university of groningen

GBB / Molecular Dynamics

"Sweet" Martini 3

Guidelines for a transferable sugar model in Martini 3

Mats Punt

S3206327

July 5, 2021

Supervisors

F. Grünewald, P. Vainikka

Examiner

prof. S.J. Marrink

"I always feel dumb... there is an uneasy feeling called confusion... and I'm like a monkey who is trying to put the two sticks together to get the banana"

— Richard Feynman

Acknowledgements

When writing a thesis or report of something you worked on for quite some time and poured quite some energy in, the effort is never yours alone and this thesis is no exception. Here's an attempt to thank the people involved:

Fabian - for helping out a clueless molecular biologist and lending me your time and experience. From not understanding to how to make an index file to slowly getting a glimpse of insight into what goes on in Martini and MD in general, it's been a fun ride is all I can say

Petteri - for the great help, plethora of (somehow always relevant) fun anecdotes, interesting conversations and for supplying us cold beverages on Fridays

Alex - for being an engaging teacher during my bachelor and bringing my attention to MD as a research technique

Siewert Jan - for running a cool group with interesting people and research subjects and for allowing me to work on the sugar project, which was quite the adventure to say the least

Paulo - for all the help in the background. Although I haven't gotten to know you very well, I still very much appreciate the time you took to look at our problems and the advice you gave in the Sugar Taskforce.

The MD group - for the nice chats and never-ending (!) supply of pie which were a welcome addition to the many months of quarantaine!

Abstract

Carbohydrates (saccharides) play a key role in a large number of biological processes. Precise experimental determination of structure-function relationships of carbohydrate systems through NMR or cryo-EM methods remains problematic. Molecular dynamics (MD) approaches, by use of force fields like Martini, can bridge this gap and help to study complex carbohydrate systems in detail. The original extension of the Martini force field to carbohydrates, released over ten years ago, succesfully reproduced complex properties of saccharides such as the transition between different cellulose allomorphs, the mechanical properties of bacterial peptidoglycan and β-cyclodextrin mediated cholesterol extraction. However, reports of unphysical aggregation behaviour in i.e. proteins and carbohydrates, colloquially referred to as the "sticky effect", raised questions on the accuracy of carbohydrate interactions in Martini 2. With the recent release of the Martini 3 force field, the underlying interaction matrix has been completely reparameterized, including more interaction levels and size dependent cross-interactions. In this work we propose mapping guidelines for carbohydrates in Martini 3, based on an extensive top down approach. We find good agreement with newly experimentally determined free energies of transfer, and experimental osmotic pressures, included to give a measure for aggregation propensity of the carbohydrate model. Retaining the building block principle of Martini, we improve on the extensively adapted Martini 2 carbohydrates with a more specific representation of chemical variety and glycosidic linkages and define clear mapping guidelines that can be applied to any carbohydrate of interest.

Contents

1 | Introduction

1.1 Scope

The widely adopted Martini 2 coarse-grained force field has had many successes in applications to lipids, proteins and carbohydrates. Overestimated aggregation propensities, as showcased in (membrane) proteins[1,2] and (complex) carbohydrates[3] were addressed in the recently published Martini 3 force field[4]. Martini already offers tools to readily study protein systems[5], but lacks a transferable, up-to-date sugar model that can be applied to e.g. glycosylated protein tails (glycans), an important part of cell regulation. By taking a measure for aggregation through osmotic pressures into account, we have parameterized an updated sugar model. This thesis aims to give an overview of the problems, strategies and solutions that have emerged while parameterizing such a Martini carbohydrate model using the newly released Martini 3 force field.

1.1.1 Carbohydrates

Carbohydrates (sugars) are a fundamental class of biomolecules and play a systematic role in cell metabolism[6], signalling pathways[7], in addition to acting as structural building blocks for many biological structures such as nucleic acids and cell walls[8]. The chemical diversity of carbohydrates allows for a large variety in which monosaccharides can be linked together to form complex, branched structures, commonly referred to as glycans. Present in glycosidic tails of proteins or extensions of the cellular membrane, the large heterogeneity of glycans form challenges for experimental methods to elucidate their structure and functioning. In X-ray crystallography, the main technique of solving protein structures, protein crystals can commonly only be formed after removal of their attached glycans. Other experimental approaches suffer from similar problems[9]. Recent advances in cryo-EM

methods[10] do show promising results but require expensive equipment and cannot be easily used in high-throughput assays. To tackle some of these problems computational approaches are becoming increasingly more popular. By developing models based on empirical criteria which approximate complex biological systems, some of the limitations of experimental methods can be circumvented. One such method, with broad applications, is that of molecular dynamics (MD)[11].

1.1.2 Carbohydrates and aromatic interactions

The hydrophilic character of carbohydrates promotes hydrogen bonding in the active site of a protein environment but another feature, attributable to the specific ring-like composition of carbohydrates, is their favoured interaction with aromatic residues[12]. Specific interactions between -CH groups of the carbohydrate and the π electron density of the aromatic lead to temporary (weak) dipoles that form stabilizing interactions through van der Waals and electrostatic interactions[13]. A result of the nature of the interaction is that it is highly dependent on geometry and orientation of the carbohydrate with respect to the aromatic residue. Analysis of carbohydrate libraries found a distinct preference for the axial orientation of C-H bonds of aromatic residues in carbohydrate binding sites[14]. Indeed, studies with carbohydrates that differ only in one axial and equatorial placement (e.g. the isomers β-glucose and β-galactose which only differ in the placement of the -OH group at the C4 carbon) measured a significant difference in aromatic orientation and binding affinity[15]. At least for negatively charged saccharides, favourable interaction is suggested to be dependent on the protonation state of the binding site residues[12], where either protonated or deprotonated residues affect the binding affinity. Studies on variations in carbohydrate length and substituted moieties find an important role for the overall shape and volume of the glycosidic moieties, as substitution potentially reduces binding affinity significantly[12,15]. Experimental measures of the enthalpy change for a carbohydrate interacting with a single aromatic residue amount to 1.5 - 2 kcal mol^{-1}, owing a significant contribution of binding affinity to CH-π interactions. The

highly conserved nature and large number of aromatic residues in carbohydrate binding sites underlines this significance[15] (see *Figure 1.1*).

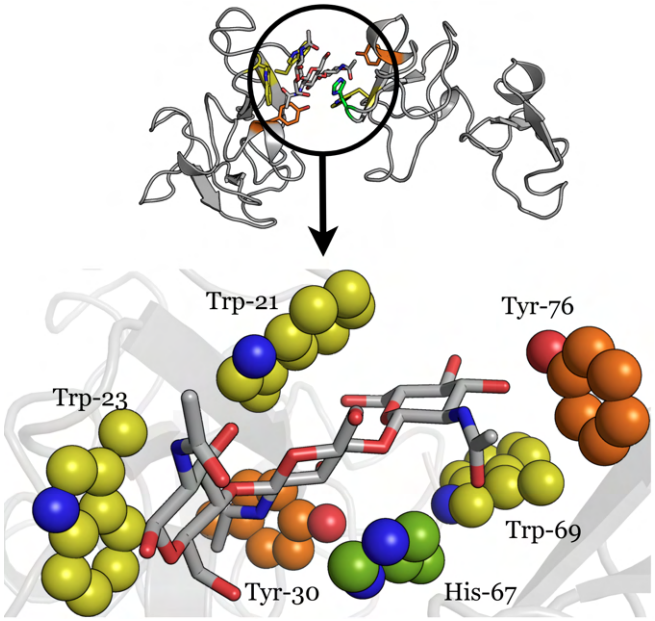

Figure 1.1: The aromatic residues in the binding site of lectin UDA-VI and its ligand (N,N′,N″-Triacetylchitotriose) are depicted. The high number of aromatics, commonly present in triads, participate in stacking with the sugar rings and contribute significantly to ligand recognition and binding affinity (PDB: 1EHH[16]).

1.2 Molecular Dynamics

1.2.1 Computational approaches and approximations

Biological systems are characterized by a high degree of complexity and need to be approximated to offer realistic insight at relevant time scales. In molecular dynamics (MD), particles are treated according to classical mechanics and electronic effects are modelled implicitly, neglecting bond formation and breakage. Via this approach, atomistic resolution and dynamics can easily be obtained at nanosecond (ns) timescales or greater, making detailed observations of complex systems possible. Even with limitations like the use of periodic boundary conditions (PBC) and sampling issues, MD has been shown to reproduce thermodynamic properties and large scale dynamics of a wide range of systems quite well[11]. Some examples of recent work include the study of high-throughput protein-ligand binding[17] or the

elucidation of glycoprotein-membrane structure-function relationships[18].

1.2.2 Force fields

Since the dynamics in MD are largely represented by classical mechanics, a ruleset for the atomistic interactions is required, which is defined in the force field. Bonded interactions are commonly captured by harmonic potentials (see *Section 7.2.2*) which are computationally efficient to solve[11]. The non-bonded interactions are captured by two components. The electrostatic effect is modelled by the Coulomb potential (*eq 1.1*)

$$V_{coulomb} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1} \cdot \frac{1}{r_{ij}} \tag{1.1}$$

where q_i and q_j are the charges of two particles, ε_0 and ε_1 the permitittivity of vacuum and the dielectric constant of the solvent respectively and r_{ij} the distance between two interacting particles: The Lennard-Jones (*eq 1.2*) potential

$$V_{LJ}(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.2}$$

features a repulsive part (r^{12} term) which models the overlap of electron clouds following the Pauli exclusion principle. The attractive part (r^6 term) aims to capture the dispersion force, where r_{ij} is the distance between two interacting particles, ε the depth of the potential well and σ_{ij} the distance at which the particle-particle potential energy V_{LJ} is zero:

Long ranged interactions are generally assumed to degrade to zero after a certain cutoff range which is also unique to a specific force field. The choice and parameterization of a force field largely depend on the objective and validation strategies employed[11]. The OPLS all-atom force field[19], for example, was specifically parameterized to reproduce experimental properties of liquids - such as density and heat of vaporization - making it not suitable for the study of protein dynamics.

1.2.3 Ensemble and sampling

The particles in an MD simulation are subject to the laws of statistical thermodynamics. In a "real" system, excess energy is dissipated to the environment, but in a computer simulation artificial methods need to be introduced to deal with this physical process. Through inclusion of thermostats and barostats, which maintain temperature and pressure respectively, experimental conditions can be closely approximated. The choice of environmental conditions (e.g. constant pressure, constant volume or constant temperature), reflected in the ensemble, is connected to the sampling of the simulation. Since ensemble averages, which are dependent on the position and momentum/force on a particle, are taken as averages over time, the question is thus when sufficient sampling has occurred. The larger the number of degrees of freedom in the system, the harder this problem becomes.

Although developments in computer hardware and algorithm efficiency have drastically improved the performance of molecular dynamics simulations[11], simulations that treat all atoms (AA) explicitly are inherently limited to nanosecond (ns) - microsecond (μs) ranges, while many relevant biological processes only unfold at larger timescales. Techniques to improve the sampling efficiency and to speed up simulations are thus crucial to obtain large scale dynamics.

1.3 Coarse-graining

1.3.1 Advantages of coarse-graining

A widely applied approach to improve the sampling speed and efficiency in MD simulations is to artifically reduce the degrees of freedom through coarse-graining (CG). In coarse-graining, multiple atoms in a molecule may be represented by a single, larger pseudo-atom (often referred to as a "bead"), with the goal of conserving the topology of the molecule in the best way possible (see *Figure 1.2*).

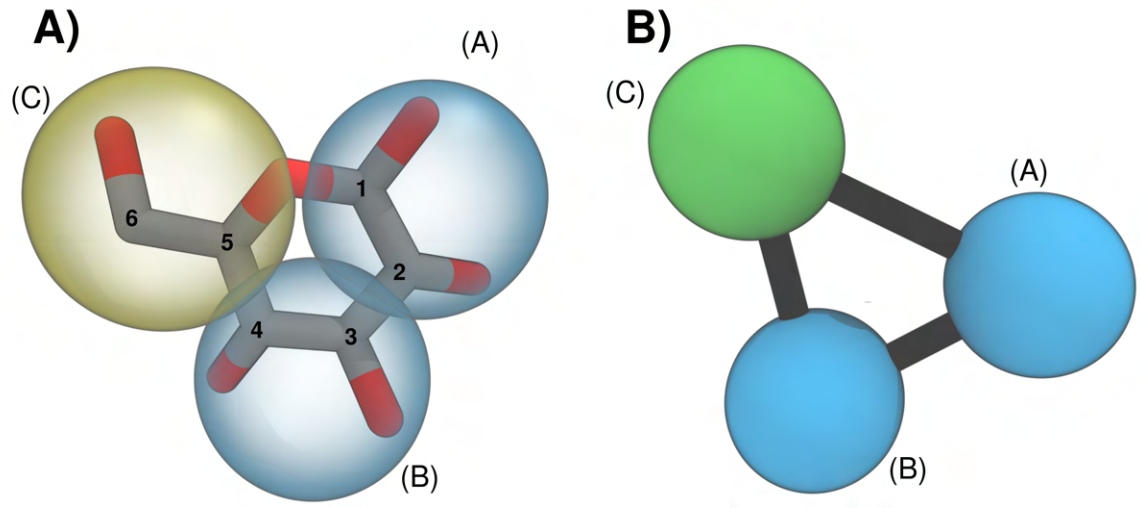

Figure 1.2: A possible coarse-graining approach to β-D-glucose is depicted. **A)** The topology of the original molecule is "mapped" to be represented by three large, pseudo-atoms (beads). **B)** The coarse-grained molecule is depicted, the original degrees of freedom have been reduced to only three atoms and three bonds while still globally capturing the β-D-glucose properties.

The advantage of coarse-graining is at least two-fold: through direct reduction of the degrees of freedom, less particles and subsequently, less bonded and non-bonded interactions need to be calculated. Due to the increase in mass of the pseudo-atoms relative to the atomistic structure, the bonded oscillation is reduced. Effectively, larger time steps and softer force constants for bond stretching and angle potentials are allowed, leading to a decrease in computational load and significant speedup of the structurally simplified model. All these simplifications open up the possibility to study complex systems at nearly biologically relevant timescales[20].

In coarse-grained methods, a clear distinction can be made between individually optimized models and so called generic (building block) approaches. The former has the potential to be a more accurate representation of the atomistic resolution, but requires extensive reparameterization when extending to new molecules. The latter focuses on transferability over full accuracy. Commonly, a standardized library of beads are supplied to the end user, which can be used to model certain chemical groups based on model-specific validation criteria. Via these approaches, the parameterization of consistent coarse-grained models is a real possibility. One of the

most popular and widely applied building block CG force fields is Martini, recently updated to its third iteration[4].

1.3.2 The Martini force field

The focus of Martini is to offer a transferable approach to parameterizing new (classes of) molecules. In Martini 3, this is achieved by defining non-bonded interactions through a Lennard-Jones (LJ) potential (*eq. 1.2*), or in case of electrostatic interactions, a (full) charge on a coarse-grained bead. By varying the parameters ε_{ij} and σ_{ij}, different levels of self-self and self-other interactions can be captured. σ_{ij} captures the size effects of a certain molecular fragment while ε_{ij} (well depth of the LJ potential) captures the relative strenght of interaction.

Three main bead sizes, regular (R) (σ=0.47 nm), small (S) (σ=0.41 nm) and tiny (T) (σ=0.34 nm), are present, parameterized to represent a certain fraction of heavy (i.e. non-hydrogen) atoms. R-beads optimally represent a 4:1 heavy atom to bead mapping and are suitable for linear, aliphatic arrangements, S-beads are recommended for aliphatic ring, or branched structures in a 3:1 ratio and T-beads in a 2:1 ratio and are optimized for stacking distances in aromatic rings. 22 levels of interaction (expressed in ε_{ij}) and size dependent interactions (expressed in σ_{ij}) give rise to a library of standardized bead types, grouped in different, global classes (e.g. Polar (P), apolar (C), charged (Q)). By reproducing experimental quantities such as free energies of transfer (see *Section 1.4*), the interaction levels are validated[21]. A full description of the parameterization process can be found in the original Martini 3 publication[4].

The interaction matrix (Supplementary Table 1, *Souza et al, 2021*[4]) is the heart of Martini. Since non-bonded interactions are taken care of through a specific bead type, a Martini model can be extended with relative ease to a new molecule. The first objective is to obtain a viable mapping where some generic rules are employed (see Section C1, *Souza et al, 2021*[4]). Here, conserving the topology and volume of the

original molecule, and preventing a mismatch in the amount of heavy atoms that are mapped per bead, are important goals. The second objective is to decide on and validate the bead assignment, i.e. to decide on which Martini beads should be used to model certain chemical groups. Validation with experimental data, or, at the least, comparisons with atomistic reference simulations should optimally be used. A highlight of general validation strategies is summarized in *Section 1.4*.

