively examined for many years using MD[328–333], Brownian dynamics[334, 335], and Monte Carlo simulations[336], as well as more mathematical approaches based on elastic sheets[337–339]. The mechanism of membrane fusion has been examined specifically in the context of virology by investigating the membrane association behaviour of fusion peptides, in particular that of influenza A virus[340–346]. It is even possible to extend simulation methods to study entire virions, both nonenveloped[148, 347–350] and enveloped[185, 202, 351], although in this latter case, the genome and capsid are often omitted and only the viral envelope with its proteins included in the simulations.

These ‘whole virion’ simulations represent a major branch of development in the simulation field, heading away from merely modelling realistic models of biomembranes, along a path that might tend towards entire cells. They move out of the realm of the planar bilayer, however refined in terms of composition and dimensions, to explore a more complex topological organisation.

Over the approximately four decades since the first atomic resolution structure of a viral capsid was solved[352], there have been a number of simulations of complete viral capsids of nonenveloped virions. Concurrently, the development of computational virology has largely been facilitated by a number of technological and methodological advances in simulation hardware and tools, including increased computer power, improved simulation algorithms, development of CG and multiscale simulation approaches, and the emergence of novel methodologies for integrating diverse lines of experimental data into theoretical models[353]). However, even with these advances, simulation of complete virions represents a considerable computational challenge. Indeed, a desire to

simulate entire virions has been cited as a key factor driving the implementation of GPUs in computer simulations[354]. More-or-less complete, high resolution structures are now available for a number of viruses. Although these may lack data for part of the genome and for surface glycans, they serve as a starting point for simulations that can be used to 'breath life' into these static structures.

The first ATMD simulation of a complete virus particle was of satellite tobacco mosaic virus (STMV) with and without an RNA core[347]. This study revealed that, consistent with the apparent importance of RNA in capsid assembly, the capsid was much more stable in the presence of RNA. This was true even though the RNA used had an artificial sequence unrelated to the native STMV genome. This work was subsequently extended to include the native genomic sequence of STMV[348] providing the first all-atom model for the complete structure of a virus using its true genomic sequence. Consistent with the results from STMV, experiments have shown that cleavage of the Brome mosaic virus (BMV) CP N-terminal tails, thought to penetrate into the RNA core, is required for capsid disassembly, and after a 5 μs CGMD simulation of the BMV capsid with these tails removed, the capsid did indeed collapse[148].

However, in contrast to the above results, ATMD simulations of southern bean mosaic virus (SBMV) found it to be relatively stable during simulations of 10-20 ns even without a genome present[349]. Furthermore, in a series of long (1-25 μs) CGMD simulations of a range of highly simplified (200:1 mapping) viral capsids[148], the capsids of three plant viruses, STMV, satellite panicum mosaic virus (SPMV) and STNV collapsed during simulations in the absence of a model for the genomic RNA, whereas the capsids of BMV, bacteriophage ϕX174, reovirus, and poliovirus were all stable in the absence of genetic material. These results suggest that the genome is not the only factor influencing capsid stability. For instance, many non-enveloped viruses have been found to possess Ca2+-binding sites at their icosahedral symmetry axes.

Simulations of SBMV reveal changes in capsid nanomechanical properties upon calcium removal which may lead to opening of exit ports for RNA release[349]. A series of ATMD simulations of (at the time) previously unprecedented length (1 μs) were carried out for the plant virus STNV, revealing that without Ca2+, the capsid transformed from a closed to open conformation, the lattice swelling due to increased water permeability[350].

Enveloped viruses present a greater computational challenge. Not only is there

a substantial increase in system size on addition of a bounding lipid bilayer, but the presence of this, envelope glycoproteins, and a matrix, in addition to the genome and capsid, adds several layers of complexity in terms of both initial model construction and subsequent simulations. It therefore becomes necessary to use CG representations and innovative approaches to both construct the virion models and subsequently simulate and analyse them.

The first microsecond-timescale simulation of a viral envelope[202] used the Martini CG forcefield to simulate a model of the influenza A virus envelope, providing insight into the influence of bilayer components and temperature on the stability of this medically important virus. By integrating data from X-ray crystallography, NMR spectroscopy, cryoEM, and lipidomics, the authors assembled a model for the viral envelope with inserted spike glycoproteins – haemmagglutinin (HA) and neuraminidase (NA) – and the M2 proton channel. Although there was no M1 protein matrix included, the presence of the ribonucleoprotein (RNP) core was treated using a shell of position-restrained particles, and the influence of the matrix was modelled by applying position restraints to the centroids of the envelope proteins. Simulations were carried out at 295 K or 323 K, with or without glycolipid replacing two thirds of the sphingomyelin in the envelope, and with or without the proteins restrained (i.e. with or without the M1 matrix considered). From these simulations, it could be seen that the glycolipid covered the virion surface (Figure 8a), and largely masked the M2 proton channel from the external surface, which could have implications for drugs such as adamantanes targeted against M2. In addition to affecting the volume of the virion surface accessible from the outside, the presence of glycolipid (and lower temperature) also appeared to reduce mobility of envelope components. Reduced mobility of envelope components is expected to increase physical robustness of the virion, which could contribute to the known stability of the virus under different environmental conditions[355, 356].

In addition to influencing translational dynamics, the glycolipid appeared to reduce clustering of the NA and HA spike glycoproteins, maintaining their roughly even surface distribution at a spacing that was suggested to be conducive to binding of bivalent IgGs, as well as the broadly neutralising antibody FI6 targeted against the HA stem. Not only are the observed spacings compatible with antibody binding, but they would also facilitate polyvalent interactions between HA and/or NA and host sialic acid residues, as would be required for

strong virus-host interaction in spite of relatively weak affinity between HA and host receptors detected in vitro, and be important in progeny virion release.

The membrane envelope of the dengue virus (DENV) has also been simulated in two independent studies, again using the Martini CG force-field. In the first study a combination of CG modelling and simulation was used to 'add back' the lipid bilayer to the cryoEM structure of the DENV envelope proteins[44]. This allowed estimation of the diffusive properties of lipids within the outer envelope of the DENV virion. The dense crowding of protein TM domains and the enclosure of the outer leaflet of the lipid bilayer within a protein shell resulted in lipid diffusive properties similar to those in the 'raft-like' influenza A virion membrane.

In a subsequent multiscale modelling study of the DENV envelope[351], a CG model of the DENV-2 envelope was constructed, and validated through comparison with cryoEM maps. This model was used to explore the roles of protein domains and protein-lipid interactions in the envelope organisation and stability. There was good agreement between the cryoEM maps and the MD simulations. The MD simulations reproduced bilayer curvature and distortions around TM helices of the proteins observed in both leaflets from cryoEM. By changing the envelope lipid composition, it was further demonstrated that the proteins make numerous contacts with anionic lipids (phosphatidylserine) in the system, and that these interactions help to stabilise the viral particles. On this basis it was suggested that changes in the electrostatic environment during the viral lifecycle would alter protein-lipid interactions. For instance, histidine residues found to interact with lipid headgroups may become protonated in the endosome and phosphatidylserine headgroup carboxylates may also change ionisation state as the local pH falls, in both cases potentially leading to rearrangements in the envelope complex to promote fusion.

In the simulations discussed so far, the model resolution is invariant throughout the simulation, and so multiscale simulations require 'back-mapping' from CG to AT resolution models[43]. However, it is in principle possible to combine multiple resolutions within the same simulation. This was done, for example, in the previously mentioned mixed-resolution study examining SeMV, TrV, and ZIKV virus-like particles using the SIRAH force field[170, 357].

While most efforts towards simulating entire virions have used CG models to allow such large systems to be simulated for reasonably long times, more recently

a model of the influenza A H1N1 2009 virus envelope has been modelled using an all-atom approach[358], albeit without glycosylation of the surface glycoproteins. The explicitly solvated envelope model was over 100 nm in diameter and included approximately 160 million particles. Given the large size of the system, it could only be simulated for about 121 ns. However, by combining the ATMD simulations with Markov state model theory, it was possible to deduce features of long timescale (microseconds) glycoprotein dynamics in the crowded environment of the envelope. Such an analysis revealed features of the NA conformational dynamics that could prove useful for future drug development efforts as an AT model of the virus envelope provides considerable opportunity for understanding the topology that any drugs against a virus might encounter and have to tackle in order to act at the external virion stage. Of relevance to events taking place during the writing of this thesis, there has been considerable effort recently towards the modelling of the envelope of the SARS-COV-2 coronavirus envelope from the same group, taking an approach similar to that for the IAV envelope.

2.8 Molecular dynamics simulations presented in this thesis

I finish this chapter by providing general details of the MD simulations presented in this thesis. The specifics for each individual case are given in Appendix C. Here, I only give a basic summary.

Simulations

All the simulations were performed using GROMACS[311].

For all CGMD simulations, I used the Martini force field[160, 187, 189, 359], either version 2.2 or the beta of version 3, as indicated, whilst for ATMD I used the CHARMM36 force field[360], having used CHARMM-GUI[92, 251, 268] to construct the AT system. I constructed CG systems either manually, or using *insane*[190]. Steric clashes in the systems were alleviated using Alchembed[289].

In all cases, after the system was constructed (including solvating waters and ions), it was subject to energy minimisation, equilibration, and then the production simulation. The equilibrations used generally the same set of simulation parameters as the production simulation, but a lower timestep. In some cases, it was necessary to carry out a series of equilibrations at gradually increasing timesteps, as indicated in the text and Appendix C.

Protein modelling

In order to model CG protein structures, I first obtained AT models for the relevant protein from the PDB`www.rcsb.org`[58]. Where these had missing regions such as short unresolved loops or no transmembrane helix, I modelled these using MODELLER[62] and the protein sequence taken from UniProt[361]. The completed AT protein was then converted to Martini 3 representation using Martinize[160, 359], and this then used as the input for glycosylation.

Data analysis and coding

Analyses and the glycosylation method used in-house proprietary code written in Python, employing primarily the modules NumPy[362], MDAnalysis[363], PyPlot[364], and Seaborn[365]. Visualisations and renderings were performed with Pymol[366].

Chapter 3

METHODOLOGICAL DEVELOPMENT OF DOUBLE BILAYER SIMULATIONS

The development of this method was helped considerably by advice from Bart Bruininks at the University of Groningen.

3.1 Introduction

All membrane-membrane associations are necessarily underpinned by the presence of two or more proximal lipid bilayers. It is therefore critical that, as part of this work, I establish means by which such systems can be modelled. As discussed in Chapter 1, this is not a novel concept in the MD simulation community, being used e.g. as a methodological tool in computational electrophysiology[47, 367], in simplified forms to examine the physical bases of membrane-based phenomena[51–57, 368], and in a smaller number of instances to look at more biologically-realistic scenarios[48, 50]. It is this latter type of double bilayer system that I am interested in simulating and, at present, there is no standard approach to set up such assemblies.

In simple conception, the construction of a double bilayer system is quite straightforward, so that the absence of simulations of this kind in the literature may be attributed in part to their relatively large size. To prepare such a system, one need do little more than position the bilayers of interest together, in parallel, within a single simulation cell at a normal separation suitable for the biological system (based on e.g. electron microscopy data). Subsequently, the simulation can be run just as would be done on either of the bilayers alone. The construction of lipid bilayers for simulations has become a fairly routine procedure[43, 205, 277], and there exist a number of more-or-less automated methods to do this[190, 369]. Then, putting two prepared bilayers in a simulation cell together can be achieved manually or with a relatively simple bit of coding. Thus, the construction of a double bilayer system is by no means a singular task.

Preparing the system in this way, however, gives rise to a notable complication: the creation of two physically separate water compartments (Figure 3.1). In an MD simulation, periodic boundary conditions are used to help counteract surface

effects of particles at the simulation cell boundaries, and they in effect mean that the system is replicated infinitely in all directions. Particles leaving through one wall of the cell, reenter through the opposing wall. When a lipid bilayer is simulated in an aqueous environment, the bilayer is generally constructed so as to span the x,y dimensions (that is, to touch the cell walls), and as a consequence if one looks at just one simulation cell, it appears as if there were two water compartments, above and below the bilayer. However, due to the periodic boundaries, these apparently separate aqueous milieux are one.

Consider stacking two bilayers, one atop the other, within the simulation cell. Looking at a single simulation cell, one can see a compartment of water in the space between the bilayers that is discontinuous with that above and below them. These two solvent compartments cannot exchange solvent or ion particles unless there is a sufficiently large and suitably hydrophilic conduit across one or both of the bilayers - that is, there are now two truly separate water compartments. In the absence of free solvent exchange, the dynamics of the two bilayers will be coupled - one cannot undergo fluctuations without the other undergoing such motions also (Figure 3.1).

If the two bilayers were bridged, for example by a pair of *trans*-interacting proteins, the situation might become even more problematic. For example, proteins such as SNAREs that help to promote fusion of their host membranes, are believed to help pull the opposing bilayers towards each other. If there is an incompressible layer of water between the two bilayers, this becomes impossible without disruption of the lipidic environment. As for the case with no proteins present, the likely result is artefactual behaviour of the system, or at least a behaviour that cannot be claimed with confidence to be representative of the true biological system.

In this chapter I discuss one potential way in which this problem may be alleviated. Effectively, the approach is to introduce a ‘pressure release valve’ into one of the bilayers. This conduit allows free movement of water and ions so that there is, in effect, one continuous solvent compartment as with a single bilayer simulation. To achieve this requires an undeniable perturbation of the system but, as I show, this can be minimised so that the benefits of avoiding spurious bilayer coupling might outweigh the costs of the added irregularity. However, the utility of the method will, necessarily, depend on the system under investigation and the nature of the questions being posed.

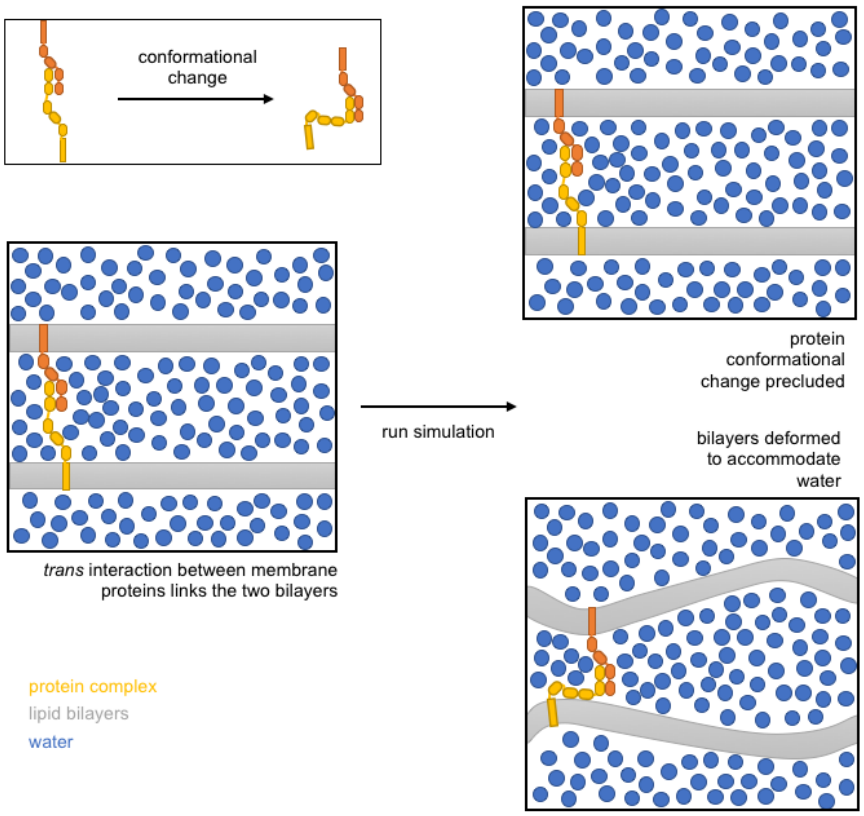


Figure 3.1: Illustration of the problem created by having two bilayers in a single simulation cell. The issue derives from the fact that the bilayers are (unless in possession of one or more open pores) impermeable to water and ions, and due to the periodic boundary conditions are essentially infinite in the xy (lateral) dimensions. This creates two incompressible water compartments between the two bilayers, which has the potential to deform the bilayers and/or impede their dynamics.

3.2 Method overview

There are a number of ways one could create a pathway for solvent passage across a bilayer. In the conceptually simplest case, merely incorporating freely permeable proteinaceous pores such as occur in certain biological membranes ought to provide a thoroughfare for the waters, and potentially ions if a non-selective channel is selected. However, it is desirable that the pores do not interfere with the proteins and lipids whose behaviour is particularly the focus of the investigation, and a protein-based pore may well interact with these components, thereby influencing the results. To avoid such inopportune associations, it is likely that one would need to either repel the proteins/lipids of interest from the pore protein and/or have a sufficiently large bilayer so as to render such interactions improb-

able. What is more, it is known that integral membrane proteins can impact on the properties of the bilayer in which they are inserted[82, 370–373]. It is a risk that the pore proteins might do just that, for instance clustering specific lipids in their vicinity or locally altering the bilayer thickness.

The solution I propose herein does not entirely eliminate the aforementioned complications, but it goes some way towards lessening them. My method, illustrated in Figure 3.2, effectively seeks to create a headgroup-lined (therefore hydrophilic) toroidal lipidic pore at a specified point in one of the bilayers, through which waters and ions can freely flow at a rate sufficient to permit alleviation of pressure in either water compartment as the bilayers move.

To do this, a flat-bottomed repulsive potential (Equation 3.1) is applied to a specified particle in the tails of all bilayer lipids. This repels the acted-on particles from a cylindrical volume centred about the specified point, but is only felt within a certain, user-defined radius of that point, attenuating rapidly at this distance. Beyond this, the particles no longer experience the repulsion and so ought to behave normally. Approaching the defined point, however, the particles are pushed away by the repulsive potential acting on them, thereby creating region devoid of lipids – the pore. The use of a flat-bottomed potential means that no force acts on the particles in the flat-bottomed region of the potential (i.e. the parts of the bilayer away from the pore), but a harmonic force acts upon them to move them to this region (i.e. away from the pore) if they stray outside it.

Because the potential acts on tail beads only, the headgroups are able to face into the pore centre so as to form the hydrophlic lining. As the repulsive potential is cylindrical about the specified point in the bilayer plane, it will act on the lipids regardless of how the bilayer moves in the orthogonal direction. The radius of the pore can be tuned based on the radius over which the repulsion acts, and may need to be varied for different lipids, depending on the size and nature of their headgroups. Additionally, it is straightforward to tune the strength of the repulsion, and the pore location within the bilayer, if either of these things is deemed necessary. Here, I use throughout the same pore strength and location for simplicity, as initial tests showed that varying these had little/no impact on the features being examined. The repulsion force constant was 1000 kJ/mol/nm^2, and a position 5 nm from the edge of the box in the x and y directions.

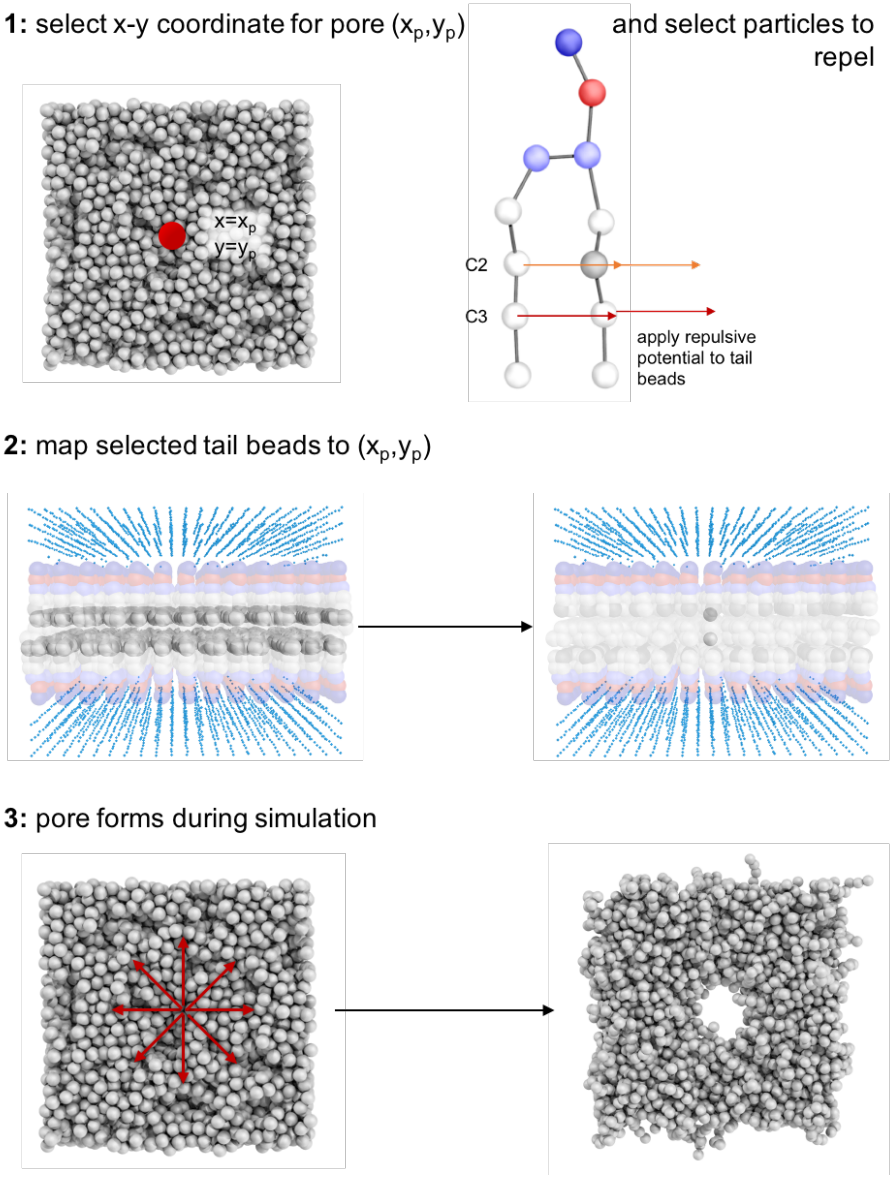


Figure 3.2: Illustration of the method used to create pores in lipid bilayers. First, the desired x-y coordinate of the pore is chosen and one or more particles in the bilayer lipids are selected e.g. the C3A and C3B beads in a Martini lipid tail (top). A reference structure file is then created in which all of these beads in the system are mapped to the desired site of the pore (middle). During the simulation, these beads are then repelled from that specified point in the bilayer plane using a flat-bottomed repulsive potential, which acts only on those lipids in close proximity to the defined site of the pore (bottom). This has the effect of inducing the lipids to form a toroidal, headgroup-lined pore at the defined position.

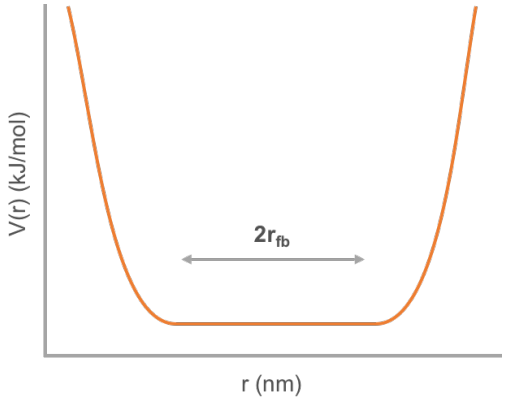


Figure 3.3: Illustration of the functional form of the flat-bottomed potential used to form pores in the bilayers. The regions where the potential increases would correspond to the pore, where the lipids feel the repulsive potential, whilst the rest of the bilayer would be the flat part of the potential.

Equations governing the pore formation:

$$V_{fb}(r_i) = \frac{1}{2} k_{fb}[d_g(r_i; \mathbf{R}_i) - r_{fb}]^2 H[d_g(r_i; \mathbf{R}_i) - r_{fb}] \tag{3.1a}$$

$$d_g(r_i; \mathbf{R}_i) = \sqrt{(x_i - X_i)^2 + (y_i - Y_i)^2} \tag{3.1b}$$

where $\mathbf{R}_i$ is the fixed reference position (the site of pore formation), r_{fb} is the distance from the centre of the flat region of the potential(Figure 3.3), k_{fb} the force constant, H the Heaviside step function[374] that effectively facilitates a switch-like behaviour of the potential (lipids are repelled or not repelled), and $d_g(r_i; \mathbf{R}_i)$ the distance of the particle i from the reference position. The subscript g refers to the geometry, which in my case is cylindrical, with the cylindrical axis parallel to the bilayer normal, and the distance calculated as shown in Equation 3.1b. By definition, the force of the potential acts towards the centre of the cylinder, and therefore the potential is given as negative when defining a pore in the topology files of the lipids to be repelled.

Although conceptually relatively simple, there are several considerations that need to be taken into account in order to employ this method effectively. These are discussed individually in the following sections.

repelled beads	repulsion radius (nm)	minimum radius (nm)	waters in pore?
D2A/C2B	2.5	1.08	yes
C3A/C3B	2.5	0.83	yes
D2A/C2B	2.0	0.62	yes
C3A/C3B	2.0	0.49	yes
C1A/C1B	1.5	0.62	yes
D2A/C2B	1.5	0.22	yes
C3A/C3B	1.5	0.10	yes
C4A/C4B	1.5	0.48	no
D2A/C2B	1.0	0.08	no
C3A/C3B	1.0	0.06	no

Table 3.1: Properties of pores formed when the repulsion was applied to different lipid tail particles or a different repulsion radius. The minimum pore radius was estimated using HOLE[375], and the presence or absence of waters in a pore (as an indicator of whether or not the pores were open) was assessed by visual inspection. In each case, the repulsion force constant was 1000 kJ mol^{-1}.

3.3 Basic pore properties

In order to investigate the properties and functionality of the pores formed using this method, I employed a 'toy system' composed of 338 palmitoyl oleoyl phosphatidylcholine (POPC) lipids and 2748 CG water particles, simulated using the beta version of the Martini 3 forcefield[160, 187, 189, 359] as, at the time of this work, the final version was not available.

I simulated this first with no pore, and then with a pore present, constructed using various different combinations of parameters, specifically altering the lipid bead on which, and the radius over which, the repulsion acts. The parameters tested are summarised in Table 3.1. Note that the radius over which the repulsive potential acts is not the same as the radius of the pore formed, and it is likely that the radius of the pore will differ to some degree depending on the precise identity of the bilayer lipids themselves.

The resulting conduits are visualised in Figure 3.4. The repulsive potential was applied initially during the equilibration phase, which was run with a short timestep of 1 fs, and then maintained throughout the production runs, which employed the usual timestep for Martini simulations of 2 fs. For pores of the sizes tested here, removing the repulsive potential even after the pore has formed leads to prompt reformation of the intact bilayer (within a few ns). This may not hold true for all systems. For instance, it has been suggested that pores may not reform for ATMD simulations [Laura John, personal communication].

An interesting observation from these data is that the shape of the water conduit

varies depending on the specific lipid tail bead to which the repulsive potential is applied, with more hourglass-shaped pores being formed with beads at position 3 in the tail, and more cylindrical ones with beads at position 2. For my purposes, the shape of the pore is not of particular interest, but given the diversity of ion channel internal organisations, the ability to alter pore geometry could be something to consider exploring in relation to other fields of investigation.

As can be seen from Table 3.1 and Figure 3.4, when the pore had a minimum radius of about 0.10 nm, it was filled with water, and I was able to visualise the passage of water through the pore in both directions, as well as movement of lipids between the two leaflets of the bilayer. As one ideally wants to minimally perturb the bilayers in a simulation, the smallest water-conductive pore possible might be considered ideal. However, if one is simulating a double bilayer system then consideration also needs to be given to the need for allowing a sufficient rate of water passage to allow the bilayers to fluctuate relative to each other. Balance must therefore be struck between minimising bilayer perturbation and ensuring sufficiently rapid passage of water from one side of the bilayer to the other.

In subsequent sections, unless otherwise indicated, I simulate a pore constructed using repulsion of radius 1.5 nm, force constant 1000 kJ mol^{-1} nm^{-2}, and acting on the C3A/C3B beads of the lipid tails.

3.4 Bilayer asymmetry

The formation of the toroidal pore renders continuous the headgroup regions of the two leaflets, allowing free movement of lipids from one to the other, thereby creating what is, in effect, a simple ‘scramblase’, and expediting loss of any asymmetry that initially existed across the bilayer. This is observed clearly if I run simulations akin to those presented in the previous section, but using an asymmetric bilayer rather than pure POPC. I simulated a bilayer composed of 164 POPC molecules in one leaflet and 164 palmitoyl oleoyl phosphatidylserine (POPS) molecules in the other – in the normal situation, that is, with no pore present, there would be no exchange of lipids between the two leaflets. Indeed, when I simulated my asymmetric bilayer in the absence of a pore, the two leaflets maintained their single-lipid compositions (Figure 3.5, top). However, upon adding a pore, the asymmetry was quickly lost (Figure 3.5, middle), a result of lipids from one leaflet diffusing into the other, and vice versa.

This loss of asymmetry may not be a problem in practice if the bilayer asymmetry

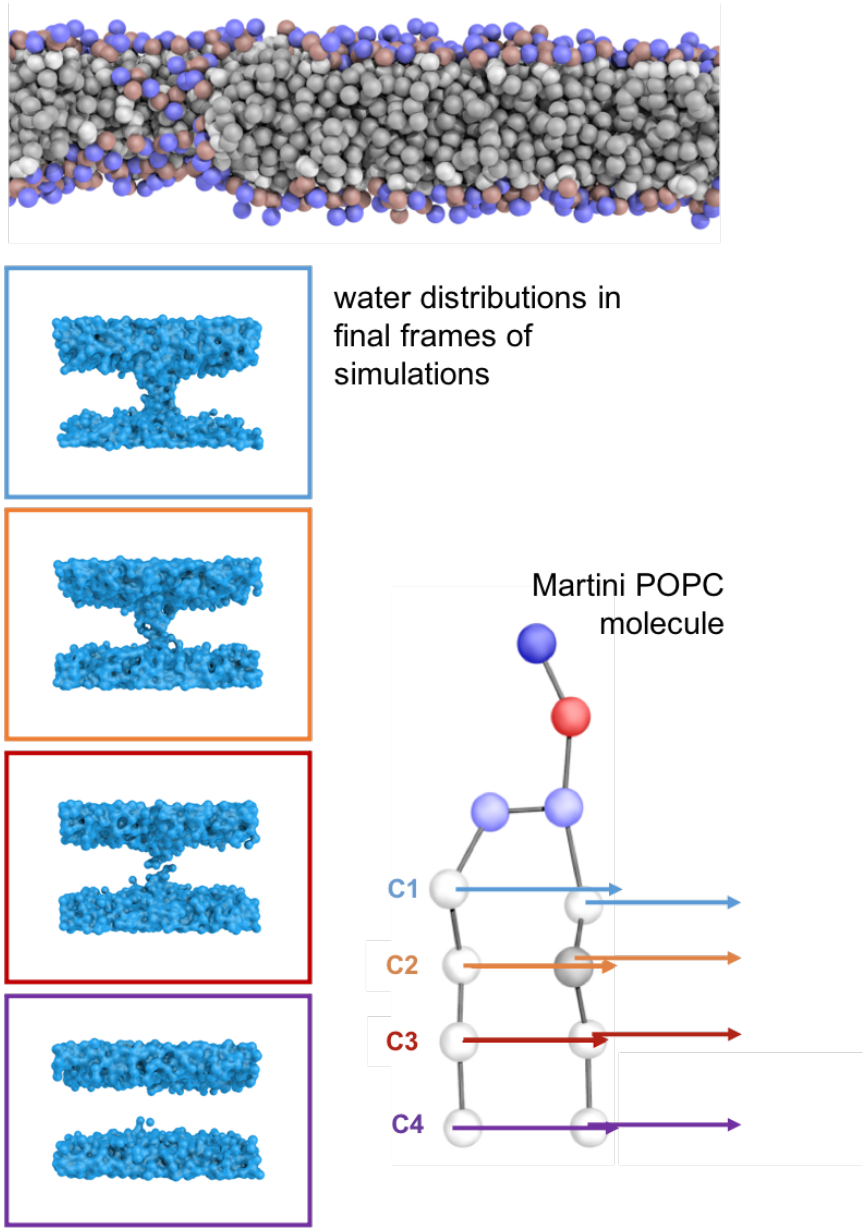


Figure 3.4: Final frames from 1 μs simulations of POPC bilayers in which pores were formed by applying repulsive potentials of force constant 1000 kJ mol^{-1} nm^{-2} and radius 1.5 nm to different tail beads of the POPC lipid. The tail beads are indicated in the right-hand image, and colour-coded with the corresponding snapshot on the right-hand-side. In each snapshot, the system is viewed along the plane of the bilayer and shows only the waters, highlighting the presence or absence of a complete water conduit. The image at the top shows a cross-section of the bilayer itself, with lipid tail beads in dark grey, the lipid headgroup choline, phosphate, and glycerol groups in blue, pink, and light grey, respectively.

is not important to address the question being posed. For instance if one simply wishes to know how the outer leaflets of two plasma membranes interact, or how the ectodomains of a protein in one bilayer bind surface-exposed lipids in another. In such circumstances, it may well be perfectly acceptable to simulate two symmetric bilayers, or even just one symmetric bilayer (possessing the pore) with the other being asymmetric, even though biologically neither membrane would be symmetric in its composition.

However, there may be instances where the bilayer asymmetry is critical. I therefore investigated a means to prevent its loss, although was unable to come to a solution without considerable drawbacks. my approach was to apply a repulsion acting orthogonally away from the midplane of the bilayer, and acting on all lipid headgroups (Figure 3.5, bottom). This hinders lipids crossing the bilayer midplane to enter the other leaflet. However, it also prevents z-motions of the bilayer in question and is expected to slightly destabilise the central region of the pore through which the repulsive plane passes.

I therefore note that although this approach is successful in maintaining bilayer asymmetry, it requires additional controls be applied to the system, and any such intervention will carry an unavoidable risk of perturbing system behaviour. It is therefore recommended that this method only be used where asymmetry is thought to be critical for the behaviours under investigation, and the bilayer containing the pore is not expected to undergo a net large perpendicular translation. In other cases, I recommend simply employing symmetric bilayers whose compositions match those of the interacting leaflets of the two bilayers e.g. both leaflets have the composition of the outer leaflet of the plasma membrane for two interacting cell surfaces.

3.5 Protein-containing bilayers

Biological membranes generally contain a somewhat notable complement of proteins, rather than being composed solely of the lipids that form the bilayer itself. In many cases, these proteins are of interest to researchers, and therefore it is generally desirable to include them in one or both bilayers of a double bilayer simulation. This creates an additional complication for the presented poration method as how the proteinaceous bilayer components might interact with the pore must now be considered.

When simulating the previously used pure POPC bilayer with a small pore and a

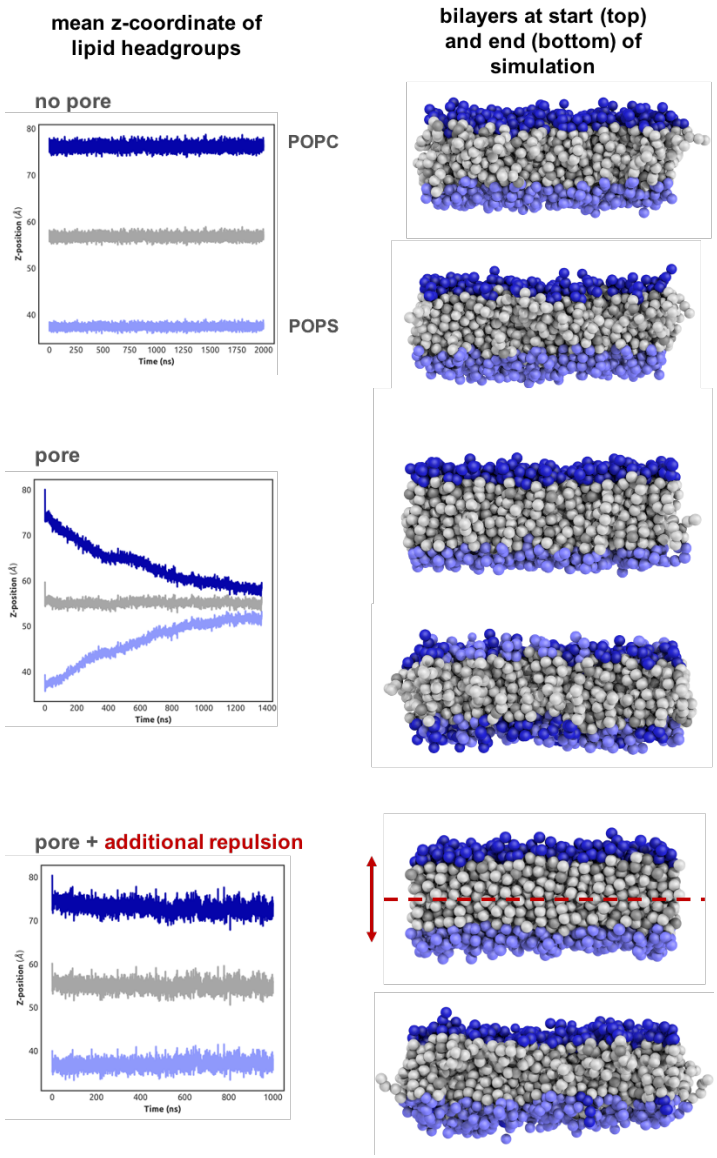


Figure 3.5: Loss of asymmetry *via* a lipidic toroidal pore. When there is no pore present in this small, asymmetryic POPC/POPS bilayer, the two leaflets remained separated (top). Upon addition of a pore, the leaflets quickly mix and asymmetry is lost (middle). Applying a repulsive potential to the lipid headgroups to discourage such mixing allows asymmetry to be maintained (bottom).

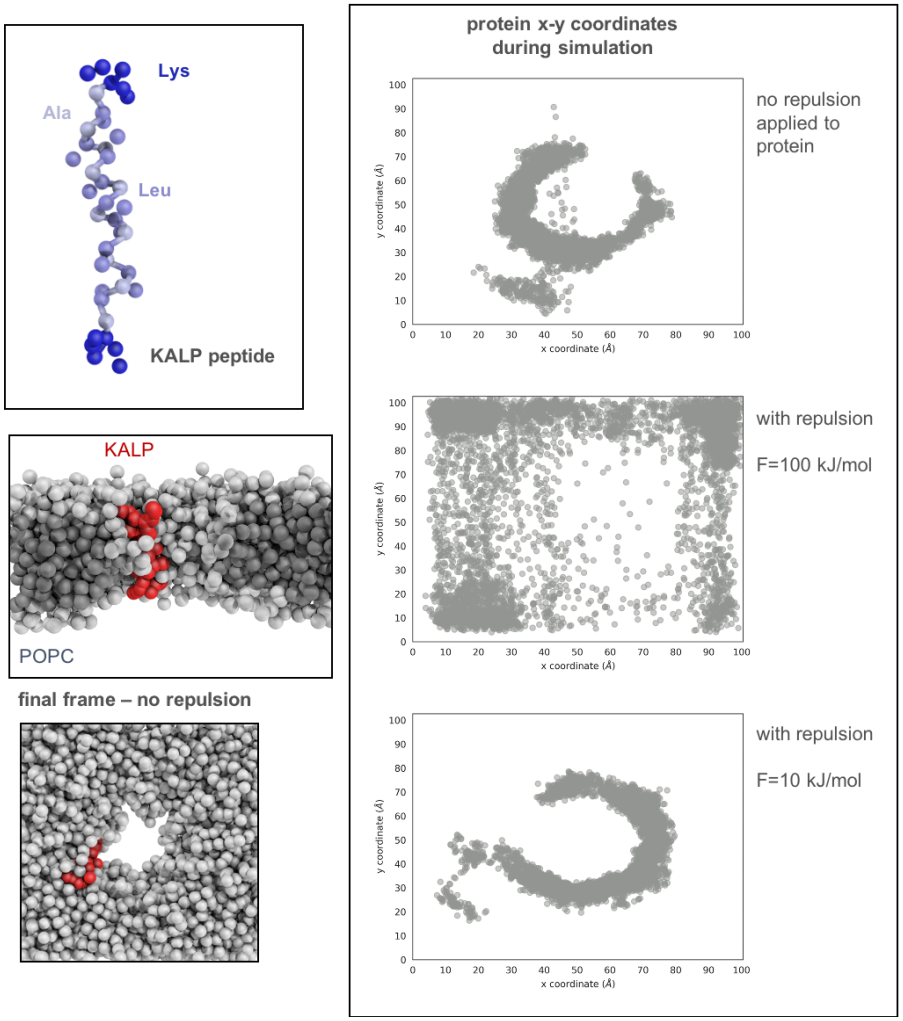


Figure 3.6: Simulations of a KALP peptide in a POPC bilayer containing a single pore. The top image shows the peptide associating with the pore edge when no repulsive potential is applied to it. This can be seen more clearly in the middle plot, showing the x-y (in-plane) protein coordinates over the course of the simulation. Each point represents the protein centre of geometry at a single frame in the simulation, and the points are translucent such that a dense colour indicates repeated occupancy of the same position. Applying a repulsive potential allowed the peptide to explore the simulation cell more freely, as can be seen in the lower plot, showing the protein to lo longer circumnavigate the pore.

single KALP peptide in it, the peptide spent the vast majority of the simulation associated with the edge of the pore (Figure 3.6). This is clearly undesirable, but intriguingly reminiscent of the proposed behaviour of some pore-forming peptides. In order to oppose such spurious associations, I applied a repulsive potential to the protein beads in exactly the same way as is done for the lipid tails, but of both larger radius and weaker force. This allowed the protein to explore the box as if the pore was not there, indeed avoiding the pore almost entirely. Logically, the bilayers must therefore be of sufficient dimensions for the protein to explore the area to a reasonable extent beyond the range of influence of the pore, therefore possibly introducing a requirement for more atoms than may be initially considered necessary for the system.

3.6 Justification

It is evident, from the previous discussion, that introducing a pore into a bilayer comes with several complications. It therefore needs to be demonstrated that there is, in fact, more than just a hypothetical benefit to be derived from simulating these pores. Previously, considerable fluctuations have been observed in lipid bilayers with in-plane dimensions on the order of 100 nm^2 [Anna Duncan, personal communication]. Therefore, I set up bilayers of a similar size in pairs, with and without pores of varying sizes. The purpose was to examine their global dynamics with and without a pore.

The individual bilayers were symmetric, with a composition modelled on a 'standard' eukaryotic plasma membrane extracellular leaflet (POPC, POPE, the ganglioside GM3, and cholesterol, in the ratio 11:2:2:5). As some of the component lipids were not available in Martini 3 beta at the point of writing, I used Martini version 2.2 for this system. The simulations were run for 5 μs, following a 1 μs equilibration and pore formation period. The total simulation cell was approximately 123 nm x 123 nm at the end of the equilibration, and consisted of 4 970 048 particles.

From visual inspection alone, there is little clear influence of the pores on the relative motions of the bilayers. Any such impact will require more extensive tests and quantitative analyses of the motions of each individual bilayer.

3.7 Concluding remarks

The coming together of two or more membranes occurs throughout biological systems, playing critical roles in both physiological and pathological processes.

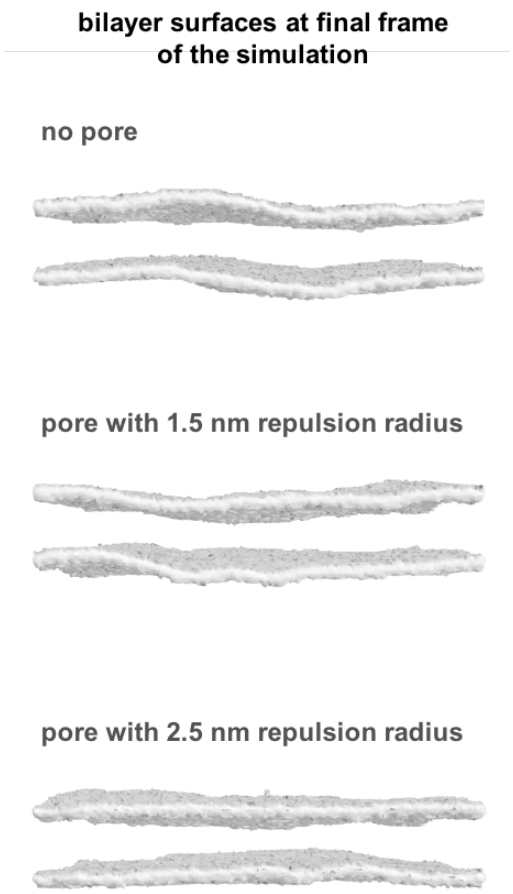


Figure 3.7: Final frames from 5 μs simulations of double bilayers with no pore (top), with a pore formed using a repulsion radius of 1.5 nm (middle), or a pore formed using a repulsion radius of 2.5 nm (bottom). In each case, the surface of each bilayer is shown in grey and the waters are ommitted.

In many cases, the interacting surfaces are expected to be adorned in glycans, creating an interface that may be densely packed with sugar chains.

Despite the biological importance of membrane-membrane associations, my understanding of how they operate is relatively limited. Most information obtained to date comes from relatively low resolution visualisations made using fluorescence or cryo-electron microscopy and tomography techniques[376–378], as well as more detailed structural and biochemical examinations of individual components[60, 61, 379–385]; Whilst these investigations have been highly informative, it is difficult to infer from them properties of the membrane-membrane complex as it occurs in a biological context. In the former case, the limited spa-

tial resolution precludes in-depth understanding of the system at a structural level, whilst the latter requires one to assume that the individual components and membranes will behave in isolation as they would in the membrane-membrane complex. There is no reason to believe that this ought to hold true, and indeed, evidence is emerging to support that it in fact does not[386].

It is challenging to study closely juxtaposed biomembranes at high resolution using available structural and experimental tools. Such a problem is, however, well suited for examination using MD simulations. One can use existing structural and biochemical data about individual membranes and their components to build a model of the juxtaposed system, and then examine an MD simulation of this model in order to ascertain details about the time-dependent behaviour of the membrane junction itself. Provided that the individual components of the simulation system, and the forcefield used to model interactions therein, are reasonably accurate, then in theory the combined assembly ought to behave in a manner that is reflective of the properties of the real system, thereby providing another layer of insight into the operation of these complex biological environments.

Here I have presented a method for introducing a pore into a lipid bilayer under the impetus of being able to simulate bilayer-bilayer juxtapositions without impeding the movement and dynamics of the bilayers. Although the approach is far from ideal in that it precludes use of an asymmetric lipid composition and requires some perturbations of the system, if one is expecting to see some change in bilayer separation and/or conformations of the proteins or lipids bridging the gap between the bilayers, it may be worth the costs to allow such changes to take place.

An important deficiency of the work hitherto presented is a lack of demonstration that the pores are, in fact, necessary. I had hoped to run my large bilayer simulations for longer, and to analyse the simulations to assess whether or not the bilayer motions are coupled in the presence or absence of a pore, and doing so will be important to demonstrate the degree to which my pores are serving their intended function. However, it has not been possible to do this analysis thus far.

It is interesting to note, as an aside, that the presence of a simple lipid toroid is sufficient to act as a scramblase[387], culminating in loss of asymmetry, and that proteinaceous species tend to associate with the wall of the toroid. Both

observations are of considerable interest from the perspective of pore-forming peptides and proteins, and merit further investigation.

Chapter 4

PARAMETERISING GLYCANS FOR MARTINI 3

For reference structures, abbreviations, and symbols for the monosaccharides mentioned in this chapter, please refer to Appendix A, and for reference structures and abbreviations of lipids discussed, please refere to Appendix B.

4.1 Introduction

The occurrence of glycans in biology

Sugars, in one form or another, play a range of important roles in cellular function, among the most well-known of which is their utility as a source of chemical energy. However, sugars participate in a great many additional processes, for instance, modulating protein activities and interactions, serving as structural elements, and participating in signalling processes[388].

Of these various non-metabolic functions, many take place topologically external to the cytosol. Indeed, the vast majority of naturally occurring cells and viruses are covered in a dense and heterogenous array of glycans, both attached to constituents of their outer layers and secreted from the cells into the surrounding matrix. Sugars and chains thereof adorn most surface-exposed or secreted proteins of eukaryotes, as well as the headgroups of certain lipids such as the gangliosides and glycosylinositolphosphoceramides (GIPCs) of animals and plants, respectively[389, 390]. Extensive chains of glycans often with repetitive motifs, make up much of the extracellular milieu of many cells creating dense, mesh-like networks that can greatly alter the outer perimeter of the cell, changing the environment encountered by other vicinal cells or approaching small molecules. Polysaccharides form a fundamental component of the cell walls belonging to plants, fungi, bacteria, and archaea, which have important roles in providing mechanical protection and structural support, in addition to a degree of filtering of solutes between the cell and the environment.

The glycans exposed on the surfaces of cells and viruses can participate in signalling, cell-cell adhesion, and self- versus non-self recognition[391–393]. The sugar coats of pathogenic microbes and viruses might have additional roles in evading elimination by the host immune system, and in adherence to host tissues

and cells. Indeed, this function of surface glycans in adherence is thought to be common across the kingdoms. An essential facet of multicellular organisms and congregations of unicellular ones is the ability of cells to interact with each other and with the surrounding environment. Being external to the cell, many glycans are well-placed to participate in the modulation or mediation of various cell-cell (inter- and intraspecies), cell-matrix, and cell-molecule interactions.

Cytosolic and nuclear protein-linked glycans are also observed commonly in eukaryotes, with roles in signalling processes within the cell. These intracellular glycans appear to bear little evolutionary or biological similarity with those found outside the cell, just as there is no evolutionary connection identified between bacterial glycans and those of eukaryotic origin[388]. Such instances of convergent evolution are testimony to the important role played by glycans in biological systems – an entire synthetic enterprise, especially one that requires considerable costs in terms of resources, is unlikely to have evolved and be maintained independently in distinct lineages if it is not of some considerable benefit to the organisms possessing it.

In order for glycans to participate in some of their biological functions, particularly those related to recognition and signalling, it is necessary that there exist cognate glycan-binding proteins, some of which may derive from the same organism, while others might belong to pathogens, symbionts, or other external agents. Some classes of glycan-binding proteins, specifically the lectins, appear to have shared evolutionary origins, or at least have been found to adopt similar structural folds. Glycan-binding proteins of this type often reveal a high degree of specificity for different glycan structures, but typically have relatively low affinities such that multivalency of both the glycan and glycan-binding protein is often necessary to achieve high avidity. Other classes of glycan-binding protein seem to have emerged through convergent evolution, as exemplified by the sulphated GAG-binding proteins.

