

1 Extending the Martini coarse-grained forcefield to 2 N-glycans

3
4 Aishwary T. Shivgan^{1,2}, Jan K. Marzinek², Roland G. Huber², Alexander Krah²,
5 Richard H. Henchman^{3,4}, Paul Matsudaira¹, Chandra S. Verma^{1,2,5,*}, Peter J.
6 Bond^{1,2,*}
7

- 8 1. National University of Singapore, Department of Biological Sciences, 14
9 Science Drive 4, Singapore 117543
- 10 2. Bioinformatics Institute (A*STAR), 30 Biopolis Street, #07-01 Matrix,
11 Singapore 138671
- 12 3. Manchester Institute of Biotechnology, The University of Manchester, 131
13 Princess Street, Manchester, M1 7DN, United Kingdom
- 14 4. Department of Chemistry, The University of Manchester, Oxford Road,
15 Manchester, M13 9PL, United Kingdom
- 16 5. School of Biological sciences, Nanyang Technological University, 50
17 Nanyang Drive, Singapore 637551

18
19 * Corresponding authors:

20
21 Dr Peter J. Bond
22 Bioinformatics Institute (A*STAR)
23 30 Biopolis Str.
24 #07-01 Matrix
25 Singapore 138671
26 Email: peterjb@bii.a-star.edu.sg
27

28
29 Dr Chandra Verma
30 Bioinformatics Institute (A*STAR)
31 30 Biopolis Str.
32 #07-01 Matrix
33 Singapore 138671
34 Email: chandra@bii.a-star.edu.sg
35

36
37
38 Keywords: molecular dynamics (MD) simulations, glycosylation, sugars,
39 carbohydrate, lectins, aggregation, umbrella sampling (US), potential of mean force
40 (PMF), coarse graining (CG)
41
42
43

44 **Abstract**

45 Glycans play a vital role in a large number of cellular processes. Their complex and
46 flexible nature hampers structure-function studies using experimental techniques.
47 Molecular dynamics (MD) simulations can help in understanding dynamic aspects of
48 glycans if the forcefield (FF) parameters used can reproduce key experimentally
49 observed properties. Here, we present optimized coarse-grained (CG) Martini FF
50 parameters for N-glycans, calibrated against experimentally derived binding affinities
51 for lectins. The CG bonded parameters were obtained from atomistic (ATM)
52 simulations for different glycan topologies including high mannose and complex
53 glycans with various branching patterns. In the CG model, additional elastic networks
54 are shown to improve maintenance of the overall conformational distribution. Solvation
55 free energies and octanol-water partition coefficients were also calculated for various
56 n-glycan disaccharide combinations. When using standard Martini non-bonded
57 parameters, we observed that glycans spontaneously aggregated in the solution and
58 required down-scaling of their interactions for reproduction of ATM model radial
59 distribution functions. We also optimised the non-bonded interactions for glycans
60 interacting with seven lectin candidates and show that scaling down the glycan-protein
61 interactions can reproduce free energies obtained from experimental studies. These
62 parameters should be of use in studying the role of glycans in various glycoproteins,
63 carbohydrate binding proteins (CBPs) as well as their complexes, while benefiting from
64 the efficiency of CG sampling.

65

66 **Introduction**

67 Glycosylation is a key post-translational modification involved in a wide range of
68 cellular processes including host-pathogen interactions^{1,2}, cell trafficking³,
69 fertilization⁴, immune system function⁵, energy storage⁶⁻⁸, and are associated with
70 disease states such as congenital disorders including cellular transport defects⁹,
71 muscular dystrophies¹⁰⁻¹² as well as cancer progression¹³. Glycans play a major role
72 in folding and stability of glycoproteins¹⁴. They are one of the key parameters to be
73 considered while developing therapeutic antibodies¹⁵⁻¹⁷. Glycans bind to carbohydrate
74 recognition domains (CRDs) of lectins¹⁸ with low affinity¹⁹, giving cells a versatile
75 system for carbohydrate-protein recognition. They are made up of monosaccharides
76 which can form a variety of anomeric ring linkages resulting in different structures and
77 associated specificities for diverse receptors. These structures range from e.g.
78 cellulose, which is a linear polymer, to cyclodextrins²⁰, a cyclic polymer. Branching of
79 glycans gives them an overall tertiary structure and in turn contributes to the
80 quaternary structures of glycoproteins²¹.

81

82 Glycans are classified according to their protein attachment sites. N-linked glycans are
83 covalently bound to asparagine (N) at the NxS/T motif, where x can be any amino acid
84 apart from proline (P). The other major type of glycans are called O-linked glycans due
85 to their attachment to the hydroxyl oxygen atom of serine (S) or threonine (T)
86 residues^{21,22}. Depending upon the sugars (mannose (Man), n-Acetylglucosamine
87 (GlcNAc) and galactose (Gal)) that extend the common core sequence, Man- α (1,6)-
88 (Man- α (1,3)-Man- β (1,4)-GlcNAc- β (1,4)-GlcNAc- β 1-N-X-S/T, N-glycans are classified
89 into three classes. In the first class, the high mannose (oligomannose) type, the core
90 is extended only by mannose sugars. The second is the complex class, in which

91 branches are extended by N-acetylglucosaminyltransferases (GlcNAcTs). The third
92 class includes the hybrid glycans, in which the Man- α (1,6) arm is attached only to
93 mannose sugars while either one or two complex branches are attached to the Man-
94 α (1,3) arm^{21,22}. Even though there are only three classes of these glycans, the number
95 of glycans found in each class is numerous, hampering structure-function studies. The
96 complexity of the multistep glycosylation pathways very often results in multiple
97 glycoforms for each glycoprotein²³. Also, the inherently flexible nature of glycosidic
98 linkages typically makes it difficult to define their precise structure by either X-ray
99 crystallography or NMR spectroscopy beyond a few monosaccharide units²⁴. The
100 requirement of highly purified samples further complicates NMR studies²³. Very often
101 glycoproteins are deglycosylated in order to reduce the microheterogeneity and
102 associated surface entropy in an attempt to obtain higher resolution crystal
103 structures²⁵. Mass spectrometry can provide structural data for small glycans but is
104 harder for larger complex glycans due to difficulties in determining the glycosidic
105 linkages^{26,27}. Hence, even though glycans are biologically very important, rapid
106 experimental characterization of their structure and delineation of their functional roles
107 remains a major challenge.

108

109 The gaps in understanding the role of glycans at the molecular level and their potential
110 impact on biological processes can potentially be filled by computational modelling,
111 and in particular, by the use of molecular dynamics (MD) simulations. The precise
112 dynamic, biophysical, and thermodynamic *in silico* properties of polysaccharides that
113 can mimic experimental observations depend upon their representation and
114 parameters defined within the FF. Several carbohydrate-specific FFs have been
115 developed in recent years²⁸⁻³¹, the choice of which depends upon the application and

116 simulation conditions desired. The MM3³² FF is useful for reproducing gas phase
117 potential energy curves while the SPASIBA³³ FF is designed to reproduce infrared and
118 Raman spectroscopy data. Other FFs such as AMBER³⁴, CHARMM³⁵, GROMOS³⁶,
119 and GLYCAM²⁸ are good choices for simulating solvated systems but might not be
120 able to reproduce crystal-phase infrared data (see e.g.^{33,37}). Excellent web based tools
121 such as CHARMM-GUI^{38,39} and GLYCAM-Web⁴⁰ make it easier to generate input data
122 for glycan simulations and have been employed in studies that seek to understand, for
123 example, their interactions with other biomolecules as well as their dynamics in
124 different environments^{41,42}. Many of these studies were carried out using ATM
125 representations, which can limit the accessible time scales that may be reached.
126 Biologically relevant complexes containing glycans, such as antibodies, carbohydrate
127 recognition domains like DC-SIGN, or mannose receptors, can encompass millions of
128 atoms, thus making these calculations very expensive and limiting time scales to the
129 sub-microsecond regime, i.e. not equivalent to those sampled in key biological
130 processes or biophysical experiments that reach microseconds to milliseconds or
131 beyond⁴³. Alternatively, a CG representation, in which groups of atoms are
132 represented as larger beads, can be helpful in overcoming the associated limitations,
133 by reducing the number of degrees of freedom and simplifying the energy landscape.
134
135 The Martini FF⁴⁴ is one of the most widely used CG models for biomolecular systems.
136 Martini was originally developed for lipids, and was later extended to proteins⁴⁵,
137 carbohydrates⁴⁶, and nucleic acids⁴⁷. In the Martini representation, approximately four
138 heavy atoms are grouped into a single bead. This represents a relatively lighter
139 coarse-graining approach which allows maintenance of the key structural details of
140 biomolecules. In Martini, non-bonded parameters for different particles have been

141 calibrated against partitioning free energies of small compounds in polar and apolar
142 solvents⁴⁴. The bonded parameters are typically derived empirically by comparing the
143 distributions with ATM simulations. An increasing number of studies have shown that
144 there is an imbalance in the non-bonded interactions, making Martini (v2.2) too “sticky”
145 which has necessitated fine-tuning of the parameters^{48–50}. Recently, an open beta
146 version of Martini v3.0b was released for phospholipids and proteins⁵¹. This version of
147 Martini adds more bead types with various interaction types that aims to solve the
148 shortcomings in the Martini v2.2 FF^{48–50}.

149

150 The Martini FF has been extended to carbohydrates⁴⁶ and includes the parameters for
151 monosaccharides such as glucose and fructose and disaccharides like sucrose,
152 trehalose, maltose, and cellobiose, whose particles have been calibrated to reproduce
153 water-octanol solvation and partitioning energies. The application of these parameters
154 to oligosaccharides such as amylose and curdlan reproduced their key structural
155 properties⁴⁶. Nevertheless, the FF still lacks bonded parameters for different types of
156 glycosidic linkages as well as branching patterns specific to N-glycans such as
157 trimannose (Man- α (1,6)-[Man- α (1,3)-]Man) and bisected N-glycans¹⁴. In addition,
158 parameters are not presently available for N-Acetylglucosamine (GlcNAc), Fucose
159 (Fuc), and Sialic acid (Neu5Ac), which are very common building blocks in many of
160 the N-glycans found in glycoproteins. Hence, there is a gap in the availability of
161 parameters covering the variety of linkages and topologies needed to model
162 biologically relevant glycans, as well as in reproducing experimentally observed
163 glycan-protein binding affinities and aggregation properties.

164

165 In this work, we have extended the Martini FF to N-glycans. As there are many
166 possible glycans, we have restricted our parameterization to the most commonly found
167 N-glycans at present. The bonded parameters for glycans with different glycosidic
168 linkages and branching patterns were obtained by comparing them to ATM simulation
169 data. Elastic networks were shown to be useful in maintaining the conformations of
170 longer glycans. We also observed that the CG glycans tend to aggregate in solution,
171 as in previous studies⁴⁸. Solvation and partitioning coefficients were calculated and
172 compared against prediction methods such as ClogP and KOWWIN⁵². Binding free
173 energies of glycans to lectins, obtained from umbrella sampling (US) calculations were
174 overestimated for all the glycans, confirming a requirement for the fine tuning of non-
175 bonded parameters. This was achieved by scaling non-bonded interactions and
176 comparing the data to binding free energies, radial distribution functions, as well as
177 second virial coefficients (B_{22})⁵³. We found that relatively modest scaling helped to
178 reproduce solution behaviour of glycans only systems and most of the experimental
179 binding affinities in the case of seven candidate lectins.