The final objective is to obtain the bonded parameters. In Martini, these can be extracted in a relatively straightforward way from sufficiently sampled atomistic reference simulations. The standard in Martini 3 over previous releases is to place the beads at the center of geometry (COG) of all the constituting atoms, including hydrogen atoms. A mapped pseudo-CG trajectory can be prepared by linking atom positions to CG beads. By matching the potential distributions with this mapped, atomistic trajectory, bonded potentials can be obtained and validated. In practice, the parameterization objectives presented here often intertwine. E.g. mapping or bead types may be changed depending on findings in certain reference systems or when volume is not accurately met.

1.3.3 On the Martini 2 carbohydrate definitions

The original extension of Martini to carbohydrates was published in 2009[22], and focused on reproducing water/octanol free energies of transfer for accurate bead assignment. Application of the proposed saccharide model to amylose fibers in water and nonane, as well as reproduction of the cryo-anhydro effect were additionally demonstrated to show the accuracy of the model. A wide range of applications, including succesful modelling of conformational changes in cellulose[23], reproduction of experimental properties such as the bending modulus in peptidoglycan[24] and replication of experimental observations such as β-cyclodextrin mediated cholesterol extraction[25,26] further showcased its efficacy.

An observation in protein and carbohydrate systems in Martini 2 was the propensity

for overaggregation (colloquially referred to as the "sticky" effect), attributable to too high solute-solute interaction levels, as characterized in multiple systems[1,3,27]. As *Schmalhorst et al* point out[3], such observations (i.e. overaggregation) are not limited to coarse-grained force fields alone. Atomistic force fields have also been shown to inaccurately represent aggregation properties of proteins[28] and (poly)saccharides[29], substantiating the need for experimental data in the validation process.

In their work, *Schmalhorst et al*[3] downscaled non-bonded sugar-sugar interactions in the Martini 2 carbohydrates through systematic scaling of the LJ parameter ε_{ij}, leading to accurate reproduction of experimental second virial coefficients of osmotic pressure (B_{22}). This approach of matching solution properties with experimental data has also succesfully been employed in matching experimental sugar binding affinities to proteins[27], or to significantly improve aggregation behaviour of carbohydrates in CHARMM36[30] and to optimize protein-protein interactions in Martini[1]. Rescaling the LJ interactions thus alleviates the sticky effect, but with the recent release of Martini 3[4] the increase in interaction levels and the inclusion of size dependent interaction offers a decidedly more robust foundation to improve the Martini carbohydrates even further.

1.3.4 On carbohydrate diversity

The chemical diversity of carbohydrates exceeds that of lipids and proteins[31]. The simplest possible carbohydrates include the five carbon ring (pyranose) structure and hemiacetal (e.g. glucose). However, four carbon rings (furanose) are also prevalent. The stereochemical configuration of the carbon substituents subtly affects the rotameric configurations of carbohydrates. Effectively, a simple carbohydrate like glucose exists in many isomeric forms and distinct rotamers. This complexity increases when looking at carbohydrates that include substituents such as carboxylic acids (e.g. D-glucuronic acid), amides (e.g. N-Acetylglucosamine) or combinations of both (e.g. N-acetylneuraminic acid). Other carbohydrates include deoxygenated species, such as D-ribose in nucleic acids, or sugars such as L-fucose and L-rhamnose

which are prevalent in N-linked glycans in glycoproteins[32]. When constructing polysaccharides from these individual building blocks, the complexity increases by an order of magnitude due to the variety of glycosidic linkages. The anomeric carbon (C1 carbon) can in principle be linked to any other carbon. If this receiving carbon is the C1 or C2 carbon, the sugar is non-reducing and cannot be extended, but from the C3-C6 carbons a wide variety of linkages can be envisioned, which offers challenges for coarse-graining force fields like Martini. Somehow, chemical specificity needs to be generalized while ideally maintaining a generic and transferable model that matches experimental data. To that end, some general validation strategies are commonly employed.

1.4 Validation strategies

In the so called top down approach, the Martini models are verified based on measurable (experimental) quantities. While the Martini force field already contains the information for non-bonded interactions in the specific bead types, the question is what bead types accurately capture the chemical specificity of carbohydrates, commonly validated through the experimental partioning between two solvent phases. Some *a priori* estimates of bead assignment follow from previous applications or are based on the guidelines in Supplementary Table 24 of *Souza et al*[4]. The following validation strategies are recommended for Martini models and were applied in this study as well.

1.4.1 SASA and molecular volume

A coarse-grained model should capture the global shape and volume of the molecule in the best way possible. The different bead sizes in Martini 3 affect volume each in a specific way. Where beads are placed at the center of geometry (COG) of its constituent atoms, the validation of volume by comparison with atomistic reference molecules is important. To that end, Solvent Accessible Surface Area (SASA)

calculations are commonly employed. The algorithm, as implemented in the *gmx sasa* package[33], comprises of probing the molecular structure using a user defined atom size which then gives an indication of the molecular volume, based on the van der Waals radii of the constituting atoms (see *Figure 1.3*). In the case of atomistic force fields, these can be derived from experimental data[34] and for Martini they are explicitly defined in the force field (*Table 7.1*). The surface of a molecule is proportionally related to its volume and scales with size. SASA (expressed in $Å^2$) is thus a computationally efficient way to estimate *relative* volume differences. A deviation of roughly 10 % between the Martini model and the atomistic structure is deemed acceptable if other properties are reproduced well.

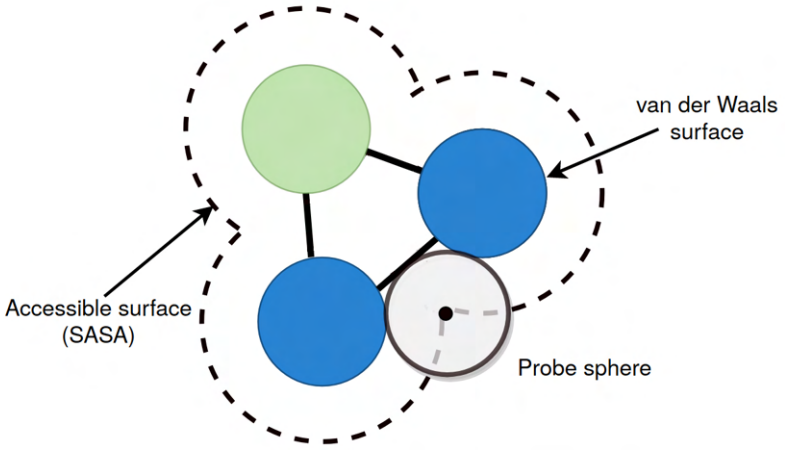

Figure 1.3: A schematic overview of the SASA methodology is presented. A probe, mimicking a solvent molecule, is rolled across the van der Waals surface of the molecule, giving an indication of the volume.

While SASA gives an indication whether the coarse-grained model approaches the atomistic volume to an acceptable degree, local deviations in volume cannot be captured as the surface is averaged over the entire molecule. To visualize areas where chemical moieties are either under- or overestimated a closely related property, the solvent-excluded (Connolly) surface can be utilized[35] . By overlaying the Connolly surface of the proposed coarse-grained model on top of an atomistic reference structure, a relative indication of volume can be utilized in the mapping procedure.

1.4.2 Free energies of transfer

The free energy of transfer (ΔG) for a solute is defined as the free energy change between two immiscible fluids, commonly an organic solvent (like octanol) and a water phase. Effectively, the transfer of the solute across the solvents gives a measure for the hydrophobicity of the molecule of interest. In the context of the overarching Martini force field, free energies of transfer are utilized in the parameterization process of the bead types as they can be validated with experimental measurements. As the first requirement in a Martini model is to formulate an effective mapping scheme, bead assignment accuracy can subsequently be scored based on reference data. Free energies of transfer are commonly expressed in partition coefficients (log P), as the ratio of unionized solutes in both solvents. In this study, experimental partition coefficients for different carbohydrates were recorded (*Table A1.6*), which can be transformed to the free energy of transfer $\Delta G_{OCO \rightarrow W}$ (kJ mol^{-1}) via the relation

$$\Delta G_{OCO \rightarrow W} = \ln(10)\, RT\, \log\, P \tag{1.3}$$

where R is the gas constant (kJ mol^{-1} K^{-1}) and T the temperature (K).

Since free energies of transfer can be estimated from MD simulations (*Section 7.4.2*) as a function of the solvation free energies ΔG_W and ΔG_{OCO}, a validation routine for bead assignment follows.

1.4.3 Osmotic pressure

The osmotic pressure is a thermodynamic solution property which describes the minimum pressure that needs to be applied to a dissolved solute to prevent flow across a semipermeable membrane. The osmotic pressure of many particle systems is given by the relationship:

$$\Pi = \phi\, c\, R\, T \tag{1.4}$$

where Π is the osmotic pressure (bar), ϕ is the (concentration dependent) osmotic coefficient, c the concentration (m) and R the gas constant (L bar K^{-1} mol^{-1}).

The osmotic coefficient represents the ratio of the solute specific osmotic pressure to that corresponding to an ideal solution and can therefore be interpreted in the following way: a value higher than 1 indicates repulsive solute-solute interactions, while a value lower than 1 indicates attractive solute-solute interactions. The osmotic coefficient or pressure that is calculated for a specific solute can then be correlated to experimentally determined values, to give a measure for aggregation qualities of the specific model.

Osmotic coefficients for many carbohydrates are ubiquitously available[36–40] and osmotic pressures may be obtained from MD simulations (see *Section 7.5.1*), as a function of molality:

$$m = \frac{\frac{N}{N_A}}{\rho_W V} \tag{1.5}$$

where N is the absolute number of solute molecules in the system, N_A is Avogadro's constant, ρ_W the density of water and V the volume of the box.

2 | Modelling carbohydrates in Martini 3

2.1 General considerations

Due to their large variety (see *Section 1.3.4*), in this study a representative group of carbohydrates was selected, with a focus on those for which physiochemical experimental data in the form of free energies of transfer or osmotic coefficients was available.

The following monosaccharides were considered in this study: β-D-glucopyranose (Glc), β-D-mannopyranose (Man) , β-D-galactopyranose (Gal), β-D-fructofuranose (Fruf), D-ribofuranose (Ribf), β-D-xylopyranose (Xyl), β-L-fucose (Fuc), β-L-rhamnose (Rha), N-acetylglucosamine (GlcNAc), N-acetylneuraminic acid (Neu5Ac) and myo-inositol (Ino).

The following disaccharides were considered in this study: lactose (Gal-β1,4-Glc) (Lac), sucrose (Glc-α1,2-Fruf)(Sucr) and trehalose (Glc-α1,1-Glc) (Treh).

The following polysaccharides were considered in this study: β-cyclodextrin (ring of seven α1,4-linked D-glucose subunits).

Monosaccharides in solution are in equilibrium between an open, aldehyde structure and closed, ring structure through the process of mutarotation[41], in which the ring form isomerizes between the α- and β-anomers through an open form intermediate. The contribution of the latter structure is minimal, at room temperature only 0.02% of glucose is in the open structure, with a 36% and 64% contribution for the α- and β-epimer respectively[41]. As the ring form is also the only biologically relevant structure[42], the open form was not considered in the parameterization procedure.

2.2 Hypotheses concerning modelling guidelines

2.2.1 Monosaccharides

Generating parameters for molecules using Martini 3 follows a set of rules as outlined previously (*Section 1.3.2*). Ultimately, parameters that yield the best match against experimental data determine the goodness of the parameters. To derive generic guidelines for parameterizing monosaccharides, we translate the generic Martini 3 rules into hypotheses, which will be proven or disproven by making comparisons to experimental data.

As was explained in *Section 1.3.2*, the first objective of a Martini model is to derive a valid mapping scheme. In the case of glucose, twelve heavy atoms (six carbons and six oxygens) can be represented by three regular (R) beads in a 4:1 mapping. However, as is pointed out in Section C1 of *SI, Souza et al*[4], R bead types are mostly parameterized for modelling linear, unbranched arrangements. Small (S) beads are generally recommended to model ring structures, making them in theory more suitable than R beads to map the carbohydrate ring. Similarly, Section C3 of *Souza et al*[4] underlines the significance of the position where the bead is placed for accurate reproduction of volume. In Martini 2, beads were placed at the center of mass (COM) of the constituting heavy atoms while Martini 3 has moved to a center of geometry (COG) approach where beads are placed in the geometrical center of the constituting atoms (hydrogen atoms included). While this is the recommended approach, it should still be explicitly verified whether this is optimal for monosaccharides. In the (preliminary) mapping scheme we attempted to follow the rules as outlined in *Souza et al*[4] as closely as possible. A consequence of the branching rule is that diols are kept together as much as possible, which is showcased in *Figure 2.1*:

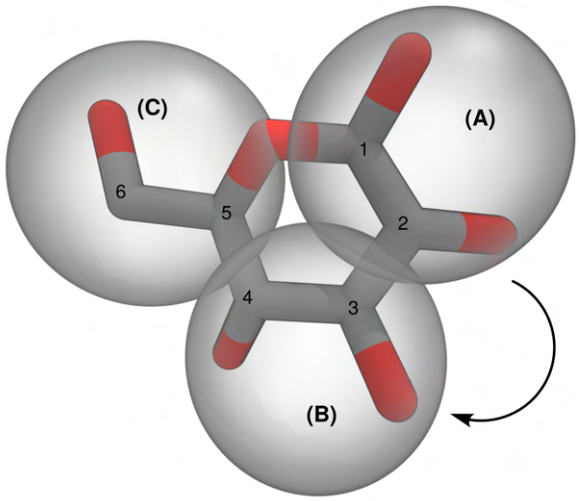

Figure 2.1: Preliminary mapping of a simple carbohydrate (β-D-glucose). The main ring is modelled by three beads, each representing four heavy atoms.

Based on these considerations, we formulated the following hypothesis:

> **Hypothesis 1 (H1):** *The best mapping scheme for monosacharides utilizes 3 S-beads, maximizes the number of diols per bead and places the beads at the center of geometry of the constituting atoms.*

Next, we have to assign the optimal bead types to the chemical groups. For the initial bead assignment, we used *Supplementary Table 24* from *Souza et al*[4] as a reference. Specifically, the diol (-CH(OH)-CH(OH)-) moieties (bead "A" and "B" in *Figure 2.1*) are similar to ethanediol, modelled by a SP3 bead. The remaining hemiacetal (bead "C") matches the P2 moiety as highlighted in the Martini 3 publication, prompting us to map it with SP2, which leads to the following hypothesis:

> **Hypothesis 2 (H2):** *SP3 is the most accurate bead assignment for modelling the diol moieties in carbohydrates. SP2 is the most accurate bead assignment for modelling the hemiacetal group*

The presence of anomers in monosaccharides (i.e. the sterical configuration of the -OH group on the anomeric (C1) carbon) leads to two distinctly different molecules,

such as α-glucose and β-glucose. Due to the anomeric effect, where the β-anomer shows higher stability in aqeuous solutions due to axial positioning of the alcohol[43], the β-anomer was deemed representative of monosaccharides. In the context of Martini as a coarse-grained force field, representing anomers explicitly would expand the sugar library twofold. To maintain transferability, we thus hypothesize the following:

Hypothesis 3 (H3): *α-anomers of monosaccharides do not significantly affect bonded parameters, with respect to β-anomers, in Martini 3.*

So called epimers, where the only difference between isomers is the axial or equatorial orientation of one -OH group on one carbon, also theoretically lend themselves for generic models. β-D-glucose, β-D-mannose and β-D-galactose fit this specific type of isomericity and could perhaps be represented by a singular, unified model. While we believe this to be possible, agreement with experimental data should still be verified to validate the accuracy of the constituting model, leading to the following hypothesis:

Hypothesis 4 (H4): *Representing carbohydrate isomers by generic models still leads to good reproduction of experimental data of the individual carbohydrates.*

As is highlighted in *Section 1.1.2*, the CH-π ring stacking between carbohydrates and aromatic residues is an important property of carbohydrates interacting with protein systems. Initial testing with a preliminary Martini 3 sugar model in a Shiga toxin (Stx)[44] found more accurate affinity for lipids by inclusion of a hydrophobic virtual site bead in the geometrical center of each monosaccharide ring. A virtual site is a massless particle that when placed in a molecule does not interact with its constituting atoms[45]. In the recent development of phosphatidylinositide (PIP) lipid parameters of Martini 3, such a virtual particle was included in the inositol carbohydrate model as well[46]. In general, we expect the inclusion of a TC4 virtual

site bead to help affinity of carbohydrates for hydrophobic environments. For example, the inclusion of cholesterol in the inner face of β-cyclodextrin[26], or the binding of carbohydrates to proteins - governed by aromatic residues - are systems we expect to behave more in line with experimental observations when using such a TC4 dummy particle. However, the agreement with reference data such as molecular volume, partitioning and osmotic pressure should still be verified, leading to the explicit hypothesis:

Hypothesis 5 (H5): *The inclusion of a TC4 virtual site in the geometric center of the monosaccharide does not significantly affect agreement with thermodynamic reference data.*

2.2.2 Disaccharides

Disaccharides are formed from a condensation reaction between two monosaccharides. A glycosidic bond is formed between the anomeric carbon of the first sugar and one of the carbons in the second sugar. Based on the specific stereochemistry of the two monosaccharides, a distinction is made between α- and β-glycosidic linkages. Due to glycosidic linkages, the total number of heavy atoms of the sum of two monosaccharides decreases by one (see *Figure A1.1*), which affects the volume of the disaccharide.