This intrinsic importance of glycans is borne out by genetic studies. In eukaryotic cell culture, genetic defects in surface-exposed and/or secreted glycans might reveal few consequences for cell viability and behaviour, but the corresponding mutations in intact multicellular organisms can have major phenotypic consequences. For instance, genetic defects in glycan synthesis, when they may lead to complete elimination of a major glycan class, often prove lethal early in development, whilst those causing abnormalities in later stages of the glycan

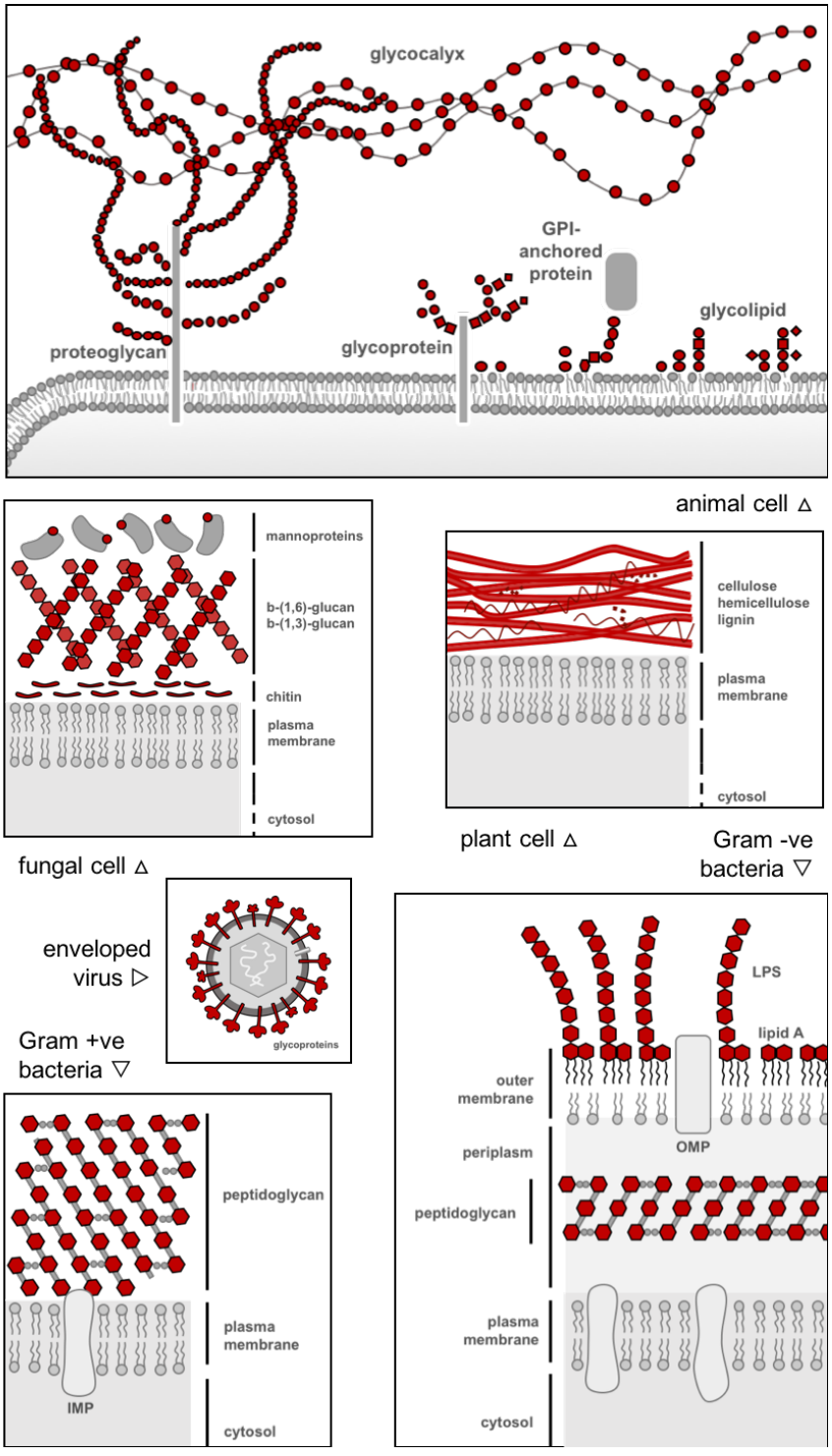


Figure 4.1: Schematic illustration of the occurrence of glycans on the surfaces of Gram positive and Gram negative bacteria, animal, plant, and fungal cells, and viruses. The glycan components of the surfaces are shown in red, whilst the other constituents (lipids, proteins, cytoplasm etc.) are in grey. Note that these representations are illustrative and, for instance, there are variations in the specific compositions, some of the glycan types shown may not always be present whilst others not indicated may be found.

synthetic process might demonstrate tissues-specific alterations glycosylation patterns with both cell-autonomous and distal biological consequences[388].

Despite the evidence in favour of their importance in a general sense, it is often extremely difficult to ascertain decidedly the function of any given glycan. This is in part owing to the aforementioned pleiotropic phenotypes observed on disruption of glycan synthetic processes, but also to the difficulties associated with targeting only one glycan or glycoconjugate for disruption or modification out of the myriad expressed, whether by use of genetic manipulation or enzymatic intervention. Furthermore, many of the purported roles of glycans relate to their impact on the nature of the environment surrounding cells or to the interaction of cells with agents in their surroundings. Outside of the native context, it can be challenging to reproduce any such activities, whilst *in vivo* examinations are generally compounded by the complexities of living systems. The situation is further complicated by the observation that some glycan functions might change depending on the tissue in which the glycan is found and/or the present state of the organism. Microheterogenity, the simultaneous occurrence of variants in length and structure of a given glycan, alongside their potential inherent complexity further compounds attempts to characterise their structures or functions experimentally.

Despite my present lack of clarity regarding the functions of glycans, what is apparent is that they form an essential component of most, if not all, biological systems. To a large degree, this lack of understanding may derive from the difficulties associated with studying glycan function both *in vitro* and *in vivo*, as well as the challenges associated with structural characterisation of glycans that hinder their analysis by traditional structural biology techniques. Computational methods offer one means to help relieve this information deficit, but only if adequate models can be constructed for the molecules in question.

The chemistry of glycans

The fundamental building blocks of all glycans are the monosaccharides, which may be salvaged from the environment or synthesised within a cell. These molecules can exist in chain (open) or cyclic (ring) forms, and it is as this latter state that they form the units of more complex carbohydrates. Cyclisation results in a chiral anomeric centre at a particular carbon (often C1) in the ring, which governs whether the monosaccharide is in an α or β conformation. These

two variants are anomers of each other, differing only in the orientation of substituents to this carbon (top panel in Figure 4.2). Differences in the relative orientations of the other substituents to the ring, where these are invariant in nature, create isomers (right panel in Figure 4.2). The pyranose (six-membered) ring can adopt various different conformations, as illustrated in the central left panel of Figure 4.2. The chair conformations are the most stable, collectively dominating the conformational ensemble at equilibrium under physiological conditions. The specific chair predominantly adopted is dependent on steric effects, as well as potential hydrogen bonding and electrostatic interactions. For example, in general, it is preferred that substituents to the ring are in equatorial positions (compare, for instance, glucose (Glc) and mannose (Man) – the C2 hydroxyl in a $^4C_1$ chair conformation is axial in Man but equatorial in Glc). In the case of Glc, most of the hydroxyls are equatorial in the 4C1 conformation to minimise steric clashes, whilst for idose (Ido) there are similar numbers of axial and equatorial hydroxyls in $^4C_1$ as in $^1C_4$, and therefore both conformations tend to be populated to a comparable extent.

Covalent linkage of two cyclic monosaccharides occurs *via* glycosidic bonds involving the anomeric hydroxyl of one residue and one of the hydroxyls belonging to the other. The lower panel of Figure 4.2 shows the glycosidic bond between two pairs of monosaccharides to form two distinct disaccharides. Bonds between monosaccharides are named according to the constituents and bonds between, and the nature of the linkages e.g. the two disaccharides shown are Glc-α 1,4-Glc (top; α -glucose on the left is in a glycosidic bond to carbon 4 of the glucose on the right) and Glc-β 1,6-Glc (bottom; β -glucose on the left is in a glycosidic bond to carbon 6 of the glucose on the right). Monosaccharide polymerisation into glycans is catalysed by glycosyltransferases, a class of enzymes that shows specificity for the glycan, protein or lipid to which they add the activated sugar moiety.

Often, monosaccharides to be linked to each other are first activated in nucleotide- or lipid-linked forms. Addition of residues into extended chains of glycosidic-linked monosaccharides forms oligosaccharides or, if a large number of repeating oligosaccharide motifs are combined, polysaccharides. In covalently linked pairs and chains of monosaccharides, the end with a free hydroxyl on the anomeric carbon (or at least one not linked to another sugar) is the reducing end, whilst the latter is the non-reducing end. Within extended chains of monosaccharides, the

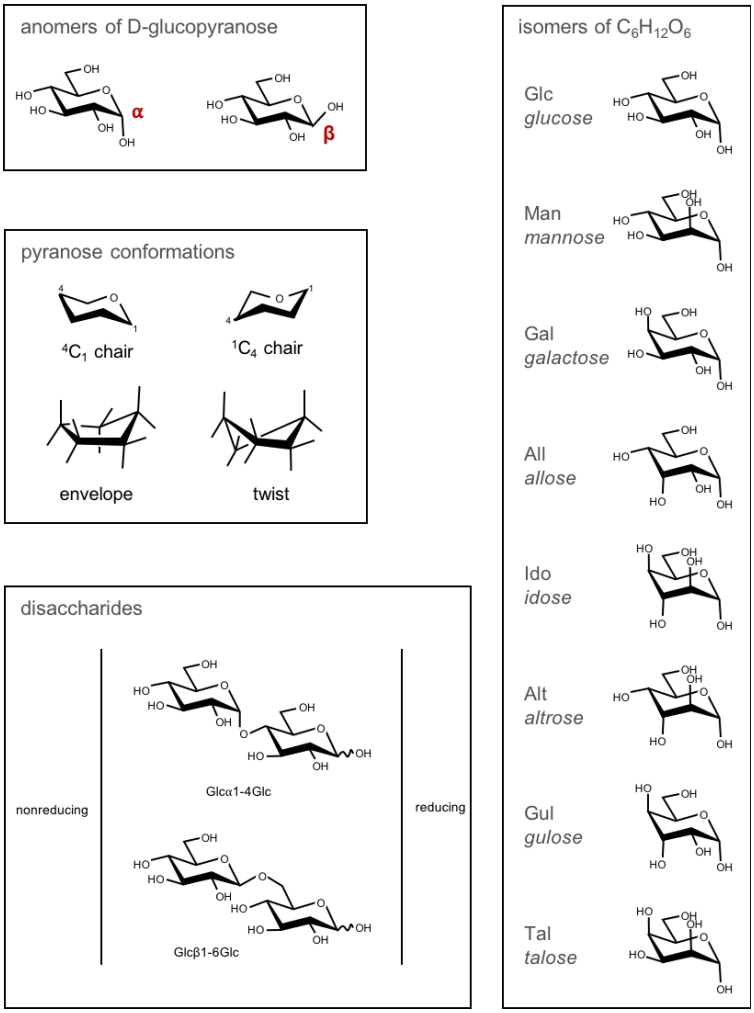


Figure 4.2: Skeletal structures illustrating some of the basic features of carbohydrate chemistry, as discussed in more detail in the text. Top and right: examples of the differences between anomers and isomers, which differ in the orientations, but not identities, of substituents about the ring. Middle left: a comparison of some of the conformations a pyranose ring might adopt in solution. Bottom left: two examples of disaccharides of Glc differing in the nature of the glycosidic bond, and highlighting the reducing and nonreducing ends of the molecule.

two types of linkage (α versus β) confer distinct structural and biological properties on sequences of monosaccharides that are in all other respects identical, as can be seen in the comparison between starch and cellulose (Figure 4.3).

The potential variation in glycan composition far exceeds that possible in nucleic acids or proteins, which may be described as linear polymers of units linked by only one type of covalent bond. Theoretically, three different amino acids or nucleotides could potentially be joined into six distinct trimers. Three different monosaccharides, however, could theoretically form anywhere between 1,056 and 27,468 unique trisaccharides. The complexity of glycans may be further increased by the incorporation of multiple different monosaccharides, of which hundreds have been identified naturally occurring in biological systems, and by covalent modification of the sugar units themselves through addition of functional groups such as sulphate and acetyl.

However, most naturally occurring glycans contain relatively little of the theoretically feasible monosaccharide repertoire, and that only in a limited number of combinations, owing to such factors as the availability and activity of the requisite biosynthetic enzymes. Many enzymes involved in glycan biosynthesis are expressed in a cell-type or tissue-specific manner, with additional temporal and/or spatial regulation superimposed upon this through, for example, compartmentalisation and gene regulation.

Experimentally, it remains a considerable challenge to characterise in much detail the structures of glycans. Their inherent flexibility generally prevents techniques such as X-ray crystallography or NMR from being able to define their structures beyond a few monosaccharide units[394]. Additionally, the frequent heterogeneity of naturally occurring glycans poses a challenge for NMR, which requires very pure samples[395], and it has often been necessary in X-ray crystallography of glycoproteins for any glycans to be either removed entirely or truncated if high resolution structures are to be obtained[396]. Mass spectrometry can be used to determine detailed information regarding the composition and connectivity of small glycans, but more complex and larger structures are less readily analysed due to difficulties associated with characterising individual glycosidic linkages[397, 398].

4.2 Parameterising the Martini monosaccharides

This work was aided in large part by Siewert Jan Marrink and Paulo de Souza, at the University

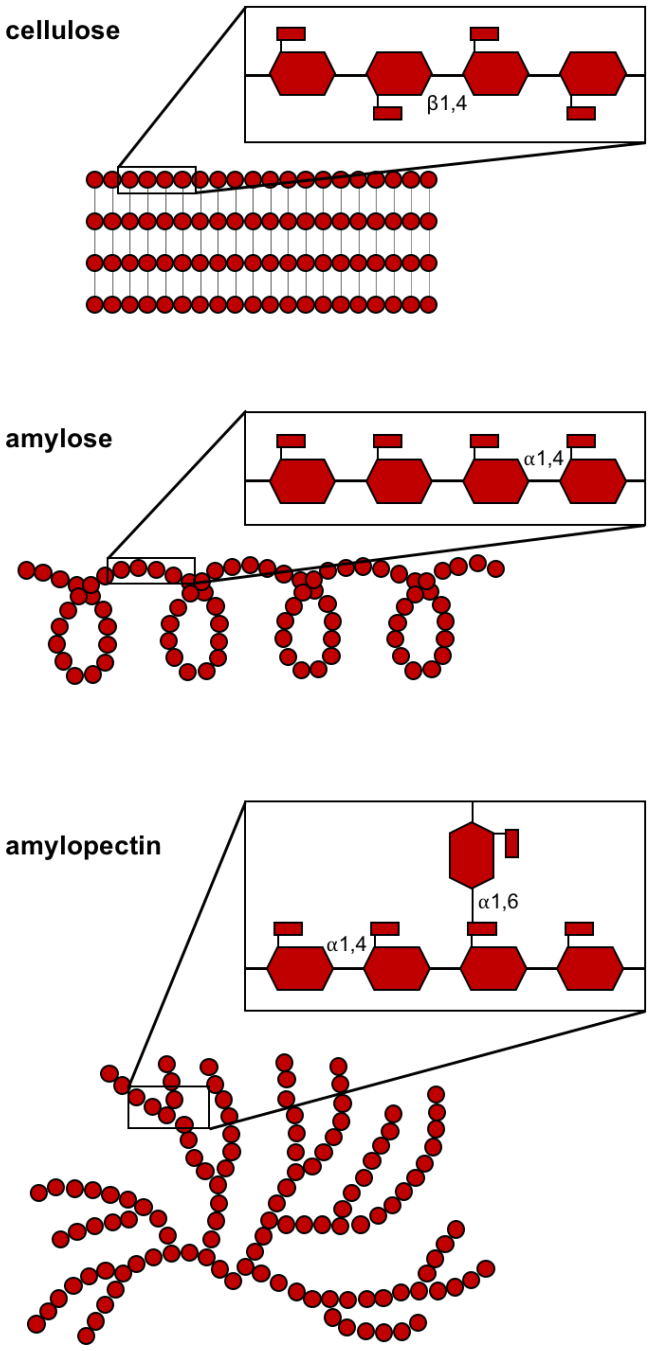


Figure 4.3: Comparison between the structures of three plant polysaccharides of glucose. Both cellulose and amylose are linear polymers, differing only in the anomeric state of their consituents. This difference may lead to the formation of helical structures in amylose, as opposed to long linear chains in cellulose. Amylopectin differs from amylose in the presence of additional branches linked to C6 of some residues. Both components of starch, the combination of these two types of organisation allow the formation of semicrystalline granules that are held together by networks of hydrogen bonds and are insoluble in water. In contrast, the linear chains of cellulose hydrogen bond with each other to form microfibrils with high tensile strength, an important feature for this major structural component of plant cell walls.

of Groningen in the Netherlands, where I spent some months working on the initial stages of developing the models for glycans. The new version of Martini 3 is being developed in their laboratory, and the information given on its particulars here is based on the state of development at the time of writing.

Previous Martini glycan models

There are several forcefields for use in ATMD simulations with compatible sugar parameters[399, 400], including the GLYCAM06 force field developed to model sugars but extended to permit its use with other biomolecules[401], and extensions of OPLS[402], GROMOS[403, 404], and CHARMM[89, 405, 406] for carbohydrates. These have found various applications, including studies the behaviour of sugars in solution and interacting with other biomolecules[407–409], as well as looking at the dynamics of glycoconjugates such as glycoproteins and glycolipids[410, 411].

Parameters have also been developed for a number of monosaccharides and glycan-containing molecules within previous iterations of the Martini forcefield[159, 162, 198, 217]. Several of these were specifically designed for particular glycan species, such as simple sugars[162], gangliosides[159] or N-glycans[412].

Among the problems encountered with these previous models was a propensity towards aggregation that was greater than experiments or ATMD simulations predicted[198, 214, 217]. This could be attributed to the solute-solute (i.e. sugar-sugar) interactions, or features of the water model. The latter has changed in Martini 3 significantly from that in Martini 2, whilst there exist no coherent set of models in either version for the monosaccharides and the assemblies they link to form.

The first extension of the Martini force field to the realm of glycobiology[162] (Figure 4.4) included parameters for the monosaccharides glucose and fructose, and a number of disaccharides selected to encompass a range of glycosidic linkages: sucrose (Glcα1–2fructose (Fru)), trehalose (Glcα1–1Glc), maltose (Glcα1–4Glc), cellobiose (Glcβ1–4Glc), kojibiose (Glcα1–2Glc), sophorose (Glcβ1–2Glc), laminarabiose (Glcβ1–3Glc), nigerose (Glcα1–3Glc). The authors did not include in their development the 1–6 glycosidic linkage, reasoning that its conformational flexibility poses a greater challenge to CG modelling. The flexibility of these linkages is important for the biological roles of glycans containing them. For instance, the occurrence of 1–6 glycosidic bonds in storage

polysaccharides may help facilitate access by hydrolytic enzymes.

The monosaccharide and disaccharide models were evaluated by their extension into the polysaccharides amylose and curdan, and the behaviours of these in CGMD simulations as compared with experimental properties and ATMD simulations. Additionally, the ability of glucose and trehalose to cryo-protect a dipalmitoylphosphatidylcholine (DPPC) bilayer modelled with the same force field was examined, along with how well the density of different concentrations of glucose solutions matched experimental data. In this latter evaluation, the agreement with experiment was better for lower concentrations, but the density was underestimated at higher concentrations, suggesting some errors in the sugar-sugar interaction potentials so that packing does not occur as efficiently as in reality. Both the monosaccharide and the disaccharide, however, were able to prevent the DPPC bilayer from transitioning from the liquid-crystalline to a gel-like phase at 270 K[162].

The most notable limitations of this early model were a direct consequence of the reduction in dimensionality and simplification of bonded interactions: it could not distinguish between different ring conformations, there was no distinction made between anomers other than that achieved through different bonded angle and dihedral interaction potentials, and only one conformation for the glycosidic linkage was represented.

More recently, after the release of the beta version of Martini 3, there was an effort to extend the force field to the modelling of *N*-linked glycans, specifically focusing on those that have been frequently detected in biological systems thus far[412]. ATMD simulations were used to parameterise the oligosaccharides. It was found that the conformations of the glycans could not be reproduced from ATMD without the application of elastic networks, in analogy to the means by which protein conformations are maintained in Martini simulations, accounting for the lack of directional hydrogen bonds.

It was noted that the CG glycan models had a tendency to aggregate in solution, analysed using RDFs and PMFs, as has been observed previously in both CG and AT simulations[189, 214, 413]. The need for tuning parameters was also confirmed in the reproduction of lectin binding free energies. The nonbonded interactions were scaled to match the behaviour of AT models. It was suggested that the loss of stereochemical features and the hydrogen bonding networks of the sugars, as well as the network of surrounding water models, may affect

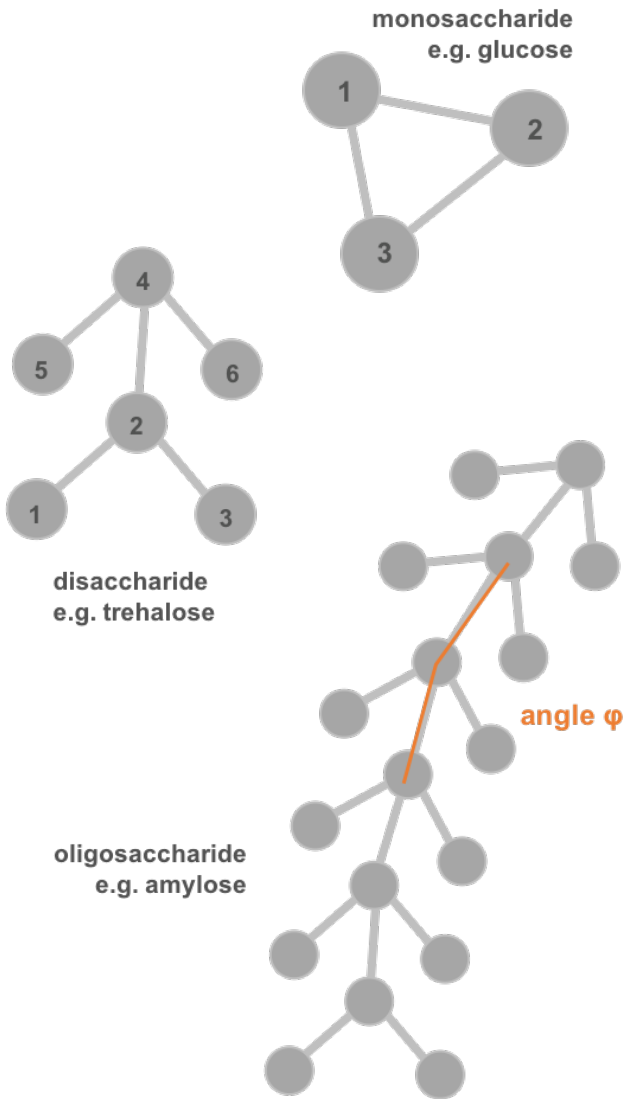


Figure 4.4: Illustration of the approach taken in the first work extending Martini to model carbohydrates. Monosaccharides such as glucose were modelled using three connected CG beads, whilst in disaccharides and oligosaccharides such as trehalose and amylose, each constituent monomer was modelled as a linear set of beads linked by a single harmonic bond with two dihedral angle terms (not shown) to define their relative orientation. In the case of sequences of three or more monosaccharide units, an additional angle term was included to better model higher order oligosaccharide conformations.

the local translational and rotational dynamics of the glycans, precipitating the requirement for an elastic network to achieve realistic conformational properties.

A similar approach had previously been taken to improve glycan models for Martini 2[214], with a focus on glucose, sucrose, the A2 glycan, and α– and β–cyclodextrin. By incorporating B_{22} into the parameterisation process, a scaling factor could be determined that, when applied to the 12–6 Lennard Jones potential, achieved better match against experimental B_{22} values.

To my knowledge, there has been no effort to comprehensively parameterise a set of glycan building blocks for use in CGMD simulations with Martini. In this chapter, I discuss work towards developing and testing a set of parameters for monosaccharides for use in Martini 3. It is my intention that these should be modular, as the amino acids and nucleotides in Martini are, so that they can be used to build up oligosaccharides and polysaccharides without any requirement for bespoke parameterisation of each new glycan structure studied.

4.3 Parameterising the glycan building blocks

The monosaccharides

The potential diversity in a cell's glycome (its total expressed glycan repertoire) is vast and capable of variations in space and time, and it is therefore important that a modular approach to modelling this class of molecule is adopted. That is, robust models for the individual monosaccharides need to be developed, along with parameters for the range of types of bonds in which they might be involved, so that there is the potential to explore the complexity of glycan systems. Furthermore, the models must be readily modifiable in light of new data that may emerge and suggest a need to refine or adjust the initial parameters to better match experimental results or biological systems. Such an approach to forcefield development is in-line with the premise of the Martini forcefield, as illustrated in its initial extension to proteins and later to nucleic acids[160, 189]. In the present work, I do not profess to devise a final model for CG glycans, but rather to make some early progress towards this end.

Hundreds of distinct monosaccharides have been identified in nature. A selection of the more commonly found is summarised in Figure 4.5, along with the associated symbol according to the Symbol Nomenclature For Glycans (SNFG)[414]. Appendix A provides a visual reference for the structures and symbols of the monosaccharides discussed most frequently in this and subsequent chapters. Of

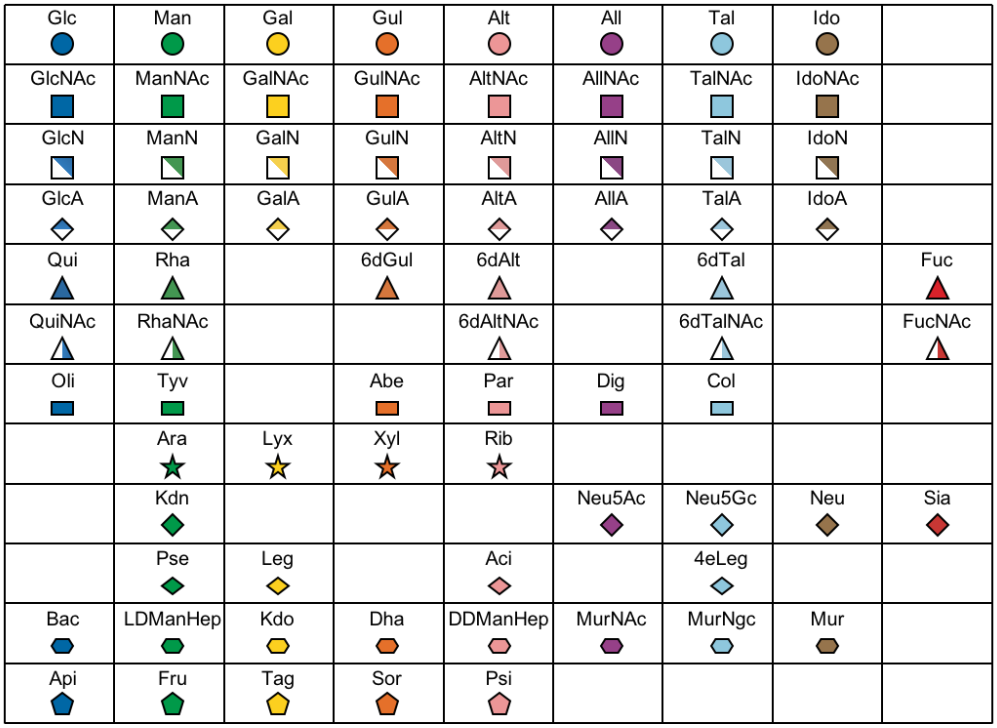


Figure 4.5: Chart of some monosaccharides commonly identified in biological systems, showing the symbol and nomenclature scheme that is standard in glycobiology[414]. The shapes and colours are consistent with stereochemisty for hexoses, hexosamines, N-acetylhexosamines, hexuronates, and pentoses (rows 1–4 and row 8) e.g. Glc is an isomer of Ido and Alt, and has the same arrangement of substituents to the ring as GlcNAc, GlcN, Qui, and GlcA. The shape alone is consistent for deoxyhexoses, deoxy-N-acetylhexosamines, dideoxyhexoses, and dideoxynonulosonates (rows 5–7 and row 10) e.g. Fuc is an isomer of Rha and Qui.

this limited set, I only endeavour in the present work to construct models of a subset of those most frequently detected in eukaryotic systems, and isomers of these that would necessarily be described using the same CG model. By selecting a small subset, I can more rigorously test and establish a suitable methodology for parameterising monosaccharides than would be possible if a more extensive coverage was to be attempted at this stage. The general approach presented might then, I hope, be applied to other monosaccharides, with particular attention perhaps being due to those belonging specifically to plants, fungi, bacteria, and archaea, which are underrepresented in the present selection.

Eukaryotic glycans are composed, for the most part, of a relatively limited number of different monosaccharides. There are infinitely many possible permutations of these, with different chain lengths, degrees of branching, and modifications to the sugar units, but the underlying monosaccharides themselves show a small degree of variation. Therefore, I only need initially develop a rela-

tively small number of distinct CG monosaccharides to facilitate the modelling of a diverse range of biologically-relevant glycans. Indeed, the set chosen to be modelled here are sufficient for subsequent work developing means to simulate at least the more commonly occurring gangliosides and glycoproteins found in animals, as well as several in other eukaryotic clades.

General approach

Parameterising molecules for a CG forcefield benefits from a general simplicity compared to the corresponding task in AT. In my case, this is clearly exemplified by the relatively small number of molecules I need to parameterise as several of the monosaccharides are similar enough (indeed, are isomeric) that a single CG model suffices to describe them all. This does, of course, engender a concomitant loss of structural detail, but that is an unavoidable consequence of coarse graining. Thus, using seven distinct CG models, I cover a total of 40 different AT monosaccharides, the details of which are subsequently discussed. However, with this comes the challenge of developing models that can in fact to reproduce the behaviour of multiple slightly different monosaccharides with a reasonable degree of accuracy.

In principle, the process of developing a CG forcefield requires two steps. The first is to decide on how the system of atoms shall be converted into a simpler system of CG interaction sites, that is, to elect upon the mapping. It is then necessary to refine the parameters of these sites, including their bonded and non-bonded interactions, in order to reproduce the behaviour of the AT molecules so far as is possible given the reduced representation. The initial mapping is done on the basis of prior knowledge of the chemical compositions of the monosaccharides and the nature of the specific bead types available in the Martini force field (Chapter 2). Once the mapping is chosen upon, I need to obtain suitable values for the bonds and angles within the monosaccharide units. This is done using data of the corresponding distances and angles from ATMD simulations.

Mapping

Overview

When choosing how to map an AT molecule to CG, one must consider the system and processes of interest. The mapping ought to preserve all features necessary to describe the phenomena under investigation and to capture all the relevant slow, large amplitude motions of the molecules involved, whilst eliminating sufficient

spurious details and high frequency motions to yield a significant increase in computational efficiency [130]. In my case, I do not want to deviate too far from the average 4:1 and chemical-group-based mapping that typifies Martini [49, 160, 359] as the glycans will be part of larger systems modelled using this forcefield. Furthermore, any finer grained model would come at greater computational cost, which would render many of the types of system of interest intractable.

In the Martini forcefield there are a number of different bead types to represent different set of chemical groups, and each of these can be in any of three sizes depending on the number and arrangement of the underlying heavy atoms [49, 160], as discussed in detail in Chapter 2. I therefore start my mapping with a clearly defined idea of the available CG beads and the types of chemical groups they can serve to represent.

Initial mapping

The first step in the mapping procedure is to decide on how to group together heavy atoms in the molecules being modelled. As all the monosaccharides are being modelled in their pyranose ring structures, the models will all centre around a coarse-grained version of this. Considering first the hexoses (Glc, Man, galactose (Gal) etc.), the molecule consists of 12 heavy atoms. A 4:1 mapping would give three CG beads, each an S-type (small) bead to account for the fact that several of the atoms are within the ring itself.

Now, I break the ring into three S-type beads: two beads cover -CH(OH)-CH(OH)- groups of carbons 1–4, and the final one accounts for the -C(CH2-OH)-O- group including carbons 5 and 6. These beads are all linked together into a triangle through bonds between the CG beads. To model the other monosaccharides, I consider variations on this scheme. For instance, xylose (Xyl) and the deoxyhexoses have effectively smaller rings due to the loss of -CH2-OH and -OH, respectively, so the size of one bead can be decreased to account for this. The *N*-acetylhexosamines and sialic acid have more extensive substituents to the ring, which will require additional CG beads to be linked to the core unit. This basic grouping is shown in the left-hand column of Figures 4.6 and 4.7 for each of the monosaccharide sets to be modelled.

With this basic grouping in mind, I can then consider what bead types ought to be used to represent each of the groups of heavy atoms so as to match the

underlying chemical properties. There are general notions of which Martini beads can be used to represent particular types of molecules, as discussed in Chapter 2, and this is the primary means by which I elect upon a preliminary mapping scheme, which is shown in the central column of Figures 4.6 and 4.7. These bead choices were based on comparison of the underlying AT chemical groups with the available Martini 3 bead types and their recommended use. For instance, SP3 is recommended for ethanediol, which is chemically quite similar to the -CH(OH)-CH(OH)- of a monosaccharide ring. For the charged sugars, Q-type beads are assigned, in both cases being given a -1 charge.

Intramolecular bonds

General approach

The next stage of the parameterisation process is the choice of intramolecular bonded parameters that hold the beads together within a monosaccharide unit. In order to do this, I carried out ATMD simulations of the corresponding monosaccharides, where models for them were readily available, and from these extracted mean values for the distances and angles between grouped atoms in order to provide initial values for the bonded parameters in my CG models.

Atomistic simulations

There are several AT forcfields with parameters for sugars. I chose to carry out my simulations using the CHARMM36, which includes parameters for at least one of the monosaccharides included withtin each CG residue model. Specifically, Gal, Glc, Man, Fuc, rhamnose (Rha), Xyl, glucuronic acid (GlcA), GlcNAc, *N*-acetylgalactosamine (GalNAc), Neu5Ac, allose (All), altose (Alt), quinovose (Qui), galacturonic acid (GalA), IdoA, and GlcN. Unfortunately, while the hexoses are well covered by the force field, there are far fewer available examples of the other groups, and so it is possible that the models developed for these will be less applicable across the range of covered AT structures than the hexoses. However, the selection of parameters available does, to some extent, reflect the biological abundance of the monosaccharides, and so the actual practical consequences of this deficiency may not be particularly detrimental.

A single copy of each of these AT monosaccharides was simulated for 500 ns in a cubic box of TIP3P water with NaCl to a concentration equivalent to 0.15 M, with appropriate numbers of each ion to equalise the system when charged

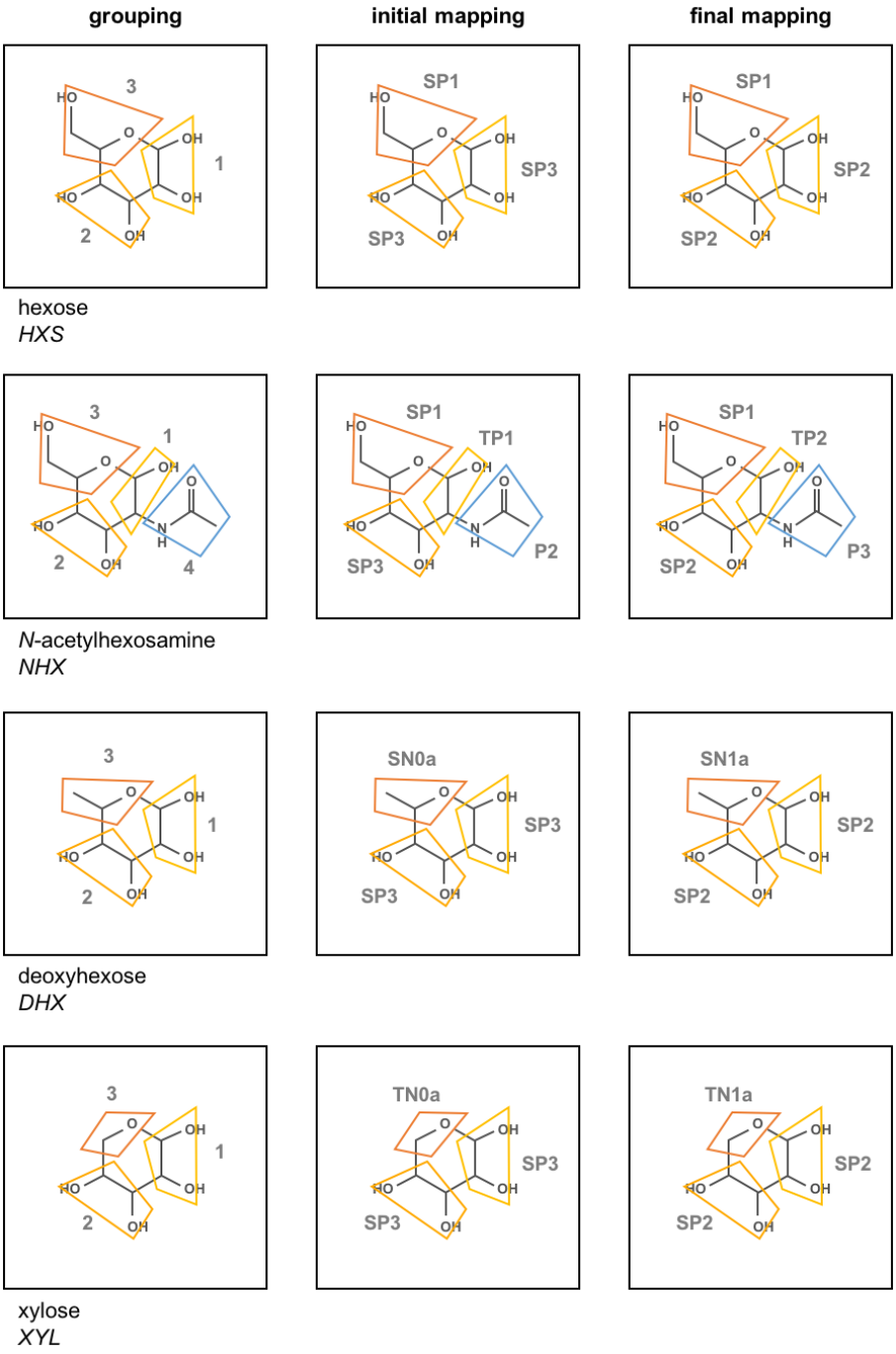


Figure 4.6: Overview of the grouping (left), preliminary (middle), and final (right) mappings of the CG monosaccharides corresponding to the hexoses (e.g. Glc), *N*-acetylhexosamines (e.g. *N*-acetylglucosamine (GlcNAc)), deoxyhexoses (e.g. fucose (Fuc)), and Xyl. In each case, a general skeletal structure for the group is shown, with the heavy atoms to be grouped into CG beads boxed together. The numbers in the left-hand column indicate the CG bead number, whilst the bead types in the middle and right-hand columns are the preliminary and final, respectively, bead types chosen for each of those groups. The name of the CG residue is given: HXS, NHX, DHX, and XYL, respectively.

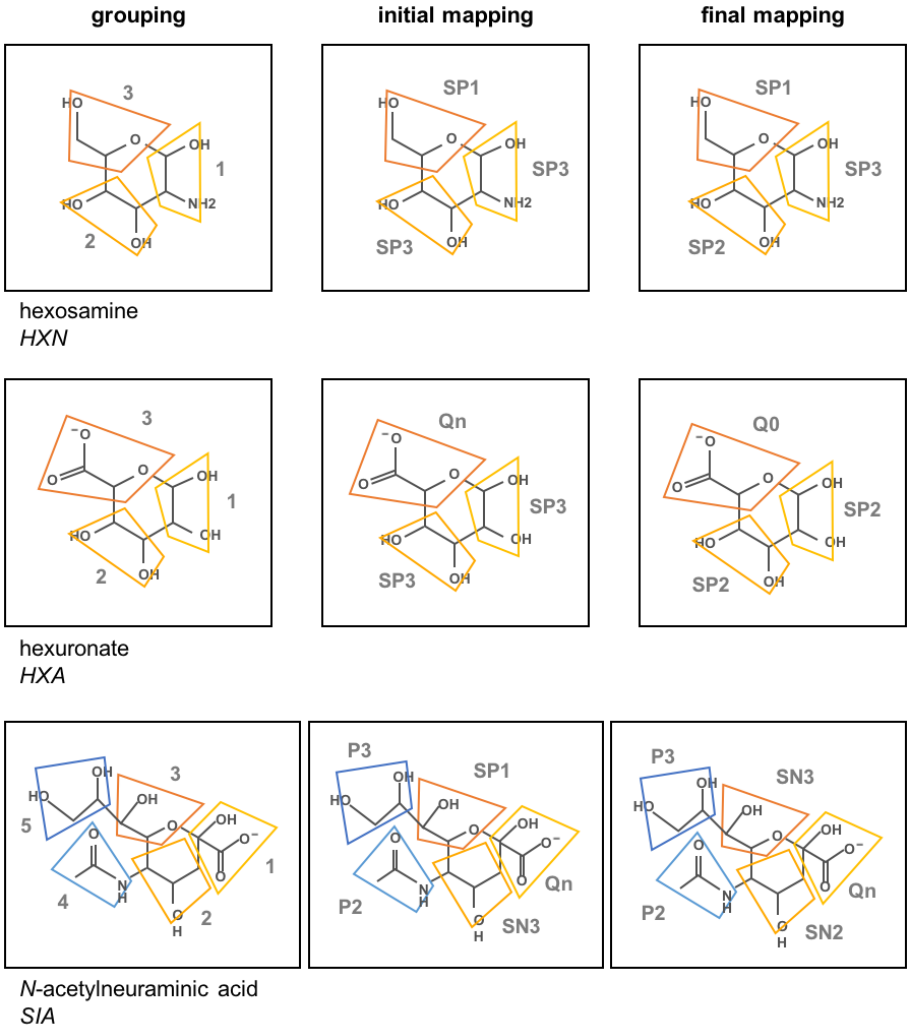


Figure 4.7: Overview of the grouping (left), preliminary (middle), and final (right) mappings of the CG monosaccharides corresponding to the hexosamines (e.g. glucosamine (GlcN)), hexuronates (e.g. iduronic acid (IdoA)), and sialic acids (specifically, *N*-acetylneuraminic acid (Neu5Ac)). In each case, a general skeletal structure for the group is shown, with the heavy atoms to be grouped into CG beads boxed together. The numbers in the left-hand column indicate the CG bead number, whilst the bead types in the middle and right-hand columns are the preliminary and final, respectively, bead types chosen for each of those groups. The name of the CG residues is given: HXN, HXA, and SIA, respectively.

monosaccharides were being simulated. From the simulations, the distances between the centres of mass of the heavy atoms grouped together to form CG beads in the corresponding models were measured, and the mean values over the trajectories calculated. For GlcNAc, GalNAc, and Neu5Ac, angles to the substituent groups were also computed. These values were averaged over the trajectories (and monosaccharides where more than one was simulated for a given CG model), and the results used as the initial bond lengths and angles in the monosaccharide models. The resulting 'bond' lengths and angles are given in Table 4.1.

For stability, and to reduce the degrees of freedom, it was decided that the bonds within the monosaccharides should be treated using constraints, essentially fixing each monosaccharide rigidly in itself. I can justify this to some extent by the very low standard deviations observed in the distances computed from ATMD simulations, suggesting that there is to be expected very little variation around the equilibrium 'bond length'. The constraints can later be changed to harmonic bonds if necessary and possible, but the impact of doing so is not addressed in the present work. The angles were treated using harmonic potentials with a force constant provisionally set at 50 kJ mol^{-1} rad^{-2}.

Evaluating the models

The water-octanol partition coefficient

One of the main factors to consider in my parameterisation is that the physico-chemical properties of the monosaccharides must be captured with reasonable accuracy using the chosen mapping scheme. In particular, I wish to recapitulate to some degree the hydrophilicity of the monosaccharides so that they partition in a physiologically feasible fashion between protein and membrane surfaces and the surrounding aqueous solvent.

The octanol-water parition coefficient is a parameter frequently used to express the lipophilicity of a compound, employed extensively within the fields of pharmacology and environmental science. A partition coefficient, generally speaking, can be used to quantify the relative solubility of a chemical species between two solvents. The octanol-water partition coefficient P more specifically reflects the relative solubility of uncharged species in water versus octanol. In a biphasic mixture of the two solvents, a more lipophilic species will tend to dissolve in the organic phase, whereas a more hydrophilic one will more readily

residue	bond	distance (nm)		angle (deg)	
		mean	*stdev*	mean	*stdev*
HXS	1-2	0.32	*0.01*		
	2-3	0.32	*0.01*		
	1-3	0.36	*0.02*		
NHX	1-2	0.33	*0.02*		
	2-3	0.33	*0.00*		
	1-3	0.33	*0.00*		
	1-4	0.31	*0.01*		
	2-1-4			81	*4*
	3-1-4			139	*3*
HXN	1-2	0.33	*0.00*		
	2-3	0.32	*0.00*		
	1-3	0.37	*0.00*		
HXA	1-2	0.31	*0.03*		
	2-3	0.32	*0.02*		
	1-3	0.40	*0.01*		
DHX	1-2	0.30	*0.02*		
	2-3	0.29	*0.01*		
	1-3	0.30	*0.02*		
XYL	1-2	0.33	*0.00*		
	2-3	0.28	*0.00*		
	1-3	0.26	*0.00*		
SIA	1-2	0.35	*0.02*		
	2-3	0.32	*0.00*		
	1-3	0.36	*0.01*		
	2-4	0.34	*0.00*		
	3-5	0.29	*0.00*		
	1-2-4			130	*1*
	3-2-4			72	*1*
	1-3-5			106	*1*
	2-3-5			149	*0*

Table 4.1: Values for the CG bond distances and angles calculated from ATMD simulations of the monosaccharides. The averages were taken over the trajectories, and over the simulations where multiple different AT monosaccharides were simulated (e.g. for HXS I averaged over all the different hexoses simulated). The numbers refer to beads in the monosaccharide ring, with the numbering as in Figures 4.6 and 4.7.

dissolve in the aqueous phase. The value of P therefore provides insight into the physicochemical properties of a molecule.

In practice, P is often expressed as a log value:

$$logP_{oct/wat} = log\left(\frac{[solute]_{octanol}}{[solute]_{water}}\right) \tag{4.1}$$

A negative logP suggests hydrophilicity, whilst positive logP suggests lipophilicity. A logP value of zero indicates the molecule partitions equally between the two phases. logP can be directly related to the free energy of transfer between octanol and water ΔG:

$$logP = \frac{-\Delta G}{RT \ln 10} \tag{4.2}$$

where R is the molar Boltzmann constant and T the temperature in K.

The value of logP depends on the choice of the organic solvent and, to a lesser degree, the conditions under which the measurement was made. It is therefore a useful parameter for helping to predict how molecules will behave in physiological and ecological settings.

Although it is possible to measure the partition coefficients of molecules experimentally using e.g. HPLC, NMR, in many cases where logP is of interest, such data may not be already available. An alternative means to estimate the value of this parameter is to predict it from the chemical structure of the molecule. There are several general schemes by which this can be done, including atom- and fragment-based methods. These work by combining the predicted contributions to logP from the constituent atoms or non-overlapping molecular fragments, respectively, often in the latter case with corrections to account for steric and electronic effects within the complete molecule.

The predicted $logP$ and ΔG values for the monosaccharide groups to be modelled are given in Table 4.2. Note that these predictions do not differ between the various isomers. In addition to the predicted values, where available, experimental results from the literature are given.

My first method of evaluating the models produced by combining the initial mapping with the bonded parameters was to assess the extent to which the CG monosaccharides matched the logP values predicted for their AT counterparts.

monosaccharide group	logP	dG (kJ mol^{-1})	experimental logP
hexose	-2.9	16.7	-3.24[415],-2.8[416]
N-acetylhexosamine	-3.2	18.4	
hexosamine	-3.0	17.23	
hexuronate	-2.6	14.93	
deoxyhexose	-1.9	10.9	
xylose	-2.5	14.1	
Neu5Ac	-3.6	20.7	

Table 4.2: Octanol-water partition data for the groups of monosaccharides to be modelled. The logP values are predictions made using ChemAxon, obtained from DrugBank[417]. The conversion to dG used a temperature of 300 K, the same temperature at which the logP values were predicted for the Martini bead types during their parameterisation.

In order to do this, I carried out alchemical free energy calculations to estimate the energy changes associated with CG monosaccharide solvation in water and hydrated octanol[418]. The partition free energy associated with logP is then obtained from the difference between these two values.

In principle, and were I to attempt to replicate what the water-octanol partition coefficient truly represents, I would simulate each solute in an equilibrium biphasic mixture of water and octanol. As the free energy difference between the solute in either phase is directly related to the relative probability of it occupying either, by measuring how often the solute occupies one phase or another over a long period of time, I could in theory obtain the desired partition free energy. However, for many of the monosaccharides, I would need to wait an unfeasibly long time for a transition from the more favoured phase to the less favoured one to occur, let alone for it to do so sufficiently many times to obtain reasonable statistics for the event.

I therefore take a different approach, in which the energy change between the solvated and vacuum states of the solute is calculated for each solvent, and then the difference between these used to extract the partition energy according to the thermodynamic cycle.

To calculate the energy difference between the vacuum and solvated state of the solute, I use a variable λ to modify the interactions between the solute and the rest of the system such that when λ is zero the solute is effectively in vacuum (no interactions with the rest of the system), whilst at λ of one, it interacts fully with the system. The difference in energy between the two states is then calculated using a method such as the Bennett Acceptance Ratio (BAR) method, which

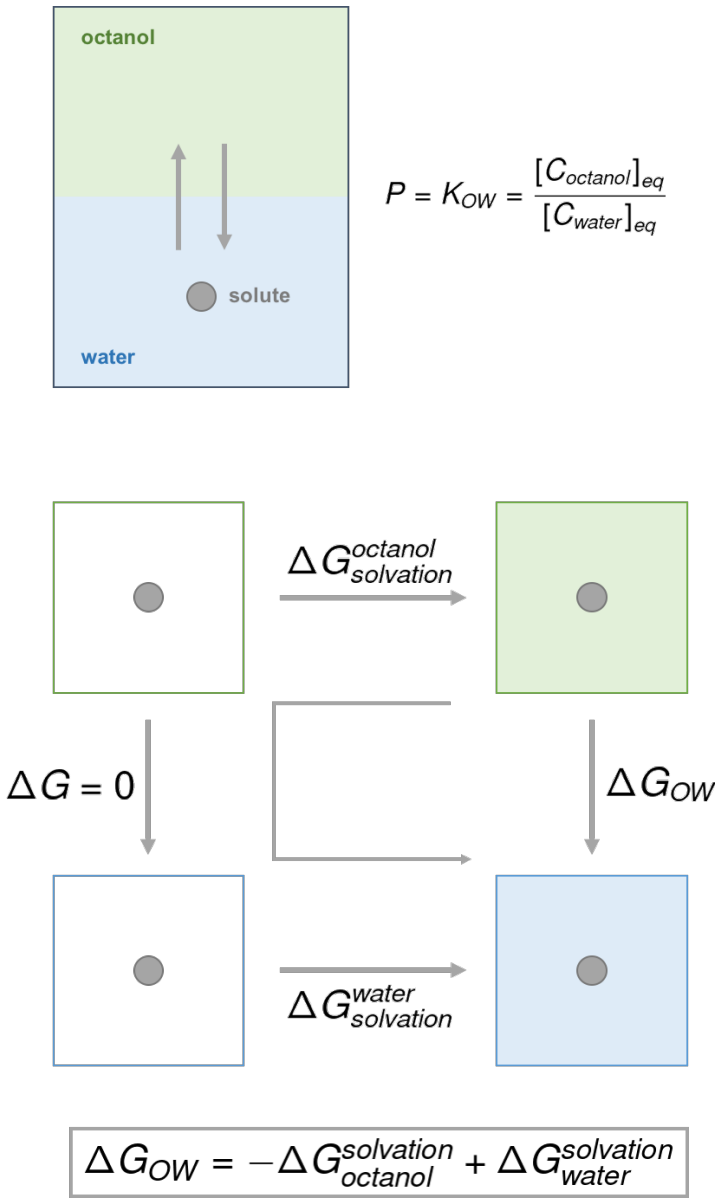


Figure 4.8: Schematic illustration of the theory behind the octanol-water partition coefficient. The parameter reflects the relative occupancy of a solute in octanol verus water phases at equilibrium (top), whilst my computational approach involves simulating in octanol or water systems and calculating free energy differences between the solute in the different solvents. This is done using the thermodynamic cycle shown (bottom).

allows the calculation of the free energy difference between pairs of simulations at λ_A and λ_B, provided these two are close enough to have some overlap in phase space.