180

181 **Methods**

182 **All atom simulations**

183 The GROMOS54a7⁵⁴ united atom (UA) FF was used for various α and β glycosidic
184 linkages with different monosaccharides including D-glucose (Glc), D-mannose (Man),
185 and D-galactose (Gal). D-Glucose parameters were used for the glucose unit of the n-
186 acetyl-D-glucosamine (GlcNAc). The bonded parameters for GlcNAc- β 1-asparagine
187 (N-) connections were derived from the extended GROMOS53a6_{GLYC}³¹ FF for
188 glycoproteins. D-fucose (Fuc) and D-sialic acid (Neu5Ac) were manually prepared
189 based on the corresponding galactose and mannose structures, respectively. Different

190 types of ATM N-Glycan structures including disaccharides, trisaccharides (Figure 1)
191 and full length glycans (Figure 2) were constructed using the GLYCAM carbohydrate
192 builder⁴⁰. Each structure was placed in a cubic box such that any atom was at least 1
193 nm away from any wall of the simulation box to avoid self-interaction. The molecules
194 were then energy minimized for ≤ 2000 steps using the steepest descent (SD)
195 algorithm with a 0.01 nm step size⁵⁵. The simulation box was solvated with explicit
196 SPC water molecules⁵⁶ and then again energy minimized using SD. Ions were added
197 to neutralize the overall system charge. The systems were equilibrated for 5 ns in the
198 NPT (constant number of atoms, pressure and temperature) ensemble. The
199 production runs were performed for 1000 ns and convergence was assessed by
200 performing block analysis on the bond, angle, and dihedral distributions. The
201 simulations were run using the velocity rescale thermostat with an additional stochastic
202 term⁵⁷ at a temperature of 310 K with a relaxation time of 0.1 ps. The Berendsen
203 barostat⁵⁸ was used to maintain the pressure at 1.0 bar with weak coupling using a
204 relaxation time of 1 ps. All bonds to hydrogen atoms were constrained using the
205 LINCS⁵⁹ algorithm with a relative geometric tolerance of 10^{-4} enabling a time step of 2
206 fs to integrate Newton's equations of motions with the leap-frog algorithm. A short-
207 range cut-off of 1.2 nm was used for electrostatics and van der Waals interactions.
208 The Particle Mesh Ewald (PME)⁶⁰ method was used for long-range electrostatics, with
209 a 1.2 nm real space cutoff. In addition to GROMOS54a7, comparative simulations
210 using the CHARMM36m⁶¹ FF with the TIP3P water model were used to assess the
211 solution behaviour of glycans, in which case similar conditions were applied along with
212 an additional force switch smoothing function from 1.0 to 1.2 nm for van der Waals
213 interactions. Atomic coordinates were saved every 0.1 ns. All AA simulations were run

214 using the GROMACS 5.1.2⁶² package on an in-house Linux cluster as well as on the
215 ASPIRE 1 supercomputer of the National Supercomputing Centre Singapore (NSCC).

216

217 **CG simulation setup**

218 We followed the Martini mapping scheme for simple monosaccharides and
219 unbranched oligosaccharides as suggested for carbohydrates⁴⁶. In this mapping
220 scheme, each sugar is modelled using three beads and the glycosidic bonds are made
221 by connecting the central beads of two monosaccharides, regardless of the type of
222 glycosidic linkage. However, this is harder to implement in the case of glycans with
223 heavy branching where as many as four monosaccharides are attached to one
224 monosaccharide, such as in bisected tetra-antennary complex glycans. Hence, for
225 these heavily branched glycans, we employed a slightly different method for
226 connecting the monosaccharides. Figure 1 shows the typical di/trisaccharide
227 combinations observed in N-glycans. In the case of Fuc- α (1,6)-GlcNAc, the α (1,6)
228 bond was represented by linking the 2nd and 5th beads. For trimannose Man-
229 α (1,6)[Man- α (1,3)]-Man, the α (1,3) bond was represented by linking the 3rd and 4th
230 beads and the same was done for other types of glycosidic linkages (Figure 1). The
231 bonds were implemented between beads containing atoms originating from the
232 glycosidic linkage in their ATM counterparts. All monosaccharides were linked using
233 the above strategy for all the glycosidic bonds including α/β (1/2, 2/3/4/6) connections.
234 The bead types used for most of the monosaccharides were chosen as suggested in
235 the original carbohydrates Martini CG study⁴⁶ and slightly modified depending upon
236 the new mapping scheme (Figure 1). P1, SP2 and P4 polar beads were used for
237 monosaccharides with a three bead model such as Man and Gal. Fuc, GlcNAc and
238 Neu5Ac models were not previously available, so a five bead model was used to

239 represent Neu5Ac using Qa, SP1, P4 and P5 beads. Fuc was parameterized using
240 SP1, P2 and P4 beads. A four bead model was used for GlcNAc in which P1, SP2, P3
241 beads were used to model the core sugar while SP1 was used to model the acetyl
242 group. The beads for the Fuc, GlcNAc and Neu5Ac were chosen based on chemical
243 intuition and by analogy with previously parametrised carbohydrate-like molecules^{46,63},
244 and further validated by calculating their partitioning data (see below). The mapping
245 schemes are shown in Figure 1, while bead type selections for each of them are given
246 in supplementary Table S2.

247

248 **CG simulation parameters**

249 The Martini FF^{44,46,64} was used for all the CG simulations performed in this study. ATM
250 glycans were mapped according to the representation in Figure 1. Each di/tri-
251 saccharide system was prepared similarly to the ATM system. Systems were solvated
252 with Martini water beads and 10% antifreeze particles. Ions were added to neutralize
253 the overall system charge before energy minimization using SD. A time step of 10 to
254 20 fs was used to integrate the equations of motion using the leap frog algorithm. A
255 constant temperature of 310 K and a constant pressure of 1 bar were maintained, via
256 the velocity rescale thermostat⁶⁵ and the Parrinello-Rahman barostat⁵⁴ with relaxation
257 times of 1 ps and 12 ps, respectively. The non-bonded interactions were truncated at
258 a distance cut-off of 1.1 nm. Electrostatics were handled using a reaction-field⁶⁶ with
259 a cut-off of 1.1 nm. Production runs were carried out for 1 μ s with coordinates saved
260 every 0.2 ns. These parameters were directly taken from the suggested mdp file
261 settings available on the Martini website
262 (http://cgmartini.nl/images/parameters/exampleMDP/martini_v2.x_new-rf.mdp).

263

264 **Parameterization of CG bonded interactions**

265 Three types of bonded harmonic potentials were used. The potential $V_{bond}(r)$ was used
266 to describes the bonds between the CG particles using:

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

267 where r_{bond} and K_{bond} are the equilibrium distance and the force constant, respectively.

268 A harmonic potential for angles was used for three consecutive beads:

$$V_{angle} = \frac{1}{2} K_{angle} (\theta - \theta_0)^2 \quad (2)$$

269 where θ_0 and K_{angle} are the equilibrium angle and force constant, respectively. When
270 the angle was found to be greater than 140° , the restricted bending potential (ReB)
271 was used in order eliminate numerical instabilities associated with dihedral angles:

$$V_{angle_ReB} = \frac{1}{2} \frac{K_{angle} (\cos\theta - \cos\theta_0)^2}{\sin^2\theta} \quad (3)$$

272 Dihedrals were described using the function:

$$V_{dihedral} = K_{dihedral} (1 + \cos(n * \theta - \theta_{dihedral})) \quad (4)$$

273 where $\theta_{dihedral}$ is the equilibrium angle between planes defined by the coordinates of
274 the atoms i, j, k and j, k, l respectively, $K_{dihedral}$ is the force constant, and n is the
275 multiplicity. Most of the dihedrals with a single minimum were fitted using a multiplicity
276 of 1, while those with two minima were fitted using a multiplicity of 2.

277

278 All the ATM trajectories for the systems in Figure 1 were converted to pseudo CG
279 trajectories. Bonds, angles and dihedral distributions were obtained from these
280 pseudo-CG ATM based trajectories and converted into potentials using the Boltzmann
281 inversion method, and fitted with the respective bonded potential functions. CG
282 simulations with these potentials were run and manually fine-tuned in an iterative

283 fashion until they matched as closely as possible to the ATM distributions. Ultimately,
284 these parameters were averaged for the molecules having the same types of bonds
285 and angles within the same disaccharide (eg. GlcNAC- β (1,4)-GlcNAC) or different
286 disaccharides (eg. Man- β (1,4)-GlcNAC and GlcNAC- β (1,4)-GlcNAC) since we
287 observed a maximum of only 5-10% variation between them. The parameters for all
288 di/tri-saccharides (Figure 1) are shown in Table S1 and the ATM vs CG distributions
289 are shown in Figure S1.

290

291 **CG non-bonded interactions**

292 In Martini, the van der Waals component of the non-bonded interactions is described
293 by the Lennard Jones (LJ) 6-12 potential energy function:

$$294 \quad V_{LJ}(r_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

$$V_{LJ}(r_{ij}) = \frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^6}{r_{ij}^6} \quad (5)$$

295 where σ_{ij} is the distance at which the potential crosses zero and epsilon ϵ_{ij} is the well
296 depth. There are a total of 19 different bead types. Beads are divided into four
297 categories according to their ϵ values: polar (P), nonpolar (N), apolar (C) and charged
298 (Q). Each main type of bead is subdivided by its hydrogen bonding properties such as
299 donor (d), acceptor (a), both (da) and none (0). The polarity of the bead ranges from
300 low (=1) to high (=5) with interaction level (ϵ) ranging from 2.0 to 5.6 kJ mol⁻¹ and an
301 interaction distance (σ) of 0.47 nm. Smaller beads are used for ring structures with
302 $\sigma=0.43$ nm and 75% of the normal ϵ value. These values were previously
303 parameterized to reproduce partition free energies of a library of small molecules⁴⁴. In
304 order to optimise the non-bonded interactions for our glycans of interest, we changed

305 the well depth of the LJ potential i.e. modified the value of ϵ using the following
306 equation,

$$\epsilon_{new} = 2 + \lambda(\epsilon_{original} - 2) \quad (6)$$

307 where λ is a scaling factor ranging from 0 to 1, whilst the value of ϵ remains unchanged
308 when $\lambda=1.0$ and becomes 2 kJ/mol when $\lambda=0$, corresponding to the lowest value of ϵ
309 for a bead in the Martini FF. A similar approach has been used in other studies to
310 correct the non-bonded FF parameters^{48,49}. Only the solute-solute (glycan-glycan or
311 glycan-protein) interactions were scaled down while solute-solvent and solvent-
312 solvent interactions were kept at their default level. This was done by adding special
313 glycan bead types, and rescaling (by λ) the C^6 and C^{12} terms (Equation 5) for their
314 interaction with other solute particles accordingly. The down-scaling of glycan-
315 glycan/glycan-protein interactions effectively makes the glycan-water interactions
316 more favourable, consistent with experiments analysing the second virial coefficient
317 (B_{22})⁴⁸.