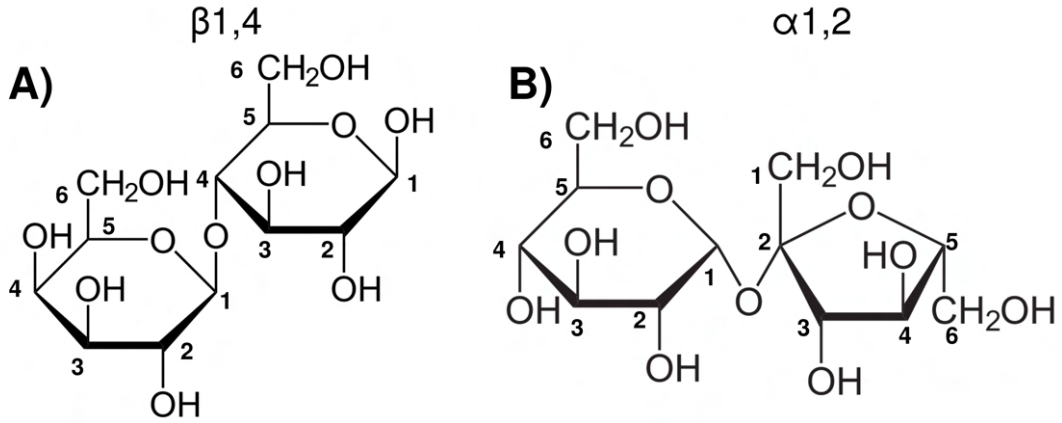

Figure 2.2: A) Lactose (Gal-β1,4-Glc) is depicted. **B)** Sucrose (Glc-α1,2-Fruf) is depicted. During glycosidic linkage (condensation reaction), disaccharides "lose" one oxygen (leading to a total of eleven heavy atoms)

Ideally, we would like to construct disaccharides from monosaccharides in a consistent manner, and then extend those to polysaccharides. To achieve this goal, we need to account for the glycosidic linkage in a specific way. As we hypothesize the core monosaccharide ring to be modelled by three S-beads (*Hypothesis 1*), we can conceivably account for the decrease in heavy atom by taking part of the sugar as a T-bead, since it will have one atoms less due to the asymmetrical nature of disaccharides. Therefore, we formulate the following hypothesis:

Hypothesis 6 (H6): *The optimal way to account for volume loss is by taking the glycosidic oxygen in an S-bead of the first monosaccharide unit, while the loss of heavy atom in the receiving monosaccharide is modelled by a T-bead.*

The nature of the glycosidic linkage (i.e. α vs β) affects the orientation and flexibility of carbohydrates[47,48], which is ideally captured by Martini. However, since Martini utilizes a coarse-graining approach, a generic model for both types offers a few distinct advantages. Firstly, fewer models (by a factor of two) need to be parameterized, promoting transferability and ease of usage. Secondly, the use of generic parameters promotes stability due to fewer bonded combinations that need to be tested:

Hypothesis 7 (H7): *It is possible to parameterize a generic carbohydrate model which combines α- and β linkages together in one, generic parameter set, similar to Martini 2.*

The inclusion of virtual sites, which was already discussed in *Hypothesis 5* is also relevant to disaccharides. The observation that virtual sites in constraint heavy systems improve stability[49] could be argued to play a role in disaccharides and polysaccharides as well. The rigidity in the monosaccharide ring, expressed in a small standard deviation of bond length (similar as in Martini 2) lends itself well for constraints. In more complex sugars a large number of constraints are thus used, promppting us to define the bonded potentials through the virtual site, as is

described in *Section C4* of *Supplementary Information, Souza et al*[4]:

> **Hypothesis 8 (H8):** *Connecting individual monosaccharide units via bonds and angle/dihedral potentials through virtual sites leads to accurate reproduction of atomistic dynamics and flexibility and promotes (numerical) stability*

The aforementioned Martini 3 model for myo-inositol in PIP lipid also serves as a frame of reference in deriving a valid mapping scheme[46]. For reproduction of volume, a 13% increase in COG bond lengths was needed in their model. Based on this consideration, our hypotheses and preliminary testing, we focused on the following models:

- **S-bead** - COG mapping using S-beads, no virtual site
- **N-bead** - COG mapping using R-beads, no virtual site
- **VS (TC4)** - COG mapping using S-beads and a TC4 virtual site
- **VS (TC4) (15%)** - mapping using S-beads, TC4 virtual site and 15% uniform scaling of the constraints in the monosaccharide ring.

As our main goal is to improve on the Martini 2 carbohydrates, we should of course draw a comparison for each validation criterion. Of the carbohydrates studied here, glucose, fructose, sucrose and trehalose were also mapped to Martini 2. For clarity it should be highlighted that in Martini 2, monosaccharides were mapped by three R-beads and consecutively, disaccharides by six R-beads.

These five models (including Martini 2) were applied to (our) experimental data, the result of which is presented in the next Chapter.

3 | Results

3.1 Molecular volume and SASA

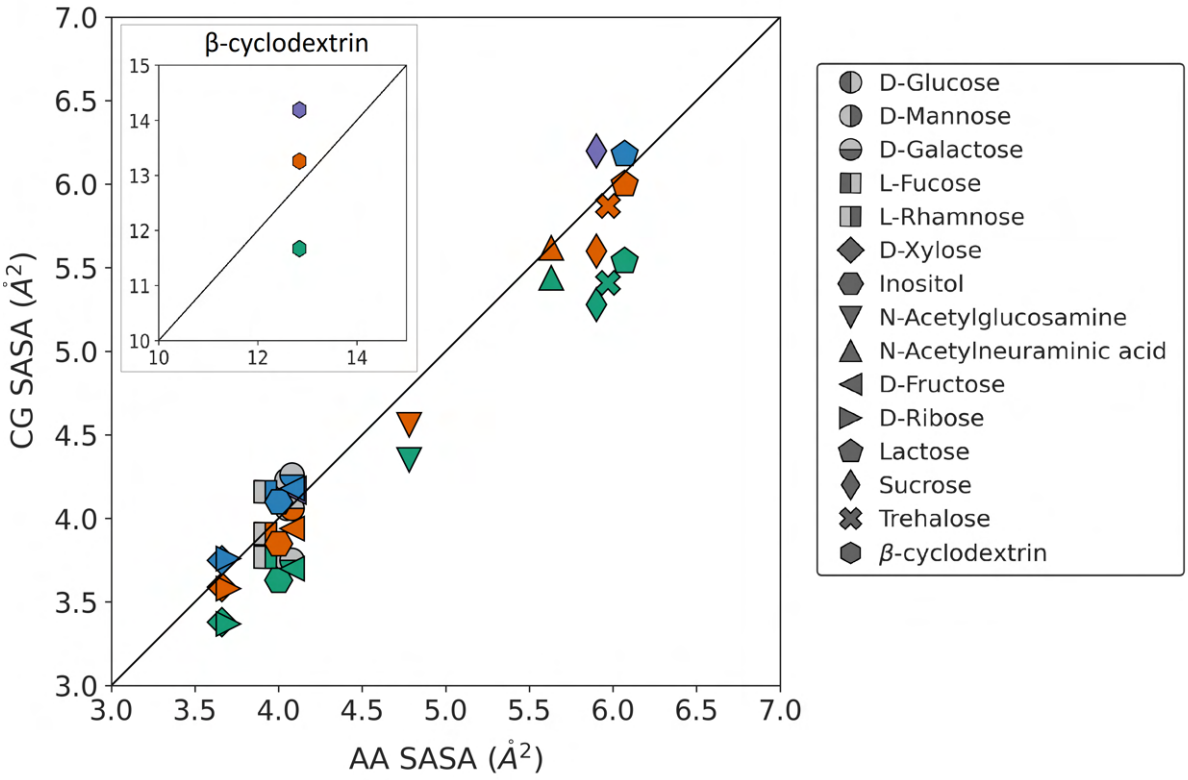

Figure 3.1: SASA values of different carbohydrates are displayed. Colour refers to different models. Green = S-bead (-7.7% AA), TC4. Orange = S-bead, TC4 15% (-1.9% AA) . Blue = R-bead (-3.6% AA). Purple = Martini 2 (R-bead).

In *Figure 3.1* the coarse-grained model SASA values are plotted with respect to the atomistic reference value. The atomistic van der Waals radii are those by Rowland and Taylor[34], as recommended in the Martini 3 publication[4] for hydrogen rich compounds.

Since SASA only gives a 2D representation of the volume, we have also calculated the Connolly (solvent-excluded) surface for β-D-glucose (*Figure 3.2*). The Connolly surface gives a more extensive view of local volume deviations.

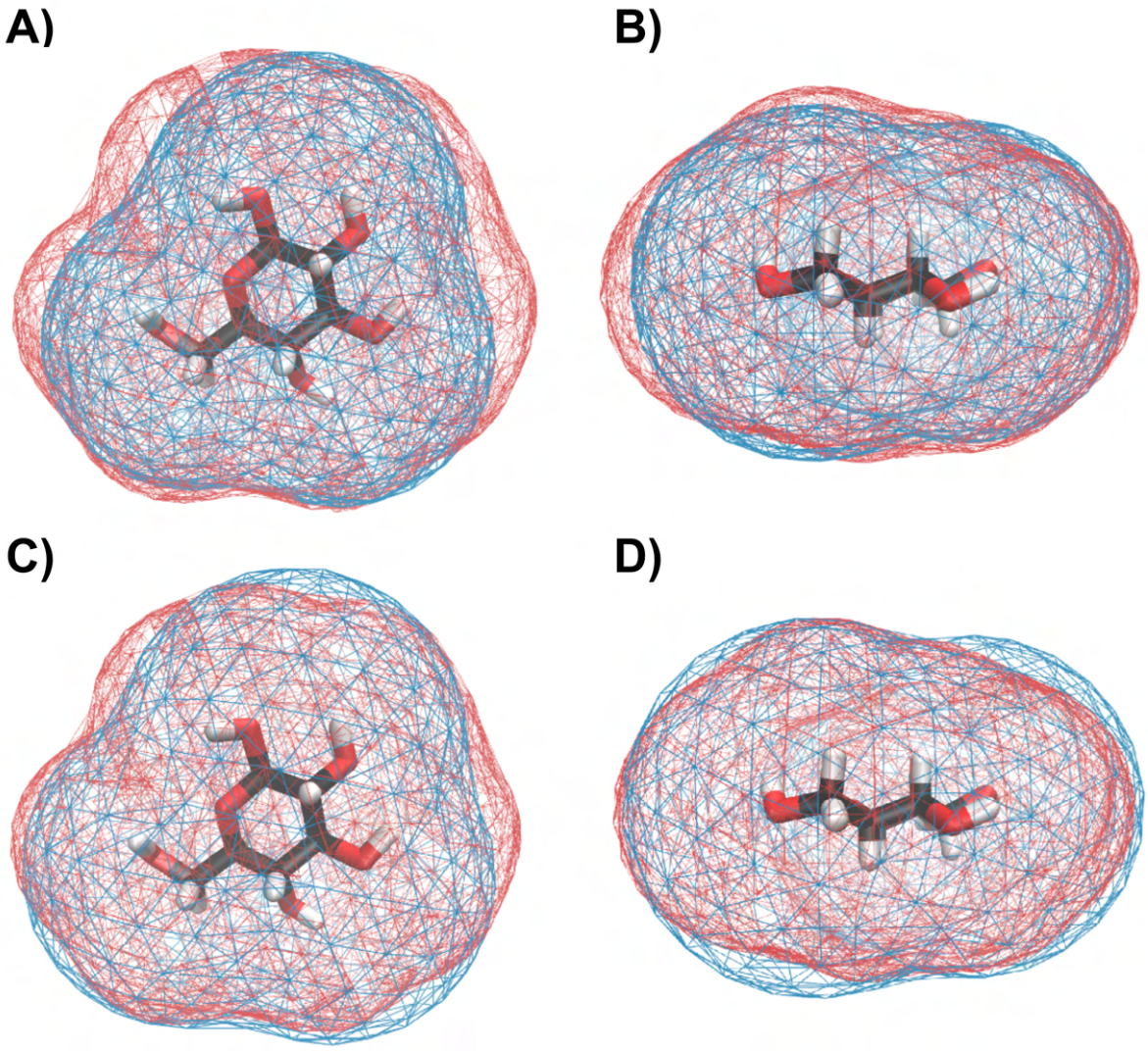

Figure 3.2: Connolly (solvent-excluded) surfaces of glucose, where in blue the atomistic surface is presented and in red the specific coarse-grained surface. **A)** R-bead model, top view. **B)** R-bead model, side view. **C)** S-bead model, top view. **D)** S-bead model, side view.

The full SASA values are collected in *Table A1.4*. In the case of monosaccharides with substituent groups, such as GlcNAc or Neu5Ac, scaling is only performed on the internal bonds, i.e. on the beads that make up the core monosaccharide ring and not on substituent beads.

3.2 Free energies of transfer

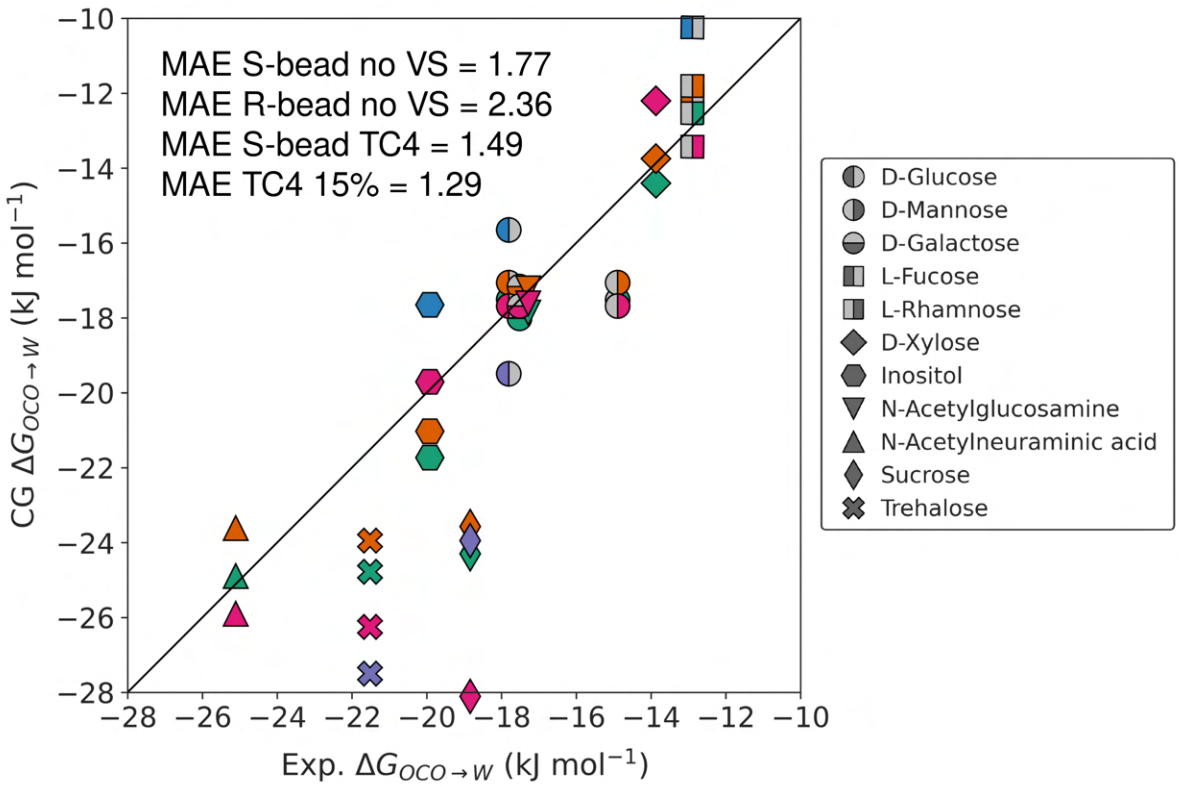

Figure 3.3: The free energies of transfer for different coarse-grained models are plotted with respect to the experimental reference. Mean absolute errors (MAE) are given in kJ mol^{-1}. Models are distinguished by colour where pink = S-bead no VS, blue = R-bead no VS, purple = Martini 2, green = S-bead TC4 and orange = S-bead TC4 15% scaled.