The obvious choices for λ are 0 and 1 i.e. the end states for the transformation from 'non-existent' to 'existent' solute. However, it is likely that there will be very little phase space overlap between these end states, the errors on the calculation would be extremely high, and my results would have little hope of convergence. I therefore need to construct a series of intermediates that take us between the end states. These are nonphysical entities between which there is sufficient phase space overlap for a free energy difference to be calculated. Thus, interactions with the system are slowly turned on over a series of simulations run at different λ values in the range 0–1. The energy difference is then calculated between neighbouring simulations at neighbouring λ values, and combining these ultimately yields the overall energy difference for the transformation. In the case of my work, as logP relates to neutral solutes only, I only turn on/ off the van der Waals interactions, ignoring the electrostatic interactions. As subsequently discussed, this is a problem when considering parameterisation of the charged monosaccharides hexuronic acid, and Neu5Ac.

This procedure must be done for both water and for octanol, and then the results are compared to get the octanol-water partition using the thermodynamic cycle shown in Figure 4.8. In comparing with the predictions in Table 4.2, I used the free energies of partition, rather than logP (the log of this energy difference), as any errors will show up more in the former than the latter.

The simulations I carried out were as follows. I ran two sets of simulations, one with CG water alone and the other with an approximately 12:1 ratio of octanol to water, a ratio based on equilibrium simulations of a biphasic mixture of the two solvents (data not shown), and chosen to represent a hydrated octanol system. Hydrated octanol was chosen because the logP is based on the partitioning of a solute between a biphasic system, within which the octanol phase will invariably have some water content. The solute (CG monosaccharide) was positioned in the simulation cell of the appropriate solvent so that it was at least 2 nm from any edge, and then a series of simulations carried out for the same duration of 50 ns at 51 different values of λ , coupled to the van der Waals interactions, and spaced at intervals of 0.02 in the region from 0.00 to 1.00. The simulations were carried out at temperature 300 K and isotropic pressure of 1 atm.

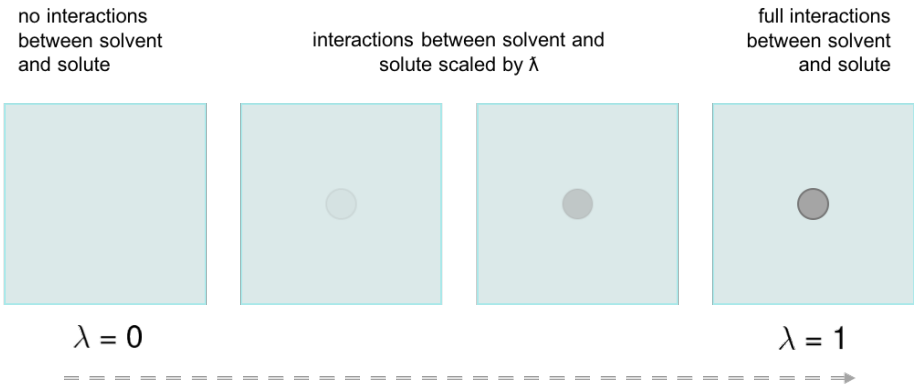


Figure 4.9: Simple illustration of the free energy calculation process, in which the solute is slowly 'phased' into existence within the given solute by increasing the value of λ over a series of simulations. The overall free energy difference between the end states at λ of zero and one, respectively, is calculated by combining the smaller energy differences between intermediate states.

The actual simulations were carried out following an energy minimisation and a subsequent equilibration run. After the equilibration was completed, bringing the system to the desired temperature and pressure, the system was split into the different values of λ , and the MD simulation run at each of these different values.

At the end of the simulations, the free energy differences between each neighbouring pair of λ values was evaluated using BAR. Ultimately, the method yields an overall free energy estimate for the transition from the solute being 'absent' from the cell to that of it being present. These two energy changes were then subtracted from one another to yield an estimate for the difference between the solute in water and octanol i.e. the partition energy.

The results are shown in Table 4.3 and Figure 4.10. The extent to which the computed values agree with the predictions varies somewhat and in no consistent manner. The initial models for HXS and DHX, for instance, appear somewhat too hydrophilic, whilst that for NHX is too hydrophobic. The two charged monosaccharides show the greatest difference between the CG and predicted AT partition energies. This is expected, as logP reflects the partitioning behaviour for uncharged solutes, and these sugars are predicted to be negatively charged at neutral pH. There is therefore uncertainty in the validity and accuracy of using logP as an estimate in these cases, particularly because the Martini model does not have the ability to be an uncharged version – a protonated carboxylate would require an entirely different bead type from a deprotonated carboxylate, while

monosaccharide	WN	OCOS	dG	predicted AT
HXS	38.2 ± 0.03	20.2 ± 0.06	18.0	16.7
NHX	48.3 ± 0.06	30.9 ± 0.04	17.4	18.4
HXN	37.7 ± 0.01	20.0 ± 0.05	17.7	17.2
XYL	31.7 ± 0.01	17.3 ± 0.05	14.4	14.1
DHX	35.3 ± 0.02	21.4 ± 0.03	13.9	10.9
HXA	50.9 ± 0.03	28.1 ± 0.05	22.8	14.9
SIA	79.4 ± 0.09	54.8 ± 0.07	24.6	20.7

Table 4.3: Summary of the octanol-water partition energies (all in kJ mol^{-1}, ± standard deviation) calculated from simulation, as compared to the values predicted for the AT structures and for the mappings based on the individual beads.

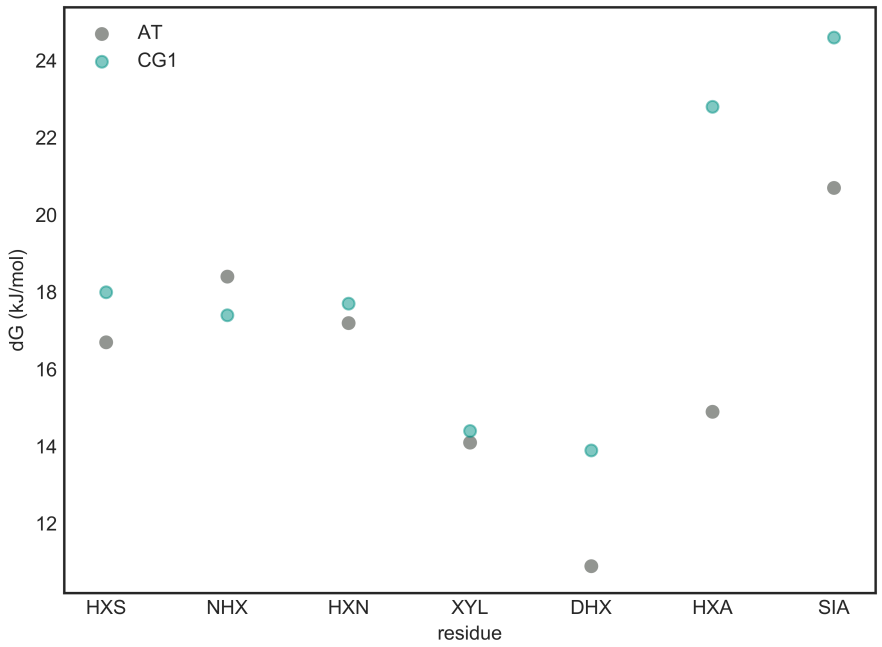


Figure 4.10: Scatter plot of the computed partition energies for my preliminary CG models (turquoise), compared with those predicted from the AT structures (grey). The standard deviations on the computed values (less than 0.1 kJ mol^{-1}) are too small to be seen on the plot.

in experiment and predictions using the AT structure, such neutralisation of the molecule can occur.

As the core of most of the monosaccharides is based upon the HXS residue, I focused my improvement of the models on changes made to this. Specifically, I adopted two approaches to reduce the hydrophilicty: reducing that of bead 3 (corresponding the the extra-ring hydroxyl and ether groups, C5–C6), and reducing those of both beads 1 and 2 (corresponding to the diols, C1–C4). In the first case (model 2), I altered the identity of bead 3, changing it from SP1 to SN3,

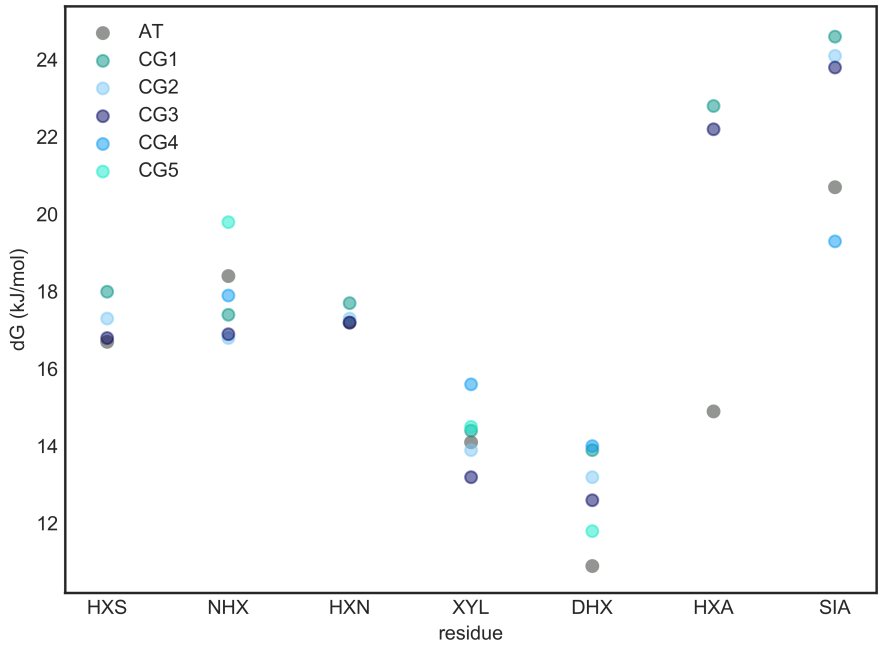


Figure 4.11: Scatter plot of the computed partition energies for my CG models (shades of blue/green), compared with those predicted from the AT structures (grey). The standard deviations on the computed values (less than 0.1 kJ mol^{-1}) are too small to be seen on the plot. The models chosen to be used for each monosaccharide are those with the closest match to the predicted partition energy. Full details on all the models tested are given in Appendix D.

whilst in the latter (model 3) I changed the other two beads from SP3 to SP2. I applied these changes to all the monosaccharides (where the corresponding chemical groups exist) with no other alterations compared with model 1. The data for these sets of models are shown in Figure 4.11.

From these plots, I can deduce the 'best' models from my trials, so far as the octanol-water partition coefficient is concerned. The final models are shown in the right-hand column of Figures 4.6 and 4.7. These are the ones that I carry forward into the second stage in my parameterisation protocol, in which I assess the capacity of these monosaccharides to reproduce experimentally observed properties of monosaccharides, specifically the density in aqueous solution at different concentrations, and the cryo-protection of lipid bilayers.

There are some experimental data on the octanol-water partitions of a small number of different sugars. For example, by an enzymatic method at 293 K, it was found that the logP value for Glc is -2.8, which gives a partition energy of 15.8 kJ mol^{-1} [416]. This is not wildly different from my computational value. However, parameterisation here was done to match the predicted logP

values given the completeness of that dataset and the incompleteness of the experimental data. Ideally, I would have had a complete set of recently obtained experimental data on sugar octanol-water partitions, all obtained under the same conditions, rather than having to rely on theoretical predictions. It would not be challenging to carry out the experiments to deduce these parameters, but as it is in itself of little direct biological or medicinal interest, it has not, to my knowledge, been done and I have yet to have the opportunity to do it.

It would also be useful to be able to do the same partition calculations for oligosaccharides, but the larger and more complex the molecule, the less reliable the calculations become and the results could therefore be misleading. I hope that further refinement and validation on test systems will reveal any errors in interactions introduced at this stage, and so the final models will not overly suffer from this limitation.

Further work

As noted previously, the parameterised monosaccharides are but a representative selection of the most common types found in eukaryotic glycans. The set covers neither the full eukaryotic repertoire, nor the vast majority of bacterial-specific monosaccharides, and is therefore incomplete. Whilst sufficient for one to start modelling and examining glycosylation, it should be extended to comprehensively cover a greater part of the range of biologically relevant monosaccharides.

Furthermore, the sugars at present do not have parameters for their modification by sulphation or phosphorylation. This should be relatively straightforward to incorporate into the CG models through the addition of an appropriately charged bead, although I will need to determine the best way to validate these modifications given the noted difficulties associated with estimating or computing partitions for charged species.

The mapping is, at present, preliminary and based on a combination of experience with the Martini forcefield and chemical knowledge regarding the nature of sugars. It does not necessarily follow that the glycans produced using this mapping will behave in a physiologically relevant manner, and so it may become necessary to subsequently modify it based on newly available data, and if application of the models to test systems reveals hitherto unidentified flaws.

Additionally, the granularity of the sugars might be modified at a later date in order to increase computational efficiency and render tractable increasingly

complex biological systems. In many cases, it seems that the precise structural details of the individual sugars is not the decisive factor in their biological function, but rather it is their overall occupancy on protein/membrane surfaces or general physicochemical features that govern their influence on membranes and proteins bearing them. If this is indeed the case, it should be possible to simplify the sugars further without loss of model accuracy as relates to the features of interest. Specifically, I have considered modelling each individual monosaccharide using a single central bead and, where relevant (e.g. GlcNAc, Neu5Ac) bonding this to substituent beads. This would greatly simplify the systems, but runs the risk of losing some biological accuracy. Testing such a simplified model will therefore be of considerable interest at a later stage in the project once I have better understanding of the behaviour of the current model.

4.4 Interresidue bonded parameters

Approach

Having developed initial models for a set of monosaccharides, the next step in developing CG glycans is determining how to best model the interactions between these residues when they are covalently linked into oligo- and polysaccharides. That is, I must determine the connectivity of the CG beads so that, once in motion, the glycans behave over time as their AT counterparts might do in a real biological system.

The general approach I took was similar to that employed for the bonds within monosaccharides viz. I carried out ATMD simulations of suitable molecules, and from these extracted data regarding the required bonded parameters. However, when modelling the bonds between monosaccharides, there are more dynamic features to consider as compared with the individual sugars, which I treated as relatively rigid within themselves. I need to choose sets of bonded parameters that allow a reasonable reproduction of the dynamics of the oligosaccharides without being so detailed as to cost computational efficiency.

The approach I take to inter-residue bonded interactions is as illustated in Figure 4.12. Briefly, for each pair of sugar residues, a harmonic bond is defined between a virtual site at the geometric centre of the three-bead ring. These virtual sites, in addition to being used for the linking of monosaccharides together, offer the potential for modifying the interactions between monosaccharides in future work, specifically offering the potential to include a description of ring-stacking

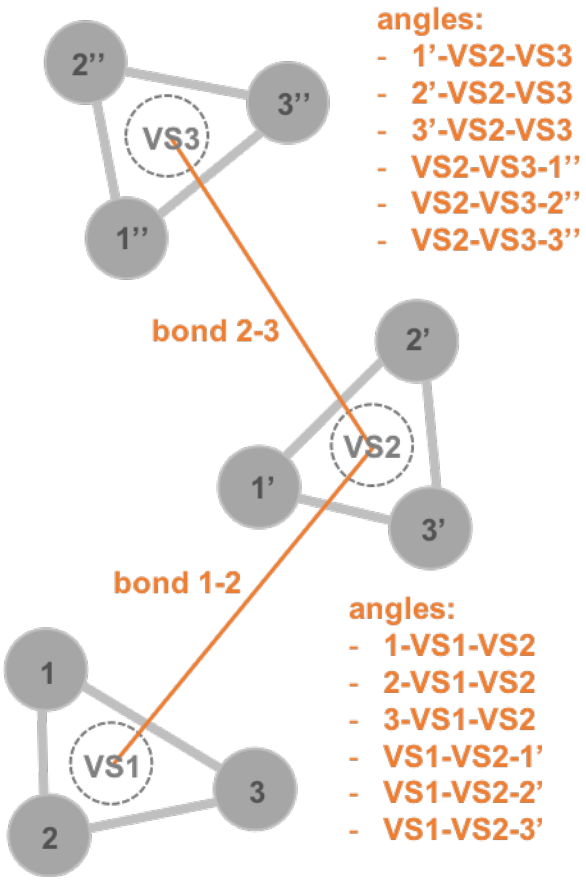


Figure 4.12: Illustration of how the bonds and angles between Martini monosaccharides are to be described. Between each pair of residues there is a single harmonic bond length potential linking virtual sites (VS) at their centre of geometry, and six harmonic bond angle potentials describing the spatial relationships of the beads from each other. Both the bond length and angles vary between different pairs of residues and bond types (α, β, number of the carbon involved).

interactions that could not be modelled using a simple three-membered ring of Martini beads alone [Paulo de Souza, personal communication]. The virtual sites do not interact with the system at all themselves, but serve only to help in my modular bonding of monosaccharides together. They are defined within the individual monosaccharide residues, being positioned equidistant from the ring beads, and so can be used for bonded interactions between residues. In addition to this harmonic bond, there are bond angles involving ring beads and virtual sites that define the geometry of the bonded interaction.

In a recent work developing models for *N*-glycans compatible with Martini 3, it was noted that the conformations of the glycans could not match those from ATMD unless an elastic network was applied[412], whilst the first extension of

Martini 2 to glycans discussed the need for an added angle term between trisaccharide units to achieve realistic conformations of amylose and curdan[162]. The use of an elastic network would be challenging in the modular method I intend to use for constructing diverse glycans, but may need to be considered at a later date if aberrant glycan conformations are observed.

Parameters from disaccharides

In order to obtain values for the bonded parameters, I carried out a set of ATMD simulations of bonded pairs of monosaccharides, covering a range of disaccharide units found in biologically occurring oligosaccharides (extracted from the database of glycan arrays available on GlyTouCan[419], discussed in more detail in Chapter 5). These provided a first set of bonded parameters, but the set does not cover the entire range of theoretically possible disaccharide types, nor does it offer even coverage of those included. I therefore may need to extend the examined set at a later date in order to further refine my models. The initial set of 85 distinct disaccharides is at least sufficient to model a good range of biologically feasible glycans. The full set of simulated molecules and data extracted are presented in the supplementary information.

The disaccharides were each simulated for 50 ns in a box of TIP3P water with neutralising NaCl at 150 mM, at 303.15 K, using the CHARMM36 force field. The CG bonded parameters were evaluated using a short Python code written for the purpose. Using the MDAnalysis[363] and NumPy[362] libraries, this program first groups heavy atoms together according to the mapping described in Section 4.2, and then calculates the mean values of distances and angles between the centres of mass of these over the course of the simulations. Where multiple simulations yielded parameters corresponding to a given type of bond, the mean over all the different simulations was taken. In my initial approach, I grouped all bonds occurring between a given CG residue pair (e.g. Glc-α3-Glc with Glc-β3-Glc and Man-α6-Man) together. The results are shown in Figures 4.13 and 4.14.

During the collection of these data, I observed that some bonds were notably longer than others within a group of paired monosaccharides, in particular, the bonds involving C6 tended to be longer than those involving the other carbons. I therefore split the bond length data up to consider how the bonded parameters might differ if I considered each type of bond between a CG residue pair sep-

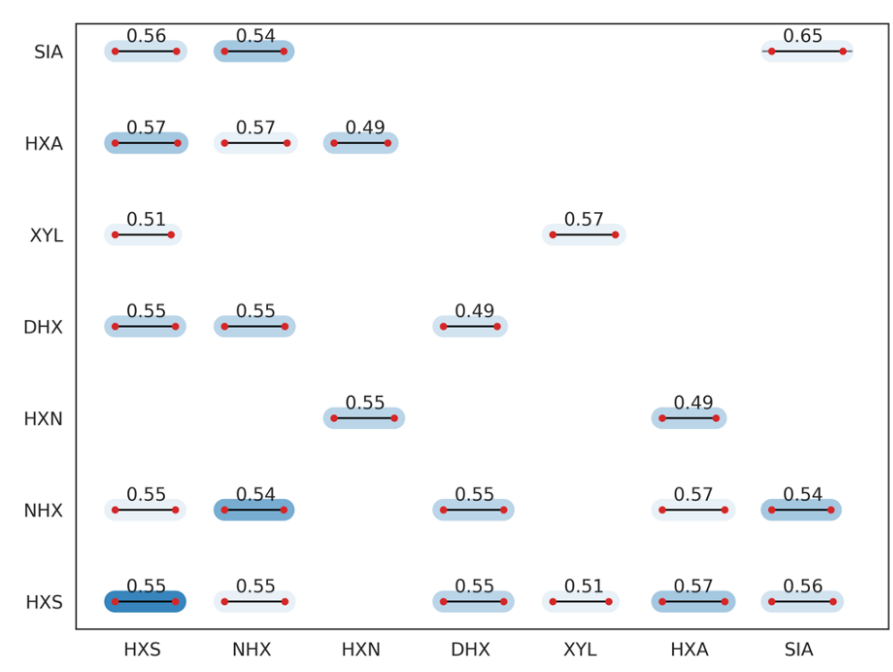


Figure 4.13: Plots of bond lengths obtained from the ATMD simulations of disaccharides. In each case, the bond length is illustrated with a horizontal bar proportional to the mean bond length, with a surrounding blue bar of length indicating the standard deviation. The transparency of this bar is proportional to the number of simulations of this particular pair, with a darker bar indicative of more simulations (i.e. higher n). The value of the mean length is given above the bar, and the name of the CG disaccharide pair shown beneath it. The data are given in full in the Supplementary Information.

arately (Figure 4.15). The data are somewhat limited and unevenly distributed across the residue pairs (for instance, only one SIA-SIA pair was simulated, compared with several HXS-HXS ones). However, it can clearly be seen that bonds involving C6 of one of the monosaccharides do indeed tend to be longer than those involving the other carbons, regardless of whether the bond is an α or β type. Because of this, and not wanting to lose to much of the AT detail, I decided to split the bonded parameters according to the nature of the bond such that the results for bond lengths shown in Figure 4.15 are those that I go on to use in later simulation work on glycans.

4.5 Model evaluation

The bonding scheme I devised is intended to allow a modular approach to glycan construction, as required to reproduce to some extent the heterogeneity so often seen in glycans. Furthermore, the use of Martini 3, even though it remains in development at the time of writing, was decided on to attempt to alleviate one of the issues with the Martini force field, namely an imbalance of solute-

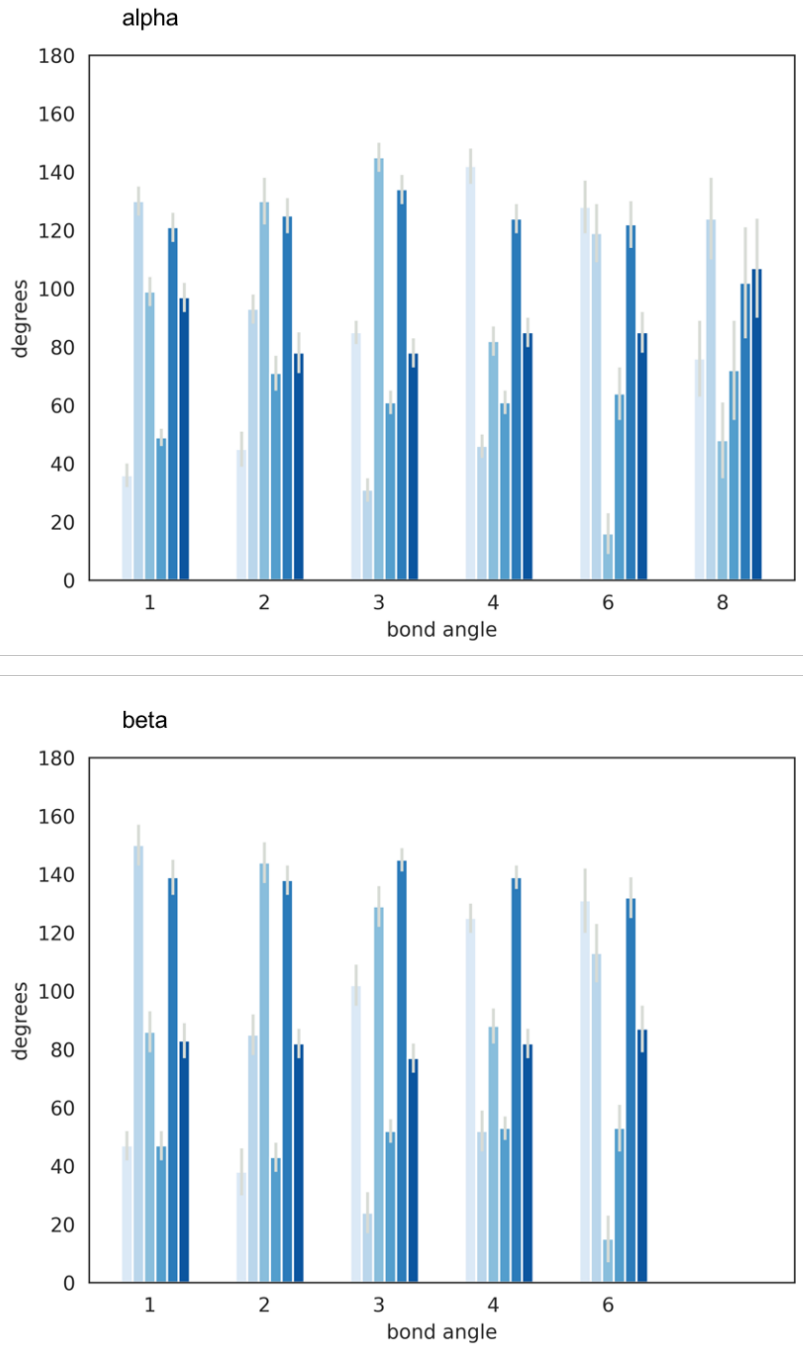


Figure 4.14: Bar blots indicating the angles associated with each type of bond. Each bar corresponds to a single angle, defined as previously discussed, and the grey thinner bar to the standard deviation for that angle.

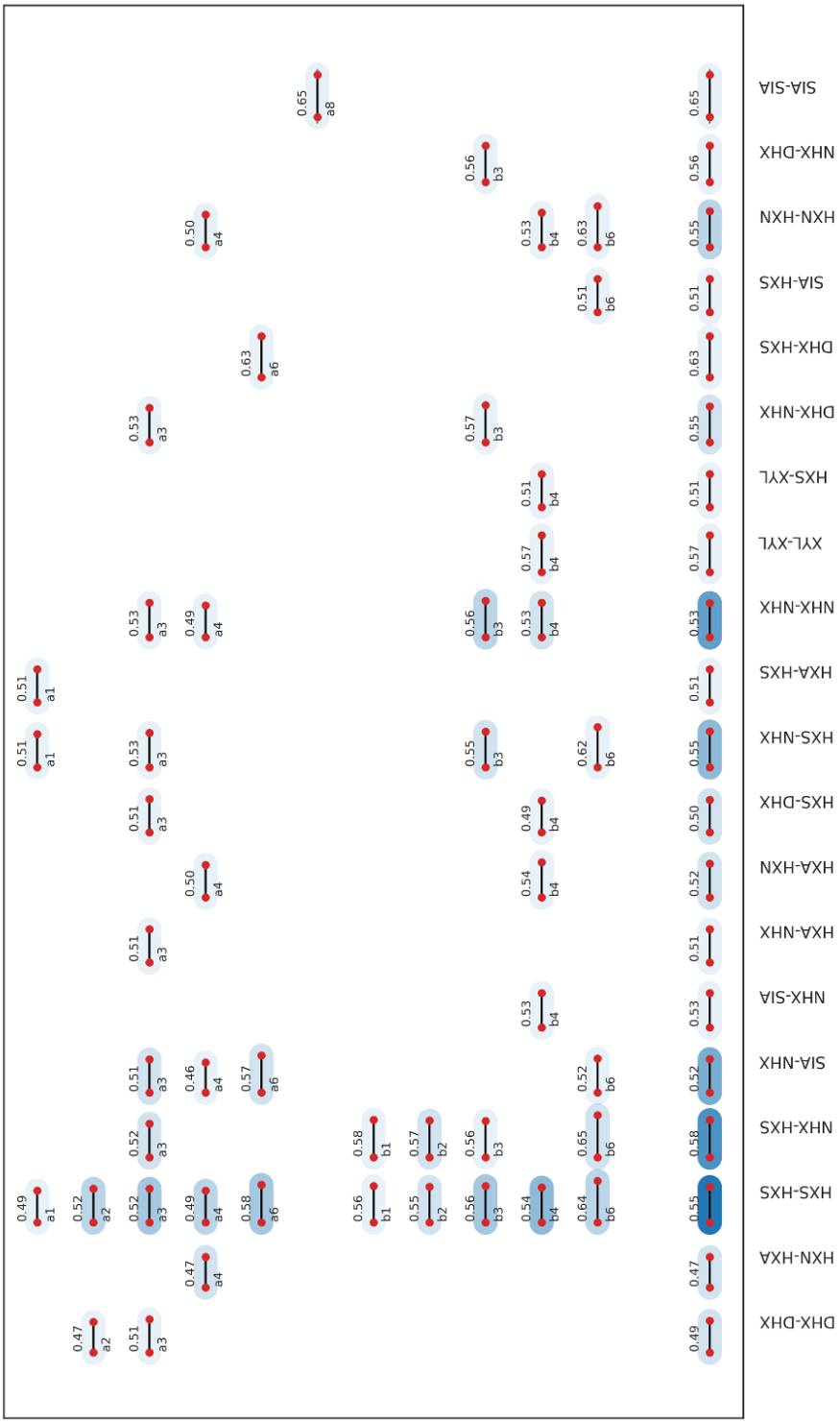


Figure 4.15: Plot indicating the values of individual bond types between pairs of CG residues. At the bottom of each column is the collected data for all pairs of that type (as shown in Figure 4.13). The data are presented as in Figure 4.13.

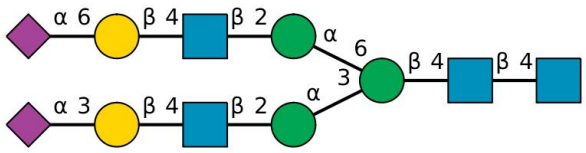


Figure 4.16: SNFG structure of the A2 glycan.

solute, solute-water, and water-water interactions, as discussed in Chapter 2. One consequence of such an imbalance may be an unrealistic propensity, or lack thereof, to cluster at certain concentrations. In this section, I briefly evaluate my glycan models in relation to this issue, as well as more generally demonstrating that it, at least, produces reasonable glycan models. The evaluation at this stage is not extensive, and there is clearly more work to be done. However, as herein discussed, the proof-of-principle appears successful, and the models at least show promise.

Glycan aggregation

In order to evaluate whether or not my model of the 'basic' monosaccharide unit (representing the hexoses) spuriously aggregates in solution, I carried out a series of simulations of this residue, HXS, in Martini water at both low and high concentrations. Each simulation was performed for 1 μs at 310 K, with a simulation cell size of 20 nm x 20 nm x 20 nm. Although I did observe some interaction between the monosaccharides during the course of the simulation, they did not aggregate together as one unit or form any associations that lasted the duration of the simulation (Supplementary Information). This is itself not entirely unexpected as unrealistic aggregation of Martini sugars has been noted as a concern for oligosaccharides, rather than monosaccharides.

I therefore applied the same approach to a simple model glycan for which there are some existing experimental data. Glycan A2 is a disialylated, bianntenary oligosaccharide model for animal *N*-linked glycans. It is commercially available, and has been used previously to evaluate Martini success in modelling sugars[214]. Biologically, this glycan has been found occurring on mammalian glycoproteins, including serum transferrin and fibrin.

Using the bonding data obtained from the ATMD simulations and my previously defined CG monosaccharides, I built the A2 glycan manually. The glycan was

then simulated first alone in a cubic cell, solvated with Martini water, and the final conformation from this simulation used to populate simulation cells of the same dimensions with 5, 10, 50, and 100 A2 molecules. The cell size and simulation parameters were as for the simulations with HXS, with the exception of the presence of neutralising counterions at a concentration of 150 mM to account for the charged SIA residues (sialic acids) in the glycan.

One of the reasons for choosing A2 for my evaluation of the models was the previous use of this particular glycan in assessing the Martini force field[214]. In this work, the authors noted that using Martini 2 parameters without modification, the glycan aggregated promptly, with all 102 molecules in the 19 x 19 x 19 cm^3 simulation cell aggregating together within a few tens of nanoseconds, with no dissociation events over the entire course of a 1 μs simulation. This problem could be alleviated by modifying the nonbonded interaction potential by a scaling factor, an approach that had previously been applied with success to prevent spurious aggregation of proteins in Martini[215].

It is worth nothing here that aggregation of sugars has also been observed in ATMD simulations, including for Glc, Man, Xyl, sucrose, and the A2 glycan[213, 214, 413]. Some of this could be corrected by changing the water model used, for instance from TIP3P to TIP5P for Glc, Man, and Xyl[413].

In my simulations of Martini 3 A2 glycan, I did not observe spurious aggregation of the glycan, even when 100 glycans were present in a 19 x 19 x 19 cm^3 box for 1 μs. There were certainly interactions between the glycan molecules, as would be expected, but there were both binding and dissociation events observed over the course of all my simulations, as can be seen from the rapid fluctuation in total number of contacts in each simulation (Figure 4.17). The radius of gyration of A2 when simulated alone was on average about 1 nm, comparable to 1.2-1.3 nm found using light scattering methods[214]. This decreased slightly in the higher concentration simulation (50 and 100 A2 molecules), but not drastically, as may take place were the glycan to cluster together excessively.

This is reassuring in light of the previous issues noted with Martini regarding too strong solute-solute interactions as it suggests that my models represent an improvement in at least this respect.

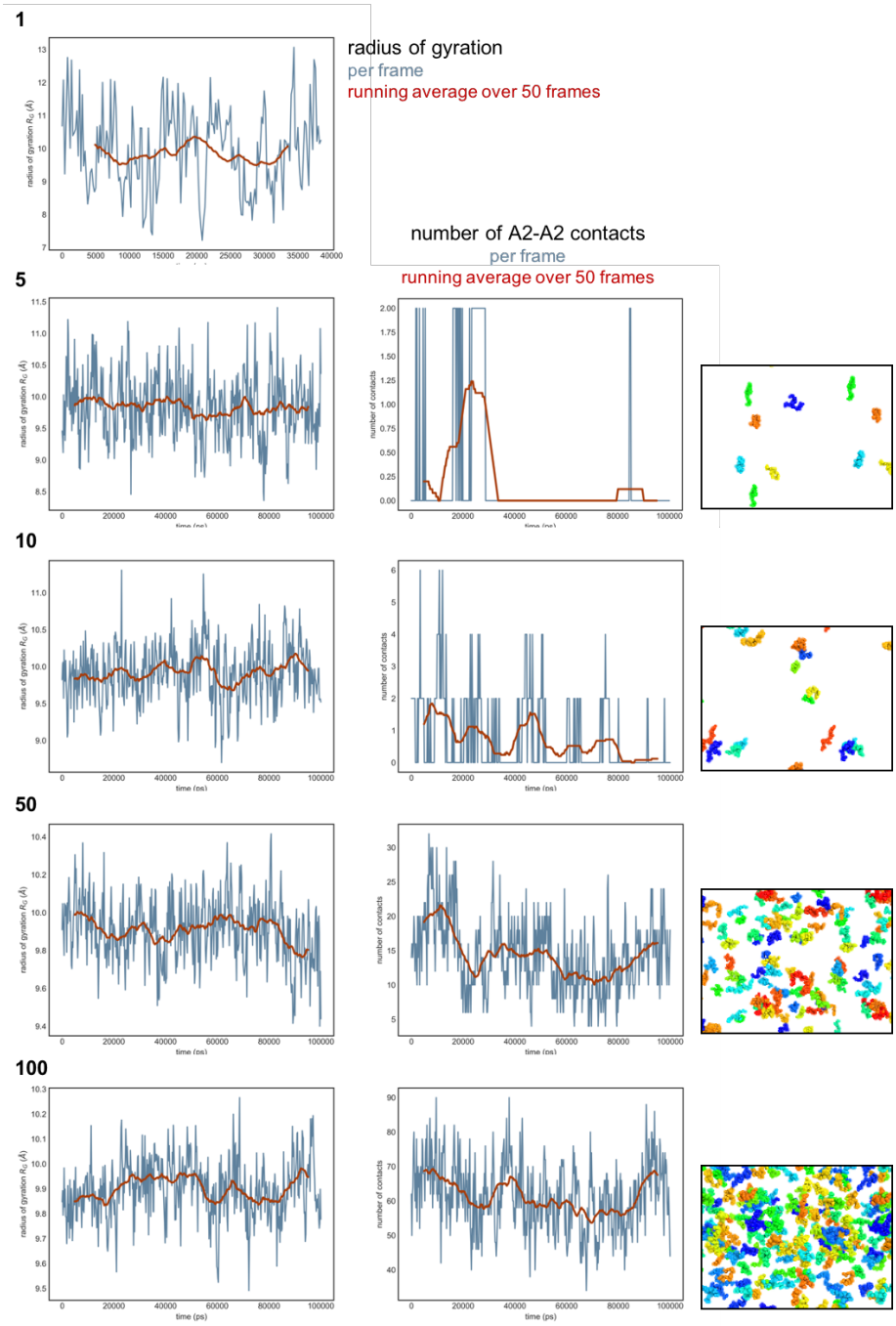


Figure 4.17: Plots of the radius of gyration, averaged over all A2 molecules present, and the total number of contacts between the A2 molecules, for simulations of 1, 5, 10, 50, or 100 and 5, 10, 50, or 100 A2 molecules, respectively. The A2 glycans in the final snapshot are shown in coloured spheres in the images on the right, with the neighbouring simulation cells also shown to illustrate any interactions that may be taking place across the periodic boundaries.

Interaction with lipid bilayers

The ability of sugars to influence membrane properties is well-established, although the underlying mechanistic basis remains somewhat debated. A number of organisms, including members from all kingdoms of life, are known to undergo stages of anhydrobiosis in which the liquid water in the organism is reduced to a level where metabolism is completely and reversibly halted[420, 421]. A common physiological response to this state of dehydration is the synthesis of cryoprotective sugars (e.g. sucrose in plants, trehalose in animals), which have been found to also accumulate within cells under conditions of temperature drift, osmotic shift, and oxidative stress[421, 422].

Indeed, it has been experimentally demonstrated that small glycans can stabilise lipid bilayer structures[423, 424], and that a number of organisms use sugars of one type or another to help them withstand changes in temperature of their environment[425, 426]. Some sugars have been demonstrated to act as particularly effective cryoprotectors, for instance sucrose and trehalose[423, 427].

The precise mechanism underlying the cryo- and anhydro-protection properties of sugars is not established, with proposals including the replacement of lipid-water hydrogen bonds with lipid-sugar hydrogen bonds, entrapping of waters hydrating the lipid headgroups, and vitrification effects[428]. The experimentally observable consequence of the elusive underlying mechanism is a reduction in the main phase transition temperature of the lipid bilayer so as to stabilise the liquid crystalline phase[429]. Sugars have also been shown to influence the properties of hydrated bilayers, for instance altering the spontaneous curvature of liposomes[430], bending moduli and rigidity[431, 432], and melting temperature[433, 434].

The Martini force field has been used in many instances to investigate the phase transitions of lipid bilayers, and so the effect of sugars on such transitions ought to be one means to evaluate the accuracy of my models. Indeed, this was examined in the first iteration of Martini modelling of glycans[162]. I carried out simulations along the same lines as this earlier publication to better allow my models to be compared not only with experimental data, but also earlier models for Martini glycans.

Specifically, I ran CGMD simulations of a 128-lipid (64 lipids per leaflet) DPPC bilayer at a range of temperatures from 200 K to 400 K (200, 270, 325, and 400K), either alone or with HXS added (HXS being the model that incorporates

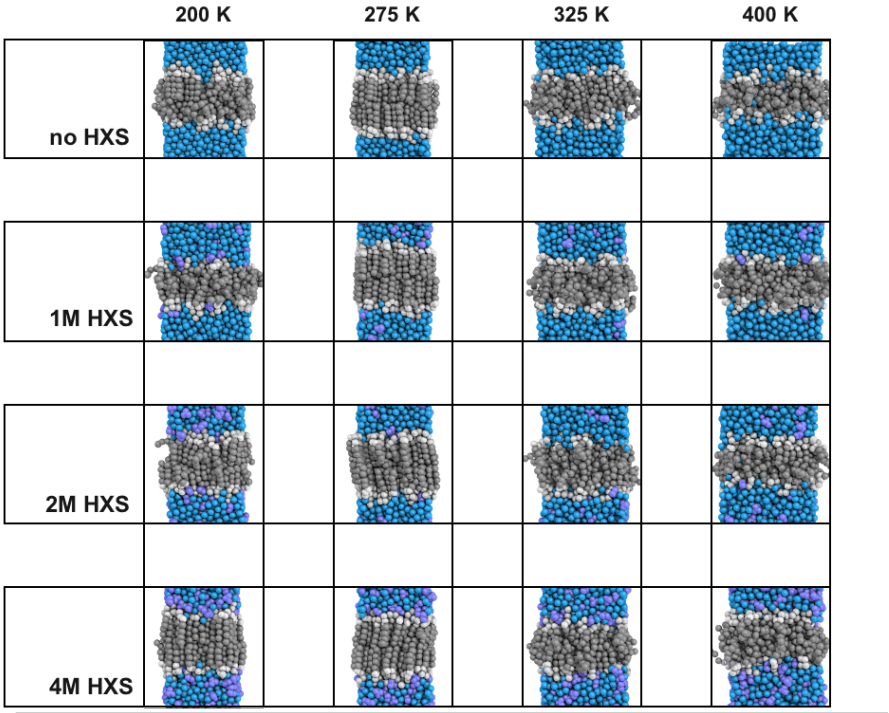
Glc) to a concentration of approximately 1M, 2M, or 4M. Each simulation was run for 1 μs, and the mean area per lipid extracted from the trajectory.

The results are shown in Figure 4.18, along with a visual representation of the bilayer with and without the sugars at a range of temperatures. It is clear from these data that there is some effect of the sugars on the bilayer at low temperatures, below the freezing temperature of Martini DPPC bilayers, but this does not seem to correlate with concentration. This unexpected results could be due to a number of reasons, including an inability of my models to correctly reproduce the properties of sugar-bilayer interactions. However, it could also reflect the small size of the bilayers, the relatively short simulation times compared with the timescales in biological systems, or my choice to use a monosaccharide in this test. Disaccharides such as trehalose are known to be more effective cryoprotectants than monosaccharides, but I chose to use a monosaccharide in my test to allow comparison with the previous Martini carbohydrate model[162]. It is possible that in attempting to avoid spurious aggregation of the sugars in my model, I have eliminated the cryoprotectant activity of monosaccharides that was observed in the earlier iteration. It would be useful to test this notion by repeating these tests using a disaccharide modelled on trehalose.

Phase miscibility

In addition to impacting the overall stability of bilayers, it has been shown using MD and fluorescence confocal microscopy that single monosaccharides (e.g. Glc, Fru), and reducing disaccharides (e.g. palatinose, maltose, gentiobiose) do not affect coexisting liquid ordered (Lo) and disordered (Ld) domains in phase-separated bilayers, whereas nonreducing disaccharides (e.g. trehalose, sucrose) can disrupt the domains and promote lipid mixing[435]. In order to evaluate whether this behaviour is reproduced by my models, I simulated a phase-separated bilayer alone, or with a disaccharide HXS-α1-HXS, modelled on trehalose.

The bilayer was composed of 828 DPPC (saturated), 540 dilauroylphosphatidylcholine (DLPC) (unsaturated), and 576 cholesterol molecules, a mixture that I have previously seen to undergo separation into Lo-like and Ld-like domains. This was first allowed to equilibrate, forming distinct phases, by simulating for 10 μs. Then, 38 molecules of trehalose were added, and the simulation continued for a further 5 μs. The results are shown in Figure 4.19.



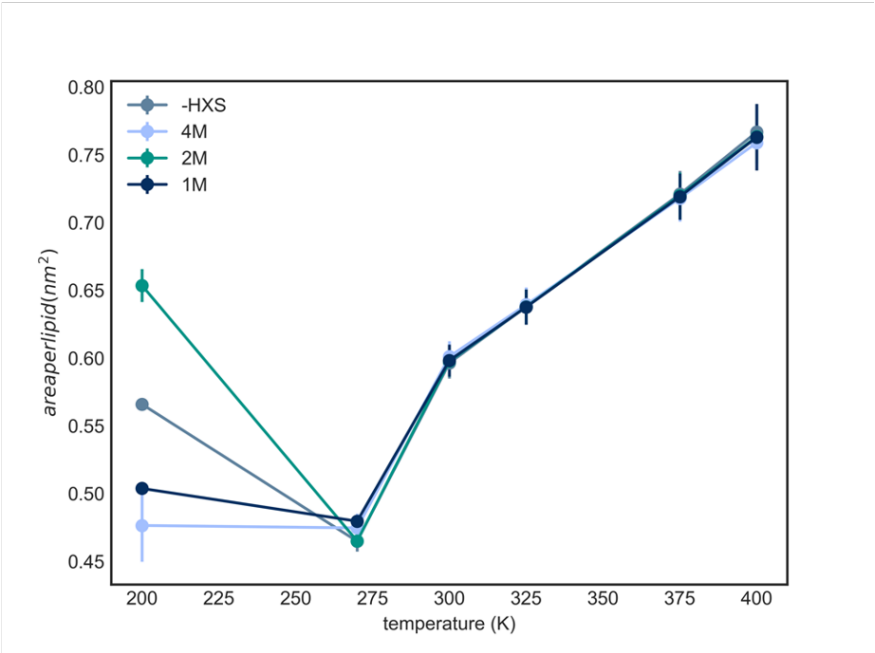


Figure 4.18: Plot of area per lipid against temperature for four systems, containing no sugars and a monosaccharide HXS at concentrations 1M, 2M, and 4M. Above the graph are images of the final frame of each simulation at temperatures 200K, 275K, 325K and 400K, with waters shown in blue, bilayer lipids in grey, and sugar molecules in slate.

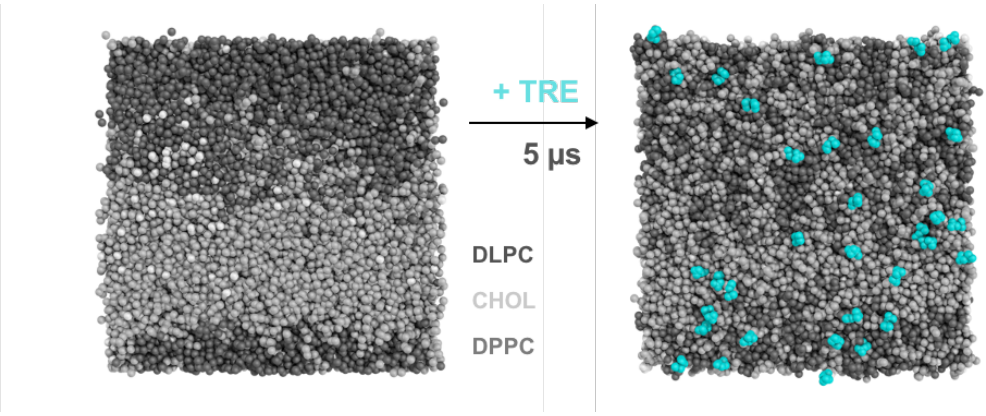


Figure 4.19: Final snapshot of a DPPC/DLPC/cholesterol bilayer after 10 µs simulation (left), and of the same bilayer simulated for a further 5 µs following addition of my trehalose model to the surrounding solution (right). The waters and ions are omitted for clarity, whilst the other components are shown as spheres (light grey and white for DPPC and cholesterol, respectively; dark grey for DLPC; cyan for trehalose.

From visual inspection, it appears that trehalose is, at least to some extent, causing a loss of the bilayer phase separation, with less of a clear distinction between the DPPC-rich (Lo-like) and DLPC-rich (Ld-like) lipid phases. The disaccharide molecules can be seen interacting with the lipid headgroups, and penetrating some depth into the headgroup region, during the simulation, although the interactions are not long-term, with many association and dissociation events with the bilayer taking place over the simulation. This is consistent with experimental and previous MD data on the effect of trehalose on lipid bilayer phases[435].

The extent to which the added sugars may indeed lead to loss of phase segregation could be better examined by extending the simulations for longer, to observe whether or not complete mixing of the lipids takes place over long timescales. It would be interesting to also examine whether or not different types of glycan (e.g. monosaccharide, disaccharide, oligosaccharide) have different impacts on the phases, although I suspect that the level of coarse-graining is such that any such differences may be minimal. In particular, I do not believe that my models would reproduce the differences observed between reducing and nonreducing disaccharides on bilayer phase separation.

Gangliosides

Biological significance

As a more extensive evaluation of my glycan parameters, applied to ‘real’ systems, I used the parameters to construct models for a selection of glycolipids,

focusing in particular on gangliosides as a well studied subset of this group.

Glycolipids are defined by the presence of one or more saccharide units attached covalently through a glycosidic bond to a hydrophobic moiety, a description that encompasses a great variety of possible molecules. Glycoconjugates of this type are found in organisms belonging to all kingdoms of life, and have been suggested to participate in diverse processes, from influencing the local environment of a cell to acting as receptors or ligands for signalling and adhesion processes.

The glycosphingolipids form a diverse subset of glycolipids, defined by the presence of sphingoid or ceramide in the hydrophobic backbone. The group is further categorised into neutral and acidic glycosphingolipids, based on the nature of the carbohydrate moieties, and in particular whether or not they include acidic monosaccharides or derivatives thereof. For example, the lactosides, galactosides, and mucosides are neutral, whilst the gangliosides are acidic due to the presence of at least one sialic acid (N-acetylneuraminic acid or N-glycolylneuraminic acid) residue in the carbohydrate chain.

The gangliosides were first characterised by Ernst Klenk, who isolated them from the human brain in 1942[436], and as of 2004, 188 distinct gangliosides have been identified in vertebrate tissues[437]. Although expressed throughout the tissues and fluids of vertebrates, they are most prominent in the nervous system. Cell surface gangliosides (those in the outer leaflet of the plasma membrane – their most common localisation, to my knowledge) are thought to participate in cell-cell recognition and adhesion, as well as signalling processes, particularly from within microdomains enriched in glycosphingolipids, sphingomyelin, and cholesterol. Additionally, gangliosides have been identified in the nuclear membranes, and proposed to participate in intracellular and intranuclear calcium homeostasis activities.

Gangliosides are primarily synthesised in the ER, with further modification in the Golgi apparatus through the sequential addition of monosaccharide units to the acceptor lipid molecule. These synthetic reactions are catalysed by a series of specific glycosyltransferases, in most cases (the exception being GM4), acting on the precusror lactosylceramide (LacCer), as illustrated in Figure 4.20.

Although ubiquitously distributed throughout the body, gangliosides are most abundant in the nervous system, where they participate in various essential acitivities, including the maintenance and repair of neurones[438], memory

formation[439], synaptic transmission[440], and the development and regeneration of neurones[441, 442]. Their distribution in the nervous system varies considerably during development and between different functional regions, at least in part owing to differential expression of synthetic enzymes. For instance, the main ganglioside in the brains of human and rodent embryos are GM3 and GD3, but as development proceeds, more complex gangliosides such as GM1, GD1a, GD1b, and GT1b become more prevalent, eventually comprising over 90% of the adult brain[443–445].