318

319 **Partitioning free energies**

320 The choice of bead types used in Figure 1 was validated by calculating partitioning
321 propensities. Solvation free energies of various disaccharides in the water (ΔG_W) and
322 octanol (ΔG_O) phases were used for calculating partition coefficients ($\log P_{OW}$). The
323 free energies of solvation ΔG_W and ΔG_O were calculated using thermodynamic
324 integration⁶⁷:

$$\Delta G_{A \rightarrow B} = G_B - G_A = \int_{\lambda=0}^{\lambda=1} d\lambda \left\langle \frac{\delta U(\lambda)}{\delta \lambda} \right\rangle_{\lambda} \quad (7)$$

325 where the potential energy change (δU) for going from state A to B is calculated as a
326 function of coupling parameter (λ) which goes from 0 (full interactions between beads

327 and solvent) to 1 (no interaction). Non-bonded interactions were scaled linearly. A soft
328 core potential was used to circumvent potential singularities which occur during
329 annihilation or creation of atoms⁶⁸. A total of 55 intermediate λ values were used,
330 including additional ones in the high curvature regions. Each λ point was subjected to
331 40 ns of simulation time with the final 20-30 ns used for analysis, depending upon
332 convergence. The free energy differences were estimated using the Bennett
333 acceptance ratio method⁶⁹ (BAR) implemented within GROMACS. The partitioning
334 free energy $\Delta\Delta G_{OW}$ is then the difference between ΔG_W and ΔG_O , from which the
335 partition coefficient ($\log P_{OW}$) may be calculated using:

$$\Delta\Delta G_{OW} = -2.3RT * \log P_{OW} \quad (8)$$

336 The P_{OW} values from simulations were compared to partition coefficient prediction
337 methods such as ClogP and KOWWIN, which have been benchmarked against a wide
338 variety of compounds⁵². The water-only simulations were composed of 1 disaccharide
339 and 1000 water molecules, while the hydrated octanol simulations were composed of
340 1 disaccharide, 43 water molecules, and 519 octanol molecules, representing a 0.255
341 water/octanol molar fraction⁷⁰. The vacuum-only simulations were composed of a
342 single disaccharide in the simulation box.

343

344 **Umbrella sampling simulations to estimate binding affinities**

345 Potential of mean force (PMF) profiles for the association of two solute molecules
346 (glycan-glycan or lectin-glycan pairs) were calculated using umbrella sampling (US)⁷¹.
347 Thus, multiple MD simulation windows were run along a pre-defined reaction
348 coordinate – the separation distance between solute groups, along the z-axis of the
349 simulation box – with an additional biasing harmonic potential. For a lectin-glycan pair,
350 the groups were the center of mass of the glycan and the center of mass of residues

351 defining the binding site in the protein. First, the two solutes were pulled away from
352 each other at a rate of 10 nm ns^{-1} , in order to generate the initial coordinates for the
353 US windows. In the case of glycan-lectin PMFs, the glycan was pulled away from the
354 lectin binding site. Both pulling simulations and US windows employed a harmonic
355 potential between the centres of mass of the two groups of interest along the z-axis
356 using a force constant of $1000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$. Each complete pulling simulation
357 corresponded to a 3 to 5 nm distance and resulted in 30 to 40 US windows with a 0.12
358 nm spacing. Additional windows were added in the regions of the minima to achieve
359 greater overlap between windows, where necessary. Each window was subjected to
360 production runs of 500 ns, leading to $40 \times 500 \text{ ns}$ ($20 \mu\text{s}$) of sampling per system per
361 replica. For each system, at least two simulation replicas were performed. Block
362 analysis was performed in order to assess the convergence. This was done by splitting
363 the PMF trajectories into 100 ns windows. The PMFs were constructed using
364 GROMACS's inbuilt Weighted Histogram Analysis Method⁷² (WHAM) algorithm. The
365 converged part of the trajectory was used for constructing the final PMF. 200 cycles of
366 bootstrapping using the *b-hist* (Bayesian Bootstrapping) method with a tolerance of
367 10^{-6} was used to estimate the standard deviation across all replicas.

368

369 **Calculation of the binding free energy with standard state correction (ΔG^0)**

370 The binding free energy (ΔG^0) for each glycan to its lectin was calculated from the one
371 dimensional PMF $W(z)$ by defining ΔW as zero at the minimum of the PMF curve
372 minus the exponential average of the PMF over the unbound region, while a correction
373 term (ΔG_V) was added to account for the standard state volume $V^0 = 1661 \text{ \AA}^3$ based
374 on the sampled unbound volume (V_u)⁷³:

$$\Delta G_{sim}^0 = \Delta W + \Delta G_V \quad (9)$$

$$\Delta W = RT \ln \left[\frac{\int_{bound} \exp\left(-\frac{W(z)}{RT}\right) dz}{\int_{unbound} dz} \right] \quad (10)$$

$$\Delta G_V = -RT \ln \left(\frac{V_u}{V^0} \right) = -RT \ln \left(\frac{l_b A_u}{V^0} \right) \quad (11)$$

375 where A_u is the area sampled by the ligand in the unbound region. The protein
 376 backbone was restrained during the PMF calculation to prevent its rotation, which also
 377 prevented the ligand forming unproductive interactions with regions distant from the
 378 binding site in umbrella windows at increasing z-values. The unbound area was
 379 approximated as the cross-sectional area of the simulation box (i.e. the xy plane),
 380 following verification of complete sampling in x and y by each ligand across all
 381 unbound 500 ns umbrella windows (Figure S2). l_b is the bound length calculated by:

$$l_b = \int_{bound} \exp\left(-\frac{W(z)}{RT}\right) dz \quad (12)$$

382 No further rotational entropy corrections were necessary, as the ligands were allowed
 383 full rotational sampling in the unbound region. The standard state free energies of
 384 binding were compared with the corresponding experimental values.

385

386 **Second virial coefficients (B_{22})**

387 Osmotic data provides information about the nature of interactions between two solute
 388 molecules. This informs on how much the simulated association deviates from
 389 experimental measurements with molar concentration (c)⁵³ and has previously been
 390 used to correct carbohydrate⁷⁴ and protein^{49,75} FFs. In particular, the second virial
 391 coefficient (B_{22}) comes from the virial expansion of pressure of many particle systems
 392 given by⁵⁷:

$$\Pi(T, c) = RT(c + B_{22}C^2 + B_{23}C^3 + \dots) \quad (13)$$

393 where B_{ij} are coefficients of virial expansion, B_{22} is the second virial coefficient, T is
394 the temperature, and R is the gas constant. $B_{22} > 0$ indicates repulsive interactions
395 between the two solutes while $B_{22} < 0$ indicates attractive interactions. The B_{22} value
396 can be experimentally determined by self-interaction chromatography⁷⁶ or diffraction
397 studies⁷⁷. McMillan and Mayer derived a method to calculate B_{22} values using MD
398 simulations⁵³ providing a powerful tool that can be used to optimise FFs. In this
399 method, the PMF $W(z)$ can be used to obtain the B_{22} value using the following
400 expression:

$$B_{22} = -2\pi N_A \int_0^{\infty} \left[\exp\left(-\frac{W(z)}{RT}\right) - 1 \right] z^2 dz \quad (14)$$

401 where N_A is the Avogadro number, R is the gas constant, T is the temperature of the
402 system, and z is the distance between the solute molecules.

403

404 **N-glycans from di/trisaccharides**

405 CG glycans such as high mannose, bi/tri/tetra-antennary and bisected complex
406 glycans were constructed from component di/trisaccharide units (Figure 1).
407 Parameters for the connecting angle between a sugar and its second neighbour were
408 missing. To obtain these parameters, corresponding ATM glycans were constructed
409 using the GLYCAM glycan builder⁴⁰. Again, a similar methodology as used for
410 di/trisaccharides was used to generate pseudo CG trajectories from 1 μ s ATM
411 simulations. As shown in Figure 2C-E, the glycosidic bonds between N-
412 acetylglucosamine and mannose in a tetra-antennary complex glycan can be
413 $\beta(1,2/4/6)$. These branch patterns were parameterized separately. The angles
414 obtained are summarized in Table 1. When constructing full length glycans these
415 parameters are added to the disaccharide parameters depending upon the topology
416 of the glycan constructed. The different types of glycans are named according to their

417 Oxford notation⁷⁸. This is based on building up N-glycan structures and it can be used
418 to denote very complex glycans: all N-glycans have two core GlcNAcs; a given number
419 of mannose sugars on the core GlcNAcs are denoted by Mx (e.g. M3, M5, M9); the
420 number of antennae on the trimannosyl core are given by Ax (e.g. A2, A3, A4); Gx is
421 the number of linked galactose units on antennae (eg. A2G1, A3G3); Sx is used to
422 denote the number sialic acids linked to galactose (e.g. A2G2S2, A3G3S1); and an F
423 start denotes the presence of fucose (e.g. FA2, FA3G2S1).

424

425 **Aggregation studies**

426 It was previously reported^{48,63} that sugars in the Martini FF have a greater tendency to
427 aggregate than observed experimentally. To investigate this further, systems
428 containing 35 to 40 molecules of glycans were set up, with a $\sim 50 \text{ g L}^{-1}$ concentration.
429 Glycans should be readily soluble at concentrations of 50 g L^{-1} , as dextran, a branched
430 polymer of glucose is soluble even at a concentration of 400 g L^{-1} ⁷⁹. The simulations
431 were performed using Martini v2.2, Martini 3.0b, GROMOS54a7 as well as
432 CHARMM36m. To estimate the aggregation propensity of these glycans, RDFs were
433 calculated. PMFs were also calculated to quantify the aggregation strength using
434 Martini v2.2, GROMOS54a7 and CHARMM36m FFs. Triplicate simulations were used
435 to construct the final PMFs at each scale factor. Scaling factors (λ) of 1.0, 0.9, 0.7, 0.5
436 and 0.3 were used to estimate second virial coefficients (B_{22}) according to equation
437 14.

438

439 **Lectin binding studies**

440 To supplement the partitioning and virial data, we pursued a complementary approach
441 to validate non-bonded interactions by calculating binding free energies of different

442 types of glycans with the lectins, which are selective for specific sugar patterns. Lectins
443 are a class of proteins which selectively bind to mono or oligosaccharide molecules
444 with specific glycosidic linkages¹⁸. This makes them good candidates for testing and
445 validating the bonded as well as non-bonded parameters of our glycans. A total of
446 seven candidate lectins, including cyanovirin-N (CVN), concanavalin-A (CONA),
447 pterocarpus anolensis (PAL), ricinus communis agglutinin (RCA), wheat germ
448 agglutinin (WGA), Maackia amurensis (MAA) and urtica dioicia agglutinin (UDA) were
449 chosen; each has available crystal structures, along with either Isothermal Calorimetry
450 (ITC) or Surface Plasmon Resonance (SPR) data for binding to various types of N-
451 glycans. For every lectin, the ATM structure was converted to CG resolution using the
452 *martinize.py* script from the Martini website. An elastic network with upper and lower
453 cut-off values of 0.5 and 0.9 nm, respectively, and a force constant of 500 kJ mol⁻¹ nm⁻²
454 was implemented in addition to secondary structural bond/dihedral terms to maintain
455 the higher order structures of all lectins. While setting up the lectin-glycan systems,
456 the sugars which were resolved in the crystal structures were aligned with the CG
457 glycan models, while non-interacting branches of the glycan were modelled pointing
458 outwards, into solvent. The lectins used in this study, their PDB IDs and their binding
459 affinities for various glycans obtained either from ITC or SPR experiments are
460 summarized in Table 3. For reasons that will become apparent below, we calculated
461 PMFs with scaling factors of 1, 0.95, or 0.9 for each of the 13 glycan-lectin pairs (Table
462 3, Figure 5).

463

464

465 **Results**

466 **Elastic network in extended N-glycans**

467 As previously reported for studies of glycolipids and some oligosaccharides^{48,63}, when
468 applying dihedral potentials in simulations of our extended CG glycans, numerical
469 instabilities limited the maximum time step to 5 fs. This was alleviated by using ReB
470 angle potentials, enabling a timestep of 10 fs. To retain the overall conformation in
471 accordance with the ATM models, elastic connections were also implemented
472 between the central bead of the first mannose of the trimannosyl (M3) core and the
473 last monosaccharide of each branch of the glycans shown in Figure 2. In addition,
474 harmonic angle potentials (Table 1) were added between the branch ends and the M3
475 core. Looking at the branch angle distributions of high mannose (M9) and complex
476 type (FA2G2S2) glycans (Figure 3), it was observed that dihedrals alone could not
477 reproduce the AA distributions. The distributions in CG simulations with only dihedrals
478 were wider compared to their ATM counterparts. The distributions with either dihedrals
479 plus elastic network or elastic network alone could, however, reproduce the ATM data
480 closely, confirming the requirement of an elastic network to maintain the overall
481 conformations of the glycans. This allowed us to run our simulations of glycans having
482 an elastic network with dihedrals switched off using a stable time step of 15 fs. All the
483 elastic network parameters are summarised in Table 1.