Free energies of transfer were calculated according to the methodology as outlined in *Section 7.4.2* and the results are displayed in *Figure 3.3*, with respect to the experimental reference value. The selection of bead assignment followed an iterative approach. Firstly, the bead assignment for the diol moiety was validated in inositol for each specific model (the structure of inositol can be seen in *Figure 6.1D*), as that sugar consists of three of such diol groups. Then, the hemiacetal bead (C5/C6 hydroxymethyl group) was validated in D-glucose, as that sugar is mapped by two diol moieties (validated in inositol) and then optimization could be done on solely the hemiacetal bead. The bead assignment followed this iterative manner until all chemical groups were explicitly verified.

3.3 Osmotic pressure

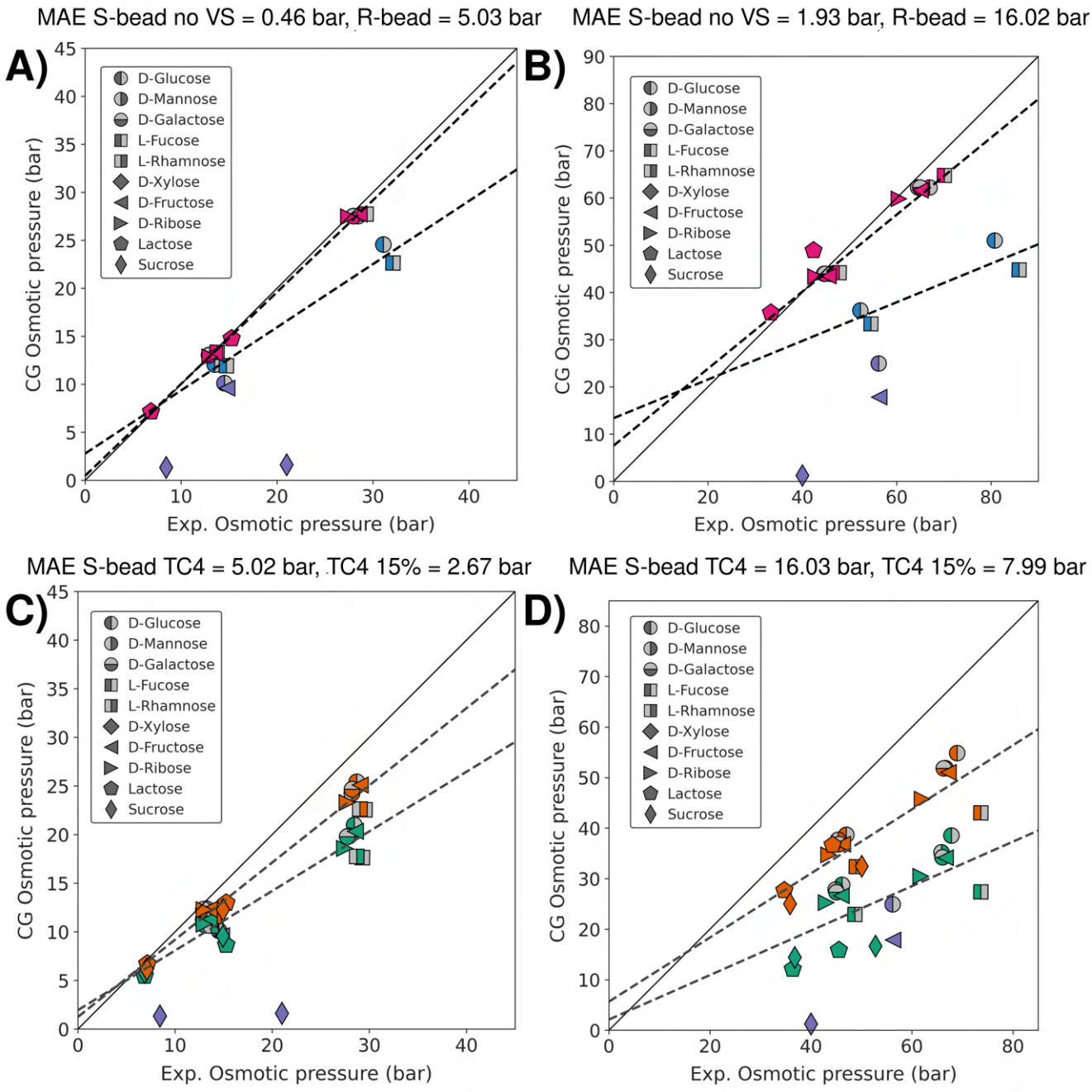

Figure 3.4: The osmotic pressure of different coarse-grained models with respect to the experimental value is shown. Models are distinguished by colour where pink = S-bead no VS, blue = R-bead no VS, purple = Martini 2, green = S-bead TC4 and orange = S-bead TC4 15% scaled. Mean absolute errors (MAE) are calculated from the range 0-1 molal (A and C) or 0-2.5 molal (B and D). **A)** <1 molal range for R-bead and S-bead model. **B)** 1-2.5 molal R-bead and S-bead model. **C)** < 1 molal for S-bead TC4 and S-bead TC4 15% scaled. **D)** 1-2.5 molal for S-bead TC4 and S-bead TC4 15% scaled.

The osmotic pressure as a function of temperature and the concentration in molality was calculated and related to experimental values derived from osmotic coefficients (see *Section 7.5.2*). The Martini 2 model, modelled by three R-beads, was also taken into account as a reference, where 10% WF (antifreeze) particles were included, as *Schmalhorst et al* indicate is necessary for optimal aggregation behaviour in Martini 2[3].

3.3.1 Lactose - receiving glycosidic bead

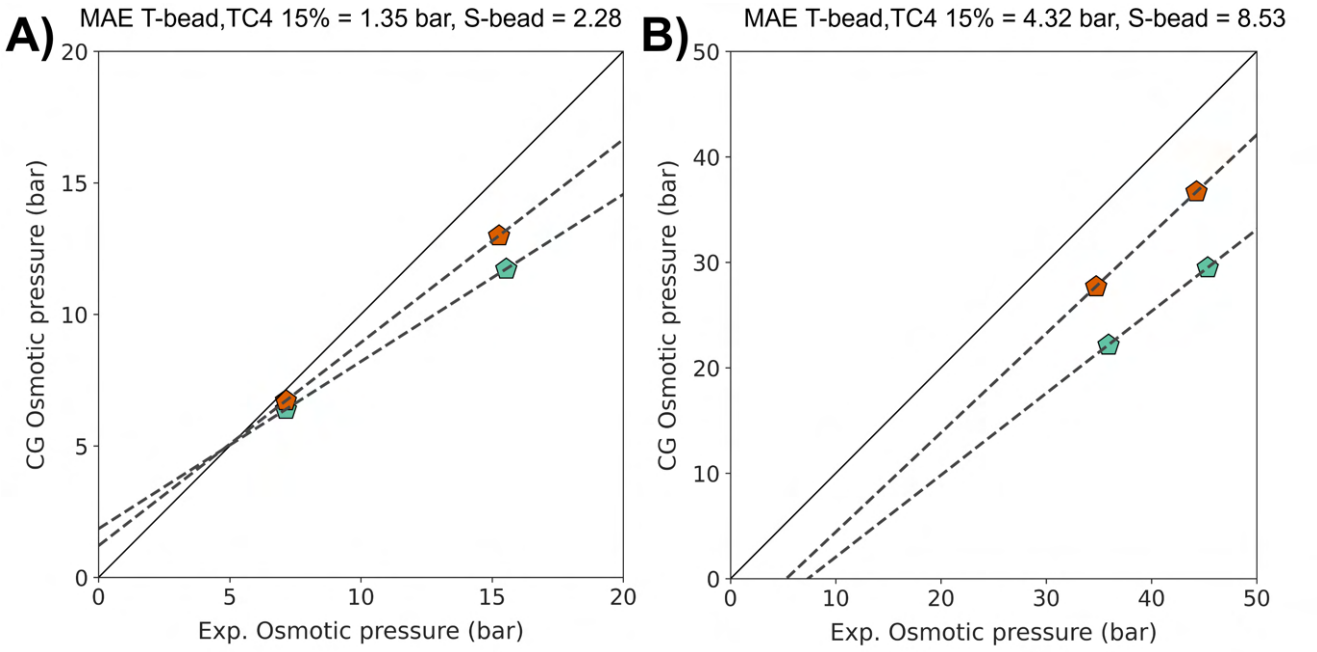

Figure 3.5: Osmotic pressures of lactose, for the S-bead TC4 15% scaled model where the receiving (i.e. not the bead that includes the glycosidic oxygen) bead in the second monosaccharide is modelled as a TP1 bead (orange) or SP1 bead (light green). **A)** < 1 molal concentration range. Mean absolute errors (MAE) are calculated from the range 0-1 molal **B)** 1-2.5 molal concentration range. Mean absolute errors (MAE) are calculated from the range 0-2.5 molal (B)

3.4 Bonded potentials

3.4.1 α-D-glucose & β-D-glucose

The anomeric influence (i.e. whether the -OH at the C1 carbon is in the axial or equatorial position) on the Martini 3 monosaccharides was related to D-glucose. The

bonded parameters are given in *Table 3.1*:

Bond	**α-D-Glc**	**β-D-Glc**	**%-diff**
A - B	0.322	0.330	2.45%
A - C	0.389	0.409	5%
B - C	0.349	0.344	1.4 %

Table 3.1: Bonded parameters of α-D-glucose and β-D-glucose. Bead names (A, B or C) refer to the mapping of monosaccharides as outlined in *Chapter 6*.

3.4.2 Virtual site & direct glycosidic bond

The construction of carbohydrates through a virtual site description was related to a direct bond approach, where bonded potentials are not drawn through virtual sites. To showcase the differences, the dihedral distributions of two relatively flexible sugars, maltose (Glc-α1,4-Glc) and sucrose (Glc-α1,2-Fruf) are showcased in *Figure 3.6*:

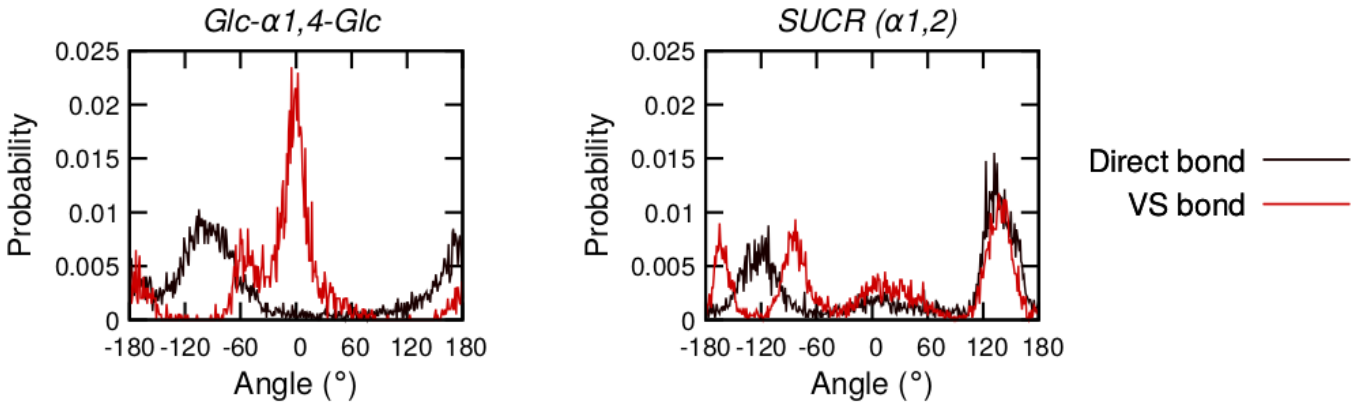

Figure 3.6: The difference in dihedral multiplicity for the virtual site bonded model and the direct bonded model is shown for two α-linked sugars, maltose (Glc-α1,4-Glc) and sucrose (Glc-α1,2-Fruf). The dihedral profile for virtual site bonds is erratic and hard to accurately capture by harmonic potentials.

3.4.3 α- & β- glycosidic linkages

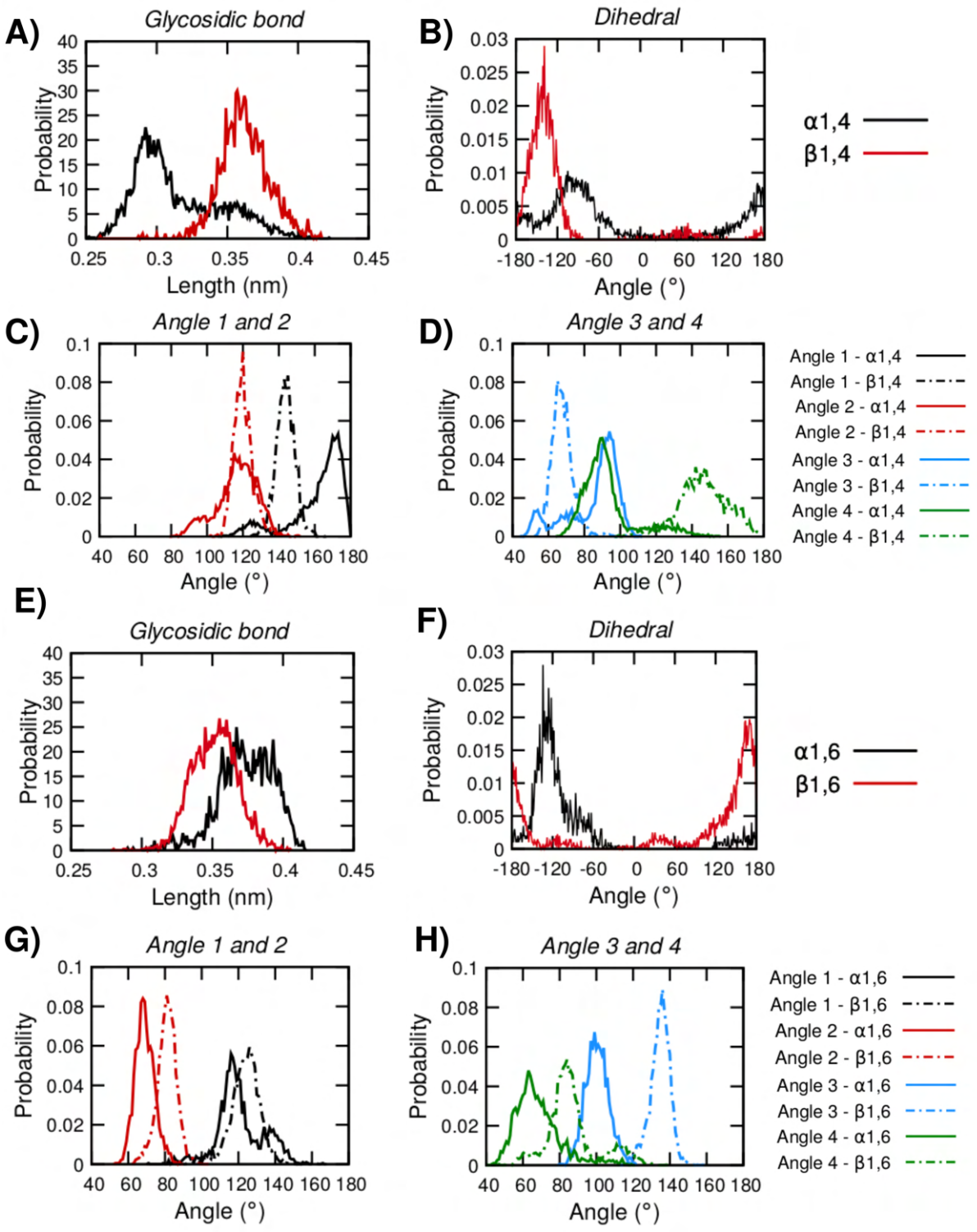

Figure 3.7: Distributions for the glycosidic bond, angles and dihedrals are shown for Glc-α1,4-Glc, Glc-β1,4-Glc, Glc-α1,6-Glc and Glc-β1,6-Glc. Angles correspond to the mapping scheme as outlined in *Section 6.2*.

4 | Discussion

4.1 Martini 3 monosaccharides are best represented by S-beads

A simple monosaccharide like glucose can be represented by three beads in a 4:1 mapping. The reflection on the choice for S-beads or R-beads depends on a few criteria. From *Figure 3.1*, it becomes apparent that the S-bead model underestimates atomistic SASA by on average 7.9%. The R-bead model in that regard is more accurate, with an average overestimation of 3.6%. However, when looking at larger polysaccharides like β-cyclodextrin, the Martini 2 model (comparable to a Martini 3 R-bead model) overshoots volume significantly (by a factor of $\tilde{1}0$ %).