Models

Rather than attempt to construct models for all the ganglioside species, I have elected to start with a small subset that includes the relatively simple species LacCer (a ganglioside precursor) and GM3, and the more complex species GM1, GD1a, and GT1b. At this stage, my interest is merely in assessing whether or not the gangliosides are stable in lipid bilayers, and compare their behaviours with the corresponding species from Martini version 2.

Each ganglioside model was constructed manually by combining the appropriate Martini 3 lipid group (a ceramide) with the required oligosaccharide chains. Each was simulated for 1 μs in an asymmetric lipid bilayer composed of 120 dioleoylphosphatidylcholine (DOPC) and 20 gangliosides in one leaflet and 144 DOPC in the other. The final arrangement of the bilayer leaflet containing the gangliosides is shown for each system in Figure 4.21 (left-hand side), along with a results from quantitative analyses of the inter-ganglioside contacts (middle) and height above the DOPC headgroups (right-hand side).

This latter evaluation is comparable with the data that might be obtained using atomic force microscopy (AFM), which could be used to evaluate my models. In an AFM study of GM1 in DPPC, a height of 2.5 nm was measured for the GM1 clusters above the surrounding phospholipid[446], as compared with 1.2 nm predicted using X-ray diffraction data[447]. In another AFM study of GM1, a modal height of 0.49 nm above the surrounding DOPC lipids was observed[448]. my GM1 model showed a modal height of about 1.0 nm above the surrounding DOPC, taken from the maximum heights of each ganglioside present above the DOPC choline groups in each frame of the trajectory. AFM of GM3 has revealed a similar variation heights including 2.0 nm[449], and essentially zero as compared with the surrounding phospholipids[450]. my data yielded a modal

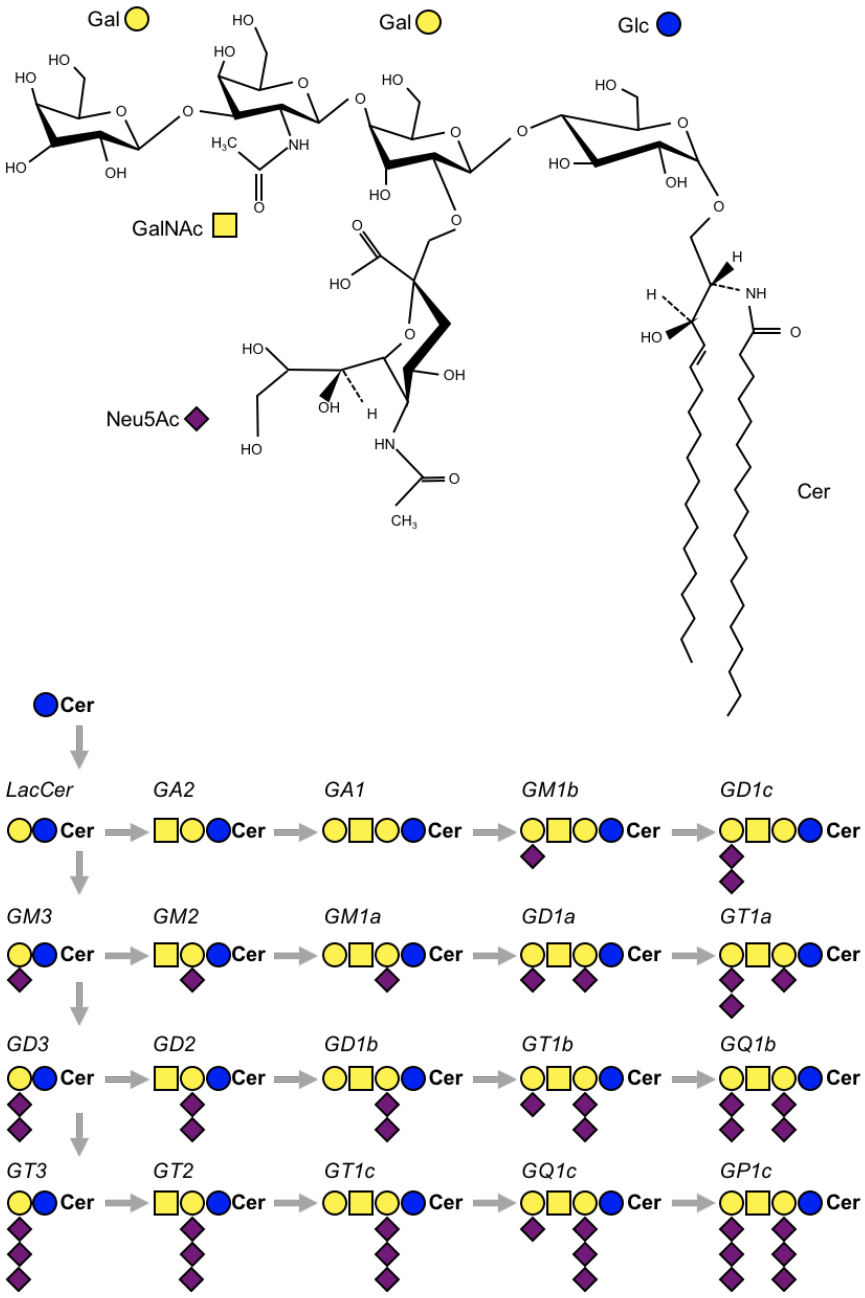


Figure 4.20: Illustration of the synthetic pathways, and molecular relationships, between some of the major brain gangliosides, starting from addition of sialic acid to LacCer to form GM3. The gangliosides may be classified as asialo-, a-, b-, and c-series depending on whether 0, 1, 2, or 3 sialic acids are linked to the inner Gal residue. When sialic acid moieties are linked to the inner GalNAc residue, the α-series gangliosides are produced.

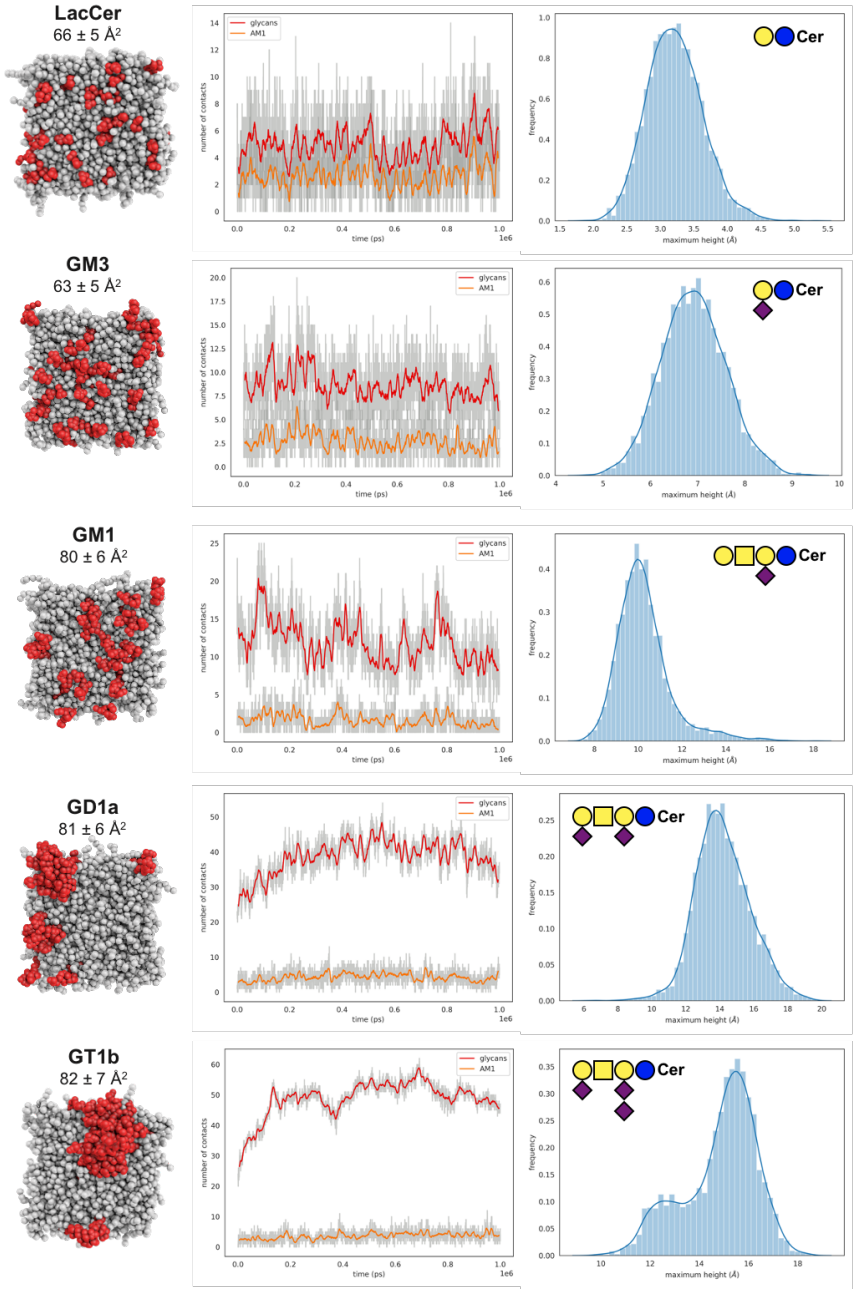


Figure 4.21: Comparison between simulations of different Martini 3 gangliosides, with the SNFG representation of each given on the right-hand side. Left: final (1 µs) snapshot, with gangliosides in red and DOPC in grey, and the mean ± standard deivation of the area per lipid given above. Middle: average number of contacts between ganglioside sugar moieties (red) and the interfacial regions of their ceramide groups (orange), with a contact being defined as any two particles belonging to the sugar/ceramide groups of a pair of gangliosides being within a 6.0 Åcutoff. Right: Distribution of maximum height differences between ganglioside headgroups and DOPC headgroups over the entire simulation. The height difference was taken between the mean height of the DOPC choline (NC3) bead and the ganglioside sugar beads, with the maximum of these differences being recorded at each timestep.

maximum height of approximately 0.7 nm for GM3. The heights showed a considerable range for both GM1 and GM3, on the order of 0.4 nm for GM3 and 0.8 nm for GM1, which may reflect the flexbility of the sugar headgroup. my data therefore yield heights not outside, but to the higher end, of the experimentally observed data. In order to better evaluate my data, it would be useful to do a systematic comparison against all available AFM data for gangliosides, carrying out simulations set up to match as best as I can the experimental procedures. Ideally, I would also have AFM data for all the gangliosides examined to allow a better comparison across them.

One feature that is apparent from my data is that the gangliosides with larger glycan moieties appear to cluster more, predominantly *via* interactions involving the sugars, as may be expected. This is evidenced by the higher number of contacts involving the sugar regions (red lines in the middle plots of Figure 4.21) as opposed to the interfacial part of the ceramide (orange lines in the same plots), and is visually apparent from the final snapshots. GT1b, for example, formed a stable cluster *via* glycan-glycan interactions within about 200 ns, and did not dissociate for the remainder of the simulation time. GM3 and GM1, however, did not all aggregate as one and indeed showed both association and dissociation events during the simulation, as evidenced by the fluctuations in the contacts over time. This is reassuring in light of the previously mentioned concerns regarding Martini overestimating polar interactions.

With limited data available on the physiological behaviour of gangliosides, particularly the more complex ones, it is not known whether or not such aggregation might take place in a biological system, particularly when the membranes in which these gangliosides naturally occur contain a great many additional species, including other lipids, proteins, and components of the extracellular milieu, which could compete for interactions with the ganglioside headgroups. Whether or not such aggregation is observed in more realistic environments will need to be a subject of further investigation.

The area per lipid (mean ± standard deviation given below each ganglioside name in Figure 4.21) also increases with increasing ganglioside complexity to some degree. However, there is a notable increase between GM3 and GM1, but not when going from GM1 to GD1a or GT1b. This could be a result of the clustering of the latter, which could effectively restrict the ganglioside motion, and/or reflect the method by which the area per lipid is evaluated. The

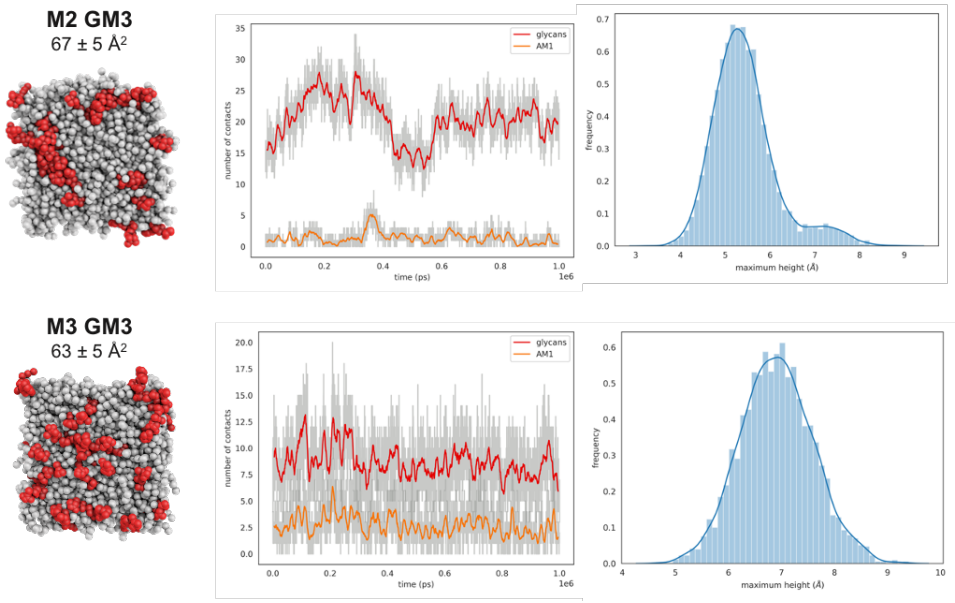


Figure 4.22: Comparison between simulations of Martini 2 (top) and Martini 3 (bottom) simulations. Left: final (1 µs) snapshot, with GM3 in red and DOPC in grey, and the mean ± standard deivation of the area per lipid given above. Middle: average number of contacts between ganglioside sugar moieties (red) and the interfacial regions of their ceramide groups (orange), with a contact being defined as any two particles belonging to the sugar/ceramide groups of a pair of gangliosides being within a 6.0 Åcutoff. Right: Distribution of maximum height differences between ganglioside headgroups and DOPC headgroups over the entire simulation. The height difference was taken between the mean height of the DOPC choline (NC3) bead and the ganglioside sugar beads, with the maximum of these differences being recorded at each timestep.

evaluation considered only the part of the lipid in the bilayer plane i.e. that part of the ganglioside contributing to the bilayer area. For these larger gangliosides, much of the greater area is further from the bilayer plane and therefore may not have such an impact on the bilayer-integral ganglioside area.

I might have expected to see some curvature of the bilayers given the presence of gangliosides in only one leaflet. However, in none of my simulations did I observe more than a small amount of bilayer bending. This may be due to the relatively small size of the bilayers and the periodic boundary conditions, which could suppress bilayer fluctuations. Whether or not this is the case could be examined by repeating these simulations with larger bilayers.

I also repeated the same simulation using Martini 2 GM3, and compared the results with the Martini 3 GM3 (Figure 4.22). These data show that the area per lipid is similar for the Martini 2 and Martini 3 models. The area per lipid of Martini 3 DOPC was found to be approximately 64 Å^2, which is not too dissimilar from the GM3 results. This may be the explanation for the lack of

bilayer bending observed in my systems despite the presence of GM3 in only one leaflet.

These data also suggest that my new model of GM3 clusters less than the Martini 2 model, with both association and disocciation events being observed over the course of the simulation and in general fewer contacts per frame. I also see that the height of the GM3 headgroups above the bilayer is greater for my models than Martini 2 GM3. However, given the lack of consistent data about how GM3 truly behaves in terms of its clustering or height difference from surrounding lipids, I cannot definitively say whether or not these differences are an improvement over the previous model.

4.6 Conclusions and further work

In CGMD, it is necessary to accept that the parameters used will never be perfect. They must simply be good enough to model the properties of interest. One could spend an infinite amount of time trying to develop the perfect model of a biological molecule, when in reality a cruder, imperfect model may well suffice for the questions being put forward.

In this chapter, I have presented preliminary models for monosaccharides and higher order glycans that appear to perform reasonably well the aspects considered. This allows us to further my overall aim of investigating bilayer-bilayer interactions, which often involve glycolipids and glycoproteins, and provides the groundwork for further development of Martini glycan models.

There remain several areas in clear need of improvement, many of which have already been noted. For instance, the set of ATMD disaccharide simulations is not sufficient to cover all potential bonds that I might want to model between monosaccharides, and the monosaccharides themselves are not comprehensive enough to include rarer sugars found in animals or those belonging to other kingdoms of life. Furthermore, my evaluation of the models is somewhat limited, being more of a proof-of-principle and demonstration that the models achieve some of my goals for them. More extensive examination of how the glycans perform in a greater range of situations would be useful, as well as if possible more comparison against experimental data rather than relying largely on ATMD simulations, which are not necessarily accurate or long enough for a reasonable representation of the real systems.

This relates to one of the major challenges I encountered during my work on

parameterising glycans: a general scarcity of data against which I could evaluate my models. Ideally, I may have experimental data that could be directly compared with simulation, but finding this proved difficult, perhaps owing to challenges associated with experiments involving sugars, but also because the simple sugars in which I am particularly interested from a parameterisation standpoint are potentially not very attractive for evaluation using experimental methods.

In many cases, parameterising CG force fields makes use of the 'gold-standard' AT force fields for the corresponding molecular species. However, it has been demonstrated that AT force fields of sugars are neither particularly accurate nor consistent in their inaccuracies. For instance, as noted previously, simulations of Glc, Man, and Xyl using the GLYCAM06 force field with TIP3P water revealed spurious aggregation that was eliminated when the water model was replaced with TIP5P[413, 451]. Similarly, overaggregation was observed in ATMD of glucose and sucrose modelled using GLYCAM06 with TIP3P water and CHARMM36 with TIP3P, that could be alleviated by modification of the Lennard-Jones potential in the case of GLYCAM06[213]. A similar problem with existing AT force fields was clearly demonstrated in the examination of the ganglioside GM1, simulated using CHARMM36, OPLS, GROMOS, and Amber99-GLYCAM-6 force fields[448]. In this study, the performance of the different force fields was variable, not only showing differences between the force fields themselves, but also differences in the degree to which they were in agreement with supporting experimental data.

With limited experimental data and no reliable AT force field parameters to aid my parameterisation and validation, it is difficult to accurately assess how good my glycan models are. However, what I have done is produced a set of modular glycan building blocks that can be used to start to look into the modelling of glycoconjugates such as glycolipids as discussed briefly in this chapter, or glycoproteins, the focus of Chapter 5.

Chapter 5

MODELLING GLYCOPROTEINS

For reference structures, abbreviations, and symbols for the monosaccharides mentioned in this chapter, please refer to Appendix A, and for reference structures and abbreviations of lipids discussed, please refere to Appendix B.

5.1 Introduction

The covalent linkage of carbohydrates to proteins is one of the most common posttranslational modifications observed in cells, being particularly well known as a critical step in the biosynthesis of eukaryotic transmembrane and extracellular proteins. Here, the added glycans serve a quality control and targeting function in the secretory pathway, helping to ensure that proteins are properly folded and, if not, that they are retained in the ER until correct folding or destruction can be achieved. This *N*-linked glycosylation of proteins, that is, the linkage of sugars to Asn side chains, forms an important point in the synthesis of almost all proteins bound for compartments topologically distinct from the cytosol in eukaryotic cells, and as such the vast majority of such proteins in these species will bear at least one covalently-linked glycan. Many glycoproteins, however, possess multiple potential *N*-linked glycosylation sites (PNGs) that may be used to various degrees.

Beyond *N*-linked glycosylation, there are several other classes of protein glycosylation, even if one looks only within the vertebrate clade (Figure 5.1). This variation increases substantially if other eukaryotic lineages, bacteria, and archaea are considered too.

Cell surface glycans influence interactions of a cell with its surrounding milieu. Motifs within the exposed oligosaccharides can serve as ligands for processes such as cell adhesion, pathogen attachment, and binding of external macromolecules[388, 453], whilst the general environment created by a surface covered in glycans can alter the overall nature of the immediate surroundings of a cell, which can influence more indirectly the ability of certain molecules to enter or leave its vicinity. For the glycoprotein itself, attached glycans might affect its functions through various mechanisms, including altering the protein

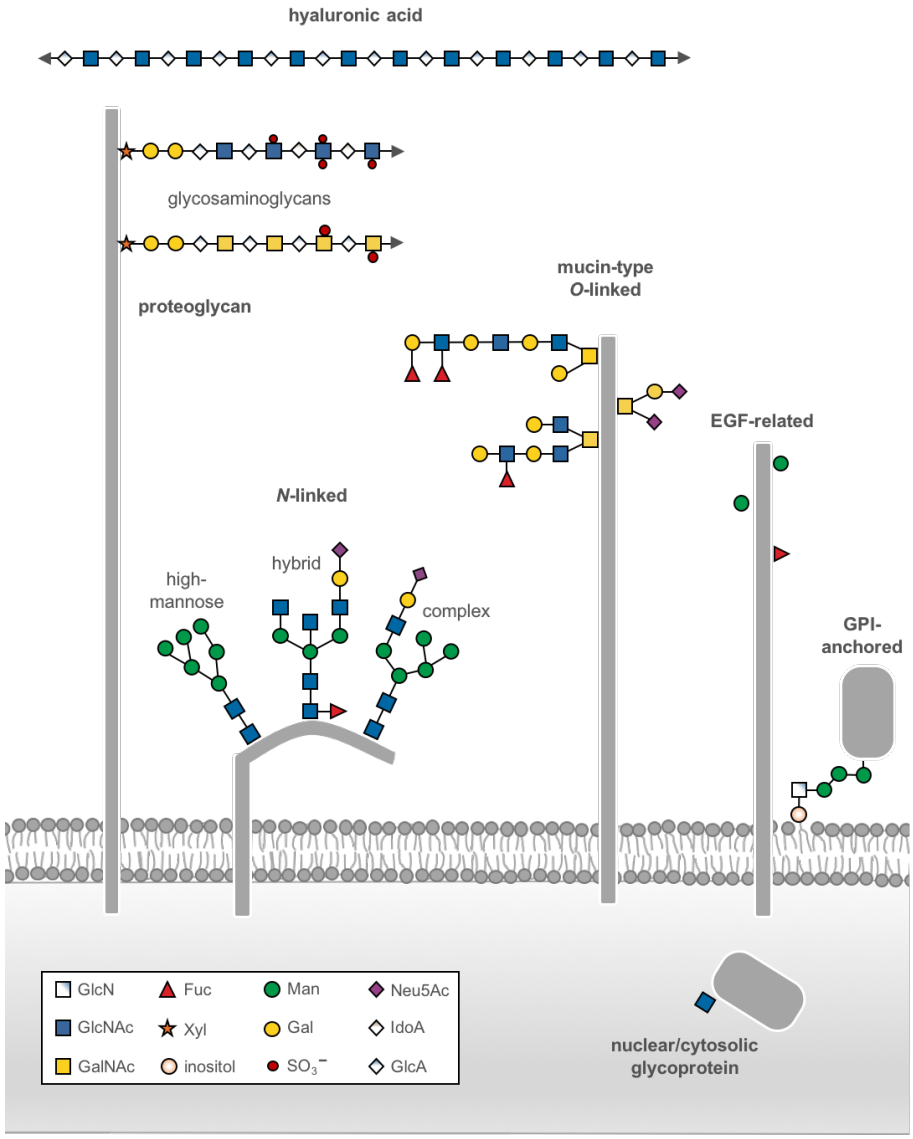


Figure 5.1: Schematic illustration of some different types of glycoprotein found in vertebrates. Glycans can be attached to proteins through e.g. amide linkages with Asn side chains (*N*-glycosylation), glycosidic linkages to Ser/Thr side chains (*O*-glycosylation), and carbon-carbon linkages with the C2 position of Trp (*C*-mannosylation). Proteoglycans are defined by the covalent attachment of long glycosaminoglycan chains of repeating disaccharide units, often *via* a tetrasaccharide core to a Ser hydroxyl. *N*-linked glycans can be classified as complex, hybrid, or oligomannose which differ in their degree of processing. Mucin-type *O*-glycans are characterised by a GalNAc linked to a protein Ser or Thr side chain, which may be extended by addition of further monosaccharides to form oligosaccharides of considerable diversity in length. Other types of *O*-glycan, such as *O*-linked Fuc and *O*-linked Man, tend to be found only on specific proteins or protein domains, such as epidermal growth factor repeats. *O*-GlcNAc is a unique type of *O*-glycosylation that occurs on Ser and Thr residues of nucleocytoplasmic proteins. glycosylphosphatidylinositol (GPI)-anchored glycoproteins are linked at the carboxyl terminus through a phosphodiester linkage to phosphoethanolamine attached to an oligosaccharide core; the GlcN residue is linked to phosphatidylinositol, which is embedded in the cell membrane[452].

cell surface half-life, affecting its solubility, and impeding or enhancing intermolecular interactions[454]. Glycans also play a role in glycoprotein synthesis, influencing protein folding, providing ligands for chaperones, and contributing to ER quality control, transit through the secretory pathway, and selective targeting therewithin[455, 456]. Glycans are not restricted to eukaryotic cells, but are additionally found on the surfaces of many bacteria and viruses, and for pathogenic species these can serve important roles in immune evasion and host cell attachment.

Given this preponderance of cell surface glycoproteins, if one hopes to model the interactions between cell surface membranes and/or viral envelopes, it is important that methods are in place to glycosylate the involved proteins. In this chapter, I discuss the method I have developed for constructing glycoproteins for use in MD simulations.

5.2 Computational glycosylation method

Intent

It is my aim to produce a general methodology that can be used to construct models of various types of glycoproteins for use in MD simulations. The method should be flexible, in that the principles behind it should be applicable to any simulation force field with sugar parameters, and extensive, such that it can carry out many different types of glycosylation without the user having to design and specify the exact nature of each glycan added. Additionally, it should be constructed in such a way that it can be improved, modified, and added to through extension of the code rather than a complete reformulation.

However, at this stage, it has been possible only to take steps towards achieving only some of these goals. In particular, the code I have developed thus far works only for Martini glycoproteins, with no other force fields having been addressed. The range of types of glycosylation is also relatively limited, especially when it comes to considering bacterial or archaeal glycans, and there is a strong bias towards vertebrate glycoproteins that will need to be addressed to allow more accurate modelling of the glycans belonging to other clades. However, I have built in some initial means for modelling plant, fungal, and invertebrate glycans, as well as covering several of the better studied classes of glycoproteins thought to be present in most, if not all, eukaryotic lineages.

In the subsequent sections, I first provide an overview of the code as a whole.

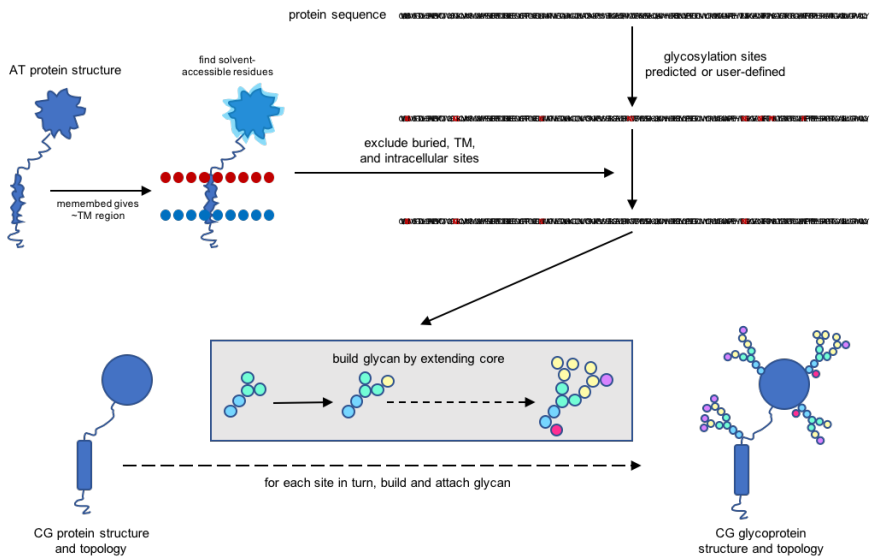


Figure 5.2: Overview of my glycosylation method. The user provides a protein structure and topology, and information regarding the type of glycosylation desired (*N*-linked, mucin-type *O*-linked, glycosaminoglycan, or monosaccharide), as well as indicating the residue numbers of glycosylation sites (if known), and whether or not the protein is membrane-embedded. The code then uses this information to construct a glycan for each site to be modified, using knowledge-based rules and a degree of randomisation to achieve physiologically plausible and heterogeneous glycan structures. Because the glycan-construction component is managed by a separate code from the glycan attachment, it is also possible to use the same tools to construct free glycans.

I then discuss in more detail the different types of glycan it can be applied to, finishing with example applications to several different types of glycoprotein. The glycosylation code is written in Python, making use of the MDAnalysis[363] and NumPy[362] modules.

Overview

A general outline of the process is illustrated in Figure 5.2.

The required and optional user-specifications are as follows:

- *-p* prefix of the topology and coordinate files of the protein [default: Protein]
- *-t* type of glycosylation; N (*N-linked*), O (mucin-type *O*-linked), mono (monosaccharide), GAG (glycosaminoglycan)
- *-s* sites to be modified; may be omitted for *N*-linked (code then predicts sites)
- *-o* name for output files [default: glycoprotein]

- *-nt* type of *N*-linked glycan; complex, hybrid, or oligomannose
- *–bilayer* is the protein transmembrane?
- *–noncanonical* include noncanonical sequons in finding *N*-linked sites?
- *–glycosaminoglycan* type of glycosaminoglycan if type is *GAG*
- *–mono* identity of monosaccharide if type is *mono*
- *–plant* does the glycoprotein belong to a plant?
- *–invertebrate* does the glycoprotein belong to an invertebrate?

For each site, whether provided (using flag -s) or identified by finding glycosylation motifs, a check is first carried out to establish whether or not that site is too sterically encumbered to accommodate a glycan or, for transmembrane proteins (flag –bilayer), whether it lies within a predicted transmembrane domain or intracellular region. In either such case, the site will be removed from the list of potential glycosylation sites.

The glycan is then built according to the user-provided specifications, as discussed in detail in Section 5.3. It is then translated to the glycosylation site and attached – that is, a bond is defined between it and the protein in the topology file, and the glycan coordinates are added to those of the protein in the structure file at an appropriate location relative to the modified side chain. It is possible to produce a single glycoprotein topology file or include the protein and glycans separately – I have not yet been able to determine if either holds particular advantages over the other, and so at present can recommend only based on personal preference and utility for a particular case.

Once the process is complete for every glycosylation site, the resulting glycoprotein will need equilibration to allow the newly constructed glycans to settle, as the construction is based on geometric translations that will differ from the favoured conformations of the molecules. This equilibration may need to be done in stages with gradually increasing timesteps, particularly for proteins with more and/or longer glycans. However, once this is done, the glycoprotein ought to be either ready for subsequent simulations, or for addition of more glycans of a different type. There ought to be no major steric clashes to be alleviated, but if some do occur they can be reduced with a quick run of Alchembed[289], in

which the protein interactions with the glycans are turned on gradually over the course of a short simulation. This allows the glycans to move out of the way if they are initially overlapping with any regions of the protein.

In the subsequent sections, I discuss in more detail certain elements of the glycosylation code.

Determining glycosylation sites

Predicting potential glycan sites

For some types of protein glycosylation, most notably eukaryotic *N*-linked glycosylation, potential sites of modification can be reliably predicted from a protein sequence alone. PNGs are defined by the consensus glycosylation motif N-x-S/T (X denotes any amino acid except Pro). Provided this sequon is exposed to the enzymatic machinery responsible for *en bloc* transfer of the precursor glycan in the ER lumen, it can in theory be subject to modification, although whether or not it is indeed modified may be influenced by additional factors such as conformational constraints arising as the nascent protein folds or altered efficiency of the reaction governed by the nature of the 'X' residue. The only way to be sure of whether or not one of the PNGs is used is through experimental detection[388]. Therefore, unless the user specifies the sites to be used, all PNGs will be found from the protein sequence and added to the list of potential sites for modification. There is some evidence for the existence of additional sequons that may also define *N*-linked glycosylation sites, such as N-x-C[388]. If the user specifies that noncanonical sequons may be used (flag –noncanonical), then the code will search also for N-x-C motifs.

Removing transmembrane/intracellular sites

If the protein being glycosylated is a transmembrane protein, as indicated by the user (flag –bilayer), then a check is carried out for any residues located within membrane-spanning or cytosolic regions of the protein. Any such residues identified are removed from the list of possible glycosites. In the present version of the code, the check is carried out using TMHMM2.0[457, 458], which uses a hidden Markov method to predict transmembrane (TM) helices from protein sequences. Example results of employing this method are illustrated in Figure 5.3, comparing the predictions against the corresponding proteins in the Orientation of Proteins in Membranes (OPM) database[459]. This database predicts the

spatial arrangement of membrane proteins relative to the lipid bilayer based on minimising the estimated energy of transfer of the protein structure from water into the hydrophobic bilayer.

In the future, the incorporation of additional topology prediction tools, and the taking of a consensus from these, may help to ensure a more accurate prediction. It should be noted that this method would not be applicable to beta barrel transmembrane proteins such as found in bacterial outer membranes.

Surface accessibility

If any of the residues in the glycosite list are found to be solvent inaccessible, they are removed from the list. I first attempted to do this using the GROMACS gmx sasa tool and a probe radius of 0.26 nm, corresponding to an approximate radius of Martini 3 water. However, I found that this approach would not take notice of e.g. lipid bilayers or glycans already present. I therefore decided to do a slower, but more robust method. Specifically, I solvate the protein structure and identify those residues that are in contact with solvent particles, defining a contact as the particles being within 0.5 nm. This works both with fully unglycosylated proteins, and with ones that are part-way through their glycosylation. The latter is important as it allows instances where an existing glycan would impede attachment of another to prevent this later modification from taking place. In such cases, where two or more glycosites potentially interfere with each other, it would be desirable to have a way of randomly using one site or the other according to some biological scheme, but this has not yet been implemented. An additional problem with this approach is that it does not reflect the manner in which glycosylation occurs biologically – it acts on the fully folded structure of a protein, and fully assembled glycans where present, rather than a nascent form of the protein and the initial glycan precursors. This may introduce a risk of excluding glycosites that would have been used in a real situation if those sites are somewhat occluded in the crystal structure, but less so during ongoing translation.

5.3 Glycan construction

Incorporating heterogeneity

The glycan structures found linked to proteins can be extremely complex, with far greater potential structural diversity than nucleic acids or polypeptides[473].

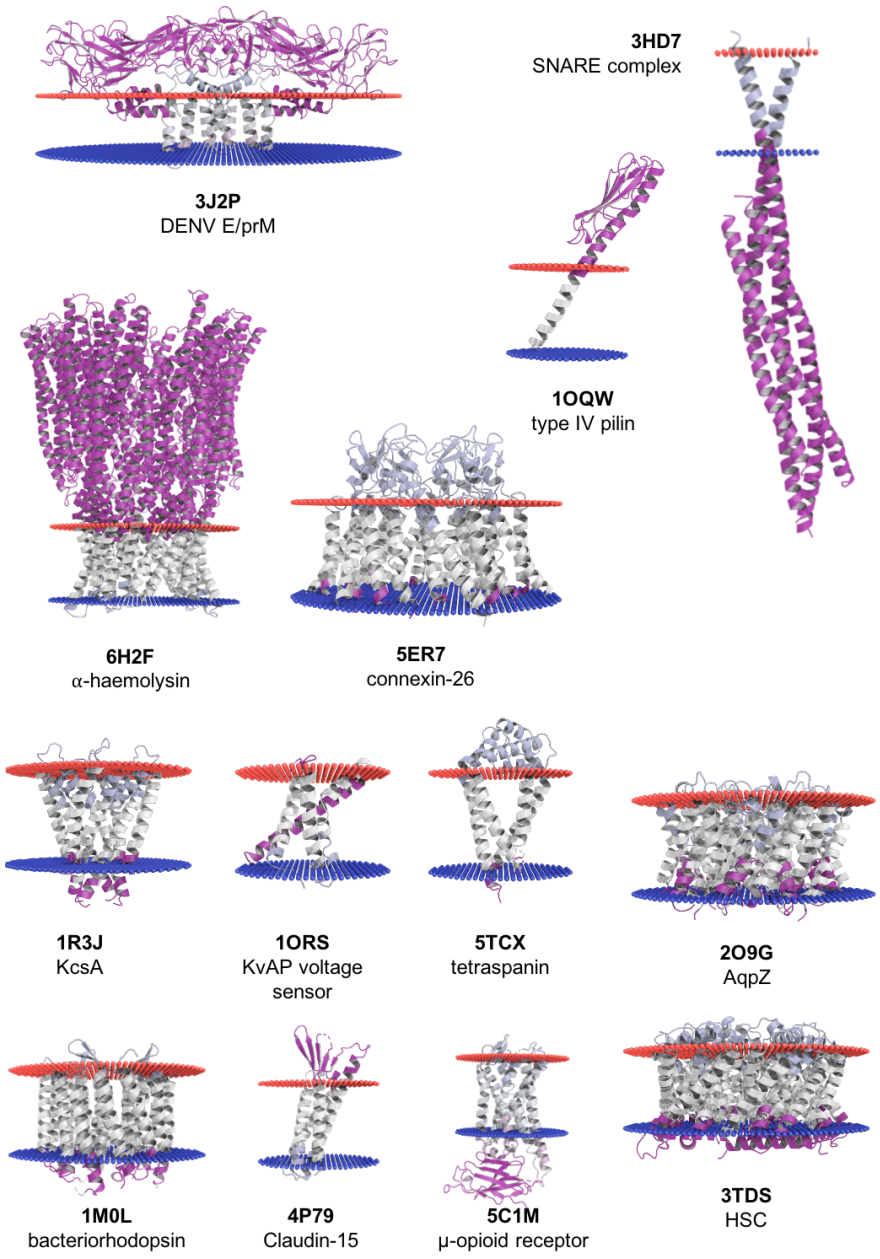


Figure 5.3: Topology predictions were carried out for 35 different proteins, belonging to 12 distinct superfamilies. One example from each superfamily is shown (with the exception of the Kv voltage sensor, which is shown along with KcsA). The cartoon representations of the proteins are coloured according to their predicted topological location, with grey for transmembrane, purple for extracellular, and light blue for intracellular. The red and blue spheres indicate the OPM-predicted positions of the bilayer headgroup region for the same proteins, with red for the outer leaflet and blue for the cytoplasmic leaflet. The proteins are: 3J2P[460], 6H2F[461], 5ER7[462], 1R3J[463], 1M0L[464], 1ORS[465], 4P79[466], 5TCX[467], 5C1M[468], 2O9G[469], 3TDS[470], 1OQW[471], and 3HD7[472].

One contribution to the diversity of many classes of glycan is the mechanism of their synthesis. The biosynthetic pathway for several types of protein-linked glycan involves an intricate interplay of competing and compartmentalised enzymatic activities[474, 475]. Combined with the relative inefficiency of the enzymatic reactions involved, this competition and localisation of activities often yields diverse ensembles of glycan structures (microheterogeneity), even at the same glycosite of otherwise identical glycoproteins[476].

The diversity of glycan structure can produce pleiotropy. The particular properties of the glycosylation on a given glycosite may alter the interactions within a specific cellular context, but in other situations may have little functional consequences or cause distinct effects. The diversity of glycan structures found on a cell or tissue sample present considerable challenges to structural characterisation[477].

Within mammals, it has been estimated that approximately 700 different proteins (including about 200 glycosyltransferases) are required to generate the full detected glycan diversity, which is believed to include at least 7,000 different structures[473]. These diverse glycans are produced from only ten monosaccharides: Fuc, Gal, Glc, GalNAc, GlcNAc, GlcA, IdoA, Man, sialic acid, and Xyl, combined in different configurations and subject to various modifications[473, 474].

It is not possible to simulate or in any way truly replicate the factors that give rise to glycan variability. Nor can I develop accurate mathematical models that determine the most likely monosaccharide additions at every stage in the biosynthetic process as (i) this is likely to vary depending on the cell and its current state, and (ii) the data required to build such a model are simply not, to my knowledge, available. I instead take an approach based on estimating a conditional probability for each monosaccharide addition, determined by combining biological knowledge with experimental glycan array data.

The conditional probability (Equation 5.1) is the probability of adding a given monosaccharide through a given bond type to the monosaccharide present at the nonreducing end of the chain (i.e. the most recently added monosaccharide). These probabilities are derived from the GlyTouCan repository of experimentally

identified glycan structures[419].

$$P(X|Y_z) = \frac{P(X \cap Y_z)}{P(X)} \tag{5.1}$$

where $P(X|Y_z)$ is the probability of monosaccharide Y being linked through bond z when monosaccharide X is at the nonreducing end of the chain; $P(X \cap Y_z)$ is the probability of monosaccharide X being linked to Y through bond z; and $P(X)$ is the probability of X occurring at all.

Whenever this is used, the code generates a random number between 1 and 0, and if it is below the conditional probability derived from the experimental repository, the addition is accepted. If not, it is rejected and another iteration of possible additions continues. If none of the possible additions are accepted, the glycan is deemed complete.

It should be noted that the premise behind this approach is not that the repository has a distribution of glycans similar to that expected on a real cell, and indeed it is unlikely to do so, but rather that all of the glycans included in it are ones that have been observed in biological systems, and certain pairs of monosaccharides are seen more often among these than others various glycan structures. Ideally I would have some way of ensuring the glycan structures I use is truly representative of the cell type in question (e.g. by giving greater weight to pairs from motifs that the cell surface glycans are known to possess), but the present approach does at least function to allow us to incorporate some randomness into my models without risking wildly non-physiological glycans being generated. Thus, in short, the approach might capture those motifs that arise frequently within glycans but may not represent the true frequency with which they occur in nature.

In the context of the glycosylation code, this probability is used within the confines of biology-based rules that ensure, for instance, that if a capping residue or bond (one that typically terminates a growing glycan chain) is encountered, no further monosaccharide additions are made, or that Xyl is not included in vertebrate *N*-linked glycans[388].

N-linked glycans

Biology

N-linked glycans in eukaryotes are among the most extensively studied form of protein glycosylation. Across the lineage, all such glycans appear to share a common core sequence of Manα1–3(Manα1–6)Manβ1–4GlcNAcβGlcNAcβ1–Asn, where the final bond is a β 1-glycosidic linkage to the nitrogen atom of the amino group of Asn, generally within the N-x-S/T consensus motif. The process of glycan synthesis begins with *en bloc* transfer of a precursor oligosaccharide to the Asn side chain, and proceeds by the progressive removal and addition of monosaccharides from/to this core, events which take place in the ER and Golgi compartments (Figure 5.4).

These processes can ultimately yield glycans defined as being complex, hybrid, or oligomannose depending on how far along the maturation pathway they progress. *N*-linked glycans that have undergone minimal processing with mannosidases are called oligomannose glycans due to the presence of only Man residues beyond the core structure. Addition of GlcNAc to the α 3 arm of an acceptor containing five Man residues produces a hybrid glycan[475, 478], which may be extended on this arm with Gal, GalNAc, Fuc, and sialic acid. Further processing of hybrid glycans produces complex glycans, which can be similarly modified with Gal, GalNAc, Fuc, and sialic acid. Within each of these general categories of glycan, there is considerable possible variation resulting from e.g. differences in the availability of activated sugar donors and enzymes in each compartment through which the glycoprotein transits as it matures.

Thus, for any given glycoprotein, it is possible in most cases only to say that a particular glycosite is predominantly occupied by a complex, hybrid, or oligomannose type glycan[479–483]. One can very rarely specify that it will always be complex/hybrid/oligomannose, or the exact nature of the glycan that it possesses.

It should be noted that the classifcation of *N*-linked glycans into hybrid, complex, and oligomannose is perhaps not applicable to non-vertebrate eukaryotes[484]. For instance, Planaria have been found to have glycans in which the $Man_3GlcNAc_2$ core is modified with an extended oligosaccharide chain linked to a Fuc branching from the first core GlcNAc[484]. It is important that a further iteration of the glycosylation method I present here includes these possible variations in glycan structure to allow modelling of glycans belonging to eukaryotes outside the vertebrate lineage.

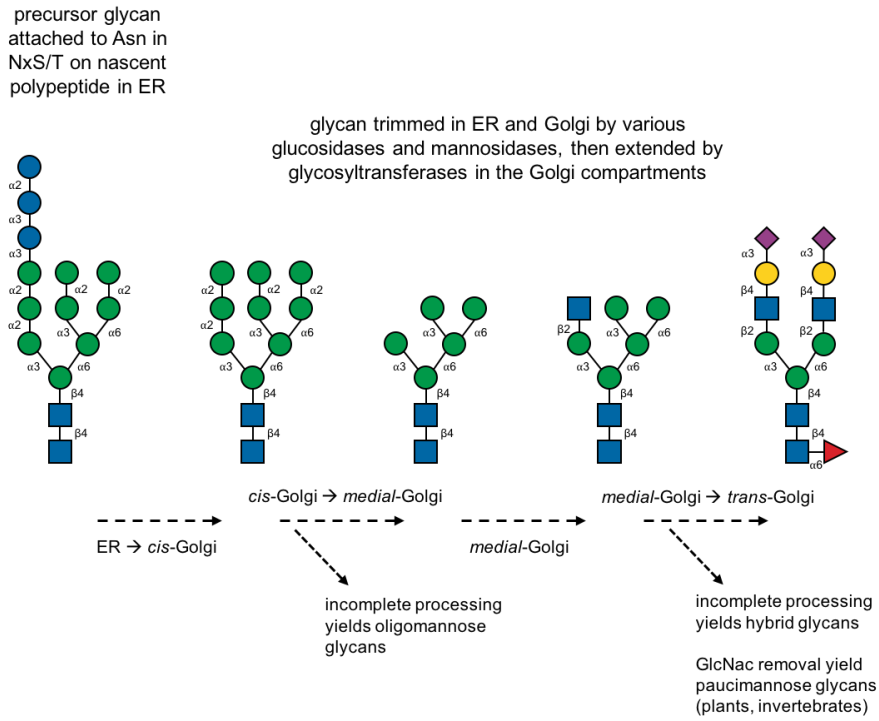


Figure 5.4: Simplified outline of the stages in eukaryotic synthesis of *N*-linked glycans, highlighting the stages at which the glycan may deviate into formation of oligomannose or hybrid, rather than complex, structures. The precursor oligosaccharide $Glc_3Man_9GlcNAc_2$ that is transferred from a dolichol-phosphate lipid precurosor to the Asn side chain is trimmed in the ER before the glycoprotein can proceed to the Golgi. The removal and reattachment of the Glc residues plays an important role in protein quality control, which is not discussed here but may be studied in detail in any standard introduction to glycobiology[388, 485]. The glycan, on transit from the ER to the Golgi, will usually contain eight or nine Man residues, and is then further trimmed to yield the $Man_5GlcNAc_2$ intermediate. If this trimming does not occur, the glycan may be released as an oligomannose type. Processing of the intermediate in the *medial*-Golgi is initiated by the addition of β1–2 GlcNAc to the α1–3Man branch in the core unit, which allows removal of the terminal Man residues from the other branch to form $GlcNAcMan_3GlcNAc_2$, and is generally followed by similar attachment of GlcNAc to the α1–6Man arm of the glycan. If this does not occur, it may lead to the formation of a hybrid glycan, but if both branches are processed to yield terminal GlcNAc residues, they may be further extended to yield a variety of complex glycans. Additional branches may be initiated at C4 of the core α3 Man, the C6 of the core α6 Man or, at least in birds and fish, at the C4 of the core α6 Man. Complex and hybrid glycans can also carry a bisecting GlcNAc linked to the βMan of the core. The extension of hybrid and complex glycan precursors occurs through further additions to the core structure, elongation of the branches, and attachment of capping residues, such as α-linked sialic acid, Fuc, Gal, GlcNAc, or sulphate. The α linkage in a branch of β linkages means that the capping residue protrudes from the branch, therefore being more readily presented for interaction with e.g. lectins or antibodies.

Code

The construction of *N*-linked glycans begins with the user-provided choice of glycan type (complex, hybrid, oligomannose). It would be a nice addition to be able to predict the suitable *N*-glycan type, if the data become available that will allow us to reliably do so. For instance, if some correlation between the local environment of a PNG and the type of glycan attached were identified. This would require considerably more glycoproteins to be examined in a site-specific manner, as has already been done for a number of viral glycoproteins[479–483].

However, at present, I require the glycan type to be provided. If the type is oligomannose, a random decision is taken between ten different possible oligomannose structures, with equal probability of choosing any of these. If the type is hybrid or complex, the starting core is defined as either possessing one terminating GlcNAc (hybrid, on the α3 arm) or two (complex), and each of these extended by further monosaccharide additions.

The identities and bonds of these extensions are governed by a combination of biology-based restrictions and empirical conditional probability, and the extension continues until a capping residue is added[388]. Specifically, if sialic acid, Fuc, Gal, or GlcNAc, are added *via* an α linkage, the branch is terminated, and unless the user specifies the glycoprotein belongs to a plant, Xyl is not allowed in the extensions.

I have not yet built in the means to modify the core (e.g. through fucosylation of the GlcNAc residues) or construct more than two branches, and these will be important requisite improvements in later developments of the code. For example, in later iterations, I can add further rules such as for plants allowing α3-linked Fuc on the first GlcNAc and β2 Xyl on the first core Man, or for invertebrates allowing α3 or α6-linked Fuc on both GlcNAcs, as opposed to vertebrates where only α6-linked Fuc is found on the first core GlcNAc[388, 484]. Such additions will futher allow us to incorporate species-specific properties into the glycans.

Example output glycans using this approach are shown in Figure 5.5 using the SNFG nomenclature[414].

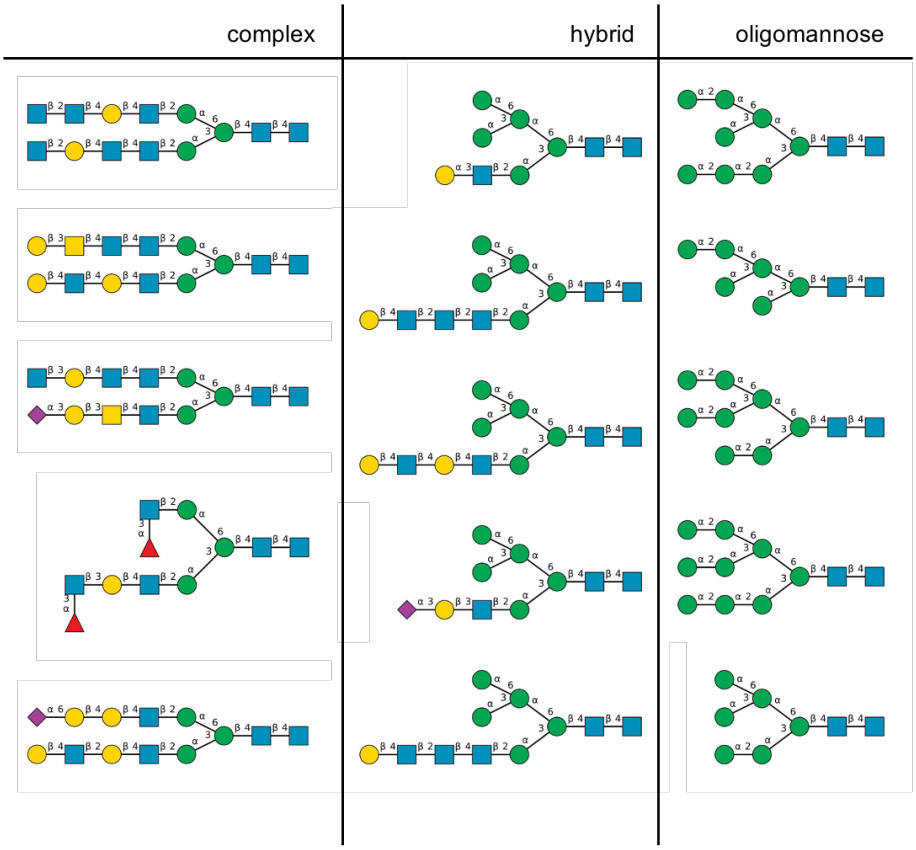


Figure 5.5: SNFG illustrations of *N*-linked glycans produced using my approach, showing the output from five independent repeats for each glycan class, complex, hybrid, or oligomannose.