484

485 **Partitioning behaviour**

486 The Martini forcefield building block parameters were derived in part according to their
487 partitioning behaviour. In previous carbohydrate development efforts, experimental
488 partitioning energies were accurately reproduced for various small sugars including
489 glucose, fructose, sucrose, maltose, cellobiose, kojibiose, sophorose, nigerose,

490 laminaraibose and trehalose⁴⁶. Although the bead type selection is very similar to
491 these sugars, new parameters were derived for variants including GlcNAc, Neu5Ac
492 and Fuc. Thus, we calculated solvation free energies of various n-glycan
493 disaccharides in water (ΔG_W) and octanol (ΔG_o) phases to obtain partition coefficients
494 ($\log P_{OW}$) for comparison with corresponding values from empirical fragment-based
495 models (Table 2). The $\log P_{OW}$ values were negative for all sugars, consistent with their
496 preference towards the water phase. Overall, the simulated values are in reasonable
497 agreement with those obtained from the empirical models, with slightly closer
498 accordance with the ClogP data compared to KOWWIN, consistent with previous
499 studies⁵².

500

501 **Aggregation of glycans**

502 Simulations of both high mannose and complex glycans at $\sim 50 \text{ g L}^{-1}$ concentration, a
503 concentration at which all of them should be soluble, led to aggregation within a few
504 hundred nanoseconds at both CG and AA resolution, except when using the
505 CHARMM36m FF (Figure 4). These aggregates remained irreversibly associated
506 even after extending the simulations to $2 \mu\text{s}$. When comparing against ATM FFs,
507 GROMOS54a7 showed very similar behaviour to Martini v2.2 and 3.0b in terms of
508 both RDFs and PMFs. However, CHARMM36m did not form clusters in any of the
509 simulations, and also resulted in a very shallow PMF well depth (Figure 4B-C, E-F).
510 Martini 3.0b is still in the early stages of its development and the new mapping scheme
511 is not yet available for simple carbohydrates. So, with the added bead types and
512 interaction levels, the final sugar mapping may still improve the results in future.
513 Similar to findings of a previous study⁴⁸, non-bonded interactions in Martini v2.2 were
514 found to be too attractive for the glycans. This is as reported for other Martini-

515 parameterized molecules like glycolipids and proteins such as lysozyme^{48,49,63}. Thus,
516 we attempted to optimize the non-bonded parameters of the glycans, firstly by tuning
517 the second virial coefficient (B_{22})⁵³ and secondly by comparing the glycan binding free
518 energies calculated from simulations with either ITC or SPR experiments. Scaling
519 down the interactions alleviated the aggregation propensity, but required a scaling
520 factor of 0.7 or below, to reach B_{22} coefficient values of $\geq 0 \text{ L mol}^{-1}$, in both cases
521 (Figure 4D,G). The experimental value of the B_{22} for the complex glycan (A2G2S2) is
522 around 40 L mol^{-1} ⁴⁸, which could not be achieved even after reducing the scaling
523 factor to as low as 0.3. Nevertheless, aggregates were not formed during $1 \mu\text{s}$
524 simulations of high mannose and complex glycans when a scaling factor of 0.85 was
525 used (Figure 4B,E). PMFs calculated with these scaling factors implemented for both
526 high mannose and complex glycan were flat for a scaling factor 0.9 or below, indicating
527 dominating water-glycan interactions as observed in the experimental conditions
528 (Figure S5). Importantly, we also observed that aggregation was dependent upon the
529 size of the glycans (Figure S6). Monosaccharides did not require any scaling while a
530 scaling of 0.95-0.9 was required for disaccharides. For sugars bigger than
531 disaccharides, a maximum scaling factor of 0.85 was required to reproduce ATM RDFs
532 including high mannose and complex glycans.

533

534 **Binding of glycans with different lectins**

535 Based on the results of PMF calculations, described in detail below, every lectin-glycan
536 pair overpredicted binding free energies from unscaled simulations. Scaling by 0.95
537 was sufficient for most of the pairs, except for the MAA and PAL lectins, where 0.9
538 scaling was required. The calculated binding free energies across all systems are

539 summarised in the Table 3, with detailed information about the binding pocket in given
540 in Table S3.

541

542 **1. Mannose binding lectins**

543 **1.1 Concanavalin A (CONA)**

544 The mannose binding specificity of CONA is dependent on the trimannoside core
545 found in most N-glycans. The crystal structure⁸⁰ (Figure S7A) shows tetrameric CONA
546 interacting with four core trimannoside Man- α (1,6)-[Man- α (1,3)]Man or M3. The
547 interaction site includes residues Y12, N14, T15, D16, G98, L99, Y100, A207, D208,
548 G227 and R228. ITC experiments with GlcNAc- β (1,2)-Man- α (1,6)[GlcNAc- β (1,2)-
549 Man- α (1,3)]Man-OH glycan, which will be referenced as A2-nocore hereafter, yielded
550 a ΔG of -8.4 ± 0.1 kcal mol⁻¹⁸¹. In our studies, the M3 sugars of the glycan were aligned
551 with the crystal structure. Our simulations predicted a binding free energy of -9.5 ± 0.2
552 kcal mol⁻¹ (Figure 5A) for $\lambda=1.0$, which is an overprediction of ~13%.

553

554 **1.2 Cyanovirin-N (CVN)**

555 CVN is a widely studied 110 kDa lectin because of its role in inactivation of many
556 strains of Human Immunodeficiency virus (HIV)⁸². CVN preferentially binds to high-
557 mannose oligosaccharides⁸³. ITC experiments show that two mannose
558 oligosaccharides M8 and M9 bind to CVN with binding affinities of -8.7 ± 0.1 and -9.2
559 ± 0.3 kcal mol⁻¹ respectively⁸⁴. The crystal structure for CVN with M9 (Figure S7B)
560 reveals a binding interface of three stacked α (1,2) linked mannose sugars interacting
561 with residues L1, G2, K3, T7, E23, N93, D95 and E101 of CVN while the rest of the
562 chain is exposed to solvent⁸⁵. In our studies, both M8 and M9 glycans were used to
563 estimate the binding free energy. M8 and M9 have a terminal α (1,2) linked mannose

564 which is important for the interaction. Similar behaviour to CONA was observed in the
565 case of the CVN lectin. The calculated PMFs predict binding free energies of $-9.8 \pm$
566 0.7 and -10.9 ± 0.4 kcal mol⁻¹ (Figure 5B & 5C) for M8 and M9 respectively, which are
567 overpredicted by ~12% and ~18% respectively.

568

569 **1.3 Pterocarpus Anolensis lectin (PAL)**

570 PAL is a Mannose/Glucose specific lectin and multiple crystal structures showing its
571 interactions with mono, di and trisaccharides are available^{86,87}. Recent studies
572 revealed PAL's interactions with M9 and M5 high mannose glycans⁸⁸ (Figure S7C). It
573 was observed crystallographically that PAL interacts with M5 and M9 in the same
574 unique way in which the Man- α (1,2)-Man- α (1,6)-[Man- α (1,3)-]Man- α (1,-) motif binds
575 to PAL via a combination of van der Waal's contacts and hydrogen bonds⁸⁸. The glycan
576 interacts with residues D36, N83, G106, D136, N138, E221. Both M5 and M9 interacts
577 strongly with PAL, with binding affinities of -5.2 and -5.8 kcal mol⁻¹ respectively, as
578 shown by ITC⁸⁸. Our simulations led to an overestimation of the interactions in the
579 non-scaled systems for both glycans (Figure 5D & 5E). In the case of M5 and M9 the
580 binding free energy was overestimated by ~77% and ~155% respectively. To achieve
581 reasonable agreement with the experimental binding affinities, the interactions needed
582 to be scaled by 0.9.

583

584 **2. Galactose/N-acetylgalactosamine binding lectins**

585 **2.1 Ricinus Communis Agglutinin (RCA₁₂₀)**

586 RCA₁₂₀ is a hemagglutinin and is tetrameric in nature. It has two α and two β subunits
587 that are 29.5 and 37 kD in size, respectively. Out of the two types of subunits, it has
588 been shown that the oligosaccharides interact only with the β subunits of the lectin⁸⁹.

589 RCA₁₂₀ specifically recognises Gal-β(1,4) with very similar affinities for Gal-β(1,4)-
590 GlcNAc, Gal-β(1,4)-Glc and Gal-β(1,4)-Man terminal residues⁹⁰. SPR studies were
591 performed with bi (A2G2), tri (A3G3) and tetra-antennary (A4G4) complex glycans
592 (Table 3) and resulted in binding affinities of -7.7, -7.3, and -7.0 kcal mol⁻¹,
593 respectively⁹¹. Although there are no crystal structures showing direct interactions with
594 any of the A2G2, A3G3 or A4G4 glycans, there is a crystallographic study showing
595 interactions with two terminal galactoses (PDB 1RZO). In the structure (Figure S7D)
596 of RCA₁₂₀, the first galactose interacts with D22, G25, E26, Q35, K40 and N46, while
597 the other galactose interacts with N95 and Y125. Considering the higher number of
598 interactions of the first GlcNAc with polar and charged residues, it was used for the
599 PMF calculations. For all branched glycans, including bi, tri, and tetra-antennary
600 glycans, we found that the FF overestimates the binding free energy without any
601 scaling of interactions (Figure 5F, 5G & 5H). The energies obtained from the simulation
602 with a scaling factor of 0.9 were -6.1 ± 0.7, -6.3 ± 0.7 and -6.3 ± 0.3 kcal mol⁻¹ with
603 20%, 14% and 10% deviation from the experimental values for A2G2, A3G3 and A4G4
604 glycans, respectively.

605

606 **3. Sialic acid/N-acetylglucosamine binding lectins**

607 **3.1 Wheat Germ agglutinin (WGA)**

608 WGA is an Neu5Ac and GlcNAc specific lectin which is antifungal in nature and has
609 three isoforms^{92,93}. The crystal structure reveals a stable dimer with each polypeptide
610 showing four hevein domains responsible for GlcNAc recognition⁹⁴, though not all
611 eight binding sites were observed to be occupied in a single crystal⁹⁵. In this structure
612 (Figure S7G), the first GlcNAc occupied the region defined by residues D86, S105,
613 F109 from monomer 1 and A71, E72 of monomer 2, while the region defined by

614 residues S62, Y64, Y66, E72, Y73 of monomer 1 and S114, E115 of monomer 2 were
615 occupied by the second GlcNAc. Binding studies revealed that WGA has the highest
616 affinity (5.8 kcal mol⁻¹) for (GlcNAc)₅ and decreases to -3.7 kcal mol⁻¹ as the number
617 of sugars reduces from (GlcNAc)₅ to (GlcNAc)₁⁸¹. ITC experiments of single domain
618 recombinant WGA with (GlcNAc)₃ and (GlcNAc)₄ resulted in a 10-fold lower binding
619 constant than the wild type oligomer, emphasizing the importance of the dimer
620 interface in the binding of oligosaccharides⁹⁶. Considering the selectivity of (GlcNAc)₅,
621 it was used for the calculation of PMFs, based on the first dimeric interface from the
622 crystal structure (Figure S7G). In our studies, the binding free energies for (GlcNAc)₅
623 were overpredicted and scaling 0.95 was required to obtain a binding free energy of -
624 5.6 kcal mol⁻¹ (Figure 5I) which represents a 4% deviation from experiment.