The risk of overmapping is also reflected in the lack of agreement with partitioning data and aggregation behaviour of the R-bead model. The S-bead, no VS model has a mean absolute error of 1.77 kJ mol^{-1}, while the R-bead, no VS model has an error of 2.36 kJ mol^{-1}. This increase in hydrophobicity also makes the R-bead model more prone to aggregation, reflected in osmotic pressure (see *Figure 3.4* A & B, blue projection). S-bead carbohydrates have a MAE of only 1.93 bar for the 0-2.5 molal range, while R-bead carbohydrates are significantly worse with a MAE of 16.02 bar. We therefore opted to map the base monosaccharide (pyranose) ring with three S-beads, in a 4:1 mapping. The agreement with experimental data for this model is better, and it offers us the flexibility of scaling COG bonds uniformly if volume becomes an important consideration in later stages of the parameterization process.

4.2 Diols are mapped by SP3 beads and hemiacetals by SP1

In the choice for bead assignment, we worked in an iterative manner. We started in inositol, which has three diol moieties that should thus have the same bead assignment. When opting for SP3, the free energy of transfer is -19.71 kJ mol^{-1}, which is only 0.21 off the experimental value. One can imagine that all SP2 or all SP4 beads would thus worsen the agreement, validating SP3 for the diol moieties.

However, when looking at monosaccharides with two diol moieties and a hemiacetal group (e.g. glucose), we use two SP3 beads and then validate the hemiacetal. When we take SP2, the free energy of transfer is -18.89 kJ mol^{-1} (+1.08 kJ mol^{-1} from experimental). When we move one polarity level down, to SP1, our agreement becomes much better with a value of -17.68 kJ mol^{-1} and deviation of only -0.13 kJ mol^{-1}. The electronegative ether oxygen in the hemiacetal group could perhaps explain this change in two polarity levels. The change from SP2 to SP1 also improves the agreement with reference osmotic pressure data for glucose. SP3-SP3-SP2 mapping has a MAE of 3.8 bar for the 0-2.5 molal range, while SP3-SP3-SP1 mapping has a MAE of 1.9 bar (*data not shown*).

4.3 α- and β-monosaccharides are highly similar in Martini 3

As can be seen from *Table 3.1*, there are slight influences of anomeric positioning of the C1 alcohol on the COG bond lengths. The maximum bond length difference is only 5% which is barely significant. Here, the goal of simplicity and transferability should probably have priority over explicit representation of α- and β-monosaccharides in Martini 3. We have currently opted to take the final bonded parameters of monosaccharides as simply being the β-configuration. A possible

improvement to the model proposed here might be to average the bonded parameters of the α- and β anomers to give a more accurate reflection, while still limiting the amount of explicit representation needed.

4.4 Generic models can be used to limit the complexity of the Martini 3 carbohydrate library

While not shown in this thesis, we have built generic models out of the average bonded parameters of β-D-glucose, β-D-mannose and β-D-galactose. The difference in bonded parameters only leads to a very small effect in partitioning agreement, as can be seen from the values in *Table A1.5*, as well as the osmotic pressure agreement, as can be seen from the clustering of monosaccharides in *Figure 3.4*. In this study, we have mostly focused on the the validation of the mapping concepts and therefore have opted to treat each sugar explicitly for now. However, from our generic models (*data not shown*) and from inspecting the validation criteria, we see no reason why generic models should not be utilized. One, perhaps, interesting exception to this is D-mannose ($\Delta G_{OCO \rightarrow W}$ = -14.9 kJ mol^{-1}), which is decidedly more hydrophobic than D-glucose ($\Delta G_{OCO \rightarrow W}$ = -17.81 kJ mol^{-1}), while sharing the same chemical structure. Because of this, this effect cannot really be captured by Martini as the bead assignment is exactly the same as in D-glucose. Small changes in bonded parameters cannot significantly make the coarse-grained mannose more hydrophobic, perhaps requiring this molecule to be represented in a worse way than either D-glucose, or D-galactose, to maintain transferability.

4.5 Inclusion of a virtual site significantly affects the chemical behaviour of carbohydrates in Martini 3

The inclusion of a TC4 dummy particle in the center of each monosaccharide ring severely affects the aggregation behaviour of the carbohydrates. As can be seen from

Figure 3.4C & D and as is reflected in the mean absolute eror (MAE), when including a TC4 virtual site the agreement with osmotic pressure worsens. When comparing the S-bead model, COG mapped, with no virtual site (MAE = 1.93 bar), to the model with a TC4 virtual site, COG mapped (MAE = 16.03 bar), the effect of the TC4 bead on aggregation behaviour becomes apparent.

To compensate for this effect, we have opted to scale the internal/ring bonds uniformly by 15%, effectively helping both SASA agreement and aggregation behaviour. When we scale by 15% our CG SASA deviation is reduced to only an average error of 1.9% from 7.7 % and our osmotic pressure MAE improves from 16.03 bar to 7.99 bar.

The advantage of transferability, intrinsic to Martini is still maintained with this approach. Still, even after scaling the TC4 model is significantly worse with regard to aggregation, especially for the disaccharides (sucrose and lactose). Because of preliminary observations in complex systems, we have still decided to take the final carbohydrate model to be based on the TC4 model, with uniform bond scaling of 15% to both improve volume agreement and to help osmotic pressure agreement. However, the highest priority after finalizing the carbohydrate definitions is to finally validate this application in complex test systems.

4.6 Disaccharides cannot simply be constructed from monosaccharides

In the Martini 2 carbohydrate definitions, disaccharides were effectively directly constructed from monosaccharides, leading to a six R-bead mapping. As we have shown, the use of S-beads is preferable to model the monosaccharide ring but the question still remains on how we should deal with the glycosidic bond.

In our current approach we take the glycosidic oxygen in the bead in the first monosaccharide. The receiving glycosidic moiety is then modelled by a T-bead. In

the case of lactose, the SASA agreement moderately changes, from LAC AA 6.07 $Å^2$, CG 15% receiving T-bead 6.0 $Å^2$ and CG 15% receiving S-bead 6.05 $Å^2$. However, the effect on aggregation is more pronounced. For lactose, we go from a MAE of 4.32 bar from the 0-2.5 molal range for a receiving T-bead to a MAE of 8.53 for a receiving S-bead. The choice for T-bead is thus validated, but we can imagine that the use of a T-bead can lead to packing issues in tightly packed crystal structures such as cellulose. The use of S-bead here is in principle not incorrect, because we would still be maintaining a 3:1 heavy atom - bead mapping. Due to the chemical similarity in the glycosidic bead moiety see (*Figure 6.2*) to the hemiacetal group, we have opted to represent this as an SP1 bead as well for 1,1 - 1,4 linkages, while 1,5-1,6 linkages are represented by an SN6 bead due to the extra carbon modelled in the bead affecting the polarity. The receiving bead is modelled as a TP1 bead, which has been validated in D-ribose and N-Acetylglucosamine via free energies of transfer.

4.7 α- and β-linkages should be treated separately in Martini 3

The use of generic models is secondary to accurate models. The goal of simplifying glycosidic linkages by treating α or β linkages indiscriminately, or by some combination of individual parameters seems to overshoot its purpose. As the distributions in *Figure 3.7* show, α bonds have distinctly different bond lengths from their β counterparts. While angle distributions are fairly similar, the difference in force constant, especially between α1,4 and β1,4, is also significant. The observation that β-glycosidic linkages are more rigid match observations in larger structures such as cellulose fibers (*data not shown*). For these reasons, these types of linkages are parameterized separately in Martini 3.

4.8 A direct bond approach is more stable than a virtual site bond

Where all glycosidic bonds and bonded potentials were initially drawn between two virtual sites, this implementation led to instabilities, especially in the more flexible α-linkages[48]. This matches observations of instabilities in the simulation of the coronavirus, where a preliminary sugar model, based on a virtual site bonded description was used. The free rotation around the glycosidic bond seemed to allow for more conformational flexibility, exhibited in the dihedral distributions shown in *Figure 3.6*. These high degrees of multiplicity also directly oppose experimental findings where mostly two distinct rotameric states are dominant in glycosidic linkages[50]. Therefore, we opted to represent glycosidic linkages as explained in *Chapter 6*. Using direct bonds, a time step of 20 fs is possible without major issues, even in larger carbohydrates such as β-cyclodextrin and dextran (*data not shown*) which is in line with recommendations by *Souza et al*[4].

4.9 Bead assignment leads to generic linkages

One result of modelling the carbohydrates by three S-beads, is that in the case of pyranoses, two carbons are mapped per bead. Because of this, we have decided to group 1,1-1,2, 1,3-1,4 and 1,5-1,6 linkages together in one overarching linkage class. Our potential scheme then leads to six Martini "glycosidic" linkages (summarized in *Table 6.1*). Explicitly representing each glycosidic linkage would lead to twelve distinct parameter sets, which is possible, but a large undertaking. By averaging the bonded parameters of the three main bond classes, we expect improved behaviour relative to Martini 2, while still having a manageable amount of parameters in the carbohydrate definitions.

4.10 Comparison to Martini 2 reworked carbohydrates

The Martini 2 carbohydrate model was applied to a wide range of complex systems[23–26], but the unphysical aggregation levels hampered application in e.g. glycosylated proteins. As was explained in *Section 1.3.3*, the strategy of scaling the ε_{ij} parameter between the sugar-sugar interactions alleviated this issue significantly, allowing for good reproduction of experimental second virial coefficients (see *Table 2, Schhmalhorst et al*[3]) and accurate aggregation properties of complex carbohydrates like A2 glycan. However, in the case of glucose a systematic *overestimation* of B_{22} is reported, while we are underestimating the agreement with experimental osmotic pressure.

Another thing of note is *Schmalhorst et al*'s reported standard deviations which are relatively high (e.g. for glucose a B_{22} value of 0.22 L mol^{-1} with a standard deviation of 0.23 is reported, relative to an experimental value of 0.117 L mol^{-1}), making it questionable as to how exactly the carbohydrates are scoring relative to our model. The following distinct advantages from the Martini 3 carbohydrates compared to Martini 2 become at least apparent. Firstly, in this study more monosaccharides were mapped (in Martini 2, only glucose and fructose were mapped) and related to free energies of transfer, allowing for validation of a higher variety of chemical moieties. Secondly, a measure of aggregation through osmotic pressures was taken into account. From *Figure 3.4* it can be clearly observed that the changes in Martini 3 improve aggregation markedly, especially when bond scaling is introduced. Thirdly, our description of glycosidic linkages is more detailed. We represent α- and β-linkages explicitly, which was not the case in Martini 2. In the Martini 2 carbohydrate definitions, representation of 1,6 linkages was omitted due to problems in matching the dihedral distributions, which is not a problem in our potential scheme.

4.11 Considerations for further parameterization

The current model revolves around a virtual site description, where a hydrophobic TC4 bead is included in the center of each monosaccharide, and constituent bead assignments are scaled in polarity accordingly. To improve the aggregation behaviour and to improve the SASA agreement, we have opted to scale the internal ring bonds by 15%. This decision is based on the rigid structure of the core, three ring beads. The glycosidic bond and the bond of substituent groups on the monosaccharides are comparatively much more dynamic.

The decision for TC4 as being the central bead is also not fully explored. While we have tested some other options, the application in complex systems should tell us more than agreement with experimental data currently can.

To improve agreement with aggregation data, we could also opt for an 'r' label to reduce the self-interaction level by one (see *Section A4, Supplementary Information, Souza et al*[4]). Effectively, the interaction with octanol/water is not affected, leading to similar bead assignment, but theoretically this reduced self-interaction could lead to more realistic aggregation behaviour.

Some of the limitations of the carbohydrate model proposed here include intrinsic limitations to MD approaches. The modelling of ring puckering, i.e. boat and chair conformation is also limited in MD. Similar to Martini 2, only the $^4C_1$ chair conformation is considered which is fortunately the dominant puckering state[51].

The absence of complex test cases (such as protein-carbohydrate binding, lipid affinity and other applications) is the next goal in the optimization of the model, although the current mapping approach is mature enough that it can be applied to any carbohydrate system of interest.

5 | Conclusion

Here, we show the parameterization and mapping guidelines of an updated carbohydrate model in the overarching Martini 3 force field. By experimental determination of free energies of transfer from external collaborators, we are able to validate bead assignment of many different chemical moieties in a more accurate way than in Martini 2. Additionally, we show that our final model, revolving around a hydrophobic virtual site description, improves on the Martini 2 carbohydrates where aggregation is concerned. The use of R-beads over S-beads in the previous model seems to be a large contributing factor to the aggregation problems in Martini 2, as we show that R-beads are significantly more hydrophobic. Our model also improves in the level of detail by which glycosidic linkages are described. The Martini 3 carbohydrate model distinguishes between α and β-linkages, and can describe all glycosidic linkages without any problems. The addition of a hydrophobic virtual site should lead to better affinity for systems such as the inclusion of cholesterol in the inner face of β-cyclodextrin[26], or the binding of carbohydrates to proteins. We have yet to fully explore complex systems, which is the next step now that the mapping definitions and parameterization are finished.

6 | Final modelling guidelines

6.1 Monosaccharides

6.1.1 Mapping principles

Based on the validation principles, the following general mapping rules for monosaccharides are suggested:

- Mapping starts at the lowest molecular weight substituent next to the ether oxygen, which becomes bead "A"
- Mapping always goes in the direction of the ether oxygen, making that the last bead of the ring structure, and thus always bead "C"
- A focus is put on matching diols going in a circle
- Functional groups are kept together, whenever possible
- In the case of substituent groups on the base five/six carbon ring, the ring is mapped first, moving on to substituents in the order of the ring mapping. I.e. N-acetylneuraminic acid (*Figure 6.1H*), has bead "D" connected to bead "A", "E" to "B", et cetera
- Internal bonds (i.e. between bead "A", "B" and "C") are defined as constraints while substituent beads are characterized by bonds and simple angle potentials that allow for flexibility of substituent groups
- Nrexcl = 3 is used, meaning the exclusion of non-bonded interactions between atoms that are no further than 3 bonds away
- A TC4 virtual site is included in the geometrical center of the ring and internal constraints are scaled by 15%

Representative monosaccharides and bead ordering are presented in *Figure 6.1*:

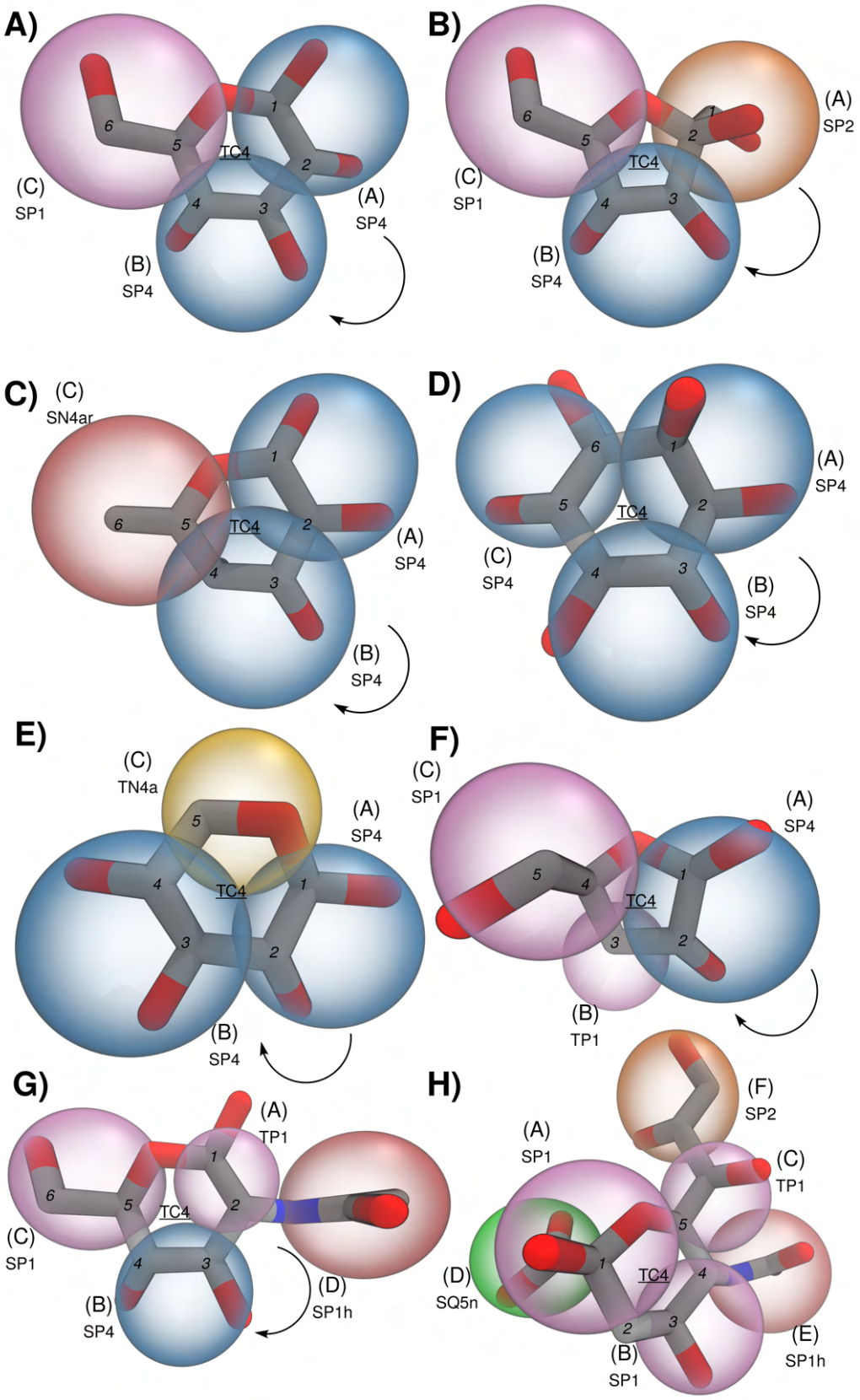

Figure 6.1: The mapping approach for representative monosaccharides as modelled in this study are showcased. **A)** D-glucose. **B)** D-fructose. **C)** L-fucose. **D)** Myo-inositol. **E)** D-xylopyranose. **F)** D-ribose. **G)** N-Acetylglucosamine. **H)** N-Acetylneuraminic acid

6.2 Disaccharides & polysaccharides

6.2.1 Mapping principles

As has been explained in *Chapter 3*, three main types of glycosidic linkages can be distinguished where α- and β-linkages are treated separately (see *Table 6.1*). The loss of a heavy atom through glycosidic linkage affects the mapping scheme. The glycosidic oxygen is taken in a bead with the first monosaccharide (named "A", "B" or "C", depending on which linkage class is present) as an S-bead while the recipient bead is modelled by a T-bead. Through a difference in ordering of the mapping scheme (*Figure 6.1*), a clear distinction between the three linkage classes is possible.