Mucin-type *O*-linked

Biology

There are a range of distinct glycosylation modes that can be classed as *O*-linked, sharing the common theme of glycosidic linkage between a glycan and a hydroxyl moiety, usually that of a Ser or Thr side chain. In many cases, the monosaccharide immediately attached to the amino acid is GlcNAc or GalNac[486]. GalNAc-linked glycans, also called mucin-type glycans, are abundantly present on the surface of many extracellular proteins/protein domains, including their namesake mucins[487].

The GalNAc in mucin-type glycans may be present on its own, or it may be extended with Gal, GlcNAc, or GalNAc to produce eight different core structures[488] (Figure 5.6), each of which can be subject to further additions to form a variety of linear and branched chains. The resulting oligosaccharides can include potentially over 20 residues, and typically include GalNAc, Gal, GlcNAc, Fuc, and sialic acid[388]. The sialic acid residues may be modified

by *O*-acetylation, whilst the Gal and GlcNAc residues can be sulphated, therefore further increasing the variability and range of possible properties of these chains[388]. The terminating residues may be Fuc or sialic acid in α linkages, sulphate, or Gal, GalNAc, and GlcNAc in either α or β linkages.

Synthesis of mucin-type glyans is initiated by polypeptide-GalNAc transferases, which differ in their specificity for amino acid motifs, adding a degree of regulation in the attachment of *O*-glycans. However, these enzymes appear in many cases to be relatively promiscuous. Subsequent non-templated addition of monosaccharides such as Gal, GlcNAc, sialic acid, and Fuc to the initial GalNAc results in the formation of a diverse set of oligosaccharide structures[388].

Other *O*-linked glycans are classified on the basis of their initiating monosaccharide. Specific Cys-rich protein domains may include two types of *O*-linked Fuc and one type of *O*-linked Glc glycan, in addition to *C*-Man residues on Trp side chains[388, 489, 490]. Man is the initiating residue of the *O*-Man glycans, some of which may be extended by a disaccharide repeat unit[491]. In addition to attachment to Ser and Thr side chains, the hydroxyl groups of hydroxylysine in collagen can be modified with Gal, which may be extended by the addition of an α1,2-Glc residue[492], while Glc may be linked to a Tyr side chain of glycogenin, and this extended and branched with additional Glc to form glycogen[493]. Other than those types that involve only addition of a monosaccharide, these other forms of *O*-linked glycosylation are not currently implemented in my code.

Code

Mucin-type *O*-linked glycosylation has no known consensus sequon defining where it takes place. Therefore, if this is the desired glycosylation type, the user must provide a list of residue numbers to be glycosylated, all of which must correspond to Ser or Thr (which the code checks as a safety measure, terminating with an error message should any be found not to conform to this requirement). For each site to be modified, the glycan that is to be attached is built starting from a randomly chosen core (with a bias towards cores 1 and 2, as these are the most commoly observed core types), which is then extended in a manner similar to that described for the *N*-linked glycans.

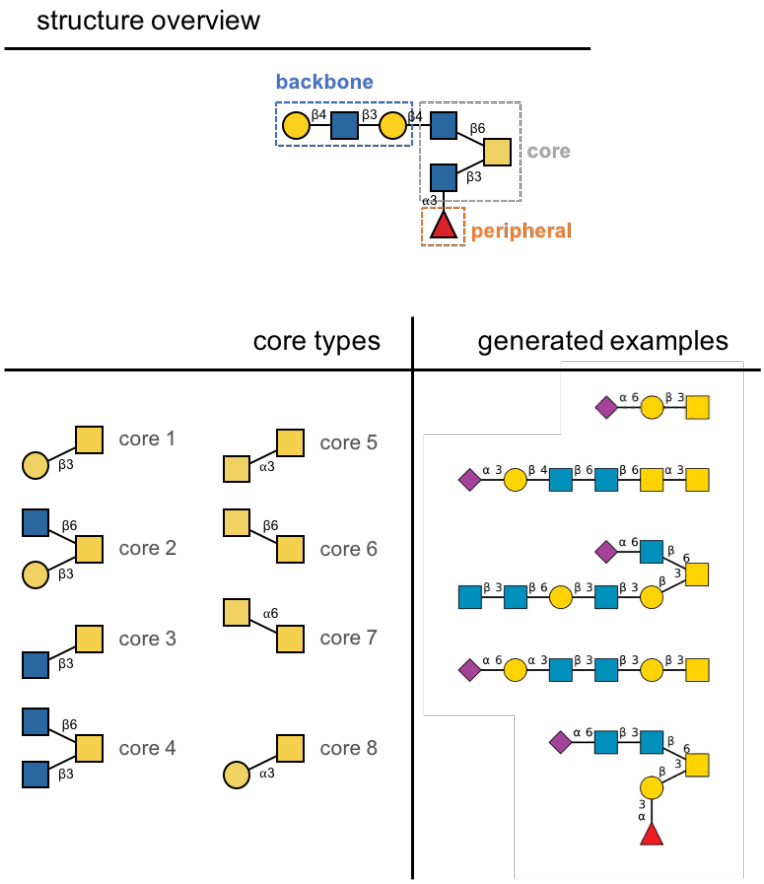


Figure 5.6: Overview of the general structure of mucin-type glycans, the eight core structures, and example glycans generated using my code. In the latter case, the output from five independent runs of the code are shown. All glycans are shown using the SNFG representation.

Glycosaminoglycans

Biology

Glycosaminoglycans (GAGs) are linear polysaccharides containing, in addition to a core region, a repeating disaccharide motif composed of an amino monosaccharide (GlcN, GlcNAc) and a uronic acid (GlcA, IdoA) or Gal. These glycan chains can contain over 80 monosaccharides, in stark contrast to the typical 5-12 residues observed in *N*-linked glycans[388].

In many cases, the GAGs are linked to a core protein to form a proteoglycan, which may possess one or a great many such glycans, often in addition to mucin-type *O*-linked and *N*-linked glycans. Relatively few different proteins have been identified that carry GAGs, but within these there is considerable structural variation[388]. For instance, even in cases where many possible sites are present at which GAGs might be attached, not all may be used in a single protein, and

some sites may be used preferentially to others. A given proteoglycan may bear one or two different types of GAG, and each of these can vary in length and sulphation of the chains, both within and between cell types. As a result, any given proteoglycan, as defined by the core protein, in fact represents a range of structurally distinct molecules.

The commonly observed GAGs are hyaluronic acid (HA), dermatan sulphate (DS), keratan sulphate (KS), chondroitin sulphate (CS), and heparan sulphate (HS). Among these, HA is not found covalently linked to proteins, but rather associates with HA-binding motifs.

Code

As with mucin-type glycans, the user must define the sites at which any GAG chains are to be added, as well as the type of GAG.

GAG chains may be modified by sulphation and epimerisation (GlcA conversion to IdoA), therefore producing considerable heterogeneity from the disaccharide repeats[494]. Thus, the major points of variability that need to be considered in my code are in the total length of the oligosaccharide and the sulphation of the component monosaccharides. Without additional knowledge to govern my choices, I assign an equal probability to every biologically possible degree/ position of sulphation of the monosaccharides in the chain, and randomly choose a length in the range of 20 to 100 residues. I note, however, that very long chains would require very large simulation cells, and may therefore be practically challenging. While it is possible to produce an effectively 'infinite' glycan chain by having it continuous across the periodic boundaries, and I have succesfully applied this to HA (data not shown), such a thing would not be possible when one end of the GAG is linked to a protein.

Monosaccharide

Biology

There are some instances of glycosylation that involve the attachment of only a single sugar residue that is not extended further. For example, the attachment of Man to Trp side chains, *O*-linked Fucosylation of some Ser or Thr side chains, or the glycosylation of nucleocytoplasmic proteins[388].

Whilst glycosylation is best understood as a modification present on extracellular

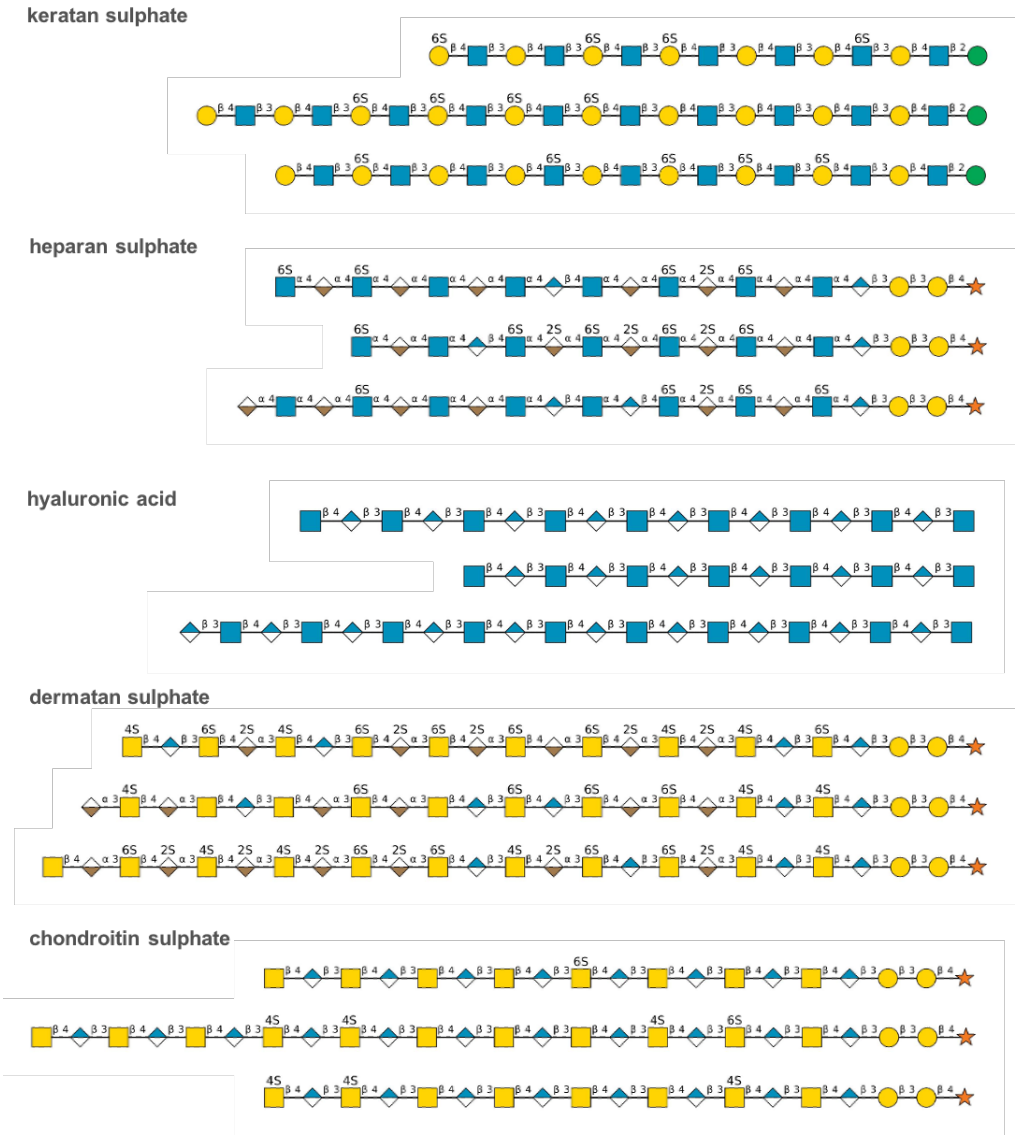


Figure 5.7: Examples of glycosaminoglycan chains produced using my code. For each GAG chain, the outputs from three independent runs of the code are shown in SNFG representation.

proteins and domains, there exist stable glycoprotein residents of the cytoplasm, nucleus, and plastids of eukaryotes, and indeed it appears that such modifications are ubiquitous and essential at least in metazoans. The most well-studied such cases are various proteins found inside eukaryotic cells bearing β-*O*-GlcNAc that is generally not elongated or elaborated. This modification shows a dynamic interplay with, and a similar subcellular distribution to, Ser/Thr phosphorylation, and is mediated by different enzymes from the glycosylation of extracellular/cell surface glyocproteins[495]. Furthermore, unlike the other glycosylation types discussed, this is highly dynamic, with the modifying enzymes carrying out rapid cycles of GlcNAc addition and removal in order to effect regulatory activity, based both on the modification itself and the competition of this modification with phosphorylation[495].

Code

Addition of single monosaccharides to protein side chains is relatively straightforward. The only additional element worth noting here is that a check is carried out to ensure the monosaccharide and residue type are matched. For instance, a single Man may be added to Trp or to Ser/Thr, but GlcNAc may not be added to the former as *C*-GlcNAcylation has not been known to take place. This check serves against, for instance, changes in residue numbering that can occur when handling downloaded or modified structures.

5.4 Example applications

Having presented the general workings of my glycosylation code, in this section I give examples of its application to a range of different glycoproteins. These serve principally to demonstrate that the code is functional, rather than to seek the investigation of biological phenomena. For full details of the simulations discussed here, please see Appendix C.

Myelin-associated glycoprotein

Myelin associated glycoprotein (MAG) is a single-pass transmembrane glycoprotein found as a component of the periaxonal myelin membranes in both the central and peripheral nervous systems. It is one of a number of different proteins that regulates the interaction between myelin and axons, and MAG has been shown to play an important role in the formation and maintenance of myelinated axons[496, 497].

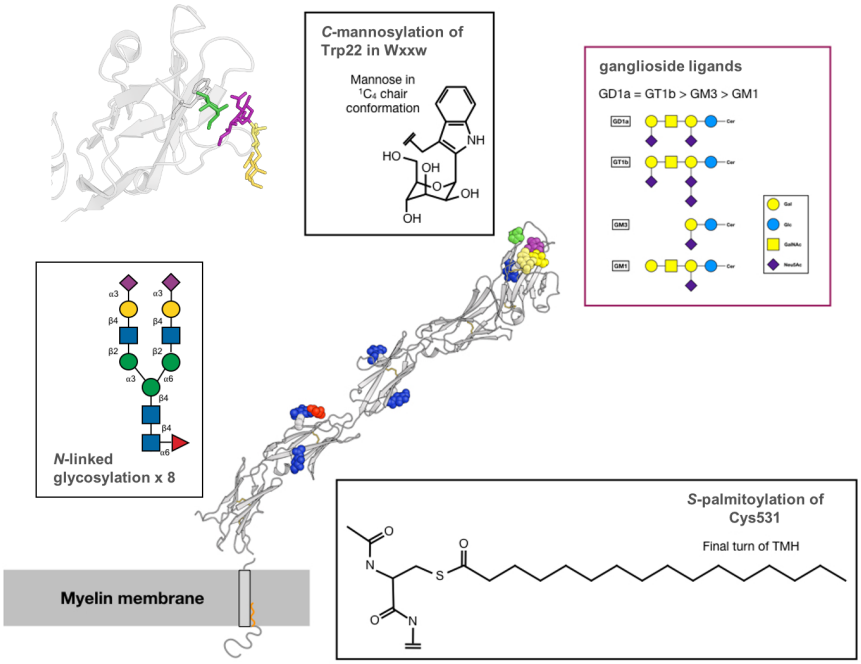


Figure 5.8: Structure and posttranslational modifications of MAG. The crystal structure (PDB ID 5LFU[38]) is shown in grey cartoon representations, with resolved sugar moieties as coloured spheres. Individual posttranslational modifications are indicated in the boxes. Note that the protein includes both Trp-mannosylation, and *N*-linked glycosylation, the former being in the vicinity of the ligand-binding site as shown in the top-left inset with the Man in sticks, and the ligand sugars resolved in the crystal structure as spheres. The ligands are gangliosides, with a preference towards GD1a and GT1b, as indicated in the top-right.

Structurally, MAG comprises an extracellular domain containing five immunoglobulin (Ig)-like domains, a transmembrane helix and a purportedly unstructured intracellular domain that may be long or short, depending on the MAG isoform (Figure 5.8). The N-terminal V-type Ig domain is involved in recognition of the MAG ligands, specifically found to be neuronal gangliosides bearing the Neu5Ac-α2,3-Gal-β1,3-GalNAc motif, such as GT1b and GD1a[498–502]. Through this interaction, MAG is able to bridge the space between the axonal membrane containing these gangliosides and the myelin membrane expressing MAG[503, 504]. This adhesion helps to maintain the myelin-axon spacing, and it has been suggested that *cis*-dimerisation of MAG can help to stabilise a 10 nm spacing in mature myelinated axons, whereas the monomer could span periaxonal distances of up to 16 nm that may exist during myelin formation[38]. In addition to this adhesive role, MAG also participates in bidirectional signalling between the axon and myelin, helping to regulate axon growth and myelin formation and integrity[496, 497].

Mass spectrometry and structural studies have demonstrated eight *N*-linked glycosites in human MAG[38, 505, 506], and a single site of *C*-mannosylation on Trp22 in the ligand-binding Ig1 domain[38].

Recently, structures were solved for the complete MAG ectodomain, both fully glycosylated and deglycosylated, as well as with and without mutations to alter its oligomeric state[38]. Accompanying these structural data, the authors carried out a series of small angle X-ray scattering (SAXS) studies to provide insight into the structural features underlying MAG-mediated adhesion and the role of dimerisation in MAG function. The availability of SAXS data for both the glycosylated and deglycosylated protein in multiple oligomeric states presents an excellent opportunity to evaluate my glycosylation method.

I therefore applied my method to the ectodomains of the same MAG variants as used in this study, summarised in Table 5.1. In each case, I built the CG protein using the AT-to-CG conversion tool, Martinize[160, 359], and the crystal structures 5LFU[38] for the full ectodomains ($MAG_{1\text{-}5}$) and 5LFR[38] for the truncated variants ($MAG_{1\text{-}3}$). In the single mutants (N406Q and I473E), the appropriate mutation was made before Martinization, using MODELLER[62]. The resulting CG proteins were then glycosylated using my method, with the *N*-linked glycans being added first (type selected as complex) and the Man residues added to Trp22 after.

In those cases where the experimental data had indicated the construct to be primarily monomeric, I considered it as a monomer, whilst in the cases where it was found to be mostly dimeric, I modelled a dimeric state. Ideally, I might run enough/sufficiently long simulations to permit dimerisation to potentially occur during the simulations. However, this was not possible at this stage and I therefore chose to take the dominant oligomeric state to ensure a better comparison with the likely state of the protein when it was examined using SAXS. For the dimers, each of the chains was glycosylated separately to ensure heterogeneity in the *N*-linked glycans.

The glycoproteins were incorporated into a cubic simulation cell with at least 2.0 nm distance to any edge, solvated with Martini water particles, and the system neutralised by addition of 0.15 M monovalent ions (NA and CL). Following this, energy minimisation and a short low timestep (1 fs) equilibration were performed, and the output from this used as the starting configuration for a 1 μs MD simulation. The results of these simulations were compared in terms of

MAG variant	glycosylation state	oligomeric state
WT-$MAG_{1\text{-}5}$	glycosylated	dimer
WT-$MAG_{1\text{-}5}$	deglycosylated	dimer
WT-$MAG_{1\text{-}3}$	glycosylated	monomer
WT-$MAG_{1\text{-}3}$	deglycosylated	monomer
N406Q-$MAG_{1\text{-}5}$	glycosylated	dimer
N406Q-$MAG_{1\text{-}5}$	deglycosylated	dimer
I473E-$MAG_{1\text{-}5}$	glycosylated	monomer
I473E-$MAG_{1\text{-}5}$	deglycosylated	monomer

Table 5.1: Summary of the constructs studied in [38], and used in my MAG ectodomain simulations. The subscript refers to the number of Ig-like domains included in the construct. The oligomeric state refers to the predominant state observed experimentally for each construct, and to the oligomeric state used in my simulations. Note that, as dimerisation was shown to be *via* the C-terminal Ig-like domains, the truncated mutants were monomeric. The N406Q mutant eliminates one of the *N*-linked glycosylation sites, which was found to enhance dimerisation[38], whilst dimerisation was abolished by the I473E mutation[38].

radius of gyration and maximum interatomic distance against the experimental data (Figure 5.9).

These results show that, although there is a discrepancy between the experimental and simulation data, the overall pattern (that is, the changes in radius of gyration and maximum interatomic distance) is similar. An exact match between the experimental and simulation results is not necessarily expected owing, for instance, to the presence of a population that includes monomers and dimers in the experiments as opposed to the single oligomeric states in my simulations, as well as the relatively short simulation times. One interesting feature of these data is that the glycans in both experiment and simulation do not make a considerable difference to either of the parameters quantified here. It would be both interesting and valuable to investigate the dimerisation potential of all my constructs, in comparison with the oligomeric data from the experiments, as well as to try to construct a hypothetical SAXS curve from my simulations for direct comparison against the SAXS data. This is, in principle, possible and indeed has been done previously[507]. However, it requires AT systems and I do not, at present, have a means to backmap my glycans to AT representation.

I note also that the experimental study presented some very interesting results that would be exciting to investigate computationally. Although time constraints and unforeseen challenges prevented us from doing so thus far, I have used the MAG-mediated adhesion as my proof-of-principle test system (Chapter 6). However, this too is just a preliminary model of the system that I hope can be

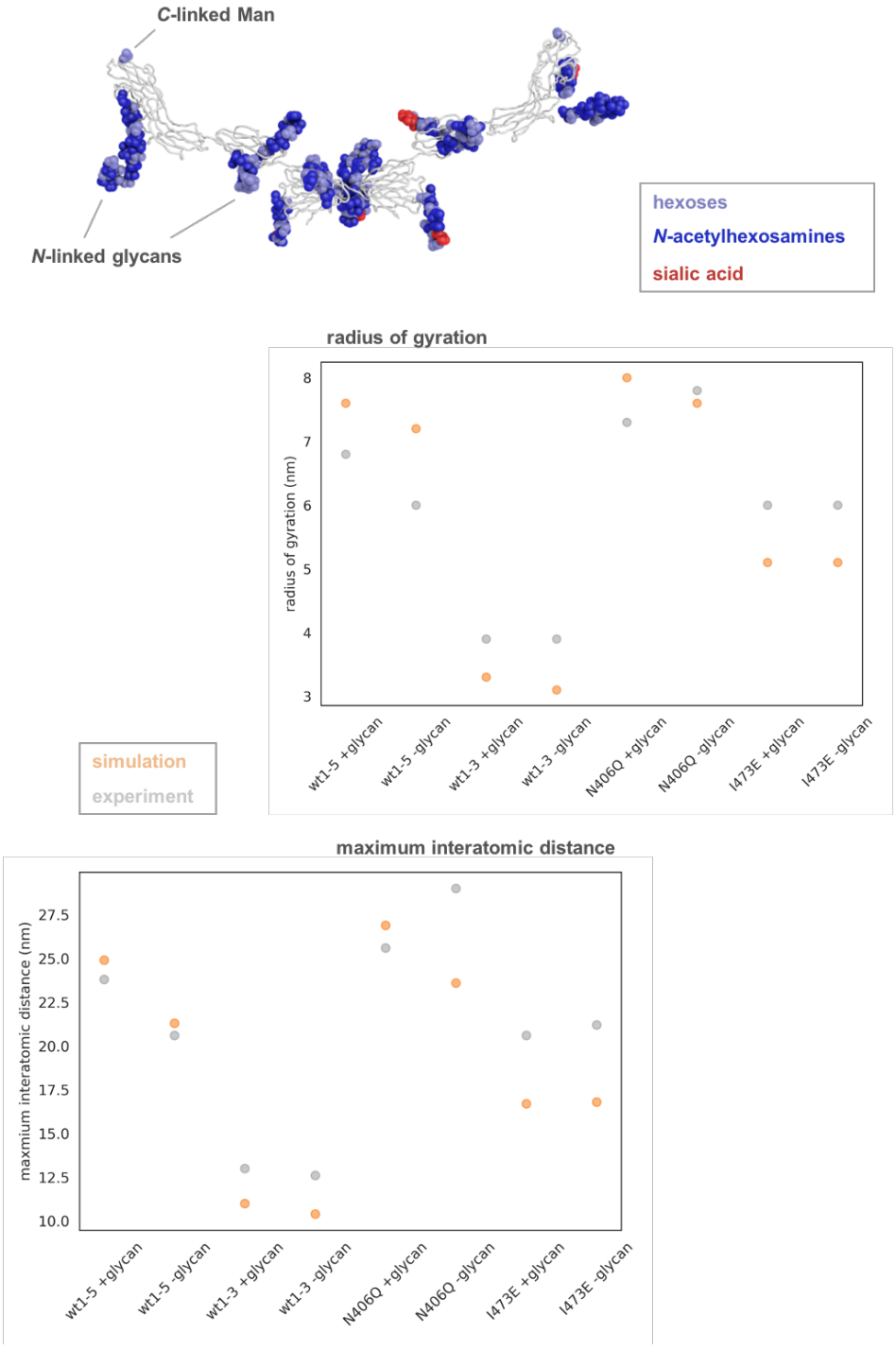


Figure 5.9: Results from the MAG simulations. Top: final frame from the CGMD of a glycosylated WT MAG dimer, showing the protein backbone as a grey tube and the glycans as spheres. Waters and ions have been omitted for clarity. Bottom: plots of mean radius of gyration and maximum interatomic distance obtained from simulation (orange) compared with the corresponding data from SAXS experiments (grey).

improved upon to allow a more in-depth examination of the MAG-mediated myelin-axon interface.

'Simple' glycoproteins

In order to further examine my method, I applied it to a set of glycoproteins with relatively few glycosylation sites, and a range of different types of glycosylation.

The approach in all cases was as follows. The protein structure, downloaded from the www.rcsb.org[58], was first modified using MODELLER[62] where necessary to incorporate missing loops or, in the case of ACE2, to include the transmembrane helix that is absent from the crystal structure[508]. The completed AT protein was then converted to Martini 3 representation using Martinize[160, 359], and the resulting CG protein then used as the input into the glycosylation code. As all the proteins were membrane-integral, I then incorporated the glycoprotein into a palmitoyloleoylphosphatidylcholine (POPC) bilayer of dimensions sufficient to accomodate the protein with no part of it within 2.0 nm of the simulation cell edge. The system (glycoprotein plus bilayer) was solvated with CG water, neutralised with 0.15 M NaCl, and this then equilibrated by short simulations (10 000 steps) at two low timesteps (1 fs, and 10 fs). Once this was completed, each was simulated for 1 μs with a 20 fs timestep.

The glycoproteins were stable for the duration of the 1 μs simulation, and the glycans were mobile rather than sticking to the protein surface, as was seen in earlier iterations of my tested models (data not shown). Brief descriptions of each of the test cases are given in the following sections, and the final frames from each are shown in Figure 5.10.

Smoothened

Smoothened is a class F GPCR involved in the Hedgehog signalling pathway, which plays critical roles in embryonic development and maintenance of adult tissues. There are three PNGs in the Smoothened protein sequence, one of which occurs in the intracellular region, and therefore should not be glycosylated. When I ran my code without the bilayer flag, all three were identified and glycosylated, whereas when the flag was included, only the two extracellular sites were used (Figure 5.11), demonstrating that my topology prediction method can, at least in this case, work as expected.

In the subsequent 1 μs simulation of Smoothened with glycans in a POPC bilayer,

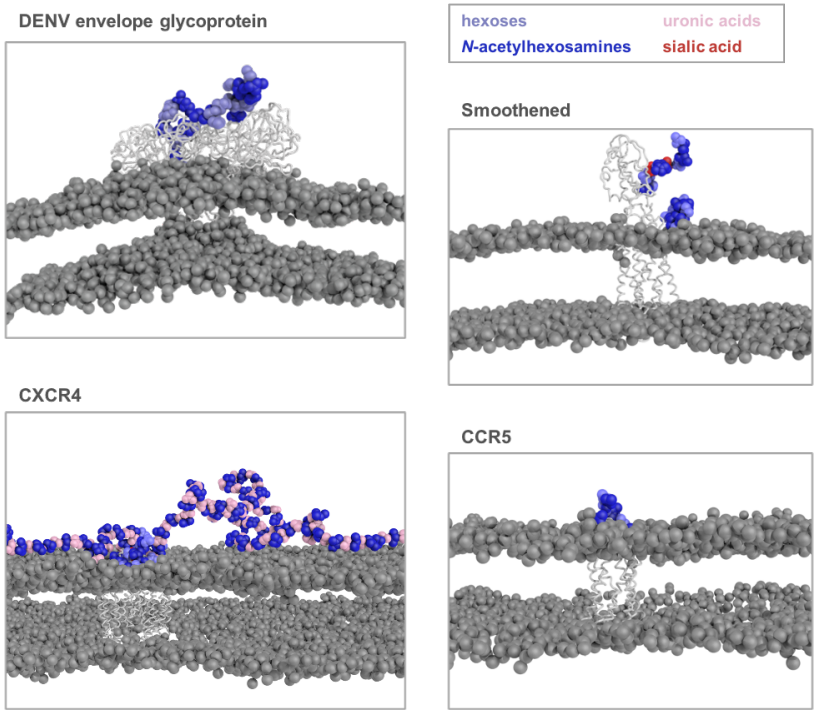


Figure 5.10: Final (1 µs) snapshots from four different simulations of glycoproteins in bilayers. In each case, the lipid headgroups (phosphate and choline beads) are shown in dark grey, the protein backbone as a light grey tube, and the glycans as spheres coloured according to CG monosaccharide type, as indicated in the key. For CXCR4, the GAG chain crossed the periodic boundary, and so appears to be coming back in from the other side of the simulation cell.

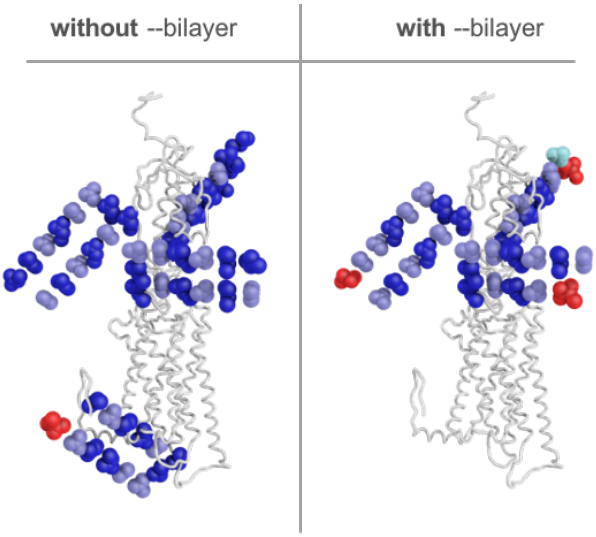


Figure 5.11: Illustration of the product Smoothened glycoproteins when the -bilayer flag is not (left) and is (right) included, showing the exclusion in the latter case of the intracellular PNG. The protein backbone is shown as a grey tube and the glycans as spheres, coloured according to CG residue type, with sialic acid in red, the *N*-acetyl hexosamines in dark blue, the hexoses in slate, and deoxyhexose in cyan.

the glycoprotein was stable, with no overtly aberrant behaviours observed. As this, and the subsequent glycoproteins discussed in this section, was just used as a proof-of-principle, I did not carry out any further analyses into the glycans or their effect on the protein.

HIV-1 coreceptors CCR5 and CXCR4

In order to gain entry into host cells, the human immunodeficiency virus HIV-1 surface glycoprotein gp120 must interact with both a receptor (CD4) and a coreceptor (CCR5 or CXCR4), expressed on the surface of host cells. The latter interaction is necessary to promote the conformational changes in gp120 that culminate in its dissociation into fusogen and receptor-binding parts, and is thought to stabilise the CD4-induced conformation of the viral glycoprotein[509–511]. The choice of coreceptor has been suggested to play a role in determining viral tropism, and the transition from CCR5 to CXCR4 usage appears to be associated with accelerated disease progression[512, 513].

Both CCR5 and CXCR4 are class A GPCRs, functioning physiologically as chemokine receptors. In addition to the classic GPCR 7TM fold, both possess extracellular, glycosylated N-terminal domains that participate in both chemokine and gp120 binding[511]. Despite having similar overall folds and being bound by the same viral glycoprotein, these two GPCRs exhibit very different glycosylation patterns. CCR5 is subject to mucin-type *O*-linked glycosylation on at least two sites, Ser6[514] and Ser7[515], whilst CXCR4 is *N*-glycosylated at Asn11[516], and *O*-glycosylated with a CS glycosaminoglycan chain at Ser18[517]. It has been shown that *N*-glycosylation of CXCR4 can reduce its use as a coreceptor for HIV-1 by directly impeding its binding to the viral glycoprotein[516]. The sites of *O*-linked glycosylation in CCR5, however, have been suggested to participate in gp120-binding as both are part of the N-terminal extension structurally shown to interact with the viral glycoprotein[511]. However, as only one residue of the *O*-linked glycan on Ser6 and none on Ser7 could be resolved in the cryo-EM structure of gp120-bound CCR5[511], and to my knowledge the experiments with and without glycosylation have not been done, it is not yet clear to what extent glycans may or may not contribute to gp120 binding by CCR5.

I used the structures 4MBS (CCR5)[518] and 3OE6 (CXCR4)[519] as the starting points for my models of these GPCRs, producing glycosylated Martini models and simulating them in POPC bilayers as described previously. As CXCR4

is believed to exist as a dimer, I modelled it as such, doing the glycosylation separately for each chain.

Both glycoproteins were stable in the bilayer, although in order to simulate CXCR4 I needed to run a longer equilibration, specifically equilibrating at more gradually increasing timesteps (1 fs, 2 fs, 5 fs, and 10 fs). I believe this to be due to the GAG chain, which with far more residues than the *N*-linked glycans, has considerably more opportunity for steric clashes or geometric instabilities. The fact that CXCR4 was modelled as a dimer would exacerbate this issue as there are two long GAG chains to contend with. This issue may suggest that there is some element in my oligosaccharide models that produces abnormal conformations of longer glycans. I suspect that the problem may arise in the method I use to build the oligosaccharides in my code, as it involves the same geometric positioning of all new residues relative the preceeding one, regardless of residue or bond type.

ACE2

Angiotensin converting enzyme 2 (ACE2) has become something of a hot topic over the past year owing to the demonstration that it acts as the receptor for the SARS-CoV-2 Spike glycoprotein[520]. Physiologically speaking, it is a carboxypeptidase with important roles in cardiovascular homeostasis. It is a single-pass transmembrane glycoprotein with a globular ectodomain that is subject to *N*-linked glycosylation on at least six of the seven PNGs identified in its sequence[508]. Interestingly, glycosylation of Asn90 has been found to reduce the usage of ACE2 as a receptor by the SARS-CoV-2 Spike, and binding affinity of the Spike receptor binding domain was significant enhanced when this glycosylation site was disrupted[521].

I used the structure 1R42[508] as my starting point, and modelled the glycosylated version as described previously, including adding the missing TM helix, and incorporated the structure into a POPC bilayer. I repeated this process for the deglycosylated protein and for the glycosylated protein with no glycan on Asn90, and compared the solvent accessibility of the residues in each case over the course of the 1 μs simulation. The results are shown in Figure 5.12, where the average accessibility of each residues to solvent is indicated by a number between zero and one, with one suggesting the residue is accessible to solvent throughout the simulation, and zero that it is never exposed to solvent.

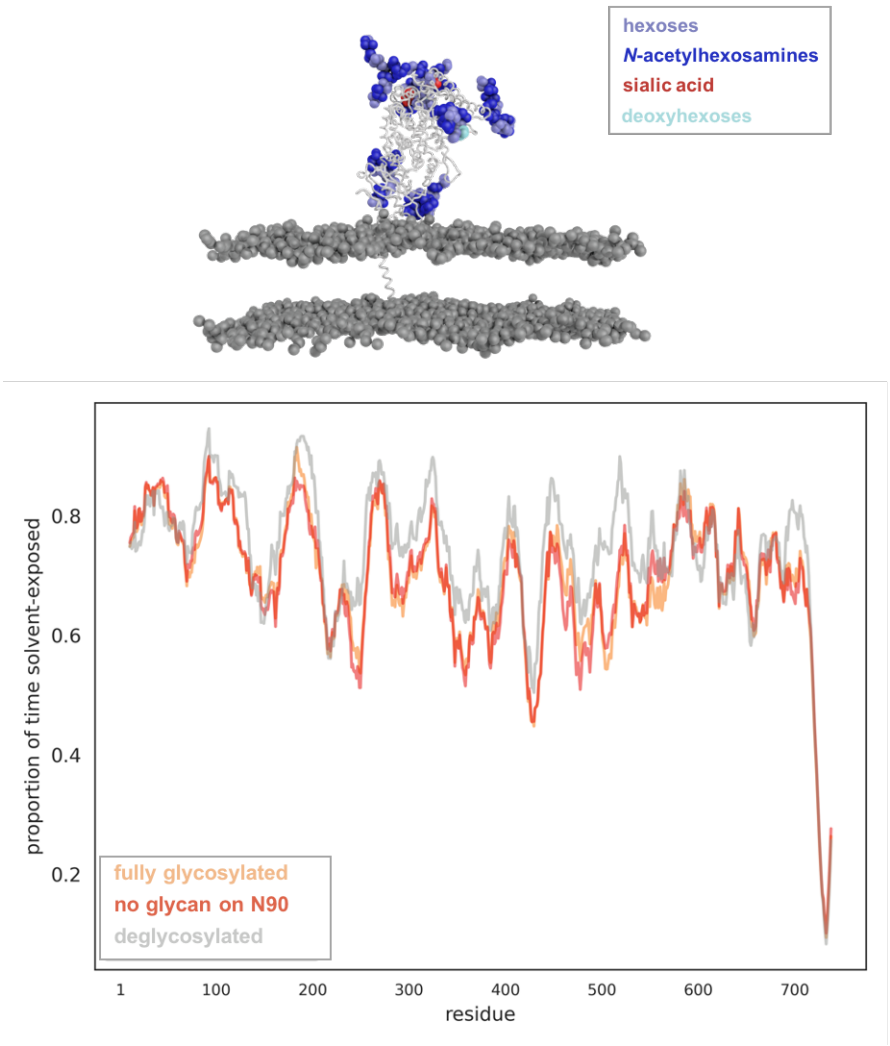


Figure 5.12: ACE2 simulation results. Top: final frame from CGMD of fully glycosylated ACE2 in a POPC bilayer, with the protein backbone as a grey tube and the glycans as spheres, coloured by residue type as indicated in the key. Waters and ions have been omitted for clarity. Bottom: plots of the proportion of the simulation each residue spends exposed to solvent, comparing fully glycosylated, fully deglycosylated, and partially deglycosylated (missing the glycan on Asn90) simulations.

From these data, there is a slight difference observable between the solvent exposure of the glycosylated and deglycosylated variants, but very little difference in any region between the variants with and without glycosylation on Asn90. This could suggest that the presence or absence of a glycan on this residue does not alter the accessibility of epitopes on ACE2 to potential binding partners, and therefore that any effect on binding is more subtle than this. However, my data is very preliminary and the simulations too short and inextensive to permit any conclusions to be drawn from them.

Dengue virus glycoproteins

Dengue virus (DENV) is an enveloped, arthropod-borne virus that causes dengue fever, usually associated with symptoms such as headache, high fever, and skin rash, but which can in rarer cases develop into a far more severe haemorrhagic fever. The mature viral envelope contains 180 copies of E and prM glycoproteins, of which only E is exposed on the surface[522]. The E protein, which binds to host cell receptors and acts as the viral fusogen to achieve cell entry[522–524], is subject to *N*-linked glycosylation at two sites, N67 and N153[525–527]. The nature of the attached glycans is influenced by the viral host, and therefore its stage in the transmission cycle. Specifically, DENV produced in insect cells shows a high level of oligomannose and paucimanose structures[527, 528], whilst those from primary dendritic cells presented more complex type glycans[528]. This has implications for cellular tropism, and highlights the importance of being able to model different types of glycosylation. PrM is also glycosylated at N69, a modification which was suggested to be important in proper protein folding[529].

I modelled the DENV envelope proteins using the cryo-EM model of the E-prM heterotetramer 3J2P[460] and glycosylated the identified modification sites as described previously with oligomannose-type glycans. During the equilibration stage of the simulation, the POPC bilayer adopted a distinct curvature owing to the protein, and this was retained throughout the simulation. A cursory visualisation of the trajectories revealed little of further note – the glycoproteins were stable and the glycans dynamic. It would be interesting to use this glycoprotein in further simulations to investigate whether or not my oligomannose-type glycans are recognised by DC-SIGN, which is known to be a DENV receptor and to bind oligomannose-type glycans[530].

More 'complex' cases

Having demonstrated the capacity of my method to model glycoproteins with relatively few attached glycans, I moved onto attempting to glycosylate proteins that are more extensively adorned with sugars. These proved considerably more challenging, and generally were ultimately intractable. I focus here on just one of the attempted systems to illustrate the problem, which I have yet to find a viable solution to.

SARS-CoV-2 glycoprotein

The enveloped coronavirus SARS-CoV-2 is the causative agent of a pandemic that has dominated the news and took centre stage of scientific research for much of this year, having spread rapidly across the globe, with over 26 million cases having been recorded worldwide as of 3/09 2020. The viral receptor-binding and fusogenic activities are attributed to the Spike glycoprotein, a trimeric transmembrane protein that is heavily *N*-glycosylated with predominantly complex type glycans[479]. This had potential as a very useful case to test my glycosylation method due to the considerable structural and experimental data becoming available at present. It would have been too ambitious to hope that my method, still very much in the developmental stage, could have provided any useful results from the point of view of understanding the virus.

I attempted to construct a CG model of the Spike glycoprotein, and although I was able to do so and include both complex and oligomannose glycans at approproate sites, the resulting model was unstable and could not be simulated for more than a few nanoseconds, and this only at a low timestep (10 fs), even after lengthy equilibrations at gradually increasing timesteps. This was true regardless of whether I considered the full-length glycoprotein or just the ectodomain, and the unglycosylated variants of both were successfully simulated for 1 μs each.

One possible explanation for the instabilities might have been the high density of glycans clashing with each other, a notion that I examined by repeating the tests but including only a subset of all the identified glycosylation sites (Figure 5.13). The removal of some of the glycosylation sites did seem to help matters, in that the system has proceeded further than the fully glycosylated test and not yet crashed. However, the simulation is still ongoing in this case and it is entirely possible that it will crash at a later stage.

Even if the partially glycosylated protein can be successfully simulated for the

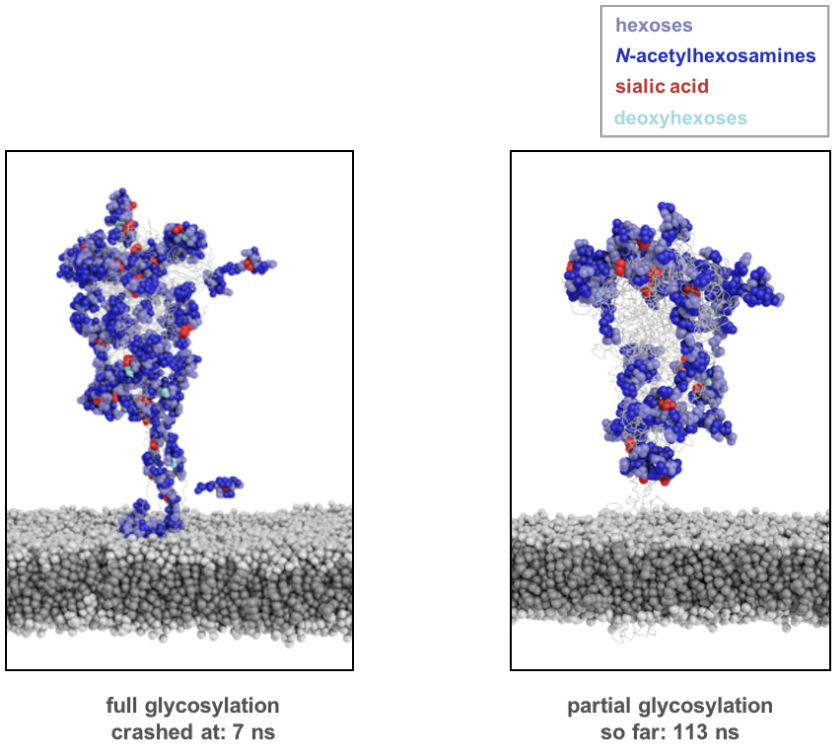


Figure 5.13: Comparison of the SARS-CoV-2 Spike glycoprotein simulations with all (left) and with a subset (right) of the glycosites modified. Top: starting configurations. Bottom: configurations at the longest time the simulation got to/has been run thus far, as indicated. The protein is shown as grey tubes, the bilayer as grey spheres, and the glycans as spheres coloured according to the key.

full time, omitting glycosites is not a suitable solution – I cannot justify simply ignoring sites known to be glycosylated as doing so defeats the logic of glycosylating the proteins in the first place, that is, to model a more biologically-relevant construct. I expect it will take some considerable work and systematic investigation to identify the sources of the instability and develop a solution.

5.5 Future work

In its present state, the code described is capable of producing glycoproteins that exhibit micro- and macroheterogeneity, and which can be used in CGMD simulations. However, the output glycoproteins appear to be unreliable in their stability, particularly when there are multiple glycans present, in such cases often crashing after a few hundred nanoseconds. This suggests that there is something in the topology that renders the molecule ultimately unstable. As isolated glycans are stable for long simulation times, it is likely that the source of the instability is either how I am linking the glycans to the protein or how

I am extending the cores in the script. In this latter case, the problem may lie in the way the glycans are geometrically assembled, which does not take into account the preferred spatial arrangement of the sugars in the oligosaccharide. Determining the specific problem will require picking apart the programme carefully and examining the success of each individual step.

Once this issue is fixed, it will be important to incoporate into my method several additional features, particularly the ability for oligosaccharides to contain more than one branch point. Currently, branching is limited to that already present within the core structure, whilst in biology glycans might contain up to four branch points. my code can only yield relatively small, bifurcated glycans, and is therefore unable to capture a large part of oligosaccharide space. Furthermore, I will need to build in the possibility of applying modifications to the monosaccharides such as phosphorylation and sulphation. Another extension of the approach that would be useful would be to be able to define specific glycan epitopes to be included, as for some proteins it is known that their glycans include, for instance, blood group determinant motifs or polysialic acid regions. Additionally, if domain-specific glycans could be added as a possible type (alongside *N*-linked, mucin-type *O*-linked, GAG, and monosaccharide), this would make it easier to model specific glycoproteins about which there are some existing data on the glycan components.

Finally, it will be necessary to thoroughly validate the behaviour of the resulting glycoproteins against experimental results. This is rendered more difficult by the limited availability of suitable experimental data for such a purpose. The utility of having suitable experimental data is illustrated in the case of MAG, discussed in this chapter, where my models could be assessed against SAXS data. Similar types of experiment for other glycoproteins would prove excellent for refining and validating any method of simulating glycoproteins.

As a posttranslational moficiation, protein glycosylation is thought to be ubiquitous across all clades. However, I focus my development here on eukaryotic glycosylation, and in particular that of vertebrates. This is not due to a belief that the glycosylation of proteins in other organisms is insignificant, but rather because it was the simplest place to start. The wealth of variety in glycosylation that exists in the biological world is extensive, and it will take quite some time to build the glycosylation types belonging to other clades.

Chapter 6

PUTTING IT ALL TOGETHER

6.1 Introduction

In Chapters 3, 4, and 5, I have presented work that aims to establish means to simulate individual elements that frequently compose membrane-membrane associations. It was my hope that in this concluding chapter I might present a set of proof-of-principle simulations, illustrating the coming together of all these elements and applying them to a thorough investigation of at least one membrane-membrane interface. However, it has not been possible to simulate any such association for a long enough time for any conclusions to be drawn or interesting observations made, let alone consider variants upon a system or additional complexes. I have, however, set up a system whose simulation is ongoing, and in this final chapter I present the approach taken to setting this system up and observations made during this process, which may be of use in improving on the individual methods previously discussed.

6.2 The system

The membrane-membrane association I elected to model as my proof-of-principle is the myelin associated glycoprotein (MAG)-mediated association between myelin and axonal membranes in the central and peripheral nervous systems, as discussed in Chapter 5. As mentioned in that chapter, the experimental study I used to help evaluate my simulation data included some interesting observations and, among other things, proposed that dimerisation of MAG might be enhanced by clustering of its ganglioside ligands[38]. This seemed to be the sort of suggestion that is very much amenable to evaluation using MD simulations.

For instance, one could assemble systems in which MAG in one bilayer interacts with a ganglioside in the other. Doing so with two copies of monomeric MAG, and then repeating the simulation without the ganglioside ligand present (a single bilayer or the double bilayer with a non-ligand ganglioside), might provide qualitative insight into the effect of ligand binding on dimerisation. Here, however, I only construct the most basic model of the system as a starting point to help establish an approach to construct double-bilayer systems for simulation.

6.3 Approach taken

My general approach to system construction is illustrated in Figure 6.1. First, I prepared a bilayer using *insane*[190]. The composition of this bilayer was based on a previous simulation of the neuronal plasma membrane in which the composition was derived from the consensus of several lipidomic studies[152], although my bilayer was somewhat simplified compared to this. Specifically, POPC: DOPC:PUPC:PAPC:DPPC:POPE:PUPE:PAPE:DPCE:PNCE:PGSM:PN SM:DPSM:GM3:CHOL in the ratio 9:3:2:5:5:1:3:5:5:1:1:1:6:4:44. For reference, as some of the CG lipid naming is nonstandard, the headgroups are phosphatidylcholine (PC), phosphatidylethanolamine (PE), ceramide (CE), and sphingomyelin (SM), and the lipid tail groups corresponding to these species are as follows:

- PO – C16:0/18:1
- DO – di-C16:1-C18:1
- PA – C16:0/20:4
- PU – C16:0/22:6
- DP – di-C16:0-C18:0
- PN – C(d18:1/24:1)
- PG – C(d18:1/22:1)

After constructing the bilayer, I replaced the gangliosides (GM3) in this with my GT1b model, discussed in Chapter 4. This bilayer was then duplicated to create the two bilayers to be combined. In this particular system, I chose to use bilayers of the same composition as this is primarily a proof-of-principle application and I was not certain of having sufficient time to equilibrate two different bilayers. However, the plasma membranes involved are in fact likely to be different in composition. For instance, myelin membranes are known to be enriched in galactosylceramide and its sulphated derivative[531]. In general, it would be straightforward to construct two different bilayers if necessary using the various available methods such as those discussed in Chapter 2.