625

626 **3.2 Urtica Dioicia agglutinin (UDA)**

627 UDA is a chitin and is a GlcNAc oligomer specific lectin derived from plants^{97,98}. UDA
628 is speculated to be antifungal and insecticidal in nature^{99,100}. Binding experiments
629 suggest that the lectin has two binding sites with a preference for (GlcNAc)₅, with an
630 affinity of -5.9 kcal mol⁻¹ that decreases to -3.9 kcal mol⁻¹ upon a reduction in chain
631 length to (GlcNAc)₂.⁸¹ The crystal structure of UDA isolectin VI is available, revealing
632 its interactions with (GlcNAc)₃¹⁰¹ (Figure S7F). The GlcNAc oligomer interacts with
633 UDA at residues S19, C24 and Y30. In our simulations, the binding free energy with
634 the (GlcNAc)₅ oligosaccharide – modelled based on the (GlcNAc)₃ coordinates for the
635 middle three GlcNAc groups – was overpredicted by ~25% in the absence of scaling.
636 A scaling of 0.95 yielded close agreement with the experimental value of -5.9 kcal mol⁻¹
637 (Figure 5J) with only a 7 % deviation.

638

639 **3.3 Maackia Amurensis (MAA)**

640 Maackia Amurensis has two isolectins, hemagglutinin and leukoagglutinin (MAA),
641 which were identified by their agglutination activity against different blood cell types
642 and their binding properties with either O-linked or N-linked oligosaccharides^{102,103}.
643 Subsequently, it was shown that MAA is specific towards Neu5Ac units, especially
644 towards the NeuAc- α (2,3)-Gal- β (1,4)-GlcNAc/Glc oligosaccharide^{104,105}. SPR binding
645 studies of MAA with different N-glycans such as sialylated tri-antennary (A3G3S2),
646 fully sialylated tri-antennary (A3G3S3) and fully sialylated tetra-antennary (A4G4S4)
647 yielded similar affinities of -4.7, -5.7 and -5.5 kcal mol⁻¹ respectively, suggesting a slight
648 preference for the NeuAc- α (2,3)-Gal- β (1,4)-GlcNAc/Glc β motif¹⁰⁶. The crystal
649 structure of MAA is dimeric with each monomer folding into large β -pleated sheets¹⁰⁷.
650 Each monomer shows interactions with sialyllactose at residues T45, D87, S104,
651 L107, T131, T136 and T221 (Figure S7E). The sialyllactose coordinates were used as
652 the initial coordinates for modelling all three glycans (A3G3S2, A3G3S3 and A4G4S4)
653 with other branches pointing towards the solvent. PMF calculations revealed that sialic
654 acid containing glycans result in overprediction of binding free energies by ~125-200 %
655 (Figure 5K, 5L & 5M). Scaling the interactions by 0.9 was required to obtain reasonable
656 agreement with the experimental values.

657

658 **Discussion**

659 The highly flexible nature of glycans limits detailed structure-dynamics-function
660 studies using experimental techniques, but this gap may be supplemented by MD
661 simulations. The behaviour of molecules in these simulations depends upon the FF
662 used. In this work, we have extended the Martini parameters towards N-glycans. A
663 slightly different mapping scheme was used where bonds were made between beads

664 originating from ATM models compared to the one used by Lopez et al⁴⁶ in which only
665 the central beads were connected to describe the glycosidic linkage. Parameterization
666 of disaccharides using this scheme was convenient for making highly branched
667 patterns of N-glycan like bisected tetra-antennary complex glycans. Although all the
668 N-glycan glycosidic linkages were parameterized using this mapping scheme, one
669 should note that coarse-graining still leads to reduced accuracy. Firstly, it results in a
670 loss of the explicit stereochemical nature and the hydrogen bonding network of sugars,
671 which is important for the specific interactions with carbohydrate binding proteins
672 (CBPs)^{88,92,101,107}. The water network around the sugars is lost, and in turn can affect
673 the local translational and rotational dynamics¹⁰⁸. Although coarse-graining leads to
674 loss of information, the Martini CG approach^{45,109–112} nevertheless implicitly maintains
675 chemical identities by using appropriate polar bead types that have been
676 experimentally validated against water-octanol partitioning free energies⁴⁶.

677

678 Monosaccharides undergo chair-to-boat and chair-to-chair transitions, also referred to
679 as “ring puckering”, making the choice of FF for calibration important. The ATM FF
680 used in this study, GROMOS54a7¹¹³, reproduces these conformations very well. But
681 in the Martini CG model, this effect is neglected. Although the ring can transition
682 between ⁴C₁ and ¹C₄ conformations in the ATM model, the overall shape of the sugar
683 remains the same in the CG model as a result of the grouping of atoms into unified
684 particles, making the sugar effectively linear (Figure 1). Therefore, the CG model
685 should not be affected by the preference of the ATM FF towards a specific
686 conformation showing the average structure of these puckering transitions. Also, the
687 more common dextrorotatory (D) form of sugars was considered for
688 parametrization^{114–116}.

689

690 The CG approach also affects the degree to which one can distinguish between α and
691 β anomers, which were not considered here as the anomeric form of most of the
692 sugars is already fixed for N-glycans. Multiple rotameric states of the hydroxymethyl
693 group and its preference towards $gg(-60^\circ)$ and $gt(60^\circ)$ ^{117,118} conformations was
694 observed in the ATM simulations. We modelled the bimodal distributions using a single
695 harmonic potential by fitting them to the most populated conformation observed. It
696 should be noted that in future, they could be modelled via tabulated potentials¹¹⁹. The
697 glycosidic linkages in di/tri-saccharides were represented by using dihedral potentials
698 which orient the monosaccharides relative to each other. But in the case of N-glycans,
699 dihedrals resulted in numerical instabilities due to geometric tension between the
700 glycosidic bonds, as observed in other studies^{48,63}. The problem could either be solved
701 by using a smaller timestep, which cripples the efficiency of the CG approach, or by
702 not using the dihedral potentials at all. The first approach has been used for
703 glycolipids⁶³ while the latter has been used for some oligosaccharides⁴⁸. A timestep of
704 5 fs or less is manageable but seriously limits the benefits of the CG approach,
705 becoming computationally inefficient for larger biologically relevant systems such as
706 glycoprotein-antibody complexes. This was partially alleviated here by using ReB
707 angle potentials, which improves the associated numerical instabilities. Furthermore,
708 an elastic network proved beneficial in maintaining the overall conformations of the
709 glycans (Figure 3). This ultimately allowed the simulations to be run using a timestep
710 of 15 fs. An alternative approach to the elastic network that could be investigated in
711 future would be to introduce virtual sites, within the trimannose core, the
712 disaccharides, or at other branch sites, in order to maintain key conformations via
713 effective dihedrals¹²⁰.

714

715 A first step towards validating the non-bonded interactions in new Martini parameters
716 involves calculating partition coefficients for the building blocks of the molecules of
717 interest. We observed that the selection of bead types for our di/saccharide
718 combinations correlated well with the partition data from empirical prediction
719 methods⁴⁶. However, constructing full length glycans from these building blocks
720 showed some serious issues in terms of self-aggregation. Comparing the effect of the
721 elastic network on the self-interaction energies of M9 and A2G2 complex type glycans
722 suggests that the elastic network does result in slightly higher binding free energies
723 (Figure S4). It has recently been shown that this is likely because of weak force
724 constants in the elastic network resulting in a short bond length effect which creates
725 “superinteraction” centres¹²¹. This observed effect is small in our model as we have a
726 maximum of just three elastic bonds, compared to 24 in the polyleucine model used in
727 that study. But even without using an elastic network, the predicted self-binding
728 energies for the glycans were observed to be negative, which resulted in the well-
729 documented “sticky” behaviour. Similar observations were made in a related study on
730 few oligosaccharides⁴⁸. Thus, Schmalhorst et al⁴⁸ showed that carbohydrates
731 including glucose (monosaccharide), sucrose (disaccharide), α/β -cyclodextrin (cyclic),
732 and sialylated biantennary glycan (A2G2S2) spuriously aggregate within a few
733 hundred nanoseconds and proposed a 50% scaling down of non-bonded interactions
734 for oligosaccharides. Likewise, for our glycan models including A2G2S2 and high
735 mannose (M9) (Figure 4), carbohydrate molecules aggregated within a few hundred
736 nanoseconds. The calculated B_{22} values were negative (Figure 4) consistent with
737 interactions between sugars being attractive in the Martini representation, compared
738 to positive and hence repulsive experimental B_{22} values⁴⁸. This sticky behaviour is

739 suggested to be in part the result of mixing smaller sized beads with regular (R) beads
740 creating artificial energy barriers¹²¹. The effect is significant with the “tiny” (T type)
741 beads while much lower with “small” beads (S type). So, depending upon the type
742 and length of the glycan, the added artificial barriers will potentially aggravate the
743 sticky nature of Martini FF.

744

745 The overly attractive behaviour observed for oligosaccharides in the Martini FF calls
746 for adjustments in the underlying non-bonded interactions. A simple way of solving this
747 issue is to make the water-glycan interactions stronger or make glycan-glycan
748 interactions weaker. The latter approach is generally preferred^{48,49} so as to keep
749 central properties of the Martini FF constant such as the partitioning behaviour
750 between water and apolar solvents, hence avoiding the need for complete
751 reparameterization of the entire FF. Thus, the solvent-solvent and solute-solvent
752 interactions were not changed. Weakening the glycan-glycan interactions partly solved
753 the issues (Figure 4). Scaling by as much as $\lambda=0.7$ was required to reach B_{22}
754 coefficients of 0 L mol⁻¹ for the high mannose and complex type glycans. The
755 experimental B_{22} value for A2G2S2 glycan was not attainable by scaling down the
756 interactions drastically. Visual inspection and examination of RDFs suggested that
757 aggregation was reversible when a scaling factor of 0.85 or lower was used,
758 suggesting that B_{22} is a somewhat problematic choice of parameter for optimization of
759 non-bonded interactions (Figure 4). It is also noteworthy that the scaling factor
760 required for reversing the aggregation is dependent upon the size of the glycan studied
761 (Figure S6).

762

763 While solution properties of the glycans are important, their interactions with other
764 biomolecules are equally critical. The binding of N-glycans to lectins, carbohydrate
765 binding proteins that make specific interactions with terminal sugars of glycan
766 chains^{92,97,99,101–105,122}, are crucial in many biological phenomena. Thus, optimizing the
767 binding properties of these glycans with proteins is important for their applicability in
768 multi-protein complexes, and reproduction of experimental affinities of glycans with
769 proteins represents a useful way of optimizing the non-bonded parameters, as shown
770 here. Binding free energies obtained for a total of thirteen candidate lectin-glycan pairs
771 showed that our glycan models can reproduce the experimental binding affinities
772 without the need for drastic corrections in non-bonded interactions (Figure 5).

773

774 It should be noted that for many of the systems, the exact binding mode of the whole
775 glycan was often unresolved. The partially resolved sugars in available crystal
776 structures were thus used for aligning and constructing the whole glycan which can
777 result in multiple initial conformations. Replicates ensured that the initial structure bias
778 was reduced. Cluster analysis showed that these glycans could maintain the overall
779 binding pose during umbrella sampling simulations (Figure S10). A simple glycan such
780 as (GlcNAc)₅ could distinguish between a favourable as well random surfaces on a
781 UDA lectin (Figure S11). Unrestrained simulations of high affinity CVN+M9 maintained
782 a very similar binding pose when compared to ATM simulations, whereas in the case
783 of the low affinity PAL and M5 pair, the glycan was more dynamic and could drift from
784 the pocket in both ATM and CG representations, consistent with the weak binding of
785 the ligand (Figure S12, S13). All these observations support the applicability of the
786 newly derived CG N-glycan models for specifically quantifying energetics with given
787 protein-ligand pairs.