Class	Glycosidic linkage
Class 1	α1,1 & α1,2
Class 1	β1,1 & β1,2
Class 2	α1,3 & α1,4
Class 2	β1,3 & β1,4
Class 3	α1,5 & α1,6
Class 3	β1,5 & β1,6

Table 6.1: The six groups of glycosidic linkages and their mapping approach (showcased in the class type) are highlighted. Due to the more flexible nature of α-glycosidic linkages, they are treated separately from β-linkages. Further simplification follows from the division of the six different carbon linkages into three main types of classes. By changing the bead assignment, each class is easily mapped and recognized.

All internal constraints (so not the glycosidic bond) are scaled by 15%. The mapping direction follows the opposite direction of the glycosidic linkage, in accordance with how atomistic glycans are constructed. In practice this means that e.g. β1,4 linkages are mapped in the $4 \longrightarrow 1$ direction (see *Figure 6.2*)

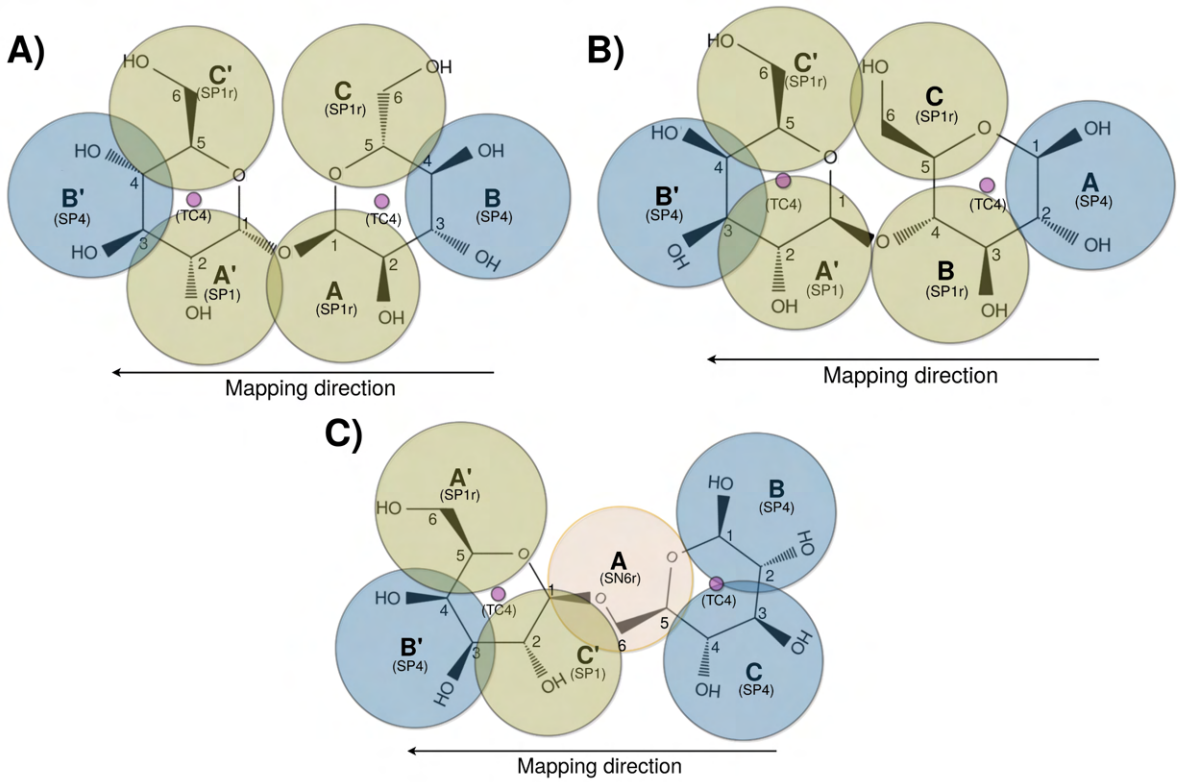

Figure 6.2: The three main types of linkages in the Martini 3 carbohydrate model are presented here. **A)** Class 1 encompasses the x1,1/x1,2 linkages where glycosidic bonds are drawn between the neighbouring beads A and A'. **B)** Class 2 encompasses the x1,3/x1,4 linkages where bonds are drawn between the neighbouring beads B and A'. **C)** Class 3 encompasses the x1,5/x1,6 linkages where bonds are drawn between beads A and C'

6.2.2 Bonded potentials

While internal monosaccharide bonds are defined as constraints in GROMACS, glycosidic bonds are intrinsically more flexible, and need to be explicitly modelled as such. Furthermore, a neighbour exclusion of 3 is defined, excluding non-bonded interactions between atoms that are no further than three bonds away. A simple bonded potential scheme is suggested, where four angle potentials (GROMACS function 1) and one proper dihedral potential (GROMACS function 1) seem to define the conformational space of carbohydrates well, with no major instabilities reported. Using a Class 3 linkage as an example, *Figure 6.3* highlights which potential controls each specific rotation. It should be noted that for polysaccharides, an extra dihedral potential spanning from n to the n + 2 neighbour is required that controls this type of

rotation.

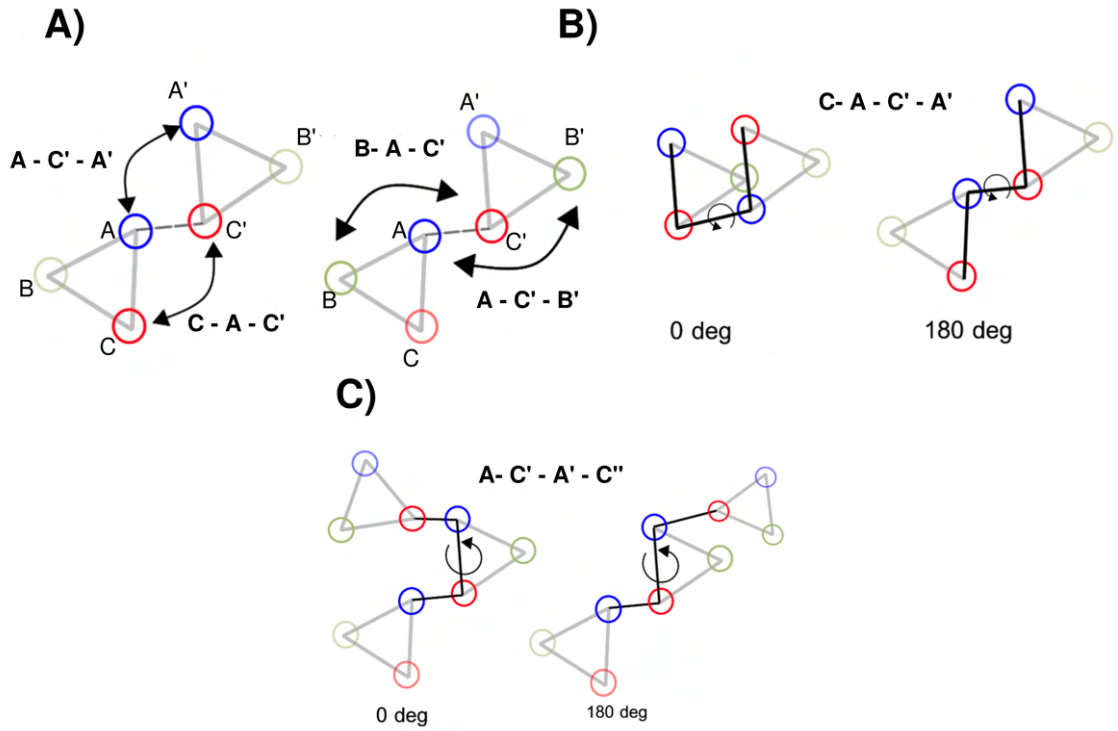

Figure 6.3: All the bonded potentials required to cover the conformational space for Class 3 linkages are presented. **A)** The four angle potentials control rotation of the two sugar rings with respect to each other. **B)** The dihedral potential captures flipping of sugar beads across the glycosidic bond. **C)** The additional dihedral potential captures rotation from n with respect to the $n + 2$ neighbour.

By extrapolating the potential scheme as outlined in *Figure 6.3*, the potentials for the other linkage classes logically follow. These are presented in *Table 6.2* for additional clarity. It should be further noted that in the case of Class 1 potentials, homogenous polysaccharides do not exist due to all monosaccharides having their reducing end already bound. In the case of heterogeneous polysaccharides/glycans, a generic dihedral is desirable that will require explicit testing in complex systems, such as glycosylated proteins.

	Class 1	Class 2	Class 3
Type	**Potential**	**Potential**	**Potential**
Bond	A - A′	B - A′	A - C′
Ang 1	B - A - A′	A - B - A′	C - A - C′
Ang 2	A - A′ - B′	B - A′ - B′	A - C′ - A′
Ang 3	A - A′ - C′	B - A′ - C′	A - C′ - B′
Ang 4	C - A - A′	C - B - A′	B - A - C′
Dih 1	B - A - A′ - C′	A - B - A′ - B′	C - A - C′ - A′
Dih 2	-	B - A′ - B′ - A″	A - C′ - A′ - C″

Table 6.2: The potential scheme suggested here, applied to the main three classes of linkages. Due to specific mapping of each different linkage type, different linkages can quickly be distinguished.

7 | Methods

7.1 Simulation details

All simulations were performed with GROMACS (2018.8)[52,53]. Atomistic simulations (AA) were performed with the GLYCAM06[54], CHARMM36[55] and GROMOS54a7[56] force fields. Coarse-grained (CG) simulations were performed with the Martini 3 force field[4], or Martini 2 force field[57]. Visualization of the trajectories was done in VMD[58] and PyMOL (DeLano Scientific, Palo Alto, CA,USA)

7.1.1 GLYCAM06 setup

Each sugar was placed in the middle of a box of size 2.4 x 2.4 x 2.4 nm^3 and then solvated using the TIP3P water model[58]. Periodic boundary conditions were used and a 1 nm cutoff for nonbonded interactions was employed. Simulations were performed at constant pressure of 1 atm and temperature of 310 K, using the Parrinello-Rahman barostat[59,60] , and velocity-rescaling algorithm[61], respectively. Each simulation was ran for 200 ns, following an equilibration stage with position restraints applied on all sugar atoms. The LINCS algorithm[62] was used to solve the constraints. Glucose, mannose, N-Acetylglucosamine (GlcNAc) and N-Acetylneuraminic acid (Neu5Ac) were simulated using this force field.

7.1.2 Basic CHARMM36 setup

Each CHARMM36 simulation used the following settings, unless specified otherwise. The CHARMM-GUI[63] Glycan reader[64] was employed to prepare the simulations. The TIP3P[65] water model was used to solvate the structure in such a way that each atom was at least 10 Å removed from the box boundaries. Periodic boundary conditions were used and for treatment of long-range electrostatic interactions, the particle-mesh Ewald method[66,67] was employed. Electrostatic

interactions were cutoff at 1.2 nm using the Verlet buffer scheme. The twin range cutoff scheme with a cutoff value of 1.2 nm was used for short range van der Waals interactions. Simulations were performed at constant pressure of 1 atm and a temperature of 303.15 K, using the Parrinello-Rahman barostat[59,60] , and Nosé-Hoover thermostat[68,69], respectively. All systems were equilibrated in the NPT ensemble using position restraints for 1.25 ns, which were removed in the production simulation. All systems were ran for at least 200 ns, with an integration timestep of 2 fs. The LINCS algorithm[62] was used to solve the constraints. All carbohydrates not mentioned in GLYCAM06 and GROMOS54a7 were simulated using these settings.

7.1.3 GROMOS54a7 setup

Structures were prepared using the ATB[56]. The Verlet cutoff scheme[70] and reaction field method[71], were employed, both with a cut-off of 1.4 nm and relative permittivity ε_r set to 61. The SPC water model[72] was used to solvate the structure in such a way that each atom was at least 10 Å removed from the box boundaries. Simulations were performed at constant pressure of 1 atm and a temperature of 298.15 K, using the Parrinello-Rahman barostat[59,60], and Nosé-Hoover thermostat[68,69], respectively. All systems were ran for at least 200 ns, with an integration timestep of 2 fs. The LINCS algorithm[62] was used to solve the constraints. Sucrose and inositol were simulated using this force field.

7.1.4 Basic Martini 3 setup

Each Martini 3 simulation followed the standard simulation settings as outlined in the main publication[4] and the "new" set of parameters as outlined by *de Jong et al, 2016*[73], unless specified otherwise. The Verlet cutoff scheme[70] using a cutoff of 1.1 nm was used for the Lennard-Jones terms. Electrostatic interactions were treated with reaction field[71], with $\varepsilon_r = 15$ and a cutoff value of 1.1 nm. The velocity scaling thermostat[61] and Parrinello-Rahman barostat were used to maintain temperature

and pressure respectively. Equilibration was performed in two stages, 2.5 ns long each, first using an integration timestep of 10 fs and then 20 fs.

7.2 Parameterization of bonded interactions

7.2.1 Mapping approach

Mapping files were prepared using pycgbuilder (https://github.com/marrink-lab/pycgbuilder). Mapping always followed the general guidelines as outlined in *Chapter 6*. In the case of monosaccharides, bead assignment was iteratively validated with partition data until the best possible agreement could be found. In case of known accurate assignment, only unknown chemical moieties were touched.

Initial bonded parameters were derived using pycgmap (https://github.com/fgrunewald/pycgmap), which employs the following summarized methodology: a simulation trajectory and mapping files may be supplied. The coarse-grained beads are then placed at the center of geometry of the constituting atoms (H-atoms included), following the Martini 3 guidelines[4]. Using Boltzmann inversion (BI)[74], an initial estimate of the bonded parameters could be made. The parameters were then iteratively optimized for close agreement of their potential distributions with the atomistic reference simulation.

7.2.2 Modelled potentials

Three types of bonded interactions were modelled. Bonds were described by a harmonic potential as shown in *eq. 7.1*:

$$V_{bond} = \frac{1}{2} K_{bond} (r - r_{bond})^2 \tag{7.1}$$

where r_{bond} and K_{bond} are the equilibrium distance and force constant respectively. Internal bonds are constrained at the equilibrium length while glycosidic bonds are treated explicitly by V_{bond}. Angles were defined over three beads that crossed the

glycosidic bond, described by the potential V_{angle} (eq. *7.2*).

$$V_{angle} = \frac{1}{2} K_{angle} (\theta - \theta_{angle})^2 \tag{7.2}$$

where θ_{angle} and K_{angle} are the equilbrium angle and force constant respectively. Proper dihedrals (GROMACS functions 1 and 9) were described using the potential $V_{dihedral}$ (eq. *7.3*)

$$V_{dihedral} = K_{dihedral}(1 + cos(n \cdot \theta - \theta_{dihedral})) \tag{7.3}$$

where $K_{dihedral}$ is the force constant, n is the multiplicity and $\theta_{dihedral}$ is the equilibrium angle between planes defined by the beads i, j, k and j, k, l respectively.