One of these was equilibrated gradually while introducing a pore using the method presented in Chapter 3, with repulsion radius of 2.5 nm and strength

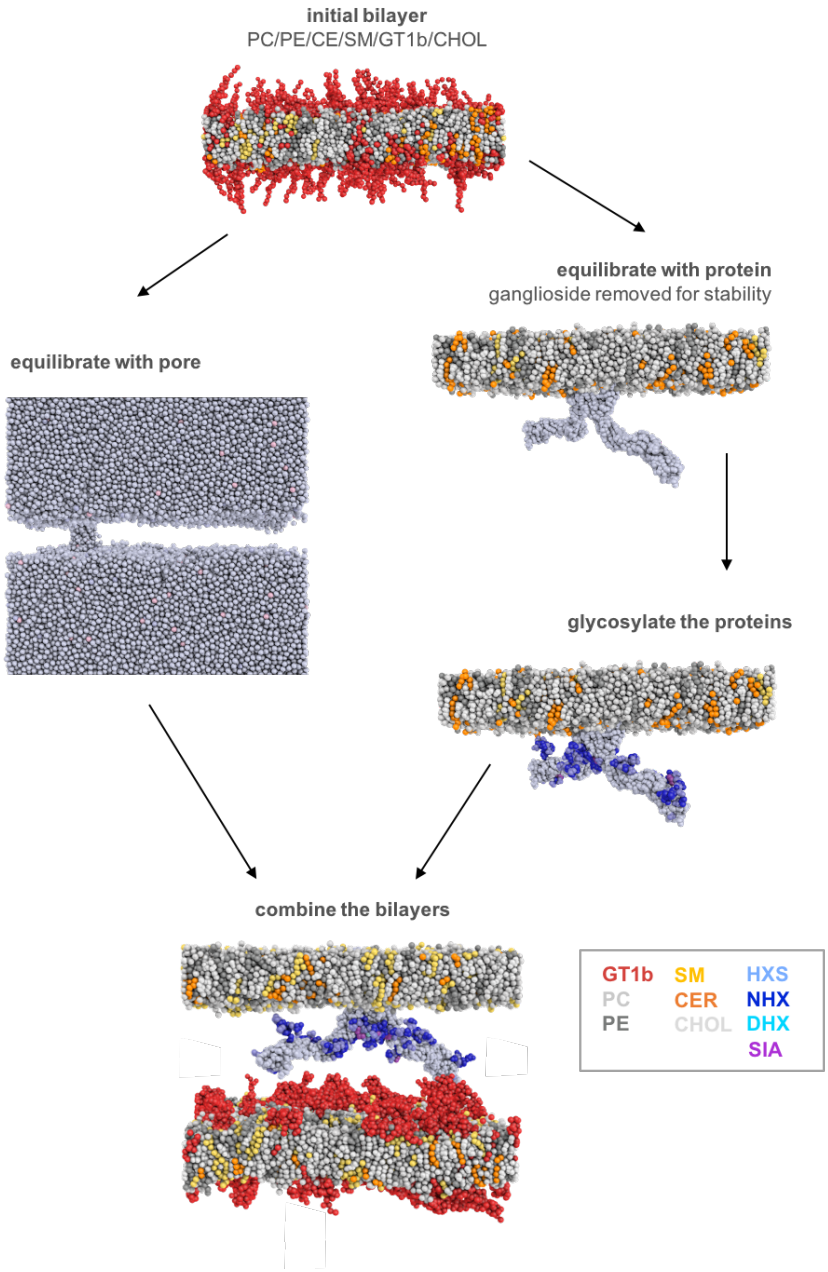


Figure 6.1: In order to assemble a double bilayer system, I started by constructing and equilibrating each of the bilayers separately. A pore was incorporated into one during equilibration, and the unglycosylated proteins into the other. After equilibration of this latter, the proteins were glycosylated. The bilayers were then combined by translating them relative to each other, the system solvated and neutralised, thereby yielding the complete system.

1000 kJ mol^{-1} applied to all bilayer constitutents. To account for the complexity of the bilayer, I carried out the equilibration in stages, with 100 000 steps of simulation at 1 fs, 2 fs, 5 fs, and 10 fs timesteps, which had not been necessary for the simple bilayer presented in Chapter 3. Attempting to use a too high timestep immediately caused the simulation to crash.

The other bilayer was equilibrated without a pore, following the addition of full-length, unglycosylated MAG molecules. These were constructed by first building a full-length AT protein using MODELLER[62] and the crystal structure of the full MAG ectodomain[38],and converted to Martini 3 using Martinize[359]. The proteins were incorporated into the bilayer using Alchembed[289], and the system equilibrated as for the pore-containing bilayer. This equilibration was done in the same stages as for the pore-containing bilayer to ensure consistency in how each bilayer was prepared.

Only once the bilayer had been equilibrated with the deglycosylated proteins present did I glycosylate them using the method discussed in Chapter 5 at the sites of glycosylation indicated in Figure 5.8, including both the *N*-linked glycans and the Trp-mannosylation. I considered the MAG dimer, which is proposed to be the dominant state in the mature association, and glycosylated each of the protein chains separately to ensure the potential for heterogeneity between them.

Following this glycosylation, I found the system was unstable and could not be simulated on its own, even after relatively long equilibration at gradually increasing timesteps. I was able to equilibrate it and run a short simulation at a 20 fs timestep after removing the gangliosides from that bilayer. For the time being, I found it necessary to make this undesirable change in order to proceed with the system assembly, but it points to either a serious issue with my models or a problem with my preparation approach. Had I time to repeat the process, I would equilibrate the bilayer slowly with the already glycosylated proteins to assess whether or not this prevents such problems, or if the issues is in the models themselves.

The next step was to add one bilayer to the other. This was in itself relatively straightforward, being achieved through simple translation of one bilayer along the z-direction (normal to the bilayer) and elongation of the simulation cell in this direction. I had to remove all solvent in this process and resolvate the system, followed by readdition of the ions. The system was then energy minimised and in theory ready for further simulation (Figure 6.2). In this case, I did not define

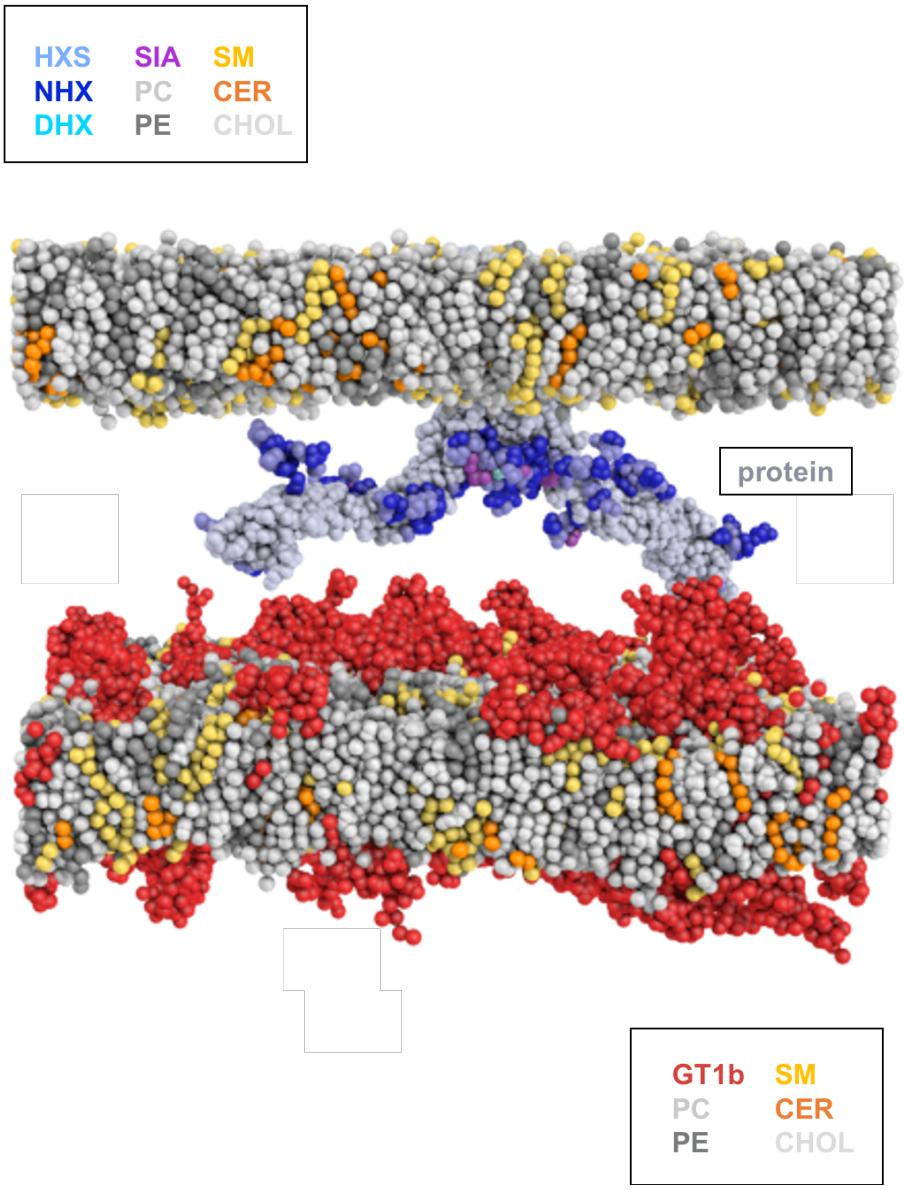


Figure 6.2: The starting configuration of the MAG-bridged myelin-axon membrane-membrane association system simulated. The lipid, glycan, and protein components are shown as spheres, and waters and ions omitted for clarity. The components are coloured as indicated in the keys.

MAG as being already bound to its ligands as I wish to see whether or not they are able to do so during the simulation, although it would be possible to start with a bound configuration given the binding site for the ganglioside ligands in MAG is known.

This system was energy minimised, and equilibrated in stages of 100 000 steps simulation at increasing timesteps 1 fs, 2 fs, 5 fs, and 10 fs. I intend to simulate it for 5 µs production run, and this has been started, but have not got more than

a few hundred nanoseconds thus far.

6.4 Concluding remarks and future directions

In this thesis, I have presented my approach to developing a robust methodology for using MD simulations to study membrane-membrane interactions. Although I do not believe that I have succeeded in producing such a methodology at this stage, there being too many aspects incomplete in their evaluation and remaining stability issues, I have certainly made progress.

The use of a pore in one of the interacting bilayers offers the potential to facilitate greater relative motion of the bilayers than might otherwise be possible, which would be beneficial not only in hopefully allowing more realistic bilayer dynamics, but of particularly evident importance in situations where the bilayers may undergo overall motion towards or away from each other during the simulation.

Glycans form a major constituent of the extracellular milieu, and are believed to play an important role in a number of membrane-membrane interactions. It is therefore critical that I can model the glycan constituents of the interacting bilayers. I have presented my parameterisation of glycans for use in CGMD simulations using the Martini force field, taking a modular approach to their construction that is in keeping with the Martini philosophy[49]. I have then used these models as a basis to develop a protein glycosylation protocol that is, in its basic design, applicable to other force fields. In this, heterogeneity of the glycans is achieved through a biology-based quasi-random approach to sugar additions for *N*- and mucin-type *O*-linked glycans. The glycan structures this scheme yields are, I think, not quite representative of the real situation and therefore work will need to be done to improve this. For instance, at present only biantennary glycans are possible, and no modifications to the glycan cores such as fucosylation of the GlcNAc residues in the *N*-glycan core are considered. Additionally, the glycans are essentially restricted to vertebrate-types, and this will need to be amended in future developments. The greatest issue, however, remains the instability of glycoproteins that are particularly densely glycosylated, as described for the SARS-CoV-2 Spike glycoprotein. As there is a likelihood that this instability arises in the way in which I construct the glycoproteins, it is important that solving this problem takes priority over extensions of the glycosylation code.

Despite these limitations, and other discussed in more detail in the respective

chapters, I believe that my work does at least provide the basis for allowing the construction and simulation of double-bilayer systems for simulation using the Martini force field. I have shown in this chapter that there are further considerations to be made in how those systems are actually constructed, including whether to glycosylate constituent proteins before or after equilibration of the bilayers in which they are found, and the best approach to equilibrating what is ultimately a complex system. A systematic investigation of different approaches to prepare the double bilayer systems would be useful, as would developing code that can at least in part automate this process to ensure e.g. that the bilayers are separated by the correct distance and are of the same x-y dimension following pore-formation and protein insertion.

Thus, to briefly conclude this work, I have successfully made progress towards designing a general approach that could be used to simulate double-bilayer systems, but in doing so have identified a considerable number of issues that need to be tackled before I can reliably prepare such simulations and have reasonable confidence in the results I obtain. As typical for science, in attempting to solve one problem, I have uncovered a veritable wealth of others.

ACRONYMS

ACE2 angiotensin converting enzyme 2.

AFM atomic force microscopy.

All allose.

Alt altose.

AT atomistic.

BAR Bennett Acceptance Ratio.

CE ceramide.

CG coarse-grained.

CLEM correlative light electron microscopy.

CS chondroitin sulphate.

DENV dengue virus.

DLPC dilauroylphosphatidylcholine.

DOPC dioleoylphosphatidylcholine.

DPPC dipalmitoylphosphatidylcholine.

DS dermatan sulphate.

EM electron microscopy.

Fru fructose.

Fuc fucose.

GAG glycosaminoglycan.

Gal galactose.

GalA galacturonic acid.

GalNAc N-acetylgalactosamine.

Glc glucose.

GlcA glucuronic acid.

GlcN glucosamine.

GlcNAc N-acetylglucosamine.

GPI glycosylphosphatidylinositol.

HA hyaluronic acid.

HS heparan sulphate.

Ido idose.

IdoA iduronic acid.

Ig immunoglobulin.

KS keratan sulphate.

MAG myelin associated glycoprotein.

Man mannose.

MD molecular dynamics.

MIFC multispectral imaging flow cytometry.

Neu5Ac N-acetylneuraminic acid.

OPM Orientation of Proteins in Membranes.

PBC periodic boundary conditions.

PC phosphatidylcholine.

PE phosphatidylethanolamine.

PG phosphatidylglycerol.

PNG potential N-linked glycosylation site.

POPC palmitoyloleoylphosphatidylcholine.

PS phosphatidylserine.

Qui quinovose.

Rha rhamnose.

SAXS small angle X-ray scattering.

SG supra-coarse-grained.

SM sphingomyelin.

SNFG Symbol Nomenclature For Glycans.

TM transmembrane.

Xyl xylose.

GLOSSARY

anomers monosaccharides that differ from each other only in the configuration of the anomeric carbon of the ring structure.

atomistic a version of a system in which all the heavy atoms are included.

barostat algorithm used to maintain the average pressure of a simulation system at a defined value.

chondroitin sulphate glycosaminoglycan defined by the disaccharide unit [GalNAcβ1–4GlcAβ1–3], which may be sulphated at certain positions.

coarse-grained a version of a system in which the number of degrees of freedom are reduced i.e. a simplified representation of the system.

dermatan sulphate glycosaminoglycan with a [GalNAcβ1–4GlcAβ1–3] disaccharide repeat, which may be sulphated at certain sites.

disaccharide molecule composed of two monosaccharides covelently linked together *via* a glycosidic bond.

ensemble the set of possible states of a system in phase space.

force field set of potential functions and parameterised interactions that can be used to model a system of interacting particles.

ganglioside glycosphingolipids containing one or more sialic acid groups.

ganglioside a glycosphingolipid with one or more sialic acid groups.

glycan generic term for sugars, which may be free or covalently linked to another molecule.

glycoconjugate a molecule in which one or more glycan units are covalently linked to a non-glycan entity.

glycolipid lipid with a covalently attached glycan moiety.

glycoprotein protein with one or more covalently linked glycans.

glycosaminoglycan linear chains of repeating disaccharide units, each composed of a hexosamine and a hexose or hexuronic acid.

glycosidic bond covalent bond between one glycan to another molecule *via* the anomeric hydroxyl group.

glycosphingolipid class of glycolipid composed of ceramide covalently linked to a glycan moiety.

glycosylation the enzymatic linkage of a glycan to polypeptide, lipid, nucleic acid, or other organic compound.

GROMACS molecular dynamics simulation software package, used throughout this work for running simulations.

Hamiltonian an operator, which is the sum of the potential and kinetic energies for all the particles in a system.

heparan sulphate glycosaminoglycan defined by disaccharide repeats of [GlcNAcα1–4 IdoA/GlcAβ1–4], with epimerisation and sulphation at various positions.

hexose a six-carbon monosaccharide, typically with an aldehyde/potential aldehyde at the C1 position and hydroxyls at the others e.g. glucose, mannose, galactos.

hyaluronic acid glycosaminoglycan with a GlcNAcβ1–4GlcAβ1–3 disaccharide repeat, that is generally neither sulphated nor covalently linked to a protein.

keratan sulphate glycosaminoglycan with a [Galβ1–4GlcNAcβ1–3] disaccharide repeat, which may be sulphated at certain sites.

Martini coarse-grained force field for molecular dynamics simulations of biomolecular systems; not to be confused with the alcoholic drink of the same name.

microheterogeneity structural variations in the glycans found at any given glycosite on a glycoprotein.

monosaccharide carbohydrate that cannot be hydrolysed into simpler carbohydates.

non-reducing end in a glycoconjugate, the outermost end of a glycan chain (i.e. that not attached to the aglycone component).

oligosaccharide molecule composed of several monosaccharide residues covalently bonded *via* glycosidic bonds into linear or branched chains.

phase space for a system of N particles, a 6N dimensional space, within which atomic positions and momenta are coordinates; a molecular dynamics simulation generates a series of points in phase space.

polysaccharide glycan, generally over ten residues in length, composed of repeating monosaccharides.

proteoglycan a protein with one or more covalently linked glycosaminoglycan chains.

pyranose six-membered ring form of a monosaccharide.

reducing end in an glycoconjugate, the end of a glycan that is attached to the aglycone component by a glycosidic bond.

thermostat algorithm used to maintain the average temperature of a simulation system at a defined value.

timestep the interval in a simulation between evaluations of the potential.

trajectory in molecular dynamics, the time evolution of the system of particles being simulated.

BIBLIOGRAPHY

[1] Belardi B, Son S, Felce J H, Dustin M L and Fletcher D A 2020 *Nature Reviews Molecular Cell Biology* **21** 750–764 ISSN 1471-0072 URL http://www.nature.com/articles/s41580-020-00298-7

[2] Gumbiner B M 1996 *Cell* **84** 345–357 ISSN 00928674 URL http://linkinghub.elsevier.com/retrieve/pii/S0092867400812799

[3] Batista F D and Dustin M L 2013 *Immunological Reviews* **251** 7–12 ISSN 01052896 URL http://doi.wiley.com/10.1111/imr.12025

[4] Weiner J A, Jontes J D and Burgess R W 2013 *Frontiers in Molecular Neuroscience* **6** ISSN 1662-5099 URL http://journal.frontiersin.org/article/10.3389/fnmol.2013.00012/abstract

[5] Chao D L, Ma L and Shen K 2009 *Nature Reviews Neuroscience* **10** 262–271 ISSN 1471-003X URL http://www.nature.com/articles/nrn2594

[6] Haselkorn R 2008 Cell-cell communication in filamentous cyanobacteria

[7] A Garnett J and Matthews S 2012 *Current Protein and Peptide Science* **13** 739–755 ISSN 13892037 URL http://openurl.ingenta.com/content/xref?genre=article&issn=1389-2037&volume=13&issue=8&spage=739

[8] Dunne W M 2002 *Clinical Microbiology Reviews* **15** 155–166 ISSN 0893-8512 URL https://cmr.asm.org/content/15/2/155

[9] Harrison S C 2015 *Virology* **479-480** 498–507 ISSN 10960341 URL http://linkinghub.elsevier.com/retrieve/pii/S004268221500183Xhttps://www.sciencedirect.com/science/article/pii/S004268221500183Xhttps://linkinghub.elsevier.com/retrieve/pii/S004268221500183X

[10] Riethmüller J, Riehle A, Grassmé H and Gulbins E 2006 *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1758** 2139–2147 ISSN 00052736 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273606002938

[11] Helle S C, Kanfer G, Kolar K, Lang A, Michel A H and Kornmann B 2013 *Biochimica et Biophysica Acta - Molecular Cell Research* **1833** 2526–2541 ISSN 01674889 URL https://linkinghub.elsevier.com/retrieve/pii/S0167488913000438

[12] Steinberg M S 2007 *Current Opinion in Genetics and Development* **17** 281–286 ISSN 0959437X URL https://linkinghub.elsevier.com/retrieve/pii/S0959437X07001062

[13] Guillot C and Lecuit T 2013 *Science* **340** 1185–1189 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1235249

[14] Fletcher A G, Osterfield M, Baker R E and Shvartsman S Y 2014 *Biophysical Journal* **106** 2291–2304 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349513057949

[15] Maître J L, Berthoumieux H, Krens S F G, Salbreux G, Jülicher F, Paluch E and Heisenberg C P 2012 *Science* **338** 253–256 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1225399

[16] Bazellières E, Conte V, Elosegui-Artola A, Serra-Picamal X, Bintanel-Morcillo M, Roca-Cusachs P, Muñoz J J, Sales-Pardo M, Guimerà R and Trepat X 2015 *Nature Cell Biology* **17** 409–420 ISSN 14764679 URL http://www.nature.com/articles/ncb3135

[17] Basiji D A, Ortyn W E, Liang L, Venkatachalam V and Morrissey P 2007 *Clinics in Laboratory Medicine* **27** 653–670 ISSN 02722712 URL https://linkinghub.elsevier.com/retrieve/pii/S0272271207000534

[18] Galbraith C G and Galbraith J A 2011 *Journal of Cell Science* **124** 1607–1611 ISSN 00219533 URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.080085

[19] Gomez G A, McLachlan R W, Wu S K, Caldwell B J, Moussa E, Verma S, Bastiani M, Priya R, Parton R G, Gaus K, Sap J and Yap A S 2015 *Molecular Biology of the Cell* **26** 1249–1262 ISSN 19394586 URL https://www.molbiolcell.org/doi/10.1091/mbc.E14-07-1223

[20] Truong Quang B A, Mani M, Markova O, Lecuit T and Lenne P F 2013 *Current Biology* **23** 2197–2207 ISSN 09609822 URL https://linkinghub.elsevier.com/retrieve/pii/S0960982213011317

[21] Wu E L, Cheng X, Jo S, Rui H, Song K C, Dávila-Contreras E M, Qi Y, Lee J, Monje-Galvan V, Venable R M, Klauda J B and Im W 2014 *Journal of Computational Chemistry* **35** 1997–2004 ISSN 1096987X URL http://doi.wiley.com/10.1002/jcc.23702

[22] Angres B, Barth A and Nelson W J 1996 *Journal of Cell Biology* **134** 549–557 ISSN 00219525 URL https://rupress.org/jcb/article/134/2/549/15330/Mechanism-for-transition-from-initial-to-stable

[23] Yap A S, Brieher W M, Pruschy M and Gumbiner B M 1997 *Current Biology* **7** 308–315 ISSN 09609822 URL https://linkinghub.elsevier.com/retrieve/pii/S0960982206001540

[24] Werner J 2019 *Introduction to Flow Cytometry* **45** 1–322 ISSN 0021-9746 URL http://jcp.bmj.com/cgi/doi/10.1136/jcp.45.3.275-d

[25] Elliott G 2009 *Combinatorial Chemistry & High Throughput Screening* **12** 849–859 ISSN 13862073 URL http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1386-2073&volume=12&issue=9&spage=849

[26] Okagaki L H, Strain A K, Nielsen J N, Charlier C, Baltes N J, Chrétien F, Heitman J H, Dromer F and Nielsen K N 2010 *PLoS Pathogens* **6** e1000953 ISSN 15537366 URL https://dx.plos.org/10.1371/journal.ppat.1000953

[27] Moreau H D, Lemaître F, Terriac E, Azar G, Piel M, Lennon-Dumenil A M and Bousso P 2012 *Immunity* **37** 351–363 ISSN 10747613 URL https://linkinghub.elsevier.com/retrieve/pii/S1074761312002270

[28] Chojnacki J and Müller B 2013 *Traffic* **14** 15–24 ISSN 13989219 URL http://doi.wiley.com/10.1111/tra.12006

[29] Van Engelenburg S B, Shtengel G, Sengupta P, Waki K, Jarnik M, Ablan S D, Freed E O, Hess H F and Lippincott-Schwartz J 2014 *Science* **343** 653–656 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1247786

[30] Chojnacki J, Staudt T, Glass B, Bingen P, Engelhardt J, Anders M, Schneider J, Müller B, Hell S W and Kräusslich H G 2012 *Science* **338** 524–528 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1226359

[31] PALADE G E 1953 *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* **1** 188–211 ISSN 00221554 URL http://journals.sagepub.com/doi/10.1177/1.4.188

[32] Buckley C D, Tan J, Anderson K L, Hanein D, Volkmann N, Weis W I, Nelson W J and Dunn A R 2014 *Science* **346** 1254211–1254211 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1254211

[33] Fernández-Busnadiego R, Saheki Y and De Camilli P 2015 *Proceedings of the National Academy of Sciences of the United States of America* **112** E2004–E2013 ISSN 10916490 URL http://www.pnas.org/lookup/doi/10.1073/pnas.1503191112

[34] Collado J and Fernández-Busnadiego R 2017 *Biochimica et Biophysica Acta - Molecular Cell Research* **1864** 1507–1512 ISSN 18792596 URL https://linkinghub.elsevier.com/retrieve/pii/S0167488917300630

[35] Risco C, De Castro I F, Sanz-Sánchez L, Narayan K, Grandinetti G and Subramaniam S 2014 *Annual Review of Virology* **1** 453–473 ISSN 23270578 URL http://www.annualreviews.org/doi/10.1146/annurev-virology-031413-085351

[36] Bykov Y S, Cortese M, Briggs J A and Bartenschlager R 2016 *FEBS Letters* **590** 1877–1895 ISSN 18733468 URL http://doi.wiley.com/10.1002/1873-3468.12153

[37] Kukulski W, Schorb M, Welsch S, Picco A, Kaksonen M and Briggs J A 2011 *Journal of Cell Biology* **192** 111–119 ISSN 00219525 URL https://rupress.org/jcb/article/192/1/111/36311/Correlated-fluorescence-and-3D-electron-microscopy

[38] Pronker M F, Lemstra S, Snijder J, Heck A J, Thies-Weesie D M, Pasterkamp R J and Janssen B J 2016 *Nature Communications* **7** 13584 ISSN 20411723 URL http://www.nature.com/articles/ncomms13584

[39] Yan R, Zhang Y, Li Y, Xia L, Guo Y and Zhou Q 2020 *Science* **367** 1444–1448 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.abb2762

[40] Harrison O J, Jin X, Hong S, Bahna F, Ahlsen G, Brasch J, Wu Y, Vendome J, Felsovalyi K, Hampton C M, Troyanovsky R B, Ben-Shaul A, Frank J, Troyanovsky S M, Shapiro L and Honig B 2011 *Structure* **19** 244–256 ISSN 09692126 URL https://linkinghub.elsevier.com/retrieve/pii/S0969212611000037

[41] Chavent M, Duncan A L and Sansom M S 2016 *Current Opinion in Structural Biology* **40** 8–16 ISSN 1879033X URL http://linkinghub.elsevier.com/retrieve/pii/S0959440X16300641http://www.ncbi.nlm.nih.gov/pubmed/27341016http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5404110https://linkinghub.elsevier.com/retrieve/pii/S0959440X16300641

[42] Hospital A, Goñi J R, Orozco M and Gelpí J L 2015 *Advances and Applications in Bioinformatics and Chemistry* **8** 37–47 ISSN 11786949 URL http://www.ncbi.nlm.nih.gov/pubmed/26604800http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4655909https://www.dovepress.com/molecular-dynamics-

simulations-advances-and-applications-peer-reviewed-article-AABC

[43] Stansfeld P J and Sansom M S 2011 *Structure* **19** 1562–1572 ISSN 09692126 URL http://www.ncbi.nlm.nih.gov/pubmed/22078556http://linkinghub.elsevier.com/retrieve/pii/S0969212611003364https://linkinghub.elsevier.com/retrieve/pii/S0969212611003364

[44] Reddy T and Sansom M S 2016 *Structure* **24** 375–382 ISSN 18784186 URL http://www.ncbi.nlm.nih.gov/pubmed/26833387http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4780862https://linkinghub.elsevier.com/retrieve/pii/S0969212616000046

[45] Koyama H, Okumura H, Ito A M, Otani T, Nakamura K, Kato K and Fujimori T 2019 *bioRxiv* 812198 URL http://biorxiv.org/content/early/2019/10/20/812198.abstract

[46] McCammon J A, Gelin B R and Karplus M 1977 *Nature* **267** 585–590 ISSN 00280836 URL http://www.ncbi.nlm.nih.gov/pubmed/301613http://www.nature.com/articles/267585a0

[47] Kutzner C, Grubmüller H, De Groot B L and Zachariae U 2011 *Biophysical Journal* **101** 809–817 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349511007065

[48] Wan S, Flower D R and Coveney P V 2008 *Molecular Immunology* **45** 1221–1230 ISSN 01615890 URL http://linkinghub.elsevier.com/retrieve/pii/S0161589007007705https://linkinghub.elsevier.com/retrieve/pii/S0161589007007705

[49] Marrink S J, Risselada H J, Yefimov S, Tieleman D P and De Vries A H 2007 *Journal of Physical Chemistry B* **111** 7812–7824 ISSN 15206106 URL //000247761600015http//pubs.acs.org/doi/pdfplus/10.1021/jp071097finternal-pdf://Marrink-2007-TheMARTINIforcefi-1884373298/Marrink-2007-TheMARTINIforcefi.pdfhttps://pubs.acs.org/doi/abs/10.1021/jp071097f?journalCode=jpcbfk&quickLinkVolume=111&quickL

[50] Hsu P C, Samsudin F, Shearer J and Khalid S 2017 *Journal of Physical Chemistry Letters* **8** 5513–5518 ISSN 19487185 URL https://pubs.acs.org/doi/10.1021/acs.jpclett.7b02432

[51] Kasson P M and Pande V S 2007 *PLoS Computational Biology* **3** e220 ISSN 1553-7358 URL http://dx.plos.org/10.1371/journal.pcbi.0030220https://dx.plos.org/10.1371/journal.pcbi.0030220

[52] Kasson P M, Lindahl E and Pande V S 2010 *PLoS Computational Biology* **6** 1–11 ISSN 1553734X URL http://dx.plos.org/10.1371/journal.pcbi.1000829https://dx.plos.org/10.1371/journal.pcbi.1000829

[53] Marrink S J and Mark A E 2003 *Journal of the American Chemical Society* **125** 11144–11145 ISSN 00027863 URL http://www.ncbi.nlm.nih.gov/pubmed/16220905

[54] Katsov K, Müller M and Schick M 2004 *Biophysical Journal* **87** 3277–3290 ISSN 00063495 URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15326031%5Cnhttp://ac.els-cdn.com/S0006349504737959/1-s2.0-S0006349504737959-main.pdf?_tid=1fa0057e-3b43-11e2-903f-00000aab0f27&acdnat=1354317205_c87132a86f00cd54

[55] Risselada H J, Bubnis G and Grubmuller H 2014 *Proceedings of the National Academy of Sciences* **111** 11043–11048 ISSN 0027-8424 URL http://www.ncbi.nlm.nih.gov/pubmed/25024174http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4121774http://www.pnas.org/cgi/doi/10.1073/pnas.1323221111

[56] Müller M, Katsov K and Schick M 2003 *Biophysical Journal* **85** 1611–1623 ISSN 00063495 URL http://www.ncbi.nlm.nih.gov/pubmed/12944277http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1303336http://www.ncbi.nlm.nih.gov/pubmed/12944277%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1303336https://linkinghub.elsevie

[57] Kawamoto S and Shinoda W 2014 *Soft Matter* **10** 3048 ISSN 1744-683X URL http://xlink.rsc.org/?DOI=c3sm52344f

[58] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Research* **28** 235–242 ISSN 03051048 URL http://nar.oxfordjournals.org/content/28/1/235https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235

[59] Krummenacher C, Supekar V M, Whitbeck J C, Lazear E, Connolly S A, Eisenberg R J, Cohen G H, Wiley D C and Carfí A 2005 *EMBO Journal* **24** 4144–4153 ISSN 02614189 URL http://www.ncbi.nlm.nih.gov/pubmed/16292345http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1356314http://www.ncbi.nlm.nih.gov/pubmed/16292345%

5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1356314http://emboj.embopress.org

[60] Chowdary T K, Cairns T M, Atanasiu D, Cohen G H, Eisenberg R J and Heldwein E E 2010 *Nature Structural and Molecular Biology* **17** 882–888 ISSN 15459993 URL http://www.nature.com/doifinder/10.1038/nsmb.1837http://www.nature.com/articles/nsmb.1837

[61] Zhang N, Yan J, Lu G, Guo Z, Fan Z, Wang J, Shi Y, Qi J and Gao G F 2011 *Nature Communications* **2** 577 ISSN 20411723 URL http://www.nature.com/doifinder/10.1038/ncomms1571http://dx.doi.org/10.1038/ncomms1571http://www.nature.com/articles/ncomms1571

[62] Šali A and Blundell T L 1993 *Journal of Molecular Biology* **234** 779–815 ISSN 00222836 URL http://www.ncbi.nlm.nih.gov/pubmed/8254673%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/8254673%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/8254673?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSumhttps://li

[63] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer F T, De Beer T A, Rempfer C, Bordoli L, Lepore R and Schwede T 2018 *Nucleic Acids Research* **46** W296–W303 ISSN 13624962 URL https://academic.oup.com/nar/article/46/W1/W296/5000024

[64] Valverde J R 2001 *Briefings in Bioinformatics* **2** 199–200 ISSN 1467-5463 URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/2.2.199

[65] Allen M P and Tildesley D J 2017 *Computer simulation of liquids: Second edition* vol 1 (Oxford University Press) ISBN 9780198803195 URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198803195.001.0001/oso-9780198803195

[66] Binder K, Horbach J, Kob W, Wolfgang P and Fathollah V 2004 *Journal of Physics Condensed Matter* **16** 66–66 ISSN 09538984 URL http://scitation.aip.org/content/aip/magazine/physicstoday/article/50/7/10.1063/1.881812%0Ahttp://physicstoday.scitation.org/doi/10.1063/1.881812http://physicstoday.scitation.org/doi/10.1063/1.881812

[67] Berendsen H J, van der Spoel D and van Drunen R 1995 *Computer Physics Communications* **91** 43–56 ISSN 00104655 URL https://linkinghub.elsevier.com/retrieve/pii/001046559500042E

[68] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B and Lindah E 2015 *SoftwareX* **1-2** 19–25 ISSN 23527110 URL https://linkinghub.elsevier.com/retrieve/pii/S2352711015000059

[69] Hess B, Bekker H, Berendsen H J and Fraaije J G 1997 *Journal of Computational Chemistry* **18** 1463–1472 ISSN 01928651 URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H

[70] Ryckaert J P, Ciccotti G and Berendsen H J 1977 *Journal of Computational Physics* **23** 327–341 ISSN 10902716 URL https://linkinghub.elsevier.com/retrieve/pii/0021999177900985

[71] Peliti L 2019 *Statistical Mechanics in a Nutshell* (Princeton University Press) ISBN 9781400839360 URL http://www.jstor.org/stable/10.2307/j.ctvcm4h83

[72] Pathria R K and Beale P D 2011 *Statistical Mechanics* (Elsevier) ISBN 9780123821881 URL https://linkinghub.elsevier.com/retrieve/pii/C20090623102

[73] Halgren T A 1992 *Journal of the American Chemical Society* **114** 7827–7843 ISSN 0002-7863 URL https://pubs.acs.org/doi/abs/10.1021/ja00046a032

[74] Desgranges C and Delhommelle J 2014 *The Journal of Chemical Physics* **140** 104109 ISSN 0021-9606 URL http://aip.scitation.org/doi/10.1063/1.4867498

[75] Lorentz H A 1881 *Annalen der Physik* ISSN 15213889

[76] Berthelot D 1898 *Comptes rendus hebdomadaires des séances de l'Académie des Sciences* **126** 1703–1855

[77] DELHOMMELLE J and MILLIÉ P 2001 *Molecular Physics* **99** 619–625 ISSN 0026-8976 URL http://www.tandfonline.com/doi/abs/10.1080/00268970010020041

[78] Song W, Rossky P J and Maroncelli M 2003 *Journal of Chemical Physics* ISSN 00219606

[79] Waldman M and Hagler A 1993 *Journal of Computational Chemistry* **14** 1077–1084 ISSN 0192-8651 URL http://doi.wiley.com/10.1002/jcc.540140909

[80] Kong C L 1973 *The Journal of Chemical Physics* **59** 2464–2467 ISSN 0021-9606 URL http://aip.scitation.org/doi/10.1063/1.1680358

[81] Lopes P E, Guvench O and Mackerell A D 2015 Current status of protein force fields for molecular dynamics simulations *Methods in Molecular Biology* vol 1215 pp 47–71 URL http://link.springer.com/10.1007/978-1-4939-1465-4_3

[82] Pluhackova K and Böckmann R A 2015 *Journal of Physics Condensed Matter* **27** 323103 ISSN 1361648X URL https://iopscience.iop.org/article/10.1088/0953-8984/27/32/323103

[83] Botan A, Favela-Rosales F, Fuchs P F, Javanainen M, Kanduč M, Kulig W, Lamberg A, Loison C, Lyubartsev A, Miettinen M S, Monticelli L, Määttä J, Ollila O H, Retegan M, Róg T, Santuz H and Tynkkynen J 2015 *Journal of Physical Chemistry B* **119** 15075–15088 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.5b04878

[84] Cino E A, Choy W Y and Karttunen M 2012 *Journal of Chemical Theory and Computation* **8** 2725–2740 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct300323g

[85] Lindorff-Larsen K, Maragakis P, Piana S, Eastwood M P, Dror R O and Shaw D E 2012 *PLoS ONE* **7** e32131 ISSN 19326203 URL https://dx.plos.org/10.1371/journal.pone.0032131

[86] Bayly C I, Merz K M, Ferguson D M, Cornell W D, Fox T, Caldwell J W, Kollman P A, Cieplak P, Gould I R and Spellmeyer D C 1995 *Journal of the American Chemical Society* **117** 5179–5197 ISSN 15205126 URL http://pubs.acs.org/doi/abs/10.1021/ja00124a002https://pubs.acs.org/doi/abs/10.1021/ja00124a002

[87] MacKerell A D, Bashford D, Bellott M, Dunbrack R L, Evanseck J D, Field M J, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F T, Mattos C, Michnick S, Ngo T, Nguyen D T, Prodhom B, Reiher W E, Roux B, Schlenkrich M, Smith J C, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D and Karplus M 1998 *Journal of Physical Chemistry B* **102** 3586–3616 ISSN 15206106 URL http://www.ncbi.nlm.nih.gov/pubmed/24889800http://pubs.acs.org/doi/10.1021/jp973084fhttps://pubs.acs.org/doi/10.1021/jp973084f

[88] Jorgensen W L and Tirado-Rives J 1988 *Journal of the American Chemical Society* **110** 1657–1666 ISSN 15205126 URL https://pubs.acs.org/doi/abs/10.1021/ja00214a001

[89] Klauda J B, Venable R M, Freites J A, O'Connor J W, Tobias D J, Mondragon-Ramirez C, Vorobyov I, MacKerell A D and Pastor R W 2010 *Journal of Physical Chemistry B* **114** 7830–7843 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp101759q

[90] Vanommeslaeghe K and Mackerell A D 2015 *Biochimica et Biophysica Acta - General Subjects* **1850** 861–871 ISSN 18728006 URL https://linkinghub.elsevier.com/retrieve/pii/S0304416514002736

[91] Venable R M, Brown F L and Pastor R W 2015 *Chemistry and Physics of Lipids* **192** 60–74 ISSN 18732941 URL https://linkinghub.elsevier.com/retrieve/pii/S0009308415300190

[92] Jo S, Kim T, Iyer V G and Im W 2008 *Journal of Computational Chemistry* **29** 1859–1865 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.20945

[93] Jo S, Cheng X, Lee J, Kim S, Park S J, Patel D S, Beaven A H, Lee K I, Rui H, Park S, Lee H S, Roux B, MacKerell A D, Klauda J B, Qi Y and Im W 2017 *Journal of Computational Chemistry* **38** 1114–1124 ISSN 1096987X URL http://doi.wiley.com/10.1002/jcc.24660

[94] Best R B, Zhu X, Shim J, Lopes P E, Mittal J, Feig M and MacKerell A D 2012 *Journal of Chemical Theory and Computation* **8** 3257–3273 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct300400x

[95] Venable R M, Sodt A J, Rogaski B, Rui H, Hatcher E, MacKerell A D, Pastor R W and Klauda J B 2014 *Biophysical Journal* **107** 134–145 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349514005724

[96] Leonard A N, Wang E, Monje-Galvan V and Klauda J B 2019 *Chemical Reviews* **119** 6227–6269 ISSN 15206890 URL https://pubs.acs.org/doi/10.1021/acs.chemrev.8b00384

[97] Dickson C J, Madej B D, Skjevik A, Betz R M, Teigen K, Gould I R and Walker R C 2014 *Journal of Chemical Theory and Computation* **10** 865–879 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct4010307

[98] Madej B D, Gould I R and Walker R C 2015 *Journal of Physical Chemistry B* **119** 12424–12435 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.5b04924

[99] Jämbeck J P and Lyubartsev A P 2012 *Journal of Chemical Theory and Computation* **8** 2938–2948 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct300342n

[100] Ermilova I and Lyubartsev A P 2016 *Journal of Physical Chemistry B* **120** 12826–12842 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b05422

[101] Schmid N, Eichenberger A P, Choutko A, Riniker S, Winger M, Mark A E and Van Gunsteren W F 2011 *European Biophysics Journal* **40** 843–856 ISSN 01757571 URL http://link.springer.com/10.1007/s00249-011-0700-9

[102] Chiu S W, Pandit S A, Scott H L and Jakobsson E 2009 *Journal of Physical Chemistry B* **113** 2748–2763 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp807056c

[103] Kukol A 2009 *Journal of Chemical Theory and Computation* **5** 615–626 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct8003468

[104] Poger D, Van Gunsteren W F and Mark A E 2010 *Journal of Computational Chemistry* **31** 1117–1125 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.21396

[105] Lopes P E, Roux B and MacKerell A D 2009 *Theoretical Chemistry Accounts* **124** 11–28 ISSN 1432881X URL http://link.springer.com/10.1007/s00214-009-0617-x

[106] Klesse G, Klesse G, Rao S, Tucker S J, Tucker S J and Sansom M S 2020 *Journal of the American Chemical Society* **142** 9415–9427 ISSN 15205126 URL https://pubs.acs.org/doi/10.1021/jacs.0c02394

[107] Inakollu V S, Geerke D P, Rowley C N and Yu H 2020 *Current Opinion in Structural Biology* **61** 182–190 ISSN 1879033X URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X19301514

[108] Huang J, Lopes P E, Roux B and MacKerell A D 2014 *Journal of Physical Chemistry Letters* **5** 3144–3150 ISSN 19487185 URL https://pubs.acs.org/doi/10.1021/jz501315h

[109] Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder J W and Ren P 2013 *Journal of Chemical Theory and Computation* **9** 4046–4063 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct4003702

[110] Laury M L, Wang L P, Pande V S, Head-Gordon T and Ponder J W 2015 *Journal of Physical Chemistry B* **119** 9423–9437 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp510896n

[111] Grossfield A, Ren P and Ponder J W 2003 *Journal of the American Chemical Society* **125** 15671–15682 ISSN 00027863 URL https://pubs.acs.org/doi/10.1021/ja037005r

[112] Chu H, Peng X, Li Y, Zhang Y, Min H and Li G 2018 *Molecular Physics* **116** 1037–1050 ISSN 13623028 URL https://www.tandfonline.com/doi/full/10.1080/00268976.2018.1436201

[113] Fritz D, Herbers C R, Kremer K and Van Der Vegt N F 2009 *Soft Matter* **5** 4556–4563 ISSN 1744683X URL http://xlink.rsc.org/?DOI=b911713j

[114] Schlick T, Collepardo-Guevara R, Halvorsen L A, Jung S and Xiao X 2011 *Quarterly Reviews of Biophysics* **44** 191–228 ISSN 00335835 URL http://www.ncbi.nlm.nih.gov/pubmed/21226976http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3700731http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3700731&tool=pmcentrez&rendertype=abstracthttps://www.cambridge.org/core/product/id

[115] Noid W G 2013 *Journal of Chemical Physics* **139** 090901 ISSN 00219606 URL http://aip.scitation.org/toc/jcp/139/9http://aip.scitation.org/doi/10.1063/1.4818908

[116] Mullinax J W and Noid W G 2009 *Journal of Chemical Physics* **131** 104110 ISSN 00219606 URL http://scitation.aip.org/content/aip/journal/jcp/131/10/10.1063/1.3220627

[117] Voth G a 2008 *Coarse-Graining of Condensed Phase and Biomolecular Systems* (CRC Press) ISBN 9780429137297 URL https://www.taylorfrancis.com/books/9781420059564http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/1420059556

[118] Nielsen S O, Lopez C F, Srinivas G and Klein M L 2004 *Journal of Physics Condensed Matter* **16** R481–R512 ISSN 09538984 URL http://stacks.iop.org/0953-8984/16/i=15/a=R03?key=crossref.4b8cb84e1aa73326460570b19153d778https://iopscience.iop.org/article/10.1088/0953-8984/16/15/R03

[119] Ayton G S, Blood P D and Voth G A 2007 *Biophysical Journal* **92** 3595–3602 ISSN 00063495 URL http://www.ncbi.nlm.nih.gov/pubmed/17325001http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1853153http://linkinghub.elsevier.com/retrieve/pii/S000634950771161Xhttps://linkinghub.elsevier.com/retrieve/pii/S000634950771161X

[120] Clementi C 2008 *Current Opinion in Structural Biology* **18** 10–15 ISSN 0959440X URL https://www.sciencedirect.com/science/article/pii/S0959440X07001753https://linkinghub.elsevier.com/retrieve/pii/S0959440X07001753

[121] Matysiak S and Clementi C 2008 *Archives of Biochemistry and Biophysics* **469** 29–33 ISSN 00039861 URL https://www.sciencedirect.

com/science/article/pii/S0003986107004262https://linkinghub.elsevier.com/retrieve/pii/S0003986107004262

[122] Peter C, Delle Site L and Kremer K 2008 *Soft Matter* **4** 859–869 ISSN 17446848 URL http://xlink.rsc.org/?DOI=b717324e

[123] Sherwood P, Brooks B R and Sansom M S 2008 *Current Opinion in Structural Biology* **18** 630–640 ISSN 0959440X URL https://www.sciencedirect.com/science/article/pii/S0959440X08001097https://linkinghub.elsevier.com/retrieve/pii/S0959440X08001097

[124] Murtola T, Karttunen M and Vattulainen I 2009 *Journal of Chemical Physics* **131** 055101 ISSN 00219606 URL http://scitation.aip.org/content/aip/journal/jcp/131/5/10.1063/1.3167405

[125] Peter C and Kremer K 2009 *Faraday Discussions* **144** 9–24 ISSN 13596640 URL http://xlink.rsc.org/?DOI=B919800H

[126] Trylska J 2010 *Journal of Physics Condensed Matter* **22** 453101 ISSN 09538984 URL http://stacks.iop.org/0953-8984/22/i=45/a=453101?key=crossref.6d4a3951516f24048bdec5cff8753733https://iopscience.iop.org/article/10.1088/0953-8984/22/45/453101

[127] Hyeon C and Thirumalai D 2011 *Nature Communications* **2** 487 ISSN 20411723 URL http://www.nature.com/articles/ncomms1481

[128] Kamerlin S C, Vicatos S, Dryga A and Warshel A 2011 *Annual Review of Physical Chemistry* **62** 41–64 ISSN 0066426X URL http://www.annualreviews.org/doi/10.1146/annurev-physchem-032210-103335

[129] Takada S 2012 *Current Opinion in Structural Biology* **22** 130–137 ISSN 0959440X URL https://www.sciencedirect.com/science/article/pii/S0959440X12000280https://linkinghub.elsevier.com/retrieve/pii/S0959440X12000280

[130] Riniker S, Allison J R and Van Gunsteren W F 2012 *Physical Chemistry Chemical Physics* **14** 12423–12430 ISSN 14639076 URL http://xlink.rsc.org/?DOI=c2cp40934h

[131] Saunders M G and Voth G A 2012 *Current Opinion in Structural Biology* **22** 144–150 ISSN 0959440X URL https://www.sciencedirect.com/science/article/pii/S0959440X1200005Xhttps://linkinghub.elsevier.com/retrieve/pii/S0959440X1200005X

[132] Zhang J and Muthukumar M 2009 *Journal of Chemical Physics* **130** 035102 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.3050295

[133] Dama J F, Sinitskiy A V, McCullagh M, Weare J, Roux B, Dinner A R and Voth G A 2013 *Journal of Chemical Theory and Computation* **9** 2466–2480 ISSN 15499618 URL http://pubs.acs.org/doi/10.1021/ct4000444https://pubs.acs.org/doi/10.1021/ct4000444

[134] Gay J G and Berne B J 1981 *The Journal of Chemical Physics* **74** 3316–3319 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.441483

[135] Babadi M, Everaers R and Ejtehadi M R 2006 *Journal of Chemical Physics* **124** 174708 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.2179075

[136] Tschöp W, Kremer K, Hahn O, Batoulis J and Bürger T 1998 *Acta Polymerica* **49** 75–79 ISSN 03237648 URL http://doi.wiley.com/10.1002/%28SICI%291521-4044%28199802%2949%3A2/3%3C75%3A%3AAID-APOL75%3E3.0.CO%3B2-5http://doi.wiley.com/10.1002/%28SICI%291521-4044%28199802%2949%3A2%2F3%3C75%3A%3AAID-APOL75%3E3.3.CO%3B2-X

[137] Canutescu A A, Shelenkov A A and Dunbrack R L 2003 *Protein Science* **12** 2001–2014 ISSN 09618368 URL http://doi.wiley.com/10.1110/ps.03154503

[138] Maupetit J, Gautier R and Tufféry P 2006 *Nucleic Acids Research* **34** W147–W151 ISSN 03051048 URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl289

[139] Rotkiewicz P and Skolnick J 2008 *Journal of Computational Chemistry* **29** 1460–1465 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.20906

[140] Gopal S M, Mukherjee S, Cheng Y M and Feig M 2010 *Proteins: Structure, Function and Bioinformatics* **78** 1266–1281 ISSN 08873585 URL http://doi.wiley.com/10.1002/prot.22645

[141] Maciejczyk M, Spasic A, Liwo A and Scheraga H A 2010 *Journal of Computational Chemistry* **31** 1644–1655 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.21448

[142] Ayton G S and Voth G A 2009 *Current Opinion in Structural Biology* **19** 138–144 ISSN 0959440X URL https://www.sciencedirect.com/science/article/pii/S0959440X09000359https://linkinghub.elsevier.com/retrieve/pii/S0959440X09000359

[143] Gohlke H and Thorpe M F 2006 *Biophysical Journal* **91** 2115–2120 ISSN 00063495 URL http://www.ncbi.nlm.nih.gov/pubmed/16815893http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1557556https://linkinghub.elsevier.com/retrieve/pii/S0006349506719269

[144] Zhang Z, Lu L, Noid W G, Krishna V, Pfaendtner J and Voth G A 2008 *Biophysical Journal* **95** 5073–5083 ISSN 15420086 URL http://www.ncbi.nlm.nih.gov/pubmed/18757560http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2586547https://linkinghub.elsevier.com/retrieve/pii/S0006349508789343

[145] Sinitskiy A V, Saunders M G and Voth G A 2012 *Journal of Physical Chemistry B* **116** 8363–8374 ISSN 15205207 URL http://pubs.acs.org/doi/10.1021/jp2108895https://pubs.acs.org/doi/10.1021/jp2108895