788

789 Out of 13 different lectin-glycan systems, every one of them overpredicted the binding
790 free energies calculated during unscaled simulations. A 0.95 scaling was enough to
791 reduce the gap between the predicted and experimental binding free energies by a
792 large factor. It was observed that charged complex glycans – which contain an explicit
793 charge as well as a higher number of S type beads – required a higher scaling of 0.9.
794 In the PAL lectin, which also required 0.9 scaling, the glycan interacts with the lectin
795 via a relatively high number of polar residues in the binding pocket compared to any
796 other lectins (Table S3). Thus, the scaling required appears to be in part glycan type
797 dependent, particularly when electrostatics and a greater number of small type beads
798 are involved.

799

800 The scaling approach for Martini was first proposed by Stark et al⁴⁹ due to the
801 imbalance of the non-bonded interactions in the Martini FF for protein-protein systems,
802 and has since been used for glycan-glycan⁴⁸ interactions as well. In studies of
803 dimerization of different receptors such as ErbB1 and EphA1, binding free energies
804 were again overestimated^{123–126}. Javanainen et al⁵⁰ predicted the binding free
805 energies of dimerization of TM domains of five candidate receptor tyrosine kinases
806 (RTKs) and suggested a relatively modest correction of 10% in the well depth (ϵ) to
807 achieve better agreement with FRET studies⁵⁰ compared to the 60% correction
808 suggested by Stark et al⁴⁹ where they compared their PMFs against the B_{22} coefficient.
809 In the present work, we also obtained data pointing towards the imbalance in the
810 Martini FF, but this was not drastic and was alleviated by scaling the non-bonded
811 interactions by a relatively small value.

812

813 **Conclusions**

814 In summary, we have extended the Martini CG model parameters to N-glycans with
815 various branching patterns. An elastic network was found to be advantageous in
816 maintaining the conformations of branched glycans. The spurious self-aggregation of
817 glycans could be alleviated by scaling the non-bonded interaction, and when working
818 with glycans in solution, we recommend a scaling factor of 0.85. On the other hand,
819 when protein-binding is involved, free energy calculations with a wide variety of lectins
820 revealed that only modest scaling was needed to achieve experimental ΔG values
821 from SPR or ITC experiments. Thus, in initial studies of novel carbohydrates, we would
822 in general recommend that the N-glycan parameters developed herein should be
823 implemented with a non-bonded scaling factor of 0.9 for charged and/or highly polar
824 complex type glycans whereas 0.95 is sufficient for the simpler high mannose type
825 glycans. The parameters presented here should be useful for others interested in
826 studying the role of glycans in the dynamics of various large glycoproteins and
827 glycoprotein complexes which would benefit from a CG representation. Although the
828 open beta version of Martini 3 has been released for phospholipid bilayers and
829 proteins⁵¹, the bonded parameters and mapping schemes outlined in this study should
830 be sufficiently robust for the future optimization of new compatible N-glycan
831 parameters.

832

833 **Acknowledgments**

834 ATS acknowledges NUS Research Scholarship funded by Ministry of Education (AcRF
835 grant R-154-100-580-112). CSV, PJB, and PM acknowledge funding by the Ministry of
836 Education of Singapore (MOE2012-T3-1-008). PJB and JKM further acknowledge
837 funding from the National Research Foundation (NRF2017NRF-CRP001-027). We

838 wish to thank the Singapore National Supercomputing Centre (<https://www.nscg.sg>)
839 for computational resources.

840

841

842 **Figures & Tables**

843

844 **Table 1:** Extra angles and elastic network parameters for N-glycans shown in Figure
 845 2. Each bead is defined by its name as shown in Figure 2; in some cases, a
 846 superscript is used when there are two or more similar types of connections. i.e. G2a
 847 is a G2 bead belonging to the 'a' typed sugar in Figure 2. For angles greater than
 848 140°, the restricted bending potential (ReB) was used.

849

850

Glycan	Elastic Bonds	Rbond (nm)	Kbond (kJ mol ⁻¹ nm ⁻²)	Angles	Θ (°)	Kangle (kJ mol ⁻¹)
High-Mannose (M9) (A)	G2a-M2f	1.36	75	A2ASN-G2a-G2b	172	110
	G2a-M2h	1.03	75	G2a-G2b-M2c	160	40
	G2a-M2k	1.90	75	M2f-M2d-M2h	120	15
				M2f-M2c-M2k	78	10
				M2h-M2c-M2k	105	10
Biantennary complex glycan (FA2G2S2) (B)	M2c-GL2f	1.32	200	A2ASN-G2a-G2b	172	110
	M2c-GL2j	1.35	200	G2-G2-M2	160	40
				M3-G2-GL2	170	95
				G2-GL2-S1	112	30
				GL2-M2c-GL2	125	30
Tetraantennary complex glycan (FA4G4) (C)	M2a-GL2d	1.43	300	M1a-M2b-G2c	130	65
	M2a-GL2f	1.32	200	M2-G2-GL2	160	170
				M3-G2-GL2	178	70
Tetraantennary complex glycan (FA4G4) (D)	M2a-GL2d	1.43	300	M3a-M2b-G2c	138	120
	M2a-GL2f	1.32	200	M2-G2-GL2	162	170
				M3-G2-GL2	178	70
Tetraantennary complex glycan (FA4G4) (E)	M2a-GL2d	1.41	350	M1-G2-GL2	166	180
	M2a-GL2f	1.35	420	M3-G2-GL2	178	50

851

852

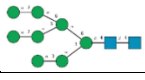
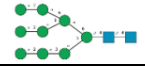

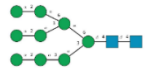
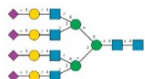

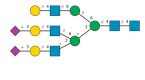
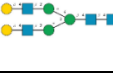
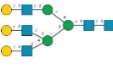


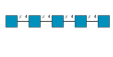

853

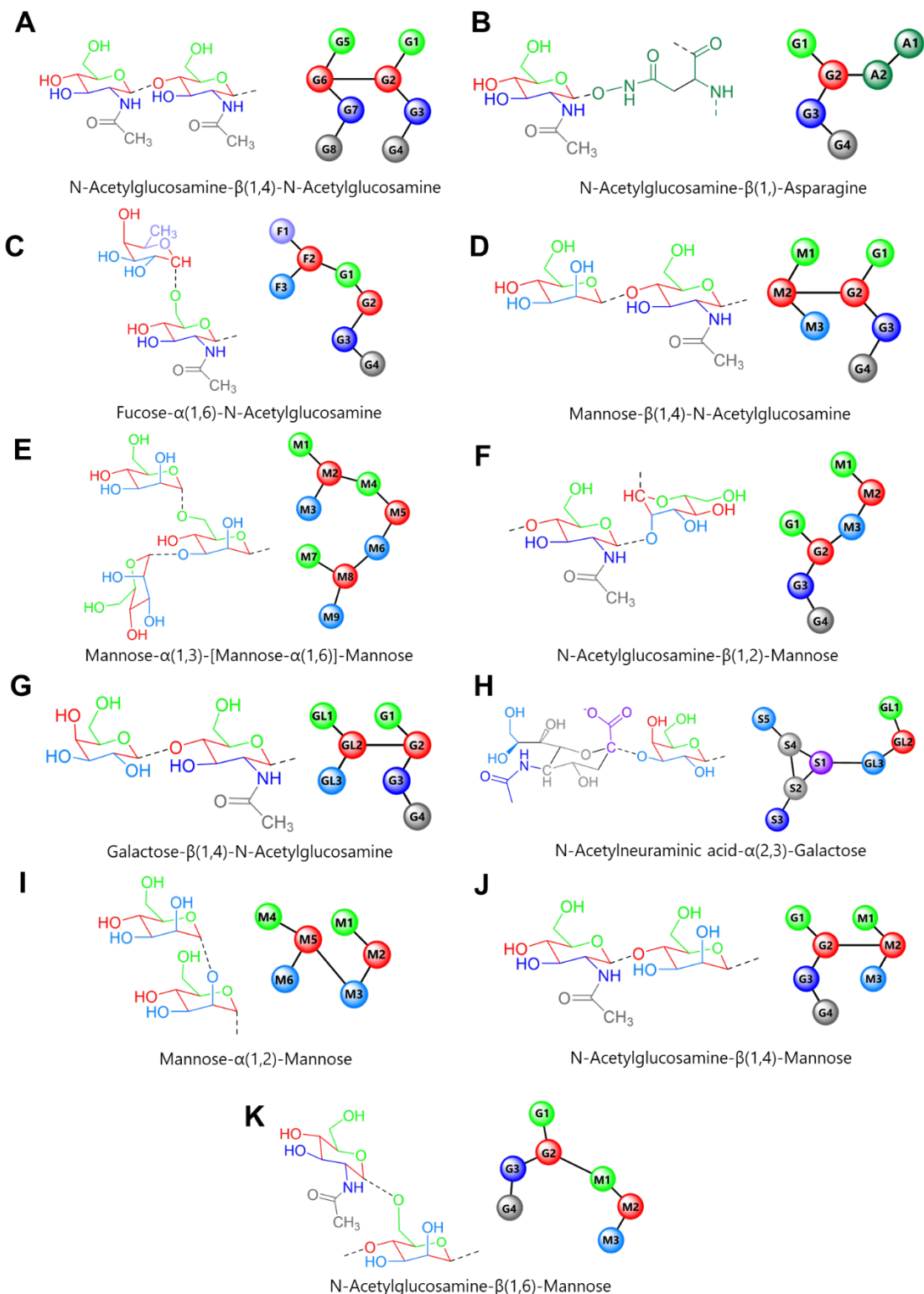
854 **Table 2:** Thermodynamic properties of CG n-glycan di/trisaccharides. Free energies
 855 of solvation and partition coefficients for various disaccharides compared to
 856 predictions methods ClogP and KOWWIN. The errors estimated for the solvation free
 857 energies in water (ΔG_W) and in octanol (ΔG_O) for obtaining the partitioning free
 858 energies ($\Delta\Delta G_{OW}$) were incorporated into the final reported partition coefficients
 859 ($\log P_{OW}$). These were compared against empirical predictions of $\log P_{OW}$ obtained
 860 from ClogP and KOWWIN⁵².
 861

Molecule	ΔG_W (kcal mol ⁻¹)	ΔG_O (kcal mol ⁻¹)	$\Delta\Delta G_{OW}$ (kcal mol ⁻¹)	$\log P_{OW}$	$\log P_{OW}$ CLogP	$\log P_{OW}$ KOWWIN
Fuc- α 16-GlcNAc	-34.9 \pm 0.2	-29.1 \pm 0.1	5.8 \pm 0.3	-4.1 \pm 0.3	-3.2	-4.0
GlcNAc- β 14-GlcNAc	-36.2 \pm 0.1	-31.0 \pm 0.2	5.2 \pm 0.3	-3.7 \pm 0.2	-4.1	-4.1
Man- β 14-GlcNAc	-32.8 \pm 0.2	-27.2 \pm 0.1	5.6 \pm 0.3	-4.0 \pm 0.2	-4.1	-4.5
Man- α 16-[Man- α 13-]Man	-41.7 \pm 0.2	-33.5 \pm 0.2	8.2 \pm 0.4	-5.8 \pm 0.3	-5.9	-6.5
Man- α 12-Man	-28.9 \pm 0.1	-23.5 \pm 0.1	5.4 \pm 0.2	-3.8 \pm 0.1	-4.0	-4.0
Gal- β 14-GlcNAc	-32.1 \pm 0.1	-26.8 \pm 0.2	5.3 \pm 0.3	-3.7 \pm 0.1	-4.1	-4.5
Neu5Ac- α 23-Gal	-35.4 \pm 0.2	-28.2 \pm 0.4	7.2 \pm 0.6	-5.1 \pm 0.4	-5.3	-6.0