7.3 SASA and volume calculations

7.3.1 SASA

Solvent Accessible Surface Area (SASA) calculations were performed on PBC-corrected, at least 200 ns long, trajectories using the *gmx sasa*[33] package, as implemented in GROMACS. Experimental van der Waals radii[34] were used for the atomistic trajectory. For the Martini model, custom radii were defined via the definition:

$$r_{vdW} = r_{min} = \frac{d_{min}}{2} = \frac{\sqrt[6]{2}\sigma}{2} \tag{7.4}$$

where σ is the radius of a certain bead type.

The resulting radii are collected in *Table 7.1*. The probe size was kept at 0.191 nm in all calculations as that is the radius of a T-bead and thus the smallest possible particle that interacts in Martini.

Bead type	VDW radius (nm)
T-bead	0.191
S-bead	0.230
N-bead	0.264

Table 7.1: The van der Waals radii for Martini beads as employed in SASA calculations

7.4 Free energies of transfer

7.4.1 Experimental methodology

Partition coefficients of the free energies of transfer $\Delta\Delta G_{OCO\rightarrow W}$ were collected via the methodology as outlined by *Virtanen et al, 2020*, at a temperature of 298.15 K and 4% octanol hydration. For a full description on the methodology, see *Appendix A1.3.1*. Log P values can be transformed to the free energy of transfer $\Delta\Delta G_{OCO\rightarrow W}$ (kJ mol^{-1}) via the relation

$$\Delta\Delta G_{OCO\rightarrow W} = 2.203\, R\, T \log P \tag{7.5}$$

where R is the gas constant (kJ mol^{-1} K^{-1}) and T the temperature (K). Final partition values are collected in *Table A1.6*.

7.4.2 MD methodology

The free energy of transfer ($\Delta\Delta G_{OCO\rightarrow W}$) for a compound is defined as the difference in the solvation free energy of the water (ΔG_W) phase and 5w% octanol (ΔG_{OCO}) phase, which describes the free energy change from vacuum to solvent.

The solvation free energies were calculated as the free energy difference ΔF of the sugar in vacuum (state A) and the solvated phase (state B), using the thermodynamic integration (TI)[75] procedure (eq. *7.6*)

$$\Delta F_{BA} = \Delta F_B - \Delta F_A = \int_{\lambda_\mathrm{A}}^{\lambda_\mathrm{B}} d\lambda \left\langle \frac{\partial U_{\mathrm{uv}}(\lambda)}{\partial \lambda} \right\rangle_\lambda \tag{7.6}$$

The potential energy function ∂U_{uv} describes the interactions between sugar and solvent and is averaged over a specific simulation trajectory. λ is defined as a coupling parameter that controls the level of ∂U_{uv}, which varies from no ($\lambda_A = 0$) to full ($\lambda_B = 1$) interaction.

The calculations were performed using 19 λ values (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.55, 0.6 0.65, 0.7, 0.75, 0.775, 0.8, 0.825, 0.85, 0.875, 0.9, 0.95 and 1.0), at 298.15 K. For each window, 12 ns of simulation trajectory was used. The free energy difference was estimated using Bennett's Acceptance Ratio (BAR)[76] as implemented in the *gmx bar* package.

7.5 Osmotic pressures

7.5.1 Pressure calculations

The osmotic coefficient was computed from simulations adopting the protocol as originally proposed by Luo and Roux[77]. A rectangular box was created in which the solute molecules were confined in the z-direction by a flat-bottomed potential to the center of the box. At a distance of 2.52078 nm from the center of the box, a harmonic force with a force constant of 1000 kJ mol^{-1} nm^{-2} was applied to the solute molecules. The box dimensions were taken to be 10.08312 nm and 5.04156 nm, in the z dimension and $x + y$ direction respectively. Previous to each run, a random configuration of solute and solvent molecules was created with Polyply (https://github.com/marrink-lab/polyply_1.0) placing solute molecules only in the center of the box, and the solvent in the entire box.

After energy minimization, this setup was subjected to 10 ns of equilibration using the Berendsen barostat[78]. Production simulations were run for 500 ns at a pressure of 1 bar and coupled to a temperature of the respective experimental reference values using the velocity rescaling algorithm[61]. Following Sauter and Grafmüller[29], the pressure was coupled only in the z-dimensions using the Parrinello-Rahman

barostat[59]. In the original implementation, the simulation was performed in the NVT ensemble. The particle-mesh Ewald method[66,67] was used in all simulations. The osmotic pressure was computed from the trajectory by recalculating the total force exerted by the solute particles onto the flat-bottomed potential averaged over the two potentials. Subsequently, that force was divided by the xy area of the box. The ensemble average as well as an error were computed from the time-series of the osmotic pressure.

The osmotic pressures were calculated as a function of the concentration in molality, determined by the expression

$$m = \frac{\frac{N}{N_A}}{\rho_W V} \tag{7.7}$$

where N is the absolute number of solute molecules in the system, N_A is Avogadro's constant, ρ_W the density of water and V the volume of the box.

7.5.2 Experimental osmotic pressures

Experimental osmotic coefficients (ϕ) for many representative carbohydrates may be found in literature and are commonly expressed a function of molality. Relating the osmotic coefficient to osmotic pressure follows the expression

$$\Pi = \phi\, c\, R\, T \tag{7.8}$$

where Π is the osmotic pressure in bar, ϕ the osmotic coefficient, c the solute concentration (m), R the gas constant (L bar K^{-1} mol^{-1}) and K the temperature (K).

References

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

[2] Javanainen, M., Martinez-Seara, H., & Vattulainen, I. (2017). Excessive aggregation of membrane proteins in the Martini model. *PLOS ONE*, *12*(11), e0187936.

[3] Schmalhorst, P. S., Deluweit, F., Scherrers, R., Heisenberg, C.-P., & Sikora, M. (2017). Overcoming the Limitations of the MARTINI Force Field in Simulations of Polysaccharides. *Journal of Chemical Theory and Computation*, *13*(10), 5039–5053. PMID: 28787166.

[4] Souza, P. C. T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B. M. H., Wassenaar, T. A., Kroon, P. C., Melcr, J., Nieto, V., Corradi, V., Khan, H. M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., Best, R. B., Vattulainen, I., Monticelli, L., Periole, X., Tieleman, D. P., de Vries, A. H., & Marrink, S. J. (2021). Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods*, *18*(4), 382–388.

[5] de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., & Marrink, S. J. (2012). Improved Parameters for the Martini Coarse-Grained Protein Force Field. *Journal of Chemical Theory and Computation*, *9*(1), 687–697.

[6] Maughan, R. (2009). Carbohydrate metabolism. *Surgery (Oxford)*, *27*(1), 6–10. Hepatopancreatobiliary I.

[7] Rini, J. M., & Leffler, H. (2010). Chapter 13 - Carbohydrate recognition and signaling. In *Handbook of Cell Signaling (Second Edition)*, (pp. 85–91). San Diego: Academic Press, second ed.

[8] Ruiz-Herrera, J., & Ortiz-Castellanos, L. (2019). Cell wall glucans of fungi. a review. *The Cell Surface*, *5*, 100022.

[9] Hang, I., wei Lin, C., Grant, O. C., Fleurkens, S., Villiger, T. K., Soos, M., Morbidelli, M., Woods, R. J., Gauss, R., & Aebi, M. (2015). Analysis of site-specific N-glycan remodeling in the endoplasmic reticulum and the Golgi. *Glycobiology*, *25*(12), 1335–1349.

[10] Lee, J. H., Ozorowski, G., & Ward, A. B. (2016). Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. *Science*, *351*(6277), 1043–1048.

[11] van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glättli, A., Hünenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F. A., & Yu, H. B. (2006). Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie International Edition*, *45*(25), 4064–4092.

[12] Asensio, J. L., Ardá, A., Cañada, F. J., & Jiménez-Barbero, J. (2012). Carbohydrate-Aromatic Interactions. *Accounts of Chemical Research*, *46*(4), 946–954.

[13] Houser, J., Kozmon, S., Mishra, D., Hammerová, Z., Wimmerová, M., & Koča, J. (2020). The CH-π interaction in Protein-Carbohydrate Binding: Bioinformatics and In Vitro Quantification. *Chemistry - A European Journal*, *26*(47), 10769–10780.

[14] Fernández-Alonso, M. d. C., Cañada, F. J., Jiménez-Barbero, J., & Cuevas, G. (2005). Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate-aromatic interactions. *Journal of the American Chemical Society*, *127*(20), 7379–7386. PMID: 15898786.

[15] Hudson, K. L., Bartlett, G. J., Diehl, R. C., Agirre, J., Gallagher, T., Kiessling, L. L., & Woolfson, D. N. (2015). Carbohydrate-Aromatic Interactions in Proteins. *Journal of the American Chemical Society*, *137*(48), 15152–15160.

[16] Harata, K., & Muraki, M. (2000). Crystal structures of Urtica dioica agglutinin and its complex with tri-N-acetylchitotriose. *Journal of Molecular Biology*, *297*(3), 673–681.

[17] Souza, P. C. T., Thallmair, S., Conflitti, P., Ramírez-Palacios, C., Alessandri, R., Raniolo, S., Limongelli, V., & Marrink, S. J. (2020). Protein–ligand binding with the coarse-grained Martini model. *Nature Communications*, *11*(1).

[18] Domicevica, L., Koldsø, H., & Biggin, P. C. (2018). Multiscale molecular dynamics simulations of lipid interactions with P-glycoprotein in a complex membrane. *Journal of Molecular Graphics and Modelling*, *80*, 147–156.

[19] Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236.

[20] Ingólfsson, H. I., Melo, M. N., van Eerden, F. J., Arnarez, C., Lopez, C. A., Wassenaar, T. A., Periole, X., de Vries, A. H., Tieleman, D. P., & Marrink, S. J. (2014). Lipid organization of the plasma membrane. *Journal of the American Chemical Society*, *136*(41), 14554–14559.

[21] Alessandri, R., Souza, P. C. T., Thallmair, S., Melo, M. N., de Vries, A. H., & Marrink, S. J. (2019). Pitfalls of the Martini model. *Journal of Chemical Theory and Computation*, *15*(10), 5448–5460. PMID: 31498621.

[22] López, C. A., Rzepiela, A. J., de Vries, A. H., Dijkhuizen, L., Hünenberger, P. H., & Marrink, S. J. (2009). Martini Coarse-Grained Force Field: Extension to Carbohydrates. *Journal of Chemical Theory and Computation*, *5*(12), 3195–3210. PMID: 26602504.

[23] López, C. A., Bellesia, G., Redondo, A., Langan, P., Chundawat, S. P. S., Dale, B. E., Marrink, S. J., & Gnanakaran, S. (2015). MARTINI Coarse-Grained Model for Crystalline Cellulose Microfibers. *The Journal of Physical Chemistry B*, *119*(2), 465–473. PMID: 25417548.

[24] Vaiwala, R., Sharma, P., Puranik, M., & Ayappa, K. G. (2020). Developing a Coarse-Grained Model for Bacterial Cell Walls: Evaluating Mechanical Properties and Free Energy Barriers. *Journal of Chemical Theory and Computation*, *16*(8), 5369–5384.

[25] López, C. A., de Vries, A. H., & Marrink, S. J. (2011). Molecular Mechanism of Cyclodextrin Mediated Cholesterol Extraction. *PLOS Computational Biology*, *7*(3), 1–11.

[26] López, C. A., de Vries, A. H., & Marrink, S. J. (2013). Computational microscopy of cyclodextrin mediated cholesterol extraction from lipid model membranes. *Scientific Reports*, *3*(1).

[27] Shivgan, A. T., Marzinek, J. K., Huber, R. G., Krah, A., Henchman, R. H., Matsudaira, P., Verma, C. S., & Bond, P. J. (2020). Extending the Martini Coarse-Grained Force Field to N-Glycans. *Journal of Chemical Information and Modeling*, *60*(8), 3864–3883. PMID: 32702979.

[28] Best, R. B., Zheng, W., & Mittal, J. (2014). Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation*, *10*(11), 5113–5124.

[29] Sauter, J., & Grafmüller, A. (2015). Solution Properties of Hemicellulose Polysaccharides with Four Common Carbohydrate Force Fields. *Journal of Chemical Theory and Computation*, *11*(4), 1765–1774.

[30] Lay, W. K., Miller, M. S., & Elcock, A. H. (2016). Optimizing Solute–Solute Interactions in the GLYCAM06 and CHARMM36 Carbohydrate Force Fields

Using Osmotic Pressure Measurements. *Journal of Chemical Theory and Computation*, *12*(4), 1401–1407.

[31] Hölemann, A., & Seeberger, P. H. (2004). Carbohydrate diversity: synthesis of glycoconjugates and complex carbohydrates. *Current Opinion in Biotechnology*, *15*(6), 615–622.

[32] Becker, D. J., & Lowe, J. B. (2003). Fucose: biosynthesis and biological function in mammals. *Glycobiology*, *13*(7), 41R–53R.

[33] Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., & Scharf, M. (1995). The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, *16*(3), 273–284.

[34] Rowland, R. S., & Taylor, R. (1996). Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *The Journal of Physical Chemistry*, *100*(18), 7384–7391.

[35] Connolly, M. L. (1983). Analytical molecular surface calculation. *Journal of Applied Crystallography*, *16*(5), 548–558.

[36] Jamehbozorg, B., & Sadeghi, R. (2018). Evaluation of the effect of carbohydrates as renewable, none-charged and non-toxic soluting-out agents on the ionic-liquid-based ABS implementation. *Journal of Molecular Liquids*, *255*, 476–491.

[37] Miyajima, K., Sawada, M., & Nakagaki, M. (1983). Studies on Aqueous Solutions of Saccharides. I. Activity Coefficients of Monosaccharides in Aqueous Solutions at 25 °C. *Bulletin of the Chemical Society of Japan*, *56*(6), 1620–1623.

[38] Stokes, R. H., & Robinson, R. A. (1966). Interactions in Aqueous Nonelectrolyte Solutions. I. Solute-Solvent Equilibria. *The Journal of Physical Chemistry*, *70*(7), 2126–2131.

[39] Barone, G., Cacace, P., Castronuovo, G., Elia, V., & Iappelli, F. (1981). Excess thermodynamic properties of aqueous solutions of l-fucose and l-rhamnose at 25°. *Carbohydrate Research*, *93*(1), 11–18.

[40] Ebrahimi, N., & Sadeghi, R. (2016). Osmotic properties of carbohydrate aqueous solutions. *Fluid Phase Equilibria*, *417*, 171–180.

[41] Silva, A. M., da Silva, E. C., & da Silva, C. O. (2006). A theoretical study of glucose mutarotation in aqueous solution. *Carbohydrate Research*, *341*(8), 1029–1040.

[42] Bruce Grindley, T. (2001). *Structure and Conformation of Carbohydrates*, (pp. 3–51). Berlin, Heidelberg: Springer Berlin Heidelberg.

[43] Juaristi, E., & Cuevas, G. (1992). Recent studies of the anomeric effect. *Tetrahedron*, *48*(24), 5019–5087.

[44] Pezeshkian, W., Hansen, A. G., Johannes, L., Khandelia, H., Shillcock, J. C., Kumar, P. B. S., & Ipsen, J. H. (2016). Membrane invagination induced by Shiga toxin B-subunit: from molecular structure to tube formation. *Soft Matter*, *12*(23), 5164–5171.

[45] Feenstra, K. A., Hess, B., & Berendsen, H. J. C. (1999). Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *Journal of Computational Chemistry*, *20*(8), 786–798.

[46] Borges-Araújo, L., Souza, P., Fernandes, F., & Melo, M. N. Improved Parameterization of Phosphatidylinositide Lipid Headgroups for the Martini 3 Coarse Grain Force Field.

[47] Hardy, B. J. (1997). The glycosidic linkage flexibility and time-scale similarity hypotheses. *Journal of Molecular Structure: THEOCHEM*, *395-396*, 187–200.

[48] Pendrill, R., Säwén, E., & Widmalm, G. (2013). Conformation and Dynamics at a Flexible Glycosidic Linkage Revealed by NMR Spectroscopy and Molecular

Dynamics Simulations: Analysis of β-l-Fucp-(1$\rightarrow$6)-α-d-Glcp-OMe in Water Solution. *The Journal of Physical Chemistry B*, *117*(47), 14709–14722.

[49] Melo, M. N., Ingólfsson, H. I., & Marrink, S. J. (2015). Parameters for Martini sterols and hopanoids based on a virtual-site description. *The Journal of Chemical Physics*, *143*(24), 243152.

[50] Höög, C., Landersjö, C., & Widmalm, G. (2001). Oligosaccharides Display Both Rigidity and High Flexibility in Water as Determined by 13C NMR Relaxation and 1H, 1H NOE Spectroscopy: Evidence of *anti*-Φ and *anti*-Ψ Torsions in the Same Glycosidic Linkage. *Chemistry - A European Journal*, *7*(14), 3069–3077.

[51] Kabayama, M. A., & Patterson, D. (1958). THE THERMODYNAMICS OF MUTAROTATION OF SOME SUGARS: II. THEORETICAL CONSIDERATIONS. *Canadian Journal of Chemistry*, *36*(3), 563–573.