[146] Potestio R, Pontiggia F and Micheletti C 2009 *Biophysical Journal* **96** 4993–5002 ISSN 15420086 URL http://www.ncbi.nlm.nih.gov/pubmed/19527659http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2712024https://linkinghub.elsevier.com/retrieve/pii/S0006349509007814

[147] Harmandaris V A, Reith D, Van Der Vegt N F and Kremer K 2007 *Macromolecular Chemistry and Physics* **208** 2109–2120 ISSN 10221352 URL http://doi.wiley.com/10.1002/macp.200700245

[148] Arkhipov A, Freddolino P L and Schulten K 2006 *Structure* **14** 1767–1777 ISSN 09692126 URL http://www.sciencedirect.com/science/article/pii/S0969212606004023http://www.ncbi.nlm.nih.gov/pubmed/17161367https://ac.els-cdn.com/S0969212606004023/1-s2.0-S0969212606004023-main.pdf?_tid=c433ad84-e00c-11e7-84f2-00000aacb360&acdnat=1513173289_b521f3134

[149] Arkhipov A, Yin Y and Schulten K 2008 *Biophysical Journal* **95** 2806–2821 ISSN 15420086 URL http://www.ncbi.nlm.nih.gov/pubmed/18515394http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2527247https://linkinghub.elsevier.com/retrieve/pii/S0006349508784248

[150] Zhang Z and Voth G A 2010 *Journal of Chemical Theory and Computation* **6** 2990–3002 ISSN 15499618 URL http://pubs.acs.org/doi/abs/10.1021/ct100374ahttps://pubs.acs.org/doi/10.1021/ct100374a

[151] Brini E, Algaer E A, Ganguly P, Li C, Rodríguez-Ropero F and Van Der Vegt N F 2013 *Soft Matter* **9** 2108–2119 ISSN 17446848 URL http://xlink.rsc.org/?DOI=C2SM27201F

[152] Ingólfsson H I, Lopez C A, Uusitalo J J, de Jong D H, Gopal S M, Periole X and Marrink S J 2014 *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4** 225–248 ISSN 17590884 URL http://doi.wiley.com/10.1002/wcms.1169

[153] Smit B, Hilbers P A, Esselink K, Rupert L A, Van Os N M and Schlijper A G 1990 *Nature* **348** 624–625 ISSN 00280836 URL http://www.nature.com/articles/348624a0

[154] Goetz R and Lipowsky R 1998 *Journal of Chemical Physics* **108** 7397–7409 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.476160

[155] Bennun S V, Hoopes M I, Xing C and Faller R 2009 *Chemistry and Physics of Lipids* **159** 59–66 ISSN 00093084 URL https://linkinghub.elsevier.com/retrieve/pii/S0009308409000607

[156] Lyubartsev A P and Rabinovich A L 2011 *Soft Matter* **7** 25–39 ISSN 1744683X URL http://xlink.rsc.org/?DOI=C0SM00457J

[157] Cascella M and Vanni S 2016 Toward accurate coarse-graining approaches for protein and membrane simulations *Chemical Modelling* vol 12 pp 1–52 ISBN 9781782621157 URL http://ebook.rsc.org/?DOI=10.1039/9781782622703-00001

[158] Bradley R and Radhakrishnan R 2013 *Polymers* **5** 890–936 ISSN 20734360 URL http://www.mdpi.com/2073-4360/5/3/890

[159] Lopez C A, De Vries A H and Marrink S J 2013 *Scientific Reports* **3** 2071 ISSN 20452322 URL http://www.nature.com/articles/srep02071

[160] Monticelli L, Kandasamy S K, Periole X, Larson R G, Tieleman D P and Marrink S J 2008 *Journal of Chemical Theory and Computation* **4** 819–834 ISSN 15499618 URL http://pubs.acs.org/doi/abs/10.1021/ct700324x

[161] De Jong D H, Lopez C A and Marrink S J 2012 *Faraday Discussions* **161** 347–363 ISSN 13645498 URL http://xlink.rsc.org/?DOI=C2FD20086D

[162] López C A, Rzepiela A J, de Vries A H, Dijkhuizen L, Hünenberger P H and Marrink S J 2009 *Journal of Chemical Theory and Computation* **5** 3195–3210 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct900313w

[163] Uusitalo J J, Ingólfsson H I, Marrink S J and Faustino I 2017 *Biophysical Journal* **113** 246–256 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349517306124

[164] Uusitalo J J, Ingólfsson H I, Akhshi P, Tieleman D P and Marrink S J 2015 *Journal of Chemical Theory and Computation* **11** 3932–3945 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.5b00286

[165] Yesylevskyy S O, Schäfer L V, Sengupta D and Marrink S J 2010 *PLoS Computational Biology* **6** 1–17 ISSN 1553734X URL http://dx.plos.org/10.1371/journal.pcbi.1000810https://dx.plos.org/10.1371/journal.pcbi.1000810

[166] Shinoda W, Devane R and Klein M L 2010 *Journal of Physical Chemistry B* **114** 6836–6849 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp9107206

[167] Shinoda W, DeVane R and Klein M L 2012 *Current Opinion in Structural Biology* **22** 175–186 ISSN 0959440X URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X12000292

[168] Seo S and Shinoda W 2019 *Journal of Chemical Theory and Computation* **15** 762–774 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.8b00987

[169] Orsi M and Essex J W 2011 *PLoS ONE* **6** e28637 ISSN 19326203 URL https://dx.plos.org/10.1371/journal.pone.0028637

[170] Darré L, Machado M R, Brandner A F, González H C, Ferreira S and Pantano S 2015 *Journal of Chemical Theory and Computation* **11** 723–739 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct5007746

[171] Brandner A, Schüller A, Melo F and Pantano S 2018 *Biochemical and Biophysical Research Communications* **498** 319–326 ISSN 10902104 URL https://linkinghub.elsevier.com/retrieve/pii/S0006291X17318570

[172] Machado M R, Barrera E E, Klein F, Sónora M, Silva S and Pantano S 2019 *Journal of Chemical Theory and Computation* **15** 2719–2733 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.9b00006

[173] Bereau T and Deserno M 2009 *Journal of Chemical Physics* **130** 235106 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.3152842

[174] Wang Z J and Deserno M 2010 *Journal of Physical Chemistry B* **114** 11207–11220 ISSN 15205207 URL http://pubs.acs.org/doi/abs/10.1021/jp102543jhttps://pubs.acs.org/doi/10.1021/jp102543j

[175] Bereau T, Wang Z J and Deserno M 2014 *Journal of Chemical Physics* **140** 115101 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4867465

[176] Arnarez C, Uusitalo J J, Masman M F, Ingólfsson H I, De Jong D H, Melo M N, Periole X, De Vries A H and Marrink S J 2015 *Journal of Chemical Theory and Computation* **11** 260–275 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct500477k

[177] Berger O, Edholm O and Jähnig F 1997 *Biophysical Journal* **72** 2002–2013 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349597788453

[178] Cordomí A, Caltabiano G and Pardo L 2012 *Journal of Chemical Theory and Computation* **8** 948–958 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct200491c

[179] Tieleman D P, MacCallum J L, Ash W L, Kandt C, Xu Z and Monticelli L 2006 *Journal of Physics Condensed Matter* **18** S1221–S1234 ISSN 09538984 URL https://iopscience.iop.org/article/10.1088/0953-8984/18/28/S07

[180] Han Y, Ahn J, Concel J, Byeon I J L, Gronenborn A M, Yang J and Polenova T 2010 *Journal of the American Chemical Society* **132** 1976–1987 ISSN 00027863 URL http://pubs.acs.org/doi/abs/10.1021/ja908687khttp://www.ncbi.nlm.nih.gov/pubmed/20092249http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2829833https://pubs.acs.org/doi/10.1021/ja908687k

[181] Wassenaar T A, Ingólfsson H I, Prieß M, Marrink S J and Schäfer L V 2013 *Journal of Physical Chemistry B* **117** 3516–3530 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp311533p

[182] Bashford D and Case D A 2000 *Annual Review of Physical Chemistry* **51** 129–152 ISSN 0066426X URL http://www.annualreviews.org/doi/10.1146/annurev.physchem.51.1.129

[183] Im W, Feig M and Brooks C L 2003 *Biophysical Journal* **85** 2900–2918 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349503747122

[184] Tanizaki S and Feig M 2005 *Journal of Chemical Physics* **122** 124706 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.1865992

[185] MacHado M R, González H C and Pantano S 2017 *Journal of Chemical Theory and Computation* **13** 5106–5116 ISSN 15499626 URL http://www.ncbi.nlm.nih.gov/pubmed/28876928http://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00659https://pubs.acs.org/doi/10.1021/acs.jctc.7b00659

[186] Stansfeld P J, Hopkinson R, Ashcroft F M and Sansom M S 2009 *Biochemistry* **48** 10926–10933 ISSN 00062960 URL https://pubs.acs.org/doi/10.1021/bi9013193

[187] Marrink S J, De Vries A H and Mark A E 2004 *Journal of Physical Chemistry B* **108** 750–760 ISSN 15206106 URL http://pubs.acs.org/doi/abs/10.1021/jp036508ghttps://pubs.acs.org/doi/abs/10.1021/jp036508g?journalCode=jpcbfk&quickLinkVolume=108&quickLinkPage=750&selectedTab=citation&volume=108https://pubs.acs.org/doi/10.1021/jp036508g

[188] Schäfer L V, De Jong D H, Holt A, Rzepiela A J, De Vries A H, Poolman B, Killian J A and Marrink S J 2011 *Proceedings of the National Academy of Sciences of the United States of America* **108** 1343–1348 ISSN 00278424 URL http://www.pnas.org/cgi/doi/10.1073/pnas.1009362108

[189] Marrink S J and Tieleman D P 2013 *Chemical Society Reviews* **42** 6801–6822 ISSN 14604744 URL http://xlink.rsc.org/?DOI=c3cs60093ahttp://www.ncbi.nlm.nih.gov/pubmed/23708257

[190] Wassenaar T A, Ingólfsson H I, Böckmann R A, Tieleman D P and Marrink S J 2015 *Journal of Chemical Theory and Computation* **11** 2144–2155 ISSN 15499626 URL http://pubs.acs.org/doi/abs/10.1021/acs.jctc.5b00209https://pubs.acs.org/doi/10.1021/acs.jctc.5b00209

[191] Gu R X, Ingólfsson H I, De Vries A H, Marrink S J and Tieleman D P 2017 *Journal of Physical Chemistry B* **121** 3262–3275 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b07142

[192] Lee H and Pastor R W 2011 *Journal of Physical Chemistry B* **115** 7830–7837 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp2020148

[193] Grunewald F, Rossi G, De Vries A H, Marrink S J and Monticelli L 2018 *Journal of Physical Chemistry B* **122** 7436–7449 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.8b04760

[194] Dahlberg M 2007 *Journal of Physical Chemistry B* **111** 7194–7200 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp071954f

[195] Dahlberg M and Maliniak A 2010 *Journal of Chemical Theory and Computation* **6** 1638–1649 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct900654e

[196] Bulacu M, Périole X and Marrink S J 2012 *Biomacromolecules* **13** 196–205 ISSN 15257797 URL https://pubs.acs.org/doi/10.1021/bm201454j

[197] Ma H, Cummins D D, Edelstein N B, Gomez J, Khan A, Llewellyn M D, Picudella T, Willsey S R and Nangia S 2017 *Journal of Chemical Theory and Computation* **13** 811–824 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.6b00856

[198] Ma H, Irudayanathan F J, Jiang W and Nangia S 2015 *Journal of Physical Chemistry B* **119** 14668–14682 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.5b07122

[199] Van Oosten B and Harroun T A 2016 *Journal of Molecular Graphics and Modelling* **63** 125–133 ISSN 18734243 URL https://linkinghub.elsevier.com/retrieve/pii/S1093326315300899

[200] Hsu P C, Jefferies D and Khalid S 2016 *Journal of Physical Chemistry B* **120** 11170–11179 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b06615

[201] Melo M N, Ingólfsson H I and Marrink S J 2015 *Journal of Chemical Physics* **143** 243152 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4937783

[202] Reddy T, Shorthouse D, Parton D L, Jefferys E, Fowler P W, Chavent M, Baaden M and Sansom M S 2015 *Structure* **23** 584–597 ISSN 18784186 URL http://www.ncbi.nlm.nih.gov/pubmed/25703376http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4353694http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4353694&tool=pmcentrez&rendertype=abstracthttps://linkinghub.elsevier.com/retrieve/

[203] Stansfeld P J, Jefferys E E and Sansom M S 2013 *Structure* **21** 810–819 ISSN 09692126 URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3746155&tool=pmcentrez&rendertype=abstracthttps://linkinghub.elsevier.com/retrieve/pii/S0969212613000877

[204] Van Eerden F J, Van Den Berg T, Frederix P W, De Jong D H, Periole X and Marrink S J 2017 *Journal of Physical Chemistry B* **121** 3237–3249 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b06865

[205] Ingólfsson H I, Arnarez C, Periole X and Marrink S J 2016 *Journal of Cell Science* **129** URL http://jcs.biologists.org/content/129/2/257

[206] Yuan D, Liu Z, Kaindl J, Maeda S, Zhao J, Sun X, Xu J, Gmeiner P, Wang H W and Kobilka B K 2020 *Nature Chemical Biology* **16** 507–512 ISSN 15524469 URL http://www.nature.com/articles/s41589-020-0492-2

[207] Periole X, Cavalli M, Marrink S J and Ceruso M A 2009 *Journal of Chemical Theory and Computation* **5** 2531–2543 ISSN 15499618 URL http://pubs.acs.org/doi/abs/10.1021/ct9002114https://pubs.acs.org/doi/10.1021/ct9002114

[208] Petrov D and Zagrovic B 2014 *PLoS Computational Biology* **10** e1003638 ISSN 15537358 URL https://dx.plos.org/10.1371/journal.pcbi.1003638

[209] Best R B, Zheng W and Mittal J 2014 *Journal of Chemical Theory and Computation* **10** 5113–5124 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct500569b

[210] Henriques J, Cragnell C and Skepö M 2015 *Journal of Chemical Theory and Computation* **11** 3420–3431 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct501178z

[211] Miller M S, Lay W K and Elcock A H 2016 *Journal of Physical Chemistry B* **120** 8217–8229 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b01902

[212] Abriata L A and Dal Peraro M 2015 *Scientific Reports* **5** 10549 ISSN 20452322 URL http://www.nature.com/articles/srep10549

[213] Lay W K, Miller M S and Elcock A H 2016 *Journal of Chemical Theory and Computation* **12** 1401–1407 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.5b01136

[214] Schmalhorst P S, Deluweit F, Scherrers R, Heisenberg C P and Sikora M 2017 *Journal of Chemical Theory and Computation* **13** 5039–5053 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.7b00374

[215] Javanainen M, Martinez-Seara H and Vattulainen I 2017 *Scientific Reports* **7** 1143 ISSN 20452322 URL http://www.nature.com/articles/s41598-017-01247-9

[216] Stark A C, Andrews C T and Elcock A H 2013 *Journal of Chemical Theory and Computation* **9** 4176–4185 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct400008p

[217] Kapla J, Stevensson B and Maliniak A 2016 *Journal of Physical Chemistry B* **120** 9621–9631 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b06566

[218] Luo Y and Roux B 2010 *Journal of Physical Chemistry Letters* **1** 183–189 ISSN 19487185 URL https://pubs.acs.org/doi/10.1021/jz900079w

[219] Yoo J and Aksimentiev A 2012 *Journal of Physical Chemistry Letters* **3** 45–50 ISSN 19487185 URL https://pubs.acs.org/doi/10.1021/jz201501a

[220] Yoo J and Aksimentiev A 2016 *Journal of Chemical Theory and Computation* **12** 430–443 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.5b00967

[221] Ploetz E A and Smith P E 2011 *Physical Chemistry Chemical Physics* **13** 18154–18167 ISSN 14639076 URL http://xlink.rsc.org/?DOI=c1cp21883b

[222] Karunaweera S, Gee M B, Weerasinghe S and Smith P E 2012 *Journal of Chemical Theory and Computation* **8** 3493–3503 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct300079v

[223] Miller M S, Lay W K, Li S, Hacker W C, An J, Ren J and Elcock A H 2017 *Journal of Chemical Theory and Computation* **13** 1812–1826 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.6b01059

[224] McMillan W G and Mayer J E 1945 *The Journal of Chemical Physics* **13** 276–305 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.1724036

[225] Stigter D 1960 *Journal of Physical Chemistry* **64** 118–124 ISSN 00223654

[226] Tessier P M, Vandrey S D, Berger B W, Pazhianur R, Sandler S I and Lenhoff A M 2002 *Acta Crystallographica Section D: Biological Crystallography* **58** 1531–1535 ISSN 09074449 URL http://scripts.iucr.org/cgi-bin/paper?S0907444902012775

[227] George A, Chiang Y, Guo B, Arabshahi A, Cai Z and Wilson W W 1997 Second virial coefficient as predictor in protein crystal growth *Methods in Enzymology* vol 276 pp 100–110 URL https://linkinghub.elsevier.com/retrieve/pii/S007668799776052X

[228] Blanco M A, Sahin E, Robinson A S and Roberts C J 2013 *Journal of Physical Chemistry B* **117** 16013–16028 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp409300j

[229] Nikoubashman A, Mahynski N A, Capone B, Panagiotopoulos A Z and Likos C N 2015 *Journal of Chemical Physics* **143** 243108 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4931410

[230] Grünberger A, Lai P K, Blanco M A and Roberts C J 2013 *Journal of Physical Chemistry B* **117** 763–770 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp308234j

[231] Calero-Rubio C, Saluja A and Roberts C J 2016 *Journal of Physical Chemistry B* **120** 6592–6605 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b04907

[232] Lelimousin M, Limongelli V and Sansom M S 2016 *Journal of the American Chemical Society* **138** 10611–10622 ISSN 15205126 URL https://pubs.acs.org/doi/10.1021/jacs.6b05602

[233] Domański J, Hedger G, Best R B, Stansfeld P J and Sansom M S 2017 *Journal of Physical Chemistry B* **121** 3364–3375 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.6b08445

[234] Mori T, Miyashita N, Im W, Feig M and Sugita Y 2016 *Biochimica et Biophysica Acta - Biomembranes* **1858** 1635–1651 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273615004381

[235] Fathizadeh A and Elber R 2018 *Journal of Chemical Physics* **149** 072325 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.5027078

[236] Cardelli C, Barducci A and Procacci P 2018 *Biochimica et Biophysica Acta - Biomembranes* **1860** 1480–1488 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273618301317

[237] Nitschke N, Atkovska K and Hub J S 2016 *Journal of Chemical Physics* **145** 125101 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4963192

[238] Elber R 2017 *Quarterly Reviews of Biophysics* **50** e8 ISSN 14698994 URL https://www.cambridge.org/core/product/identifier/S0033583517000063/type/journal_article

[239] Salari R, Joseph T, Lohia R, Hénin J and Brannigan G 2018 *Journal of Chemical Theory and Computation* **14** 6560–6573 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.8b00447

[240] Darden T, York D and Pedersen L 1993 *The Journal of Chemical Physics* **98** 10089–10092 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.464397

[241] Tironi I G, Sperb R, Smith P E and Van Gunsteren W F 1995 *The Journal of Chemical Physics* **102** 5451–5459 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.469273

[242] Castro-Román F, Benz R W, White S H and Tobias D J 2006 *Journal of Physical Chemistry B* **110** 24157–24164 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp064746g

[243] Hünenberger P H and McCammon J A 1999 *Journal of Chemical Physics* **110** 1856–1872 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.477873

[244] Engle R D, Skeel R D and Drees M 2005 *Journal of Computational Physics* **206** 432–452 ISSN 00219991 URL https://linkinghub.elsevier.com/retrieve/pii/S0021999104005212

[245] Andersen H C 1980 *The Journal of Chemical Physics* **72** 2384–2393 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.439486

[246] Berendsen H J, Postma J P, Van Gunsteren W F, Dinola A and Haak J R 1984 *The Journal of Chemical Physics* **81** 3684–3690 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.448118

[247] Morishita T 2000 *Journal of Chemical Physics* **113** 2976–2982 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.1287333

[248] Braun E, Moosavi S M and Smit B 2018 *Journal of Chemical Theory and Computation* **14** 5262–5272 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.8b00446

[249] Nosé S 1984 *Molecular Physics* **52** 255–268 ISSN 13623028 URL http://www.tandfonline.com/doi/abs/10.1080/00268978400101201http://www.tandfonline.com/loi/tmph20%5Cnhttp://dx.doi.org/10.1080/00268978400101201%5Cnhttp://www.tandfonline.com/

[250] Parrinello M and Rahman A 1981 *Journal of Applied Physics* **52** 7182–7190 ISSN 00218979 URL http://aip.scitation.org/doi/10.1063/1.328693

[251] Brooks B R, Brooks C L, Mackerell A D, Nilsson L, Petrella R J, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner A R, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R W, Post C B, Pu J Z, Schaefer M, Tidor B, Venable R M, Woodcock H L, Wu X, Yang W, York D M and Karplus M 2009 *Journal of Computational Chemistry* **30** 1545–1614 ISSN 1096987X URL http://www.ncbi.nlm.nih.gov/pubmed/19444816http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2810661https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21287

[252] Salomon-Ferrer R, Case D A and Walker R C 2013 *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3** 198–210 ISSN 17590876 URL http://doi.wiley.com/10.1002/wcms.1121

[253] Case D A, Cheatham T E, Darden T, Gohlke H, Luo R, Merz K M, Onufriev A, Simmerling C, Wang B and Woods R J 2005 *Journal of Computational Chemistry* **26** 1668–1688 ISSN 01928651 URL http://www.ncbi.nlm.nih.gov/pubmed/16200636http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1989667http://doi.wiley.com/10.1002/jcc.20290

[254] Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kalé L and Schulten K 2005 *Journal of Computational Chemistry* **26** 1781–1802 ISSN 1096987X URL http://www.ncbi.nlm.nih.gov/pubmed/16222654http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2486339http://doi.wiley.com/10.1002/jcc.20289

[255] Eastman P, Swails J, Chodera J D, McGibbon R T, Zhao Y, Beauchamp K A, Wang L P, Simmonett A C, Harrigan M P, Stern C D, Wiewiora R P, Brooks B R and Pande V S 2017 *PLoS Computational Biology* **13** e1005659 ISSN 15537358 URL https://dx.plos.org/10.1371/journal.pcbi.1005659

[256] Plimpton S 1995 *Journal of Computational Physics* **117** 1–19 ISSN 00219991 URL https://linkinghub.elsevier.com/retrieve/pii/S002199918571039X

[257] Halverson J D, Brandes T, Lenz O, Arnold A, Bevc S, Starchenko V, Kremer K, Stuehn T and Reith D 2013 *Computer Physics Communications* **184** 1129–1149 ISSN 00104655 URL https://linkinghub.elsevier.com/retrieve/pii/S0010465512004006

[258] Shaw D E, Deneroff M M, Dror R O, Kuskin J S, Larson R H, Salmon J K, Young C, Batson B, Bowers K J, Chao J C, Eastwood M P, Gagliardo

J, Grossman J P, Ho C R, Lerardi D J, Kolossváry I, Klepeis J L, Layman T, McLeavey C, Moraes M A, Mueller R, Priest E C, Shan Y, Spengler J, Theobald M, Towles B and Wang S C 2008 *Communications of the ACM* **51** 91–97 ISSN 00010782 URL https://dl.acm.org/doi/10.1145/1364782.1364802

[259] Shaw D E, Grossman J P, Bank J A, Batson B, Butts J A, Chao J C, Deneroff M M, Dror R O, Even A, Fenton C H, Forte A, Gagliardo J, Gill G, Greskamp B, Ho C R, Ierardi D J, Iserovich L, Kuskin J S, Larson R H, Layman T, Lee L S, Lerer A K, Li C, Killebrew D, Mackenzie K M, Mok S Y H, Moraes M A, Mueller R, Nociolo L J, Peticolas J L, Quan T, Ramot D, Salmon J K, Scarpazza D P, Ben Schafer U, Siddique N, Snyder C W, Spengler J, Tang P T P, Theobald M, Toma H, Towles B, Vitale B, Wang S C and Young C 2014 Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer *International Conference for High Performance Computing, Networking, Storage and Analysis, SC* vol 2015-Janua (IEEE) pp 41–53 ISBN 978-1-4799-5500-8 ISSN 21674337 URL http://ieeexplore.ieee.org/document/7012191/

[260] Higashi Y, Kutchan T M and Smith T J 2011 *Journal of Biological Chemistry* **286** 6532–6541 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M110.168633

[261] Martí-Renom M A, Stuart A C, Fiser A, Sánchez R, Melo F and Šali A 2000 *Annual Review of Biophysics and Biomolecular Structure* **29** 291–325 ISSN 10568700 URL http://www.annualreviews.org/doi/10.1146/annurev.biophys.29.1.291

[262] Cavasotto C N and Phatak S S 2009 *Drug Discovery Today* **14** 676–683 ISSN 13596446 URL https://linkinghub.elsevier.com/retrieve/pii/S1359644609001469

[263] Bashford D 1997 An object-oriented programming suite for electrostatic effects in biological molecules: An experience report on the MEAD project *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 1343 pp 233–240 ISBN 354063827X URL http://link.springer.com/10.1007/3-540-63827-X_66

[264] Bas D C, Rogers D M and Jensen J H 2008 *Proteins: Structure, Function and Genetics* **73** 765–783 ISSN 08873585 URL http://doi.wiley.com/10.1002/prot.22102

[265] Alexov E G and Gunner M R 1997 *Biophysical Journal* **72** 2075–2093 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349597788519

[266] Alexov E G and Gunner M R 1999 *Biochemistry* **38** 8253–8270 ISSN 00062960 URL https://pubs.acs.org/doi/10.1021/bi982700a

[267] Madura J D, Briggs J M, Wade R C, Davis M E, Luty B A, Ilin A, Antosiewicz J, Gilson M K, Bagheri B, Scott L R and McCammon J A 1995 *Computer Physics Communications* **91** 57–95 ISSN 00104655 URL https://linkinghub.elsevier.com/retrieve/pii/001046559500043F

[268] Lee J, Cheng X, Swails J M, Yeom M S, Eastman P K, Lemkul J A, Wei S, Buckner J, Jeong J C, Qi Y, Jo S, Pande V S, Case D A, Brooks C L, MacKerell A D, Klauda J B and Im W 2016 *Journal of Chemical Theory and Computation* **12** 405–413 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.5b00935

[269] Ghahremanpour M M, Arab S S, Aghazadeh S B, Zhang J and Van Der Spoel D 2014 *Bioinformatics* **30** 439–441 ISSN 13674803 URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt680

[270] Knight C J and Hub J S 2015 *Bioinformatics* **31** 2897–2899 ISSN 14602059 URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv292

[271] Bovigny C, Tamò G, Lemmin T, Maïno N and Dal Peraro M 2015 *Journal of Chemical Information and Modeling* **55** 2491–2499 ISSN 15205142 URL https://pubs.acs.org/doi/10.1021/acs.jcim.5b00501

[272] Jo S, Lim J B, Klauda J B and Im W 2009 *Biophysical Journal* **97** 50–58 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349509007917

[273] Martinez L, Andrade R, Birgin E G and Martínez J M 2009 *Journal of Computational Chemistry* **30** 2157–2164 ISSN 1096987X URL http://doi.wiley.com/10.1002/jcc.21224

[274] Durrant J D and Amaro R E 2014 *PLoS Computational Biology* **10** e1003720 ISSN 15537358 URL https://dx.plos.org/10.1371/journal.pcbi.1003720

[275] Boyd K J, Alder N N and May E R 2018 *Biophysical Journal* **114** 2116–2127 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349518304363

[276] Martinez-Seara H and Róg T 2013 *Methods in Molecular Biology* **924** 407–429 ISSN 10643745

[277] Stansfeld P J, Goose J E, Caffrey M, Carpenter E P, Parker J L, Newstead S and Sansom M S 2015 *Structure* **23** 1350–1361 ISSN 18784186 URL http://www.sciencedirect.com/science/article/pii/S0969212615001847https://linkinghub.elsevier.com/retrieve/pii/S0969212615001847

[278] Koldsø H and Sansom M S 2015 *Journal of the American Chemical Society* **137** 14694–14704 ISSN 15205126 URL https://pubs.acs.org/doi/10.1021/jacs.5b08048

[279] Antipina A Y and Gurtovenko A A 2015 *Journal of Physical Chemistry B* **119** 6638–6645 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.5b01256

[280] Woolf T B and Roux B 1994 *Proceedings of the National Academy of Sciences of the United States of America* **91** 11631–11635 ISSN 00278424 URL http://www.pnas.org/cgi/doi/10.1073/pnas.91.24.11631

[281] Edholm O, Berger O and Jähnig F 1995 *Journal of Molecular Biology* **250** 94–111 ISSN 00222836 URL https://linkinghub.elsevier.com/retrieve/pii/S0022283685703610

[282] Shen L, Bassolino D and Stouch T 1997 *Biophysical Journal* **73** 3–20 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349597780421

[283] Faraldo-Gómez J D, Smith G R and Sansom M S 2002 *European Biophysics Journal* **31** 217–227 ISSN 01757571 URL http://link.springer.com/10.1007/s00249-002-0207-5

[284] Kandt C, Ash W L and Peter Tieleman D 2007 *Methods* **41** 475–488 ISSN 10462023 URL https://linkinghub.elsevier.com/retrieve/pii/S1046202306001873

[285] Schmidt T H and Kandt C 2012 *Journal of Chemical Information and Modeling* **52** 2657–2669 ISSN 15499596 URL https://pubs.acs.org/doi/10.1021/ci3000453

[286] Wolf M G, Hoefling M, Aponte-SantamaríA C, Grubmüller H and Groenhof G 2010 *Journal of Computational Chemistry* **31** 2169–2174 ISSN 1096987X URL http://doi.wiley.com/10.1002/jcc.21507

[287] Staritzbichler R, Anselmi C, Forrest L R and Faraldo-Gómez J D 2011 *Journal of Chemical Theory and Computation* **7** 1167–1176 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct100576m

[288] Javanainen M 2014 *Journal of Chemical Theory and Computation* **10** 2577–2582 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct500046e

[289] Jefferys E, Sands Z A, Shi J, Sansom M S and Fowler P W 2015 *Journal of Chemical Theory and Computation* **11** 2743–2754 ISSN 15499626 URL http://pubs.acs.org/doi/abs/10.1021/ct501111d#notes-1https://pubs.acs.org/doi/10.1021/ct501111d

[290] Onufriev A V and Izadi S 2018 *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8** e1347 ISSN 17590884 URL http://doi.wiley.com/10.1002/wcms.1347

[291] Lyman E, Hsieh C L and Eggeling C 2018 *Biophysical Journal* **115** 595–604 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349518308166

[292] Nickels J D, Smith J C and Cheng X 2015 *Chemistry and Physics of Lipids* **192** 87–99 ISSN 18732941 URL https://linkinghub.elsevier.com/retrieve/pii/S0009308415300177

[293] Heberle F A, Myles D A and Katsaras J 2015 *Chemistry and Physics of Lipids* **192** 41–50 ISSN 18732941 URL https://linkinghub.elsevier.com/retrieve/pii/S0009308415300256

[294] Sezgin E and Schwille P 2012 *Molecular Membrane Biology* **29** 144–154 ISSN 09687688 URL http://www.tandfonline.com/doi/full/10.3109/09687688.2012.700490

[295] Marty M T, Hoi K K and Robinson C V 2016 *Accounts of Chemical Research* **49** 2459–2467 ISSN 15204898 URL https://pubs.acs.org/doi/10.1021/acs.accounts.6b00379

[296] Deleu M, Crowet J M, Nasir M N and Lins L 2014 *Biochimica et Biophysica Acta - Biomembranes* **1838** 3171–3190 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273614003058

[297] Van Der Wel P C 2014 Lipid dynamics and protein-lipid interactions in integral membrane proteins: Insights from solid-state NMR *eMagRes* vol 3 (Chichester, UK: John Wiley & Sons, Ltd) pp 111–118 URL http://doi.wiley.com/10.1002/9780470034590.emrstm1356

[298] Maity P C, Yang J, Klaesener K and Reth M 2015 *Biochimica et Biophysica Acta - Molecular Cell Research* **1853** 830–840 ISSN 18792596 URL https://linkinghub.elsevier.com/retrieve/pii/S0167488914003991

[299] Lee H, De Vries A H, Marrink S J and Pastor R W 2009 *Journal of Physical Chemistry B* **113** 13186–13194 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp9058966

[300] Van Der Ploeg P and Berendsen H J 1982 *The Journal of Chemical Physics* **76** 3271–3276 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.443321

[301] Marrink S J and Berendsen H J 1996 *Journal of Physical Chemistry* **100** 16729–16738 ISSN 00223654 URL https://pubs.acs.org/doi/10.1021/jp952956f

[302] Edholm O and Nyberg A M 1992 *Biophysical Journal* **63** 1081–1089 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349592816788

[303] Damodaran K V, Merz K M and Gaber B P 1995 *Biophysical Journal* **69** 1299–1308 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349595799970

[304] Lindahl E and Edholm O 2000 *Biophysical Journal* **79** 426–433 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349500763041

[305] Marrink S J, Lindahl E, Edholm O and Mark A E 2001 *Journal of the American Chemical Society* **123** 8638–8639 ISSN 00027863 URL https://pubs.acs.org/doi/10.1021/ja0159618

[306] Risselad H J and Marrink S J 2008 *Proceedings of the National Academy of Sciences of the United States of America* **105** 17367–17372 ISSN 00278424 URL http://www.pnas.org/cgi/doi/10.1073/pnas.0807527105

[307] Tieleman D P, Marrink S J and Berendsen H J 1997 *Biochimica et Biophysica Acta - Reviews on Biomembranes* **1331** 235–270 ISSN 03044157 URL https://linkinghub.elsevier.com/retrieve/pii/S0304415797000087

[308] Pastor R W 1994 *Current Opinion in Structural Biology* **4** 486–492 ISSN 0959440X URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X94902097

[309] Marrink S J, de Vries A H and Tieleman D P 2009 *Biochimica et Biophysica Acta - Biomembranes* **1788** 149–168 ISSN 00052736 URL http://www.ncbi.nlm.nih.gov/pubmed/19013128http://linkinghub.elsevier.com/retrieve/pii/S0005273608003325https://linkinghub.elsevier.com/retrieve/pii/S0005273608003325

[310] Gumbart J, Wang Y, Aksimentiev A, Tajkhorshid E and Schulten K 2005 *Current Opinion in Structural Biology* **15** 423–431 ISSN 0959440X URL https://linkinghub.elsevier.com/retrieve/pii/S0959440X05001284

[311] Lindahl E and Sansom M S 2008 *Current Opinion in Structural Biology* **18** 425–431 ISSN 0959440X URL http://www.ncbi.nlm.nih.gov/pubmed/18406600http://linkinghub.elsevier.com/retrieve/pii/S0959440X08000304https://linkinghub.elsevier.com/retrieve/pii/S0959440X08000304

[312] Wang X and Deserno M 2015 *Journal of Chemical Physics* **143** 164109 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4933074

[313] Risselada H J and Marrink S J 2009 *Physical Chemistry Chemical Physics* **11** 2056–2067 ISSN 14639076 URL http://xlink.rsc.org/?DOI=b818782g

[314] Risselada H J, Kutzner C and Grubmüller H 2011 *ChemBioChem* **12** 1049–1055 ISSN 14394227 URL http://doi.wiley.com/10.1002/cbic.201100020

[315] Lin C M, Li C S, Sheng Y J, Wu D T and Tsao H K 2012 *Langmuir* **28** 689–700 ISSN 07437463 URL https://pubs.acs.org/doi/10.1021/la203755v

[316] Barragán Vidal I A, Rosetti C M, Pastorino C and Müller M 2014 *Journal of Chemical Physics* **141** 194902 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.4901203

[317] Harmandaris V A and Deserno M 2006 *Journal of Chemical Physics* **125** 204905 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.2372761

[318] Boyd K J, Alder N N and May E R 2017 *Langmuir* **33** 6937–6946 ISSN 15205827 URL https://pubs.acs.org/doi/10.1021/acs.langmuir.7b01185

[319] Elías-Wolff F, Lindén M, Lyubartsev A P and Brandt E G 2018 *Journal of Chemical Theory and Computation* **14** 1643–1655 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.7b00878

[320] Noguchi H 2011 *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **83** 061919 ISSN 15393755 URL https://link.aps.org/doi/10.1103/PhysRevE.83.061919

[321] Koldsø H, Shorthouse D, Hélie J and Sansom M S 2014 *PLoS Computational Biology* **10** e1003911 ISSN 15537358 URL https://dx.plos.org/10.1371/journal.pcbi.1003911

[322] Flinner N and Schleiff E 2015 *PLoS ONE* **10** e0133999 ISSN 19326203 URL https://dx.plos.org/10.1371/journal.pone.0133999

[323] Ingólfsson H I, Carpenter T S, Bhatia H, Bremer P T, Marrink S J and Lightstone F C 2017 *Biophysical Journal* **113** 2271–2280 ISSN 15420086 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349517311323

[324] Van Eerden F J, De Jong D H, De Vries A H, Wassenaar T A and Marrink S J 2015 *Biochimica et Biophysica Acta - Biomembranes* **1848** 1319–1330 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273615000656

[325] Guigas G and Weiss M 2016 *Biochimica et Biophysica Acta - Biomembranes* **1858** 2441–2450 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273615004277

[326] Corradi V, Mendez-Villuendas E, Ingólfsson H I, Gu R X, Siuda I, Melo M N, Moussatova A, Degagné L J, Sejdiu B I, Singh G, Wassenaar T A, Delgado Magnero K, Marrink S J and Tieleman D P 2018 *ACS Central Science* **4** 709–717 ISSN 23747951 URL https://pubs.acs.org/doi/10.1021/acscentsci.8b00143

[327] Holdbrook D A, Huber R G, Piggot T J, Bond P J and Khalid S 2016 *PLoS ONE* **11** e0156963 ISSN 19326203 URL https://dx.plos.org/10.1371/journal.pone.0156963

[328] Nguyen H D, Reddy V S and Brooks C L 2007 *Nano Letters* **7** 338–344 ISSN 15306984 URL http://pubs.acs.org/doi/abs/10.1021/nl062449hhttp://www.ncbi.nlm.nih.gov/pubmed/17297998https://pubs.acs.org/doi/abs/10.1021/nl062449h?journalCode=nalefd&quickLinkVolume=7&quickLinkPage=338&selectedTab=citation&volume=7https://pubs.acs.org/doi/10.1021/

[329] Grime J M, Dama J F, Ganser-Pornillos B K, Woodward C L, Jensen G J, Yeager M and Voth G A 2016 *Nature Communications* **7** 11568 ISSN 20411723 URL http://www.nature.com/doifinder/10.1038/ncomms11568http://www.nature.com/articles/ncomms11568

[330] Chen B and Tycko R 2011 *Biophysical Journal* **100** 3035–3044 ISSN 15420086 URL http://www.ncbi.nlm.nih.gov/pubmed/21689538http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3123931http://linkinghub.elsevier.com/

retrieve/pii/S0006349511005959https://linkinghub.elsevier.com/retrieve/pii/S0006349511005959

[331] Grime J M and Voth G A 2012 *Biophysical Journal* **103** 1774–1783 ISSN 00063495 URL http://www.sciencedirect.com/science/article/pii/S0006349512010193http://www.ncbi.nlm.nih.gov/pubmed/23083721http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3475334http://linkinghub.elsevier.com/retrieve/pii/S0006349512010193https://linki

[332] Johnson K N, Tang L, Johnson J E and Ball L A 2004 *Journal of Virology* **78** 11371–11378 ISSN 0022-538X URL http://www.ncbi.nlm.nih.gov/pubmed/15452258http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC521806http://jvi.asm.org/cgi/doi/10.1128/JVI.78.20.11371-11378.2004https://jvi.asm.org/content/78/20/11371

[333] Elrad O M and Hagan M F 2008 *Nano Letters* **8** 3850–3857 ISSN 15306984 URL http://pubs.acs.org/doi/abs/10.1021/nl802269ahttp://www.ncbi.nlm.nih.gov/pubmed/18950240http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2742690https://pubs.acs.org/doi/10.1021/nl802269a

[334] Hagan M F and Chandler D 2006 *Biophysical Journal* **91** 42–54 ISSN 00063495 URL http://linkinghub.elsevier.com/retrieve/pii/S0006349506717064http://www.ncbi.nlm.nih.gov/pubmed/16565055http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1479078http://www.sciencedirect.com/science/article/pii/S0006349506717064https://linki

[335] Perlmutter J D, Qiao C and Hagan M F 2013 *eLife* **2013** e00632 ISSN 2050084X URL https://elifesciences.org/articles/00632http://www.ncbi.nlm.nih.gov/pubmed/23795290http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3683802http://elifesciences.org/lookup/doi/10.7554/eLife.00632

[336] Wilber A W, Doye J P, Louis A A, Noya E G, Miller M A and Wong P 2007 *Journal of Chemical Physics* **127** 085106 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.2759922http://www.ncbi.nlm.nih.gov/pubmed/17764305

[337] Hicks S D and Henley C L 2006 *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **74** 031912 ISSN 15502376 URL https://link.aps.org/doi/10.1103/PhysRevE.74.031912

[338] Yu Z, Dobro M J, Woodward C L, Levandovsky A, Danielson C M, Sandrin V, Shi J, Aiken C, Zandi R, Hope T J and Jensen G J 2013 *Journal of Molecular Biology* **425** 112–123 ISSN 10898638 URL http://www.sciencedirect.com/science/article/pii/S0022283612008157http://www.sciencedirect.com/science/article/pii/S0022283612008157?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S0022283612008157

[339] Levandovsky A and Zandi R 2009 *Physical Review Letters* **102** 198102 ISSN 00319007 URL https://link.aps.org/doi/10.1103/PhysRevLett.102.198102

[340] Epand R M 2003 *Biochimica et Biophysica Acta - Biomembranes* **1614** 116–121 ISSN 00052736 URL http://linkinghub.elsevier.com/retrieve/pii/S000527360300169Xhttps://linkinghub.elsevier.com/retrieve/pii/S000527360300169X

[341] Larsson P and Kasson P M 2013 *PLoS Computational Biology* **9** e1002950 ISSN 1553734X URL http://dx.plos.org/10.1371/journal.pcbi.1002950https://dx.plos.org/10.1371/journal.pcbi.1002950

[342] Mirjanian D, Dickey A N, Hoh J H, Woolf T B and Stevens M J 2010 *The Journal of Physical Chemistry B* **114** 11061–11068 ISSN 1520-6106 URL http://pubs.acs.org/doi/abs/10.1021/jp1055182https://pubs.acs.org/doi/10.1021/jp1055182

[343] Bonnafous P and Stegmann T 2000 *Journal of Biological Chemistry* **275** 6160–6166 ISSN 00219258 URL http://www.ncbi.nlm.nih.gov/pubmed/10692407http://www.jbc.org/lookup/doi/10.1074/jbc.275.9.6160

[344] Donald J E, Zhang Y, Fiorin G, Carnevale V, Slochower D R, Gai F, Klein M L and DeGrado W F 2011 *Proceedings of the National Academy of Sciences of the United States of America* **108** 3958–3963 ISSN 00278424 URL http://www.pnas.org/cgi/doi/10.1073/pnas.1019668108http://www.ncbi.nlm.nih.gov/pubmed/21321234http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3054033

[345] Légaré S and Lagüe P 2014 *Proteins: Structure, Function and Bioinformatics* **82** 2118–2127 ISSN 10970134 URL http://doi.wiley.com/10.1002/prot.24568

[346] Victor B L, Lousa D, Antunes J M and Soares C M 2015 *Journal of Chemical Information and Modeling* **55** 795–805 ISSN 15205142 URL http://pubs.acs.org/doi/abs/10.1021/ci500756vhttps://pubs.acs.org/doi/10.1021/ci500756v

[347] Freddolino P L, Arkhipov A S, Larson S B, McPherson A and Schulten K 2006 *Structure* **14** 437–449 ISSN 09692126 URL http://www.ncbi.nlm.nih.gov/pubmed/16531228https://linkinghub.elsevier.com/retrieve/pii/S0969212606000608

[348] Zeng Y, Larson S B, Heitsch C E, McPherson A and Harvey S C 2012 *Journal of Structural Biology* **180** 110–116 ISSN 10478477 URL http://www.sciencedirect.com/science/article/pii/S1047847712001888?via%3Dihubhttps://linkinghub.elsevier.com/retrieve/pii/S1047847712001888

[349] Zink M and Grubmüller H 2010 *Biophysical Journal* **98** 687–695 ISSN 15420086 URL http://linkinghub.elsevier.com/retrieve/pii/S0006349509017202http://www.sciencedirect.com/science/article/pii/S0006349509017202https://linkinghub.elsevier.com/retrieve/pii/S0006349509017202

[350] Larsson D S, Liljas L and van der Spoel D 2012 *PLoS Computational Biology* **8** e1002502 ISSN 1553734X URL http://dx.plos.org/10.1371/journal.pcbi.1002502https://dx.plos.org/10.1371/journal.pcbi.1002502

[351] Marzinek J K, Holdbrook D A, Huber R G, Verma C and Bond P J 2016 *Structure* **24** 1410–1420 ISSN 18784186 URL http://www.ncbi.nlm.nih.gov/pubmed/27396828https://linkinghub.elsevier.com/retrieve/pii/S0969212616301216

[352] Harrison S C, Olson A J, Schutt C E, Winkler F K and Bricogne G 1978 *Nature* **276** 368–373 ISSN 00280836 URL http://www.nature.com/doifinder/10.1038/276368a0http://www.nature.com/articles/276368a0

[353] Huber R G, Marzinek J K, Holdbrook D A and Bond P J 2017 *Progress in Biophysics and Molecular Biology* **128** 121–132 ISSN 00796107 URL http://www.sciencedirect.com/science/article/pii/S0079610716300815#bib51https://linkinghub.elsevier.com/retrieve/pii/S0079610716300815

[354] Stone J E, Phillips J C, Freddolino P L, Hardy D J, Trabuco L G and Schulten K 2007 *Journal of Computational Chemistry* **28** 2618–2640 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.20829

[355] Mihai M E, Tecu C, Ivanciuc A E, Necula G, Lupulescu E and Onu A 2011 *Roumanian archives of microbiology and immunology* **70** 78–84 ISSN 12223891 URL http://www.ncbi.nlm.nih.gov/pubmed/22106514

[356] Stallknecht D E, Kearney M T, Shane S M and Zwank P J 1990 *Avian diseases* **34** 412–418 ISSN 00052086 URL

http://www.ncbi.nlm.nih.gov/pubmed/2142421https://www.jstor.org/stable/1591429?origin=crossref

[357] Barrera E E, Frigini E N, Porasso R D and Pantano S 2017 *Journal of Molecular Modeling* **23** 259 ISSN 09485023 URL http://link.springer.com/10.1007/s00894-017-3426-5

[358] Durrant J D, Kochanek S E, Casalino L, Ieong P U, Dommer A C and Amaro R E 2020 *ACS Central Science* **6** 189–196 ISSN 23747951 URL https://pubs.acs.org/doi/10.1021/acscentsci.9b01071

[359] De Jong D H, Singh G, Bennett W F, Arnarez C, Wassenaar T A, Schäfer L V, Periole X, Tieleman D P and Marrink S J 2013 *Journal of Chemical Theory and Computation* **9** 687–697 ISSN 15499618 URL http://pubs.acs.org/doi/10.1021/ct300646ghttps://pubs.acs.org/doi/10.1021/ct300646g

[360] Huang M L and Godula K 2016 *Glycobiology* **26** 797–803 ISSN 14602423 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cww022

[361] Bateman A 2019 *Nucleic Acids Research* **47** D506–D515 ISSN 13624962 URL https://academic.oup.com/nar/article/47/D1/D506/5160987

[362] Oliphant T E 2007 *Computing in Science and Engineering* **9** 10–20 ISSN 15219615 URL http://ieeexplore.ieee.org/document/4160250/

[363] Michaud-Agrawal N, Denning E J, Woolf T B and Beckstein O 2011 *Journal of Computational Chemistry* **32** 2319–2327 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.21787

[364] Hunter J D 2007 *Computing in Science and Engineering* **9** 90–95 ISSN 15219615 URL http://ieeexplore.ieee.org/document/4160265/

[365] Waskom M 2018 *Sphinx 1.7.4*

[366] Chakraborty A, Ghosh S, Mukhopadhyay P, Dinara S M, Bag A, Mahata M K, Kumar R, Das S, Sanjay J, Majumdar S and Biswas D 2014 *MRS Proceedings* **XXXIII** 81–87 ISSN 0717-6163 URL http://www.americanbanker.com/issues/179_124/which-city-is-the-next-big-fintech-hub-new-york-stakes-its-claim-1068345-1.html%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/15003161%5Cnhttp://cid.oxfordjournals.org/lookup/doi/10.1093/cid/cir991%5Cnhttp://www.scielo

[367] Kutzner C, Köpfer D A, Machtens J P, De Groot B L, Song C and Zachariae U 2016 *Biochimica et Biophysica Acta - Biomembranes* **1858** 1741–1752 ISSN 18792642 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273616300360

[368] Kanduč M, Schlaich A, De Vries A H, Jouhet J, Maréchal E, Demé B, Netz R R and Schneck E 2017 *Nature Communications* **8** 14899 ISSN 20411723 URL http://www.nature.com/articles/ncomms14899

[369] Sansom M S, Scott K A and Bond P J 2008 *Biochemical Society Transactions* **36** 27–32 ISSN 03005127 URL https://portlandpress.com/biochemsoctrans/article/36/1/27/83627/Coarsegrained-simulation-a-highthroughput

[370] Sharma D, Priyadarshini P and Vrati S 2015 *Journal of Virology* **89** 1489–1501 ISSN 0022-538X URL http://www.ncbi.nlm.nih.gov/pubmed/25428870http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4300767http://jvi.asm.org/lookup/doi/10.1128/JVI.02027-14https://jvi.asm.org/content/89/3/1489

[371] Martínez-Ballesta M d C and Carvajal M 2016 *Frontiers in Plant Science* **7** ISSN 1664462X URL http://journal.frontiersin.org/Article/10.3389/fpls.2016.01322/abstract

[372] D'Agostino M, Risselada H J and Mayer A 2016 *EMBO reports* **17** 1590–1608 ISSN 1469-221X URL http://embor.embopress.org/lookup/doi/10.15252/embr.201642209http://www.ncbi.nlm.nih.gov/pubmed/27644261https://onlinelibrary.wiley.com/doi/abs/10.15252/embr.201642209

[373] Andersen O S and Koeppe R E 2007 *Annual Review of Biophysics and Biomolecular Structure* **36** 107–130 ISSN 10568700 URL http://www.annualreviews.org/doi/10.1146/annurev.biophys.36.040306.132643

[374] Milter D 1964 *Mathematical functions* vol 32

[375] Smart O S, Neduvelil J G, Wang X, Wallace B A and Sansom M S 1996 *Journal of Molecular Graphics* **14** 354–360 ISSN 02637855 URL https://linkinghub.elsevier.com/retrieve/pii/S026378559700009X

[376] Tang H, Chang H, Dong Y, Guo L, Shi X, Wu Y, Huang Y and He Y 2018 *Proceedings of the National Academy of Sciences of the United States of America* **115** 9246–9251 ISSN 10916490 URL http://www.pnas.org/lookup/doi/10.1073/pnas.1801810115