862
 863

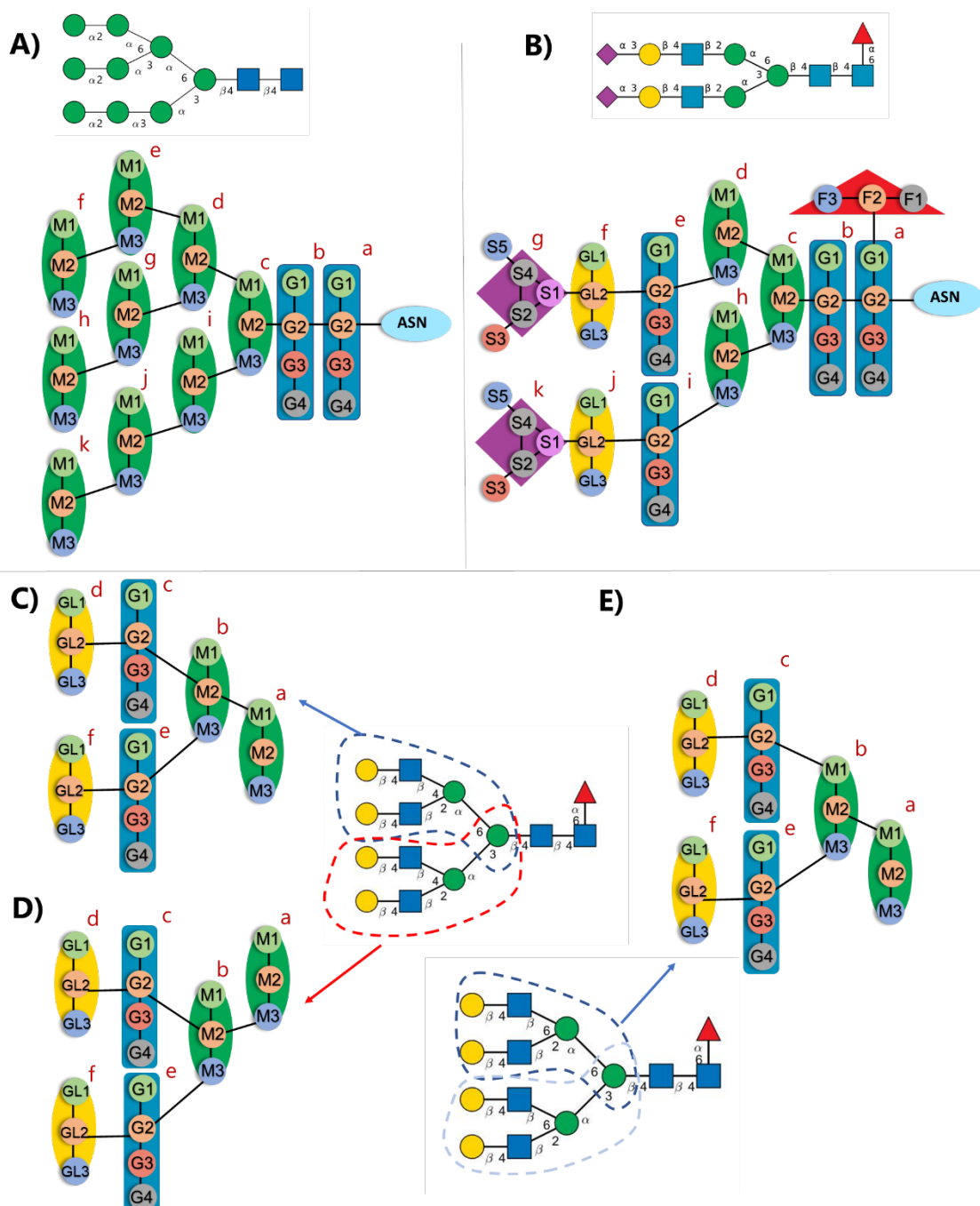
864 **Table 3:** Lectins used for calculating the binding free energies (ΔG^0_{sim}) of various
865 glycans to lectins. The binding affinities were obtained from either SPR or ITC
866 experiments (ΔG^0_{Expt}). Errors calculated from PMFs (ΔG_{PMF}) were obtained from 200
867 cycles of bootstrapping. The binding free energy ΔG^0_{sim} was calculated upon addition
868 to the ΔG_{PMF} of a correction term (ΔG_V) to convert to standard state volume, for
869 comparison with the experimental binding free energy ΔG^0_{Expt} . Monosaccharides
870 present in the glycans are represented by their symbolic representation, including
871 mannose (green circle), N-acetylglucosamine (blue square), galactose (yellow
872 circle), and Neu5Ac/sialic acid (purple diamond).
873

Lectin	Glycan	Scaling factor (λ)	ΔG_{PMF} (kcal/mol)	ΔG_V (kcal/mol)	ΔG^0_{sim} (kcal/mol)	ΔG^0_{Expt} (kcal/mol)
Cyanovirin-N ⁸⁵ (CVN) (PDB: 3GXZ)		1.0	-8.2 ± 0.7	-1.6	-9.8 ± 0.7	-8.7 ± 0.1 ⁸⁴
		0.95	-5.6 ± 0.4	-1.9	-7.5 ± 0.4	
		1.0	-9.3 ± 0.4	-1.6	-10.9 ± 0.4	-9.2 ± 0.3 ⁸⁴
		0.95	-6.7 ± 0.4	-1.9	-8.6 ± 0.4	
Pterocarpus angolensis ⁸⁸ (PAL) (PDB: 2PHW)		1.0	-8.2 ± 0.6	-1.1	-9.3 ± 0.6	-5.2 ⁸⁸
		0.95	-5.3 ± 0.8	-1.0	-6.3 ± 0.8	
		0.9	-3.4 ± 0.6	-1.5	-4.9 ± 0.6	
		1.0	-13.5 ± 1.1	-1.3	-14.8 ± 1.1	-5.8 ⁸⁸
		0.95	-8.6 ± 1.0	-1.4	-10.0 ± 1.0	
		0.9	-6.1 ± 1.2	-1.5	-7.6 ± 1.2	
Maackia Amurensis ¹⁰⁷ (MAA) (PDB: 1DBN)		1.0	-13.4 ± 1.7	-1.5	-14.9 ± 1.7	-5.5 ¹⁰⁶
		0.95	-8.8 ± 1.5	-1.6	-10.4 ± 1.5	
		0.9	-6.1 ± 0.9	-1.8	-7.9 ± 0.9	
		1.0	-11.6 ± 1.6	-1.3	-12.9 ± 1.6	-5.7 ¹⁰⁶
		0.95	-6.5 ± 0.9	-1.8	-8.3 ± 0.9	
		0.9	-5.4 ± 0.9	-1.7	-7.1 ± 0.9	
		1.0	-13.0 ± 1.6	-1.5	-14.5 ± 1.6	-4.7 ¹⁰⁶
		0.95	-8.6 ± 0.7	-1.7	-10.3 ± 0.7	
		0.9	-5.3 ± 0.8	-1.9	-7.2 ± 0.8	
Ricin communis ¹²⁷ (RCA) (PDB: 1RZO)		1.0	-12.1 ± 1.1	-1.5	-13.6 ± 1.1	-7.7 ± 0.1 ⁹¹
		0.95	-9.3 ± 0.8	-1.2	-10.5 ± 0.8	
		0.9	-4.1 ± 0.7	-2.0	-6.1 ± 0.7	
		1.0	-14.6 ± 0.3	-1.3	-15.9 ± 0.3	-7.3 ⁹¹
		0.95	-8.7 ± 0.9	-1.5	-10.2 ± 0.9	
		0.9	-4.3 ± 0.7	-2.0	-6.3 ± 0.7	
		1.0	-12.7 ± 0.2	-1.0	-13.7 ± 0.2	-7.0 ⁹¹
		0.95	-8.0 ± 0.3	-1.2	-9.2 ± 0.3	
		0.9	-4.9 ± 0.3	-1.4	-6.3 ± 0.3	
Concanavalin A ⁸⁰ (CONA) (PDB: 1CVN)		1.0	-8.4 ± 0.2	-1.1	-9.5 ± 0.2	-8.4 ± 0.1 ⁸¹
		0.95	-5.9 ± 0.9	-1.3	-7.2 ± 0.9	
Wheat Germ Agglutinin ⁹⁵ (WGA) (PDB: 2UVO)		1.0	-6.7 ± 0.5	-1.8	-8.5 ± 0.5	-5.8 ⁸¹
		0.95	-3.4 ± 0.5	-2.2	-5.6 ± 0.5	
Urtica Dioica Agglutinin ¹⁰¹ (UDA) (PDB: 1EHH)		1.0	-6.1 ± 0.8	-1.3	-7.4 ± 0.8	-5.9 ⁸¹
		0.95	-4.0 ± 0.4	-1.5	-5.5 ± 0.4	



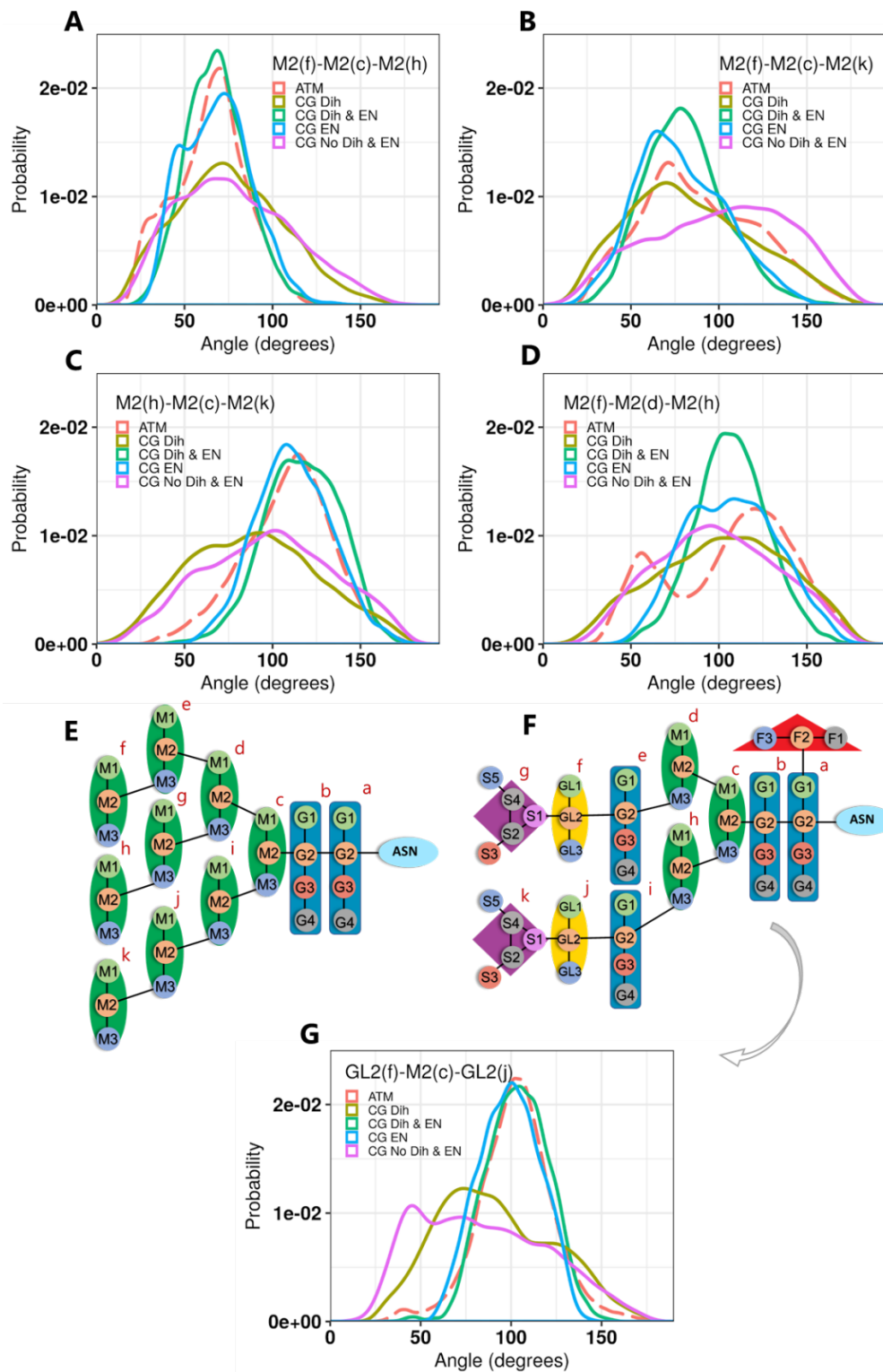
875
876
877
878
879
880
881
882

Figure 1: Disaccharide/trisaccharides used for developing parameters for N-glycans. Each image shows the atomistic representation of the saccharide (left) with mapped martini representation (right). The atoms which are mapped together are shown with the same colour as the beads in the CG model. The parameters for these di/trisaccharides are summarised in Table S1.