[52] Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., & Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, *29*(7), 845–854.

[53] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25.

[54] Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., & Woods, R. J. (2007). GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, *29*(4), 622–655.

[55] Guvench, O., Mallajosyula, S. S., Raman, E. P., Hatcher, E., Vanommeslaeghe, K., Foster, T. J., Jamison, F. W., & MacKerell, A. D. (2011). CHARMM Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in

Polysaccharide and Carbohydrate–Protein Modeling. *Journal of Chemical Theory and Computation*, *7*(10), 3162–3180.

[56] Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., Oostenbrink, C., & Mark, A. E. (2011). An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *Journal of Chemical Theory and Computation*, *7*(12), 4026–4037. PMID: 26598349.

[57] Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & de Vries, A. H. (2007). The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B*, *111*(27), 7812–7824. PMID: 17569554.

[58] Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38.

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

[60] Nosé, S., & Klein, M. (1983). Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, *50*(5), 1055–1076.

[61] Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, *126*(1), 014101.

[62] Hess, B., Bekker, H., Berendsen, H. J. C., & Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, *18*(12), 1463–1472.

[63] Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, *29*(11), 1859–1865.

[64] Park, S.-J., Lee, J., Qi, Y., Kern, N. R., Lee, H. S., Jo, S., Joung, I., Joo, K., Lee, J., &

Im, W. (2019). CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, *29*(4), 320–331.

[65] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*(2), 926–935.

[66] Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092.

[67] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, *103*(19), 8577–8593.

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

[69] Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, *31*, 1695–1697.

[70] Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, *159*(1), 98–103.

[71] Tironi, I. G., Sperb, R., Smith, P. E., & van Gunsteren, W. F. (1995). A generalized reaction field method for molecular dynamics simulations. *The Journal of Chemical Physics*, *102*(13), 5451–5459.

[72] Toukan, K., & Rahman, A. (1985). Molecular-dynamics study of atomic motions in water. *Physical Review B*, *31*(5), 2643–2648.

[73] de Jong, D. H., Baoukina, S., Ingólfsson, H. I., & Marrink, S. J. (2016). Martini straight: Boosting performance using a shorter cutoff and GPUs. *Computer Physics Communications*, *199*, 1–7.

[74] Moore, T. C., Iacovella, C. R., & McCabe, C. (2014). Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *The Journal of Chemical Physics*, *140*(22), 224104.

[75] van Gunsteren, W. F., & Berendsen, H. J. C. (1987). Thermodynamic cycle integration by computer simulation as a tool for obtaining free energy differences in molecular chemistry. *Journal of Computer-Aided Molecular Design*, *1*(2), 171–176.

[76] Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, *22*(2), 245–268.

[77] Luo, Y., & Roux, B. (2009). Simulation of Osmotic Pressure in Concentrated Aqueous Salt Solutions. *The Journal of Physical Chemistry Letters*, *1*(1), 183–189.

[78] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, *81*(8), 3684–3690.

[79] Virtanen, V., & Karonen, M. (2020). Partition Coefficients (logP) of Hydrolysable Tannins. *Molecules*, *25*(16), 3691.

A1 | Appendix

A1.1 Monosaccharides

A1.1.1 Bonded parameters

The final bead assignment of the monosaccharides is summarized in *Table A1.2*

Sugar	**Code**	**Bond A-B**	**Bond A-C**	**Bond B-C**	**Bond A-D**	**ANG C-A-D**
D-glucose	GLC	0.380	0.470	0.396	-	-
D-mannose	MAN	0.373	0.450	0.394	-	-
D-galactose	GAL	0.361	0.478	0.381	-	-
L-fucose	LFUC	0.359	0.454	0.361	-	-
L-rhamnose	LRHA	0.367	0.432	0.367	-	-
Fructose	FRU	0.376	0.454	0.357	-	-
D-ribose	RIB	0.320	0.400	0.324	-	-
D-xylose	XYL	0.380	0.362	0.315	-	-
Inositol	INO	0.369	0.375	0.386	-	-
GlcNAc/GalNAc	GYN	0.392	0.427	0.397	0.339 | 4700	144 | 100
Sugar	**Code**	**Bond A-B**	**Bond A-C**	**Bond B-C**	**Bond A-D**	
Neu5Ac	NMC	0.331	0.424	0.397	0.251	
		Bond B-E	**Bond C-F**	**Ang B-A-D**	**Ang B-C-F**	**Ang A-B-E**
		0.336 | 6000	0.273 | 11500	96 | 350	144 | 100	141 | 150

Table A1.1: Parameters belonging to monosaccharides. Bonds that make up the main monosaccharide ring (mapped by beads "A", "B" and "C") are defined as constraints. Subsituent groups are defined as bonds with angles to prevent flipping of the bead across the bond. Units: length (nm), angle (°), force constant (kJ mol^{-1} nm^{-2}). All internal constraints are scaled by 15%.

A1.1.2 Bead assignment final model (TC4)

Code	Mapping
GLC	SP4-SP4-SP1-TC4
MAN	SP4-SP4-SP1-TC4
GAL	SP4-SP4-SP1-TC4
FRUF	SP2-SP4-SP1-TC4
LFUC	SP4-SP4-SN4ar-TC4
LRHA	SP4-SP4-SN4ar-TC4
RIBF	SP4-TP1-SP1-TC4
XYL	SP4-SP4-TN4a-TC4
INO	SP4-SP4-SP4-TC4
GYN	TP1-SP4-SP1-SP1h-TC4
NMC	SP1-SP1-TP1-SQ5n-SP1h-SP2-TC4

Table A1.2: Final bead assignment for the Martini 3 monosaccharides in this study, with 15% scaling of internal bonds.

A1.1.3 Bead assignment of models considered.

Code	S-bead no VS	S-bead TC4	S-bead TC4 (15%)	N-bead no VS
GLC	SP3-SP3-SP1	SP3-SP3-SP2-TC4	SP4-SP4-SP1-TC4	P3-P3 -P2
MAN	SP3-SP3-SP1	SP3-SP3-SP2-TC4	SP4-SP4-SP1-TC4	P3-P3-P2
GAL	SP3-SP3-SP1	SP3-SP3-SP2-TC4	SP4-SP4-SP1-TC4	P3-P3-P2
FRUF	SP2-SP3-SP1	SP2-SP3-SP1-TC4	SP2-SP4-SP1-TC4	-
FUC	SP3-SP3-SN3ar	SP3-SP3-SN3ar-TC4	SP4-SP4-SN4ar-TC4	P3-P3-N3ar
RIBF	SP3-TP1-SP1	SP3-TP1-SP1-TC4	SP4-TP1-SP1-TC4	-
XYL	SP3-SP3-TN2a	SP3-SP3-TN4a-TC4	SP4-SP4-TN4a-TC4	P3-P3-TN3ar
INO	SP3-SP3-SP3	SP3-SP3-SP3-TC4	SP4-SP4-SP4-TC4	P3-P3-P3
GYN	TP1-SP3-SP1-P2	TP1-SP3-SP2-P2-TC4	TP1-SP4-SP1-SP1h-TC4	-
NMC	SP1-TP1-TP1-SP2-P2-SP2	SP1-TP1-TP1-SP2-P2-SP2-TC4	SP1-SP1-TP1-SP2-SP1h-SP2-TC4	-

Table A1.3: Final bead assignments of the four main models applied to the different monosaccharides, corresponding to the free energies of transfer as outlined in *Table A1.5*. Bead assignment follows the order"A"-"B"-"C" etc.

A1.2 SASA values

Sugar	Code	AA SASA	S-beads	S-beads (15%)	N-beads	M2
D-glucose	GLC	4.08	3.75	4.06	4.26	4.22
D-mannose	MAN	4.05	3.75	4.06	4.23	
D-galactose	GAL	4.08	3.75	4.06	4.26	
GlcNAc	GYN	4.78	4.35	4.56		
Neu5Ac	NMC	5.63	5.44	5.62		
L-fucose	FUC	3.92	3.77	3.91	4.16	
L-rhamnose	RHA	3.92	3.77	3.91	4F.16	
D-xylose	XYL	3.66	3.38	3.59	3.75	
Inositol	INO	4	3.63	3.85	4.1	
D-fructose	FRUF	4.08	3.7	3.94	4.18	4.17
D-ribose	RIBF	3.7	3.37	3.58	3.76	
Trehalose	TREH	5.97	5.41	5.87		
Sucrose	SUCR	5.9	5.28	5.6		6.2
Lactose	LAC	6.07	5.54	6	6.18	

Table A1.4: SASA values of different models, given in $Å^2$. "AA SASA" refers to the reference atomistic model (see *Section 7.1* for a full overview of which carbohydrate was modelled by which force field)

.

A1.3 Free energies of transfer

A1.3.1 Experimental methodology

Measurements were performed by researchers at the University of Turku, following a similar methodology as outlined in *Virtanen et al, 2020*[79]. Measurements were done with an UPLC-DAD-HESI-Orbitrap-MS instrument. The column in the UPLC was an Aquity BEH Phenyl (100 × 2.1 mm i.d., 1.7 µm) and the mobile phase consisted of acetonitrile (A) and 0.1% aqueous formic acid (B). The elution gradient was carried out with a constant flow rate of 0.65 mL/min as follows: 0–0.1 min: 3% A; 0.1–3.0 min: 3.0–45.0% A (linear gradient); 3.0–3.1 min: 45.0–90.0% A (linear gradient); 3.1–4.0 min: 90% A; 4.0–4.1 min: 90.0–3.0% A (linear gradient); 4.1–4.2 min: 3.0% A. The ionization mode (negative/positive) of the mass spectrometer that was used for each compound depended on their ionization efficiency in either negative or positive mode and the one where each compound ionized more effectively in the test samples was then used for quantitative measurements.

All measurements were done in triplicate and quantitation for each compound was done from extracted ion chromatograms (EICs) from full scan MS analysis with a specific *m/z*-range for each compound. Integrated EIC areas were converted to concentrations before partition coefficient calculations with a calibration series done with a dilution series of each compound. Both the calibration series samples and the actual K_{ow} samples mass responses (integrated EIC areas) were normalized with an external standards mass response so that the possible variation in the mass spectrometers performance during the measurements and on different days could be taken into account.

A1.3.2 Partition values of final Martini models

Sugar	Code	S-bead, no VS	N-bead, no VS	TC4	TC4 15%
D-glucose	GLC	-17.68	-15.65	-17.51	-17.06
D-mannose	MAN	-17.68	-15.65	-17.52	-17.06
D-galactose	GAL	-17.68	-15.65	-18.01	-17.16
N-acetylgalactosamine	GalNAc	-17.61	-	-17.86	-17.21
N-acetylglucosamine	GlcNAc	-17.61	-	-17.86	-17.21
N-acetylneuraminic acid	NMC	-25.9	-	-24.89	-23.61
D-fucose	FUC	-25.9	-10.24	-12.19	-12.04
L-rhamnose	RHA	-13.43	-10.24	-12.53	-11.8
D-xylose	XYL	-12.2	-11.8	-14.4	-13.75
Myo-inositol	INO	-19.71	-17.65	-21.73	-21.03
Trehalose	TREH	-26.25	-	-24.78	-23.95
Sucrose	SUCR	-28.11	-	-24.30	-23.57

Table A1.5: Final partition values of the different models tested, given in kJ mol^{-1}. Standard deviation of each Martini model was within ± 0.25 kJ mol^{-1}. Note that for NMC, the free energy was calculated for the protonated form (modelled by SP2 bead), where the final deprotonated model is modelled by a SQ5n bead.

A1.3.3 Experimental octanol/water partition values

Sugar	Code	Log P	SD	$\Delta G_{OCO \rightarrow W}$	SD
D-glucose	GLC	-3.12	0.1	-17.81	0.57
D-mannose	MAN	-2.61	0.03	-14.9	0.17
D-galactose	GAL	-3.07	0.03	-17.52	0.17
N-acetylgalactosamine	GYN	-3.06	0.02	-17.47	0.11
N-acetylglucosamine	GYN	-3.03	0.06	-17.29	0.34
N-acetylneuraminic acid	NMC*	-4.4	0.08	-25.11	0.46
D-glucosamine	GCA	-4.03	0.09	-23	0.51
D-glucuronic acid	GLA*	-3.26	0.02	-18.61	0.11
D-fucose	FUC	-2.26	0.06	-12.9	0.34
L-rhamnose	RHA	-2.26	0.04	-12.9	0.23
D-xylose	XYL	-2.43	0.02	-13.87	0.11
Myo-inositol	INO	-3.49	0.04	-19.92	0.23

Table A1.6: The experimentally determined partition values for the carbohydrates analyzed in this study. Log P values were translated to $\Delta G_{OCO \rightarrow W}$, as outlined in *Section 7.4.1*. The error is given as the standard deviation (n=3), for both log P and $\Delta G_{OCO \rightarrow W}$ (kJ mol^{-1}). Note that for NMC and GLA the log P is given for the protonated solute while at physical pH the majority of species are deprotonated.

A1.4 Distributions disaccharides

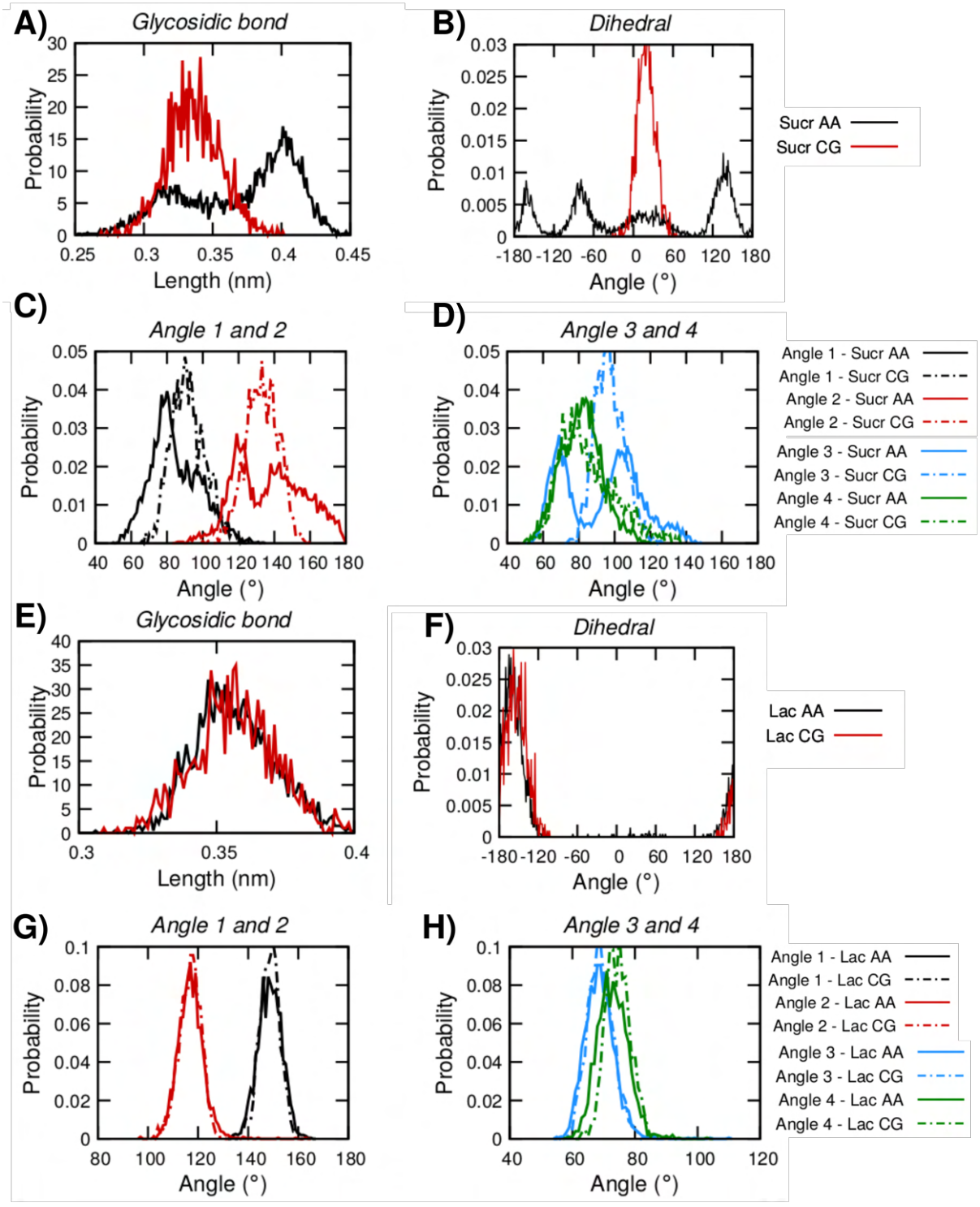

Figure A1.1: The distributions of the AA simulations and final CG models for the glycosidic bond, dihedral and four angle potentials are shown for the disaccharides sucrose and lactose.