[377] Yap A S, Michael M and Parton R G 2015 *F1000Research* **4** 273 ISSN 1759796X URL https://f1000research.com/articles/4-273/v1

[378] Elgass K D, Smith E A, LeGros M A, Larabell C A and Ryan M T 2015 *Journal of Cell Science* **128** 2795–2804 ISSN 14779137 URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.169136

[379] Buckle A M and Borg N A 2018 *Frontiers in immunology* **9** 2898 ISSN 16643224 URL https://www.frontiersin.org/article/10.3389/fimmu.2018.02898/full

[380] Boggon T J, Murray J, Chappuis-Flament S, Wong E, Gumbiner B M and Shapiro L 2002 *Science* **296** 1308–1313 ISSN 00368075 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1071559

[381] Koch A W, Manzur K L and Shan W 2004 *Cellular and Molecular Life Sciences* **61** 1884–1895 ISSN 1420682X URL http://link.springer.com/10.1007/s00018-004-4006-2

[382] Xu K, Tzvetkova-Robev D, Xu Y, Goldgur Y, Chan Y P, Himanen J P and Nikolov D B 2013 *Proceedings of the National Academy of Sciences of the United States of America* **110** 14634–14639 ISSN 10916490 URL http://www.pnas.org/lookup/doi/10.1073/pnas.1311000110

[383] Eisenberg R J, Atanasiu D, Cairns T M, Gallagher J R, Krummenacher C and Cohen G H 2012 *Viruses* **4** 800–832 ISSN 19994915 URL http://www.ncbi.nlm.nih.gov/pubmed/22754650http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3386629http://www.mdpi.com/1999-4915/4/5/800

[384] Carfí A, Willis S H, Whitbeck J C, Krummenacher C, Cohen G H, Eisenberg R J and Wiley D C 2001 *Molecular Cell* **8** 169–179 ISSN 10972765 URL https://linkinghub.elsevier.com/retrieve/pii/S1097276501002982

[385] Di Giovine P, Settembre E C, Bhargava A K, Luftig M A, Lou H, Cohen G H, Eisenberg R J, Krummenacher C and Carfi A 2011 *PLoS Pathogens* **7** e1002277 ISSN 1553-7374 URL http://dx.plos.org/10.1371/journal.ppat.1002277https://dx.plos.org/10.1371/journal.ppat.1002277

[386] Belardi B and Fletcher D A 2018 *Nature Reviews Molecular Cell Biology* **19** 747–748 ISSN 14710080 URL http://www.nature.com/articles/s41580-018-0057-3

[387] Bevers E M, Comfurius P, Dekkers D W and Zwaal R F 1999 *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* **1439** 317–330 ISSN 13881981 URL https://linkinghub.elsevier.com/retrieve/pii/S1388198199001109

[388] Varki A 2017 *Glycobiology* **27** 3–49 ISSN 14602423 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cww086

[389] Schengrund C L 2015 *Trends in Biochemical Sciences* **40** 397–406 ISSN 13624326 URL https://linkinghub.elsevier.com/retrieve/pii/S0968000415000481

[390] Gronnier J, Germain V, Gouguet P, Cacas J L and Mongrand S 2016 *Plant Signaling and Behavior* **11** e1152438 ISSN 15592324 URL http://www.tandfonline.com/doi/full/10.1080/15592324.2016.1152438

[391] Johnson J M, Tang J, Nyame Y, Willits D, Young M J and Zlotnick A 2005 *Nano Letters* **5** 765–770 ISSN 15306984 URL http://pubs.acs.org/doi/10.1021/nl050274qhttp://pubs.acs.org/doi/abs/10.1021/nl050274qhttps://pubs.acs.org/doi/10.1021/nl050274q

[392] Langer M D, Guo H, Shashikanth N, Pierce J M and Leckband D E 2012 *Journal of Cell Science* **125** 2478–2485 ISSN 00219533 URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.101147

[393] Moremen K W, Tiemeyer M and Nairn A V 2012 *Nature Reviews Molecular Cell Biology* **13** 448–462 ISSN 14710072 URL http://www.nature.com/articles/nrm3383

[394] Säwén E, Stevensson B, Östervall J, Maliniak A and Widmalm G 2011 *Journal of Physical Chemistry B* **115** 7109–7121 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/jp2017105

[395] Yamaguchi T, Sakae Y, Zhang Y, Yamamoto S, Okamoto Y and Kato K 2014 *Angewandte Chemie - International Edition* **53** 10941–10944 ISSN 15213773 URL http://doi.wiley.com/10.1002/anie.201406145

[396] Jo S, Qi Y and Im W 2015 *Glycobiology* **26** 19–29 ISSN 14602423 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cwv083

[397] Harbison A and Fadda E 2020 *Glycobiology* **30** 407–414 ISSN 14602423 URL https://academic.oup.com/glycob/article/30/6/407/5673732

[398] Hang I, Lin C W, Grant O C, Fleurkens S, Villiger T K, Soos M, Morbidelli M, Woods R J, Gauss R and Aebi M 2015 *Glycobiology* **25** 1335–1349 ISSN 14602423 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cwv058

[399] Park S J, Lee J, Qi Y, Kern N R, Lee H S, Jo S, Joung I, Joo K, Lee J and Im W 2019 *Glycobiology* **29** 320–331 ISSN 14602423 URL https://academic.oup.com/glycob/article/29/4/320/5301306

[400] Kirschner K N, Yongye A B, Tschampel S M, González-Outeiriño J, Daniels C R, Foley B L and Woods R J 2008 *Journal of Computational Chemistry* **29** 622–655 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.20820

[401] Tessier M B, DeMarco M L, Yongye A B and Woods R J 2008 *Molecular Simulation* **34** 349–364 ISSN 10290435 URL http://www.tandfonline.com/doi/abs/10.1080/08927020701710890

[402] Damm W, Frontera A, Tirado-Rives J and Jorgensen W L 1997 *Journal of Computational Chemistry* **18** 1955–1970 ISSN 01928651 URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199712)18:16%3C1955::AID-JCC1%3E3.0.CO;2-L

[403] Ott K H and Meyer B 1996 *Journal of Computational Chemistry* **17** 1068–1084 ISSN 01928651 URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199606)17:8%3C1068::AID-JCC14%3E3.0.CO;2-A

[404] Spieser S A, Van Kuik J A, Kroon-Batenburg L M and Kroon J 1999 *Carbohydrate Research* **322** 264–273 ISSN 00086215 URL https://linkinghub.elsevier.com/retrieve/pii/S0008621599002281

[405] Guvench O, Greenr S N, Kamath G, Brady J W, Venable R M, Pastor R W and Mackerell A D 2008 *Journal of Computational Chemistry* **29** 2543–2564 ISSN 01928651 URL http://doi.wiley.com/10.1002/jcc.21004

[406] Guvench O, Hatcher E, Venable R M, Pastor R W and MacKerell A D 2009 *Journal of Chemical Theory and Computation* **5** 2353–2370 ISSN 15499618 URL https://pubs.acs.org/doi/10.1021/ct900242e

[407] Feng T, Li M, Zhou J, Zhuang H, Chen F, Ye R, Campanella O and Fang Z 2015 *Innovative Food Science and Emerging Technologies* **31** 1–13 ISSN 14668564 URL https://linkinghub.elsevier.com/retrieve/pii/S1466856415001253

[408] Xiong X, Chen Z, Cossins B P, Xu Z, Shao Q, Ding K, Zhu W and Shi J 2015 *Carbohydrate Research* **401** 73–81 ISSN 1873426X URL https://linkinghub.elsevier.com/retrieve/pii/S0008621514004017

[409] Sauter J and Grafmüller A 2017 *Journal of Chemical Theory and Computation* **13** 223–236 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/acs.jctc.6b00613

[410] Patel R Y and Balaji P V 2007 *Biochimica et Biophysica Acta - Biomembranes* **1768** 1628–1640 ISSN 00052736 URL https://linkinghub.elsevier.com/retrieve/pii/S0005273607000697

[411] Jedlovszky P, Sega M and Vallauri R 2009 *Journal of Physical Chemistry B* **113** 4876–4886 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp808199p

[412] Shivgan A T, Marzinek J K, Huber R G, Krah A, Henchman R H, Matsudaira P, Verma C S and Bond P J 2020 *Journal of Chemical Information and Modeling* **60** 3864–3883 ISSN 15205142 URL http://biorxiv.org/content/early/2020/05/10/2020.05.08.085399.abstract

[413] Sauter J and Grafmüller A 2015 *Journal of Chemical Theory and Computation* **11** 1765–1774 ISSN 15499626 URL https://pubs.acs.org/doi/10.1021/ct500924f

[414] Varki A, Cummings R D, Aebi M, Packer N H, Seeberger P H, Esko J D, Stanley P, Hart G, Darvill A, Kinoshita T, Prestegard J J, Schnaar R L, Freeze H H, Marth J D, Bertozzi C R, Etzler M E, Frank M, Vliegenthart J F, Lütteke T, Perez S, Bolton E, Rudd P, Paulson J, Kanehisa M, Toukach P, Aoki-Kinoshita K F, Dell A, Narimatsu H, York W, Taniguchi N and Kornfeld S 2015 *Glycobiology* **25** 1323–1324 ISSN 14602423 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cwv091

[415] Sangster J 1989 *Journal of Physical and Chemical Reference Data* **18** 1111–1229 ISSN 15297845 URL http://aip.scitation.org/doi/10.1063/1.555833

[416] Mazzobre M F, Román M V, Mourelle A F and Corti H R 2005 *Carbohydrate Research* **340** 1207–1211 ISSN 00086215 URL https://linkinghub.elsevier.com/retrieve/pii/S0008621505000960

[417] Wishart D S, Feunang Y D, Guo A C, Lo E J, Marcu A, Grant J R, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, MacIejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C and Wilson M 2018 *Nucleic Acids Research* **46** D1074–D1082 ISSN 13624962 URL http://academic.oup.com/nar/article/46/D1/D1074/4602867

[418] Franks N P, Abraham M H and Lieb W R 1993 *Journal of Pharmaceutical Sciences* **82** 466–470 ISSN 15206017 URL https://linkinghub.elsevier.com/retrieve/pii/S0022354915491385

[419] Tiemeyer M, Aoki K, Paulson J, Cummings R D, York W S, Karlsson N G, Lisacek F, Packer N H, Campbell M P, Aoki N P, Fujita A, Matsubara M, Shinmachi D, Tsuchiya S, Yamada I, Pierce M, Ranzinger R, Narimatsu H and Aoki-Kinoshita K F 2017 *Glycobiology* **27** 915–919 ISSN 14602423 URL https://academic.oup.com/glycob/article/27/10/915/4079955

[420] Andrews S S, Addy N J, Brent R and Arkin A P 2010 *PLoS Computational Biology* **6** e1000705 ISSN 1553734X URL https://dx.plos.org/10.1371/journal.pcbi.1000705

[421] Crowe L M 2002 *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology* **131** 505–513 ISSN 10956433 URL https://linkinghub.elsevier.com/retrieve/pii/S1095643301005037

[422] Rebecchi L, Altiero T and Guidetti R 2007 *Invertebrate Survival Journal* **4** 65–81 ISSN 1824307X

[423] Crowe L M, Reid D S and Crowe J H 1996 *Biophysical Journal* **71** 2087–2093 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349596794079

[424] Crowe J H, Crowe L M, Oliver A E, Tsvetkova N, Wolkers W and Tablin F 2001 *Cryobiology* **43** 89–105 ISSN 00112240 URL https://linkinghub.elsevier.com/retrieve/pii/S0011224001923533

[425] Dashnau J L, Sharp K A and Vanderkooi J M 2005 *Journal of Physical Chemistry B* **109** 24152–24159 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp0543072

[426] Yancey P H, Clark M E, Hand S C, Bowlus R D and Somero G N 1982 *Science* **217** 1214–1222 ISSN 00368075 URL https://www.sciencemag.org/lookup/doi/10.1126/science.7112124

[427] Crowe J H, Crowe L M, Carpenter J F and Aurell Wistrom C 1987 *The Biochemical journal* **242** 1–10 ISSN 02646021 URL https://portlandpress.com/biochemj/article/242/1/1/22808/Stabilization-of-dry-phospholipid-bilayers-and

[428] Lenné T, Garvey C J, Koster K L and Bryant G 2009 *Journal of Physical Chemistry B* **113** 2486–2491 ISSN 15206106 URL https://pubs.acs.org/doi/10.1021/jp808670t

[429] Harrigan P R, Madden T D and Cullis P R 1990 *Chemistry and Physics of Lipids* **52** 139–149 ISSN 00093084 URL https://linkinghub.elsevier.com/retrieve/pii/000930849090157M

[430] Döbereiner H G, Selchow O and Lipowsky R 1999 *European Biophysics Journal* **28** 174–178 ISSN 01757571 URL http://link.springer.com/10.1007/s002490050197

[431] Genova J, Zheliaskova A and Mitov M D 2006 *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **282-283** 420–422 ISSN 09277757 URL https://linkinghub.elsevier.com/retrieve/pii/S0927775705009064

[432] Nagle J F, Jablin M S, Tristram-Nagle S and Akabori K 2015 *Chemistry and Physics of Lipids* **185** 3–10 ISSN 18732941 URL https://linkinghub.elsevier.com/retrieve/pii/S0009308414000474

[433] Strauss G, Schurtenberger P and Hauser H 1986 *BBA - Biomembranes* **858** 169–180 ISSN 00052736 URL https://linkinghub.elsevier.com/retrieve/pii/0005273686903032

[434] Crowe L M and Crowe J H 1991 *BBA - Biomembranes* **1064** 267–274 ISSN 00052736 URL https://linkinghub.elsevier.com/retrieve/pii/000527369190311U

[435] Moiset G, López C A, Bartelds R, Syga L, Rijpkema E, Cukkemane A, Baldus M, Poolman B and Marrink S J 2014 *Journal of the American Chemical Society* **136** 16167–16175 ISSN 15205126 URL https://pubs.acs.org/doi/10.1021/ja505476c

[436] Yamakawa T 1996 *Glycoconjugate Journal* **13** 123–126 ISSN 02820080 URL http://link.springer.com/10.1007/BF00731485

[437] Yu R K, Yanagisawa M and Ariga T 2007 Glycosphingolipid Structures *Comprehensive Glycoscience: From Chemistry to Systems Biology* vol 1-4 (Elsevier) pp 73–122 ISBN 9780444519672 URL https://linkinghub.elsevier.com/retrieve/pii/B9780444519672000039

[438] Ohmi Y, Tajima O, Ohkawa Y, Yamauchi Y, Sugiura Y, Furukawa K and Furukawa K 2011 *Journal of Neurochemistry* **116** 926–935 ISSN 00223042 URL http://doi.wiley.com/10.1111/j.1471-4159.2010.07067.x

[439] Rahmann H 1995 *Behavioural Brain Research* **66** 105–116 ISSN 01664328 URL https://linkinghub.elsevier.com/retrieve/pii/016643289400131X

[440] Palmano K, Rowan A, Guillermo R, Guan J and McJarrow P 2015 *Nutrients* **7** 3891–3913 ISSN 20726643 URL http://www.mdpi.com/2072-6643/7/5/3891

[441] Mocchetti I 2005 *Cellular and Molecular Life Sciences* **62** 2283–2294 ISSN 1420682X URL http://link.springer.com/10.1007/s00018-005-5188-y

[442] Ceccarelli B, Aporti F and Finesso M 1976 Effects of brain gangliosides on functional recovery in experimental regeneration and reinnervation. *Advances in experimental medicine and biology* vol 71 pp 275–293 URL http://link.springer.com/10.1007/978-1-4614-4614-9_17

[443] Yuki N 1997 *Journal of Infectious Diseases* **176** S150–S153 ISSN 00221899 URL https://academic.oup.com/jid/article-lookup/doi/10.1086/513800

[444] Vajn K, Viljetić B, Degmečić I V, Schnaar R L and Heffer M 2013 *PLoS ONE* **8** e75720 ISSN 19326203 URL https://dx.plos.org/10.1371/journal.pone.0075720

[445] Kanda T, Yoshino H, Ariga T, Yamawaki M and Yu R K 1994 *Journal of Cell Biology* **126** 235–246 ISSN 00219525 URL http://www.jcb.org/cgi/doi/10.1083/jcb.126.1.235

[446] Yuan C and Johnston L J 2001 *Biophysical Journal* **81** 1059–1069 ISSN 00063495 URL https://linkinghub.elsevier.com/retrieve/pii/S0006349501757633

[447] McIntosh T J and Simon S A 1994 *Biochemistry* **33** 10477–10486 ISSN 15204995 URL https://pubs.acs.org/doi/abs/10.1021/bi00200a032

[448] Owen M C, Karner A, Šachl R, Preiner J, Amaro M and Vácha R 2019 *Journal of Physical Chemistry B* **123** 7504–7517 ISSN 15205207 URL https://pubs.acs.org/doi/10.1021/acs.jpcb.9b05095

[449] Iijima K, Soga N, Matsubara T and Sato T 2009 *Journal of Colloid and Interface Science* **337** 369–374 ISSN 00219797 URL https://linkinghub.elsevier.com/retrieve/pii/S0021979709006626

[450] Matsubara T, Shibata R and Sato T 2016 *Frontiers in Microbiology* **7** ISSN 1664302X URL http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00468/abstract

[451] Mahoney M W and Jorgensen W L 2000 *Journal of Chemical Physics* **112** 8910–8922 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.481505

[452] Spiro R G 2002 *Glycobiology* **12** 43R–56R ISSN 09596658 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/12.4.43R

[453] Haltiwanger R S and Lowe J B 2004 *Annual Review of Biochemistry* **73** 491–537 ISSN 00664154 URL http://www.annualreviews.org/doi/10.1146/annurev.biochem.73.011303.074043

[454] Dennis J W, Lau K S, Demetriou M and Nabi I R 2009 *Traffic* **10** 1569–1578 ISSN 13989219 URL http://doi.wiley.com/10.1111/j.1600-0854.2009.00981.x

[455] Hoseki J, Ushioda R and Nagata K 2010 *Journal of Biochemistry* **147** 19–25 ISSN 0021924X URL https://academic.oup.com/jb/article-lookup/doi/10.1093/jb/mvp194

[456] Kollmann K, Pohl S, Marschner K, Encarnação M, Sakwa I, Tiede S, Poorthuis B J, Lübke T, Müller-Loennies S, Storch S and Braulke T 2010 *European Journal of Cell Biology* **89** 117–123 ISSN 01719335 URL https://linkinghub.elsevier.com/retrieve/pii/S0171933509003227

[457] Krogh A, Larsson B, Von Heijne G and Sonnhammer E L 2001 *Journal of Molecular Biology* **305** 567–580 ISSN 00222836 URL https://linkinghub.elsevier.com/retrieve/pii/S0022283600943158

[458] Sonnhammer E L, von Heijne G and Krogh A 1998 *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **6** 175–182 ISSN 15530833 URL http://www.ncbi.nlm.nih.gov/pubmed/9783223

[459] Lomize M A, Pogozheva I D, Joo H, Mosberg H I and Lomize A L 2012 *Nucleic Acids Research* **40** D370–D376 ISSN 03051048 URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr703

[460] Zhang X, Ge P, Yu X, Brannan J M, Bi G, Zhang Q, Schein S and Hong Zhou Z 2013 *Nature Structural and Molecular Biology* **20** 105–110 ISSN 15459993 URL http://www.nature.com/articles/nsmb.2463

[461] Wilson J S, Churchill-Angus A M, Davies S P, Sedelnikova S E, Tzokov S B, Rafferty J B, Bullough P A, Bisson C and Baker P J 2019 *Nature Communications* **10** 2900 ISSN 20411723 URL http://www.nature.com/articles/s41467-019-10777-x

[462] Bennett B C, Purdy M D, Baker K A, Acharya C, Mcintire W E, Stevens R C, Zhang Q, Harris A L, Abagyan R and Yeager M 2016 *Nature Communications* **7** 8770 ISSN 20411723 URL http://www.nature.com/articles/ncomms9770

[463] Zhou Y and MacKinnon R 2003 *Journal of Molecular Biology* **333** 965–975 ISSN 00222836 URL https://linkinghub.elsevier.com/retrieve/pii/S002228360301132X

[464] Schobert B, Cupp-Vickery J, Hornak V, Smith S O and Lanyi J K 2002 *Journal of Molecular Biology* **321** 715–726 ISSN 00222836 URL https://linkinghub.elsevier.com/retrieve/pii/S0022283602006812

[465] Jiang Y, Lee A, Chen J, Ruta V, Cadene M, Chait B T and MacKinnon R 2003 *Nature* **423** 33–41 ISSN 00280836 URL http://www.nature.com/articles/nature01580

[466] Suzuki H, Nishizawa T, Tani K, Yamazaki Y, Tamura A, Ishitani R, Dohmae N, Tsukita S, Nureki O and Fujiyoshi Y 2014 *Science* **344** 304–307 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1248571

[467] Zimmerman B, Kelly B, McMillan B J, Seegar T C, Dror R O, Kruse A C and Blacklow S C 2016 *Cell* **167** 1041–1051 ISSN 10974172 URL https://linkinghub.elsevier.com/retrieve/pii/S0092867416313800

[468] Huang W, Manglik A, Venkatakrishnan A J, Laeremans T, Feinberg E N, Sanborn A L, Kato H E, Livingston K E, Thorsen T S, Kling R C, Granier S, Gmeiner P, Husbands S M, Traynor J R, Weis W I, Steyaert J, Dror R O and Kobilka B K 2015 *Nature* **524** 315–321 ISSN 14764687 URL http://www.nature.com/articles/nature14886

[469] Savage D F and Stroud R M 2007 *Journal of Molecular Biology* **368** 607–617 ISSN 00222836 URL https://linkinghub.elsevier.com/retrieve/pii/S0022283607002628

[470] Czyzewski B K and Wang D N 2012 *Nature* **483** 494–497 ISSN 00280836 URL http://www.nature.com/articles/nature10881

[471] Craig L, Taylor R K, Pique M E, Adair B D, Arvai A S, Singh M, Lloyd S J, Shin D S, Getzoff E D, Yeager M, Forest K T and Tainer J A 2003 *Molecular Cell* **11** 1139–1150 ISSN 10972765 URL https://linkinghub.elsevier.com/retrieve/pii/S1097276503001709

[472] Stein A, Weber G, Wahl M C and Jahn R 2009 *Nature* **460** 525–528 ISSN 00280836 URL http://www.nature.com/articles/nature08156

[473] Cummings R D 2009 *Molecular BioSystems* **5** 1087–1104 ISSN 1742206X URL http://xlink.rsc.org/?DOI=b907931a

[474] Nairn A V, York W S, Harris K, Hall E M, Pierce J M and Moremen K W 2008 *Journal of Biological Chemistry* **283** 17298–17313 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M801964200

[475] Schachter H 2000 *Glycoconjugate Journal* **17** 465–483 ISSN 02820080

[476] Schachter H 1991 *Glycobiology* **1** 453–461 ISSN 09596658 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/1.5.453

[477] Zaia J 2010 *OMICS A Journal of Integrative Biology* **14** 401–418 ISSN 15362310 URL http://www.liebertpub.com/doi/10.1089/omi.2009.0146

[478] Kornfeld R and Kornfeld S 1985 *Annual Review of Biochemistry* **VOL. 54** 631–664 ISSN 00664154

[479] Watanabe Y, Allen J D, Wrapp D, McLellan J S and Crispin M 2020 *Science* **369** 330–333 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.abb9983

[480] Cao L, Diedrich J K, Kulp D W, Pauthner M, He L, Park S K R, Sok D, Su C Y, Delahunty C M, Menis S, Andrabi R, Guenaga J, Georgeson E, Kubitz M, Adachi Y, Burton D R, Schief W R, Yates J R and Paulson J C 2017 *Nature Communications* **8** 14954 ISSN 20411723 URL http://www.nature.com/articles/ncomms14954

[481] Cruz E, Cain J, Crossett B and Kayser V 2018 *Human Vaccines and Immunotherapeutics* **14** 508–517 ISSN 2164554X URL https://www.tandfonline.com/doi/full/10.1080/21645515.2017.1377871

[482] Bagdonaite I and Wandall H H 2018 *Glycobiology* **28** 443–467 ISSN 14602423 URL https://academic.oup.com/glycob/article/28/7/443/4951691

[483] Cipollo J F and Parsons L M 2020 *Mass Spectrometry Reviews* **39** 371–409 ISSN 10982787 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21629

[484] Schiller B, Hykollari A, Yan S, Paschinger K and Wilson I B 2012 *Biological Chemistry* **393** 661–673 ISSN 14316730 URL http://www.degruyter.com/view/j/bchm.2012.393.issue-8/hsz-2012-0150/hsz-2012-0150.xml

[485] Krasnova L and Wong C H 2016 *Annual Review of Biochemistry* **85** 599–630 ISSN 15454509 URL http://www.annualreviews.org/doi/10.1146/annurev-biochem-060614-034420

[486] Brockhausen I 2002 Glycodynamics of mucin biosynthesis in gastrointestinal tumor cells *Advances in Experimental Medicine and Biology* vol 535 pp 163–188 ISBN 00652598 (ISSN) URL http://link.springer.com/10.1007/978-1-4615-0065-0_11

[487] Bennett E P, Mandel U, Clausen H, Gerken T A, Fritz T A and Tabak L A 2012 *Glycobiology* **22** 736–756 ISSN 09596658 URL https://academic.oup.com/glycob/article-lookup/doi/10.1093/glycob/cwr182

[488] Tian E and Ten Hagen K G 2009 *Glycoconjugate Journal* **26** 325–334 ISSN 02820080 URL http://link.springer.com/10.1007/s10719-008-9162-4

[489] Luther K B and Haltiwanger R S 2009 *International Journal of Biochemistry and Cell Biology* **41** 1011–1024 ISSN 13572725 URL https://linkinghub.elsevier.com/retrieve/pii/S1357272508004093

[490] Doucey M A, Hess D, Cacan R and Hofsteenge J 1998 *Molecular Biology of the Cell* **9** 291–300 ISSN 10591524 URL https://www.molbiolcell.org/doi/10.1091/mbc.9.2.291

[491] Praissman J L and Wells L 2014 *Biochemistry* **53** 3066–3078 ISSN 15204995 URL https://pubs.acs.org/doi/10.1021/bi500153y

[492] Issoglio F M, Carrizo M E, Romero J M and Curtino J A 2012 *Journal of Biological Chemistry* **287** 1955–1961 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M111.287813

[493] Schegg B, Hülsmeier A J, Rutschmann C, Maag C and Hennet T 2009 *Molecular and Cellular Biology* **29** 943–952 ISSN 0270-7306 URL https://mcb.asm.org/content/29/4/943

[494] Esko J D and Selleck S B 2002 *Annual Review of Biochemistry* **71** 435–471 ISSN 00664154 URL http://www.annualreviews.org/doi/10.1146/annurev.biochem.71.110601.135458

[495] Hart G W, Slawson C, Ramirez-Correa G and Lagerlof O 2011 *Annual Review of Biochemistry* **80** 825–858 ISSN 00664154 URL http://www.annualreviews.org/doi/10.1146/annurev-biochem-060608-102511

[496] Quarles R H 2007 *Journal of Neurochemistry* **100** 1431–1448 ISSN 00223042 URL http://doi.wiley.com/10.1111/j.1471-4159.2006.04319.x

[497] Lopez P H H 2014 Role of Myelin-Associated Glycoprotein (Siglec-4a) in the Nervous System pp 245–262 URL http://link.springer.com/10.1007/978-1-4939-1154-7_11

[498] Trapp B D and Quarles R H 1982 *Journal of Cell Biology* **92** 877–882 ISSN 00219525 URL https://rupress.org/jcb/article/92/3/877/20679/Presence-of-the-myelinassociated-glycoprotein

[499] Trapp B D, Quarles R H and Suzuki K 1984 *Journal of Cell Biology* **99** 594–606 ISSN 00219525 URL https://rupress.org/jcb/article/99/2/594/56849/Immunocytochemical-studies-of-quaking-mice-support

[500] Li C, Tropak M B, Gerlai R, Clapoff S, Abramow-Newerly W, Trapp B, Peterson A and Roder J 1994 *Nature* **369** 747–750 ISSN 00280836 URL http://www.nature.com/articles/369747a0

[501] Pan B, Fromholt S E, Hess E J, Crawford T O, Griffin J W, Sheikh K A and Schnaar R L 2005 *Experimental Neurology* **195** 208–217 ISSN 00144886 URL https://linkinghub.elsevier.com/retrieve/pii/S0014488605001640

[502] Kelm S, Pelz A, Schauer R, Filbin M T, Tang S, Bellard M E d, Schnaar R L, Mahoney J A, Hartnell A, Bradfield P and Crocker P R 1994 *Current Biology* **4** 965–972 ISSN 09609822 URL https://linkinghub.elsevier.com/retrieve/pii/S0960982200002207

[503] Collins B E, Yang L J, Mukhopadhyay G, Filbin M T, Kiso M, Hasegawa A and Schnaar R L 1997 *Journal of Biological Chemistry* **272** 1248–1255 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.272.2.1248

[504] Vinson M, Strijbos P J, Rowles A, Facci L, Moore S E, Simmons D L and Walsh F S 2001 *Journal of Biological Chemistry* **276** 20280–20285 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M100345200

[505] Burger D, Pidoux L and Steck A J 1993 *Biochemical and Biophysical Research Communications* **197** 457–464 ISSN 0006291X URL https://linkinghub.elsevier.com/retrieve/pii/S0006291X83725015

[506] Sedzik J, Jastrzebski J P and Grandis M 2015 *Journal of Neuroscience Research* **93** 1–18 ISSN 10974547 URL http://doi.wiley.com/10.1002/jnr.23462

[507] Larsen A H, Wang Y, Bottaro S, Grudinin S, Arleth L and Lindorff-Larsen K 2020 *PLoS Computational Biology* **16** e1007870 ISSN 15537358 URL https://dx.plos.org/10.1371/journal.pcbi.1007870

[508] Towler P, Staker B, Prasad S G, Menon S, Tang J, Parsons T, Ryan D, Fisher M, Williams D, Dales N A, Patane M A and Pantoliano M W 2004 *Journal of Biological Chemistry* **279** 17996–18007 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M311191200

[509] Harrison S C 2008 *Nature Structural and Molecular Biology* **15** 690–698 ISSN 15459993 URL http://www.nature.com/articles/nsmb.1456

[510] Julien J P, Cupo A, Sok D, Stanfield R L, Lyumkis D, Deller M C, Klasse P J, Burton D R, Sanders R W, Moore J P, Ward A B and Wilson I A 2013 *Science* **342** 1477–1483 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1245625

[511] Shaik M M, Peng H, Lu J, Rits-Volloch S, Xu C, Liao M and Chen B 2019 *Nature* **565** 318–323 ISSN 14764687 URL http://www.nature.com/articles/s41586-018-0804-9

[512] Verhofstede C, Nijhuis M and Vandekerckhove L 2012 Correlation of coreceptor usage and disease progression

[513] Berger E A, Murphy P M and Farber J M 1999 *Annual Review of Immunology* **17** 657–700 ISSN 07320582 URL http://www.annualreviews.org/doi/10.1146/annurev.immunol.17.1.657

[514] Bannert N, Craig S, Farzan M, Sogah D, Santo N V, Choe H and Sodroski J 2001 *Journal of Experimental Medicine* **194** 1661–1673 ISSN 00221007 URL https://rupress.org/jem/article/194/11/1661/39090/Sialylated-OGlycans-and-Sulfated-Tyrosines-in-the

[515] Farzan M, Mirzabekov T, Kolchinsky P, Wyatt R, Cayabyab M, Gerard N P, Gerard C, Sodroski J and Choe H 1999 *Cell* **96** 667–676 ISSN 00928674 URL https://linkinghub.elsevier.com/retrieve/pii/S0092867400805772

[516] Chabot D J, Chen H, Dimitrov D S and Broder C C 2000 *Journal of Virology* **74** 4404–4413 ISSN 0022-538X URL https://jvi.asm.org/content/74/9/4404

[517] Farzan M, Babcock G J, Vasilieva N, Wright P L, Kiprilov E, Mirzabekov T and Choe H 2002 *Journal of Biological Chemistry* **277** 29484–29489 ISSN 00219258 URL http://www.jbc.org/lookup/doi/10.1074/jbc.M203361200

[518] Tan Q, Zhu Y, Li J, Chen Z, Han G W, Kufareva I, Li T, Ma L, Fenalti G, Li J, Zhang W, Xie X, Yang H, Jiang H, Cherezov V, Liu H, Stevens R C, Zhao Q and Wu B 2013 *Science* **341** 1387–1390 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.1241475

[519] Wu Z, Cui Q and Yethiraj A 2010 *Journal of Physical Chemistry B* **114** 10524–10529 ISSN 15205207 URL http://pubs.acs.org/doi/abs/10.1021/jp1019763https://pubs.acs.org/doi/10.1021/jp1019763

[520] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens T S, Herrler G, Wu N H, Nitsche A, Müller M A, Drosten C and Pöhlmann S 2020 *Cell* **181** 271–280 ISSN 10974172 URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420302294

[521] Chan K K, Dorosky D, Sharma P, Abbasi S A, Dye J M, Kranz D M, Herbert A S and Procko E 2020 *Science* **369** 1261–1265 ISSN 10959203 URL https://www.sciencemag.org/lookup/doi/10.1126/science.abc0870

[522] Kuhn R J, Zhang W, Rossmann M G, Pletnev S V, Corver J, Lenches E, Jones C T, Mukhopadhyay S, Chipman P R, Strauss E G, Baker T S and Strauss J H 2002 *Cell* **108** 717–725 ISSN 00928674 URL https://linkinghub.elsevier.com/retrieve/pii/S0092867402006608

[523] Allison S L, Schalich J, Stiasny K, Mandl C W and Heinz F X 2001 *Journal of Virology* **75** 4268–4275 ISSN 0022-538X URL https://jvi.asm.org/content/75/9/4268

[524] Huang C Y, Butrapet S, Moss K J, Childers T, Erb S M, Calvert A E, Silengo S J, Kinney R M, Blair C D and Roehrig J T 2010 *Virology* **396** 305–315 ISSN 00426822 URL https://linkinghub.elsevier.com/retrieve/pii/S0042682209006515

[525] Chambers T J, Hahn C S, Galler R and Rice C M 1990 *Annual Review of Microbiology* **44** 649–688 ISSN 00664227 URL http://www.annualreviews.org/doi/10.1146/annurev.mi.44.100190.003245

[526] Johnson A J, Guirakhoo F and Roehrig J T 1994 *Virology* **203** 241–249 ISSN 00426822 URL https://linkinghub.elsevier.com/retrieve/pii/S0042682284714814

[527] Hacker K, White L and de Silva A M 2009 *Journal of General Virology* **90** 2097–2106 ISSN 00221317 URL https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.012120-0

[528] Dejnirattisai W, Webb A I, Chan V, Jumnainsong A, Davidson A, Mongkolsapaya J and Screaton G 2011 *Journal of Infectious Diseases* **203** 1775–1783 ISSN 00221899 URL https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jir173

[529] Courageot M P, Frenkiel M P, Duarte Dos Santos C, Deubel V and Desprès P 2000 *Journal of Virology* **74** 564–572 ISSN 0022-538X URL https://jvi.asm.org/content/74/1/564

[530] Yap S S, Nguyen-Khuong T, Rudd P M and Alonso S 2017 *Frontiers in Microbiology* **8** ISSN 1664302X URL http://journal.frontiersin.org/article/10.3389/fmicb.2017.01415/full

[531] Fitzner D, Schneider A, Kippert A, Möbius W, Willig K I, Hell S W, Bunt G, Gaus K and Simons M 2006 *EMBO Journal* **25** 5037–5048 ISSN 02614189 URL http://emboj.embopress.org/cgi/doi/10.1038/sj.emboj.7601376

[532] Bussi G, Donadio D and Parrinello M 2007 *Journal of Chemical Physics* **126** 014101 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.2408420

[533] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *The Journal of Chemical Physics* **79** 926–935 ISSN 00219606 URL http://aip.scitation.org/doi/10.1063/1.445869

[534] Hoover W G 1985 *Physical Review A* **31** 1695–1697 ISSN 10502947 URL https://link.aps.org/doi/10.1103/PhysRevA.31.1695

[535] Darden T, Perera L, Li L and Lee P 1999 *Structure* **7** ISSN 09692126

Appendix A

STRUCTURES AND ABBREVIATIONS OF MONOSACCHARIDES DISCUSSED IN THE TEXT

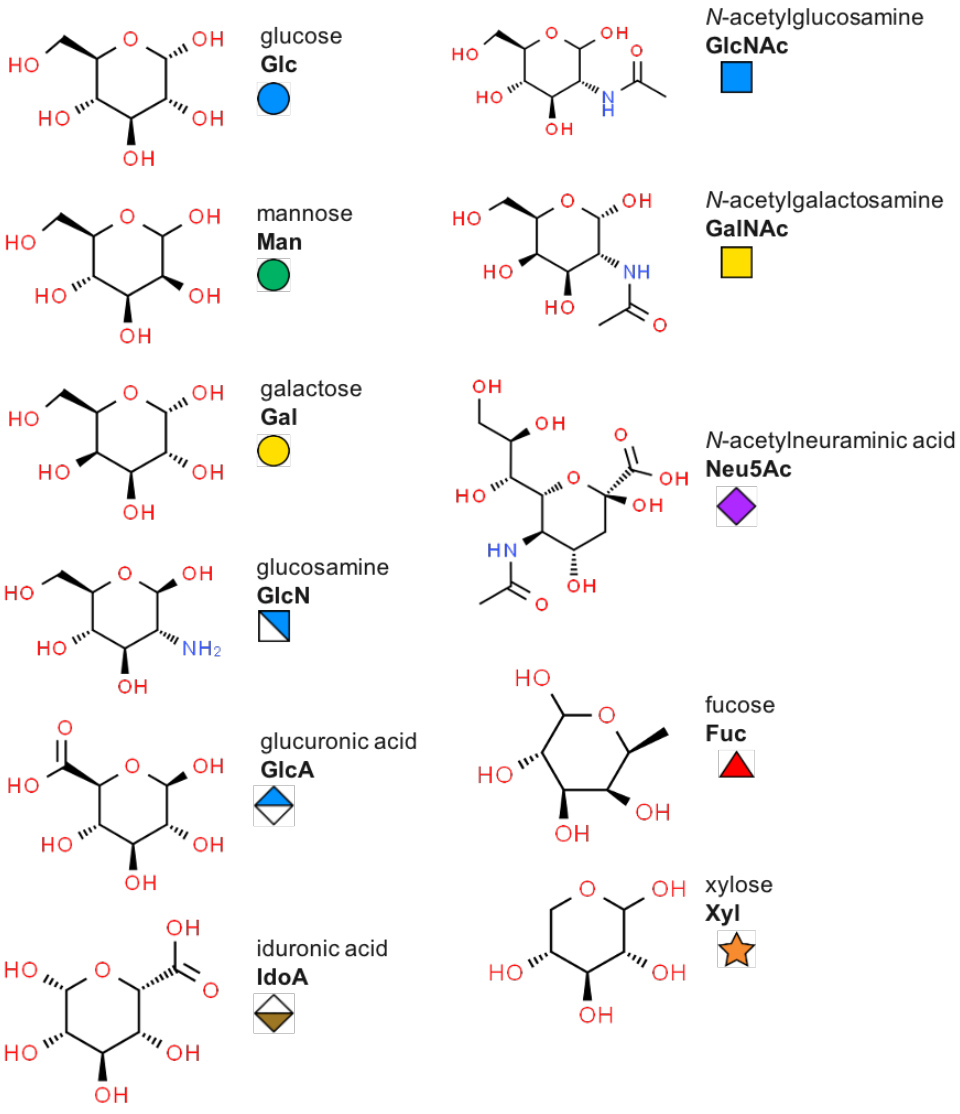


Figure A.1: Skeletal structures, name, abbreviation, and SNFG symbol for the monosaccharides most frequently discussed in this work.

Appendix B

STRUCTURES OF LIPIDS MOST OFTEN USED IN THIS WORK

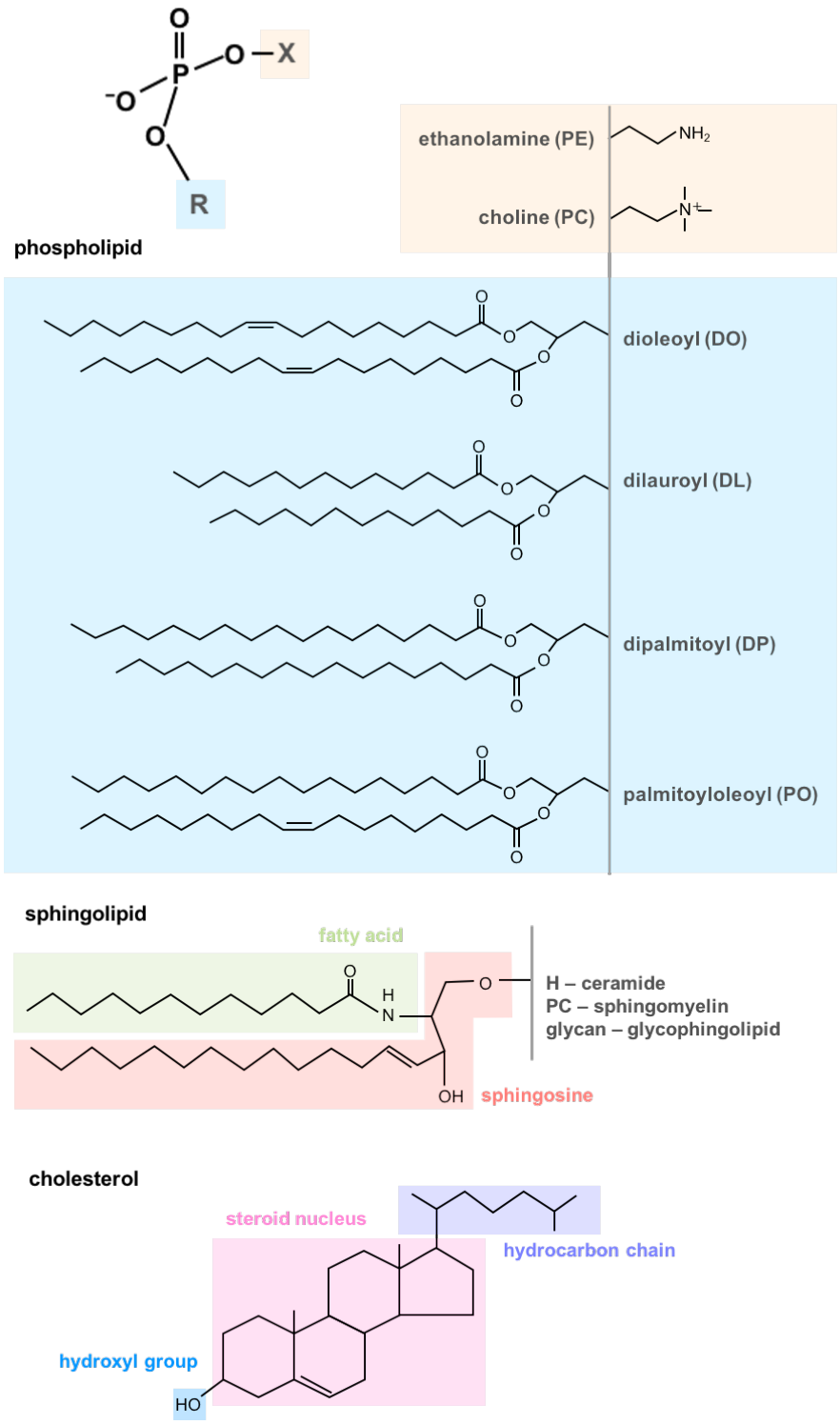


Figure B.1: Skeletal structures and overviews of the lipid types most often used in this work. Top: the phospholipids, showing the tails dipalmitoyl (DP), palmitoyloleoyl (PO), dilauroyl (DL), and DO (dioleoyl), and headgroups ethanolamine (PE) and choline (PC); these are used in the simulations as e.g. DPPC, POPE. Middle: sphingolipids, which are composed of a sphingosine and fatty acid linked to a headgroup to form e.g. ceramide (H), sphingomyelin (choline), or glycosphingolipids (glycans). Bottom: cholesterol, composed of a steroid core, hydrocarbon tail, and polar hydroxyl headgroup.

Appendix C

METHODOLOGICAL DETAILS

Here, I provide the technical details for the simulations and analyses discussedin this thesis. All the simulations were performed using GROMACS[311], and in most cases run locally. The only exceptions are the large bilayer simulations at the end of Chapter 3, which were run using the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk).

CGMD

For all CGMD simulations, I used the Martini force field[160, 187, 189, 359], either version 2.2 or the beta of version 3, as indicated. After the CG simulation system was constructed, it was subject to energy minimisation, equilibration, and production simulations as follows.

Energy minimisation Performed using 5000 steps of steepest descent.

Equilibration and production simulations In both the equilibration and production simulations, the temperature was coupled at 310 K using the velocity-rescale algorithm[532], with a relaxation time of 1.0 ps.

The pressure was coupled semiisotropically or isotropically (depending on whether or not a bilayer was present, respectively) at 1 atm using the Parrinello-Rahman barostat[359], with a compressibility of $3x10^{-4}$ ber^{-1} and relaxation time of 12.0 ps.

Nonbonded interactions were switched to zero from 0.9 to 1.1 nm, with a pair list update frequency of every 10 steps. Electrostatics shifted to zero from 0 nm to 1.1 nm, with a relative dielectric constant of 15.

The timestep was 20 fs for production simulations, and 10 fs for equilibrations, unless otherwise indicated in the text. The leap-frog integrator was used.

ATMD

All AT simulations were set up using CHARMM-GUI[92, 251, 268], and employed the CHARMM36m force field[360] with TIP3P[533] water.

After being constructed, the simulations were first energy minimised using 5000 steps steepest descent. They were then equilibrated for 125 000 steps at a 1 fs timestep, and the output from this used as the starting point for the production simulation with

a 2 fs timestep. In both equilibration and production simulations, the temperature was coupled using the Nosé -Hoover thermostat[249, 534] with relaxation time of 1.0 ps and a reference T 303.15 K. In the production simulations, pressure was coupled isotropically using the Parrinello-Rahman[250] barostat with relaxation time 5.0 ps and compressibility 4.5×10^{-5}.

All bonds to hydrogen atoms were constrained using the LINCS algorithm[69]. A short range cutoff of 1.2 nm was used for van der Waals interactions, with a force switch smoothing function applied from 1.0 to 1.2 nm. Electrostatics were treated using the PME algorithm[535], with a 1.2 nm real space cutoff.

Protein modelling

In order to model CG protein structures, I first obtained AT models for the relevant protein from the PDB`www.rcsb.org`[58]. Where these had missing regions such as short unresolved loops or no transmembrane helix, I modelled these using MODELLER[62] and the protein sequence taken from UniProt[361]. The completed AT protein was then converted to Martini 3 representation using Martinize[160, 359].

Data analysis

Analyses were performed using in-house proprietary code written in Python, employing primarily the modules NumPy[362] and MDAnalysis[363]. Visualisations and renderings were performed with Pymol[366]. The codes for glycosylation and the more complicated analyses are available on GitHub at `https://github.com/TheMissingMask`.

Specific details

Here I give specific details for individual sections where these deviate or include additional steps to those previously described.

logP calculations

The logP calculations were carried out as described in the main text, with most of the simulation parameters as for the other non-bilayer CGMD simulations. However, the equations of motion were integrated using the leap-frog stochastic dynamics integrator, and the interactions between the monosaccharide solute and the rest of the system were scaled according to a value of λ between 0 and 1.

Glycan A2

The A2 glycan was constructed manually using my monosaccharide models and translating them onto the corresponding residues in an AT structure of the glycan that I built

using the CHARMM-GUI server[92]. The molecule was added to a cubic simulation cell of dimensions 20 nm x 20 nm x 20 nm, solvated with CG water and 0.15 M NA and CL beads, and then energy minimised, equilibrated and simulated for 1 μs. The final conformation of A2 from this was used to produce systems with 5, 10, 50, or 100 A2 molecules in a 20 nm x 20 nm x 20 nm simulation cell using the gromacs insert-molecules tool. Each system was solvated and ions added as before, prior to energy minimisation, equilibration, and production simulation for 1 μs.

Gangliosides

The preliminary bilayers for the ganglioside simulations were constructed using *insane*[190] to construct an asymmetric bilayer comprising 144 POPC molecules in one leaflet, and 120 POPC and 20 GM3 molecules in the other. Using in-house code, I then replaced each of the Martini 2 GM3 molecules in this initial system with my Martini 3 ganglioside model by manually translating the tail groups of my ganglioside models to the tails of the Martini 2 GM3 molecules, and then deleting the Martini 2 GM3 from the system. As the number of beads in Martini 2 POPC and Martini 3 POPC are the same, I did not need to alter these in any way. I then solvated the bilayers with CG water and added neutralising monovalent (NA and CL) charged beads.

The system was then energy minimised, and subject to a short (1000 steps) Alchembed[289] run to remove any steric clashes that might have arisen when replacing the gangliosides. During this step, the ganglioside interactions with each other and the rest of the system were coupled to λ, which was increased from 0 to 1 over the simulation, thereby allowing the gangliosides to be 'phased' into existence. The system was equilibrated, and the output from this used as the starting point for the 1 μs production MD simulation.

Appendix D

MARTINI 3 MONOSACCHARIDE TRIALS

Each of the following tables gives the water-octanol partition energies calculated as described in Chapter 4 for different mappings tested for the monosaccharides. The mappings themselves are noted in the captions of the tables where they differ from those illustrated in Figures 4.6 and 4.7. All energies are in kJ mol^{-1}, with computed energies given as the energy ± standard deviation.

monosaccharide	mapping	water-octanol partition energy (kJ mol^{-1})	
		computed	predicted
HXS	SP3-SP3-SP1	18.0 ± 0.06	16.7
	SP3-SP3-SN3	17.3 ± 0.05	
	SP2-SP2-SP1	16.8 ± 0.07	
NHX	TP1(P2)-SP3-SP1	17.4 ± 0.06	18.4
	TP1(P2)-SP3-SN3	16.8 ± 0.05	
	TP1(P2)-SP2-SP1	16.9 ± 0.09	
	TP1(P3)-SP2-SP1	17.9 ± 0.06	
	TP1(P4)-SP2-SP1	19.8 ± 0.07	
HXN	SP3-SP3-SP1	17.7 ± 0.05	17.2
	SP3-SP3-SN3	17.3 ± 0.04	
	SP3-SP2-SP1	17.2 ± 0.03	
DHX	SP3-SP3-SN0a	13.9 ± 0.05	10.9
	SP3-SP3-SC6	13.2 ± 0.04	
	SP2-SP2-SN0a	12.6 ± 0.07	
	SP2-SP2-SN1a	14.0 ± 0.04	
	SP1-SP1-SN1a	10.5 ± 0.05	
XYL	SP3-SP3-TN0a	14.4 ± 0.03	14.1
	SP3-SP3-TC6	13.9 ± 0.03	
	SP2-SP2-TN0a	13.2 ± 0.03	
	SP2-SP2-TN2a	15.6 ± 0.04	
	SP1-SP1-TN1a	14.5 ± 0.05	
HXA	SP3-SP3-Q0	22.8 ± 0.05	14.9
	SP2-SP2-Q0	22.2 ± 0.04	
SIA	Qn-SN3(P2)-SP1(P3)	24.6 ± 0.05	20.7
	Qn-SN3(P2)-SN3(P3)	24.1 ± 0.09	
	Qn-SN2(P2)-SN3(P3)	23.8 ± 0.05	