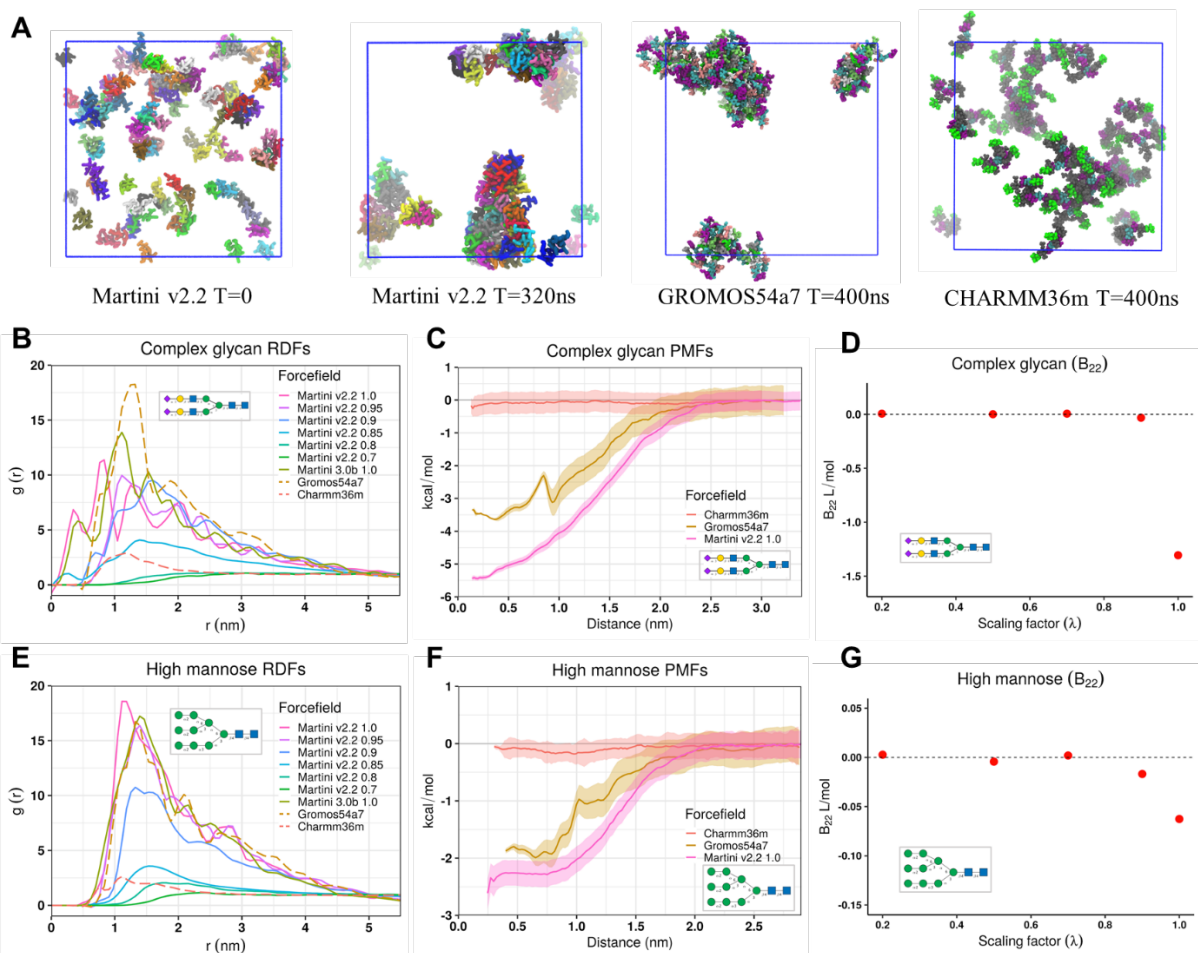
883
 884
 885
 886
 887
 888
 889
 890
 891
 892
 893
 894
 895

Figure 2: Parametrization of N-glycans constructed from disaccharides: (A) high mannose (M9) glycan; (B) sialylated bi-antennary (FA2G2S2) complex glycan; and (C) – (E) parts of tetra-antennary complex glycans parameterized separately for various linkages shown with dashed lines. All bonded parameters required to maintain the conformation of the glycans are summarised in the Table 1. Monosaccharides present in the glycans are represented by their symbolic representation: mannose (green circle), N-acetylglucosamine (blue square), galactose (yellow circle), and Neu5Ac/sialic acid (purple diamond).

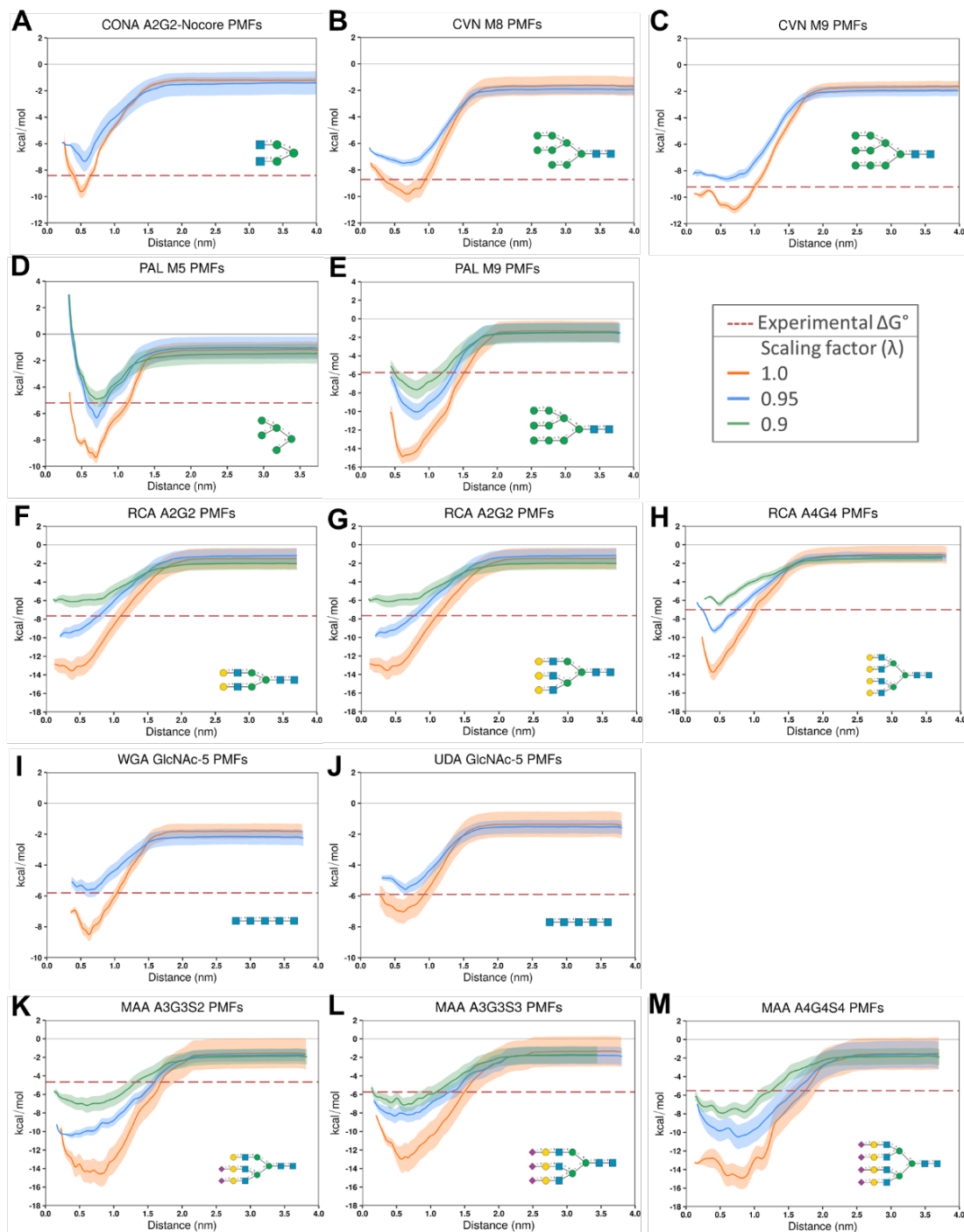


896
897
898
899
900
901
902
903
904

Figure 3: Branch angle distributions for M9 and FA2G2S2 type glycans. Distributions compare data from ATM simulations versus those for CG with dihedrals (CG Dih), CG with dihedrals and elastic network (CG Dih + EN), CG with elastic network only (CG EN), or CG with neither dihedrals or elastic network (CG No Dih & EN). The angles in plots A, B, C and D are for M9 as illustrated in (E). The angle in plot G is for FA2G2S2 complex type glycan as illustrated in (F).



905
 906 **Figure 4: Aggregation propensity for complex and high mannose glycans using**
 907 **a range of FFs.** (A) Initial and aggregated stages of glycan simulations for Martini
 908 v2.2, GROMOS54a7 and CHARMM36m forcefield. (B), (E) Radial distribution
 909 functions (RDFs) of glycans with various FFs. (C), (F) Potential mean of forces (PMFs)
 910 for glycans with various FFs. (D), (G) Partial virial coefficients (B_{22}) for Martini v2.2.
 911 Monosaccharides present in the glycans are represented by their symbolic
 912 representation: mannose (green circle), N-acetylglucosamine (blue square), galactose
 913 (yellow circle), and Neu5Ac/sialic acid (purple diamond).



914
 915
 916
 917
 918
 919
 920
 921
 922
 923
 924
 925
 926

Figure 5: PMFs for different lectin-glycan systems. The glycan used for binding studies in each case is given in the right bottom corner of each plot. Error estimates are shown with the shaded region obtained from 200 cycles of bootstrapping. (A) Concanavalin A (CONA); B), (C) Cyanovirin-N (CVN); (D), (E) Pterocarpus Anolensis (PAL); (F), (G), (H) Ricinus Communis Agglutinin (RCA); (I) Wheat Germ Agglutinin (WGA); (J) Urtica Dioicia Agglutinin (UDA); and (K), (L), (M) Maackia Amurensis (MAA). Monosaccharides present in the glycans are represented by their symbolic representation: mannose (green circle), N-acetylglucosamine (blue square), galactose (yellow circle), and Neu5Ac/sialic acid (purple diamond). The volume corrections ΔG_V were added to the total PMF in order to visually compare with the experimental data (dashed lines).